THEORIE UND NUMERIK ELLIPTISCHER RANDWERTPROBLEME

(Sommersemester 2006)

G. Lube
Georg-August-Universität Göttingen, NAM

21. Juli 2006

Inhaltsverzeichnis

Kapitel 0

Einleitung

Die Vorlesung *Theorie und Numerik elliptischer Differentialgleichungen* steht am Anfang eines Einführungszyklus in die Theorie und Numerik partieller Differentialgleichungen. Schwerpunktmäßig werden *elliptische Randwertprobleme* behandelt, die einfache stationäre Modelle der mathematischen Physik umfassen. Die Vorlesung setzt die Kenntnis der Anfängervorlesungen und möglichst (jedoch nicht zwingend) einer Vorlesung über lineare Funktionalanalysis voraus. Sie wendet sich an StudentInnen mittlerer Semester der Mathematik und auch der Physik. Die Vorlesung besteht aus vier Teilen.

In **Teil I** stellen wir grundlegende Aussagen elliptischer Randwertprobleme (RWP) und Grundideen ihrer numerischen Lösung vor. Dies erfolgt zunächst exemplarisch am Beispiel von *Zweipunkt-RWP* gewöhnlicher Differentialgleichungen. Anschließend erweitern wir die Darstellung auf den mehrdimensionalen Fall und klassifizieren partielle Differentialgleichungen 2. Ordnung. Dann besprechen wir genauer Randwertprobleme der Poisson-Gleichung als Prototyp elliptischer Probleme. Neben analytischen Lösungsmethoden behandeln wir auch einführend die Finite-Differenzen-Methode. Es schließen sich Aussagen zur klassischen und verallgemeinerten Lösungstheorie elliptischer RWP an.

Im **Teil II** der Vorlesung widmen wir uns dann genauer Finite-Elemente-Methoden (FEM) zur Lösung elliptischer RWP. Nach Darlegung des abstrakten Zugangs werden detailliert praktische Fragen der FEM wie Konstruktion von diskreten Unterräumen und Generierung der diskreten Probleme angesprochen. Ein wesentlicher neuer Aspekt dieser Lehrveranstaltung ist, daß vor allem in den Übungen mit dem neuen Programmsystem **FEMLAB** der praktische Umgang mit einem Finite-Elemente-Programm erlernt werden soll. Dieses vielseitige Programmsystem basiert wesentlich auf MATLAB, ist relativ modern und recht gut dokumentiert. Einen weiteren Schwerpunkt von Teil II der Vorlesung bilden Fragen der numerischen Analysis konformer Verfahren. Untersucht werden vorwiegend mit Mitteln der Funktionalanalysis die Existenz, Eindeutigkeit und Konvergenz diskreter Lösungen. Weiterhin diskutieren wir die Erweiterung auf nichtkonforme Verfahren sowie die adaptive Gitterverfeinerung.

Im **Teil III** der Vorlesung werden schließlich ausgehend von einer kritischen Betrachtung der grundlegenden direkten und iterativen Lösungsmethoden *schnelle Lösungsverfahren* für die entstehenden linearen Gleichungssysteme (z.B. CG-artige Verfahren, Mehrgitter-Verfahren) behandelt.

Der abschließende **Teil IV** befaßt sich mit Erweiterungen, zum Beispiel auf Probleme mit dominanter Konvektion sowie gemischte Probleme.

Im Rahmen dieser Vorlesung orientiere ich mich vorwiegend an folgenden Texten zu Finite-Elemente-Verfahren:

[4] S.C. Brenner, L.R. Scott: *The Mathematical Theory of Finite Elements*, Springer-Verlag, Berlin - Heidelberg - New York 2002

[5] D. Braess: *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elasitizitätstheorie*, Springer-Verlag, Berlin - Heidelberg - New York 2003,

bzw.

[10] Ch. Großmann, H.G. Roos: *Numerische Behandlung partieller Differentialgleichungen* Teubner, Stuttgart 2005.

[9] H. Goering, , H.G. Roos, L. Tobiska: *Finite-Element-Methoden. Eine Einführung,* Akademie-Verlag, Berlin 1993.

Hinsichtlich einer Darstellung der klassischen und verallgemeinerten Lösungstheorie elliptischer Randwertprobleme 2. Ordnung verweise ich auf

[1] H.W. Alt: *Lineare Funktionalanalysis*, Springer-Lehrbuch Berlin - Heidelberg - New York 1999,

[7] D. Gilbarg, N.S. Trudinger: *Elliptic partial differential equations of second order,* Springer-Verlag, Berlin - Heidelberg - New York 1998,

[14] R. Kreß: *Linear Integral Equations*, Springer, Berlin-Heidelberg 1999,

[24] W.S. Wladimirow: *Gleichungen der mathematischen Physik*, Verlag der Wissenschaften, Berlin 1972.

Ergänzend kann man auch heranziehen:

[11] W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen.* Teubner, Stuttgart 1986.

Die Übungen zur Vorlesung dienen einerseits der Vertiefung der theoretischen Aussagen, andererseits sollen exemplarisch grundlegende Rechenfertigkeiten vermittelt werden. Wie schon oben betont soll auch sehr intensiv mit dem FEM-Programm **FEMLAB** gearbeitet werden.

Die Vorlesung wird in nachfolgenden Semestern fortgesetzt mit der numerischen Behandlung zeitabhängiger partieller Differentialgleichungen und nichtlinearer Probleme.

Wichtige Anmerkung:
Im Sommersemester 2006 fallen leider einige Vorlesungstermine durch Feiertage aus. Daher werden einige Kapitel des Skriptes entweder gar nicht behandelt oder nur kurz im Überblick gestreift.

- Kapitel 1 zu Zweipunkt-Randwertaufgaben wurde bereits weitgehend in Teil I der Vorlesung Numerische Mathematik II im Wintersemester 2004/05 behandelt.

Teil I

Elliptische Randwertaufgaben

Kapitel 1

Zweipunkt-Randwertaufgaben

Als Einführung in die Problematik elliptischer Randwertaufgaben (RWP) 2. Ordnung behandeln wir in diesem Kapitel den einfachsten Fall. Im eindimensionalen Fall führt das auf sogenannte Zweipunkt-Randwertprobleme für gewöhnliche Differentialgleichungen 2. Ordnung. Dabei gehen wir vom klassischen Lösungsbegriff aus, betrachten die Approximation der Lösung des kontinuierlichen Problems mit Hilfe einer Finite-Differenzen-Diskretisierung und führen einige Grundbegriffe der Numerischen Analysis derartiger Verfahren ein.

1.1 Einführendes Beispiel. Definitionen

Die Modellierung wichtiger Vorgänge in Naturwissenschaft und Technik führt auf nichtlineare partielle Differentialgleichungen. Bei starker Vereinfachung kann man mitunter das Modell auf die Bestimmung einer wesentlichen Modellvariablen als Lösung einer gewöhnlichen Differentialgleichung reduzieren (vgl. dazu Beispiel 1.2).

Wir geben jetzt eine *Klassifikation* nichtlinearer gewöhnlicher Differentialgleichungen 2. Ordnung an.

Definition 1.1. *Eine gewöhnliche Differentialgleichung 2. Ordnung der Form*

$$F(x, u(x), u'(x), u''(x)) = 0. \tag{1.1}$$

für eine gesuchte Funktion $u = u(x)$ heißt

- quasilinear, *falls* $F(x,u,u',u'') := -u'' + B(x,u)u' + C(x,u) = 0,$
- semilinear, *falls* $F(x,u,u',u'') := -u'' + b(x)u' + C(x,u) = 0,$
- *bzw.* linear, *falls* $F(x,u,u',u'') := -u'' + b(x)u' + c(x)u - f(x) = 0.$

Die eindeutige Festlegung der Lösung gewöhnlicher Differentialgleichungen erfolgt bei zahlreichen Anwendungen durch Vorgabe von Zusatzbedingungen an einem Anfangspunkt, sogenannten *Anfangsbedingungen.* Derartige *Anfangswertprobleme* sind jedoch nicht Gegenstand dieser Vorlesung. Vielmehr betrachten wir hier die Vorgabe von Zusatzbedingungen an die Lösung am Rand des betrachteten Lösungsintervalls, sogenannte *Randbedingungen.* Für die Differentialgleichung (1.1) erhält man *Zweipunkt-Randwertaufgaben.*

Beispiel 1.2. Wir betrachten einen isothermen rotationssymmetrischen Strömungsreaktor der Länge L mit kontinuierlicher Zu- bzw. Abfuhr der Reaktionsmasse bzw. des -produktes. Die gesuchte Konzentrationsverteilung $c(x_1, x_2, x_3, t)$ im Reaktor ergibt sich aus der Stoffbilanzgleichung

$$\frac{\partial c}{\partial t} + \sum_{i=1}^{3} \frac{\partial}{\partial x_i}(w_i c) - \sum_{i=1}^{3} \frac{\partial}{\partial x_i}\left(D \frac{\partial c}{\partial x_i}\right) + r(c) = 0$$

Gegebene Daten seien dabei das Geschwindigkeitsfeld $\mathbf{w} = (w_i)_i$ der Strömung im Reaktor, der Diffusionskoeffizient D und der Reaktionsterm $r(c)$. Vereinfachend nehmen wir einen stationären Reaktorbetrieb, d.h. $\frac{\partial c}{\partial t} = 0$, konstante Diffusionskonstante D und ein konstantes Geschwindigkeitsfeld $\mathbf{w} = (w, 0, 0)$ an. Ferner sollen Änderungen der Konzentration c nur in axialer Richtung x des rotationssymmetrischen Reaktors betrachtet werden. Dann vereinfacht sich die Stoffbilanzgleichung zur gewöhnlichen Differentialgleichung 2. Ordnung

$$-D\frac{d^2c}{dx^2} + w\frac{dc}{dx} + r(c) = 0, \quad 0 < x < L.$$

Entdimensionierung mittels $\xi := \frac{x}{L}, u := \frac{c}{c_0}$ mit der Ausgangskonzentration c_0 ergibt

$$-\frac{1}{P}\frac{d^2u}{d\xi^2} + \frac{du}{d\xi} + R(u) = 0, \quad 0 < \xi < 1.$$

mit der Peclet–Zahl $P := \frac{wL}{D}$, die als wichtige Kennzahl das Verhältnis von konvektiver zu diffusiver Änderung der Konzentration (bezogen auf die Reaktorlänge L) beschreibt. (Diese Zahl spielt auch in der Numerischen Analysis geeigneter Diskretisierungsverfahren eine zentrale Rolle.)

