

Finite-Elemente-Verfahren und schnelle Löser

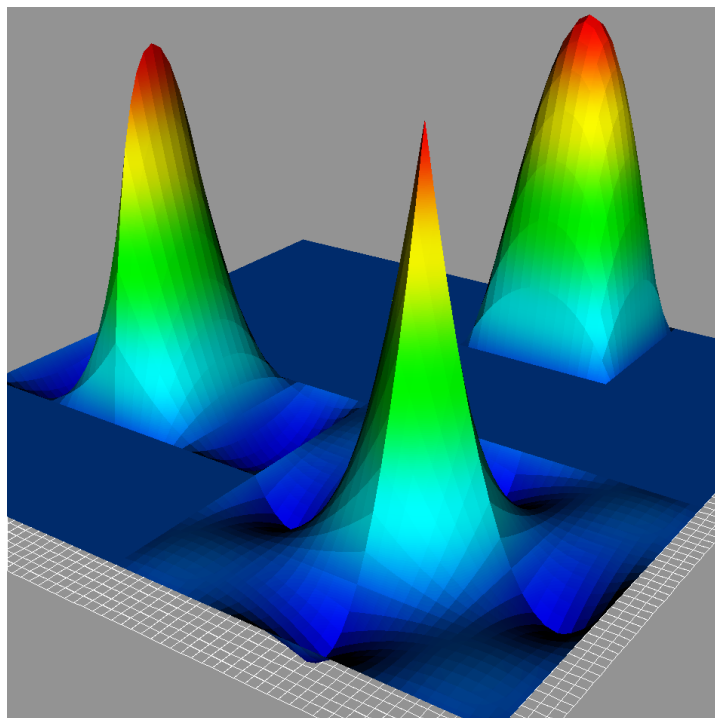
PETER BASTIAN

Universität Stuttgart, Institut für Parallele und Verteilte Systeme

Universitätsstraße 38, D-70569 Stuttgart

email: Peter.Bastian@ipvs.uni-stuttgart.de

1. April 2009



Inhaltsverzeichnis

1	Einführung in die Problemstellung	7
1.1	Grundlegende Begriffe	7
1.2	Modellierung der Wärmeströmung	7
1.3	Typeinteilung	10
1.4	Elliptische Gleichungen	11
1.5	Grundwasserströmung	12
2	Variationsformulierung	15
2.1	Aufgabenstellung	15
2.2	Charakterisierungssatz	16
2.3	Darstellung der Randwertaufgabe als Variationsproblem	18
2.4	Dirichlet'sches Prinzip	20
3	Sobolev-Räume	23
3.1	Der Raum $L_2(\Omega)$	23
3.2	Der Raum $H^m(\Omega)$	25
3.3	Poincaré-Friedrichsche Ungleichung	27
4	Lösbarkeit des Variationsproblems	31
4.1	Der Satz von Lax-Milgram	31
4.2	Anwendung auf das Dirichletproblem	32
4.3	Anwendung auf die Neumann'sche Randwertaufgabe	35
5	Ritz-Galerkin Verfahren	37
5.1	Die Idee	37
5.2	Eigenschaften der diskreten Lösung	38
5.3	Finite Elemente in einer Raumdimension	40
6	Gebräuchliche Finite Elemente	43
6.1	Eigenschaften der Zerlegung	43
6.2	Konforme Finite-Elemente-Räume	46
6.3	Ein Beispiel in zwei Raumdimensionen	52
6.4	Allgemeiner Aufbau des linearen Gleichungssystems	56
7	Approximationssätze	59
7.1	Bramble-Hilbert Lemma	59
7.2	Approximationssatz	61
7.3	Transformationssatz für allgemeine Dreiecke	63
8	Fehlerabschätzungen	71
8.1	Regularitätssätze	71
8.2	Fehlerabschätzung in der Energienorm	72
8.3	Fehlerabschätzung in der L_2 -Norm	73

9 Adaptive Gittersteuerung	75
9.1 Einführung	75
9.2 Duale Fehlerschätzung	75
9.3 Energienormfehlerschätzer	76
9.4 Verfeinerungsstrategie	78
10 Mehrgitterverfahren	81
10.1 Spektralverhalten einfacher Iterationsverfahren	81
10.2 Gitterhierarchie	82
10.3 Zweigitterverfahren	83
10.4 Mehrgitterverfahren	85
11 Konvergenz des Mehrgitterverfahrens	87
11.1 Vorbereitung	87
11.2 Analyse der Grobgitterkorrektur	89
11.3 Glättungseigenschaft	91
11.4 Zweigitterkonvergenz	91
11.5 Mehrgitterkonvergenz	92
11.6 Komplexität	93
Literatur	97

Vorwort

Dieses Skript basiert auf einer Ausarbeitung aus dem Sommersemester 2007. Es folgt in weiten Teilen dem Buch von Braess [Bra91]. Für die Erfassung des Textes in L^AT_EX danke ich Frau Sumeyra Abidin recht herzlich. Herrn Alexander Lauser recht herzlichen Dank für die Überlassung des Beispiels mit der lokal verfeinerten einspringenden Ecke.

Alle verbleibenden Fehler (und das sind im Moment noch so einige) gehen natürlich auf mein Konto.

Stuttgart, im April 2008

Peter Bastian

1 Einführung in die Problemstellung

1.1 Grundlegende Begriffe

Finite-Elemente-Methoden dienen der numerischen Lösung partieller Differentialgleichungen (PDGL) und wurden vor allem im Bereich der Strukturmechanik seit den späten 1950er Jahren entwickelt [TCMT56, Arg57]. Erste mathematische Beiträge gehen sogar auf das Jahr 1943 zurück [Cou43]. In dieser Vorlesung werden wir uns ausschließlich auf sogenannte elliptische PDGL beschränken.

Partielle Differentialgleichungen legen eine Funktion in mehreren Variablen durch Bedingungen an die Ableitungen fest. Partielle Differentialgleichungen treten vor allem im Rahmen kontinuumsmechanischer Modellierung auf. In der Kontinuumsmechanik abstrahiert man von der molekularen oder mikroskopischen Struktur eines Materials und weist jedem mathematischen Punkt $x = (x_1, \dots, x_n)$ des Körpers eine Eigenschaft zu. Als Beispiele seien genannt

$$\text{Temperatur } T(x), \quad \text{Druck } p(x), \quad \text{Geschwindigkeit } u(x), \dots$$

Das interessierende Material ist dabei üblicherweise räumlich begrenzt, d.h. wir können den Körper geometrisch als $\Omega \subseteq \mathbb{R}^n$ betrachten. Genauer definieren wir:

Definition 1.1. Ein Gebiet Ω ist eine offene und zusammenhängende Teilmenge des \mathbb{R}^n .

Die gesuchten Lösungsfunktionen haben also die Signatur $T : \Omega \rightarrow \mathbb{R}$. Oft setzt man noch weitere Differenzierbarkeitseigenschaften voraus, etwa $T \in C^k(\Omega)$ wobei $C^k(\Omega)$ die Menge der k -mal stetig (partiell-) differenzierbaren Funktionen bezeichnet.

1.2 Modellierung der Wärmeströmung

Um die Konzepte klar zu machen, betrachten wir als konkretes Beispiel die Modellierung des Wärmetransportes in einem Körper oder Fluid. Die hier gezeigte Technik lässt sich auf viele weitere Anwendungen verallgemeinern! Eine ausführliche Darstellung findet man etwa in [Fey70, p. 2-8, p. 3-4].

Sei $\Omega \subset \mathbb{R}^3$ ein Gebiet und $\Sigma = (a, b]$ ein Zeitintervall (die spezielle Wahl der Grenzen wird später klar und ist jetzt unwichtig). Gesucht ist die Temperatur $T(x, t)$ für jeden Punkt $(x, t) \in \Omega \times \Sigma$ des Körpers (also in Raum und Zeit).

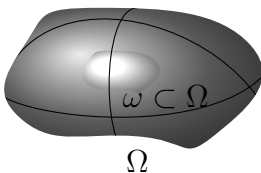
Zusätzlich sei die Temperatur am Anfang:

$$T(x, a) = T_a(x)$$

(Angangswert) und am Rand

$$T(x, t) = g(x, t), \quad x \in \partial\Omega$$

(Randwert) vorgegeben.



Ω

Energie

Wir betrachten jetzt ein beliebiges Teilgebiet $\omega \subseteq \Omega$ zum Zeitpunkt $t \in \Sigma$. In ω befindet sich zur Zeit t eine Menge an Wärmeenergie $Q_\omega(t)$. Bei gegebenem $T(x, t)$ berechnet sich diese als:

$$Q_\omega(t) = \int_{\omega} c(x)\rho(x)T(x, t)dx \quad (1.1)$$

Dabei ist

- $T(x, t)$: Temperatur in Grad Kelvin $[K]$,
- $c(x)$: Spezifische Wärmekapazität in $\frac{J}{K \cdot kg}$,
- $\rho(x)$: Massendichte in $\frac{kg}{m^3}$.

Somit hat $Q_\omega(t)$ die Einheit Joule $[J]$.

Energieerhaltung

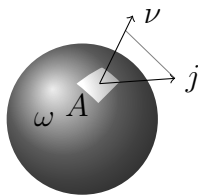
Nun betrachten wir die zeitliche Änderung von $Q_\omega(t)$ im beliebigen Zeitintervall $[t, t + \Delta t]$. Das Prinzip der Energieerhaltung besagt:

$$\underbrace{Q_\omega(t + \Delta t) - Q_\omega(t)}_{\text{Änderung der Wärmeenergie in } \omega \text{ in } [t, t + \Delta t]} = \{ \text{Energieeinspeisung/Verluste im Gebiet } \omega \} + \{ \text{Wärmefluss über den Rand } \partial\omega \}$$

In Formeln kann man dies ausdrücken als:

$$\int_{\omega} c(x)\rho(x)T(x, t + \Delta t)dx - \int_{\omega} c\rho(x)T(x, t)dx = \int_t^{t+\Delta t} \int_{\omega} f(x, t)dx dt - \int_t^{t+\Delta t} \int_{\partial\omega} q(x, t) \cdot \nu(x)ds dt.$$

Hierbei sind



- $f(x, t)$ Quell/Senkenterm in $[\frac{J}{s m^3}]$,
- $q(x, t)$ gerichteter Wärmefluss in $[\frac{J}{s m^2}]$ und
- $\nu(x)$ die nach außen gerichtete Einheitsnormale in $x \in \partial\omega$.

Nun linearisieren wir das Zeitintegral mittels

$$\int_t^{t+\Delta t} r(t)dt = \Delta t \cdot r(t) + O(\Delta t^2)$$

und erhalten damit nach Teilen durch Δt :

$$\int_{\omega} \underbrace{\frac{c(x)\rho(x)T(x, t + \Delta t) - c(x)\rho(x)T(x, t)}{\Delta t}}_{=\frac{\partial(c\rho T)}{\partial t}(x, t) \text{ für } \Delta t \rightarrow 0} dx = \int_{\omega} f(x, t) dx - \underbrace{\int_{\partial\omega} q(x, t) \cdot \nu(x) ds}_{=\int_{\omega} \nabla \cdot q(x, t) dx} + O(\Delta t).$$

Nun bildet man den Grenzwert $\Delta t \rightarrow 0$ und wendet rechts den Gaußschen Integralsatz an und erhält:

$$\int_{\omega} \left[\frac{\partial(c\rho T)}{\partial t}(x, t) + \nabla \cdot q(x, t) - f(x, t) \right] dx = 0$$

Da $\omega \subseteq \Omega$ beliebig gewählt war, folgert man, dass schon der Integrand identisch verschwinden muss und erhält die partielle Differentialgleichung

$$\frac{\partial(c\rho T)}{\partial t}(x, t) + \nabla \cdot q(x, t) = f(x, t) \quad \forall x \in \Omega, t \in \Sigma \quad (1.2)$$

Dies ist die mathematische Formulierung der Energieerhaltung. In analoger Weise kann man dies auf andere Erhaltungsgrößen wie Masse oder Impuls übertragen.

Wärmefluss

In (1.2) fehlt nun noch eine Modellierung des Flusses $q(x, t)$. Dieser setzt sich aus zwei Anteilen zusammen: konduktiver und konvektiver Wärmefluss.

Konduktion Auf molekularer Ebene ist Wärmeenergie gleich Bewegungsenergie der Moleküle. Der Übergang von Bewegungsenergie auf benachbarte Atome/Moleküle heißt Konduktion. (Es gibt hier mehrere Prozesse, Stöße, freie Valenzelektronen ...).

Auf der Kontinuumsskala macht man die (heuristische) Annahme, die Wärme fließt in Richtung des größten Temperaturunterschieds. Wegen des zweiten Hauptsatzes der Thermodynamik natürlich in Richtung kleinerer Werte. Dieses Modell bezeichnet man als Fourier'sches Gesetz (1822):

$$q^c(x, t) = -\lambda \nabla T(x, t). \quad (1.3)$$

Dabei ist

- λ : Wärmeleitfähigkeit in $[\frac{J}{smK}] = [\frac{W}{mK}]$,
- $\nabla T(x, t) = \left(\frac{\partial T}{\partial x_1}(x, t), \dots, \frac{\partial T}{\partial x_n}(x, t) \right)^T$ der Gradient von T . Dieser steht senkrecht auf den Höhenlinien $C(c, t) = \{x \mid T(x, t) = c\}$ und zeigt in Richtung des größten Ausstiegs. Die Einheit ist $[\frac{K}{m}]$.

Der Materialparameter λ ist im allgemeinen eine Matrix (Tensor 2. Stufe), der die Wärmeleitfähigkeit richtungsabhängig beschreibt (z.B. Metalle). In einem geeigneten Koordinatensystem kann die Wärmeleitfähigkeit in jede Richtung durch eine Diagonalmatrix beschrieben werden. Im allgemeinen gilt also

$$\lambda(x) = R^T(x)D(x)R(x)$$

mit einer Diagonalmatrix D , $d_{ii} > 0$ und einer Rotationsmatrix R . Von einem isotropen Tensor spricht man dann, wenn $\lambda(x) = k(x)I$ (I : Einheitsmatrix). Ist $\lambda(x)$ vom Ort abhängig so spricht man von heterogener Wärmeleitfähigkeit, sonst heißt λ homogen.

Konvektion: In Fluiden wird Wärme auch mit der *Masse* mitbewegt. Dies bezeichnet man als konvektiven Wärmetransport:

$$q^t(x, t) = c(x)\rho(x)T(x, t)u(x, t). \quad (1.4)$$

Dabei ist $u(x, t)$ die Geschwindigkeit des Fluides an der Stelle (x, t) mit der Einheit $[\frac{m}{s}]$. Der konvektive Fluss hat wie der konduktive Fluss die Einheit $[\frac{J}{sm^2}]$.

Der Gesamtfluss ergibt sich durch Addition der beiden Einzelflüsse:

$$q(x, t) = q^c(x, t) + q^t(x, t). \quad (1.5)$$

Wärmeleitungsgleichung

Einsetzen von (1.5) in (1.2) liefert die Wärmeleitungsgleichung:

$$\frac{\partial(c\rho T)}{\partial t} + \nabla \cdot \{c\rho uT - \lambda \nabla T\} = f \quad \text{in } \Omega \times \Sigma. \quad (1.6)$$

Dies ist eine lineare partielle Differentialgleichung zweiter Ordnung, da höchstens zweite Ableitungen vorkommen und diese Ableitungen mit Termen multipliziert werden, die höchstens von x , nicht aber von T abhängen.

Zu dieser Gleichung benötigt man noch die Anfangsbedingung

$$T(x, a) = T_a(x) \quad (1.7)$$

und eine Vorgabe auf dem Rand $\partial\Omega$ des Berechnungsgebietes. Hier gibt es verschiedene Möglichkeiten:

$$T(x, t) = g(x, t) \quad \text{für } x \in \Gamma_D \subseteq \partial\Omega \quad (\text{Temperatur, Dirichlet RB}) \quad (1.8)$$

$$q(x, t) \cdot \nu(x) = Q(x, t) \quad \text{für } x \in \Gamma_F = \partial\Omega \setminus \Gamma_D \quad (\text{Flussvorgabe, Neumann RB}) \quad (1.9)$$

1.3 Typeinteilung

Partielle Differentialgleichungen erlauben keine einheitliche Theorie wie gewöhnliche Differentialgleichungen. Stattdessen kann man eine Klassifikation in Typen vornehmen. Innerhalb eines Typs ist dann eine einheitliche Theorie möglich.

Definition 1.2 (Typeinteilung). Gegeben sei die allgemeine lineare partielle Differentialgleichung zweiter Ordnung in zwei Raumdimensionen:

$$a(x_1, x_2) \frac{\partial^2 T}{\partial x_1^2} + 2b(x_1, x_2) \frac{\partial^2 T}{\partial x_1 \partial x_2} + c(x_1, x_2) \frac{\partial^2 T}{\partial x_2^2} + d(x_1, x_2) \frac{\partial T}{\partial x_1} + e(x_1, x_2) \frac{\partial T}{\partial x_2} + f(x_1, x_2) = 0. \quad (1.10)$$

(1.10) heisst elliptisch im Punkt $x = (x_1, x_2) \in \Omega$ falls $a(x)c(x) - b^2(x) > 0$, hyperbolisch im Punkt x falls $a(x)c(x) - b^2(x) < 0$ und parabolisch im Punkt x falls $a(x)c(x) - b^2(x) = 0$. Man spricht von einer elliptischen (hyperbolischen, parabolischen) Gleichung falls der Typ in jedem Punkt $x \in \Omega$ elliptisch (hyperbolisch, parabolisch) ist.

Verallgemeinerungen dieser Definition auf $n > 2$ Raumdimensionen sind möglich (und sinnvoll), allerdings ist die Klassifikation nicht vollständig (d. h. es gibt lineare PDGL in drei Raumdimensionen, die weder elliptisch, hyperbolisch noch parabolisch sind).

Je nach Typ benötigt man für eine Gleichung unterschiedliche Rand- und/oder Anfangswerte, eine andere Lösungstheorie und ander numerische Verfahren. Allerdings kann man ein Verfahren für elliptische Gleichungen in der Regel sehr einfach auf parabolische Gleichungen erweitern.

Die Wärmeleitungsgleichung ist vom parabolischem Typ.

1.4 Elliptische Gleichungen

Im *stationären* Zustand ist in der Wärmeleitungsgleichung $T(x, t) = T(x)$, also T nicht mehr von der Zeit abhängig. Ist zusätzlich der konvektive Wärmetransport vernachlässigbar, so reduziert sich (1.6) auf

$$-\nabla \cdot \{\lambda \nabla T\} = f \quad \text{in } \Omega, \quad (1.11a)$$

$$T = g \quad \text{auf } \Gamma_D \subseteq \partial\Omega, \quad (1.11b)$$

$$-\lambda \nabla T \cdot \nu = Q \quad \text{auf } \Gamma_F = \partial\Omega \setminus \Gamma_D. \quad (1.11c)$$

Dies ist eine *elliptische* PDGL mit ortsabhängigem Leitfähigkeitstensor $\lambda(x)$. Wir beschränken uns in dieser Vorlesung ausschließlich auf diesen Typ von Gleichung! Elliptische Gleichungen heißen auch *Randwertprobleme*, da in jedem Punkt $x \in \Omega$ eine Randwertvorgabe gemacht werden muss.

Ist zusätzlich noch $\lambda(x) = I$, so reduziert sich die Gleichung weiter auf die *Poisson-Gleichung*:

$$-\Delta T = f \quad \text{in } \Omega, \quad (1.12a)$$

$$T = g \quad \text{auf } \Gamma_D, \quad (1.12b)$$

$$-\frac{\partial T}{\partial \nu} = Q \quad \text{auf } \Gamma_F = \partial\Omega \setminus \Gamma_D \quad (1.12c)$$

wobei $\Delta = \nabla \cdot \nabla$ der Laplace-Operator ist. Ausgeschrieben lautet der Laplace-Operator

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \dots + \frac{\partial^2}{\partial x_n^2}.$$

Ist schließlich noch $f \equiv 0$, so erhält man die *Laplace-Gleichung*:

$$-\Delta T = 0 \quad \text{in } \Omega \quad (1.13a)$$

$$T = g \quad \text{auf } \Gamma_D \quad (1.13b)$$

$$-\frac{\partial T}{\partial \nu} = Q \quad \text{auf } \Gamma_F = \partial\Omega \setminus \Gamma_D \quad (1.13c)$$

Dirichlet- und Neumann-Problem Oben wurde der Gebietsrand in zwei Teile für Dirichlet und Neumann Randbedingungen partitioniert:

$$\partial\Omega = \Gamma_D \cup \Gamma_F \quad \text{und} \quad \Gamma_D \cap \Gamma_F = \emptyset.$$

Gilt $\Gamma_D = \partial\Omega$ (und folglich $\Gamma_F = \emptyset$) so spricht man von *Dirichlet-Problem*, gilt hingegen $\Gamma_F = \partial\Omega$ (und folglich $\Gamma_D = \emptyset$) so spricht man vom Neumann-Problem.

Beim reinen Neumann-Problem ist die Lösung nur bis auf eine Konstante festgelegt, da mit $T(x)$ auch $T'(x) = T(x) + c$ für jedes $c \in \mathbb{R}$ die Gleichung erfüllt. Zudem müssen die Randwertvorgabe und der Quell/Senkenterm die Kompatibilitätsbedingung

$$\int_{\Omega} f \, dx = - \int_{\Omega} \Delta T \, dx = - \int_{\partial\Omega} \nabla T \cdot \nu \, ds = \int_{\partial\Omega} Q \, ds \quad (1.14)$$

erfüllen.

1.5 Grundwasserströmung

Um die Allgemeinheit dieses Ansatzes zu zeigen, betrachten wir als ein weiteres Beispiel die Strömung in einem voll gesättigten porösen Medium (z.B Sandstein).

Das Wasser bewege sich mit der (vektoriellen) Geschwindigkeit $u(x, t)$. Dann entspricht (1.2) der Massenerhaltungsgleichung:

$$\frac{\partial(\Phi(x)\rho(x, t))}{\partial t} + \nabla \cdot \underbrace{\{\rho(x, t)u(x, t)\}}_{\text{Massenfluss}} = \underbrace{f(x, t)}_{\text{Quellen/Senken}}. \quad (1.15)$$

Dabei bedeutet

- $\Phi(x)$: Porosität des porösen Mediums $\in [0, 1]$. Die Porosität trägt keine Einheit.
- $\rho(x, t)$: Massendichte in $\frac{kg}{m^3}$,
- $u(x, t)$: Filtergeschwindigkeit in $\frac{m}{s}$.

Für die Filtergeschwindigkeit fand Darcy 1856 den Zusammenhang

$$u(x, t) = -\frac{K}{\mu}(\nabla p - \rho G) \quad (1.16)$$

mit den folgenden Größen

- $p(x, t)$: Druck in $[Pa]$ (Pascal). Es gilt $[Pa] = [\frac{N}{m^2}] = [\frac{kg}{ms^2}]$.
- $K(x)$: absolute Permeabilität (Leitfähigkeit) in $[m^2]$. Wie die Wärmeleitfähigkeit ist dies eine symmetrisch positiv definite Matrix.
- μ : Dynamische Viskosität der Flüssigkeit in $[Pa \cdot s]$.
- G : Vektor, der in Richtung der Gravitation zeigt und als Betrag die Erdbeschleunigung besitzt, also $G = (0, \dots, -9.81)$ mit der Einheit $[\frac{m}{s^2}]$.

Im *inkompressiblen* Fall ($\rho = \text{const}$) erhält man die elliptische Gleichung:

$$\nabla \cdot \{\rho u\} = f, \quad u(x, t) = -\frac{K}{\mu}(\nabla p - \rho G) \quad \text{in } \Omega, \quad (1.17a)$$

$$p = g \quad \text{auf } \Gamma_D, \quad (1.17b)$$

$$u \cdot \nu = U \quad \text{auf } \Gamma_F = \partial\Omega \setminus \Gamma_D. \quad (1.17c)$$

Obwohl $\nabla \cdot \{\rho^2 \frac{K}{\mu} G\} = 0$ darf man den Term nicht einfach streichen, da dann die Randbedingungen nicht mehr passen! Bei Grundwasserproblemen formuliert man oft in die „Piezometerhöhe“ $h(x) = \frac{p}{\rho \|G\|} + x_n$ um (dann sind auch die Randbedingungen entsprechend zu transformieren).

Geothermie Hiermit sind wir nun in der Lage ein Modell für eine Geothermieanlage zu formulieren. Dabei sind nun sowohl die Bewegung des Wassers als auch der Transport der Wärme zu berechnen. Es handelt sich also um ein System von zwei gekoppelten Gleichungen:

$$\nabla \cdot u = f, \quad u = -\frac{K}{\mu}(\nabla p - \rho_w(T(x, t))G) \quad \text{in } \Omega \times \Sigma, \quad (1.18)$$

$$\frac{\partial(c_e \rho_e T)}{\partial t} + \nabla \cdot q + g^- T = g^+, \quad q = c_w \rho_w(T(x, t))uT - \lambda \nabla T \quad \text{in } \Omega \times \Sigma. \quad (1.19)$$

Die obere Gleichung beschreibt die Bewegung des Wassers und die untere Gleichung den Transport der Wärme. Die Geschwindigkeit des Wassers u geht in den konvektiven Transport der Wärme ein. Berücksichtigt man die Temperaturabhängigkeit der Dichte ρ , so ist auch die obere mit der unteren Gleichung gekoppelt. Allerdings wurde hier diese Abhängigkeit nur im Auftriebsterm berücksichtigt (das nennt man Boussinesq-Approximation).

Die Gleichungen sind zu ergänzen um die Rand- und Anfangsbedingungen:

$$p = \varphi \quad \text{auf } \Gamma_D^W(t), \quad u \cdot \nu = U \quad \text{auf } \Gamma_F^W(t), \quad (1.20)$$

$$T = \psi \quad \text{auf } \Gamma_D^H(t), \quad q \cdot \nu = Q(T, x, t) \quad \text{auf } \Gamma_F^H(t), \quad T(x, a) = T_a(x). \quad (1.21)$$

Helmholtz-Term In Gleichung (1.19) taucht der zusätzliche Term $g^-T(x, t)$ in der Energieerhaltung auf, der folgendermaßen zustande kommt. Wird an einer Stelle Wasser abgesaugt (Senke) so verschwindet damit auch die Wärmeenergie dieses Wassers. Die mit dem Abfluss dieses Wassers verschwindende Energie hängt natürlich von dessen Temperatur ab und beträgt

$$g^-(x, t)T(x, t) = c(x)\rho(x)r(x, t)T(x, t) \quad (1.22)$$

wobei $r(x, t)$ die Stärke des Abflusses in der Einheit $[s^{-1}]$ beschreibt. Da dieser Term auch in der sog. Helmholtz-Gleichung auftaucht nennen wir ihn Helmholtz-Term.

2 Variationsformulierung

2.1 Aufgabenstellung

Wir betrachten in dieser Vorlesung die elliptische partielle Differentialgleichung

$$-\nabla \cdot \{A\nabla u\} + a_0 u = - \sum_{i,k=1}^n \partial_i(a_{ik}(x)\partial_k u) + a_0(x)u = f \quad \text{in } \Omega \subset \mathbb{R}^n \quad (2.1a)$$

$$u = g \quad \text{auf } \Gamma_D \subseteq \partial\Omega \quad (2.1b)$$

$$-(A\nabla u) \cdot \nu = q \quad \text{auf } \Gamma_F = \partial\Omega \setminus \Gamma_D \quad (2.1c)$$

mit der unbekanntem Funktion $u : \Omega \rightarrow \mathbb{R}$. Dabei haben wir $\partial_i = \frac{\partial}{\partial x_i}$ als Abkürzung verwendet. $A(x)$ für $x \in \Omega$ bezeichnet eine positive $n \times n$ -Matrix mit den Komponenten $a_{ik}(x)$ und $a_0 : \Omega \rightarrow \mathbb{R}$ bezeichnet eine nichtnegative Funktion.

Als *klassische Lösung* dieser Gleichung bezeichnet man Funktionen

$$u \in C^2(\Omega) \cap C^0(\overline{\Omega}) \quad \text{falls } \Gamma_D = \partial\Omega \text{ (reines Dirichlet Problem), bzw.}$$

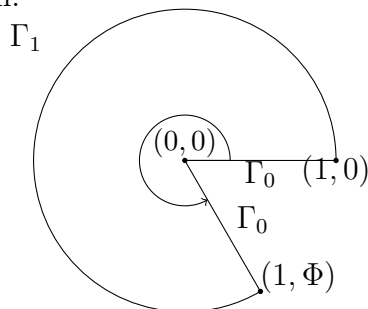
$$u \in C^2(\Omega) \cap C^1(\overline{\Omega}) \quad \text{falls } \Gamma_F = \partial\Omega \text{ (reines Neumann Problem),}$$

welche (2.1a) in jedem Punkt $x \in \Omega$ sowie die Randbedingung (2.1b) oder (2.1c) in jedem Punkt $x \in \partial\Omega$ identisch erfüllen.

Finite Differenzen Verfahren, siehe [Hac86, Bas08], zur numerischen Lösung basieren auf Taylorreihenentwicklung und erfordern die noch höhere *Regularität* $u \in C^4(\overline{\Omega})$, um Konvergenz mit einer ausreichenden Qualität zu garantieren.

Dass diese Regularität im allgemeinen nicht gegeben ist, zeigt:

Beispiel 2.1 (Einspringende Ecke). Betrachte das (parameterabhängige) Gebiet $\Omega_\Phi = \{(r, \varphi) \mid 0 < r < 1 \wedge 0 < \varphi < \Phi\}$ für $0 < \Phi \leq 2\pi$. Hierbei sind (r, φ) Polarkoordinaten.



Die Funktion $u(r, \varphi) = r^{\frac{\pi}{\Phi}} \cdot \sin(\varphi \frac{\pi}{\Phi})$ löst die Gleichung

$$\Delta u = 0 \quad \text{in } \Omega,$$

$$u = \sin(\varphi \frac{\pi}{\Phi}) \quad \text{auf } \partial\Omega.$$

Abbildung 1 weiter unten zeigt die Funktion u für $\Phi = \frac{3}{2}\pi$. Für $\pi < \Phi \leq 2\pi$ (nicht konvexes Gebiet) gilt $\frac{1}{2} \leq \frac{\pi}{\Phi} < 1$ und damit $\frac{\partial u}{\partial r}(0, \varphi) = \infty$, somit also $u \notin C^1(\overline{\Omega})$! Nichtsdestotrotz ist u eine klassische Lösung wie oben eingeführt.

Beweis: Transformation der Gleichung auf Polarkoordinaten und einsetzen, siehe Übungsaufgabe. \square

Das Beispiel zeigt ausserdem: Die Regularität der Lösung hängt von der Form des Gebietes Ω ab.

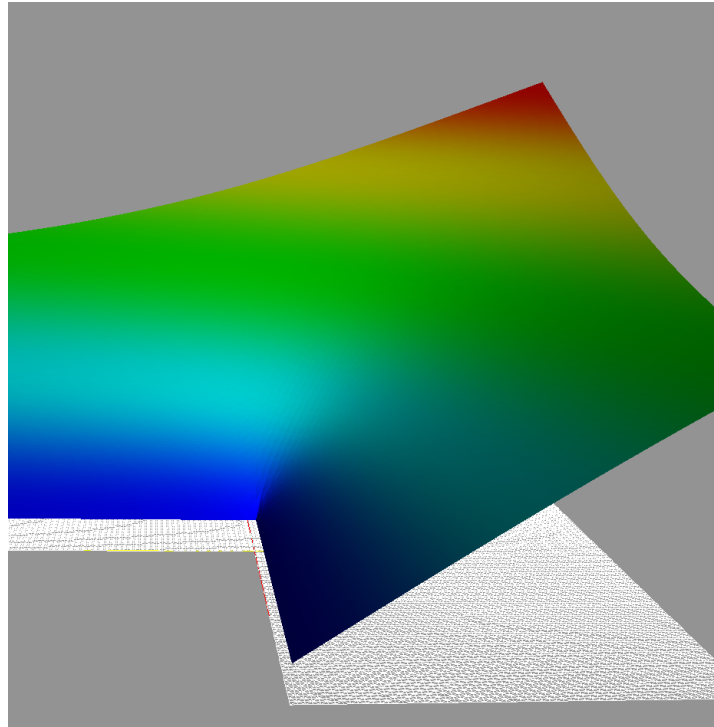


Abbildung 1: Die Singularitätenfunktion für $\Phi = \frac{3}{2}\pi$.