Die Lösung wird vereinfachend durch folgende Randbedingungen festgelegt:

$$u(0) - \frac{1}{P}\frac{du(0)}{d\xi} = 1, \quad \frac{du(1)}{d\xi} = 0.$$

Offenbar ist die im Beispiel 1.1 betrachtete Gleichung semilinear. Linearisiert man den Reaktionsterm $R(u)$ an der Stelle $u = u_0$, so entsteht die lineare Differentialgleichung

$$-\frac{1}{P}\frac{d^2u}{d\xi^2} + \frac{du}{d\xi} + R'(u_0)u = f, \quad 0 < \xi < 1$$

mit der Notation $f(\xi) := R'(u_0)u_0 - R(u_0)$. □

Die Randbedingungen sind im allgemeinen Fall

$$G_i(a, b, u(a), u(b), u'(a), u'(b)) = 0, \quad i = 1, 2$$

nichtlinear und gekoppelt. In Anwendungen ist es oft ausreichend, Randbedingungen in linearer und entkoppelter Form zu betrachten. Dies vereinfacht auch die Untersuchung entsprechender Randwertprobleme (RWP) erheblich.

Definition 1.3. *Lineare und entkoppelte Randbedingungen der Form*

$$u(a) = \alpha, \qquad u(b) = \beta \tag{1.2}$$
$$u'(a) = \alpha, \qquad u'(b) = \beta \tag{1.3}$$
$$c_1u(a) + u'(a) = \alpha, \quad c_2u(b) + u'(b) = \beta \tag{1.4}$$

heißen Randbedingungen 1. Art (oder vom Dirichlet-Typ), 2. Art (oder vom Neumann-Typ) bzw. 3. Art (oder vom Robin-Typ).

Man spricht von *gemischten* Randbedingungen, wenn wie im Beispiel 1.1 bei $x = a$ und $x = b$ unterschiedliche Typen von Randbedingungen gestellt werden.

Bei den weiteren Betrachtungen werden wir vereinfachend **lineare RWP 1. Art**

$$(Lu)(x) := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad a < x < b \tag{1.5}$$
$$u(a) = \alpha, \quad u(b) = \beta \tag{1.6}$$

betrachten. In der Mathematischen Physik beschreibt Gleichung (1.5) in Verbindung mit geeigneten Randbedingungen in stark vereinfachter Form Vorgänge vom *Diffusions-Konvektions-Reaktions-Typ*, vgl. auch Beispiel 1.1. Man kann die Lösung u zum Beispiel als Konzentration oder Temperatur interpretieren.

Mittels

$$u(x) = v(x) + \alpha\frac{x-b}{a-b} + \beta\frac{x-a}{b-a}$$

kann man die Untersuchung auf den Fall *homogener* Randbedingungen $\alpha = \beta = 0$ zurückführen. Über $x = (b-a)\xi + a$ transformiert man das RWP oft auf das Einheitsintervall, d.h. ggf. nach Umbezeichnung $\xi \mapsto x$, $\tilde{u}(\xi) \mapsto u(x)$, $\tilde{L} \mapsto L$ usw.

$$(Lu)(x) := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad 0 < x < 1 \tag{1.7}$$

$$u(0) = u(1) = 0. \tag{1.8}$$

1.2 Klassische Lösbarkeit des RWP 1. Art

Für Zahlen $m \in \mathbf{N_0}$ sei $C^m(0,1)$ der Raum der auf $(0,1)$ $m-$fach stetig differenzierbaren Funktionen und $C^m[0,1]$ die Teilmenge aus $C^m(0,1)$ mit stetig bis auf die Randpunkte $x = 0$ und $x = 1$ fortsetzbaren Ableitungen bis zur Ordnung m.

Wir setzen von jetzt an in diesem Kapitel voraus, daß für die Daten gilt

$$b,\ c,\ f \in C[0,1].$$

Eine Funktion

$$u \in C^2(0,1) \cap C[0,1]$$

heißt *klassische Lösung* des RWP (1.7) - (1.8), wenn diese Gleichungen jeweils punktweise erfüllt sind.

Das nachfolgende Beispiel zeigt nun, daß RWP 2. Ordnung nicht in jedem Fall lösbar oder eindeutig lösbar sind.

Beispiel 1.4. Die allgemeine Lösung der sogenannten Schwingungsgleichung

$$-u''(x) - u(x) = 0, \qquad a < x < b$$

hat die Form $u(x) = c_1 \cos x + c_2 \sin x$. Bei Auferlegung der Randbedingung (1.6) sind die beiden Konstanten c_1, c_2 aus dem Gleichungssystem

$$\cos(a)c_1 + \sin(a)c_2 = \alpha, \quad \cos(b)c_1 + \sin(b)c_2 = \beta,$$

zu bestimmen. Dies führt natürlich im Fall einer nichtverschwindenden Determinante $D := \cos(a)\sin(b) - \sin(a)\cos(b) \neq 0$ zu einer eindeutigen Lösung. Schwierigkeiten entstehen bei verschwindender Determinante. Wir betrachten drei Fälle

a) $u(0) = u(1) = 1$, b) $u(0) = 1, u(\pi) = -2$, c) $u(0) = 1, u(\pi) = -1$.

Dabei zeigt sich, daß das entstehende lineare Gleichungssystem im Fall a) eine eindeutige, im Falle b) keine und schließlich bei c) unendlich viele Lösungen hat. □

Eine hinreichende Existenz- und Eindeutigkeitsaussage klassischer Lösungen für das lineare RWP (1.5)-(1.6) gibt der

Satz 1.5. *(i) Hat das dem RWP (1.5)-(1.6) zugeordnete homogene RWP (d.h. $f(x) \equiv 0,\ \alpha = \beta = 0$) nur die triviale Lösung, so hat das (inhomogene) RWP (1.5)-(1.6) eine und nur eine Lösung in*

$$X := \{v \in C^2(0,1) \cap C[0,1] :\ v(0) = \alpha,\ v(1) = \beta\}.$$

(ii) Ist $c(x) \geq 0$, so hat das (1.5)-(1.6) zugeordnete homogene RWP nur die triviale Lösung.

Die tieferliegende Existenzaussage (i) beweist man im Rahmen einer Lösbarkeitstheorie in Hölder-Räumen (vgl. auch Kap. 2.2) mittels des Fredholmschen Alternativsatzes (vgl. z.B. [7], Th. 6.15. Wir kommen

darauf in wesentlich allgemeinerer Form im Kapitel 4 der Vorlesung zurück.

Zum Beweis des Eindeutigkeitsresultat (ii) verwendet man das folgende wichtige *Maximum-Minimum Prinzip.* Es besagt in Anwendung auf Diffusions-Konvektions-Reaktionsvorgänge, daß die Temperatur oder Konzentration u bei Abwesenheit von äußeren Quellen ($f \equiv 0$) und bei endothermer Reaktion ($c \geq 0$) ihre Maximalwerte nur am Rand des Gebietes annehmen kann. Besonders anschaulich ist der Fall des Diffusionsproblems

$$-u''(x) = 0, \quad 0 < x < 1; \qquad u(0) = \alpha, \quad u(1) = \beta.$$

Die Lösung $u(x) = (\beta - \alpha)x + \alpha$ nimmt offenbar die Extremwerte am Rand an.

Lemma 1.6. *Gelte $b, c \in C(0,1)$ sowie $c(x) \geq 0$. Dann gelten für $u \in C[0,1] \cap C^2(0,1)$ die Aussagen:*

(i) $(Lu)(x) \leq 0$ in $(0,1) \implies u(x) \leq \max\{0; u(0), u(1)\}$

(ii) $(Lu)(x) \geq 0$ in $(0,1) \implies u(x) \geq \min\{0; u(0), u(1)\}$.

Beweis: (1) Für den Differentialoperator $\tilde{L}u := -u'' + bu'$, d.h. $c \equiv 0$, beweisen wir zuerst die Aussagen

(i') $(\tilde{L}u)(x) \leq 0$ in $(0,1) \implies u(x) \leq \max\{u(0), u(1)\}$

(ii') $(\tilde{L}u)(x) \geq 0$ in $(0,1) \implies u(x) \geq \min\{u(0), u(1)\}$.

Wir beschränken uns auf den Nachweis von (i').

(i_1') Sei $(\tilde{L}u)(x) < 0$ in $(0,1)$. Wir nehmen an, daß u ein Maximum in $x_0 \in (0,1)$ annimmt. Wegen $u'(x_0) = 0$ folgt

$$(\tilde{L}u)(x_0) = -u''(x_0) < 0$$

im Widerspruch zur Bedingung $u''(x_0) \leq 0$ für ein Maximum.

(i_2') Sei nun $(\tilde{L}u)(x) \leq 0$ in $(0,1)$. Für die Hilfsfunktion $v(x) := \delta \exp(\lambda x)$ mit $\delta > 0$ gilt

$$(\tilde{L}v)(x) = \lambda(b - \lambda)\delta e^{\lambda x} < 0$$

für geeignetes λ. Wegen $\tilde{L}(u + v)(x) < 0$ ergibt (i_1')

$$(u + v)(x)) \leq \max\{(u + v)(0), (u + v)(1)\}.$$

Im Grenzfall $\delta \to 0$ folgt die gesuchte Aussage.

(2) Sei jetzt $c(x) \geq 0$ in $(0,1)$ Die Punktmenge

$$G^+ := \{x \in (0,1) : \ u(x) > 0\}$$

ist wegen $u \in C[0,1]$ offen und komponentenweise zusammenhängend. Ferner ist nach Voraussetzung

$$(\tilde{L}u)(x) \leq -c(x)u(x) \leq 0 \quad \text{auf } G^+.$$

Anwendung von Aussage (1) auf jeder Zusammenhangskomponente G_i von G^+ zeigt

$$u(x) \leq \max_{x \in \partial G_i} u(x), \quad \forall x \in G^+.$$

Dabei ist ∂G_i der Rand von G_i. Nach Definition von G^+ impliziert dies die gesuchte Aussage

$$u(x) \leq \max\{0; u(0), u(1)\}.$$

(3) Die Minimumaussage (ii) wird analog bewiesen. □

Als Folgerung beweisen wir folgendes Resultat über die *Stabilität* der Lösung bezüglich der Problemdaten f, α, β.

Lemma 1.7. *Seien $b, c, f \in C[0,1]$ und $c(x) \geq 0$. Für Lösungen $u \in C^2(0,1) \cap C[0,1]$ des RWP*

$$(Lu)(x) = f(x), \quad x \in (0,1); \qquad u(0) = \alpha, \;\; u(1) = \beta$$

gilt

$$\|u\|_{C[0,1]} \leq C\|f\|_{C[0,1]} + \max\{|u(0)|, |u(1)|\}.$$

Beweis: Für die Hilfsfunktion

$$v(x) := A - Be^{\lambda x}, \qquad A, B \geq 0$$

mit hinreichend großer Konstante $\lambda > 0$ gilt

$$(Lv)(x) = -Be^{\lambda x}\{c(x) + b(x)\lambda - \lambda^2\} + c(x)A \geq Be^{\lambda x}\{\lambda^2 - \lambda b(x) - c(x)\} \geq B.$$

Daraus folgern wir mit $B := \|f\|_{C[0,1]}$, daß

$$L(v \pm u)(x) \geq B \pm f(x) \geq B - \|f\|_{C[0,1]} = 0.$$

Ferner gilt für die Randwerte $x = 0$ und $x = 1$

$$(v \pm u)(x) = A - Be^{\lambda x} \pm u(x) \;\; \geq \;\; A - Be^{\lambda} - \max\{|u(0)|, |u(1)|\} = 0,$$

sofern $A := \max\{|u(0)|, |u(1)|\} + Be^{\lambda}$. Wegen $L(v \pm u) \geq 0$ in $(0,1)$ und $v \pm u \geq 0$ für die Randpunkte $x = 0$ und $x = 1$ erhalten wir nach dem Lemma 1.6

$$(v \pm u)(x) \geq 0, \quad x \in (0,1).$$

Das ergibt die Behauptung wegen

$$\begin{aligned} |u(x)| &\leq v(x) \\ &\leq A - B \\ &\leq \max\{|u(0)|, |u(1)|\} + B(e^{\lambda} - 1) \\ &\leq \max\{|u(0)|, |u(1)|\} + (e^{\lambda} - 1)\|f\|_{C[0,1]}. \end{aligned}$$ □

Beweis von Satz 1.5. (ii): Die Aussage des Lemmas 1.7 impliziert nun die Eindeutigkeit der Lösung, d.h. die Aussage von Satz 1.5. (ii). □

Nach Aussage von Teil (i) des Satzes 1.5 ergibt sich daraus auch eine Existenzaussage in $C^2(a,b) \cap C[a,b]$ für das RWP (1.5)-(1.6).

1.3 Finite-Differenzen-Verfahren

Im vorliegenden Abschnitt besprechen wir das *klassische Finite Differenzen Verfahren* (FDM) zur Lösung von Zweipunkt–Randwertaufgaben. Bei der FDM ersetzt man Ableitungen in der Differentialgleichung durch Differenzenquotienten. Dies führt dann zu einem linearen Gleichungssystem für Näherungswerte an die gesuchten Werte der Lösung an vorgegebenen Knotenpunkten.

Ausgangspunkt ist das lineare Randwertproblem (RWP)

$$-u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad 0 < x < 1 \tag{1.9}$$

$$u(0) = u(1) = 0. \tag{1.10}$$

Wir betrachten vereinfachend eine *äquidistante* Zerlegung durch die Knotenpunkte $x_i = ih, i = 0, ..., n+1$ mit der Schrittweite $h = \frac{1}{n+1}$ mit $n \in \mathbf{N}$. Zur Approximation der ersten Ableitung $u'(x_i)$ betrachten wir drei Varianten, die auf dem sogenannten *Dreipunktestern* $\{x_{i-1}, x_i, x_{i+1}\}$ basieren.

- Vorwärtsdifferenzen-Quotient: $D^+u(x_i) := \frac{u(x_{i+1})-u(x_i)}{h}$
- Rückwärtsdifferenzen-Quotient: $D^-u(x_i) := \frac{u(x_i)-u(x_{i-1})}{h}$
- Zentraler Differenzen-Quotient: $D^0u(x_i) := \frac{u(x_{i+1})-u(x_{i-1})}{2h}$.

Zur Approximation der zweiten Ableitung $u''(x_i)$ nutzen wir den *zentralen Differenzenquotienten 2. Ordnung*

$$D^+D^-u(x_i) := \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}.$$

Fortan bezeichne U_i eine (zu berechnende) Approximation der kontinuierlichen Lösung $u(\cdot)$ an der Stelle $x = x_i, i = 0, ..., n+1$. Wir erhalten nun bei Approximation der ersten und zweiten Ableitungen in der Differentialgleichung (1.9) durch die zentralen Differenzenquotienten 1. bzw. 2. Ordnung das folgende System

$$-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + b(x_i)\frac{U_{i+1} - U_{i-1}}{2h} + c(x_i)U_i = f(x_i)$$

bzw. mit den Bezeichnungen

$$b_i := b(x_i), \quad c_i := c(x_i), \quad f_i := f(x_i)$$

das folgende System von Differenzengleichungen

$$\frac{1}{h^2}\left[-\left(1 + \frac{b_ih}{2}\right)U_{i-1} + \left(2 + c_ih^2\right)U_i - \left(1 - \frac{b_ih}{2}\right)U_{i+1}\right] = f_i, \quad i = 1, ..., n. \tag{1.11}$$

Hinzu kommen wegen der Randbedingungen (1.10) die Forderungen

$$U_0 = U_{n+1} = 0. \tag{1.12}$$

Mit den Bezeichnungen $U = (U_1, ..., U_n)^*, F = (f_1, ..., f_n)^*$ und

$$A := \frac{1}{h^2}\text{tridiag}\left\{-\left(1 + \frac{b_ih}{2}\right); (2 + c_ih^2); -\left(1 - \frac{b_ih}{2}\right)\right\}$$

ergibt sich aus (1.11)-(1.12) das lineare Gleichungssystem

$$AU \;=\; F. \tag{1.13}$$

Von Interesse sind nun folgende Fragen:

- Lösbarkeit des diskreten Problems (1.13)
- Konvergenz der Lösung von (1.13) für $h \to 0$ gegen die Lösung der Zweipunkt–Randwertaufgabe (1.9)-(1.10).

Eine hinreichende Lösbarkeitsbedingung für das diskrete Problem (1.13) gibt

Satz 1.8. *Für das Problem (1.9)-(1.10) gelte*

$$c_i = c(x_i) \geq 0, \qquad \left|\frac{b_ih}{2}\right| \leq 1, \quad i = 1, ..., n. \tag{1.14}$$

Dann hat das zugehörige klassische Finite-Differenzen Schema (1.11)-(1.12) bzw. (1.13) eine und nur eine Lösung $U = (U_1, ..., U_n)^*$.

Bemerkung 1.9. Im Fall $b_i \neq 0$ impliziert Bedingung (1.14) eine Beschränkung der Schrittweite h. Die Größe $P_i := \left|\frac{b_ih}{2}\right|$ heißt auch Gitter-Peclet-Zahl (vgl. Bezeichnung in Beispiel 1.2). □

Beweis von Satz 1.8.: Die Invertierbarkeit von A impliziert die eindeutige Lösbarkeit von (1.13). Die Matrix A ist unter Voraussetzung (1.14) schwach diagonal-dominant, denn

$$|2 + c_i h^2| \geq \left|1 + \frac{b_i h}{2}\right| + \left|1 - \frac{b_i h}{2}\right| = 2, \quad i = 1, ..., n.$$

Im Fall $c(x) \geq c_0 > 0$ ist die Matrix A sogar strikt diagonal-dominant und damit invertierbar.
Im allgemeinen Fall $c(x) \geq 0$ impliziert die schwache Diagonaldominanz mit der Irreduzibilität die Invertierbarkeit von A. Die Irreduzibilität erhält man, da nicht gleichzeitig die Nebendiagonalelemente $1 \pm \frac{b_i h}{2}$ verschwinden können. □

Unter den Voraussetzungen von Satz 1.8 ist das diskrete Problem durch die einfachsten iterativen Verfahren (wie Gesamt–und Einzelschritt Verfahren, SOR) lösbar. Ein derartiger Zugang ist auch beim allgemeineren Problem von Randwertaufgaben bei partiellen Differentialgleichungen für die dort entstehenden sehr großen und schwachbesetzten linearen Gleichungssysteme oft erforderlich.

Aufgrund der sehr speziellen Tridiagonalstruktur der Matrix A erweist sich aber hier die direkte Lösung mittels $LR-$Zerlegung als wesentlich efffizienter. Wir betrachten dazu allgemeiner Tridiagonalmatrizen

$$A = \text{tridiag}\,(B_i, A_i, C_i)_{i=1}^n, \quad B_1 = C_n = 0.$$

Für die $LR-$Zerlegung setzen wir an

$$A = LR, \quad L = \text{tridiag}(B_i; \alpha_i; 0), \quad R = \text{tridiag}(0; 1; \gamma_i).$$

Ausmultiplizieren auf der Hauptdiagonalen ergibt die Beziehungen

$$A_1 = \alpha_1; \quad A_i = \alpha_i + B_i \gamma_{i-1}, \;\; i = 2, ..., n,$$

auf der oberen Nebendiagonalen entsteht

$$C_i = \gamma_i \alpha_i, \;\; i = 1, ..., n-1.$$

Dies ermöglicht eine rekursive Berechnung der Größen α_i und γ_i über

$$\alpha_1 = A_1; \quad \gamma_{i-1} = \frac{C_{i-1}}{\alpha_{i-1}}, \;\; \alpha_i = A_i - B_i \gamma_{i-1}, \quad i = 2, ..., n.$$

Die Realisierbarkeit dieses Verfahrens ist bei $\alpha_i \neq 0, \;\; i = 1, ..., n$ gesichert (vgl. Lemma 1.10).