2.2 Charakterisierungssatz

Wir stellen zunächst einen allgemeinen, abstrakten Rahmen für die unten folgende Variationsformulierung des Randwertproblems her.

Sei V ganz allgemein ein linearer Raum über dem Körper \mathbb{R} (auch Vektorraum genannt. In unserer Anwendung sind das Funktionen, der Charakterisierungssatz gilt aber allgemein). In V ist also die Addition von Elementen sowie die Multiplikation mit einer reellen Zahl (Skalar) erklärt und die Menge ist abgeschlossen unter diesen Operationen.

Weiter führen wir eine sogenannte *Bilinearform*

$$a : V \times V \rightarrow \mathbb{R} \quad (2.2)$$

als stetige Funktion mit den Eigenschaften

$$a(u + v, w) = a(u, w) + a(v, w) \quad a(u, v + w) = a(u, v) + a(u, w) \quad u, v, w \in V \quad (2.3)$$

$$a(ku, v) = ka(u, v) \quad a(u, kv) = ka(u, v) \quad u, v \in V, k \in \mathbb{R}. \quad (2.4)$$

ein.

Gilt für a zusätzlich

$$a(u, u) > 0 \quad \text{für alle } u \in V, u \neq 0 \text{ (Positivität)}, \quad (2.5)$$

$$a(u, v) = a(v, u) \quad \text{für alle } u, v \in V \text{ (Symmetrie)}. \quad (2.6)$$

so nennt man die Bilinearform symmetrisch und positiv (definit).

Ein Beispiel für eine solche Bilinearform ist das Skalarprodukt in einem Vektorraum. Schließlich heißt eine stetige Funktion

$$l : V \rightarrow \mathbb{R}$$

lineares Funktional (oder Linearform) auf V , falls gilt

$$l(u, v) = l(u) + l(v) \quad \forall u, v \in V \quad \text{und} \quad l(ku) = kl(u) \quad \forall u \in V, k \in \mathbb{R}.$$

Man kann zeigen, dass die Menge aller linearen Funktionalen über V selbst wieder ein linearer Raum ist. Statt $l(v)$ verwendet man häufig auch die Schreibweise $\langle l, v \rangle$.

Damit sind wir in der Lage, den Charakterisierungssatz zu formulieren.

Satz 2.2 (Charakterisierungssatz). Gegeben sei eine symmetrische und positive Bilinearform $a : V \times V \rightarrow \mathbb{R}$ sowie eine Linearform $l : V \rightarrow \mathbb{R}$. Die Größe (Funktional)

$$J(v) := \frac{1}{2}a(v, v) - l(v) \tag{2.7}$$

nimmt in V ihr Minimum genau dann bei u an, wenn

$$a(u, v) = l(v) \text{ für alle } v \in V. \tag{2.8}$$

Außerdem gibt es höchstens eine Minimallösung.

Beweis: Gegeben seien $u, v \in V$ und $t \in \mathbb{R}$. Dann gilt

$$\begin{aligned} J(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - l(u + tv) \\ &= \frac{1}{2}[a(u, u) + 2ta(u, v) + t^2a(v, v)] - l(u) - tl(v) \\ &= \frac{1}{2}a(u, u) - l(u) + t[a(u, v) - l(v)] + \frac{1}{2}t^2a(v, v) \\ &= J(u) + t[a(u, v) - l(v)] + \frac{1}{2}t^2a(v, v). \end{aligned} \tag{2.9}$$

„ \Rightarrow “: $u \in V$ (Annahme!) sei Minimum des Funktionals J . Dann muss für jedes $v \neq 0$ die Ableitung der Funktion $h(t) = J(u + tv)$ bei $t = 0$ notwendig verschwinden. Wegen (2.9) gilt

$$\frac{dh}{dt} = \frac{d}{dt}[J(u) + t[a(u, v) - l(v)] + \frac{1}{2}t^2a(v, v)] = a(u, v) - l(v) + ta(v, v)$$

und damit ist zu fordern:

$$\left. \frac{dh}{dt} \right|_{t=0} = a(u, v) - l(v) \stackrel{!}{=} 0,$$

also (2.8). Es liegt ein Minimum vor, da

$$\left. \frac{d^2h}{dt^2} \right|_{t=0} = a(v, v) > 0 \quad \text{für } v \neq 0 \text{ (Positivität).}$$

„ \Leftarrow “: u erfülle (2.8). Dann gilt nach Einsetzen in (2.9)

$$J(u + tv) = J(u) + \frac{1}{2}t^2 a(v, v) > J(u) \quad \forall t \neq 0, v \neq 0,$$

also liegt bei u ein Minimum vor.

Die Eindeutigkeit ergibt sich durch Rückführung auf einen Widerspruch. Angenommen es liegt ein weiteres Minimum bei $u' \neq u$ vor. Mit $v = u' - u \neq 0$ gilt

$$J(u') = J(u' + u - u) = J(u + (u' - u)) > J(u),$$

und somit ist u' kein Minimum im Widerspruch zur Annahme. \square

Bemerkung 2.3. Für den Charakterisierungssatz ist nur die Vektorraumsstruktur erforderlich. Er gilt also in \mathbb{R}^N genauso wie in Funktionenräumen.

Bemerkung 2.4. Der Satz sagt **nicht**, dass es immer ein $u \in V$ gibt, welches $J(v)$ minimiert. Er sagt nur

1. u ist Minimum von $J(v) \iff a(u, v) = l(v) \forall v \in V$.
2. Wenn es ein Minimum gibt dann ist es eindeutig.

2.3 Darstellung der Randwertaufgabe als Variationsproblem

Wir werden nun das Randwertproblem (2.1a) in eine äquivalente Minimierungsaufgabe überführen.

Reduktion auf homogene Randbedingungen

Dies wird jedoch nur für homogene Dirichlet-Randbedingungen $g = 0$ gelingen. Dies ist jedoch keine Einschränkung, wie folgende Überlegung zeigt.

Es sei das Randwertproblem

$$\begin{aligned} -\nabla \cdot \{A\nabla u\} + a_0 u &= f && \text{in } \Omega \\ u &= g && \text{auf } \partial\Omega \end{aligned}$$

zu lösen.

Es sei ausserdem eine Funktion $u' \in C^2(\Omega) \cap C^0(\bar{\Omega})$ mit $u'|_{\partial\Omega} = g$ (g : Randwerte) bekannt (Ob dies möglich ist hängt von g ab).

Mit dem Ansatz $u = u' + w$ ergibt sich aufgrund der Linearität

$$\begin{aligned} -\nabla \cdot \{A\nabla(u' + w)\} + a_0(u' + w) &= f && \text{in } \Omega \\ \Leftrightarrow -\nabla \cdot \{A\nabla w\} + a_0 w &= f + \nabla \cdot \{A\nabla u'\} - a_0 u', \end{aligned}$$

also eine Gleichung der selben Form für w mit einer neuen rechten Seite $f' = f + \nabla \cdot \{A\nabla u'\} - a_0 u'$. Für die Randwerte von w gilt

$$u' + w = g \quad \Leftrightarrow \quad w = g - u' = 0 \quad \text{auf } \partial\Omega.$$

Variationsformulierung

Im folgenden benötigen wir die Green'sche Formel, im Prinzip die Verallgemeinerung der partiellen Integration in mehrere Raumdimensionen. Für beliebige Funktionen $v, w \in C^1(\Omega) \cap C^0(\bar{\Omega})$ gilt

$$\int_{\Omega} \partial_i w v dx = - \int_{\Omega} w \partial_i v dx + \int_{\partial\Omega} w v \nu_i ds \quad (2.10)$$

wobei ν_i die i -te Komponente der äußeren Einheitsnormale an das Gebiet Ω ist.

Es sei nun $v \in C^1(\Omega) \cap C^0(\bar{\Omega})$ eine sogenannte *Testfunktion* mit $v = 0$ auf $\partial\Omega$. Multiplikation von (2.1a) auf jeder Seite mit v und Integration über das Gebiet liefert:

$$\int_{\Omega} (Lu)v dx = \int_{\Omega} [-\nabla \cdot \{A\nabla u\} + a_0 u] v dx = \int_{\Omega} f v dx.$$

Anwendung der Green'schen Formel liefert dann

$$\begin{aligned} & \int_{\Omega} [-\nabla \cdot \{A\nabla u\} + a_0 u] v dx = \int_{\Omega} f v dx \\ \Leftrightarrow & \int_{\Omega} - \sum_{i=1}^n \partial_i \left(\sum_{k=1}^n a_{ik} \partial_k u \right) v + a_0 u v dx = \int_{\Omega} f v dx \\ \Leftrightarrow & - \sum_{i=1}^n \sum_{k=1}^n \int_{\Omega} \partial_i (a_{ik} \partial_k u) v dx + \int_{\Omega} a_0 u v dx = \int_{\Omega} f v dx \\ \Leftrightarrow & - \sum_{i=1}^n \sum_{k=1}^n \left\{ - \int_{\Omega} a_{ik} \partial_k u \partial_i v dx + \int_{\partial\Omega} a_{ik} \partial_k u v \nu_i ds \right\} + \int_{\Omega} a_0 u v dx = \int_{\Omega} f v dx \\ \Leftrightarrow & \int_{\Omega} \sum_{i=1}^n \sum_{k=1}^n a_{ik} \partial_k u \partial_i v + a_0 u v dx = \int_{\Omega} f v dx \\ \Leftrightarrow & \int_{\Omega} (A\nabla u) \cdot \nabla v + a_0 u v dx = \int_{\Omega} f v dx. \end{aligned} \quad (2.11)$$

Hier haben wir benutzt, dass $v = 0$ auf $\partial\Omega$.

Der Ausdruck

$$a(u, v) = \int_{\Omega} \sum_{i=1}^n \sum_{k=1}^n a_{ik} \partial_k u \partial_i v + a_0 u v dx \quad (2.12)$$

stellt eine symmetrische und positive Bilinearform dar. Die Symmetrie sieht man unmittelbar, die Positivität werden wir weiter unten zeigen.

Darüberhinaus ist die rechte Seite

$$l(v) = \int_{\Omega} f v dx \quad (2.13)$$

eine Linearform.

Mit diesem Wissen folgt dann der

Satz 2.5. Wir betrachten den Raum der Funktionen

$$V = \{v \in C^2(\Omega) \cap C^0(\bar{\Omega}) \mid v = 0 \text{ auf } \partial\Omega\}.$$

Es sei $u \in V$ Lösung des Randwertproblems

$$\begin{aligned} Lu = -\nabla \cdot \{A\nabla u\} + a_0 uv &= f && \text{in } \Omega, \\ u &= 0 && \text{auf } \partial\Omega. \end{aligned}$$

Dann ist u auch Lösung des Variationsproblems

$$J(v) = \frac{1}{2}a(v, v) - l(v) \rightarrow \min$$

unter allen Funktionen in V .

Beweis: Wegen (2.11) gilt:

$$a(u, v) - l(v) = \int_{\Omega} [-\nabla \cdot \{A\nabla u\} + a_0 u - f]v \, dx = 0$$

für alle $v \in V$. Da a symmetrisch und positiv ist der Charakterisierungssatz 2.2 anwendbar und es folgt, dass u Minimum des zugehörigen Variationsproblems ist. \square

Bemerkung 2.6. Es gilt auch die Umkehrung. Sei $u \in V$ Lösung des Variationsproblems, dann ist u eine klassische Lösung des Randwertproblems.

Beweis: u löst das Variationsproblem, also $a(u, v) - l(v) = 0 \, \forall v \in V$. Mit (2.11) folgt dann $\int_{\Omega} [Lu - f]v \, dx = 0$ und damit $Lu = f$ in Ω . \square

Bemerkung 2.7. Dirichlet-Randbedingung muss man bei der Variationsformulierung explizit in den Funktionenraum einbauen. Sie heißen daher auch essentielle Randbedingungen (engl.: essential boundary conditions). Unten werden wir sehen, dass Neumann-Randbedingungen sich in der Variationsformulierung ganz einfach behandeln lassen. Sie heißen daher auch natürliche Randbedingungen (engl.: natural boundary conditions). \square

2.4 Dirichlet'sches Prinzip

Am Minimum $u \in V$ gilt für das Funktional

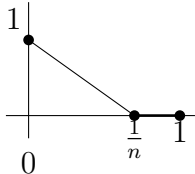
$$J(u) = \frac{1}{2}a(u, u) - l(u) = \frac{1}{2}l(u) - l(u) = -\frac{1}{2}l(u)$$

da ja $a(u, v) = l(v)$ insbesondere auch für u (aus V !) gilt.

Dirichlet argumentierte dann so: Da $J(v)$ nach unten beschränkt ist nimmt das Funktional sein Minimum für ein $u \in V$ an. V ist die Funktionenmenge über die minimiert wird. Dies ist aber im allgemeinen falsch wie folgendes Beispiel zeigt:

Beispiel 2.8. Es sei nun $J(v) = \int_0^1 v^2(t)dt$, also ein viel einfacheres Funktional. J sei zu minimieren über der Menge $V = \{v \in C^0[0, 1] \mid v(0) = 1 \wedge v(1) = 0\}$.

Wegen $J(v) \geq 0$ ist J offensichtlich nach unten durch 0 beschränkt. Wir betrachten die Folge

$$v_n(t) = \begin{cases} 1 - n \cdot t & t \leq \frac{1}{n} \\ 0 & t > \frac{1}{n} \end{cases}$$


Für diese Folge gilt

1. $v_n \in V$ für alle $n \in \mathbb{N}$
2. $J(v_n) < J(v)$ für alle $n > m$
3. $\lim_{n \rightarrow \infty} v_n \notin V$, denn es ist

$$v_\infty(x) = \lim_{n \rightarrow \infty} v_n = \begin{cases} 1 & x = 0 \\ 0 & \text{sonst} \end{cases}$$

Mit einer Folge $v_n \in V$ muss also nicht unbedingt auch der Grenzwert $\lim_{n \rightarrow \infty}$ in V liegen. Räume, für die dies doch der Fall ist, nennt man *vollständig*. So sind etwa die reellen Zahlen \mathbb{R} vollständig, die rationalen Zahlen \mathbb{Q} jedoch nicht. Vollständige Räume kann man dadurch erzeugen, dass man die Grenzwerte aller möglichen Folgen zu einem unvollständigen Raum hinzufügt. Diese Konstruktion nennt man *Vervollständigung*.

In obigem Beispiel ist das Problem, dass der Raum $C^0([0, 1])$ nicht vollständig ist bezüglich der Norm $\|u\| = \int_0^1 u^2(t)dt$. $C^0([0, 1])$ ist hingegen vollständig bezüglich der Norm $\|u\|_{C^0(\Omega)} = \sup_{x \in \Omega} |u(x)|$. Es kommt also auf die richtige Kombination von Funktionenraum und Norm an.

3 Sobolev-Räume

Das Dirichlet'sche Prinzips führte auf folgendes Problem:

Bei der Minimierung von

$$J(v) = \int_0^1 v^2(t) dt \quad \text{über} \quad V = \{v \in C^0[0, 1] \mid v(0) = 1 \wedge v(1) = 0\}$$

gibt es Minimalfolgen $(v_n)_{n \in \mathbb{N}}$, so dass

- $J(v_n) < J(v_m)$ für $n > m$ und
- $\lim_{n \rightarrow \infty} J(v_n) = J^*$, da $J(v) \geq 0$ (nach unten beschränkt).
- Aber $v^* := \lim_{n \rightarrow \infty} v_n \notin V$.

Dies liegt daran, dass der Funktionenraum V nicht „vollständig“ bezüglich einer mit dem Funktional kompatiblen Norm ist. Die Situation ist analog zu Folgen rationaler Zahlen, deren Grenzwerte in $\mathbb{R} \setminus \mathbb{Q}$ sein können.

Die Lösung des Problems liegt darin, geeignete Funktionenräume zu wählen, die bezüglich einer kompatiblen Norm vollständig sind.

Die für unsere Zwecke geeigneten sogenannten Sobolev-Räume werden wir in diesem Abschnitt einführen.

3.1 Der Raum $L_2(\Omega)$

Basis der Sobolev-Räume ist der Raum der quadratintegrierbaren Funktionen:

$$L_2(\Omega) = \left\{ v : \Omega \rightarrow \mathbb{R} \mid \int_{\Omega} |v(x)|^2 dx < \infty \right\}.$$

Die Integration wird hier im Sinne von Lebesgue (statt Riemann) verstanden (der Buchstabe L steht zu Ehren von Lebesgue). Dabei werden zwei Funktionen u, v identifiziert, falls sie sich nur auf einer Nullmenge (oder auch Menge vom Maß 0) $M \subset \Omega$ unterscheiden. Im \mathbb{R}^3 sind z.B. abzählbare Punktmengen, Linien und Flächen Nullmengen.

Mit dem Skalarprodukt

$$(u, v)_{L_2} = \int_{\Omega} u(x)v(x) dx \tag{3.1}$$

wird $L_2(\Omega)$ ein Hilbertraum mit der Norm

$$\|u\|_{L_2} = \sqrt{(u, u)_{L_2}}. \tag{3.2}$$

Hilberträume sind insbesondere vollständig. Die Vollständigkeit eines Raumes wird über Cauchy-Folgen definiert.

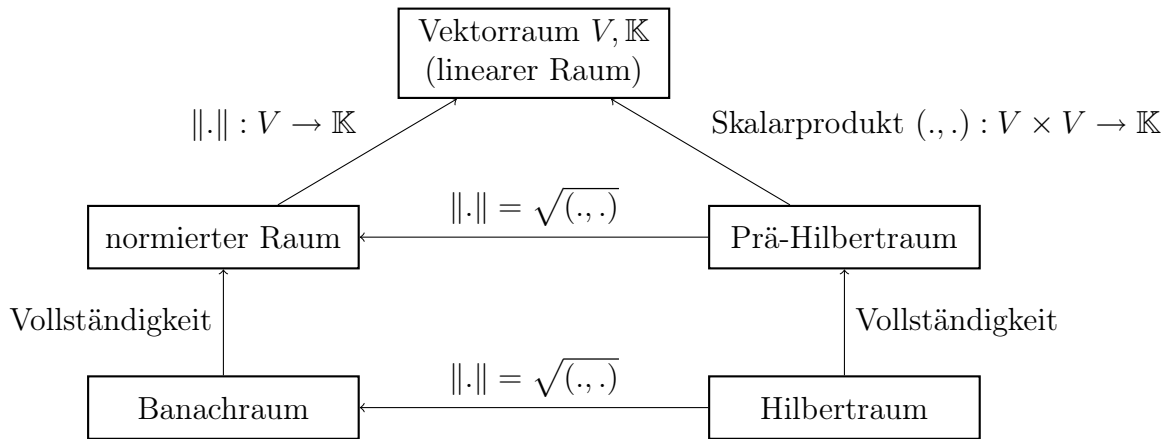


Abbildung 2: Zusammenhang der Definition verschiedener Räume in der Funktionalanalysis. Die Pfeile sind im Sinne einer „ist-ein“-Relation zu verstehen.

Sie V ein normierter Raum, d. h. ein Vektorraum mit der Norm $\|\cdot\|_V$. Dann heißt eine Folge $(v_i)_{i \in \mathbb{N}}$ Cauchy-Folge, genau dann wenn

$$\forall \epsilon > 0 : \exists N \in \mathbb{N} : \forall m, n \geq N : \|v_m - v_n\|_V < \epsilon$$

Die Besonderheit dieser Definition ist, dass der Grenzwert dieser Folge in der definition nicht auftaucht. Insbesondere kann es vorkommen, dass der Grenzwert zwar existiert, aber nicht in V ist.

Bei einem vollständigen Raum kann dies nicht passieren, denn man definiert: Der Raum V heißt vollständig bezüglich der Norm $\|\cdot\|_V$, falls jede Cauchy-Folge in V auch einen Grenzwert in V besitzt.

Die Abbildung 2 zeigt den Zusammenhang zwischen verschiedenen Eigenschaften von Räumen. Ausgangspunkt ist der Vektorraum über einem Körper \mathbb{K} , also einer Menge mit den Operationen $+$: $V \times V \rightarrow V$ und \cdot : $\mathbb{K} \times V \rightarrow V$, die gewisse Eigenschaften erfüllen. Ist auf dem Vektorraum V zusätzlich ein Skalarprodukt definiert, ist V auch ein Prä-Hilbertraum; ist auf V eine Norm definiert, so ist V ein normierter Raum. Da jedes Skalarprodukt in kanonischer Weise mittels $\|\cdot\| = \sqrt{(\cdot, \cdot)}$ eine Norm definiert, ist jeder Prä-Hilbertraum auch ein normierter Raum. Ist V normiert und vollständig, so ist V Banachraum. Hat man auf V ein Skalarprodukt und ist V vollständig, so ist V ein Hilbertraum.

Beispiel 3.1. Beispiele für Hilberträume sind

- \mathbb{R}^n mit dem Euklidischen Skalarprodukt $(x, y) = \sum_{i=1}^n x_i y_i$.
- $L_2(\Omega)$ mit dem Skalarprodukt $(u, v)_{L_2} = \int_{\Omega} u \cdot v \, dx$.

Hingegen sind die Funktionenräume $C^k(\Omega)$ keine Hilberträume (aber Banachräume). \square

Lemma 3.2. In jedem Vektorraum V mit Skalarprodukt $(\cdot, \cdot)_V$ und Norm $\|\cdot\|_V = \sqrt{(\cdot, \cdot)_V}$ gilt die Cauchy-Schwarzsche Ungleichung

$$|(x, y)_V| \leq \|x\|_V \|y\|_V.$$

3.2 Der Raum $H^m(\Omega)$

Die uns interessierenden Funktionenräume erfordern zusätzlich die Existenz partieller Ableitungen. Diese werden allerdings in spezieller Art und Weise definiert.

Definition 3.3 (Schwache Ableitung). Die Funktion $u \in L_2(\Omega)$ hat die schwache Ableitung $v = \partial^\alpha u \in L_2(\Omega)$, falls

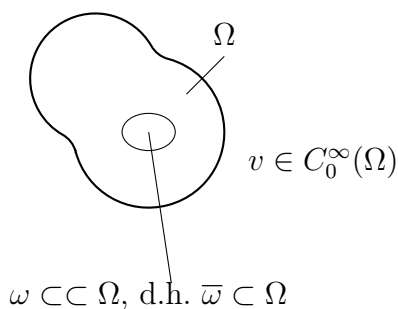
$$(\phi, v)_{L_2} = (-1)^{|\alpha|} (\partial^\alpha \phi, u)_{L_2} \quad \text{für alle } \phi \in C_0^\infty(\Omega). \quad (3.3)$$

Hierbei ist

- $\alpha = (\alpha_1, \dots, \alpha_n)$, $\alpha_i \in \mathbb{N}_0$ ein Multiindex und es werden folgende Abkürzungen eingeführt:

$$\partial^\alpha = \partial_1^{\alpha_1} \dots \partial_n^{\alpha_n} = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \dots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}, \quad |\alpha| = \alpha_1 + \dots + \alpha_n.$$

- $C^\infty(\Omega)$ ist die Menge der beliebig oft differenzierbaren Funktionen und
- $C_0^\infty(\Omega) \subset C^\infty(\Omega)$ ist der Unterraum von Funktionen, die nur auf einer kompakten Teilmenge von Ω von Null verschiedene Funktionswerte annehmen. Die Menge $\text{Tr}(u) = \{x \in \Omega \mid u(x) \neq 0\}$ heißt auch Träger von u und C_0^∞ die Menge der Funktionen mit kompaktem Träger. Im \mathbb{R}^n ist eine Teilmenge kompakt wenn sie abgeschlossen und beschränkt ist. Folgende Abbildung illustriert das Konzept einer kompakten Teilmenge.



□

Die Definition ist folgendermaßen motiviert. Für eine im klassischen Sinn differenzierbare Funktion u und eine Funktion $\phi \in C_0^\infty(\Omega)$ gilt (partielle Integration):

$$\int_{\Omega} \partial_1 u \phi \, dx = - \int_{\Omega} u \partial_1 \phi \, dx + \int_{\partial\Omega} u \phi \nu_1 \, ds$$

Der Randterm verschwindet, da $\phi = 0$ auf $\partial\Omega$ wegen dem kompakten Träger. Oben wird nun v die partielle Ableitung von u genannt, falls

$$\int_{\Omega} v\phi \, dx = - \int_{\Omega} u\partial_1\phi \, dx \quad \forall \phi \in C_0^\infty(\Omega).$$

Obige Definition ergibt sich dann durch mehrfache Anwendung.

Nimmt man (3.3) als Definition der Ableitung so spielt eine „Nichtdifferenzierbarkeit“ auf einer Nullmenge keine Rolle mehr. Z.B besitzt eine stückweise lineare Funktion eine schwache Ableitung. An den Knickstellen ist der Wert der Ableitung egal, da die Abänderung der Funktion auf dieser endlichen Menge von Punkten im Lebesgue-Integral keine Rolle spielt (siehe Abbildung 3).

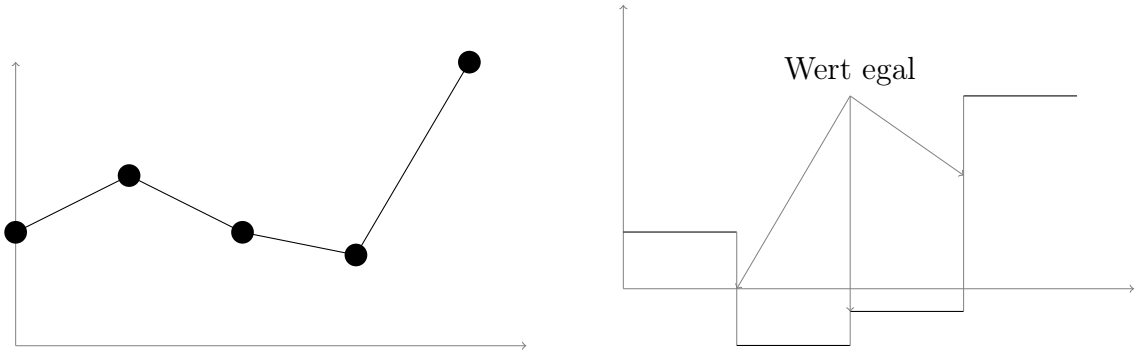


Abbildung 3: Schwache Ableitung einer stückweise linearen Funktion.

Nun sind wir in der Lage, die Sobolev-Räume zu definieren.

Definition 3.4 (Sobolev-Räume). Für ganzzahliges $m \geq 0$ bezeichnet

$$H^m(\Omega) = \{u \in L_2(\Omega) \mid \partial^\alpha u \in L_2(\Omega) \text{ existiert } \forall |\alpha| \leq m\}$$

den Sobolevraum der Ordnung m .

Auf $H^m(\Omega)$ definiert man weiter das Skalarprodukt

$$(u, v)_m := \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_{L_2(\Omega)} \quad (3.4)$$

und die zugehörige Norm lautet:

$$\|u\|_m := \sqrt{(u, u)_m} = \sqrt{\sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{L_2(\Omega)}^2}. \quad (3.5)$$

Neben der Norm betrachtet man auch die Größe

$$|u|_m := \sqrt{\sum_{|\alpha|=m} \|\partial^\alpha u\|_{L_2(\Omega)}^2}, \quad (3.6)$$

welche alle Eigenschaften einer Norm außer $|u| = 0 \Rightarrow u = 0$ erfüllt und daher Seminorm genannt wird.

$H^m(\Omega)$ ist ein Hilbertraum. Der Buchstabe H steht zu Ehren von David Hilbert. \square

Speziell für $m = 1$ lautet das Sobolev-Skalarprodukt

$$(u, v)_1 = \int_{\Omega} u \cdot v \, dx + \sum_{i=1}^n \int_{\Omega} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx.$$

Dies entspricht genau unserer Bilinearform $a(u, v)$ für $A = I$ und $a_0 = 1$.

Man kann alternativ die Sobolev-Räume auch über den Prozess der „Vervollständigung“ definieren.

Satz 3.5. Sei $\Omega \subset \mathbb{R}^n$ offen mit stückweise glattem Rand und es sei $m \geq 0$. Dann ist $C^\infty(\Omega) \cap H^m(\Omega)$ dicht in $H^m(\Omega)$. \square

Die Vervollständigung von $C^\infty(\Omega) \cap H^m(\Omega)$ bezüglich der Norm $\|\cdot\|_m$ gibt gerade wieder den $H^m(\Omega)$ bei beschränktem Ω .

Hat man Funktionen, die Nullrandbedingungen erfüllen, kann man folgende Vervollständigung betrachten.

Definition 3.6. Die Vervollständigung von $C_0^\infty(\Omega)$ bezüglich der Norm $\|\cdot\|_m$ ergibt den Raum $H_0^m(\Omega)$. \square

$H_0^m(\Omega)$ ist ein abgeschlossener Unterraum von $H^m(\Omega)$. Die Funktionen in $H_0^m(\Omega)$, $m > 0$ sind im verallgemeinerten Sinne Null auf $\partial\Omega$. Es ergibt sich das folgende Bild:

$$\begin{array}{ccccccc} L_2(\Omega) & = & H^0(\Omega) & \supset & H^1(\Omega) & \supset & H^2(\Omega) & \supset & \dots \\ & & \parallel & & \cup & & \cup & & \\ & & H_0^0(\Omega) & \supset & H_0^1(\Omega) & \supset & H_0^2(\Omega) & \supset & \dots \end{array}$$

In $H_0^m(\Omega)$ ist die Seminorm $|\cdot|_m$ sogar eine Norm, da aus $|u|_m = 0$ wegen der Nullrandbedingungen $u = 0$ folgt.

3.3 Poincaré-Friedrichsche Ungleichung

Zwei Normen, hier $|\cdot|_m$ und $\|\cdot\|_m$, heißen äquivalent, falls es Zahlen $a, b \in \mathbb{R}$; $a, b > 0$ gibt mit:

$$b\|u\|_m \leq |u|_m \leq a\|u\|_m \quad \forall u \in H_0^1(\Omega).$$

Wegen $\|u\|_m^2 = \sum_{|\alpha| \leq m} \|\partial^\alpha u\|_0^2$, ist die rechte Ungleichung mit $a = 1$ sofort erfüllt.

Im Spezialfall $m = 1$ wäre für die linke Ungleichung zu zeigen:

$$b^2 \underbrace{\left(\|u\|_0^2 + \sum_{i=1}^n \|\partial_1 u\|_0^2 \right)}_{\|u\|_1^2} \leq \underbrace{\sum_{i=1}^n \|\partial_1 u\|_0^2}_{=|u|_1^2}.$$

Insbesondere ist also $\|u\|_0$ durch $|u|_1$ abzuschätzen. Hierzu dient

Satz 3.7 (Poincaré-Friedrichsche Ungleichung). Sei Ω in einem n -dimensionalen Würfel der Kantenlänge s enthalten. Dann gilt

$$\|v\|_0 \leq s|v|_1 \text{ für alle } v \in H_0^1(\Omega). \quad (3.7)$$

Beweis: $C_0^\infty(\Omega)$ ist dicht in $H_0^1(\Omega)$, es genügt daher die Aussage für $v \in C_0^\infty(\Omega)$ zu zeigen. Nach Voraussetzung ist $\Omega \subset W = \{(x_1, \dots, x_n) \mid 0 < x_i < s\}$. Weiter kann $v = 0$ auf $W \setminus \Omega$ mit Null fortgesetzt werden.