Wir erhalten damit den folgenden *Thomas–Algorithmus:*

1. $LR-$Zerlegung von A, d.h. Bestimmung von α_i, γ_i
2. Löse das gestaffelte System $LZ = F$ durch *Vorwärtseinsetzen*
   $Z_1 = \frac{f_1}{\alpha_1}, \;\; Z_i = \frac{1}{\alpha_i}(f_i - B_i Z_{i-1}), \;\; i = 2, ..., n$
3. Löse das gestaffelte System $RU = z$ durch *Rückwärtseinsetzen*
   $U_n = Z_n, \;\; U_i = Z_i - \gamma_i U_{i+1}, \;\; i = n-1, ..., 1.$

Eine hinreichende Lösbarkeitsbedingung liefert das

Lemma 1.10. *Für die Matrix $A =$ tridiag (B_i, A_i, C_i) gelte*

$$|A_1| > |C_1| > 0; \quad |A_i| \geq |B_i| + |C_i| > 0, \quad i = 2, ..., n-1. \tag{1.15}$$

Dann ist die Matrix A nichtsingulär. Für die Koeffizienten der $LR-$Zerlegung gilt

$$|\gamma_i| < 1, \;\; i = 1, ..., n-1; \quad \alpha_i \neq 0, \;\; i = 1, ..., n.$$

Beweis: Für den Beweis mittels vollständiger Induktion ist der Induktionsanfang erfüllt wegen

$$|\gamma_1| = \left|\frac{C_1}{A_1}\right| < 1, \quad |\alpha_1| = |A_1| > 0.$$

Wir nehmen nun an, daß gilt

$$|\gamma_{i-1}| < 1, \quad \alpha_{i-1} \neq 0.$$

Über die Rekursionsbeziehungen und Voraussetzungen des Verfahrens ergibt sich

$$|\alpha_i| = |A_i - B_i\gamma_{i-1}| \geq |A_i| - |B_i||\gamma_{i-1}| > |A_i| - |B_i| \geq |C_i| > 0$$

sowie

$$|\gamma_i| = \left|\frac{C_i}{A_i - B_i\gamma_{i-1}}\right| < \frac{|C_i|}{|A_i| - |B_i|} \leq 1.$$

Damit ist die $LR-$Zerlegung realisierbar und die Matrix A regulär. □

Bemerkung 1.11. Für den Thomas-Algorithmus benötigt man $0(n)$ wesentliche Operationen, d.h. der Rechenaufwand ist asymptotisch für $n \to \infty$ optimal. □

1.4 Stabilitäts- und Konvergenzanalyse

Wir führen hier die für die Fehleranalyse des Verfahrens wesentlichen Begriffe ein. Sie sind so allgemein gehalten, daß sich die Analyse auf Differenzen-Verfahren von Randwertaufgaben im mehrdimensionalen Fall übertragen läßt.

Seien $\omega_h := \{x_1, ..., x_n\}$ die Menge der inneren Knotenpunkte im Intervall $(0, 1)$ und $\gamma_h := \{x_0, x_{n+1}\}$ die Randpunkte. $R_h v$ bezeichne die Einschränkung von $v \in C[0, 1]$ auf ω_h und L den Differentialoperator des Randwertproblems. $u \in C^2(0, 1) \cap C[0, 1]$ bzw. $U \in \mathbf{R}^n$ sind die Lösungen des Randwertproblems bzw. des diskreten Problems. Dann gilt für den Diskretisierungsfehler $R_h u - U$

$$A(R_h u - U) = AR_h u - AU = AR_h u - F = AR_h u - R_h Lu. \tag{1.16}$$

Der letzte Term wird auch als *Defekt* bezeichnet.

Zur Fehlerabschätzung sind nun sowohl eine Abschätzung des Defekts nach oben (*Konsistenzanalyse*) als auch eine Abschätzung des links stehenden Terms nach unten (*Stabilitätsanalyse*) in einer geeigneten Norm erforderlich. Bei unseren Untersuchungen verwenden wir die Maximum–Norm

$$\|V\|_{\infty,\omega_h} := \max_{i=1,...,n} |v_i| \quad \text{für } V = (v_1, ..., v_n)^*. \tag{1.17}$$

Dies führt auf die

Definition 1.12. (i) *Eine FDM heißt* konsistent in der Maximum-Norm , *falls*

$$\lim_{h\to 0} \|AR_h u - R_h Lu\|_{\infty,\omega_h} = 0. \tag{1.18}$$

(ii) Die FDM hat die Konsistenzordnung p, *falls mit einer von h unabhängigen Konstanten $C_K > 0$ gilt*

$$\|AR_h u - R_h Lu\|_{\infty,\omega_h} \leq C_K h^p. \tag{1.19}$$

Der Konsistenzbegriff beschreibt, wie gut der Differentialoperator durch das Differenzenverfahren approximiert wird.

Definition 1.13. *Eine FDM heißt* stabil in der Maximum-Norm, *falls für jeden Vektor W mit*

$$AW = F \ \text{ in } \ \omega_h, \quad W = 0 \ \text{ in } \ \gamma_h$$

die Existenz einer von h und W unabhängigen Konstanten C_S folgt mit

$$\|W\|_{\infty,\omega_h} = \|A^{-1}F\|_{\infty,\omega_h} \leq C_S\|F\|_{\infty,\omega_h}. \tag{1.20}$$

Definition 1.14. (i) *Eine FDM heißt* konvergent in der Maximum-Norm , *falls*

$$\lim_{h\to 0} \|R_h u - U\|_{\infty,\omega_h} = 0. \tag{1.21}$$

(ii) Die FDM hat die Konvergenzordnung p, *falls mit einer von h unabhängigen Konstanten $M > 0$ gilt*

$$\|R_h u - U\|_{\infty,\omega_h} \leq Mh^p. \tag{1.22}$$

(i) *Konsistenz der klassischen FDM*

Die Abschätzung des Konsistenzfehlers der klassischen FDM (1.11)-(1.12) für das RWP (1.9),(1.10) erfolgt über den Satz von Taylor. Zunächst betrachten wir die Genauigkeit der Approximation der auftretenden Ableitungen durch zentrale Differenzenquotienten.

Lemma 1.15. *Es gilt*

$$(i) \quad (D^0u)(x) = u'(x) + h^2R(x), \quad |R(x)| \leq \frac{1}{6}\|u^{(3)}\|_{C[0,1]}, \quad \textit{falls } u \in C^3[0,1]$$

bzw.

$$(ii) \quad (D^+D^-u)(x) = u''(x) + h^2R(x), \quad |R(x)| \leq \frac{1}{12}\|u^{(4)}\|_{C[0,1]}, \quad \textit{falls } u \in C^4[0,1].$$

Beweis: Aus der Taylor-Entwicklung an der Stelle x folgt

$$\begin{aligned} u(x \pm h) &= u(x) \pm hu'(x) + h^2\frac{u''(x)}{2} \pm h^3R_3^{\pm} \\ u(x \pm h) &= u(x) \pm hu'(x) + h^2\frac{u''(x)}{2} \pm h^3\frac{u^{(3)}(x)}{6} + h^4R_4^{\pm} \end{aligned}$$

mit

$$\begin{aligned} R_3^{\pm} &= \frac{1}{h^3}\int_x^{x\pm h} [u''(t) - u''(x)]\,(x \pm h - t)dt \\ R_4^{\pm} &= \frac{1}{h^4}\int_x^{x\pm h} \left[u^{(3)}(t) - u^{(3)}(x)\right]\frac{(x \pm h - t)^2}{2}dt. \end{aligned}$$

Dann ergibt sich die Aussage (i) aus

$$(D^0u)(x) = \frac{u(x+h) - u(x-h)}{2h} = u'(x) + \frac{1}{2}h^2\left(R_3^+ - R_3^-\right)$$

und einer Abschätzung der Restglieddifferenz. Aussage (ii) beweist man analog. □

Damit finden wir

Lemma 1.16. *Unter der Voraussetzung $u \in C^4[0,1]$ an die Lösung des RWP (1.9)-(1.10) hat die FDM (1.11)-(1.12) die Konsistenzordnung 2.*

Beweis: Aus (1.11)-(1.12) bzw. (1.9)-(1.10) haben wir unter Beachtung der oben eingeführten Bezeichnungen

$$\begin{aligned} (AR_hu - R_hLu)(x_i) &= \left[-\frac{u(x_i-h) - 2u(x_i) + u(x_i+h)}{h^2} + b_i\frac{u(x_i+h) - u(x_i-h)}{2h} + c_iu(x_i)\right] \\ &\quad - \left[-u''(x_i) + b_iu'(x_i) + c_iu(x_i)\right]. \end{aligned}$$

Lemma 1.15 ergibt daraus

$$|(AR_h u - R_h L u)(x_i)| \leq \frac{1}{12} h^2 \|u^{(4)}\|_{C[0,1]} + \frac{1}{6} h^2 \|b\|_{C[0,1]} \|u^{(3)}\|_{C[0,1]}.$$

Maximumbildung über alle Gitterpunkte x_i liefert die Behauptung. □

Bemerkung 1.17. Die Voraussetzung an die Lösung u des RWP ist in der Regel nicht realistisch. Eine sorgfältige Abschätzung zeigt

$$\|AR_h u - R_h L u\|_{\infty,\omega_h} \leq C \begin{cases} h^s, & u \in C^{2;s}[0,1] \\ h^{1+s}, & u \in C^{3;s}[0,1] \end{cases}$$

mit $0 \leq s \leq 1$ und den *Hölder-Räumen*

$$C^{k;s}[0,1] := \left\{ v \in C^k[0,1] : \sup_{\substack{x,y \in 0,1) \\ x \neq y}} \frac{|v^{(k)}(x) - v^{(k)}(y)|}{\|x-y\|^s} < \infty \right\}.$$

Durch diese Aussage kann man in gewisser Weise die Lücke zwischen den nach Lemma 1.16 geforderten Lösungen aus $C^4[0,1]$ und Lösungen u, die lediglich in $\in C^2(0,1) \cap C[0,1]$ liegen, schließen. □

(ii) *Stabilität der klassischen FDM*

Für die Stabilitätsdefinition in Definition 1.13 ist hinreichend, daß

$$\|A^{-1}\|_\infty \leq C_S \quad \text{mit } \|B\|_\infty := \max_{i=1,\ldots,n} \sum_{j=1}^{n} |b_{ij}|.$$

Bei den weiteren Untersuchungen nutzen wir die Halbordnungsrelation $x \geq 0$ für Vektoren x, falls komponentenweise gilt $x_i \geq 0$. Entsprechend gilt $x \geq y$, falls $x - y \geq 0$. Ferner schreiben wir für Matrizen $A \geq 0$, falls komponentenweise gilt $a_{ij} \geq 0$.

Definition 1.18. *Eine Matrix $A = (a_{ij}) \in \mathbf{R}^{n \times n}$ heißt*

(i) L_0-Matrix, *falls* $a_{ij} \leq 0,\ i,j = 1,\ldots,n,\ i \neq j$,
(ii) L-Matrix, *falls A L_0-Matrix ist mit* $a_{ii} > 0,\ i = 1,\ldots,n$,
(iii) M-Matrix, *falls A L_0-Matrix ist und die inverse Matrix A^{-1} mit $A^{-1} \geq 0$ existiert*,
(iv) inversmonoton, *falls aus der Halbordnungsrelation $Ax \leq Ay$ auch $x \leq y$ folgt.*

Die Matrix A für die klassische FDM aus Abschnitt 1.3 ist unter den Voraussetzungen (1.14) von Satz 1.8 eine L-Matrix, sie ist sogar schwach diagonaldominant und irreduzibel. Insbesondere ist die Bedingung an die Gitter-Peclet-Zahl $P_i := \frac{h}{2}|b_i| \leq 1$ wesentlich.

Wir wollen weiter untersuchen, wann A sogar M-Matrix bzw. inversmonoton ist. Zur Inversmonotonie von A ist die Existenz von A^{-1} mit $A^{-1} \geq 0$ äquivalent.

Lemma 1.19. *Unter den Voraussetzungen von Satz 1.8 ist $A^{-1} \geq 0$, d.h. A ist inversmonoton.*

Beweis: Wir betrachten die iterative Lösung des Gleichungssystems $Az = r$ mit dem Gesamtschritt-Verfahren. Aus der Zerlegung $A = D + A_L + A_R$ mit der Diagonalmatrix D und den strikten unteren bzw. oberen Dreiecksmatrizen A_L und A_R ergibt sich die Iterationsvorschrift

$$z_{m+1} = -D^{-1}(A_L + A_R) z_m + D^{-1} r, \quad m = 0, 1, \ldots. \tag{1.23}$$

Für die schwach diagonaldominante und irreduzible. Matrix A konvergiert das Verfahren (1.23).

Für die Spalten der inversen Matrix $A^{-1} = (a_1, \ldots, a_n)$ gilt $Aa_i = e_i,\ i = 1,\ldots,n$ mit den kartesischen Einheitsvektoren e_i. Damit entsteht a_i als Grenzelement von (1.23) mit $r = e_i$ und Startvektor $z_0 = 0$.

Nach den Voraussetzungen von Satz 1.8 sind die Elemente von D^{-1} und $-D^{-1}(A_L + A_R)$ nichtnegativ.

Daraus folgt die Aussage $A^{-1} \geq 0$. □

Nun ist die Stabilitätskonstante C_S abzuschätzen. Wir nutzen dazu

Lemma 1.20. *(M-Kriterium)*
Sei A L_0-Matrix. Dann ist A inversmonoton genau dann, wenn ein Vektor $e > 0$ existiert mit $Ae > 0$. Ferner gilt dann die Abschätzung

$$\|A^{-1}\|_\infty \leq \frac{\|e\|}{\min_k (Ae)_k}.$$

Beweis: (i) Sei A inversmonoton. Dann wähle man $e = A^{-1}(1, 1, \cdots, 1)^*$.
(ii) *Übungsaufgabe !* □

Die gesuchte Abschätzung der Stabilitätskonstanten C_S gelingt nun bei geeigneter Wahl eines majorisierenden Vektors e zur Matrix A gemäß Lemma 1.20.

Lemma 1.21. *Seien die Voraussetzungen von Satz 1.8 an die Matrix A erfüllt.*
(i) *Unter der Voraussetzung $c(x) \geq c^* > 0$ gilt*

$$\|A^{-1}\|_\infty \leq \frac{1}{\min_k \left(a_{kk} - \sum_{j \neq k} |a_{jk}| \right)}.$$

(ii) Bei $c(x) \geq 0$ existiert eine Konstante $C_S > 0$ (vgl. Beweis) mit

$$\|A^{-1}\|_\infty \leq C_S.$$

Beweis: (i) Bei $c(x) \geq c^* > 0$ ist A streng diagonaldominant. Die Behauptung folgt aus Lemma 1.20 mit $e = (1, 1, ..., 1)^*$.
(ii) Sei $e(x)$ Lösung des RWP

$$-w''(x) + b(x)w'(x) = 1, \;\; 0 < x < 1; \quad w(0) = w(1) = 0.$$

Aus dem Maximumprinzip (vgl. Lemma 1.6) - in einer verschärften Version mit strikten Ungleichungen - folgt $e(x) > 0, \;\; 0 < x < 1$. Ferner ist nach Konstruktion $(Le)(x) \geq 1, \;\; 0 < x < 1$. Nun wählen wir den Vektor

$$e := (e(x_1), ..., e(x_n))^*.$$

Aus Konsistenzgründen ist $Ae \geq \frac{1}{2}$ für $h \leq \tilde{h}_0$, denn in der Darstellung

$$Ae = AR_h e = (AR_h - R_h L)e + R_h Le$$

konvergiert der erste Term der rechten Seite nach Lemma 1.16 gegen 0. Für den zweiten Term ist $R_h Le \geq 1$. Die Behauptung folgt über Lemma 1.20. □

Die Abbildung 1.1 zeigt die diskrete Lösung des RWP

$$-u''(x) + 100u'(x) = 100, \quad 0 < x < 1; \qquad u(0) = u(1) = 0$$

mit der klassischen FDM auf einem äquidistanten Gitter mit $h = 0.2,\ 0.1,\ 0.01$ und $h = 0.001$ bei linearer Interpolation. Man erkennt Oszillationen der diskreten Lösungen für die groben Gitterweiten $h = 0.2$ und $h = 0.1$, offenbar ist das Maximumprinzip im diskreten Fall nicht erfüllt. Für die feineren h-Werte wird die exakte Lösung gut approximiert. Im Fall der Oszillationen genügt die Gitter-Peclet-Zahl nicht der Bedingung $P_i \leq 1$, insofern ist die Bedingung scharf (vgl. auch Übungsaufgabe, Serie 2).

(iii) *Konvergenz der klassischen FDM*

Wir kombinieren die Ergebnisse des Abschnitts zum

Satz 1.22. *Unter den Voraussetzungen von Satz 1.8 liege die Lösung u des RWP (1.9)-(1.10) in $C^4[0,1]$. Ferner sei ggf. h hinreichend klein. Dann gilt für den Diskretisierungsfehler der FDM (1.11)-(1.12)*

$$\|R_h u - U\|_{\infty,\omega_h} = \max_i |u(x_i) - U_i| \leq Mh^2,$$

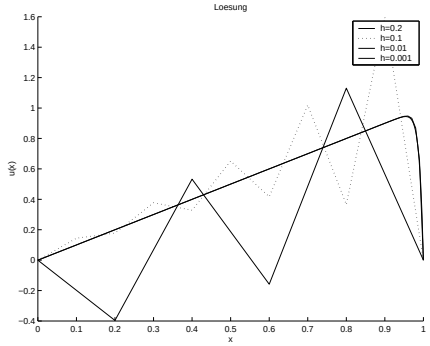


Abbildung 1.1: Lösung von $-u''(x) + 100u'(x) = 100$ für $h = 0.2,\ 0.1,\ 0.01$ und $h = 0.001$

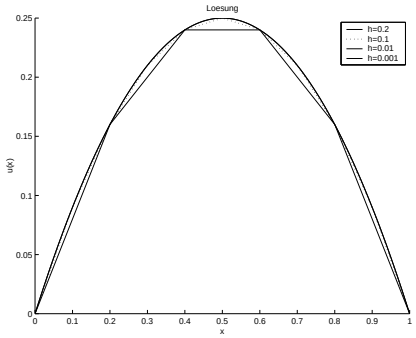


Abbildung 1.2: FDM-Lösung von $-u''(x) + \sin(\pi x)u(x) = 2 + \sin(\pi x)x(1-x)$ bei $h = \frac{1}{2}, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$

d.h. das Verfahren hat die Konvergenzordnung 2.

Beweis: Nach Konstruktion ist $R_h u - U = 0$ auf γ_h. Nach Lemma 1.16 ist ferner

$$\|AR_h u - R_h Lu\|_{\infty,\omega_h} \leq C_K h^2.$$

Mittels Lemma 1.21 folgt

$$C_S^{-1}\|R_h u - U\|_{\infty,\omega_h} \leq \|AR_h u - R_h Lu\|_{\infty,\omega_h} \leq C_K h^2$$

und damit die Konvergenzaussage mit $M = C_S C_K$. □

Die Abbildung zeigt die diskrete Lösung des RWP

$$-u''(x) + \sin(\pi x)u(x) = 2 + \sin(\pi x)x(1-x), \quad 0 < x < 1; \qquad u(0) = u(1) = 0$$

mittels klassischer FDM auf einem äquidistanten Gitter mit den Schrittweiten $h = 0.2,\ 0.1,\ 0.01$ und $h = 0.0001$. Die Knotenwerte wurden linear interpoliert. Man erkennt die Konvergenz der diskreten Lösung für $h \to 0$.

1.5 Vorgriff auf Finite-Elemente-Verfahren

Die Konvergenzanalyse der klassischen FDM erforderte sehr hohe Glattheitsforderungen an die Lösung des Zweipunkt-RWP. Abhilfe schafft hier die Abschwächung des bisher verwendeten "klassischen" Lösungsbegriff. Dies erlaubt zugleich einen natürlichen Zugang zur Finite-Elemente Methode (FEM), die wir hier

im Vorgriff auf spätere genauere Untersuchungen skizzieren. Dabei beschränken wir uns lediglich auf eine formale Einführung, da genauere Betrachtungen im weiteren Verlauf dieser Vorlesung folgen.

Vereinfachend betrachten wir speziell das 1. Randwertproblem der eindimensionalen Poisson-Gleichung:

$$-u''(x) = f(x), \quad 0 < x < 1 \tag{1.24}$$

$$u(0) = u(1) = 0. \tag{1.25}$$

Wir beginnen mit einer ersten *Abschwächung* des klassischen Lösungsbegriffs, d.h. von $u \in C^2(0,1) \cap C[0,1]$. Multiplikation von Gleichung (1.24) mit einer beliebigen *Testfunktion*

$$v \in \tilde{X} := \{w \in C^1(0,1) \cap C[0,1] : \ \ w(0) = w(1) = 0\} \tag{1.26}$$

und Integration über $(0,1)$ ergibt

$$\int_0^1 (-u'' \ v) \ dx = \int_0^1 f v \ dx.$$

Partielle Integration des Terms $-\int_0^1 u'' v \ dx$ liefert unter Beachtung der Randwerte $v(0) = v(1) = 0$

$$\int_0^1 u'v' \ dx = \int_0^1 f v \ dx \quad \forall v \in \tilde{X}. \tag{1.27}$$