Der Hauptsatz der Differential- und Integralrechnung sagt (etwas umgestellt)

$$\underbrace{v(x_1, x_2, \dots, x_n)}_x = \underbrace{v(0, x_2, \dots, x_n)}_{= 0 \text{ da } v = 0 \text{ außerhalb } \Omega} + \int_0^{x_1} \partial_1 v(t, x_2, \dots, x_n) dt.$$

Die Cauchy-Schwarzsche Ungleichung liefert:

$$|v(x)|^2 = \left(\int_0^{x_1} 1 \cdot \partial_1 v(t, x_2, \dots, x_n) dt \right)^2 \leq \underbrace{\int_0^{x_1} 1^2 dt}_{\|1\|_0^2} \cdot \underbrace{\int_0^{x_1} |\partial_1 v(t, x_2, \dots, x_n)|^2 dt}_{\|\partial_1 v\|_0^2} \leq s \cdot \underbrace{\int_0^x |\partial_1 v(t, x_2, \dots, x_n)|^2 dt}_{\text{unabhängig von } x_1!}$$

Schließlich ergibt sich unter Verwendung von diesem Zwischenresultat:

$$\begin{aligned} \|v\|_0^2 &= \int_W |v(x)|^2 dx = \int_0^s \dots \int_0^s \underbrace{|v(x)|^2}_{\text{von oben}} dx_1 \dots dx_n \\ &\leq \overbrace{\int_0^s \dots \int_0^s}^{n+1 \text{ Integrationen}} s \int_0^s |\partial_1 v(t, x_2, \dots, x_n)|^2 dt dx_1 \dots dx_n \\ &= \int_0^s \dots \int_0^s \int_0^s |\partial_1 v(t, s_2, \dots, s_n)|^2 dt \cdot s \cdot \underbrace{\int_0^s 1 dx_1 dx_2 \dots dx_n}_s \\ &= s^2 \int_0^s \dots \int_0^s |\partial_1 v(x_1, x_2, \dots, x_n)|^2 dx_1 dx_2 \dots dx_n \\ &= s^2 \int_W |\partial_1 v|^2 dx \leq s^2 |v|_1^2. \end{aligned}$$

□

Das Ganze lässt sich per Induktion auch auf $m > 1$ erweitern, d.h. auch $|\cdot|_m$ und $\|\cdot\|_m$ sind auf $H_0^m(\Omega)$ äquivalent.

Bemerkung 3.8. Es genügt für den Beweis, dass die Funktion u nur auf einem Teil des Randes Nullrandwerte annimmt. Hier sind das die Punkte $\{x \in \partial\Omega \mid \forall y \in \partial\Omega : x_1 \leq y_1\}$, im allgemeinen bei stückweise glattem Rand auf einer Menge mit positivem $n - 1$ -dimensionalem Maß. \square

4 Lösbarkeit des Variationsproblems

4.1 Der Satz von Lax-Milgram

Der Nachweis der Existenz und Eindeutigkeit von Lösungen des Variationsproblems aus Satz 2.2 gelingt nun in den geeigneten Funktionenräumen.

Wir zeigen zunächst ein allgemeines Resultat, das wir dann auf unser konkretes Variationsproblem anwenden werden.

Definition 4.1 (Eigenschaften der Bilinearform). Sei H ein Hilbertraum mit der Norm $\|\cdot\|$ und $a : H \times H \rightarrow \mathbb{R}$ eine Bilinearform.

a heißt stetig, wenn mit einem $C > 0$ gilt:

$$|a(u, v)| \leq C\|u\|\|v\| \text{ für alle } u, v \in H.$$

Ein symmetrisches, stetiges a heißt H -elliptisch (kurz: elliptisch, koerziv), falls mit einem $\alpha > 0$ gilt:

$$a(v, v) \geq \alpha\|v\|^2 \text{ für alle } v \in H. \quad (4.1)$$

Für stetige, symmetrische und elliptische Bilinearformen sind $\|\cdot\|_H$ und $\|\cdot\|_a = \sqrt{a(\cdot, \cdot)}$ äquivalente Normen. \square

Damit können wir nun den entscheidenden Satz formulieren:

Satz 4.2 (Lax-Milgram). Sei V eine abgeschlossene, konvexe Menge in einem Hilbertraum H . $a : H \times H \rightarrow \mathbb{R}$ sei eine elliptische (und damit stetige, symmetrische) Bilinearform. Für jedes lineare Funktional $l : H \rightarrow \mathbb{R}$ hat das Variationsproblem

$$J(v) := \frac{1}{2}a(v, v) - l(v) \rightarrow \min!$$

dann genau eine Lösung in V .

Beweis: Zunächst zeigen wir, dass J nach unten beschränkt ist. Für Linearformen gilt $l(v) \leq L\|v\| \forall v \in H$ (die Norm ist immer $\|\cdot\|_H$), also zusammen mit der Elliptizität:

$$J(v) \geq \frac{1}{2}\alpha\|v\|^2 - L \cdot \|v\| = \frac{1}{2\alpha}(\alpha\|v\| - L)^2 - \frac{L^2}{2\alpha} \geq -\frac{L^2}{2\alpha}$$

Setze $c_1 = \inf\{J(v) \mid v \in V\}$. c_1 ist also die größte untere Schranke. Sei $(v_n)_{n \in \mathbb{N}}$ eine Minimalfolge in V (d.h. $J(v_n) < J(v_m) \forall n > m$) Dann ist

$$\begin{aligned} \alpha\|v_n - v_m\|^2 &\leq a(v_n - v_m, v_n - v_m) = a(v_n, v_m) - 2a(v_n, v_m) + a(v_m, v_m) \\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m) - 4l(v_n) - 4l(v_m) + 4l(v_n + v_m) \\ &= 4J(v_n) + 4J(v_m) - \underbrace{[a(v_n + v_m, v_n + v_m) - 4l(v_n + v_m)]}_{4a(\frac{v_n+v_m}{2}, \frac{v_n+v_m}{2}) - 8l(\frac{v_n+v_m}{2})} \\ &= 4J(v_n) + 4J(v_m) - 8J(\frac{v_n + v_m}{2}) \\ &\leq 4J(v_n) + 4J(v_m) - 8c_1. \end{aligned}$$

Da (v_n) Minimalfolge, gilt $J(v_n), J(v_m) \rightarrow c_1$ und somit also $\|v_n - v_m\| \rightarrow 0$. Also ist (v_n) eine Cauchy-Folge in H und es existiert $u = \lim_{n \rightarrow \infty} v_n$ in H da H vollständig ist. Da V abgeschlossen ist auch $u \in V$. Da J stetig ist, gilt auch $J(u) = \lim_{n \rightarrow \infty} J(v_n) = \inf_{v \in V} J(v) = c_1$.

Eindeutigkeit: Seien $u_1, u_2, u_1 \neq u_2$ zwei Lösungen des Minimierungsproblems. Die Folge $u_1, u_2, u_1, u_2, \dots$ ist eine Minimalfolge. Da jede Minimalfolge eine Cauchy-Folge ist, muss $\|u_1 - u_2\| = 0$ sein, also $u_1 = u_2$. \square

Bemerkung 4.3. Wir betrachten den Spezialfall $a(u, v) = (u, v)_H$, wobei $(u, v)_H$ das Skalarprodukt des Hilbertraumes ist, und setzen $V = H$.

Zu jedem linearen Funktional $l(v)$ existiert nach Lax-Milgram eine eindeutige Lösung u_l , so dass dann nach dem Charakterisierungssatz

$$(u_l, v)_H = l(v) \quad \forall v \in V = H.$$

Jedes lineare Funktional über einem Hilbertraum lässt sich also mittels

$$l(v) = (u_l, v) \text{ darstellen.}$$

Dies ist der Darstellungssatz von Riesz. Umgedreht ist natürlich für jedes $u \in H$ $l(v) := (u, v)_h$ ein lineares Funktional. Jedes lineare Funktional auf H kann also mit einem Element des Hilbertraumes identifiziert werden. Man schreibt deshalb oft auch $\langle l, v \rangle$ statt $l(v)$.

Schließlich kann der Satz von Lax-Milgram auch auf unsymmetrische Bilinearformen verallgemeinert werden. Dies erfordert jedoch eine andere Beweistechnik. \square

4.2 Anwendung auf das Dirichletproblem

Wir weisen nun die Voraussetzungen des Satzes von Lax-Milgram für unsere elliptische Differentialgleichung mit Dirichlet-Randbedingungen nach.

Zunächst führen wir für die Lösung des Variationsproblems den Begriff der schwachen Lösung ein.

Definition 4.4 (Schwache Lösung). Eine Funktion $u \in H_0^1(\Omega)$ heißt schwache Lösung der Randwertaufgabe mit homogenen Dirichletrandbedingungen

$$\begin{aligned} -\nabla \cdot \{A\nabla u\} &= f & \text{in } \Omega \\ u &= 0 & \text{auf } \partial\Omega, \end{aligned} \tag{4.2}$$

wenn mit der Bilinearform $a(u, v) = \int_{\Omega} (A\nabla u) \cdot \nabla v \, dx$ die Gleichungen

$$a(u, v) = (f, v)_0 \quad \forall v \in H_0^1(\Omega)$$

gelten. Der $H_0^1(\Omega)$ ist erforderlich, da die Nullrandbedingungen in den Ansatzraum einzubauen sind. \square

Um die Existenz einer schwachen Lösung zu beweisen, benötigen wir noch eine Voraussetzung an die Koeffizienten des elliptischen Randwertproblems.

Definition 4.5 (Gleichmäßige Elliptizität). Ein elliptischer Differentialoperator der Gestalt $\mathcal{L}u = -\nabla \cdot \{A\nabla u\}$ heißt gleichmäßig elliptisch, wenn es eine Zahl $\alpha \geq 0$ gibt, so dass

$$\xi^T A(x)\xi \geq \alpha \|\xi\|^2 \quad \text{für alle } \xi \in \mathbb{R}^n \text{ und } x \in \Omega. \quad (4.3)$$

Dies bedeutet, dass die Eigenwerte von A auch gegen den Rand hin von Null weg beschränkt bleiben müssen. \square

Außerdem setzen wir voraus, dass die Einträge von A beschränkte Funktionen auf Ω sind.

Satz 4.6 (Existenzsatz). Der Differentialoperator in (4.2) sei gleichmäßig elliptisch. Dann existiert stets eine eindeutige schwache Lösung $u \in H_0^1(\Omega)$ von (4.2). Diese ist auch Lösung des Variationsproblems.

$$\frac{1}{2}a(v, v) - (f, v)_0 \rightarrow \min. \quad (4.4)$$

Beweis: Nach dem Charakterisierungssatz ist (4.4) äquivalent zum Variationsproblem. Dessen Lösbarkeit sichert der Satz von Lax-Milgram. Da $H_0^1(\Omega)$ ein Hilbertraum ist, sind als weitere Voraussetzungen nur noch die Stetigkeit und Elliptizität der Bilinearform a nachzuweisen.

Stetigkeit der Bilinearform. Wir schätzen ab

$$\begin{aligned} |a(u, v)| &= \left| \sum_{i,k} \int_{\Omega} a_{i,k} \partial_i u \partial_k v \, dx \right| \leq \sum_{i,k} \left| \int_{\Omega} a_{i,k} \partial_i u \partial_k v \, dx \right| \\ &\leq c \sum_{i,k} \int_{\Omega} |\partial_i u \partial_k v| \, dx \\ &\leq c \cdot \sum_{i,k} \left[\int_{\Omega} (\partial_i u)^2 \, dx \int_{\Omega} (\partial_k v)^2 \, dx \right]^{\frac{1}{2}} \\ &= c \sum_i \left\{ \left[\int_{\Omega} (\partial_i u)^2 \, dx \right]^{\frac{1}{2}} \cdot \sum_k \left[\int_{\Omega} (\partial_k v)^2 \, dx \right]^{\frac{1}{2}} \right\} \\ &= c \sum_i \left[\int_{\Omega} (\partial_i u)^2 \, dx \right]^{\frac{1}{2}} \cdot \sum_k \left[\int_{\Omega} (\partial_k v)^2 \, dx \right]^{\frac{1}{2}} \\ &\leq c \left(\sum_i \int_{\Omega} (\partial_i u)^2 \, dx \right)^{\frac{1}{2}} \cdot n^{\frac{1}{2}} \cdot \left(\sum_R \int_{\Omega} (\partial_k v)^2 \, dx \right)^{\frac{1}{2}} \cdot n^{\frac{1}{2}} \\ &\leq C |u|_1 |v|_1 \end{aligned}$$

Wegen $|u|_1 \leq \|u\|_1$ gilt also

$$|a(u, v)| \leq C \|u\|_1 \|v\|_1 \quad \text{für alle } u, v \in H_0^1(\Omega).$$

Elliptizität der Bilinearform. Betrachte Punkt $x \in \Omega$, dann gelten wegen der gleichmäßigen Elliptizität:

$$(A(x)\nabla v(x)) \cdot \nabla v(x) \geq \alpha \nabla v(x) \cdot \nabla v(x).$$

Integration liefert dann

$$a(v, v) = \int_{\Omega} (A\nabla v) \cdot \nabla v \, dx \geq \alpha \int_{\Omega} \nabla v \cdot \nabla v \, dx = \alpha |v|_1^2.$$

Mit der Poincaré-Ungleichung $\|v\|_0^2 \leq s^2 |v|_1^2$ für $v \in H_0^1(\Omega)$ gilt dann

$$\|v\|_1^2 = \|v\|_0^2 + |v|_1^2 \leq s^2 |v|_1^2 + |v|_1^2 = (s^2 + 1) |v|_1^2$$

also

$$\frac{1}{\sqrt{s^2 + 1}} \|v\|_1 \leq |v|_1.$$

Damit erhalten wir dann die Elliptizität der Bilinearform:

$$a(v, v) \geq \frac{\alpha}{1 + s^2} \|v\|_1^2.$$

□

Bemerkung 4.7. Obiger Beweis lässt sich auf die Gleichung

$$\begin{aligned} -\nabla \cdot \{A\nabla u\} + a_0(x)u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

für beschränktes und nicht negatives $a_0(x)$ verallgemeinern. Für den zusätzlichen Term in der Stetigkeit gilt

$$\left| \int_{\Omega} a_0 uv \, dx \right| \leq c \int_{\Omega} |u| |v| \, dx \leq c \|u\|_0 \|v\|_0 \leq c \|u\|_1 \|v\|_1.$$

Für den zusätzlichen Term in der Elliptizität erhält man

$$\int_{\Omega} a_0 v^2 \, dx \geq 0.$$

□

4.3 Anwendung auf die Neumann'sche Randwertaufgabe

Wie wollen nun die Gleichung

$$\begin{aligned} -\nabla \cdot \{A(x)\nabla u\} + a_0(x)u &= f && \text{in } \Omega, \\ -(A(x)\nabla u) \cdot \nu &= g && \text{auf } \partial\Omega, \end{aligned} \quad (4.5)$$

also Flussrandbedingung, betrachten. Im Gegensatz zu oben sei $a_0(x) \geq \alpha_0 > 0$ für alle $x \in \Omega$.

Mit dieser Annahme wird die Bilinearform auf ganz $H^1(\Omega)$ elliptisch:

$$\begin{aligned} \int_{\Omega} (A\nabla v) \cdot \nabla v + a_0 v v \, dx &\geq \alpha \int_{\Omega} \nabla v \cdot \nabla v \, dx + \alpha_0 \int_{\Omega} v \cdot v \, dx \\ &\geq \min(\alpha, \alpha_0) \int_{\Omega} v v + \nabla v \cdot \nabla v \, dx = \min(\alpha, \alpha_0) \|v\|_1^2. \end{aligned} \quad (4.6)$$

Die Linearform des zu (4.5) gehörigen Variationsproblems lautet

$$l(v) = \int_{\Omega} f v \, dx - \int_{\partial\Omega} g v \, ds. \quad (4.7)$$

Dass mit dem darin enthaltenen Randintegral ein stetiges lineares Funktional erklärt wird, sichert der Spursatz:

Satz 4.8 (Spursatz). Sei Ω beschränkt und habe einen stückweise glatten Rand. Ferner erfülle Ω die Kegelbedingung. Dann gibt es genau eine beschränkte lineare Abbildung

$$\gamma : H^1(\Omega) \rightarrow L_2(\partial\Omega)$$

mit

$$\|\gamma(v)\|_{0,\partial\Omega} \leq c \|v\|_{1,\Omega},$$

so dass $(\gamma v)(x) = v(x)$ für alle $v \in C^1(\overline{\Omega})$.

Beweis: Siehe [Bra91, S. 45]. □

Dieser Satz besagt, dass die Auswertung einer H^1 -Funktion auf dem Rand eine L_2 -Funktion ergibt. Nun können wir den Existenzsatz für das Neumann-Problem formulieren.

Satz 4.9. Das Gebiet erfülle die Voraussetzungen aus dem Spursatz. Die Variationsaufgabe für (4.5)

$$J(v) := \frac{1}{2} \underbrace{\int_{\Omega} (A\nabla v) \cdot \nabla v + a_0 v v \, dx}_{=a(u,v)} - \int_{\Omega} f v \, dx + \int_{\partial\Omega} g v \, ds \rightarrow \min$$

hat genau eine Lösung in $H^1(\Omega)$.

Beweis: Die Bilinearform $a(u, v)$ ist stetig (keine Änderung zum Dirichletproblem) und wegen $a_0 \geq \alpha_0 > 0$ auf ganz $H^1(\Omega)$ elliptisch wie oben bereits gezeigt (hierfür war keine Poincaré-Ungleichung erforderlich).

$\int_{\Omega} f v dx - \int_{\partial\Omega} g v ds$ ist eine stetige Linearform. Spursatz und die Annahme $g \in L_2(\partial\Omega)$ ergeben

$$\left| \int_{\partial\Omega} g v \right| \leq \underbrace{\left(\int_{\partial\Omega} g^2 dx \right)^{\frac{1}{2}}}_{\text{konst.}} \underbrace{\left(\int_{\partial\Omega} v^2 dx \right)^{\frac{1}{2}}}_{\leq c\|v\|_1 \text{ Spursatz}}.$$

Damit liefert der Satz von Lax-Milgram die eindeutige Lösung in $H^1(\Omega)$.

Das Variationsproblem zu (4.5) erhält man wie beim Dirichletproblem durch Anwendung der Green'schen Formel:

$$\begin{aligned} & \int_{\Omega} [-\nabla \cdot \{A\nabla u\} + a_0 u] v dx = \int_{\Omega} f v dx \\ \Leftrightarrow & \int_{\Omega} (A\nabla u) \cdot \nabla v + a_0 u v dx + \int_{\partial\Omega} -(A\nabla u) \cdot \nu v ds = \int_{\Omega} f v dx \\ \Leftrightarrow & \int_{\Omega} (A\nabla u) \cdot \nabla v + a_0 u v dx = \int_{\Omega} f v dx - \int_{\partial\Omega} g v ds. \end{aligned}$$

□

Man kann auch noch zeigen, dass die Lösung der Variationsaufgabe genau dann in $C^2(\Omega) \cap C^1(\bar{\Omega})$ enthalten ist, wenn eine klassische Lösung von 4.5 existiert.

Neumann-Problem ohne Helmholtz-Term

Für die Aufgabe

$$\begin{aligned} -\nabla \cdot \{A\nabla u\} &= f \quad \text{in } \Omega \\ -(A\nabla u) \cdot \nu &= g \quad \text{auf } \partial\Omega \end{aligned} \tag{4.8}$$

hatten wir auf Seite 12 erläutert, dass die Lösung nur bis auf eine additive Konstante bestimmt ist und zusätzlich die Kompatibilitätsbedingung (1.14) gelten muss.

In unserer Theorie ergibt sich das Problem, dass die Bilinearform $a(u, v) = \int_{\Omega} (A\nabla u) \cdot \nabla v dx$ nicht elliptisch auf $H^1(\Omega)$ ist.

Berücksichtigt man, dass die Lösung nur bis auf eine additive Konstante festgelegt ist und wählt den Raum

$$V = \left\{ v \in H^1(\Omega) \mid \int_{\Omega} v dx = 0 \right\} \subset H^1(\Omega), \tag{4.9}$$

so kann man die Elliptizität der Bilinearform in V nachweisen.

5 Ritz-Galerkin Verfahren

5.1 Die Idee

Wie wendet man das bisher gezeigte nun *praktisch* zur Lösung eines Randwertproblems an?

Hier die Idee: Löse das Variationsproblem in einem *endlichdimensionalen* Unterraum von $H_0^1(\Omega)$ (oder $H^1(\Omega)$).

Angenommen $S_h \subset H_0^1(\Omega)$ sei so ein Unterraum. Der Index h steht hierbei für einen „Diskretisierungsparameter“ (Gitterweite). Für $h \rightarrow 0$ geht $\dim(S_h) \rightarrow \infty$ und wir erwarten „Konvergenz“ der Methode (das werden wir unten präzisieren).

Wie löst man nun das Variationsproblem in S_h ? Nach dem Satz von Lax-Milgram hat das Variationsproblem

$$J(v) = \frac{1}{2}a(v, v) - l(v) \rightarrow \min \quad \text{in } S_h \subset H_0^1(\Omega) \quad (5.1)$$

genau eine Lösung in einer abgeschlossenen, konvexen Teilmenge V eines Hilbertraumes. Diese Teilmenge ist unser S_h .

Nach dem Charakterisierungssatz ist das Variationsproblem äquivalent zu

$$\text{Finde } u_h \in S_h : a(u_h, v) = l(v) \quad \forall v \in S_h \quad (5.2)$$

Nun ist S_h endlichdimensional mit $N_h = \dim(S_h)$. Dann gibt es eine *Basis* $\Psi_h = \{\psi_1, \dots, \psi_{N_h}\}$ von S_h und es *genügt* wegen der Linearität (5.2) für alle $v = \psi_1, \dots, \psi_{N_h}$ zu fordern:

$$a(u_h, \psi_i) = l(\psi_i) \quad \forall i = 1, \dots, N_h.$$

Schließlich ist auch $u_h \in S_h$ und wir können u_h in der Basis Ψ_h ausdrücken:

$$u_h = \sum_{k=1}^{N_h} z_k \psi_k.$$

Einsetzen liefert dann

$$a(u_h, \psi_i) = a\left(\sum_{k=1}^{N_h} z_k \psi_k, \psi_i\right) \stackrel{\text{Linearität}}{=} \sum_{k=1}^{N_h} z_k a(\psi_k, \psi_i) = l(\psi_i) \quad i = 1 \dots N_h. \quad (5.3)$$

Dies ist ein *lineares Gleichungssystem* (LGS) für die Koeffizienten z_1, \dots, z_{N_h} .

Dieses LGS schreiben wir als:

$$Az = b \quad A_{ik} = a(\psi_k, \psi_i), \quad b_i = l(\psi_i).$$

Aus den Eigenschaften der Bilinearform a erhält man sofort einige wichtige Eigenschaften des LGS:

- A ist symmetrisch, da $A_{ik} = a(\psi_k, \psi_i) = a(\psi_i, \psi_k) = A_{ki}$

- A ist positiv definit, da für beliebiges $z \in \mathbb{R}^{N_h}$ gilt:

$$\begin{aligned}
 z^T A z &= \sum_{i=1}^N z_i \sum_{k=1}^N A_{ik} z_k = a \left(\underbrace{\sum_{i=1}^N z_i \psi_i}_{u_h}, \underbrace{\sum_{k=1}^N z_k \psi_k}_{=u_h} \right) \\
 &= a(u_h, u_h) \geq \alpha \|u_h\|_1^2 \\
 &> 0 \quad \text{falls } u_h \neq 0 \Leftrightarrow z \neq 0.
 \end{aligned}$$

Jedem Vektor von Koeffizienten $z \in \mathbb{R}^{N_h}$ entspricht in eindeutiger Weise eine Funktion aus S_h über die Beziehung

$$u_h = \sum_{k=1}^N z_k \psi_k.$$

Einige Bezeichnungen

- A heißt in der Ingenieurliteratur oft Steifigkeitsmatrix und b Lastvektor.
- Unter einem Galerkin-Verfahren versteht man die Verallgemeinerung auf unsymmetrische Bilinearformen. Diese können dann nicht mehr als Minimierungsproblem interpretiert werden. In Galerkin-Verfahren sind Ansatzraum (für u_h) und Testraum (für v) identisch.
- Lässt man diese Voraussetzung fallen, so führt dies auf Petrov-Galerkin-Verfahren:

$$\text{Finde } u_h \in S_h : a(u, v) = l(v) \quad \forall v \in V_h \quad \text{mit } S_h \neq V_h.$$

Bekannte Vertreter sind sogenannte Finite-Volumen-Verfahren.

5.2 Eigenschaften der diskreten Lösung

Die Lösung des Variationsproblems erfolgt in einem passend zu den Randbedingungen gewählten Funktionenraum, etwa $H_0^1(\Omega)$ bei Dirichlet-Randbedingungen oder $H^1(\Omega)$ bei Neumann-Randbedingungen. Entsprechend ist dann S_h jeweils als endlichdimensionale Teilmenge des passenden Sobolevraumes zu wählen.

Die folgenden Resultate gelten unabhängig von der Wahl des speziellen Raumes, solange er passend zu dem Randwertproblem gewählt ist. Wir nehmen also im folgenden an, dass

$$H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$$

ein passend gewählter Sobolevraum und die Bilinearform a entsprechend V-elliptisch ist.

Lemma 5.1 (Stabilität). Es sei $S_h \subseteq V$. Für die Lösung von 5.2 gilt dann

$$\|u_h\|_1 \leq \alpha^{-1} \|l\|. \quad (5.4)$$

Beweis: Es ist $a(u_h, v) = l(v) \quad \forall v \in S_h$ (5.2). Speziell für $v = u_h$ gilt dann

$$\alpha \|u_h\|_1^2 \leq a(u_h, u_h) = l(u_h) \leq \|l\| \|u_h\|_1 \Leftrightarrow \|u_h\|_1 \leq \alpha^{-1} \|l\|$$

wobei $\|l\|$ die Konstante aus $l(v) \leq \|l\| \|v\|_1$ ist. \square

Unter Stabilität versteht man, dass kleine Änderungen in den Daten (hier die Linearform l und damit die rechte Seite f) auch entsprechend kleine Änderungen in der Lösung zur Folge haben. Dies kann man aus dem Lemma folgendermaßen sehen:

Für zwei verschiedene Funktionale l, l' erhalten wir:

$$\begin{aligned} a(u_h, v) &= l(v) \quad \forall v \in S_h, \\ a(u'_h, v) &= l'(v) \quad \forall v \in S_h, \\ a(u_h - u'_h, v) &= l(v) - l'(v) = \tilde{l}(v) \quad \forall v \in S_h. \end{aligned}$$

Für die Differenz $u_h - u'_h$ folgt damit nach dem Lemma

$$\|u_h - u'_h\|_1 \leq \alpha^{-1} \|\tilde{l}\|.$$

Da $\tilde{l}(v) = \int_{\Omega} (f - f')v \, dx$ gilt also

$$\text{„}f \rightarrow f'\text{„} \Rightarrow \|\tilde{l}\| \rightarrow 0 \Rightarrow \|u_h - u'_h\|_1 \rightarrow 0.$$

Das nun folgende Lemma ist der Ausgangspunkt für die Fehlerabschätzung für das Finite-Elemente-Verfahren.

Lemma 5.2 (Lemma von Céa). Es sei

$$\begin{aligned} u \in V, \quad a(u, v) &= l(v) \quad \forall v \in V, \\ u_h \in S_h, \quad a(u_h, v) &= l(v) \quad \forall v \in S_h \end{aligned} \quad (5.5)$$

Dann gilt

$$\|u - u_h\| \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_1.$$

Beweis: Subtraktion der Gleichungen aus (5.5) liefert wegen $S_h \subset V$:

$$a(u - u_h, v) = 0 \quad \forall v \in S_h \quad (5.6)$$

Mit einem beliebig gewähltem $v_h \in S_h$ gilt dann

$$\begin{aligned} \alpha \|u - u_h\|_1^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h + v_h - u_h) \\ &= a(u - u_h, u - v_h) + \underbrace{a(u - u_h, v_h - u_h)}_{=0 \text{ wg (5.6) und } v_h - u_h \in S_h} \\ &\leq C \|u - u_h\|_1 \|u - v_h\|_1 \\ \Leftrightarrow \|u - u_h\|_1 &\leq \frac{C}{\alpha} \|u - v_h\|_1 \quad \text{für beliebiges } v_h \in S_h, \\ \text{also } \|u - u_h\|_1 &\leq \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_1. \end{aligned} \quad (5.7)$$

□

Das Lemma von Céa besagt, dass der Fehler in der Näherung u_h durch den Approximationsfehler, d. h. der bestmöglichen Approximation der exakten Lösung durch eine Funktion aus S_h , abgeschätzt werden kann. Dadurch muss man in der Fehlerabschätzung nicht mehr berücksichtigen, wie das Variationsproblem nun genau gelöst wird.

Die Beziehung (5.6) wird als „Galerkin-Orthogonalität“ bezeichnet und spielt in vielen Beweisen eine wichtige Rolle.

5.3 Finite Elemente in einer Raumdimension

Zum Abschluss dieses Kapitels wollen wir ein erstes einfaches Beispiel einer Finite-Elemente-Methode kennenlernen.

Betrachte

$$-\frac{d^2u}{dx^2} = f \quad \text{in } (0, 1) \quad (5.8)$$

mit den Randbedingungen

$$\begin{aligned} u(0) &= 0, \\ u(1) &= 0. \end{aligned}$$

Hierbei handelt es sich eigentlich um eine gewöhnliche Differentialgleichung. Durch die Vorgabe von Bedingungen an zwei *verschiedenen* Punkten handelt es sich nicht um eine Anfangswertaufgabe, sondern um eine Zwei-Punkt-Randwertaufgabe, die besser mit Methoden für partielle Differentialgleichungen gelöst wird.