Klassische Lösungen $u \in C^2(0,1) \cap C[0,1]$ von (1.24),(1.25) lösen offenbar auch (1.27). Ebenso sind (bei hinreichend glatten Daten) nach Rückwärtsausführung der vorgenommenen Umformungen klassische Lösungen von (1.27) auch Lösungen von (1.24),(1.25). Offenbar reicht aber z.B. schon die Forderung $u \in \tilde{X}$ für die Lösungen von (1.27) aus. In diesem Sinne kann man die Aufgabe

$$\text{Finde } \ u \in \tilde{X}, \ \text{ so daß } \ a(u,v) = f(v), \quad \forall v \in \tilde{X} \tag{1.28}$$

mit

$$a(u,v) \ := \ \int_0^1 u'v' \ dx; \qquad f(v) \ := \ \int_0^1 f v \ dx \tag{1.29}$$

als verallgemeinerte Aufgabenstellung zu (1.24),(1.25) bezeichnen. Man spricht auch von einer Variationsgleichung. Mit dem Ziel einer weiteren Abschwächung des Lösungsbegriffs skizzieren wir jetzt Eigenschaften des Raumes $\tilde{X}$ (vgl. (1.26)) in Verbindung mit der Norm

$$\|u\|_{H^1} := \left\{ \int_0^1 [u'(x)]^2 \ dx + \int_0^1 [u(x)]^2 \ dx \right\}^{1/2}. \tag{1.30}$$

Der Raum $\{\tilde{X}; \|\cdot\|_{H^1}\}$ ist offenbar normierter Raum, jedoch nicht vollständiger Raum, d.h. kein Banach-Raum. Die Norm (13.10) ist jedoch noch für meßbare Funktionen u, u' sinnvoll, die quadratisch über $(0,1)$ im Lebesgue-Sinne integrierbar sind, d.h. für Funktionen im *Lebesgue-Raum*

$$L^2(0,1) := \{v : (0,1) \to \mathbf{R} \ \text{ meßbar} : \ \int_0^1 [v(x)]^2 \, dx < \infty\}. \tag{1.31}$$

Wir kommen nun zur avisierten weiteren *Abschwächung* des klassischen Lösungsbegriffs, die wesentlich auf einer Verallgemeinerung der klassischen Regel der partiellen Integration basiert: Dazu benötigen wir einige Begriffe. Es bezeichnet $cl_V(A)$ den Abschluß der Teilmenge A von V in der Topologie des Raumes V. Dann heißt

$$\text{supp } \ v := \text{cl}_{\mathbf{R}}\{x \in (0,1) : \ \ v(x) \neq 0\}$$

Träger von $v \in C[0,1]$. Sei

$$C_0^\infty(0,1) := \{v \in C^\infty(0,1) : \ \ \text{supp } \ v \subset (0,1)\},$$

d.h. Elemente dieser Menge verschwinden von beliebiger Ordnung bei $x = 0$ und $x = 1$.

Die Regel der partiellen Integration ergibt für $u \in C^1[0,1]$ und beliebige *Testfunktionen* $v \in C_0^\infty(0,1)$

$$\int_0^1 u'v \ dx = -\int_0^1 uv' \ dx. \tag{1.32}$$

Man kann zeigen, daß die Integrale in (1.32) noch Sinn ergeben für $u, u' \in L^2(0,1)$.

Definition 1.23 $w \in L^2(0,1)$ *heißt* verallgemeinerte erste Ableitung von $u \in L^2(0,1)$, *falls*

$$\int_0^1 wv \ dx = -\int_0^1 uv' \ dx, \quad \forall v \in C_0^\infty(0,1)$$

gilt. Man schreibt $w = u'$.

Ein einfaches (und für die Einführung finiter Elemente relevantes) Beispiel ist die Funktion

$$f : [-1,1] \to \mathbf{R}, \qquad f(x) = |x|.$$

Sie hat in $x = 0$ keine klassische Ableitung. Man rechnet jedoch leicht nach, daß die stückweise definierte klassische Ableitung bei Auslassung der Stelle $x = 0$ eine verallgemeinerte Ableitung in $L^2(0,1)$ ist (vgl. Übungsaufgabe.)

Wir führen nun formal Räume mit verallgemeinerten Ableitungen ein.

Definition 1.24. *Die Menge*

$$H^1(0,1) := \{v \in L^2(0,1) : \ \exists v' \in L^2(0,1)\}$$

heißt Sobolev-Raum *der Funktionen mit verallgemeinerten und quadratisch auf* $(0,1)$ *integrierbaren Ableitungen. Ferner ist*

$$H_0^1(0,1) := \mathrm{cl}_{H^1(0,1)} C_0^\infty(0,1).$$

Wir zeigen später (wesentlich allgemeiner), daß die Räume $\{H^1(0,1); \|\cdot\|_{H^1(0,1)}\}$ und $\{H_0^1(0,1); \|\cdot\|_{H^1(0,1)}\}$ Hilbert-Räume mit dem *Skalarprodukt*

$$(u,v)_{H^1} := \int_0^1 uv \ dx + \int_0^1 u'v' \ dx.$$

sind. Ferner gilt, daß durch die Halbnorm

$$\|v\|_X := (a(v,v))^{\frac{1}{2}} = \left(\int_0^1 u'(x)v'(x) \ dx\right)^{\frac{1}{2}}$$

sogar eine Norm auf dem Raum $X = H_0^1(0,1)$ erklärt wird. Hierbei sind die (verallgemeinerten) homogenen Randbedingungen wesentlich. Wir werden sehen, daß der Raum $\{X; \|\cdot\|_X\}$ der geeignete Funktionenraum ist, um eine verallgemeinerte Aufgabenstellung von (1.24)-(1.25) zu formulieren:

$$\text{Finde} \quad u \in X := H_0^1(0,1) : \ \ a(u,v) = f(v) \quad \forall v \in X. \tag{1.33}$$

Wir konstruieren nun geeignete endlich-dimensionale Unterräume $X_n \subset X$ zur Diskretisierung der verallgemeinerten Aufgabenstellung (1.33). Unter Zerlegung des Intervalls

$$[0,1] = \cup_{i=1}^{n+1} M_i, \quad K_i := [x_{i-1}, x_i]$$

mit der Gitterweite $h_i := x_i - x_{i-1}$ betrachten wir den endlich-dimensionalen Raum

$$X_n := \{v \in C[0,1] : \ \ v(0) = v(1) = 0, \ \ v|_{K_i} \in P_1(K_i), i = 1, ..., n+1\}. \tag{1.34}$$

Mittels stückweise linearer Lagrangescher Basisfunktionen (finite Elemente)

$$\phi_i(x) := \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}}, & x \in K_i \\ \frac{x_{i+1}-x}{x_{i+1}-x_i}, & x \in K_{i+1} \\ 0, & \text{sonst} \end{cases}$$

erhalten wir

$$X_n = \text{span}\{\phi_1(x), ..., \phi_n(x)\}. \tag{1.35}$$

Offenbar sind die Basisfuntionen ϕ_i an den inneren Knotenpunkten nicht klassisch differenzierbar. Man prüft jedoch leicht nach, daß sie im Sinne der Definition 1.23 verallgemeinert differenzierbar sind und die Ableitungen in $L^2(0,1)$ liegen. Man beachte, daß die Funktionen aus X_n per Konstruktion die homogene Dirichlet-Randbedingung erfüllen. Es gilt also die Inklusion $X_n \subset X$.

Lemma 1.25. *Jede Funktion $v^n \in X_n$ ist durch die Knotenwerte $v_i = v(x_i)$ eindeutig festgelegt und besitzt die Darstellung*

$$v^n(x) = \sum_{j=1}^{n} v_j \phi_j(x).$$

Die Finite-Elemente-Methode erhält man nun durch Restriktion der verallgemeinerten Aufgabenstellung (oder *Variationsgleichung*) (1.33) auf einen Unterraum $X_n \subset X$, d.h.

$$\text{Finde} \quad u^n = \sum_{j=1}^{n} u_j \phi_j \in X_n : \quad a(u^n, v^n) = F(v^n) := \int_0^1 f v^n \, dx \quad \forall v^n \in X_n. \tag{1.36}$$

Diese Aufgabe ist äquivalent zu einem linearen Gleichungssystem. Dazu kann man die Aufgabe (1.36) ersetzen durch ein System von Gleichungen, indem man nacheinander für v^n die Basisfunktionen ϕ_j, $j = 1, ...n$ einsetzt. Mit den Bezeichnungen

$$A := (A_{ij}) \in \mathbf{R}^{n\times n}, \qquad A_{ij} := a(\phi_j, \phi_i) = \int_0^1 \phi_j' \phi_i' \, dx$$

sowie mit dem Knotenvektor $U = (u_1, \ldots, u_n)^* \in \mathbf{R}^n$ und dem Vektor $F = (F(\phi_1), \ldots, F(\phi_n))^* \in \mathbf{R}^n$ erhält man das System

$$\text{Finde } U \in \mathbf{R}^n : \qquad AU = F.$$

Wir kommen nun zur Generierung der Matrix A. Wegen $\text{supp}(\phi_i) = (x_{i-1}, x_{i+1})$ ist

$$A_{ij} = \int_0^1 \phi_j' \phi_i' \, dx = 0, \quad |i-j| \geq 2,$$

d.h. A ist Tridiagonalmatrix. Für die Nichtnullelemente der Matrix A erhalten wir nach kurzer Rechnung

$$A_{i,i-1} = \frac{-1}{x_i - x_{i-1}}, \quad A_{i,i} = \frac{1}{x_i - x_{i-1}} + \frac{1}{x_{i+1} - x_i}, \quad A_{i,i+1} = \frac{-1}{x_{i+1} - x_i},$$

d.h.

$$A = \text{tridiag} \left\{ -\frac{1}{h_i}; \ \frac{1}{h_i} + \frac{1}{h_{i+1}}; \ -\frac{1}{h_{i+1}} \right\}. \tag{1.37}$$

Für die rechte Seite des Gleichungssystems folgt

$$F(\phi_i) = \int_0^1 f\phi_i \, dx = \int_{x_{i-1}}^{x_i} f\phi_i \, dx + \int_{x_i}^{x_{i+1}} f\phi_i \, dx. \tag{1.38}$$

Die Koeffizienten A_{ij} sind nur im Spezialfall wie hier exakt integrierbar. Im allgemeinen Fall interpoliert man die Daten durch Splines oder integriert mit passenden Quadraturformeln.

Die bei der klassischen Finite-Differenzen Methode entstehende Matrix A für Problem (1.28) stimmt mit der bei Nutzung stückweise linearer finiter Elemente entstehenden Matrix A im äquidistanten Fall $h = h_i, i = 1, ..., n+1$ bis auf den Skalierungsfaktor $\frac{1}{h}$ überein. Unterschiede entstehen jedoch ggf. bei der rechten Seite. Zur Lösung des linearen Gleichungssystems für die FEM können damit der Thomas-Algorithmus oder Standard-Iterationsverfahren herangezogen werden.

Es verbleibt die Ableitung einer Fehlerabschätzung, die sich von der Vorgehensweise bei der FDM erheblich unterscheidet. Dazu benötigen wir folgende Aussage, die wir später als Lemma von Cea bezeichnen.

Lemma 1.26. *Sei $u \in X$ Lösung von (1.33). Dann gilt für den Diskretisierungsfehler $u - u^n$ der Finite-Elemente-Methode (1.36) folgende quasi-optimale Abschätzung*

$$\|u - u^n\|_X \leq \inf_{v^n \in X_n} \|u - v^n\|_X. \tag{1.39}$$

Beweis: Durch Differenzbildung der Gleichungen (1.33) und (1.36) hat man die Fehlergleichung

$$a(u - u^n, \tilde{w}^n) = 0, \qquad \forall \tilde{w}^n \in X_n.$$

Andererseits gilt dann unter Beachtung der Normdefinition und der Fehlergleichung

$$\|u - u^n\|_X^2 = a(u - u^n, u - u^n) = a(u - u^n, u - w^n), \qquad \forall w^n \in X_n.$$

Man kann die rechte Seite dieser Ungleichungskette nach oben abschätzen durch

$$a(u - u^n, u - w^n) \leq \|u - u^n\|_X \, \|u - w^n\|_X,$$

damit folgt nach Kürzen von $\|u - u^n\|_X$ und Infimumbildung die Behauptung

$$\|u - u^n\|_X \leq \inf_{v^n \in X_n} \|u - v^n\|_X. \qquad \square$$

Die Fehlerabschätzung ist somit auf eine Interpolationsabschätzung im Unterrraum X_n an die (gesuchte) Lösung $u \in X$ zurückgeführt.

Satz 1.27. *Unter der Voraussetzung $u'' \in L^2(0,1)$ an die (verallgemeinerte) Lösung $u \in X$ des Problems (1.33) gilt für den Diskretisierungsfehler des Finite-Elemente-Verfahrens (1.36) die Abschätzung*

$$\|(u - u^n)'\|_{L^2(0,1)} \leq \frac{1}{2\sqrt{2}} h \|u''\|_{L^2(0,1)}. \tag{1.40}$$

Beweis: In einer Übungsaufgabe (vgl. Serie 2) wird für den Interpolationsfehler $u - u_I$ mit der Lagrange-Interpolierten $u_I \in X_n$ an $u \in X$ mit der zusätzlichen Glättevoraussetzung $u'' \in L^2(0,1)$ gezeigt, daß

$$\|u' - u_I'\|_{L^2(0,1)}^2 \leq \|u''\|_{L^2(0,1)} \|u - u_I\|_{L^2(0,1)}$$

sowie

$$\|u - u_I\|_{L^2(0,1)} \leq \frac{1}{2\sqrt{2}} h \|u' - u_I'\|_{L^2(0,1)},$$

also

$$\|u' - u_I'\|_{L^2(0,1)} \leq \frac{1}{2\sqrt{2}} h \|u''\|_{L^2(0,1)}$$

gilt. In Verbindung mit der Wahl $v^n := u_I$ in Lemma 1.26 erhält man die Aussage (1.40). $\square$

Bemerkung 1.28. (i) Man kann unter der schwächeren Voraussetzung $u' \in L^2(0,1)$ sogar zeigen:

$$\|u - u^n\|_{L^2(0,1)} \leq \frac{1}{2\sqrt{2}} h \|u'\|_{L^2(0,1)}. \tag{1.41}$$

(ii) Die Fehlerabschätzungen (1.40), (1.41) sind optimal und können nicht verbessert werden. Zur Gewinnung von (1.40) muß man zusätzlich die Existenz der verallgemeinerten zweiten Ableitung $u'' \in L^2(0,1)$ fordern. Man vergleiche jedoch die hier verwendeten Regularitätsannahmen an die Lösung des RWP mit denen für die Konvergenzanalyse bei der klassischen Finite-Differenzen-Methode in Abschnitt 1.4. $\square$

Kapitel 2

Klassifizierung partieller Differentialgleichungen

Nach der Behandlung von Zweipunkt-Randwertaufgaben in Kapitel 1 gehen wir nun zum allgemeineren Fall *partieller Differentialgleichungen* über. Gegenstand des vorliegenden Kapitels ist eine *Klassifizierung* partieller Differentialgleichungen 2. Ordnung. Der Typ einer vorliegenden Differentialgleichung bestimmt wesentlich die analytische Eigenschaften der Lösungen. Diese Eigenschaften charakterisieren Grundzüge der durch die Gleichung beschriebenen Modelle aus Naturwissenschaften und Technik. Der Typ der Differentialgleichung beeinflußt auch erheblich die Auswahl numerischer Lösungsverfahren.

Für die vorliegende Vorlesung ist es zunächst ausreichend, sich im vorliegenden Kapitel auf die Ausführungen zu Gleichungen 2. Ordnung zu konzentrieren. Eine wesentlich ausführlichere Darstellung zum allgemeinen Fall findet man zum Beispiel in der Monographie [6].

2.1 Grundbegriffe. Bezeichnungen

Seien $x = (x_1, ..., x_n)$ ein beliebiger Punkt im $\mathbf{R}^n$ und Ω ein beschränktes Gebiet im $\mathbf{R}^n$, d.h. eine offene und zusammenhängende Punktmenge. $\overline{\Omega}$ ist die abgeschlossene Hülle von Ω. Mit $\partial\Omega := \overline{\Omega} \setminus \Omega$ bezeichnen wir den *Rand des Gebietes.*

$C(\Omega)$ bzw. $C(\overline{\Omega})$ bezeichnen die Räume der auf Ω bzw. bis auf den Rand $\partial\Omega$ stetigen Funktionen. Sei nun m eine nichtnegative ganze Zahl. Einen Vektor $\alpha := (\alpha_1, ..., \alpha_n)$ mit nichtnegativen ganzen Zahlen α_i nennen wir *Multiindex* der Länge $|\alpha| := \sum_{i=1}^{n} \alpha_i$. Zur Abkürzung schreiben wir *partielle Ableitungen der Ordnung* α einer hinreichend oft differenzierbaren Funktion $u : \Omega \to \mathbf{R}$ im Punkt $x \in \Omega$ in folgender Form:

$$D^\alpha u(x) := \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} ... \partial x_n^{\alpha_n}}(x), \;\; |\alpha| \geq 1 : \qquad D^{(0,...,0)} u(x) := u(x).$$

Die Menge der $m-$fach auf Ω stetig differenzierbaren Funktionen ist

$$C^m(\Omega) := \{v : \Omega \to \mathbf{R} |\; D^\alpha u \in C(\Omega), \;\; \forall \alpha : |\alpha| \leq m\}.$$

$C^m(\overline{\Omega})$ ist die Menge aller Funktionen aus $C^m(\Omega)$ mit stetig auf $\overline{\Omega}$ fortsetzbaren Ableitungen bis zur Ordnung m.

Die nachfolgende Klassifizierung partieller Differentialgleichungen verallgemeinert die in Kapitel 1 vorgenommene Einteilung für den eindimensionalen Fall, d.h. $n = 1$. Im *allgemeinen* Fall einer *nichtlinearen partiellen Differentialgleichung* sucht man eine Funktion $u = u(x)$ als Lösung von

$$F(x, D^{\beta_1} u(x), ..., D^{\beta_k} u(x)) \;=\; 0, \quad x \in \Omega \tag{2.1}$$

bei gegebener Funktion F und Multiindizes $\beta_i, i = 1, ..., k$. Als *Ordnung* der partiellen Differentialgleichung (2.1) bezeichnet man die Zahl $m := \max_{i=1,...,k} |\beta_i|$.

Definition 2.1 *Die partielle Differentialgleichung (2.1) heißt*

(i) quasilinear, *falls die Funktion F linear in allen Ableitungen der höchsten Ordnung ist, d.h. mit partiellen Differentialausdrücken $\tilde{F}$ bzw. A_{β_i} der Ordnung $m-1$ gilt,*

$$\begin{aligned} F(x, D^{\beta_1}u(x), ..., D^{\beta_k}u(x)) &= \sum_{|\beta_i|=m} A_{\beta_i}(x, D^{\gamma_1}u(x), ..., D^{\gamma_l}u(x))D^{\beta_i}u(x) \\ &\quad + \tilde{F}(x, D^{\gamma_1}u(x), ..., D^{\gamma_l}u(x)), \end{aligned} \tag{2.2}$$

(ii) semilinear, *falls mit einem Differentialausdruck $\tilde{F}$ der Ordnung $m-1$ gilt*

$$F(x, D^{\beta_1}u(x), ..., D^{\beta_k}u(x)) = \sum_{|\beta_i|=m} a_{\beta_i}(x)D^{\beta_i}u(x) + \tilde{F}(x, D^{\gamma_1}u(x), ..., D^{\gamma_l}u(x)), \tag{2.3}$$

(iii) linear, *falls F linear in allen Ableitungen $D^{\beta_i}u$ mit $|\beta_i| \leq m$ ist, d.h.*

$$F(x, D^{\beta_1}u(x), ..., D^{\beta_k}u(x)) = \sum_{|\beta_i|\leq m} a_{\beta_i}(x)D^{\beta_i}u(x). \tag{2.4}$$

Beispiele 2.2. Im Rahmen dieser Vorlesung beschränken wir uns auf Differentialgleichungen 2. Ordnung (d.h. $m = 2$). Die allgemeine Form einer *quasilinearen* partiellen Differentialgleichung 2. Ordnung ist unter Verwendung des Gradienten $\nabla u := (\frac{\partial u}{\partial x_i})_{i=1}^n$ gegeben durch

$$\sum_{i,j=1}^{n} A_{ij}(x, u(x), \nabla u(x))\frac{\partial^2 u(x)}{\partial x_i \partial x_j} + B(x, u(x), \nabla u(x)) = 0. \tag{2.5}$$

Eine *semilineare* partielle Differentialgleichung 2. Ordnung hat die allgemeine Form

$$\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 u(x)}{\partial x_i \partial x_j} + B(x, u(x), \nabla u(x)) = 0, \tag{2.6}$$

die *lineare* partielle Differentialgleichung 2. Ordnung lautet in allgemeiner Form

$$\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 u(x)}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x)\frac{\partial u(x)}{\partial x_i} + c(x)u(x) - f(x) = 0. \tag{2.7}$$

Der Term mit den höchsten Ableitungen, d.h. der Ordnung 2, heißt jeweils *Hauptteil* der partiellen Diferentialgleichung. □

2.2 Punktweise Klassifizierung

Ausgangspunkt ist die allgemeine Form einer semilinearen partiellen Differentialgleichung 2. Ordnung (2.6). Zur *punktweisen Klassifizierung* nutzen wir die Eigenwerte der Matrix $A(x) := (a_{ij}(x))_{i,j=1}^n$ in einem beliebigen (jedoch fixierten) Punkt $x_0 \in \Omega$.