Wir geben eine mögliche Wahl für S_h an. Unterteile das Intervall $(0, 1)$ in n gleichgroße Abschnitte:

$$\begin{array}{c} | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \\ \hline \end{array} \quad x_i = i \cdot h \quad i = 0, \dots, n, \quad h = \frac{1}{n}$$

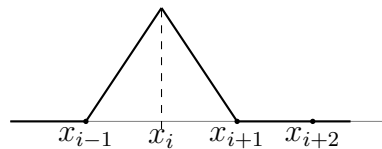
Als Funktionenraum wähle nun den Raum der stückweise linearen Funktionen

$$S_h = \{f \in C^0([0, 1]) \mid f|_{[x_i, x_{i+1}]} \text{ ist linear} \wedge f(0) = 0 \wedge f(1) = 0\}.$$

Als Basis für diesen endlichdimensionalen Funktionenraum wählt man die sogenannten Hutfunktionen

$$\psi_i \in S_h, \quad i = 1, \dots, n-1$$

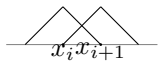
$$\psi_i(x_j) = \begin{cases} 1 & i = j \\ 0 & \text{sonst} \end{cases}$$



Die Finite-Elemente-Formulierung lautet in einer Raumdimension

$$a(u_h, v) = \int_0^1 \frac{du_h}{dx} \cdot \frac{dv}{dx} dx = \int_0^1 f v dx = l(v) \quad \forall v \in S_h.$$

Für die gewählte Basis erhalten wir dann die folgenden Matrixeinträge:

$$a(\psi_k, \psi_i) = \begin{cases} \int_{x_i-h}^{x_i} \frac{1}{h} \cdot \frac{1}{h} dx + \int_{x_i}^{x_i+h} \left(-\frac{1}{h}\right) \cdot \left(-\frac{1}{h}\right) dx = \frac{1}{h} + \frac{1}{h} = \frac{2}{h} & k = i \\ \int_{x_i}^{x_i+h} \left(-\frac{1}{h}\right) \cdot \frac{1}{h} dx = -\frac{1}{h} & k = i \pm 1 \\ 0 & \text{sonst} \end{cases}$$


Für die rechte Seite erhält man bei Darstellung der Funktion f als Funktion aus S_h (was einer numerischen Integration mit der Trapezregel entspricht):

$$b_i = h \cdot \left(\frac{1}{6}f_{i-1} + \frac{2}{3}f_i + \frac{1}{6}f_{i+1}\right) \quad \text{für} \quad f = \sum_{k=1}^n f(x_k) \cdot \psi_k.$$

Eine Zeile des linearen Gleichungssystems lautet damit

$$\frac{1}{h}(-z_{i-1} + 2z_i - z_{i+1}) = b_i$$

Die Matrix A ist somit tridiagonal.

6 Gebräuchliche Finite Elemente

Wir beschreiben nun sehr allgemein wie man Finite-Element-Räume (d.h. endlichdimensionale Funktionenräume) $S_h \subset V$ konstruiert. Hierbei ist V ein Teilraum eines Sobolevraumes $H^m(\Omega)$ (etwa der $H_0^m(\Omega)$) und Ω ein Gebiet im \mathbb{R}^n .

Gilt $S_h \subset H^m(\Omega)$, spricht man von einem *konformen* Finite-Element-Raum. Es gibt auch verallgemeinerte FE-Verfahren, für die $S_h \not\subset V$ ist. Dann spricht man von *nicht-konformen* Methoden.

Grundidee von FE-Räumen:

- Zerlegung („Triangulierung“) von Ω in Teilgebiete (= *Elemente*) einfacher Gestalt, etwa Dreiecke, Vierecke, Tetraeder, Pyramide, Prismen, Hexaeder,...
- FE-Funktionen sind *Polynome* (in mehreren Variablen) auf jedem Element.

Wichtige Merkmale von FE-Räumen sind:

- Art der Zerlegung in Elemente, welche Elemente werden verwendet, wie grenzen sie aneinander, ...
- Grad der Polynome. Polynome vom Grad t sind definiert als

$$\text{im } \mathbb{R}^2 : P_t = \left\{ u \mid u(x, y) = \sum_{i+k \leq t} c_{ik} x^i y^k \right\},$$

$$\text{im } \mathbb{R}^n : P_t = \left\{ u \mid u(\underline{x}) = \sum_{|\alpha| \leq t} c_\alpha x^\alpha \right\}.$$

bzw.

$$\text{im } \mathbb{R}^2 : Q_t = \left\{ u \mid u(x, y) = \sum_{\max\{i,k\} \leq t} c_{ik} x^i y^k \right\},$$

$$\text{im } \mathbb{R}^n : Q_t = \left\{ u \mid u(\underline{x}) = \sum_{\max \alpha_i \leq t} c_\alpha x^\alpha \right\}.$$

- Zusätzlich kann der Grad auf Kanten, Seiten, etc. der Elemente eingeschränkt sein.
- globale Stetigkeits- und Differenzierbarkeitseigenschaften. Etwa C^k -Elemente, d. h. $S_h \subset C^k(\bar{\Omega})$.

6.1 Eigenschaften der Zerlegung

- Das Gebiet Ω sei polyedrisch (= polygonal wenn $n = 2$), damit kann es in (genügend viele) Polyeder zerlegt werden.

- Wir beschränken uns im Folgenden auf $\Omega \subset \mathbb{R}^2$ und *Dreiecke*.

Definition 6.1 (Triangulierung). Eine Menge von Dreiecken $\mathcal{T}(\Omega) = \{T_1, \dots, T_M\}$ nennen wir eine Triangulierung von Ω falls die folgenden Eigenschaften erfüllt sind:

1. $\mathcal{T}(\Omega) = \{T_1, \dots, T_M\}$ zerlege Ω in Dreiecke T_i . Die T_i sind *abgeschlossene* Gebiete und es gilt:

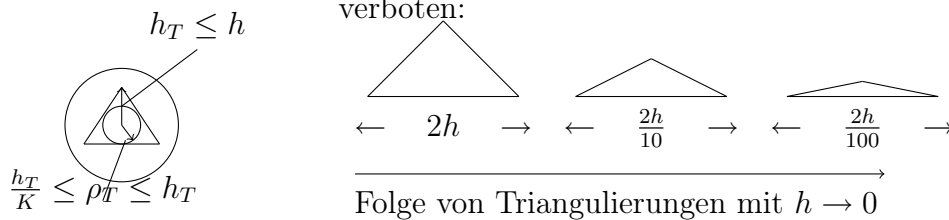
$$(i) \bigcup_{i=1}^m T_i = \overline{\Omega}$$

(ii) Ist $T_i \cap T_j = \{x\}, x \in \mathbb{R}^2$, so ist x eine Ecke von T_i und von T_j .

(iii) Ist $T_i \cap T_j = K_{ij} \subset \overline{\Omega}$ mehr als ein Punkt, so ist K_{ij} sowohl eine *Kante* von T_i als auch von T_j .

Gitter mit diesen drei Eigenschaften nennt man auch *konform*.

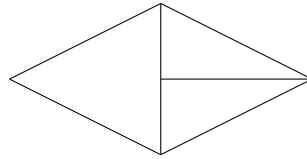
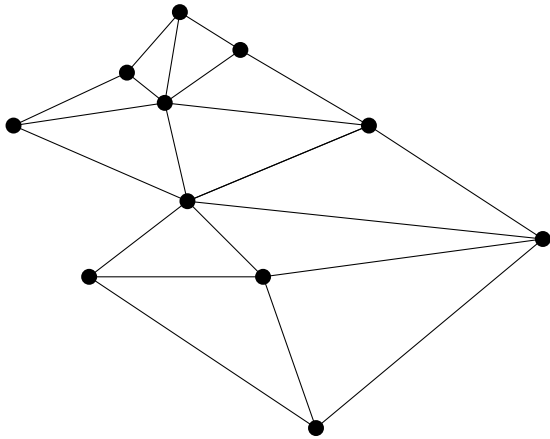
2. Wir schreiben \mathcal{T}_h falls jedes Element in einem Kreis mit Durchmesser $2h$ passt.
3. Eine Familie $\{\mathcal{T}_h\}$ von Zerlegungen heißt *quasiuniform*, wenn es eine Zahl $K > 0$ gibt, so dass jedes $T \in \mathcal{T}_h$ einen Kreis mit Radius $\rho_T \geq \frac{h_T}{K}$ enthält



Elemente in quasiuniformen Gittern können stark unterschiedlich groß sein $h_{max} = const$, $h_{max} \rightarrow 0$ ist erlaubt. Quasiuniformität verhindert allerdings, dass die Elemente beliebig spitze oder stumpfe Winkel enthalten. Daher passt der englische Begriff *shape-regular* eigentlich besser. Man überlege auch, dass ein Dreieck mit spitzen Winkeln keinen stumpfen Winkel enthalten muss, ein Dreieck mit einem stumpfen Winkel jedoch zwei spitze Winkel enthalten muss.

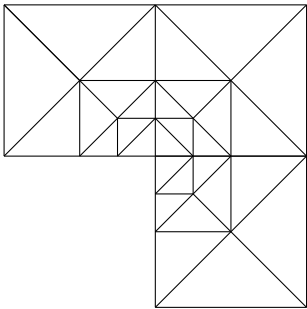
4. Eine Familie von Triangulierungen heißt *uniform*, falls jedes $T \in \mathcal{T}_h$ einen Kreis mit Radius $\rho_T \geq \frac{h}{K}$ enthält. Hier wird also das globale h verwendet. \square

Beispiele für Gitter:



nicht zulässige Triangulierung

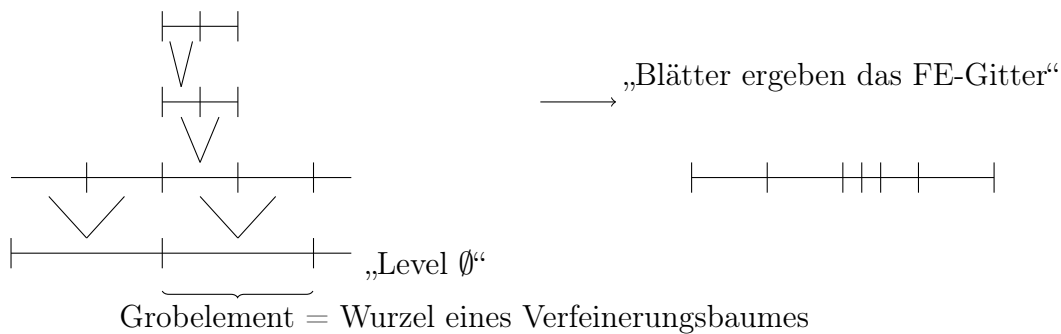
Zulässige Triangulierung



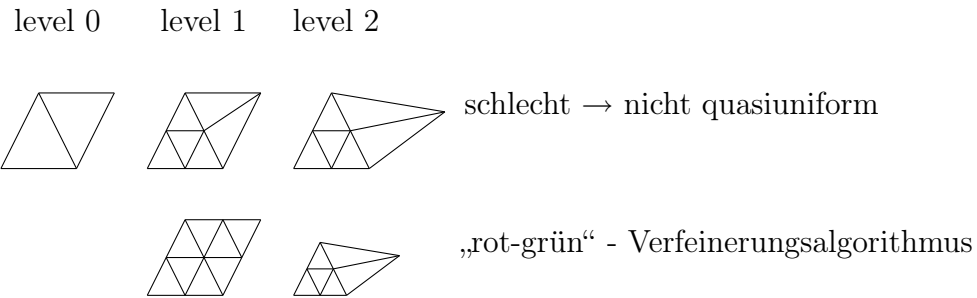
quasiuniforme, aber nicht uniforme Familie
dynamisch entwickelt.

Das automatische Erzeugen einer Triangulierung zu gegebener Geometrie nennt man „Gittergenerierung“. Je nach Geometrie ist dieses Problem sehr schwierig (siehe Vorlesung Simulationswerkzeuge).

Hierarchische Verfeinerung Eine effiziente Methode zur Erzeugung angepasster Gitter ausgehend von einem Startgitter ist die Methode der hierarchischen Verfeinerung. Wir illustrieren das Vorgehen zunächst für $n = 1$.



Ab $n = 2$ muss man spezielle Vorkehrungen treffen um die Konformität sicherzustellen:



6.2 Konforme Finite-Elemente-Räume

Zunächst ist noch gar nicht klar wie man eigentlich genau einen endlichdimensionalen Teilraum des $H^1(\Omega)$ konstruiert. Hier gibt der folgende Satz eine prinzipielle Aussage.

Satz 6.2. Sei $k \geq 1$ und Ω beschränkt. Eine elementweise beliebig oft differenzierbare Funktion $v : \bar{\Omega} \rightarrow \mathbb{R}$ gehört *genau dann* zu $H^k(\Omega)$, wenn $v \in C^{k-1}(\bar{\Omega})$ ist.

Beweis: siehe [Bra91, Satz 5.2].

Folgerung:

$$v \in C^0(\Omega) \text{ und } v \text{ elementweise Polynom} \Leftrightarrow v \in H^1(\Omega)$$

Da wir nur H^1 -Funktionen brauchen, nehmen wir demnach stetige, elementweise polynomiale Funktionen.

Dreieckselemente der Ordnung t

Sei $\mathcal{T}_h = \{T_1, \dots, T_M\}$ sei eine Triangulierung von Ω mit Dreiecken.

Mit

$$S_h^t := \{u \in C^0(\bar{\Omega}) \mid \forall j = 1, \dots, M : u|_{T_j} \in P_t\}$$

bezeichnen wir den Raum aller stetigen Funktionen, die auf jedem Dreieck ein Polynom vom Grad t sind.

Wegen

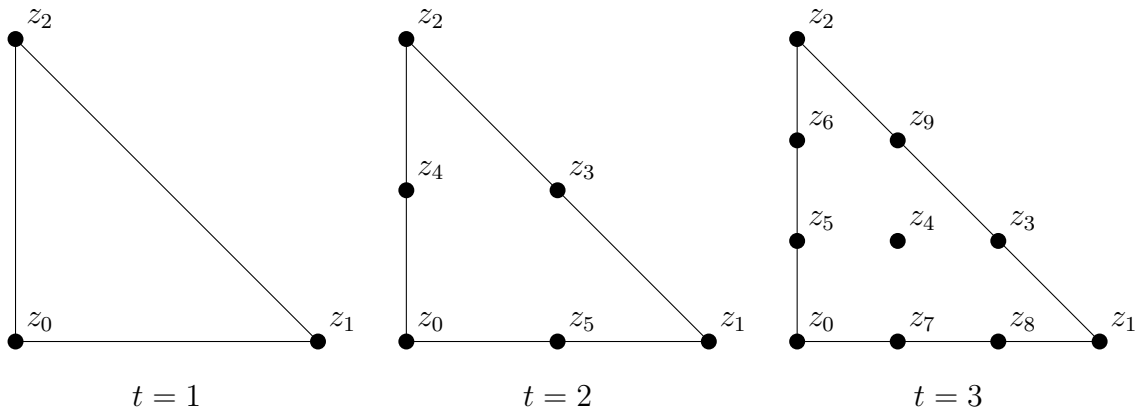
$$u|_{T_j} = \sum_{i+k \leq t} c_{ik}^j x^i y^k$$

wird $u|_{T_j}$ durch $\frac{(t+1)(t+2)}{2}$ Koeffizienten bestimmt:

$$\sum_{m=0}^t (m+1) = \frac{(t+1)(t+2)}{2} \quad (\text{arithm. Reihe}).$$

Um das FE-Verfahren wie oben beschrieben durchzuführen, benötigt man eine Basis von S_h^t . Hierzu ist die Monomdarstellung auf jedem Element nicht gut geeignet, da die Stetigkeitsbedingung nicht eingearbeitet ist.

Einfach gelingt dies mit Lagrange-Polynomen. Betrachten wir ein einzelnes Dreieck so wählt man bei Grad t die $\frac{(t+1)(t+2)}{2}$ Punkte pro Dreieck wie folgt:



Nun wählt man als Basispolynome $P_0 \dots P_{p_t-1}$, $p_t = \frac{(t+1)(t+2)}{2}$ diejenigen Polynome, die

$$P_i(z_j) = \begin{cases} 1 & i = j \\ 0 & \text{sonst} \end{cases}$$

erfüllen.

Dies ist möglich, da die Polynome in P_t durch Vorgaben an p_t Punkten eindeutig bestimmt sind. Diese Polynome heißen wie im eindimensionalen Fall Lagrange-Polynome, die daraus resultierenden Finite-Elemente-Basisfunktionen deswegen auch Lagrange-Elemente.

Die explizite Darstellung der Lagrange-Polynome vom Grad t auf dem Referenzdreieck lautet:

$$\hat{\psi}_{i,j}^t(\xi, \eta) = \prod_{\alpha=0}^{i-1} \frac{\xi - \alpha/t}{i/t - \alpha/t} \cdot \prod_{\beta=0}^{j-1} \frac{\eta - \beta/t}{j/t - \beta/t} \cdot \prod_{\gamma=i+j+1}^t \frac{\gamma/t - \xi - \eta}{\gamma/t - i/t - j/t} \quad 0 \leq i+j \leq t. \quad (6.1)$$

Für $t = 1$ erhält man demnach

$$\hat{\psi}_{0,0}^1 = 1 - \xi - \eta, \quad \hat{\psi}_{1,0}^1 = \xi, \quad \hat{\psi}_{0,1}^1 = \eta.$$

Die Abbildungen 4, 5 und 6 zeigen die Basispolynome für die Grade 1, 2 und 3 auf dem Referenzdreieck $\hat{\Omega}_P = \{(x, y) \mid x, y \geq 0 \wedge x + y \leq 1\}$. Basispolynome auf dem Referenzelement nennt man auch Shape-Funktionen.

Hat man nun eine konforme Triangulierung wie oben definiert so konstruiert man eine Basis wie folgt:

1. Bestimme in jedem Element die Interpolationspunkte
2. Nummeriere *alle* Interpolationspunkte als z_0, \dots, z_{s-1}

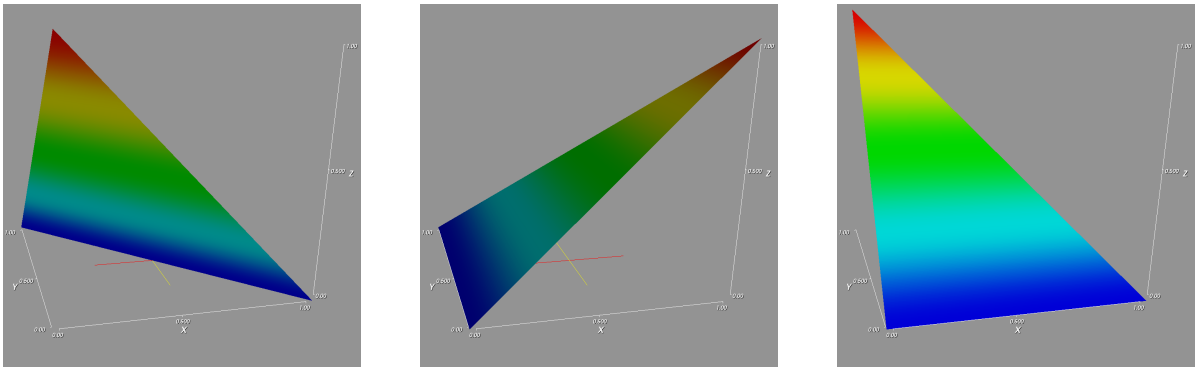


Abbildung 4: Basisfunktionen auf dem Referenzelement für P_1 .

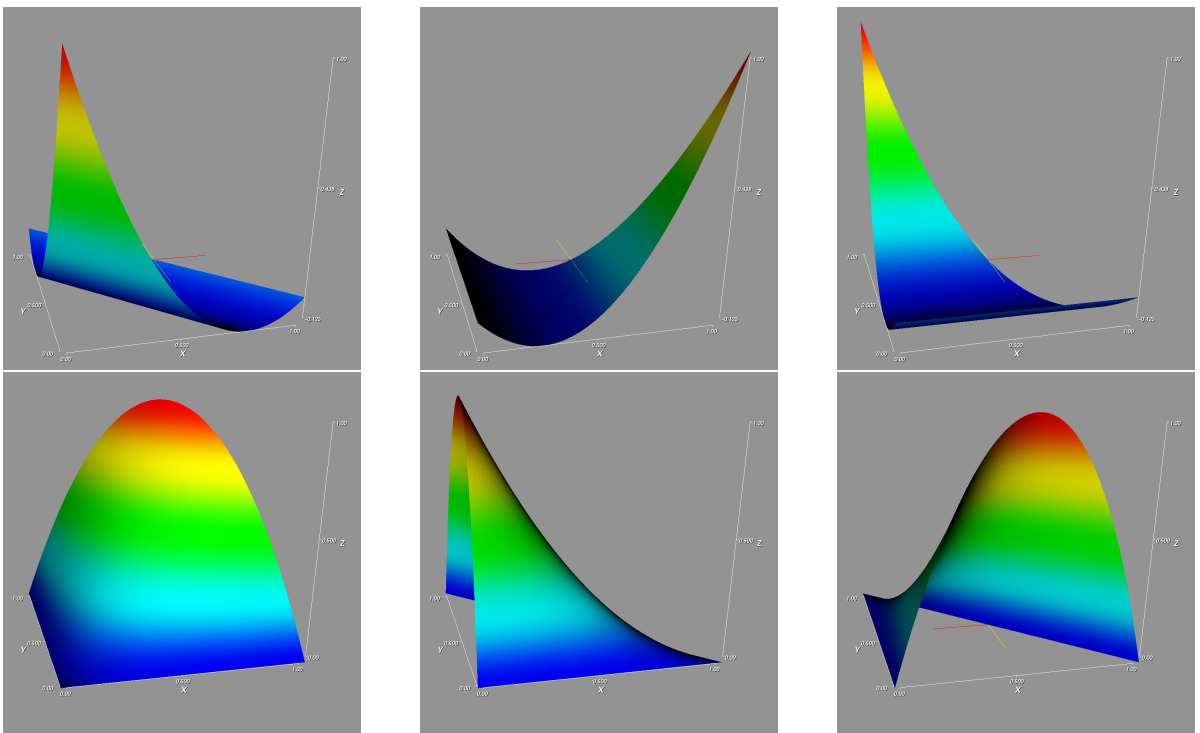


Abbildung 5: Basisfunktionen auf dem Referenzelement für P_2 .

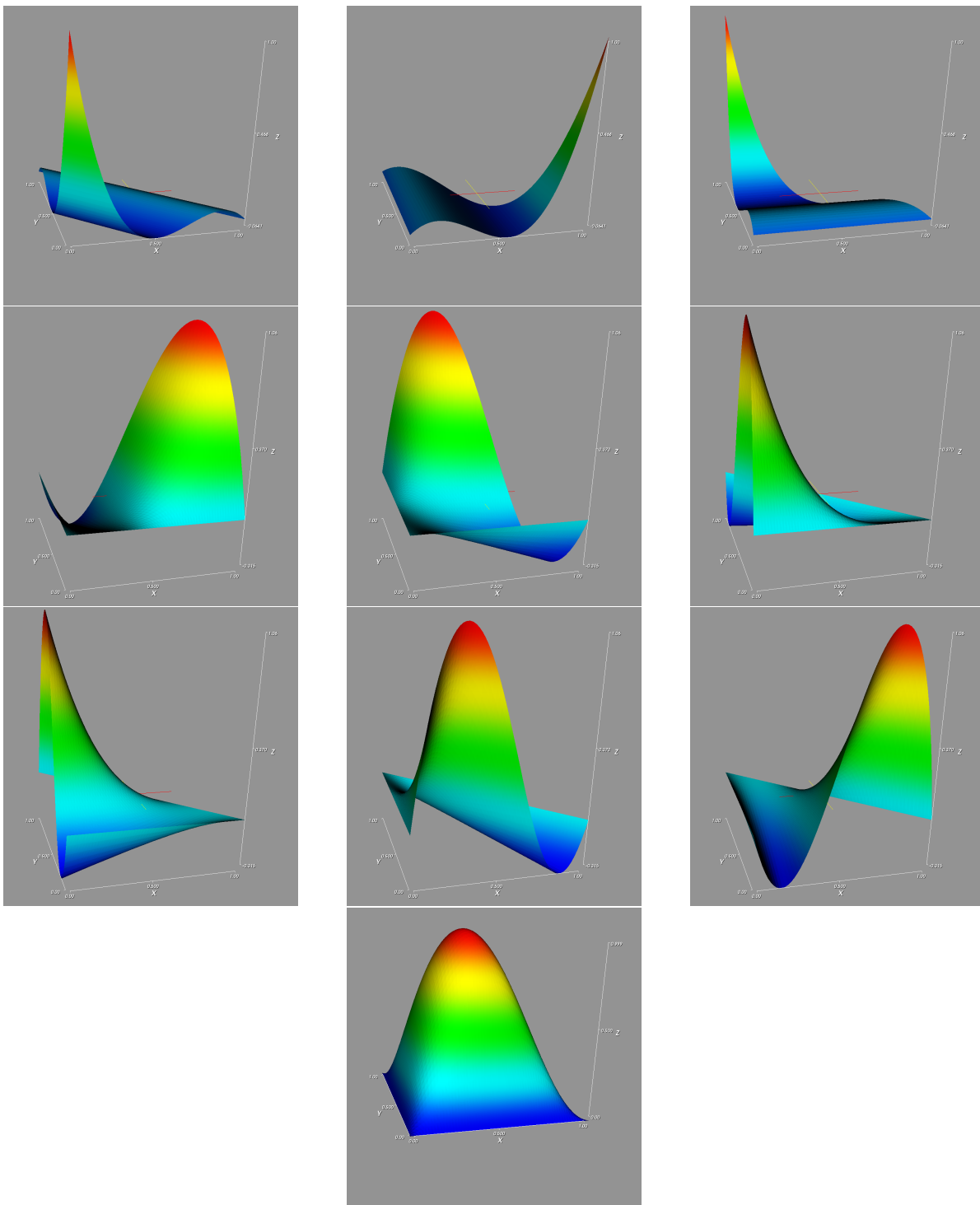
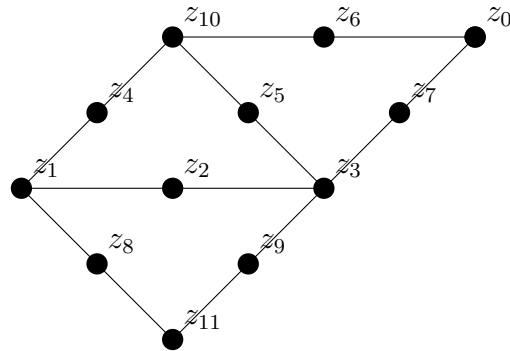


Abbildung 6: Basisfunktionen auf dem Referenzelement für P_3 .



3. $\Psi_h = \{\Psi_0, \dots, \Psi_{s-1}\}$ mit $\Psi_i \in S_h^t$ und $\Psi_i(z_j) = \delta_{ij}$.

Die globale Stetigkeit ergibt sich wie folgt:

- Auf einer Kante liegen genau $t + 1$ Punkte. Diese legen ein Polynom von Grad t in einer Dimension exakt fest. An der expliziten Darstellung der Basispolynome auf dem Referenzdreieck sieht man auch, dass alle Basispolynome, die nicht zu Punkten auf der Kante gehören auf der Kante identisch verschwinden (also nicht nur an den einzelnen Punkten).

Stimmen die beiden Polynome P_l, P_r der beiden angrenzenden Elemente in den $t + 1$ Punkten auf der Kante überein, so sind sie auf der ganzen Kante identisch.

- Stetigkeit in den Ecken der Elemente ist trivial.
- Voraussetzung ist natürlich die *Konformität* der Triangulierung.

Die Abbildung 7 zeigt zwei Basisfunktionen auf einem Gitter für das L-Gebiet.

Schließlich noch ein Hinweis zu den Randbedingung. Bei Dirichlet-Randbedingungen mit dem zugrundeliegenden Sobolevraum $H_0^1(\Omega)$ sind nur die Basisfunktionen zu den Lagrangepunkten $z_i \in \Omega$ erforderlich.

Die Konstruktion lässt sich auf Simplizes (Verallgemeinerung des Dreiecks) in n Raumdimensionen übertragen.

Viereckselemente der Ordnung t

Für Viereckselemente wählt man auf jedem Element die Polynome Q_t . Die Dimension von Q_t ist $q_t = (t + 1)^2$ in zwei Raumdimensionen.

Die entsprechenden Lagrangepunkte sind dann

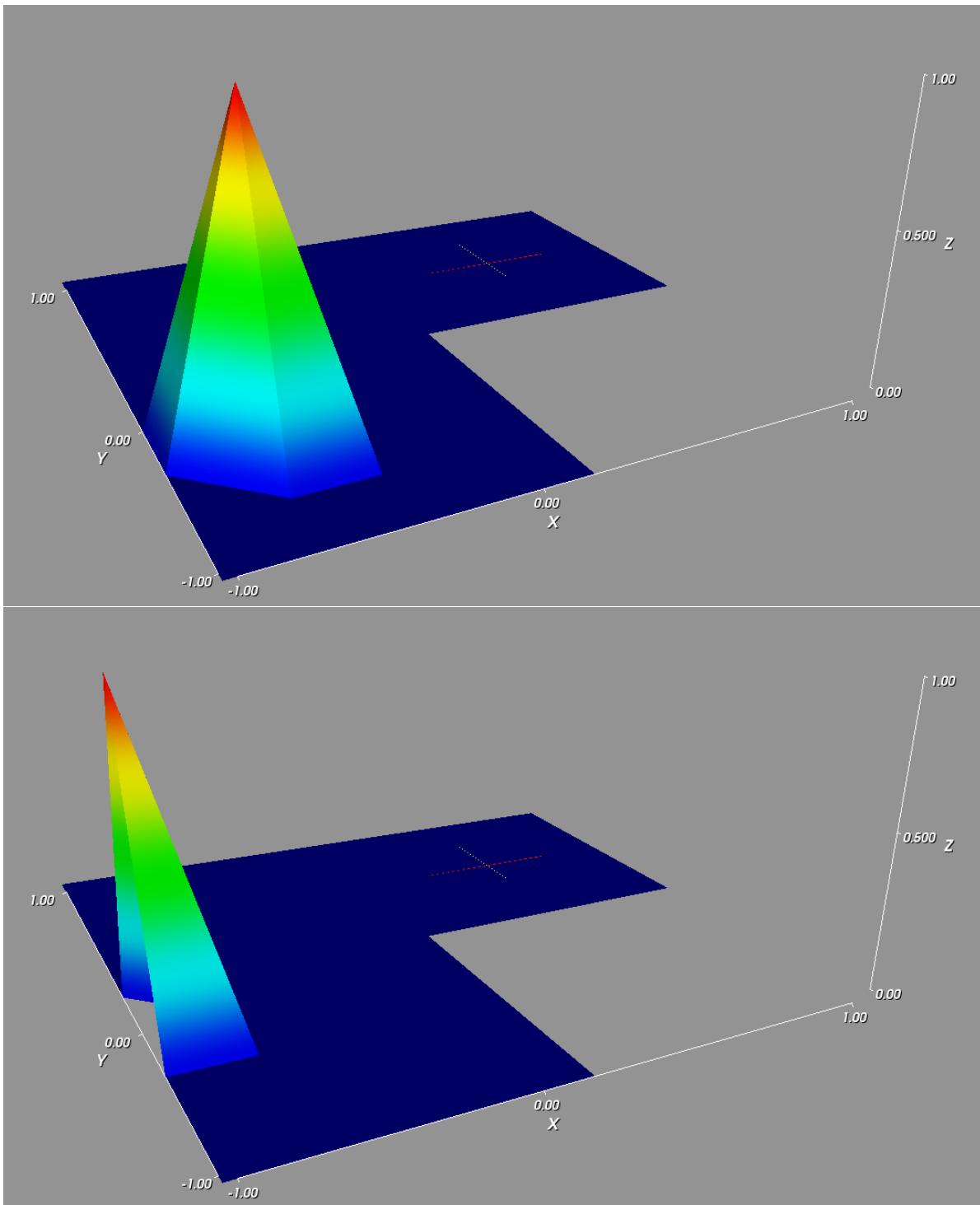
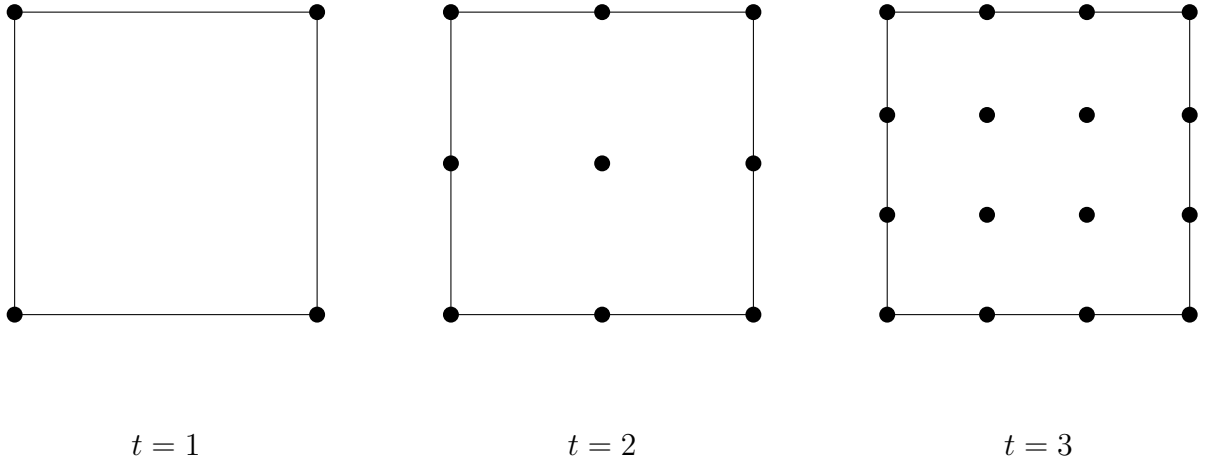


Abbildung 7: Zwei globale P_1 -Basisfunktionen im L-Gebiet.



Die entsprechenden Basisfunktionen auf dem Referenzelement $\hat{\Omega}_Q = \{(x, y) \mid 0 \leq x, y \leq 1\}$ zeigen die Abbildungen 8 für $t = 1$ und 9 für $t = 2$.

Eine explizite Darstellung der Basisfunktionen vom Grad t auf dem Referenzviereck lautet

$$\hat{\psi}_{i,j}^t(\xi, \eta) = \prod_{\alpha=0, \dots, t; \alpha \neq i} \frac{\xi - \alpha/t}{i/t - \alpha/t} \cdot \prod_{\beta=0, \dots, t; \beta \neq j} \frac{\eta - \beta/t}{j/t - \beta/t}, \quad 0 \leq i, j \leq t. \quad (6.2)$$

(Tensorprodukt der eindimensionalen Lagrange-Polynome.

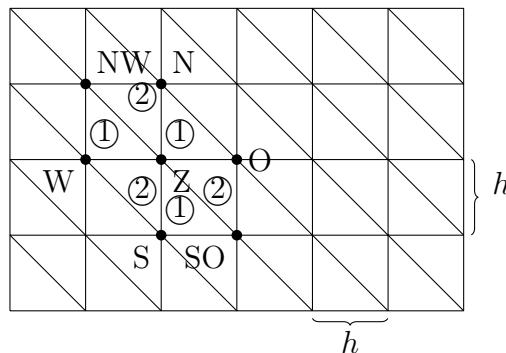
Die Konstruktion lässt sich analog auf Würfel im \mathbb{R}^n verallgemeinern. Dann hat man $q_t = (t + 1)^n$ Lagrange-Punkte pro Element.

6.3 Ein Beispiel in zwei Raumdimensionen

Wir lösen

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

mit P_1 Elementen auf dem Gitter:



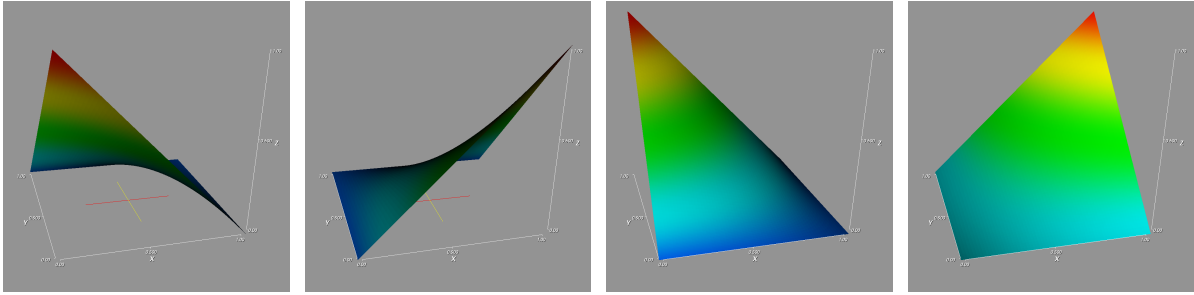


Abbildung 8: Basisfunktionen auf dem Referenzelement für Q_1 .

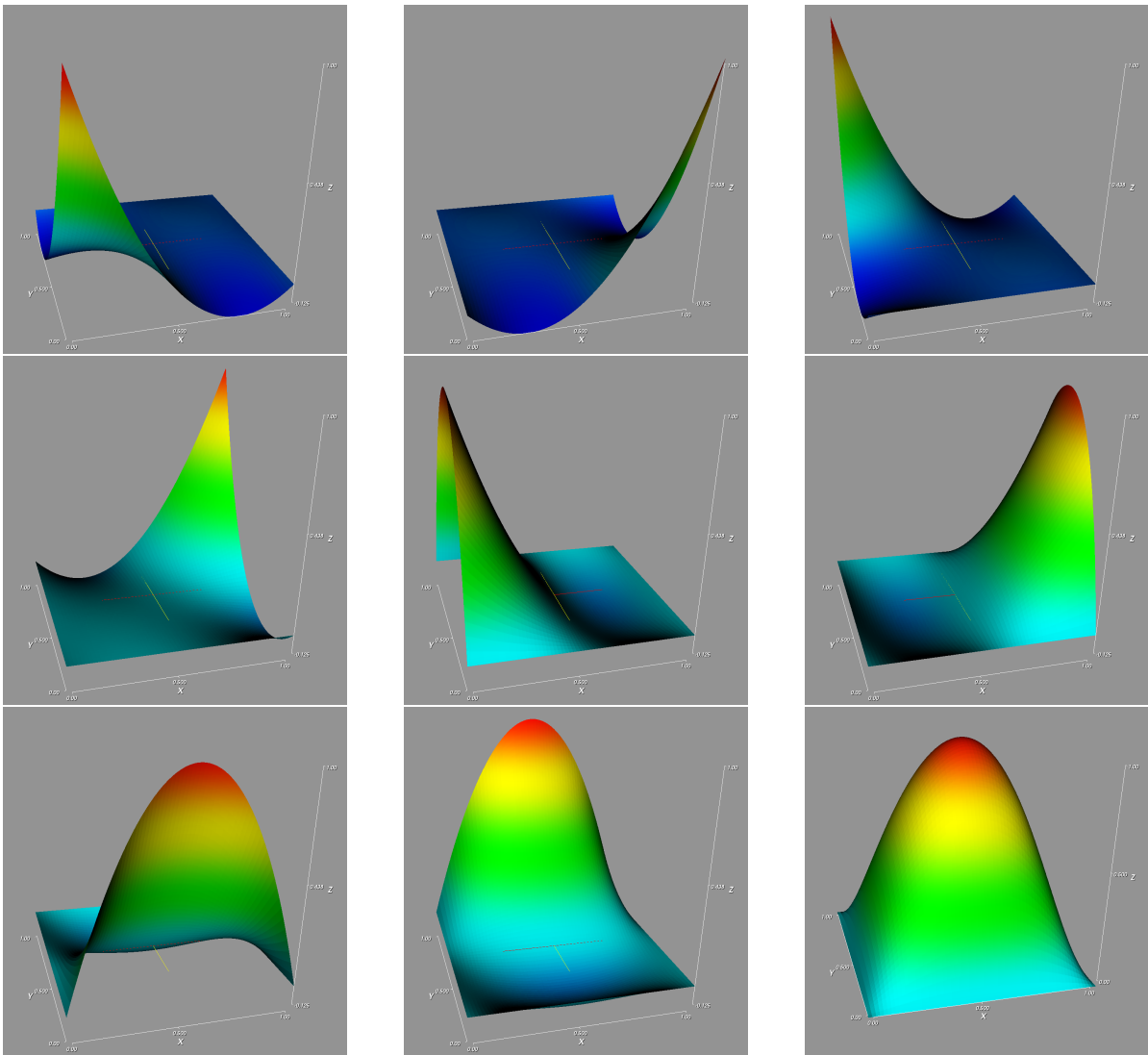

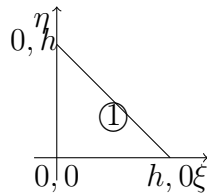


Abbildung 9: Basisfunktionen auf dem Referenzelement für Q_2 .

Die Basisfunktionen sehen an allen (inneren) Knoten gleich aus, dementsprechend sind auch die Zeilen der Steifigkeitsmatrix alle identisch (wenn man die Randknoten mit betrachtet). Es genügt deshalb nur einen Knoten Z zu betrachten. Die Nachbarn dieses Knotens bezeichnen wir relativ als N, O, SO, S, W, NW .

Es gibt zwei Sorten von Dreiecken:  und  die wir getrennt betrachten.

Typ 1 Dreieck Dort haben wir



Nodale Basisfunktionen in ξ, η Koordinaten

$$\text{Knoten 0: } \varphi_0(\xi, \eta) = \frac{h-\xi-\eta}{h} \quad \nabla \varphi_0 = h^{-1} \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$$\text{Knoten 1: } \varphi_1(\xi, \eta) = \frac{\xi}{h} \quad \nabla \varphi_1 = h^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\text{Knoten 2: } \varphi_2(\xi, \eta) = \frac{\eta}{h} \quad \nabla \varphi_2 = h^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Nun berechne für $0 \leq i, j \leq 2$:

$$a_1(\varphi_i, \varphi_j) := \int_{\textcircled{1}} \nabla \varphi_i \cdot \nabla \varphi_j \xi d\eta :$$

$$a_1(\varphi_0, \varphi_0) = \int_0^h \int_0^{h-\xi} h^{-2} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \end{pmatrix} d\xi d\eta = 2h^{-2} \int_0^h \int_0^{h-\xi} 1 d\xi d\eta = 2h^2 \cdot \frac{h^2}{2} = 1$$

entsprechend

$$a_1(\varphi_1, \varphi_1) = \int_{\textcircled{1}} h^{-2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} d\xi d\eta = h^{-2} \cdot \frac{h^2}{2} = \frac{1}{2}$$

$$a_1(\varphi_2, \varphi_2) = \int_{\textcircled{1}} h^{-2} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} d\xi d\eta = \frac{1}{2}$$

$$a_1(\varphi_0, \varphi_1) = \int_{\textcircled{1}} h^{-2} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} d\xi d\eta = -h^2 \cdot \frac{h^2}{2} = -\frac{1}{2}$$

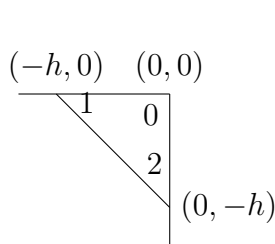
$$A_{\textcircled{1}} = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix}$$

$$a_1(\varphi_0, \varphi_2) = \int_{\textcircled{1}} h^{-2} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} d\xi d\eta = -\frac{1}{2}$$

„lokale Steifigkeitsmatrix“.

$$a_1(\varphi_1, \varphi_2) = \int_{\textcircled{1}} h^{-2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} d\xi d\eta = 0$$

Typ 2 Dreieck Dort haben wir



$$\varphi_0(\xi, \eta) = \frac{h + \xi + \eta}{h} \quad \nabla \varphi_0 = h^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\varphi_1(\xi, \eta) = -\frac{\xi}{h} \quad \nabla \varphi_1 = h^{-1} \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

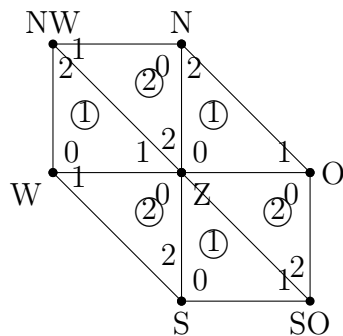
$$\varphi_2(\xi, \eta) = \frac{\eta}{h} \quad \nabla \varphi_2 = h^{-1} \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

Und man rechnet

$$a_2(\varphi_i, \varphi_j) = \int_{\textcircled{2}} \nabla \varphi_i \cdot \nabla \varphi_j d\xi, d\eta \quad A_{\textcircled{2}} = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = A_{\textcircled{1}} =: A$$

Einträge der Steifigkeitsmatrix ergeben sich dann zu

- Beachte *lokale* Nummerierung in den Dreiecken



$$A_{Z,Z} = a(\varphi_Z, \varphi_Z) = 2 \cdot A_{0,0} + 2 \cdot A_{1,1} + 2 \cdot A_{2,2} = 2 + 1 + 1 = 4$$

$$A_{Z,0} = a(\varphi_Z, \varphi_0) = A_{0,1} + A_{1,0} = -\frac{1}{2} - \frac{1}{2} = -1 = A_{Z,W}$$

$$A_{Z,N} = a(\varphi_Z, \varphi_N) = A_{0,2} + A_{2,0} = -\frac{1}{2} - \frac{1}{2} = -1 = A_{Z,S}$$

$$A_{Z,NW} = a(\varphi_Z, \varphi_{NW}) = A_{2,1} + A_{1,2} = 0 + 0 = 0 = A_{Z,SO}$$

$$\Rightarrow \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \text{ wie Finite Differenzen!}$$

Rechte Seite ergibt sich zu

$$b_i = \int_{\Omega} f \cdot \varphi_i dx. \text{ Setze } f = \sum_{k=1}^s f(z_k) \cdot \varphi_k \Rightarrow b_i = \sum_{k=1}^s f_k \int_{\Omega} \varphi_k \varphi_i dx$$

lokal auf einem Dreieck rechnet man nach:

$$\int_{\textcircled{1}, \textcircled{2}} \varphi_i \varphi_j d\xi d\eta = \begin{cases} \frac{h^2}{12} & i = j \\ \frac{h^2}{24} & i \neq j \end{cases} \Rightarrow \int_{\Omega} \varphi_i \varphi_j dx = \begin{cases} \frac{h^2}{2} & i = j \\ \frac{h^2}{12} & i \neq j \\ 0 & \text{sonst} \end{cases} \quad i, j \in T_k$$

„Als Differenzenstern“:

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} u = \frac{h^2}{12} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 6 & 1 \\ 0 & 1 & 1 \end{bmatrix} f$$

6.4 Allgemeiner Aufbau des linearen Gleichungssystems

Wir erläutern nun, wie die Einträge der Steifigkeitsmatrix und des Lastvektors für eine allgemeine Triangulierung berechnet werden können. Dazu benutzt man eine Transformation auf ein Referenzelement.

Gradientenberechnung In der Bilinearform wird über den Gradienten integriert. Betrachten wir zunächst wie man den Gradienten einer Finite-Element-Funktion an einer Stelle im Gitter berechnet.

Sei Ω_t ein Dreieck mit beliebigen Eckpunkten in dem Gitter. $\hat{\Omega}_P = \{(\xi, \eta) \mid \xi, \eta \geq 0 \wedge \xi + \eta \leq 1\}$ ist das Referenzdreieck wie oben bereits eingeführt. Die Abbildung

$$\mu_t : \hat{\Omega}_P \rightarrow \Omega, \quad \mu_t(\xi, \eta) = \begin{pmatrix} \mu_{t,x}(\xi, \eta) \\ \mu_{t,y}(\xi, \eta) \end{pmatrix}$$

ordnet jedem Punkt im Referenzelement einen Punkt im Dreieck zu. Im Referenzelement werden die Koordinaten mit (ξ, η) bezeichnet und im allgemeinen Dreieck mit (x, y) . Die Abbildung μ_t sei zudem einmal stetig differenzierbar.

Nun sei eine Funktion $u : \Omega_t \rightarrow \mathbb{R}$ auf dem allgemeinen Dreieck gegeben. Wir setzen

$$\hat{u}(\xi, \eta) = u(\mu_t(\xi, \eta)) = u(\mu_{t,x}(\xi, \eta), \mu_{t,y}(\xi, \eta))$$

und erhalten mittels der Kettenregel:

$$\begin{aligned} \frac{\partial \hat{u}}{\partial \xi}(\xi, \eta) &= \frac{\partial u}{\partial x}(\mu_t(\xi, \eta)) \frac{\partial \mu_{t,x}}{\partial \xi}(\xi, \eta) + \frac{\partial u}{\partial y}(\mu_t(\xi, \eta)) \frac{\partial \mu_{t,y}}{\partial \xi}(\xi, \eta) \\ \frac{\partial \hat{u}}{\partial \eta}(\xi, \eta) &= \frac{\partial u}{\partial x}(\mu_t(\xi, \eta)) \frac{\partial \mu_{t,x}}{\partial \eta}(\xi, \eta) + \frac{\partial u}{\partial y}(\mu_t(\xi, \eta)) \frac{\partial \mu_{t,y}}{\partial \eta}(\xi, \eta) \end{aligned}$$

was wir in Vektorform schreiben können als

$$\begin{bmatrix} \frac{\partial \hat{u}}{\partial \xi}(\xi, \eta) \\ \frac{\partial \hat{u}}{\partial \eta}(\xi, \eta) \end{bmatrix} = \begin{bmatrix} \frac{\partial \mu_{t,x}}{\partial \xi}(\xi, \eta) & \frac{\partial \mu_{t,y}}{\partial \xi}(\xi, \eta) \\ \frac{\partial \mu_{t,x}}{\partial \eta}(\xi, \eta) & \frac{\partial \mu_{t,y}}{\partial \eta}(\xi, \eta) \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial x}(\mu_t(\xi, \eta)) \\ \frac{\partial u}{\partial y}(\mu_t(\xi, \eta)) \end{bmatrix}$$

bzw. noch kompakter

$$\hat{\nabla} \hat{u}(\xi, \eta) = J_{\mu_t}^T(\xi, \eta) \nabla u(\mu_t(\xi, \eta)).$$

Hierbei ist $\hat{\nabla}$ der Gradient bezüglich der Koordinaten ξ, η , J_{μ_t} die Jacobi-Matrix bzw. Funktionalmatrix der Transformation μ und J^T bezeichnet die Transponierte einer Matrix.

Die letzte Beziehung können wir nun auflösen und erhalten

$$\nabla u(\mu_t(\xi, \eta)) = J_{\mu_t}^{-T}(\xi, \eta) \hat{\nabla} \hat{u}(\xi, \eta). \quad (6.3)$$

Wir verwenden diese Beziehung nicht für allgemeine Funktionen u sondern nur für die Basisfunktionen ψ_i . Sind die Basisfunktionen auf dem Referenzelement bekannt, d. h. $\hat{\psi}$ dann lässt sich deren Gradient auf dem allgemeinen Element aus dem Gradienten auf dem Referenzelement und der Jacobimatrix der Transformation berechnen.

Integration Die Integration einer beliebigen Funktion $f : \Omega_t \rightarrow \mathbb{R}$ führt man zunächst mittels dem Transformationssatz für Integrale auf das Referenzelement zurück:

$$\int_{\Omega_t} f(x, y) dx dy = \int_{\Omega_P} f(\mu_t(\xi, \eta)) |\det J_{\mu_t}(\xi, \eta)| d\xi d\eta. \quad (6.4)$$

Das Integral auf dem Referenzelement wird nun numerisch mit einer Quadraturformel genügend hoher Ordnung berechnet:

$$\int_{\Omega_P} f(\mu_t(\xi, \eta)) |\det J_{\mu_t}(\xi, \eta)| d\xi d\eta = \sum_{k=1}^K w_k f(\mu_t(\xi_k, \eta_k)) |\det J_{\mu_t}(\xi_k, \eta_k)| + \text{Fehler}. \quad (6.5)$$

Hierbei sind die w_k die Gewichte und die (ξ_k, η_k) die Quadraturpunkte.

In unserer Anwendung ergibt sich also speziell für die Elemente der Steifigkeitsmatrix

$$\begin{aligned} a_{ij} &= \int_{\Omega} \nabla \psi_j \cdot \nabla \psi_i dx dy = \sum_{t \in \mathcal{T}_h} \int_{\Omega_t} \nabla \psi_j \cdot \nabla \psi_i dx dy \\ &= \sum_{t \in \mathcal{T}_h} \sum_{k=1}^K w_k \left(J_{\mu_t}^{-T}(\xi_k, \eta_k) \hat{\nabla} \hat{\psi}_j(\xi_k, \eta_k) \right) \cdot \left(J_{\mu_t}^{-T}(\xi_k, \eta_k) \hat{\nabla} \hat{\psi}_i(\xi_k, \eta_k) \right) |\det J_{\mu_t}(\xi_k, \eta_k)|. \end{aligned}$$

Im Falle von Dreiecken ist die Abbildung μ affin-linear und damit die Jacobimatrix eine konstante Matrix unabhängig von ξ und η . Sei etwa ein Dreieck t mit den Eckpunkten $(x_0, y_0)^T$, $(x_1, y_1)^T$ und $(x_2, y_2)^T$ gegeben, so lautet die Abbildung μ_t wie folgt:

$$\mu_t(\xi, \eta) = (1 - \xi - \eta) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \xi \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \eta \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \quad (6.6)$$

$$= \begin{pmatrix} x_1 - x_0 & x_2 - x_0 \\ y_1 - y_0 & y_2 - y_0 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} + \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \quad (6.7)$$

Bei Polynomgrad t sind die Komponenten des Gradienten auf dem Referenzelement Polynome vom Grad $t - 1$ und somit sollte die Quadraturformel exakt für Polynome bis zum Grad $(t - 1)^2$ sein.

7 Approximationssätze

Nach dem Lemma von Céa hängt der FE-Fehler davon ab, wie gut Funktionen $v_h \in S_h$ eine gegebene Funktion $v \in H^k(\Omega)$ approximieren können. Dieser Frage widmen wir uns in diesem Abschnitt.

7.1 Bramble-Hilbert Lemma

Bezeichnung 7.1 (Elementweise Normen). Wir werden die Fehlerabschätzungen elementweise herleiten. Dazu setzen wir

$$\|v\|_{m,h} := \sqrt{\sum_{T_j \in \mathcal{T}_h} \|v\|_{m,T_j}^2} \quad \text{mit} \quad \|v\|_{m,\omega}^2 = \sum_{|\alpha| \leq m} \int_{\omega} (\partial^\alpha v)^2 dx.$$

für eine gegebene Triangulierung $\mathcal{T}_h = \{T_1, \dots, T_m\}$. Für $v \in H^m(\Omega)$ gilt $\|v\|_{m,\Omega} = \|v\|_{m,h}$. \square

Die Lagrange-Interpolation von Polynomen benutzt die punktweise Auswertung, was stetige Funktionen voraussetzt. Wann eine Sobolev-Funktion $v \in H^k(\Omega)$ stetig ist beantwortet der folgende Satz.

Satz 7.2 (Sobolev'scher Einbettungssatz). Sei $\Omega \subset \mathbb{R}^n$ ein Gebiet mit Lipschitz-stetigen Rand und Ω erfülle die Kegelbedingung. Für $k > \frac{n}{2}$ gilt $H^k(\Omega) \subset C^0(\overline{\Omega})$. Darüber hinaus ist die Einbettung stetig, d.h. $\exists c \in \mathbb{R}$, so dass $\|v\|_{H^k(\Omega)} \leq c \|v\|_{C^0(\overline{\Omega})}$.

Beweis: [RR93, S. 215]. \square

Damit erhalten wir die folgenden ganzzahligen k in Abhängigkeit von n :

$$\left. \begin{array}{r} n \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \right\} \begin{array}{l} k \geq \\ 1 \\ 2 \\ 2 \\ 3 \end{array} \Rightarrow \text{für } n \leq 3 \text{ sind } H^2(\Omega)\text{-Funktionen stetig!}$$

Zunächst untersuchen wir die Lagrange-Interpolation näher und zeigen folgenden

Hilfssatz 7.3. Sei $\Omega \subset \mathbb{R}^n$, ein Gebiet mit Lipschitz-stetigem Rand, welches eine Kegelbedingung erfüllt. Sei weiter

- $\mathbb{N} \ni t > \frac{n}{2}$
- $z_1, \dots, z_s \in \overline{\Omega}$, s Punkte mittels denen die Lagrange Interpolation $I : H^t \rightarrow \mathcal{P}_{t-1}$ durch Polynome vom Grad $t - 1$ wohldefiniert ist.

Dann gilt mit einem $c = c(\Omega, z_1, \dots, z_s)$

$$\|u - Iu\|_t \leq c|u|_t \text{ f\"ur alle } u \in H^t(\Omega).$$

Beweis: $t > \frac{n}{2}$ stellt sicher, dass nach Satz 7.2 die punktweise Auswertung sinnvoll ist. F\"ur $n \leq 3$ ist also $t \geq 2$, mithin Polynomgrad 1 eine m\"oglich Wahl.

Definiere die Norm

$$|||v||| := |v|_t + \sum_{i=1}^s |v(z_i)|.$$

Wir zeigen unten, dass $\|v\|_t \leq c|||v|||$. Mit dieser Ungleichung folgert man dann

$$\begin{aligned} \|u - Iu\|_t &\leq c|||u - Iu||| = c\{|u - Iu|_t + \underbrace{\sum_{i=1}^s |u(z_i) - (Iu)(z_i)|}_{= 0 \text{ wg. Lagrange-I.}}\} \\ &= c|u - Iu|_t \leq c|u|_t + c \underbrace{|Iu|_t}_{= 0 \text{ da Polynomgrad } t-1} = c|u|_t. \end{aligned}$$

Nun zur Ungleichung. Angenommen es gebe kein $c \in \mathbb{R}$ so dass $\|v\|_t \leq c|||v|||$ f\"ur alle $v \in H^t(\Omega)$. Dann gibt es eine Folge $(v_k) \in H^t(\Omega)$ mit

$$\|v_k\|_t = 1, \quad |||v_k||| \leq \frac{1}{k}, \quad k = 1, 2, \dots$$

(denn damit $1 \leq c \frac{1}{k}$ gilt, muss $c \geq k$ sein, also $c \rightarrow \infty$).

Aus der kompakten Einbettung von $H^t(\Omega)$ in $H^{t-1}(\Omega)$ folgt, dass eine Teilfolge der (v_k) in $H^{t-1}(\Omega)$ konvergiert. O.B.d.A. k\"onnen wir annehmen, dass dies bereits die ganze Folge ist.

Konvergenz in $H^{t-1}(\Omega)$ bedeutet insbesondere, dass (v_k) eine Cauchy-Folge in $H^{t-1}(\Omega)$ ist.

Nach unserer Annahme gilt $|||v_k||| = |v_k|_t + \sum |v_k(z_i)| \leq \frac{1}{k} \rightarrow 0$ also insbesondere auch $|v_k|_t \rightarrow 0$. Wegen

$$\begin{aligned} \|v_k - v_l\|_t^2 &= \|v_k - v_l\|_{t-1}^2 + |v_k - v_l|_t^2 \leq \underbrace{\|v_k - v_l\|_{t-1}^2}_{\rightarrow 0 \text{ wg. Cauchy-F. in } H^{t-1}} + \underbrace{(|v_k|_t + |v_l|_t)^2}_{\rightarrow 0 \text{ wg. } |||v_k||| \rightarrow 0} \end{aligned}$$

liegt sogar Konvergenz gegen ein $v^* \in H^t(\Omega)$ vor. Wegen Stetigkeit der Normen haben wir

- $(v_k) \rightarrow v^* \in H^t(\Omega)$,
- $\|v^*\|_t = 1$,
- $|||v^*||| = 0$, insbesondere hei\ss t das $|v^*|_t = 0$ und $v^*(z_i) = 0$ f\"ur alle i .

Wegen $|v^*|_t = 0$ ist $v^* \in P_{t-1}$ ein Polynom vom Grad $t-1$. Wegen $v^*(z_i) = 0, \forall i = 1 \dots s$, ist v^* das Nullpolynom! Dies ist ein Widerspruch zu $\|v^*\|_t = 1$. Die Annahme ist somit falsch. \square

Die Anwendung der Lagrange-Interpolation ist nicht zwingend. Der folgende Satz verallgemeinert die Aussage des Hilfssatzes auf beliebige Interpolationsarten.

Lemma 7.4 (Bramble-Hilbert). $\Omega \subset \mathbb{R}^n$ erfülle die Bedingung von Hilfssatz 7.3 und es sei $t > \frac{n}{2}$. $L : H^t(\Omega) \rightarrow Y$ sei eine beschränkte, lineare Abbildung in den normierten Raum Y mit $\|L\| = \sup_{v \neq 0} \frac{\|Lv\|_Y}{\|v\|_t}$. Weiter sei $\mathcal{P}_{t-1} \subseteq \ker L$, d.h. $L(v) = 0 \quad \forall v \in \mathcal{P}_{t-1}$ (z.B. oben $L = Id - I$). Dann gilt

$$\|Lv\|_Y \leq c|v|_t \quad \forall v \in H^t(\Omega).$$

Beweis:

$$\begin{aligned} \|Lv\|_Y &= \|Lv - LIv\| = \|L(v - Iv)\| \leq \|L\| \|v - Iv\|_t \\ &\leq c\|L\| |v|_t. \end{aligned}$$

□

7.2 Approximationsatz

Damit können wir den allgemeinen Approximationsatz formulieren.

Satz 7.5 (Approximation mit Polynomen). Sei $t > \frac{n}{2}$ und \mathcal{T}_h eine quasiuniforme Triangulierung von $\Omega \subset \mathbb{R}^n$. Ω erfülle die Voraussetzungen für das Bramble-Hilbert Lemma. Dann gilt für die stückweise Interpolation durch Polynome vom Grad $t - 1$:

$$\|u - I_h u\|_{m,h} \leq ch^{t-m} |u|_{t,\Omega} \quad \text{für } u \in H^t(\Omega) \text{ und } 0 \leq m \leq t.$$

Mit $I_h : H^t(\Omega) \rightarrow S_h$ wird hier die Lagrange-Interpolation in den Finite-Element-Raum bezeichnet.

Beweis: Zunächst beweisen wir einen Spezialfall für sehr einfache Gitter. Im nächsten Abschnitt werden wir dann einen allgemeinen Beweis liefern, der jedoch höheren technischen Aufwand erfordert. Jedes $T_j \in \mathcal{T}_h$ sei ein skaliertes und eventuell verschobenes Abbild von einem von zwei Referenzelementen

$$\hat{T}_1 = \{(\hat{x}_1, \hat{x}_2) \in \mathbb{R}^2 \mid 0 \leq \hat{x}_1, \hat{x}_2 \leq 1 \wedge \hat{x}_1 + \hat{x}_2 \leq 1\}$$

oder

$$\hat{T}_2 = \{(\hat{x}_1, \hat{x}_2) \in \mathbb{R}^2 \mid 0 \leq \hat{x}_1, \hat{x}_2 \leq 1 \wedge \hat{x}_1 + \hat{x}_2 \geq 1\}.$$

Die Abbildung μ_T vom Referenzelement (egal welches!) auf $T \in \mathcal{T}_h$ ist gegeben durch

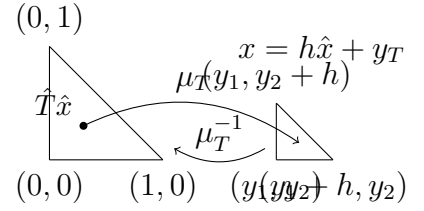
$$\mu_T(\hat{x}) = h\hat{x} + y_T.$$

Für die Rücktransformation gilt entsprechend

$$\mu_T^{-1}(x) = h^{-1}(x - y_T).$$

Sei $v \in H^t(T_h)$ gegeben und setze $\hat{v}(\hat{x}) = v(\mu_T(\hat{x}))$. Dann gilt wegen Kettenregel:

$$\hat{\partial}^\alpha \hat{v}(\hat{x}) = h^{|\alpha|} \partial^\alpha v(\mu_T(\hat{x})).$$



Damit erhält man

$$\begin{aligned} |\hat{v}|_{l, \hat{T}}^2 &= \sum_{|\alpha|=l} \int_{\hat{T}} (\partial^\alpha \hat{v}(\hat{x}))^2 d\hat{x} \\ &= \sum_{|\alpha|=l} \int_{\hat{T}} h^{2|\alpha|} (\partial^\alpha v(\mu_T(\hat{x})))^2 d\hat{x} \\ &= h^{2l} \sum_{|\alpha|=l} \int_{T_h} (\partial^\alpha v(\mu_T(\mu_T^{-1}(x))))^2 h^{-n} dx \\ &= h^{2l-n} |v|_{l, T_h}^2. \end{aligned}$$

Mit dem selben Argument erhält man für die andere Richtung $T_h \rightarrow T$

$$|v|_{l, T_h}^2 \leq h^{-2l+n} |\hat{v}|_{l, \hat{T}}^2$$

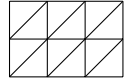
Mit diesen sogenannten Transformationsformeln folgt dann für $0 \leq l \leq t$:

$$\begin{aligned} |u - I_h u|_{l, T_h}^2 &= h^{-2l+n} |\hat{u} - I_h \hat{u}|_{l, \hat{T}}^2 && (T_h \rightarrow \hat{T}) \\ &\leq h^{-2l+n} \|\hat{u} - I_h \hat{u}\|_{l, \hat{T}}^2 && (\text{zur Norm ergänzen}) \\ &\leq h^{-2l+n} c |\hat{u}|_{t, \hat{T}}^2 && (\text{Bramble-Hilbert Lemma, } Y = H^l(\hat{T})) \\ &\leq c h^{-2l+n} h^{2t-n} |u|_{t, T_h}^2 && (\hat{T} \rightarrow T_h) \\ &= c h^{2(t-l)} |u|_{t, T_h}^2 \end{aligned}$$

jetzt über die $l = 0, \dots, m \leq t$ summieren:

$$\begin{aligned} \|u - I_h u\|_{m, T_h}^2 &= \sum_{l=0}^m |u - I_h u|_{l, T_h}^2 \\ &\leq \sum_{l=0}^m c h^{2(t-l)} |u|_{t, T_h}^2 \\ &= c |u|_{t, T_h}^2 h^{2t} \sum_{l=0}^m h^{-2l} \\ &= c |u|_{t, T_h}^2 h^{2t} (1 + h^{-2} + h^{-4} + \dots + h^{-2m}) \\ &= c |u|_{t, T_h}^2 h^{2t} h^{-2m} \underbrace{(h^{2m} + \dots + h^4 + h^2 + 1)}_{\leq C \text{ für alle } h \leq h_0 < 1 \text{ (geom. Reihe)}} \\ &= c h^{2(t-m)} |u|_{t, T_h}^2 \end{aligned}$$

Und schließlich über alle Dreiecke summieren. Alle Dreiecke sind skalierte und verschobene \hat{T}_1 oder \hat{T}_2 , also etwa ein Gitter der Form



also etwa ein Gitter der Form

$$\begin{aligned} \|u - I_h u\|_{m,h}^2 &= \sum_{T_h \in \mathcal{T}_h} \|u - I_h u\|_{m,T_h}^2 \leq \sum_{T_h \in \mathcal{T}_h} ch^{2(t-m)} |u|_{t,T_h}^2 \\ &= ch^{2(t-m)} \sum_{T_h \in \mathcal{T}_h} |u|_{t,T_h}^2 \end{aligned}$$

Wurzelziehen zeigt dann die Aussage

$$\|u - I_h u\|_{m,h} \leq ch^{t-m} |u|_{t,\Omega}.$$

□

7.3 Transformationssatz für allgemeine Dreiecke

Die weiteren Ausführungen dienen dazu den Approximationssatz für Polynome vom Grad $t - 1$ auf allgemeinen simplizialen Gittern im \mathbb{R}^n zu zeigen. Das Prinzip (sog. Skalierungsargument) ist dabei das gleiche aber die Transformationsformeln sind schwieriger zu zeigen.

Definition 7.6 (Tensorprodukt von Vektoren). Gegeben seien m Vektoren $y_k \in \mathbb{R}^{n_k}$ (die Dimension n_k darf unterschiedlich sein für jeden Vektor). Dann ist

$$\bigotimes_{k=1}^m y_k \in \mathbb{R}^N \quad \text{mit } N = \prod_{k=1}^m n_k$$

und für die Komponenten mit den Indizes aus $I_m = \{(i_1, \dots, i_m) \mid i_k \in \{1, \dots, n_k\}\}$ gilt

$$\left(\bigotimes_{k=1}^m y_k \right)_{i_1, \dots, i_m} = \prod_{k=1}^m (y_k)_{i_k}.$$

□

Für zwei Vektoren x, y gilt $(x \otimes y)_{ij} = x_i y_j$, d. h. alle Komponenten der Matrix xy^T werden in einem Vektor angeordnet.

Für die Norm eines Tensorproduktvektors gilt:

$$\begin{aligned} \left\| \bigotimes_{k=1}^m y_k \right\|_{\mathbb{R}^N}^2 &= \sum_{i_1, \dots, i_m \in I_m} \left| \prod_{k=1}^m (y_k)_{i_k} \right|^2 = \sum_{i_1=1}^{n_1} \dots \sum_{i_m=1}^{n_m} (y_1)_{i_1}^2 \cdot \dots \cdot (y_m)_{i_m}^2 \\ &= (y_1)_1^2 \sum_{i_2=1}^{n_2} \dots \sum_{i_m=1}^{n_m} (y_2)_{i_2}^2 \cdot \dots \cdot (y_m)_{i_m}^2 + \dots + (y_1)_{n_1}^2 \sum_{i_2=1}^{n_2} \dots \sum_{i_m=1}^{n_m} (y_2)_{i_2}^2 \cdot \dots \cdot (y_m)_{i_m}^2 \\ &= \left(\sum_{i_2=1}^{n_2} \dots \sum_{i_m=1}^{n_m} (y_2)_{i_2}^2 \cdot \dots \cdot (y_m)_{i_m}^2 \right) \left(\sum_{i_1=1}^{n_1} (y_1)_{i_1}^2 \right) = \prod_{k=1}^m \|y_k\|^2. \end{aligned}$$

Somit gilt also

$$\left\| \bigotimes_{k=1}^m y_k \right\|_{\mathbb{R}^N} = \prod_{k=1}^m \|y_k\|. \quad (7.1)$$

Ableitungen als Multilinearform

Es sei $v \in H^m(\Omega)$ (bzw. $C^m(\overline{\Omega})$) und $\Omega \subset \mathbb{R}^n$. Zu einem gegebenen Vektor $y \in \mathbb{R}^n$ definiere den „Differentialoperator“

$$L[y] = \sum_{i=1}^n (y)_i \partial_i.$$

Es ist also

$$L[y]v(x) = \sum_{i=1}^n (y)_i \partial_i v(x).$$

Für $y = e_i$ erhalten wir etwa

$$L[e_i]v(x) = \partial_i v(x).$$

Dies verallgemeinern wir nun auf Ableitungen der Ordnung m im \mathbb{R}^n .

Definition 7.7. Gegeben seien m Vektoren $y_1, \dots, y_m \in \mathbb{R}^n$. Dann ist

$$L[y_1, \dots, y_m] := \prod_{k=1}^m \left(\sum_{i_k=1}^n (y_k)_{i_k} \partial_{i_k} \right).$$

□

Sei etwa $m = 2$ und $n = 2$ dann ist

$$\begin{aligned} L[y_1, y_2]v(x) &= ((y_1)_1 \partial_1 + (y_1)_2 \partial_2) ((y_2)_1 \partial_1 + (y_2)_2 \partial_2) v(x) \\ &= (y_1)_1 (y_2)_1 \partial_1 \partial_1 v + (y_1)_1 (y_2)_2 \partial_1 \partial_2 v + \dots \end{aligned}$$

Wieder gilt mit $y_k = e_{i_k}$ dass

$$L[e_{i_1}, \dots, e_{i_m}]v(x) = \partial_{i_1} \dots \partial_{i_m} v(x),$$

wir können also beliebige Ableitungen der Ordnung m mit diesem Konzept darstellen.

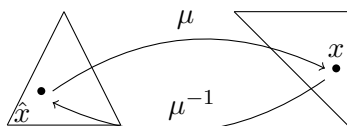
Schließlich gilt die folgende Abschätzung:

$$\begin{aligned} |L[y_1, \dots, y_m]v(x)| &= \left| \left[\prod_{k=1}^m \left(\sum_{i_k=1}^n (y_k)_{i_k} \partial_{i_k} \right) \right] v(x) \right| \\ &= |((y_1)_1 \partial_1 + \dots + (y_1)_n \partial_n) \dots ((y_m)_1 \partial_1 + \dots + (y_m)_n \partial_n) v(x)| \\ &= \left| \sum_{i_1=1}^n \dots \sum_{i_m=1}^n (y_1)_{i_1} \dots (y_m)_{i_m} \partial_{i_1} \partial_{i_2} \dots \partial_{i_m} v(x) \right| \\ &\leq \left\| \bigotimes_{k=1}^m y_k \right\| \|D^m v(x)\| = \|D^m v(x)\| \prod_{k=1}^m \|y_k\| \end{aligned} \quad (7.2)$$

Hierbei ist $D^m v(x) \in \mathbb{R}^{n^m}$ die Vektorfunktion *aller* partiellen Ableitungen $\partial_{i_1} \dots \partial_{i_m} v(x)$ der Ordnung m für $(i_1, \dots, i_m) \in I_m$ (Beachte, dass jede Permutation von Indizes in $D^m v(x)$ vorkommt obwohl die Ableitungen identisch sind). In der Abschätzung wurde nur Cauchy-Schwarz im \mathbb{R}^{n^m} verwendet.

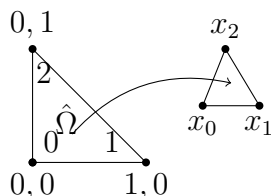
Gebietstransformation

Wir betrachten die affin lineare Transformation, die zwei simpliziale Elemente aufeinander abbildet:



$$\begin{aligned} \mu : \mathbb{R}^n &\rightarrow \mathbb{R}^n, \\ \mu(\hat{x}) &= x_0 + Bx, \quad B \text{ nicht singulär.} \end{aligned}$$

Beispiel:



Dann ist

$$\begin{aligned} \mu(\hat{x}) &= x_0 + (x_1 - x_0)(\hat{x})_1 + (x_2 - x_0)(\hat{x})_2 \\ &= x_0 + \underbrace{\begin{pmatrix} (x_1 - x_0)_1 & (x_2 - x_0)_1 \\ (x_1 - x_0)_2 & (x_2 - x_0)_2 \end{pmatrix}}_B \hat{x}. \end{aligned}$$

- B ist i. allg. nicht symmetrisch.
- B ist nicht singulär falls $x_1 - x_0, x_2 - x_0$ linear unabhängig.

Für die Transformation von Integralen auf \hat{T} bzw T gelten die Formeln;

$$\int_T v(x) dx = \int_{\hat{T}} v(\mu(\hat{x})) |\det B| d\hat{x} \quad (7.3)$$

$$\int_{\hat{T}} \hat{v}(\hat{x}) d\hat{x} = \int_T \hat{v}(\mu^{-1}(x)) |\det B^{-1}| dx \quad (7.4)$$

Die Spektralnorm der Matrizen $\|B\|, \|B^{-1}\|$ kann man folgendermaßen abschätzen. Es seien $\hat{\rho}, \hat{r}$: Inkreis- und Umkreisdurchmesser von \hat{T} , sowie ρ und r der Inkreis- bzw. Umkreisdurchmesser von T .

Damit ist

$$\|B\| = \sup_{x \neq 0} \frac{\|Bx\|}{\|x\|} = \sup_{\|x\|=\hat{\rho}} \frac{\|Bx\|}{\hat{\rho}} \leq \frac{r}{\hat{\rho}} \leq \frac{2h}{\hat{\rho}},$$

da eine Strecke in \hat{T} in dem Umkreis von T abgebildet wird. Ebenso gilt für die Rücktransformation

$$\|B^{-1}\| = \sup_{\|x\|=\rho} \frac{\|B^{-1}x\|}{\rho} \leq \frac{\hat{r}}{\rho} \leq \frac{\hat{r}K}{2h},$$

mit K der Quasiuniformitätskonstanten. Also zusammen

$$\text{cond}_2(B) = \|B\| \|B^{-1}\| \leq \frac{r\hat{r}}{\hat{\rho}\rho} \leq K \frac{\hat{r}}{\hat{\rho}}. \quad (7.5)$$

\hat{r} und $\hat{\rho}$ können für das jeweilige Referenzelement bestimmt werden.

Kettenregel

Gegeben sei eine Transformation $\mu : \hat{\Omega} \rightarrow \Omega$ und eine Funktion $v : \Omega \rightarrow \mathbb{R}$, dann definieren wir $\hat{v} : \hat{\Omega} \rightarrow \mathbb{R}$ als $\hat{v}(\hat{x}) := v(\mu(\hat{x}))$.

Wir verwenden die Bezeichnung $\hat{\partial}_i$ für die Differenziation bezüglich der i -ten Komponente von \hat{x} (also auf dem Referenzelement). ∂_i bezeichnet die Differentiation bezüglich der i -ten Komponente von x .

Dann gilt nach Kettenregel und unter Ausnutzung, dass μ affin linear:

$$\hat{\partial}_i \hat{v}(\hat{x}) = \hat{\partial}_i v(\mu(\hat{x})) = \sum_{l=1}^n \partial^l v(\mu(\hat{x})) \cdot \hat{\partial}_i \mu_l(\hat{x}) = \sum_{l=1}^n \partial^l v(\mu(\hat{x})) \cdot B_{li}. \quad (7.6)$$

Folgerung 7.8. Für den Gradienten auf dem Referenzelement gilt

$$\hat{\nabla} \hat{v}(\hat{x}) = B^T \nabla v(\mu(\hat{x})). \quad (7.7)$$

Die Formel $\nabla \psi(\mu(\hat{x})) = B^{-T} \hat{\nabla} \hat{\psi}(\hat{x})$ verwendet man, um Gradienten der Basisfunktionen vom Referenzelement \hat{T} auf das allgemeine Element T zu transformieren. \square

Wir verallgemeinern nun (7.6) auf höhere Ableitungen:

$$\begin{aligned} \hat{L}[y_1, \dots, y_m] &= \prod_{k=1}^m \left(\sum_{i_k=1}^n (y_k)_{i_k} \hat{\partial}_{i_k} \right) \\ &= \prod_{k=1}^m \left[\sum_{i_k=1}^n (y_k)_{i_k} \left(\sum_{l_k=1}^n \partial_{l_k} B_{l_k i_k} \right) \right] \\ &= \prod_{k=1}^m \left[\sum_{l_k=1}^n \partial_{l_k} \left(\sum_{i_k=1}^n B_{l_k i_k} (y_k)_{i_k} \right) \right] \\ &= L[By_1, \dots, By_m] \end{aligned} \quad (7.8)$$

Zusammen mit (7.2) gilt dann:

$$\begin{aligned} |\hat{L}[y_1, \dots, y_m] \hat{v}(\hat{x})| &= |L[By_1, \dots, By_m] v(\mu(\hat{x}))| \\ &\leq \|D^m v(\mu(\hat{x}))\| \|B\|^m \prod_{k=1}^m \|y_k\|. \end{aligned} \quad (7.9)$$

Damit zeigen wir nun

Satz 7.9 (Transformationsformel). Sei $\mu : \hat{\Omega} \rightarrow \Omega$ mit $\mu(\hat{x}) = x_0 + B\hat{x}$ eine affin lineare, nichtsinguläre Transformation sowie $\hat{v}(\hat{x}) = v(\mu(\hat{x}))$ für gegebenes $v \in H^m(\Omega)$. Dann gilt mit einer Konstanten $c = c(\hat{\Omega}, m)$:

$$|\hat{v}|_{m, \hat{\Omega}} \leq c \|B\|^m |\det B|^{-\frac{1}{2}} |v|_{m, \Omega}. \quad (7.10)$$

Beweis:

$$\begin{aligned} |\hat{v}|_{m, \hat{\Omega}}^2 &= \int_{\hat{\Omega}} \sum_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}(\hat{x})|^2 d\hat{x} \\ &\leq \int_{\hat{\Omega}} \sum_{(i_1, \dots, i_m) \in I_m} |\hat{\partial}_{i_1} \dots \hat{\partial}_{i_m} \hat{v}(\hat{x})|^2 d\hat{x} \\ &= \int_{\hat{\Omega}} \sum_{(i_1, \dots, i_m) \in I_m} |\hat{L}[e_{i_1}, \dots, e_{i_m}] \hat{v}(\hat{x})|^2 d\hat{x} \\ &\leq \int_{\hat{\Omega}} \sum_{(i_1, \dots, i_m) \in I_m} \|D^m v(\mu(\hat{x}))\|^2 \|B\|^{2m} \prod_{k=1}^m \|e_{i_k}\|^2 d\hat{x} \\ &= \|B\|^{2m} n^m \int_{\hat{\Omega}} \underbrace{\|D^m v(\mu(\hat{x}))\|^2}_{\sum_{(i_1, \dots, i_m) \in I_m} |\partial_{i_1} \dots \partial_{i_m} v(\mu(\hat{x}))|^2} d\hat{x} \\ &\leq \|B\|^{2m} n^m n^m \int_{\hat{\Omega}} \sum_{|\alpha|=m} |\partial^\alpha v(\mu(\hat{x}))|^2 d\hat{x} \\ &= \|B\|^{2m} n^{2m} \int_{\Omega} \sum_{|\alpha|=m} |\partial^\alpha v(\mu(\mu^{-1}(x)))|^2 |\det B^{-1}| dx \\ &= n^{2m} \|B\|^{2m} |\det B|^{-1} |v|_{m, \Omega}^2 \end{aligned}$$

Hier haben wir entscheidend benutzt, dass

$$|\{\alpha \in \mathbb{R}^n \mid |\alpha| = m\}| \leq |\{(i_1, \dots, i_m) \in \mathbb{R}^m \mid 1 \leq i_k \leq n\}|$$

sowie für die Rückrichtung

$$|\{(i_1, \dots, i_m) \in \mathbb{R}^m \mid 1 \leq i_k \leq n\}| \leq n^m |\{\alpha \in \mathbb{R}^n \mid |\alpha| = m\}|.$$

Wurzelziehen liefert die Behauptung. □

Bemerkung 7.10. Analog folgt für die Rückrichtung mit $\mu^{-1} : \Omega \rightarrow \hat{\Omega}$:

$$|v|_{m, \Omega} \leq n^m \|B^{-1}\|^m |\det B|^{\frac{1}{2}} |\hat{v}|_{m, \hat{\Omega}} \quad (7.11)$$

□

Damit zeigen wir nun die Aussage von Satz 7.5 im allgemeinen Fall. Zu zeigen ist $\|u - I_h u\|_{m,h} \leq ch^{t,\Omega} |u|_{t-m}$ mit $t-1$ Polynomgrad ($t \geq 2$) sowie m die Sobolevordnung in der der Fehler gemessen wird (typisch wäre für uns $m = 0, 1$).

Zunächst betrachten wir *ein* Element T (\hat{T} ist das Referenzelement) des Gitters und die Seminorm der Ordnung l :

$$\begin{aligned}
|u - I_h u|_{l,T} &\leq c_1 \|B^{-1}\|^l |\det B|^{\frac{1}{2}} |\hat{u} - I_h \hat{u}|_{l,\hat{T}} && \text{(Transformation auf } \hat{T}) \\
&\leq c_1 \|B^{-1}\|^l |\det B|^{\frac{1}{2}} \|\hat{u} - I_h \hat{u}\|_{l,\hat{T}} && \text{(zur Norm ergänzen)} \\
&\leq c_1 \|B^{-1}\|^l |\det B|^{\frac{1}{2}} c_2 |\hat{u}|_{t,\hat{T}} && \text{(Bramble-Hilbert Lemma)} \\
&\leq c_1^2 c_2 \|B^{-1}\|^l |\det B|^{\frac{1}{2}} \|B\|^t \|\det B\|^{-\frac{1}{2}} |u|_{t,T} && \text{(Rücktransformation)} \\
&= c \underbrace{(\|B\| \|B^{-1}\|)^l}_{\leq CK} \|B\|^{t-l} |u|_{t,T} \\
&\leq Ch^{t-l} |u|_{t,T}. && (C \text{ enthält } K \text{ aus (7.5)})
\end{aligned}$$

Summieren über alle Dreiecke liefert:

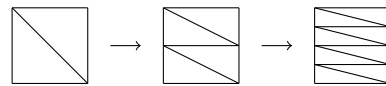
$$|u - I_h u|_{l,h}^2 = \sum_{T \in T_h} |u - I_h u|_{l,T}^2 \leq Ch^{2(t-l)} \underbrace{\sum_{T \in T_h} |u|_{t,T}^2}_{=|u|_{t,\Omega}^2}.$$

Summieren über die Ableitungsordnung liefert:

$$\begin{aligned}
\|u - I_h u\|_{m,h}^2 &= \sum_{l=0}^m |u - I_h u|_l^2 \leq c \sum_{l=0}^m h^{2(t-l)} |u|_{t,\Omega}^2 \\
&= c |u|_{t,\Omega}^2 h^{2t} \underbrace{\left(1 + \frac{1}{h^2} + \dots + \frac{1}{h^{2m}}\right)}_{\leq ch^{-2m}} \\
&\leq ch^{2(t-m)} |u|_{t,\Omega}^2
\end{aligned}$$

Wurzelziehen beweist dann die Aussage.

Bemerkung 7.11. Der hier gezeigte Approximationssatz setzt quasiuniforme Gitter voraus. Dies bedeutet, dass im Falle anisotroper Verfeinerung



$\|B\| \|B^{-1}\| \rightarrow \infty$ gilt. Eine andere Beweistechnik, siehe [BA76], zeigt jedoch, dass es nur auf den *größten* Winkel in einem Dreieck ankommt, d.h. es muss nur der maximale Winkel in einem Dreieck für h gegen 0 von π weg beschränkt bleiben. \square

Bemerkung 7.12 (inverse Abschätzung). Der Approximatinssatz lautet $\|u - I_h u\|_{m,h} \leq ch^{t-m} \|u\|_t$ mit $m \leq t$. Hier haben wir die Seminorm zur Norm ergänzt. Weiter sei $m < t$ anegenommen, damit man eine h -Potenz gewinnt.

Für die Finite-Element-Funktionen $v_h \in S_h$ zeigt man für *uniforme* Gitter die folgende Ungleichung:

$$\|v_h\|_t \leq ch^{m-t} \|v_h\|_m \quad 0 \leq m \leq t \quad (7.12)$$

Dies bezeichnet man als eine „inverse Abschätzung“ die öfters in Beweisen gebraucht wird. (7.12) kann man auch lokal auf einem Dreieck T mit $h = h_T$ verwenden. \square

Bemerkung 7.13 (Optimalität). Man kann auch zeigen [Bra91, Bem. 6.10], dass der Approximationssatz optimal bezüglich der h -Potenz ist (d. h. es gibt Funktionen für die diese Konvergenzordnung auch tatsächlich angenommen wird). \square

Bemerkung 7.14. Die Fehlerabschätzung ist auch in folgendem Sinne scharf: Die Bedingung $u \in H^t(\Omega)$ ist nicht nur hinreichend sondern auch notwendig um h^{t-m} -Konvergenz zu zeigen. Dies bedeutet, dass hoher Polynomgrad auch wirklich nur dann lohnt, wenn die Lösung u glatt genug ist. \square

8 Fehlerabschätzungen

8.1 Regularitätssätze

Nach dem Satz von Lax-Milgram 4.2 existiert stets eine schwache Lösung in $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$.

Für die Anwendung des Approximationssatzes 7.5 benötigen wir bei $m = 1$ allerdings $u \in H^t(\Omega)$ mit $t \geq 2$.

Regularitätssätze geben Auskunft unter welchen Umständen solch eine „erhöhte Regularität“, zu erwarten ist.

Definition 8.1. Sei $m \geq 1$, $H_0^m(\Omega) \subseteq V \subseteq H^m(\Omega)$ und a eine V -elliptische Bilinearform (bei uns gilt $m = 1$). Das Variationsproblem

$$a(u, v) = (f, v)_0 \quad \forall v \in V$$

heißt H^s -regulär, wenn es zu jedem $f \in H^{s-2m}$ eine Lösung $u \in H^s(\Omega)$ gibt und einer Zahl $c(\Omega, a, s)$ gilt:

$$\|u\|_s \leq c \|f\|_{s-2m}.$$

□

So bedeutet H^2 -Regularität eben $\|u\|_2 \leq c \|f\|_0$.

Bemerkung 8.2. Für die Existenz der Lösung ist „weniger“ als $f \in L_2(\Omega)$ erforderlich, da nur $(f, v)_0$ wohldefiniert sein muss. Da $v \in H^1(\Omega)$ genügt $f \in H^{-1}(\Omega)$. □

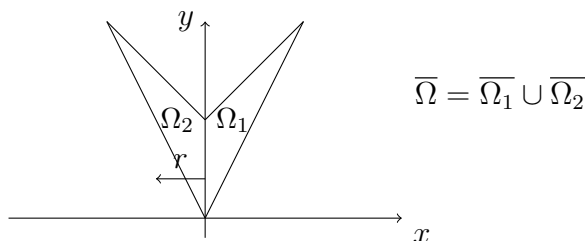
Als Beispiel für einen Regularitätssatz zitieren wir:

Satz 8.3. Sei a eine H_0^1 -elliptische Bilinearform mit hinreichend glatten Koeffizienten (z.B. $a_{\alpha\beta}$ Lipschitz-stetig - siehe [Hac86, Satz 9.1.22]).

- (1) Wenn Ω konvex ist, so ist das Dirichlet-Problem H^2 -regulär.
- (2) Sei $s \geq 2$. Wenn Ω einen C^s -Rand besitzt, so ist das Dirichlet-Problem H^s -regulär.

Beweis siehe [Hac86]. □

Beispiel 8.4. (Aus [Bra91]) Zum Neumann-Problem betrachte das achsensymmetrische, nicht-konvexe Gebiet



Da Ω nicht konvex, gilt im allgemeinen $u \notin H^2(\Omega)$. Ist die Lösung symmetrisch, d.h. $u(-x, y) = u(x, y)$ und glatt im Inneren (siehe unten) so gilt $\frac{\partial u}{\partial x} = 0$ auf $\Gamma_1 = \partial\Omega_1 \cap \{(x, y) \in \Omega | x = 0\}$.

Das Problem mit gemischten Randbedingungen

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega_1 && \text{(konvex!)} \\ \frac{\partial u}{\partial x} &= 0 && \text{auf } \Gamma_1 \\ u &= g && \text{auf } \partial\Omega_1 - \Gamma_1 \end{aligned}$$

ist somit im allgemeinen nicht H^2 -regulär, trotz konvexem Gebiet Ω_1 . □

Bemerkung 8.5. Sofern die rechte Seite f und die Koeffizienten $a_{\alpha\beta}$ genügend glatt sind, ist die Lösung im Inneren des Gebietes üblicherweise sehr viel regulärer, siehe z. B. [Hac86, Satz 9.1.26]. □

8.2 Fehlerabschätzung in der Energienorm

Satz 8.6. Sei Ω ein polygonales Gebiet und es sei eine Familie quasiuniformer Triangulierungen gegeben. Die Lösung des Variationsproblems sei H^s -regulär. Dann gilt die Fehlerabschätzung

$$\|u - u_h\|_1 \leq ch^{s-1} \|f\|_{s-2m}.$$

Beweis:

$$\begin{aligned} \|u - u_h\|_1 &\leq \frac{c}{\alpha} \inf_{v_h \in S_h} \|u - v\|_1 && \text{Lemma 5.2} \\ &\leq \frac{c}{\alpha} \|u - I_h u\|_{1,h} && \text{Abschätzen durch Interpolationsfehler} \\ &\leq \frac{c}{\alpha} ch^{s-1} |u|_{s,\Omega} && \text{Satz 7.5} \\ &\leq Ch^{s-1} \|f\|_{s-2m} && H^s\text{-Regularität.} \end{aligned}$$

□

Bemerkung 8.7. Für $s > 2$, d.h. Polynomgrad > 1 ist nach Satz 8.3 (2) ein C^3 -Rand hinreichend um $u \in H^s$ zu erreichen. Dies ist für ein polygonales Gebiet nicht möglich (die Bedingung ist aber nicht notwendig sondern nur hinreichend).

- Polynome mit Grad größer 1 lohnen nur, wenn man weiss, dass die Lösung genügend glatt ist.
- Abgesehen von speziellen Punkten, Ecken, Kanten ist sie das in der Praxis jedoch häufig. Dies führt zur sog. *hp*-Methode, bei der man in der Nähe der Stellen mit niedriger Regularität niedrigen Polynomgrad (und feines Gitter) und weg davon Polynome mit hohem Grad (und große Elemente) verwendet.
- Für $s = 2, m = 1$ gilt speziell

$$\|u - u_h\|_1 \leq Ch \|u\|_2 \leq Ch \|f\|_0.$$

□

8.3 Fehlerabschätzung in der L_2 -Norm

Satz 8.6 liefert für Polynome vom Grad 1 nur $O(h)$ Konvergenz.

Die Finite-Differenzen-Methode konvergiert (unter geg. Voraussetzungen) jedoch mit $O(h^2)$ an den Gitterpunkten.

Dies liegt natürlich an den gewählten Normen $\|\cdot\|_1$, bzw. $\|\cdot\|_{C^0(\Omega)}$, insbesondere enthält $\|\cdot\|_1$ ja auch Ableitungen der Lösung. Wir wollen hier nun eine Fehlerabschätzung in der L_2 -Norm zeigen.

Satz 8.8. Es gelten die Voraussetzungen von Satz 8.6 und es sei $s = 2$ (d.h. das Variationsproblem sei H^2 -regulär). Dann gilt die Fehlerabschätzung

$$\|u - u_h\|_0 \leq Ch^2 \|f_0\|.$$

Beweis: Wir haben die Situation

$$S_h \subset V = H_0^1(\Omega) \subset H = L_2(\Omega).$$

Für eine beliebige Funktion $g \in H = L_2(\Omega)$ definiert man das sog. „duale“ Problem:

$$a(w, \varphi_g) = (g, w)_0 \quad \forall w \in V.$$

Die Reihenfolge der Argumente ist wesentlich, wenn man unsymmetrische Probleme betrachtet!

Wir erinnern uns an

(Variationsproblem)	$a(u, v) = (f, v)_0$	$\forall v \in V,$
(FE-Problem)	$a(u_h, v) = (f, v)_0$	$\forall v \in S_h,$
(Galerkin-Orthogonalität)	$a(u - u_h, v) = 0$	$\forall v \in S_h.$

Damit erhalten wir für beliebiges $g \in H$ und Testfunktion $w = u - u_h$, sowie beliebiges $v \in S_h$:

$$\begin{aligned} (g, u - u_h)_0 &= a(u - u_h, \varphi_g) = a(u - u_h, \varphi_g - v) \\ &\leq C \|u - u_h\|_1 \|\varphi_g - v\|_1 \\ &\leq C \|u - u_h\|_1 \inf_{v \in S_h} \|\varphi_g - v\|_1. \end{aligned}$$

Rechts haben wir im Prinzip schon $O(h^2)$ stehen, wenn wir die Approximationseigenschaft für φ_g ausnutzen. Allerdings steht auf der linken Seite der Ungleichung noch ein unhandlicher Term.

Die Norm eines Elementes $w \in H$ kann man charakterisieren durch:

$$\|w\|_0 = \sup_{0 \neq g \in H=L_2(\Omega)} \frac{(g, w)_0}{\|g\|_0},$$

denn nach Cauchy-Schwarz ist $(g, w)_0 \leq \|g\|_0 \|w\|_0$ und somit

$$\frac{(g, w)_0}{\|g\|_0} \leq \|w\|_0$$

für ein beliebiges $w \in L_2(\Omega)$. Für $g = w$ gilt die Gleichheit und damit obige Charakterisierung über das Supremum.

Damit gilt nun:

$$\begin{aligned} \|u - u_h\|_0 &= \sup_{0 \neq g \in H=L_2(\Omega)} \frac{(g, u - u_h)_0}{\|g\|_0} \\ &\leq C \|u - u_h\|_1 \sup_{g \in H} \left\{ \frac{1}{\|g\|_0} \underbrace{\inf_{v \in S_h} \|\varphi_g - v\|_1}_{\substack{\varphi_g : \text{Lösung des dualen Problems. hier} \\ a(w, \varphi_g) = a(\varphi_g, w) \text{ d.h. duales = pri-} \\ \text{males Problem} = H^2\text{-regulär wg. } \varphi_g \in \\ H_0^1(\Omega) \subset L_2(\Omega) \text{ gilt } \inf_{v \in S_h} \|\varphi_g - \\ v\|_1 \leq ch \|\varphi_g\|_2 \leq ch \|g\|_0}} \right\} \\ &\leq ch^2 \|f\|_0 \end{aligned}$$

Diese Beweistechnik über das duale Problem wird auch als „Nitsche-Trick“ bezeichnet. \square

Folgerung 8.9. Mit den oben gemachten Voraussetzungen (also H^2 -Regularität) gilt auch die Abschätzung

$$\|u - u_h\|_0 \leq Ch \|u - u_h\|_1. \quad (8.1)$$

Beweis: Folgt unmittelbar aus dem Beweis oben indem man in der letzten Abschätzung nur den Supremumsterm mit ch abschätzt und $\|u - u_h\|$ stehen lässt. \square

9 Adaptive Gittersteuerung

9.1 Einführung

Ziel jeder Berechnung ist es eine Fehlerschranke

$$\|u - u_h\| \leq \text{TOL} \quad (9.1)$$

zu erfüllen. Hierbei ist die Norm zunächst nicht weiter spezifiziert. Wie garantiert man das *ohne Kenntnis* der Lösung u ?

Zweitens, wenn man den Fehler $u - u_h$ schätzen könnte, wie erreicht man (9.1) möglichst effizient (d.h. mit $|\mathcal{T}_h|$ minimal)?

Die „a priori“ Fehlerabschätzungen

$$\|u - u_h\|_1 \leq ch^{s-1} \|u\|_s \quad (\text{Satz 8.6}),$$

$$\|u - u_h\|_0 \leq ch^2 \|f\|_0 \quad (\text{Satz 8.8}),$$

nützen nicht viel, da u oder gar $\partial^\alpha u$ nicht bekannt sind und c eventuell sehr groß ist.

Ziel dieses Abschnittes ist die Herleitung sogenannter „a posteriori“ Fehlerabschätzung, die die berechnete Lösung u_h mit einbeziehen.

Fehlermaße Den Fehler bezeichnen wir mit $e_h := u - u_h$ (Vorsicht $e_h \notin S_h$ trotz Index!) wobei u und u_h die Lösungen der folgenden Variationsprobleme sind:

$$\begin{array}{lll} u \in V : & a(u, v) = (f, v)_0 & \forall v \in V = H_0^1(\Omega), \\ u_h \in S_h : & a(u_h, v) = (f, v)_0 & \forall v \in S_h \subset V. \end{array}$$

Wir betrachten hier im Übrigen nur Dirichlet-Randbedingung, d. h. $V = H_0^1(\Omega)$.

In vielen praktischen Fällen interessiert der Fehler nicht überall gleich stark, sondern man möchte den Fehler an bestimmten Stellen lokal stärker gewichten. Dazu betrachtet man Fehlerfunktionale $J : V \rightarrow \mathbb{R}$. Beispiele sind

$$\begin{array}{ll} \text{Normen:} & \|e_h\|_0, \|e_h\|_1, \\ \text{Mittelwerte:} & |(e_h, \psi)_{0,\Omega}| \text{ für } \psi \in C(\bar{\Omega}), \\ \text{Linienintegral:} & |(e_h, \psi)_{0,\partial\Omega}| \text{ für } \psi \in C(\partial\Omega), \\ \text{Punktwerte:} & |e_h(x)|, |\partial_i e_h(x)| \text{ für } x \in \Omega. \end{array}$$

9.2 Duale Fehlerschätzung

Im folgenden stellen wir die Methode der dualen Fehlerschätzung vor. Diese hat gegenüber anderen Methoden den Vorzug, dass man sehr allgemeine Funktionale des Fehlers $|J(e_h)|$ abschätzen kann. Die Darstellung der Methode folgt [Ran06].

Zu J betrachten wir das duale Problem:

$$\text{Finde } z \in V: \quad a(\varphi, z) = J(\varphi) \quad \forall \varphi \in V.$$

Für die spezielle Wahl der Testfunktion $\varphi = e_h$ ergibt sich dann

$$\begin{aligned}
J(e_h) &= a(e_h, z) = a(e_h, z - \psi_h) \quad \text{für beliebiges } \psi_h \in S_h \\
&= \sum_{T \in \mathcal{T}_h} (A \nabla e_h, \nabla(z - \psi_h))_{0,T} \\
&= \sum_{T \in \mathcal{T}_h} \{ -(\nabla \cdot (A \nabla e_h), z - \psi_h)_{0,T} + ((A \nabla e_h) \cdot \nu, z - \psi_h)_{0,\partial T} \} \\
&= \sum_{T \in \mathcal{T}_h} (f + \nabla \cdot (A \nabla u_h), z - \psi_h)_{0,T} + \sum_{e \in \Gamma_h} ((A \nabla e_h) \cdot \nu, z - \psi_h)_{0,e} \\
&+ \sum_{e \in E_h, e = T_i \cap T_j} ((A \nabla e_h|_{T_i}) \cdot \nu_i - (A \nabla e_h|_{T_j}) \cdot \nu_j, z - \psi_h)_{0,e}
\end{aligned}$$

wobei hier $E_h = \{e \mid e \text{ ist gemeinsame Kante von } T_i \text{ und } T_j\}$ die Menge der inneren Kanten und $\Gamma_h = \{e \mid e \text{ ist Randkante eines Dreiecks}\}$ die Menge der Randkanten bezeichnet.

Es bezeichne $[v](x) = \lim_{\epsilon \rightarrow 0^+} v(x + \epsilon \nu) - \lim_{\epsilon \rightarrow 0^+} v(x - \epsilon \nu)$ den „Sprung“ einer auf einer Elementkante unstetigen Funktion v . Für die Kantenterme gilt dann weiter

$$\begin{aligned}
[(A \nabla e_h) \cdot \nu]_{e \in E_h} &= \underbrace{[(A \nabla u) \cdot \nu]_{e \in E_h}}_{=0} - [(A \nabla u_h) \cdot \nu]_{e \in E_h} \\
&\Rightarrow ((A \nabla e_h) \cdot \nu, z - \psi_h)_{0,e} = -((A \nabla u_h) \cdot \nu, z - \psi_h) \quad \text{für } e \in E_h.
\end{aligned}$$

Wegen $z, \psi_h \in V = H_0^1(\Omega)$ ist $z - \psi_h = 0$ auf $\partial\Omega$, also $((A \nabla e_h) \cdot \nu, z - \psi_h)_{0,e} = 0$ für alle Kanten $e \in \Gamma_h$.

Aufteilen des Sprungtermes zu gleichen Teilen auf beide Seiten liefert

$$J(e_h) = \sum_{T \in \mathcal{T}_h} \left\{ (f + \underbrace{\nabla \cdot (A \nabla u_h)}_{=0 \text{ für } \mathcal{P}_1 \text{ und } A = \text{const auf } T}, z - \psi_h)_{0,T} - \frac{1}{2}([(A \nabla u_h) \cdot \nu], z - \psi_h)_{0,\partial T \cap \Omega} \right\}.$$

Bis hier wurden keinerlei Abschätzungen durchgeführt. Nun bildet man Normen und wendet Dreiecksungleichung bzw. Cauchy-Schwarz an:

$$\begin{aligned}
|J(e_h)| &\leq \sum_{T \in \mathcal{T}_h} |f + \nabla \cdot (A \nabla u_h), z - \psi_h)_{0,T} - \frac{1}{2}([(A \nabla u_h) \cdot \nu], z - \psi_h)_{0,\partial T \cap \Omega}| \\
&\leq \sum_{T \in \mathcal{T}_h} \{ \|f + \nabla \cdot (A \nabla u_h)\|_{0,T} \|z - \psi_h\|_{0,T} + \frac{1}{2} \|[(A \nabla u_h) \cdot \nu]\|_{0,\partial T \cap \Omega} \|z - \psi_h\|_{0,\partial T \cap \Omega} \}.
\end{aligned} \tag{9.2}$$

Nun gibt es verschiedene Möglichkeiten fortzufahren. Wir betrachten hier nur eine davon.

9.3 Energienormfehlerschätzer

Im Folgenden zeigen wir die Abschätzung des (skalierten) Energiefehlers

$$J(e_h) = \alpha |e_h|_{1,\Omega} = \alpha \|\nabla e_h\|_{0,\Omega}, \tag{9.3}$$

wobei α die Elliptizitätskonstante ist. Aufgrund der Elliptizität von a erhalten wir die Abschätzung

$$a(v, v) \geq \alpha \|v\|_{1,\Omega}^2 = \alpha (\|v\|_{0,\Omega}^2 + \|\nabla v\|_{0,\Omega}^2) \geq \alpha \|\nabla v\|_{0,\Omega}^2.$$

Nun setzen wir

$$J(\varphi) = \alpha \frac{(\nabla \varphi, \nabla e_h)_{0,\Omega}}{\|\nabla e_h\|_{0,\Omega}}$$

und erhalten unter Beachtung von $\varphi = e_h$ dann (9.3).

Die Lösung des dualen Problems Z kann man dann abschätzen durch

$$\begin{aligned} \alpha \|\nabla z\|_{0,\Omega}^2 \leq a(z, z) = J(z) &= \alpha \frac{(\nabla z, \nabla e_h)_{0,\Omega}}{\|\nabla e_h\|_{0,\Omega}} \leq \alpha \|\nabla z\|_{0,\Omega} \\ \Leftrightarrow \|\nabla z\|_{0,\Omega} &= |z|_{1,\Omega} \leq 1. \end{aligned}$$

Nun benötigt man noch folgende (lokale) Abschätzung für den Interpolationsfehler auf einem Dreieck T :

$$\|v - I_h v\|_{0,T} + h_T^{\frac{1}{2}} \|v - I_h v\|_{0,\partial T} \leq ch_T |v|_{1,\tilde{T}}$$

wobei $\tilde{T} = \{T' \in \mathcal{T}_h | T' \cap T \neq \emptyset\}$, siehe [Ran06].

Damit schätzt man nun in (9.2) die Terme $z - \psi_h$ ab indem man $\psi_h = I_h z$ einsetzt:

$$\begin{aligned} |J(e_h)| &\leq \sum_{T \in \mathcal{T}_h} \left\{ \|f + \nabla \cdot (A \nabla u_h)\|_{0,T} ch_T |z|_{1,\tilde{T}} + \frac{1}{2} h_T^{-\frac{1}{2}} \|[(A \nabla u_h) \cdot \nu]\|_{0,\partial T \cap \Omega} ch_T |z|_{1,\tilde{T}} \right\} \\ &\leq c \left(\sum_{T \in \mathcal{T}_h} h_T^2 \left\{ \|f + \nabla \cdot (A \nabla u_h)\|_{0,T} + \frac{1}{2} h_T^{-\frac{1}{2}} \|[(A \nabla u_h) \nu]\|_{0,\partial T \cap \Omega} \right\}^2 \right)^{\frac{1}{2}} \\ &\quad \underbrace{\left(\sum_{T \in \mathcal{T}_h} |z|_{1,\tilde{T}}^2 \right)^{\frac{1}{2}}}_{\leq c' (\sum_{T \in \mathcal{T}_h} |z|_{1,T}^2)^{\frac{1}{2}}, \text{ da } T \text{ in endlich vielen } \tilde{T}}. \end{aligned}$$

Für die zweite Abschätzung wurde Cauchy-Schwarz im \mathbb{R}^n verwendet. Wegen $|z|_{1,\Omega} \leq 1$ fällt der zweite Faktor komplett weg.

Das Ergebnis können wir zusammenfassen in

$$\begin{aligned} |J(e_h)| &= \alpha |e_h|_{1,\Omega} \leq C \left(\sum_{T \in \mathcal{T}_h} \eta_T^2 \right)^{\frac{1}{2}} \\ \text{mit } \eta_T &= h_T \|f + \nabla \cdot (A \nabla u_h)\|_{0,T} + \frac{1}{2} h_T^{\frac{1}{2}} \|[(A \nabla u_h) \cdot \nu]\|_{0,\partial T \cap \Omega}. \end{aligned}$$

Hierzu einige Bemerkungen:

- Die Größen η_T sind lokal pro Dreieck berechenbar.

- $(\sum_{T \in \mathcal{T}_h} \eta_T^2)^{\frac{1}{2}}$ liefert bis auf eine Konstante eine obere Schranke für $|e_h|_{1,\Omega}$.
- Es kann aber sein, dass die Fehlerabschätzung dem Fehler grob überschätzt
- Gilt auch eine Abschätzung nach unten der Form

$$\left(\sum_{T \in \mathcal{T}_h} \eta_T^2 \right)^{\frac{1}{2}} \leq C' |J(e_h)| \quad \text{für } h \rightarrow 0 \text{ und } C' \text{ unabhängig von } h$$

so heißt der Fehlerschätzer „effizient“. In diesem Fall konvergiert der Fehlerschätzer asymptotisch mit der selben Rate wie der echte Fehler.

9.4 Verfeinerungsstrategie

Die Gittersteuerung nutzt das heuristische Prinzip der „Equilibrierung“ der Fehler, d. h. ein Gitter gilt dann als optimal, wenn jedes Element T den selben Beitrag zum Gesamtfehler liefert. Um dies zu erreichen geht man wie folgt vor:

1. Wähle ein Gitter $\mathcal{T}^{(0)}$ und berechne die Finite-Elemente-Lösung $u_h^{(0)}$.
2. Berechne $E^{(m)} = \left(\sum_{T \in \mathcal{T}^{(m)}} \eta_T^2 \right)^{\frac{1}{2}}$. Falls $E^{(m)} \leq TOL$ dann sind wir fertig.
3. Ordne die Elemente nach der Größe des Fehlerbeitrages

$$\nu_{T_{i_1}} \geq \nu_{T_{i_2}} \geq \dots \geq \nu_{T_{i_{N_h}}}.$$

4. Verfeinere alle Elemente T_{i_k} mit $k \leq \omega N_h$ wobei $0 \leq \omega \leq 1$, etwa $\omega = \frac{1}{4}$. Gehe zu 1.

Beispiel 9.1. Wir betrachten die einspringende Ecke aus Beispiel 2.1. Die Abbildung 10 zeigt ein lokal verfeinertes Dreiecksgitter welches mit den in diesem Abschnitt beschriebenen Methoden erzeugt wurde. Schließlich zeigt Abbildung 11 den Fehler in Abhängigkeit der Problemgröße N_h für global verfeinerte und adaptiv verfeinerte Gitter. Deutlich ist die höhere Effizienz der adaptiven Methode zu sehen. \square

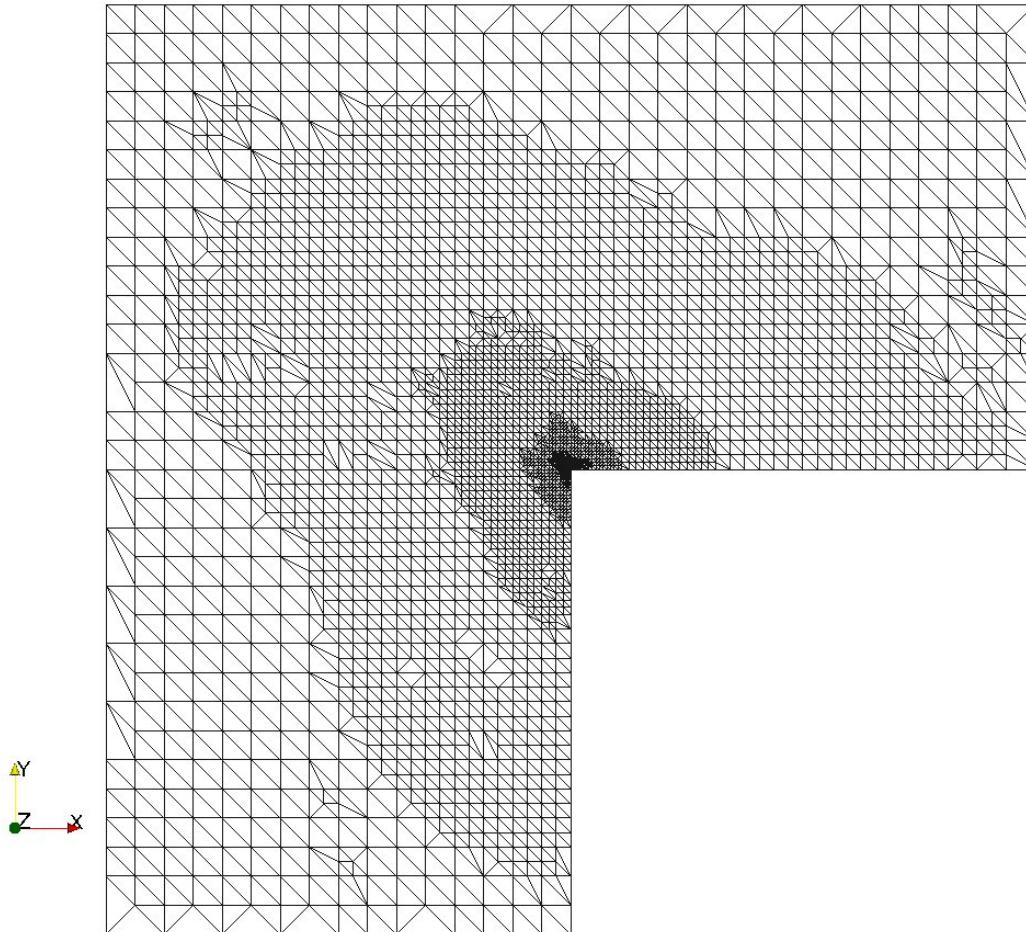


Abbildung 10: Lokal verfeinertes Gitter für die einspringende Ecke.

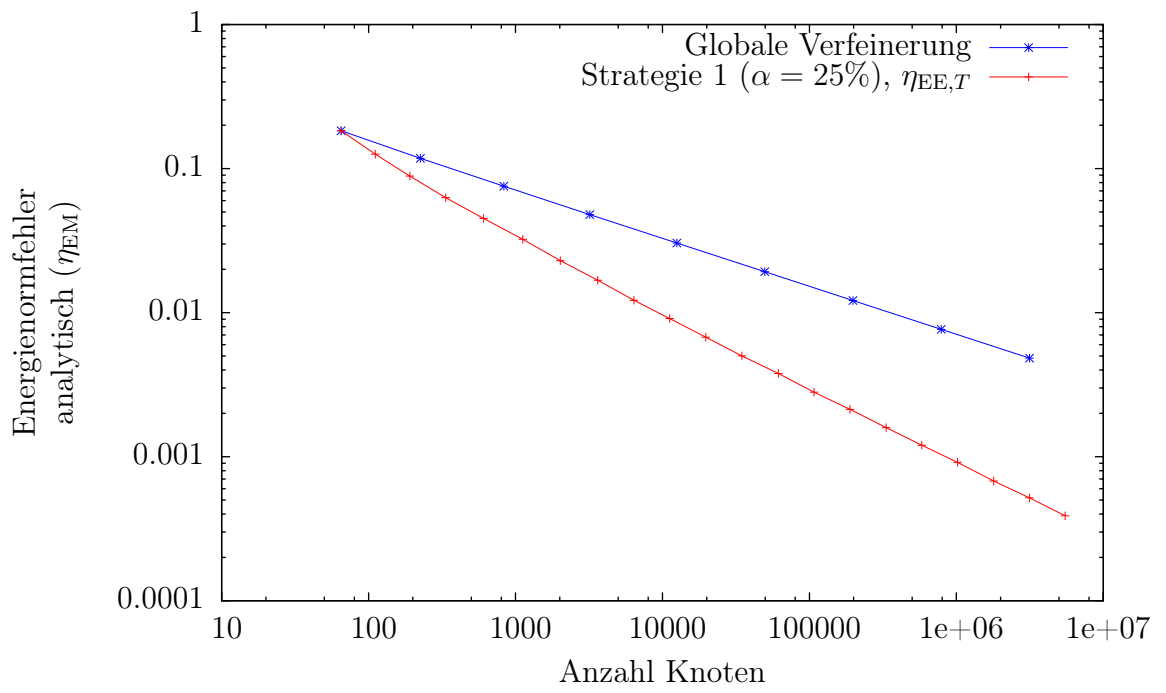


Abbildung 11: Fehler in Abhängigkeit von N_h .

10 Mehrgitterverfahren

Die Finite-Elemente-Formulierung

$$u_h \in S_h : \quad a(u_h, v) = l(v) \quad \forall v \in S_h$$

ist äquivalent zur Lösung des linearen Gleichungssystems

$$A_h x_h = b_h \quad \text{mit } b_h, x_h \in \mathbb{R}^{N_h} \text{ und } A_h \in \mathbb{R}^{N_h \times N_h}.$$

Die Matrix A_h ist symmetrisch positiv definit, d. h. sie hat ein reelles und positives Spektrum

$$\sigma(A_h) = \{\lambda_1, \dots, \lambda_{N_h}\} \subset \mathbb{R}^+, \quad 0 < \lambda_i \leq \lambda_{i+1}.$$

In diesem und dem nächsten Abschnitt geben wir eine kurze Einführung in effiziente iterative Lösungsverfahren für die linearen Gleichungssysteme.

10.1 Spektralverhalten einfacher Iterationsverfahren

Das einfachste lineare Iterationsverfahren zur Lösung linearer Gleichungssysteme ist die Richardson-Iteration gegeben durch die Vorschrift

$$x_h^{k+1} = x_h^k + \omega(b_h - A_h x_h^k).$$

Dabei ist $\omega \in \mathbb{R}$, $0 < \omega \leq 1$ der Dämpfungsfaktor.

Für den Iterationsfehler $e_h^{k+1} = x_h - x_h^{k+1}$ im $k + 1$ -ten Schritt zeigt man die Fehlerfortpflanzungsgleichung

$$e_h^{k+1} = x_h - x_h^{k+1} = x_h - x_h^k - \omega(A_h x_h^k - A_h x_h^k) = \underbrace{(I_h - \omega A_h)}_{\text{Iterationsmatrix } M_h} e_h^k,$$

wobei I_h die Einheitsmatrix bezeichnet.

Sei z_i der Eigenvektor zum Eigenwert λ_i . Aufgrund der Eigenschaften von A_h bilden die z_i eine Basis des \mathbb{R}^{N_h} . Damit kann man den Anfangsfehler mittels $e_h^0 = x_h - x_h^0 = \sum_{i=1}^{N_h} \alpha_i z_i$ in der Basis der Eigenvektoren schreiben. Dann gilt

$$e_h^n = (I_h - \omega A_h)^n \sum_{i=1}^{N_h} \alpha_i z_i = \sum_{i=1}^{N_h} (I_h - \omega A_h)^n \alpha_i z_i = \sum_{i=1}^{N_h} \alpha_i \underbrace{(1 - \omega \lambda_i)^n}_{\text{Verstärkungsfaktor}} z_i.$$

Nun wähle speziell $\omega = \frac{1}{\lambda_{max}} = \frac{1}{\lambda_{N_h}}$. Dann gilt

$$1 - \omega \lambda_i = 1 - \frac{\lambda_i}{\lambda_{N_h}} = \begin{cases} 0 & i = N_h \\ 1 - \frac{\lambda_i}{\lambda_{N_h}} \approx 1 \text{ für } \lambda_i \ll \lambda_{N_h} \end{cases}.$$

Dabei gilt $\frac{\lambda_1}{\lambda_{N_h}} = O(h^2)$ und somit liegt die Konvergenz für $h \rightarrow 0$ sehr schnell nahe 1.

Beispiel 10.1. Für einfache Fälle lassen sich die Eigenvektoren direkt angeben. Betrachte

$$u''(x) = f, \quad u(0) = u(1) = 0.$$

Dann lautet die zugehörige Matrix bis auf den Faktor $-1/h$ (den man auf die rechte Seite schaffen kann):

$$A_h = \text{tridiag}\{-1, 2, -1\}$$

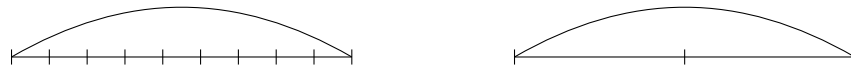
für die die Eigenvektoren z_i die folgende Form haben:

$$(z_i)_k = \sin \frac{ki\pi}{N_h} = \sin \left(\frac{k}{N_h} i\pi \right) \quad 1 \leq i, k < N_h.$$

□

10.2 Gitterhierarchie

Im letzten Beispiel waren die Eigenvektoren abgetastete Sinusfunktionen unterschiedlicher Frequenz. Dabei ist etwa die Funktion $\sin x\pi$ niederfrequent auf einem Gitter mit $N_h = 8$ Elementen, nicht aber jedoch auf einem Gitter mit $N_h = 2$ Elementen.



Idee: Kann man niederfrequente Fehler auf einem größeren Gitter berechnen wo sie wieder hochfrequent sind?

Definiere dazu eine Hierarchie von Gittern, etwa in zwei Raumdimensionen:

$$\begin{aligned} \mathcal{T}_0 : & \begin{array}{c} \square \\ \square \\ \square \end{array} \rightarrow a(u_0, v) = l(v) \quad \forall v \in S_0 \quad \rightarrow A_0 x_0 = b_0, \\ \mathcal{T}_1 : & \begin{array}{c} \square \\ \square \\ \square \end{array} \rightarrow a(u_1, v) = l(v) \quad \forall v \in S_1 \quad \rightarrow A_1 x_1 = b_1, \\ \mathcal{T}_J : & \begin{array}{c} \square \\ \square \\ \square \end{array} \rightarrow a(u_J, v) = l(v) \quad \forall v \in S_J \quad \rightarrow A_J x_J = b_J. \end{aligned}$$

Die feinste Stufe trägt den Index J und es gilt $\mathcal{T}_J = \mathcal{T}_h$, $u_J = u_h$, $x_J = x_h$.

N_0 (Größe des größten Gitters) ist nach unten durch die Komplexität der Geometrie beschränkt. Sog. „algebraische Mehrgitterverfahren“ erzeugen die Grobgittermatrizen A_{J-1}, \dots, A_0 rekursiv aus A_J auf algebraische Weise.

10.3 Zweigitterverfahren

Wir wollen zunächst die oben eingeführte Richardson-Iteration als Iteration auf Finite-Element-Funktionen in S_h statt als Iteration auf den zugehörigen Koeffizienten umschreiben.

Die Richardson-Iteration lautet:

$$x_h^{k+1} = x_h^k + \omega(b_h - A_h x_h^k)$$

wobei zu den Koeffizientenvektoren x_h^{k+1} und x_h^k die Finite-Element-Funktionen

$$u_h^{k+1} = \sum_{i=1}^{N_h} (x_h^{k+1})_i \psi_i^h \quad u_h^k = \sum_{i=1}^{N_h} (x_h^k)_i \psi_i^h$$

gehören. Nun sei $\mathcal{S}_h : S_h \rightarrow S_h$ diejenige Abbildung so dass $u_h^{k+1} = \mathcal{S}_h(u_h^k)$. \mathcal{S}_h ist gegeben durch:

$$\begin{aligned} \mathcal{S}_h(u_h^k) &= u_h^k + \omega \sum_{i=1}^{N_h} [l(\psi_i^k) - a(u_h^k, \psi_i^h)] \psi_i^h \\ &= \sum_{i=1}^{N_h} (x_h^k)_i \psi_i^h + \omega \sum_{j=1}^{N_h} \left[(b_h)_j - \sum_{i=1}^{N_h} (A_h)_{i,j} (x_h^k)_i \right] \psi_j^h \\ &= \sum_{i=1}^{N_h} \underbrace{[(x_h^k)_i + \omega(b_h - A_h x_h^k)_i]}_{(x_h^{k+1})_i} \psi_i^h. \end{aligned}$$

Damit definieren wir die sog. „Zweigitteriteration“:

Algorithmus 10.2 (Zweigitterverfahren). Sei u_h^k eine gegebene Näherung für u_h in S_h .

1. Glättungsschritt: Setze $u_h^{k,1} = \mathcal{S}^\nu(u_h^k)$, die ν -fache Anwendung einer „glättenden“ Iteration, z.B. gedämpfter Richardson. Typisch sind $\nu = 1, 2, 3$ Iterationen.
2. Grobgitterkorrektur: Bestimme ein $w \in S_{2h}$ so dass

$$a(u_h^{k,1} + w, v) = l(v) \quad \forall v \in S_{2h}.$$

Setze $u_h^{k+1} = u_h^{k,1} + w$. □

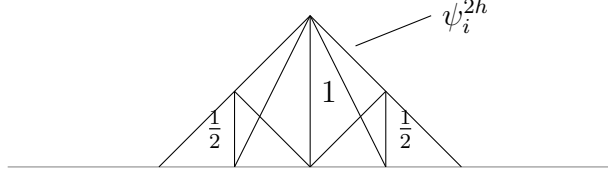
Wie realisiert man den Grobgitterkorrekturschritt? Hier hilft folgende Beobachtung: Die FE-Räume sind „geschachtelt“, d. h. $S_{2h} \subseteq S_h$ bzw. $S_l \subseteq S_{l+1}$ im Mehrgitterfall.

Für jede Funktion $v \in S_{2h}$ gilt somit auch $v \in S_h$ und insbesondere gilt das für alle Basisfunktionen $\psi_i^{2h} \in S_{2h}$. Damit gibt es Koeffizienten r_{ij} so dass:

$$\psi_i^{2h} = \sum_{j=1}^{N_h} r_{ij} \psi_j^h \quad 1 \leq i \leq N_{2h}.$$

Aufgrund der lokalen Träger der ψ_i (Knotenbasis!) überlegt man, dass nur konstant viele (abhängig von der Raumdimension) r_{ij} für ein i ungleich Null sind.

In einer Raumdimension hat man die Situation



was insbesondere zeigt, dass $r_{ij} = \psi_i^{2h}(x_j)$ (mit x_j der Position von Knoten j im Gitter).

Damit realisiert man die Grobgitterkorrektur folgendermaßen:

$$\begin{aligned}
& a(u_h^{k,1} + w, v) = l(v) && \forall v \in S_{2h} \\
\Leftrightarrow & a(w, v) = l(v) - a(u_h^{k,1}, v) && \forall v \in S_{2h} \\
\Leftrightarrow & a\left(\sum_{j=1}^{N_{2h}} (y_{2h})_j \psi_j^{2h}, \psi_i^{2h}\right) = l\left(\sum_{m=1}^{N_h} r_{im} \psi_m^h\right) - a\left(u_h^{k,1}, \sum_{m=1}^{N_h} r_{im} \psi_m^h\right) \\
& = \sum_{m=1}^{N_h} r_{im} \left[l(\psi_m^h) - a\left(\sum_{n=1}^{N_h} (x_h^{k,1})_n \psi_n^h, \psi_m^h\right) \right] && 1 \leq i \leq N_{2h} \\
\Leftrightarrow & \sum_{j=1}^{N_{2h}} (y_{2h})_j a(\psi_j^{2h}, \psi_i^{2h}) = \sum_{m=1}^{N_h} r_{im} (b_h - A_h x_h^{k,1})_m && 1 \leq i \leq N_{2h} \\
\Leftrightarrow & A_{2h} y_{2h} = R_{2h}^h (b_h - A_h x_h^{k,1}).
\end{aligned}$$

Die Matrix $R_{2h} \in \mathbb{R}^{N_{2h} \times N_h}$ ist die *Restriktionsmatrix* mit $(R_{2h})_{ij} = r_{ij}$.

Die Operation $u_h^{k+1} = u_h^{k,1} + w$ ist im Funktionenraum wegen $S_{2h} \subset S_h$ trivial. Für die Koeffizienten gilt

$$\begin{aligned}
u_h^{k+1} &= \sum_{j=1}^{N_h} (x_h^{k+1})_j \psi_j^h = \sum_{j=1}^{N_h} (x_h^{k,1})_j \psi_j^h + \sum_{i=1}^{N_{2h}} (y_{2h})_i \psi_i^{2h} \\
&= \sum_{j=1}^{N_h} (x_h^{k,1})_j \psi_j^h + \sum_{i=1}^{N_{2h}} (y_{2h})_i \sum_{j=1}^{N_h} r_{ij} \psi_j^h \\
&= \sum_{j=1}^{N_h} \left[(x_h^{k,1})_j + \sum_{i=1}^{N_{2h}} r_{ij} (y_{2h})_i \right] \psi_j^h \\
\Leftrightarrow & x_h^{k+1} = x_h^{k,1} + (R_{2h}^h)^T y_{2h}.
\end{aligned}$$

Die algebraische Formulierung des Zweigitterverfahrens in Koeffizienten lautet also:

Algorithmus 10.3. Gegeben sei x_h^k , eine Näherung für x_h .

1. $x_h^{k,0} := x_h^k$;
for $\kappa = 1, \dots, \nu$: $x_h^{k,\kappa} = x_h^{k,\kappa-1} + \omega \left(b_h - A_h x_h^{k,\kappa-1} \right)$; (Glättung)
2. $d_h = b_h - A_h x_h^{k,\nu}$; (Defektberechnung)
3. $d_{2h} = R_{2h}^h d_h$; (Restriktion)
4. Löse $A_{2h} y_{2h} = d_{2h}$; (Groggitterproblem)
5. $y_h = (R_{2h}^h)^T y_{2h}$; (Prolongation)
6. $x_h^{k+1} = x_h^k + y_h$; (Update) \square

10.4 Mehrgitterverfahren

Das Mehrgitterverfahren wendet dieses Prinzip nun rekursiv in Schritt 4 auf das Problem $A_{2h}y_{2h} = d_{2h}$ an. Zusätzlich erlaubt man meist auch noch eine weitere Glättung nach der Grobgitterkorrektur. Damit erhält man den folgenden Algorithmus.

Algorithmus 10.4. Wir formulieren das als rekursive Funktion $MGM(l, x_l, b_l)$, die auf der feinsten Stufe mit $MGM(J, x_J^k, b_J)$ aufgerufen wird und dann das x_J^{k+1} als Ergebnis liefert.

```

MGM(l, x_l, b_l)
{
  if (l == 0) {x_0 = A_0^{-1}b_0; return x_0;}           // kleines Problem exakt lösen
  for (κ = 1, ..., ν_1) x_l = x_l + ω(b_l - A_l x_l);   // Vorglättung
  d_l = (b_l - A_l x_l);                                // Defektberechnung
  d_{l-1} = R_{l-1}^l d_l;                             // Restriktion
  y_{l-1} = 0;                                         // Startwert
  for (i = 1, ..., γ) y_{l-1} = MGM(l - 1, y_{l-1}, d_{l-1}); // Rekursion
  y_l = (R_{l-1}^l)^T y_{l-1};                          // Prolongation
  x_l = x_l + y_l;                                     // Update
  for (κ = 1, ..., ν_2) x_l = x_l + ω(b_l - A_l x_l)   // Nachglättung
  return x_l;
}

```

Der Algorithmus besitzt die Parameter ν_1 , ν_2 und γ .

$\gamma = 1$ heißt V-Zyklus, $\gamma = 2$ W-Zyklus. Minimal sind die Parameter $\nu_1 = 1$, $\nu_2 = 0$ und $\gamma = 1$ zu wählen. \square

11 Konvergenz des Mehrgitterverfahrens

Ziel dieses Abschnittes ist eine Abschätzung der Konvergenzgeschwindigkeit des Mehrgitterverfahrens. Dabei wird sich herausstellen, dass das Mehrgitterverfahren unabhängig von der Gitterweite h konvergiert. Damit zählt das Mehrgitterverfahren zur Klasse der optimalen Lösungsverfahren.

11.1 Vorbereitung

Wir beschränken uns zuerst mal auf das Zweigitterverfahren ohne Nachglättung. Zur Wiederholung stellen wir nochmal die beiden Formulierungen im Funktionenraum und für die Koeffizienten gegenüber

im FE-Raum: \textcircled{A}

i) $u_h^{k,1} = \mathcal{X}^\nu u_h^k;$

ii) $a(u_h^{k,1} + w_{2h}, v) = l(v) \quad \forall v \in S_{2h};$

iii) $u_h^{k+1} = u_h^{k,2} = u_h^{k,1} + w_{2h};$

im \mathbb{R}^N für die Koeffizienten \textcircled{B}

i) $x_h^{k,0} = x_h^k; \quad x_h^{k,\frac{\kappa}{\nu}} = x_h^{k,\frac{\kappa-1}{\nu}} + \omega(b_h - A_h x_h^{k,\frac{\kappa-1}{\nu}});$

ii) $A_{2h} y_{2h} = R_{2h}^h (b_h - A_h x_h^{k,1});$

iii) $x_h^{k,1} = x_h^{k,2} = x_h^k + (R_{2h}^h)^T y_{2h};$

Als Glättungsoperation wurde hier die einfache Richardson-Iteration verwendet.

Die Konvergenzanalyse erfordert eine Rekursionsgleichung für den Iterationsfehler. Für ein lineares Iterationsverfahren der Gestalt $x^{k+1} = x^k + W(b - Ax^k)$ zur Lösung von $Ax = b$ gilt:

$$e^{k+1} = x - x^{k+1} = x - x^k - W(Ax - Ax^k) = (I - WA)(x - x^k) = (I - WA)e^k.$$

Nun will man eine Normabschätzung

$$\|e^{k+1}\| \leq \rho \|e^k\|$$

herleiten mit der Konvergenzrate $\|I - WA\| \leq \rho < 1$

Das Zweigitterverfahren besteht aus zwei hintereinandergeschalteten Iterationen:

$$e^{k+1} = (I - (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h A_h)(I - \omega A)^\nu e^k.$$

Zur Herleitung einer Schranke ρ genügt nicht alleine eine Betrachtung der involvierten Matrizen, sondern man muss geschickt den Bezug zur Formulierung \textcircled{A} im FE-Raum ausnutzen.

FE-Funktionen und Vektoren im \mathbb{R}^{N_h} sind über die Wahl der Basis gekoppelt:

$$S_h \ni u_h = \sum_{i=1}^{N_h} x_i \psi_i \quad \text{mit } x \in \mathbb{R}^{N_h}. \tag{11.1}$$

Für A sym. pos. definit kann man folgende Normen im \mathbb{R}^{N_h} definieren:

$$\|x\|_s := (x, A^s x)^{\frac{1}{2}} \tag{11.2}$$

Dabei erhält man für $s = 0, 1, 2$:

$$\begin{aligned}
s = 0 \quad |||x|||_0 &= (x, x)^{\frac{1}{2}} = \|x\| && \text{Euklidische Norm,} \\
s = 1 \quad |||x|||_1 &= (x, Ax)^{\frac{1}{2}} && \text{Energienorm,} \\
s = 2 \quad |||x|||_2 &= (x, A^2x)^{\frac{1}{2}} = (Ax, Ax)^{\frac{1}{2}} = \|Ax\| && \text{Defektnorm.}
\end{aligned}$$

Der Name für $s = 2$ beruht für $Ax = b$ auf der Identität

$$|||x - x^k|||_2 = (A(x - x^k), A(x - x^k))^{\frac{1}{2}} = (b - Ax^k, b - Ax^k)^{\frac{1}{2}} = \|b - Ax^k\|.$$

Die Norm $|||x|||_s$ kann auch für nichtganzzahliges $s \in \mathbb{R}$ definiert werden. Zu symmetrisch positiv definitem A gibt es eine unitäre Matrix Q (d.h. $Q^{-1} = Q^T$) und eine Diagonalmatrix $D = \text{diag}(\lambda_1, \dots, \lambda_N)$ so dass $A = Q^T D Q$ mit $0 < \lambda \in \mathbb{R}$.

Man setzt dann:

$$A^s := Q^T D^s Q \quad \text{mit } D^s = \text{diag}(\lambda_1^s, \dots, \lambda_N^s). \quad (11.3)$$

Damit erhält man dann allgemein

$$\begin{aligned}
|||x|||_s &= (x, A^s x)^{\frac{1}{2}} = (x, Q^T D^s Q x)^{\frac{1}{2}} \\
&= (x, Q^T D^{\frac{s}{2}} Q Q^T D^{\frac{s}{2}} Q x)^{\frac{1}{2}} = (Q^T D^{\frac{s}{2}} Q x, Q^T D^{\frac{s}{2}} Q x)^{\frac{1}{2}} \\
&= \|A^{\frac{s}{2}} x\|.
\end{aligned} \quad (11.4)$$

Nun wollen wir die Sobolevnormen $\|u_h\|_{0,\Omega}$, $\|u_h\|_{1,\Omega}$ mit entsprechenden Normen $|||x|||_0$, $|||x|||_1$ der Koeffizienten in Beziehung setzen. Wegen $|||x|||_1 = (x, A_h x)^{\frac{1}{2}} = a(u_h, u_h)^{\frac{1}{2}}$ und $\alpha \|u_h\|_1^2 \leq a(u_h, u_h) \leq C \|u_h\|_1^2$ gilt also für A_h aus der FE-Methode:

$$c_1 \|u_h\|_1 \leq |||x|||_1 \leq c_2 \|u_h\| \quad (11.5)$$

mit von h unabhängigen Konstanten c_1, c_2 für jede Basis $\Psi_h = \{\psi_1, \dots, \psi_{N_h}\}$.

Für den Fall $s = 0$ hängt der Zusammenhang von der Wahl der Basis ab.

Lemma 11.1. Es sei $\Psi_h = \{\psi_1, \dots, \psi_{N_h}\}$ die Standard-Knotenbasis auf einer Familie *uniformer* Triangulierungen \mathcal{T}_h des Gebietes $\Omega \subset \mathbb{R}^n$ sowie $u_h = \sum_{i=1}^{N_h} (x_h)_i \psi_i$. Dann gilt mit h -unabhängigen Zahlen c_1, c_2 :

$$c_1 h^{\frac{n}{2}} |||x_h|||_0 \leq \|u_h\|_{0,\Omega} \leq c_2 h^{\frac{n}{2}} |||x_h|||_0. \quad (11.6)$$

Beweis: Sei \hat{T} der n -dimensionale Referenzsimplex. Auf \hat{T} sei $\hat{\psi}_0, \dots, \hat{\psi}_n$ die entsprechende Knotenbasis. Dann gilt für $\hat{u} = \sum_{i=0}^n x_i \hat{\psi}_i$ auf dem Referenzelement:

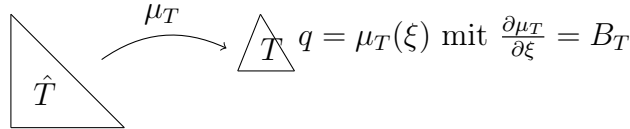
$$\begin{aligned}
\|\hat{u}\|_{0,\hat{T}}^2 &= (\hat{u}, \hat{u})_{0,\hat{T}} = \left(\sum_{i=0}^n x_i \hat{\psi}_i, \sum_{j=0}^n x_j \hat{\psi}_j \right)_{0,\hat{T}} = \sum_{i=0}^n \sum_{j=0}^n x_i x_j \underbrace{(\hat{\psi}_i, \hat{\psi}_j)_{0,\hat{T}}}_{=: \hat{M}_{ij}} \\
&= (x, \hat{M} x) = (Q^T D^{\frac{1}{2}} Q x, Q^T D^{\frac{1}{2}} Q x) = \|\hat{M}^{\frac{1}{2}} x\|^2.
\end{aligned}$$

Hierbei ist \hat{M} die *Massenmatrix* auf dem Referenzelement. \hat{M} ist symmetrisch und positiv definit und kann daher mittels $Q^T D^{\frac{1}{2}} Q$ diagonalisiert werden (mit anderen Q , D wie oben). Die Einträge von \hat{M} bzw. dessen Eigenwerte lassen sich *einmal* ausrechnen da wir uns auf dem Referenzelement befinden. Somit gilt (Raleigh-Quotienten):

$$c_1 \|x\|^2 \leq \|\hat{u}\|_{0,\hat{T}}^2 = \|\hat{M}^{\frac{1}{2}} x\|^2 \leq c_2 \|x\|^2$$

mit c_1, c_2 nur abhängig vom Referenzelement.

Nun transformieren wir wie üblich von \hat{T} auf das allgemeine Element T :



Also:

$$\begin{aligned} \|u\|_{0,T}^2 &= \int_T u^2(q) dq = \int_{\hat{T}} u^2(\mu_T(\xi)) |\det B_T| d\xi = |\det B_T| \int_{\hat{T}} \hat{u}^2(\xi) d\xi \\ &= |\det B_T| \|\hat{u}\|_{0,\hat{T}}^2 = |\det B_T| \|\hat{M}^{\frac{1}{2}} x\|^2. \end{aligned}$$

Unter der Annahme *uniformer* Gitter gilt:

$$\|u\|_{0,\Omega}^2 = \sum_{T \in \mathcal{T}_h} \|u\|_{0,T}^2 = \sum_{T \in \mathcal{T}_h} |\det B_T| \|\hat{M}^{\frac{1}{2}} x_T\|^2 \leq \max_{T \in \mathcal{T}_h} \{|\det B_T|\} c_2 \sum_{T \in \mathcal{T}_h} \|x_T\|^2 \leq ch^n \|x\|^2.$$

Bei der letzten Abschätzung wurde verwendet, dass jede Basisfunktion nur in endlich vielen Dreiecken (unabhängig von h) vorkommt. Wurzelziehen liefert die behauptete Abschätzung nach oben.

Analog erhält man die Abschätzung nach unten:

$$\|u\|_{0,\Omega}^2 = \sum_{T \in \mathcal{T}_h} \|u\|_{0,T}^2 = \sum_{T \in \mathcal{T}_h} |\det B_T| \|\hat{M}^{\frac{1}{2}} x_T\|^2 \geq \min_{T \in \mathcal{T}_h} \{|\det B_T|\} c_1 \sum_{T \in \mathcal{T}_h} \|x_T\|^2 \geq c'h^n \|x\|^2.$$

□

11.2 Analyse der Grobgitterkorrektur

Lemma 11.2 (Approximationseigenschaft). Sei $A_h x_h = b_h$ das zu lösende System, $x_h^{k,1}$ die Näherung nach der Glättung und $x_h^{k,2}$ die Näherung nach der Grobgitterkorrektur. Zusätzlich sei das zugrundeliegende Variationsproblem H^2 -regulär. Dann gilt

$$\| \|x_h - x_h^{k,2}\|_0 \| \|x_h - x_h^{k,1}\|_2 \| \leq ch^{2-n} \| \|x_h - x_h^{k,1}\|_2 \| \quad (11.7)$$

Beweis:

i) Nach Folgerung 8.9 auf Seite 74 gilt:

$$\|u - u_h\|_{0,\Omega} \leq Ch \|u - u_h\|_{1,\Omega}. \quad (11.8)$$

ii) Für die Grobgitterkorrektur gilt folgende Orthogonalitätseigenschaft:

$$\begin{aligned} a(u_h^{k,1} + w_{2h}, v) &= l(v) & \forall v \in S_{2h}, \\ a(u_h, v) &= l(v) & \forall v \in S_h, \\ a(u_h - u_h^{k,1} - w_{2h}, v) &= 0 & \forall v \in S_{2h}. \end{aligned}$$

Damit zeigt man:

$$\begin{aligned} &\alpha \|u_h - u_h^{k,2}\|_{1,\Omega}^2 \\ &\leq a(u_h - u_h^{k,1} - w_{2h}, u_h - u_h^{k,1} - w_{2h}) \\ &= a(u_h - u_h^{k,1} - w_{2h}, u_h - u_h^{k,1}) - \underbrace{a(u_h - u_h^{k,1} - w_{2h}, w_{2h})}_{=0 \text{ wg. } w_{2h} \in S_{2h}} \\ &= (x_h - x_h^{k,1} - y_{2h}, A_h(x_h - x_h^{k,1})) \\ &\leq \| \|x_h - y_h^{k,1} - y_{2h}\|_0 \| \|x - y_h^{k,1}\|_2 \\ &\leq ch^{-\frac{n}{2}} \|u_h - u_h^{k,1} - w_{2h}\|_{0,\Omega} \| \|x - x_h^{k,1}\|_2 \\ &\leq ch^{-\frac{n}{2}} h \|u_h - u_h^{k,1} - w_{2h}\|_{1,\Omega} \| \|x - x_h^{k,1}\|_2 \end{aligned}$$

Hier haben wir die Abschätzung i) auf das Problem mit der exakten Lösung $u_h - u_h^{k,1} \in S_h$ angewendet, die durch $w_{2h} \in S_{2h}$ approximiert wird (genau das macht die Grobgitterkorrektur).

Kürzen von $\|u_h - u_h^{k,1} - w_{2h}\|_{1,\Omega}$ liefert dann

$$\|u_h - u_h^{k,1} - w_{2h}\|_{1,\Omega} \leq ch^{-\frac{n}{2}} h \| \|x_h - x_h^{k,1}\|_2. \quad (11.9)$$

iii) Schließlich:

$$\begin{aligned} \| \|x_h - x_h^{k,2}\|_0 &= \| \|x_h - x_h^{k,1} - y_{2h}\|_0 \\ &\leq ch^{-\frac{n}{2}} \|u_h - u_h^{k,1} - w_{2h}\|_{0,\Omega} \\ &\leq ch^{-\frac{n}{2}} h \|u_h - u_h^{k,1} - w_{2h}\|_{1,\Omega} \\ &\leq ch^{2-n} \| \|x_h - x_h^{k,1}\|_2 \end{aligned}$$

Wie immer sind die Konstanten c generisch, d.h. an verschiedenen Stellen üblicherweise verschieden. \square

11.3 Glättungseigenschaft

Lemma 11.3. Für die Richardson-Iteration $x_h^\nu = x_h^\nu + \omega(b_h - A_h x_h^\nu)$ gilt mit $\omega = \frac{1}{\lambda_{\max}(A_h)}$

$$\| \|x_h - x_h^\nu\| \|_2 \leq \frac{\lambda_{\max}(A_h)}{\nu} \| \|x_h - x_h^0\| \|_0. \quad (11.10)$$

Beweis: Setze $e_h^\nu = x_h - x_h^\nu$ und rechne

$$\begin{aligned} \| \|e_h^\nu\| \|_2 &= \| A_h e_h^\nu \| = \| A_h (I - \omega A_h)^\nu e_h^0 \| \\ &= \| Q^T D Q (Q^T Q - \omega Q^T D Q)^\nu e_h^0 \| \\ &= \| Q^T D Q [Q^T (I - \omega D) Q]^\nu e_h^0 \| \\ &= \| Q^T D (I - \omega D)^\nu Q e_h^0 \| \\ &\leq \| Q^T \| \| D (I - \omega D)^\nu \| \| Q \| \| e_h^0 \| \\ &= \| D (I - \omega D)^\nu \| \| \|e_h^0\| \|_0. \end{aligned}$$

Hier haben wir im letzten Schritt benutzt, dass $\| Q \| = 1$ (Spektralnorm). Dies sieht man folgendermaßen. Wegen $\| Q z \|^2 = (Q z, Q z) = (z, Q^T Q z) = (z, z) = \| z \|^2$ gilt $\| Q \| = \sup_{z \neq 0} \frac{\| Q z \|}{\| z \|} = 1$. Da $D = \text{diag}\{\lambda_i\}$ mit $\lambda_i > 0$ gilt

$$D(I - \omega D)^\nu = \text{diag}\{\lambda_i(1 - \omega\lambda_i)^\nu\}$$

und

$$\begin{aligned} \| D(I - \omega D)^\nu \| &= \max_{i=1, \dots, N} \lambda_i(1 - \omega\lambda_i)^\nu \\ &= \max_{i=1, \dots, N} \left\{ \lambda_{\max} \frac{\lambda_i}{\lambda_{\max}} \left(1 - \frac{\lambda_i}{\lambda_{\max}} \right)^\nu \right\} \\ &\leq \lambda_{\max} \max_{\xi \in [0, 1]} \xi(1 - \xi)^\nu \\ &= \lambda_{\max} \frac{1}{1 + \nu} \left(\frac{\nu}{\nu + 1} \right)^\nu \\ &\leq \frac{\lambda_{\max}}{\nu}. \end{aligned}$$

□

11.4 Zweigitterkonvergenz

Satz 11.4. Das Zweigitterverfahren mit Richardson-Iteration als Vorglättung erfüllt:

$$\| \|x_h - x_h^{k+1}\| \|_0 \leq \frac{c}{\nu} \| \|x_h - x_h^k\| \|_0$$

mit c unabhängig von h . Für genügend großes ν gilt dann $\rho_1 = \frac{c}{\nu} < 1$.

Beweis: Für A_h in der Standardknotenbasis gilt:

$$(A_h)_{i,j} = a(\psi_j, \psi_i) = \int_{\Omega} \nabla \psi_j \cdot \nabla \psi_i \, dx \leq c \int_{\text{supp}(\psi_j \psi_i)} h^{-1} h^{-1} \, dx = ch^n h^{-2}.$$

Mit dem Satz von Geschgorin erhält man $\lambda_{\max}(A_h) \leq Ch^{d-2}$. Mit Lemma 11.2 und Lemma 11.3 erhält man schließlich

$$\| \|x_h - x_h^{k+1}\| \|_0 \leq ch^{2-d} h^{d-2} \frac{1}{\nu} \| \|x_h - x_h^k\| \|_0.$$

□

11.5 Mehrgitterkonvergenz

Lemma 11.5. Sei ρ_1 die Konvergenzrate des Zweigitterverfahrens. Und ρ_l die Konvergenzrate des Mehrgitterverfahrens mit l Stufen. Dann gilt bei Anwendung des μ -Zyklus ($\gamma = \mu$ in Algorithmus 10.4) die rekursive Ungleichung:

$$\rho_l \leq \rho_1 + (1 + \rho_1) \rho_{l-1}^\mu.$$

Beweis: Im Zweigitterverfahren ist $\hat{u}_l^{k,2} = u_l^{k,1} + \hat{w}_{l-1}$ die Näherung nach der Grobgitterkorrektur. Im Mehrgitterverfahren mit μ -Zyklus sei $u_l^{k,2} = u_l^{k,1} + w_{l-1}^\mu$ die Näherung nach dem μ -fachen Aufruf der Grobgitterkorrektur.

\hat{w}_{l-1} ist die Lösung des Systems

$$a(\hat{w}_{l-1}, v) = l(v) - a(u_l^{k,1}, v) \quad \forall v \in S_{2h}.$$

w_{l-1}^μ ist die genäherte Lösung dieses Systems nach μ Iteration auf Stufe $l-1$ bei Startwert $w_{l-1}^0 = 0$. Damit gilt

$$\| \hat{w}_{l-1} - w_{l-1}^\mu \|_{0,\Omega} \leq \rho_{l-1}^\mu \| \hat{w}_{l-1} - \underbrace{w_{l-1}^0}_{=0} \|_{0,\Omega} = \rho_{l-1}^\mu \| \hat{w}_{l-1} \|_{0,\Omega} = \rho_{l-1}^\mu \| \hat{u}_l^{k,2} - u_l^{k,1} \|_{0,\Omega}.$$

Damit gilt

$$\begin{aligned} \| u_l - u_l^{k+1} \|_{0,\Omega} &\leq \| u_l - u_l^{k,2} \|_{0,\Omega} = \| u_l - \hat{u}_l^{k,2} + \hat{u}_l^{k,2} - u_l^{k,2} \|_{0,\Omega} \\ &\leq \| u_l - \hat{u}_l^{k,2} \|_{0,\Omega} + \| \hat{u}_l^{k,2} - u_l^{k,2} \|_{0,\Omega} \\ &\leq \rho_1 \| u_l - u_l^k \|_{0,\Omega} + \| u_l^{k,1} + \hat{w}_{l-1} - u_l^{k,1} - w_{l-1}^\mu \|_{0,\Omega} \\ &\leq \rho_1 \| u_l - u_l^k \|_{0,\Omega} + \rho_{l-1}^\mu \| \hat{u}_l^{k,2} - u_l^{k,1} \|_{0,\Omega} \\ &= \rho_1 \| u_l - u_l^k \|_{0,\Omega} + \rho_{l-1}^\mu \| \hat{u}_l^{k,2} - u_l + u_l - u_l^{k,1} \|_{0,\Omega} \\ &\leq \rho_1 \| u_l - u_l^k \|_{0,\Omega} + \rho_{l-1}^\mu (\rho_1 \| u_l - \hat{u}_l^{k,2} \|_{0,\Omega} + \| u_l - u_l^{k,1} \|_{0,\Omega}) \\ &\leq [\rho_1 + \rho_{l-1}^\mu (1 + \rho_1)] \| u_l - u_l^k \|_{0,\Omega}. \end{aligned}$$

Diese Analyse gilt für das Mehrgitterverfahren mit Vor- und Nachglättung. Allerdings haben wir hier nur benutzt, dass die Nachglättung den Fehler nicht vergrößert. □

Satz 11.6. Ist $\rho_1 \leq \frac{1}{5}$ so gilt für den W-Zyklus (also $\mu = 2$), dass

$$\rho_l \leq \frac{1}{3} \quad \text{für } l \geq 2.$$

Beweis: Für $l = 1$ gilt $\rho_1 \leq \frac{1}{5} \leq \frac{1}{3}$. Nun gelte die Behauptung bis $l - 1$. Mit der Rekursionsformel gilt dann

$$\rho_l \leq \frac{1}{5} + \left(1 + \frac{1}{5}\right) \left(\frac{1}{3}\right)^2 = \frac{1}{5} + \frac{6}{5 \cdot 9} = \frac{9 + 6}{45} = \frac{15}{45} = \frac{1}{3}.$$

□

Bemerkung 11.7. Dies ist der simpelste Beweis für h -unabhängige Konvergenz des Zweigitter- und Mehrgitterverfahrens.

Bessere Beweise zeigen bereits h -unabhängige Konvergenz mit *einem* Glättungsschritt und V-Zyklus ($\mu = 1$) ohne Regularitätsannahmen an das Problem (d. h. die hier benutzte H^2 -Regularität ist nicht notwendig). Probleme in der Praxis sind allerdings Robustheit der Konvergenzrate gegenüber Parametern wie dem Diffusionskoeffizienten (starke Variabilität, Anisotropie), Gitteranisotropie, etc. □

11.6 Komplexität

Lemma 11.8. Sei N_l die Anzahl der Unbekannten auf Stufe l . Dann gilt für den Aufwand $A(N_l)$ für einen Zyklus des Mehrgitterverfahrens $MG(\nu_1, \nu_2, \mu)$ in n Raumdimensionen

$$A(N_l) = O(N_l),$$

wenn $n \geq 1$ für $\mu = 1$, bzw. $n \geq 2$ für $\mu = 2$.

Beweis: Bei uniformer Verfeinerung gilt $N_l/N_{l-1} = w = 2^n$ und somit

$$\begin{aligned} A(N_l) &\leq CN_l + \mu CN_{l-1} + \mu^2 CN_{l-2} + \dots \\ &= CN_l + \mu C \frac{N_l}{w} + \frac{\mu^2}{w^2} CN_l + \dots \\ &= CN_l \left(1 + \frac{\mu}{w} + \left(\frac{\mu}{w}\right)^2 + \dots\right). \end{aligned}$$

Die geometrische Reihe konvergiert, wenn $\mu/w = \mu/2^n < 1 \Leftrightarrow 2^n > \mu$. Also ab $n = 1$ für den V-Zyklus und ab $n = 2$ auch für den W-Zyklus. □

Damit ist gezeigt (hier nur für den uniformen Fall, aber eine genauere Analyse zeigt dies auch für lokal hierarchisch verfeinerte Gitter), dass der Gesamtaufwand für einen Schritt proportional zur Anzahl der Unbekannten auf dem feinsten Gitter ist. Da die Anzahl der Schritte nicht von der Gitterweite abhängt ist das Gesamtverfahren auch $O(N)$.


Ende und Ausblick

Wir haben behandelt:

- Existenz und Eindeutigkeit schwacher Lösungen elliptischer Randwertprobleme.
- Konforme Finite Elemente zur numerischen Lösung.
- A priori Fehlerabschätzungen und Approximationssätze.
- A posteriori Fehlerschätzung zur Gitteradaption.
- Effiziente Lösung der linearen Gleichungssysteme mit Mehrgitterverfahren.

Was man noch machen kann:

- nichtkonforme Finite Elemente Methoden, d. h. $v \in \tilde{S}_h$ ist *nicht* stetig. Damit ist $v \notin H_0^1(\Omega)$. Trotzdem machen solche Räume Sinn, Beispiel ist das Crouziex-

Raviart Element: 

- Gemischte Methoden

$$\nabla \cdot q = f, \quad q = -K \nabla p, \quad \text{„System 1. Ordnung“}$$

Getrennte Ansatzräume für p (skalar) und q (Vektorfunktionen!) liefern eine bessere Genauigkeit für q .

- Weitere Fehlerschätzer und Löser, etwa Teilraumkorrekturverfahren, Gebietszerlegungsverfahren sowie deren Parallelisierung.
- Anwendungen: (Strömungs-) Mechanik, gekoppelte Systeme.

Literatur

- [Arg57] ARGYRIS, J. H.: *Die Matrizenmethode der Statik*. Ingenieurarchiv, XXV:174–194, 1957.
- [BA76] BABUSKA, I. und A.K. AZIZ: *On the angle condition in the finite element method*. SIAM J. Numer. Anal., 13(2):214–226, 1976.
- [Bas08] BASTIAN, PETER: *Grundlagen der Modellbildung und Simulation*. Vorlesungsskript, 2008.
- [Bra91] BRAESS, D.: *Finite Elemente*. Springer, 1991.
- [Cou43] COURANT, R.: *Variational methods for the solution of problems of equilibrium and vibrations*. Bull. Amer. Math. Soc., 49:1–23, 1943.
- [Fey70] FEYNMAN, R. P.: *Feynman Lectures on Physics*, Band II. Addison-Wesley, 1970.
- [Hac86] HACKBUSCH, W.: *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, 1986. http://www.mis.mpg.de/scicomp/articleshackbusch_d.html.
- [Ran06] RANNACHER, R.: *Einführung in die Numerische Mathematik II (Numerik partieller Differentialgleichungen)*. <http://numerik.iwr.uni-heidelberg.de/~lehre/notes>, 2006.
- [RR93] RENARDY, M. und R. C. ROGERS: *An Introduction to Partial Differential Equations*. Springer-Verlag, 1993.
- [TCMT56] TURNER, M. J., R. M. CLOUGH, H. C. MARTIN und L. J. TOPP: *Stiffness and deflection analysis of complex structures*. J. Aeron. Sci., 23:805–823, 1956.