Wir setzen dazu die Stetigkeit der Funktionen $a_{ij}(\cdot)$ auf Ω voraus, ferner sei $u \in C^2(\Omega)$. Nach dem Satz von Schwarz ist dann bei den zweiten partiellen Ableitungen die Reihenfolge der Differentiation unwesentlich. Daher kann man ohne Beschränkung der Allgemeinheit die Symmetrie der Matrix $A(\cdot)$ annehmen, d.h.

$$a_{ij}(x) = a_{ji}(x), \quad i, j = 1, ..., n; \quad \forall x \in \Omega.$$

Für einen beliebigen Punkt $x_0 \in \Omega$ bezeichnen nun $\lambda_i(x_0),\ i = 1, ..., n$ die *Eigenwerte* der Matrix $A(x_0)$.

Definition 2.3. *Ein semilinearer Differentialoperator 2. Ordnung der Form (2.6) heißt im Punkt x_0*

- elliptisch, *falls alle Eigenwerte $\lambda_i(x_0)$ nicht verschwinden und gleiches Vorzeichen haben,*

- hyperbolisch, *falls alle Eigenwerte nicht verschwinden und wenigstens zwei Eigenwerte verschiedenes Vorzeichen haben*,
- normal-hyperbolisch, *falls der Operator hyperbolisch ist und genau ein Eigenwert ein anderes Vorzeichen als alle anderen hat*,
- parabolisch, *falls mindestens ein Eigenwert verschwindet*,
- normal- parabolisch, *falls genau ein Eigenwert verschwindet und alle anderen gleiches Vorzeichen haben.*

Man kann sich überlegen, daß die Klassifizierung in den elliptischen, hyperbolischen und parabolischen Typ im räumlich zweidimensionalen Fall ($n = 2$) erschöpfend ist. (Dies sei zur Übung empfohlen.)

Beispiele 2.4. Die ersten drei Beispiele stellen jeweils die Hauptvertreter elliptischer, normal hyperbolischer bzw. normal-parabolischer Gleichungen dar.
(i) Für die *Poisson-Gleichung*

$$-(\Delta u)(x) := -\sum_{i=1}^{n} \frac{\partial^2 u(x)}{\partial x_i^2} = f(x)$$

ist $A(x) = \text{diag}(-1)$, d.h. alle Eigenwerte sind identisch -1. Damit ist die Gleichung für alle Punkte des Definitionsbereiches elliptisch.
(ii) Die *Wämeleitungsgleichung*

$$\frac{\partial u(x)}{\partial x_n} - a^2 \sum_{i=1}^{n-1} \frac{\partial^2 u(x)}{\partial x_i^2} = f(x)$$

besitzt im Fall $a^2 = \text{konst.} > 0$ die Matrix $A(x) = \text{diag}(-a^2, ..., -a^2, 0)$, d.h. genau ein Eigenwert verschwindet. Für alle Punkte des Definitionsbereiches ist die Gleichung normal-parabolisch. Wegen der Sonderrolle der Variablen x_n bezeichnet man diese gesondert mit t, dies deutet auf die Interpretation als *zeitliche Variable* hin.
(iii) Die *Wellengleichung*

$$\frac{\partial^2 u(x)}{\partial x_n^2} - a^2 \sum_{i=1}^{n-1} \frac{\partial^2 u(x)}{\partial x_i^2} = f(x)$$

besitzt im Fall $a^2 = \text{konst.} > 0$ die Matrix $A(x) = \text{diag}(-a^2, ..., -a^2, 1)$, d.h. kein Eigenwert verschwindet und genau ein Eigenwert hat ein anderes Vorzeichen als alle anderen Eigenwerte. Für alle Punkte des Definitionsbereiches ist die Gleichung somit normal-hyperbolisch. Wegen der Sonderrolle der Variablen x_n bezeichnet man diese erneut als *zeitliche Variable* mit t.
(iv) Die sogenannte *Tricomi–Gleichung*

$$x_2 \frac{\partial^2 u(x)}{\partial x_1^2} + \frac{\partial^2 u(x)}{\partial x_2^2} = 0$$

hat die nicht-konstante Matrix $A(x) := \text{diag}(x_2, 1)$. Sie ist somit für $x_2 > 0$ punktweise elliptisch, für $x_2 = 0$ punktweise parabolisch und für $x_2 < 0$ punktweise hyperbolisch. Die Gleichung modelliert sehr vereinfachend kompressible, schallnahe Strömungen. Die Linie $x_2 = 0$ mit Typwechsel der Gleichung entspricht gerade der sogenannten Schallinie. □

2.3 Kanonische Form

Definition 2.5. *Für die semilineare partielle Differentialgleichung (2.6) seien $\lambda_i(x_0)$, $i = 1, ..., n$ die Eigenwerte der Matrix $A(x_0) := (a_{ij}(x_0))_{i,j=1}^n$ des Hauptteiles. Dann ist*

$$\sum_{i=1}^{n} \lambda_i(x_0) \frac{\partial^2 u(x_0)}{\partial x_i^2} + \tilde{B}(x_0, u(x_0), \nabla u(x_0)) = 0 \tag{2.8}$$

mit einem geeignetem Differentialausdruck $\tilde{B}$ der Ordnung 1 *die* kanonische Form *der Gleichung.*

Lemma 2.6. *Im Fall einer global konstanten und symmetrischen Matrix $A(x) \equiv A$ mit den Eigenwerten λ_i mit $i = 1, ..., n$ läßt sich die semilineare partielle Gleichung (2.6) global im Definitionsbereich auf die kanonische Form (2.8) überführen.*

Beweis: Dies folgt aus den bekannten Aussagen zur Hauptachsentransformation quadratischer Formen (vgl. Grundkurs AGLA). □

Es entsteht natürlich die Frage, ob auch im Fall einer nichtkonstanten Matrix $A(x)$ eine derartige Transformation auf die kanonische Form möglich ist. Es zeigt sich, daß im räumlich zweidimensionalen Fall ($n = 2$) eine solche Transformation zumindest lokal möglich ist.

Vereinfachend führen wir für $n = 2$ eine Umbezeichnung durch mit $(x_1, x_2) \mapsto (x, y)$ sowie $a_{11} \mapsto a$, $a_{12} = a_{21} \mapsto b$ und $a_{22} \mapsto c$. Für die Klassifizierung der entstehenden Gleichung

$$a(x,y)\frac{\partial^2 u}{\partial x^2} + 2b(x,y)\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\frac{\partial^2 u}{\partial y^2} + B(x, y; u(x,y), \nabla u(x,y)) = 0 \tag{2.9}$$

ist das Vorzeichen der Determinante $\det A(x,y) = a(x,y)c(x,y) - [b(x,y)]^2$ ausschlaggebend.

Satz 2.7. *Die Funktionen a, b, c bzw. B der semilinearen partiellen Differentialgleichung (2.9) seien in Umgebung eines Punktes (x, y) des Definitionsgebietes hinreichend glatt. In dieser Umgebung sei die Gleichung auch entweder vom elliptischen, parabolischen bzw. hyperbolischen Typ. Dann gibt es jeweils eine nichtsinguläre und hinreichend glatte Transformation $(x, y) \mapsto (\xi, \eta)$ derart, daß die Gleichung (2.9) lokal auf die kanonische Form überführt werden kann. Genauer gilt*

- im elliptischen Fall *mit* $ac - b^2 > 0$:

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} + \tilde{B}(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}) = 0,$$

- im parabolischen Fall *mit* $ac - b^2 = 0$*:*

$$\frac{\partial^2 u}{\partial \eta^2} + \tilde{B}(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}) = 0,$$

- im hyperbolischen Fall *mit* $ac - b^2 < 0$:

$$\frac{\partial^2 u}{\partial \xi^2} - \frac{\partial^2 u}{\partial \eta^2} + \tilde{B}(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}) = 0.$$

Beweis: Die Ausführung der Koordinatentransformation $(x, y) \mapsto (\xi, \eta)$ bzw. die Wahl der Transformationsfunktionen $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ basieren auf der Lösung geeigneter partieller Differentialgleichungen erster Ordnung (Charakteristiken-Gleichungen). Man vergleiche hierzu zum Beispiel die Ausführungen bei [24], S. 50 ff. □

Kapitel 3

Poisson-Gleichung als Prototyp elliptischer Gleichungen

Im weiteren Verlauf der Vorlesung behandeln wir von nun an *elliptische Randwertaufgaben 2. Ordnung*. Im Kapitel 3 betrachten wir die einfachsten Aufgaben, die sogenannte *Poisson-Gleichung* sowie die *Laplace-* oder *Potential-Gleichung*. Zielstellung ist dabei einerseits, mit der Separationsmethode (vgl. Abschnitt 3.2) und der Finite-Differenzen-Methode (vgl. Abschnitt 3.3) einige elementare Lösungsmethoden zu studieren. Die Darstellung erfolgt jeweils für den sehr einfachen Spezialfall eines quadratischen Lösungsgebietes. Andererseits gewinnt man bereits einen Eindruck über die Lösungsstruktur dieser grundlegenden Gleichungen.

Auf Fragen der Potentialtheorie gehen wir im Rahmen dieser Vorlesung nicht ein. Hinsichtlich einer detaillierteren Darstellung verweise ich auf die schon in der Einleitung genannten Lehrbücher [14, 11, 24].

3.1 Poisson- und Potential-Gleichung

Sei $x = (x_i)_{i=1}^n$ ein beliebiger Punkt im Gebiet $\Omega \subseteq \mathbf{R}^n$. Ferner bezeichnet

$$\text{grad} := \nabla = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \vec{e}_i$$

mit den kartesischen Einheitsvektoren $\vec{e}_i$ den *Gradienten-Operator*. Für eine vektorwertige Funktion $\vec{v}(x) = \sum_{i=1}^n v_i(x)\vec{e}_i$ erklärt man die Divergenz durch

$$\text{div } \vec{v}(x) := \nabla \cdot \vec{v}(x) = \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i}.$$

Speziell benutzen wir den *Laplace-Operator*

$$\Delta := \text{div grad} = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}.$$

Prototyp linearer elliptischer Gleichungen 2. Ordnung ist die *Poisson-Gleichung*

$$-(\Delta u)(x) = f(x) \qquad x \in \Omega \tag{3.1}$$

bei gegebener Funktion $f : \Omega \mapsto \mathbf{R}$ und gesuchter Funktion $u : \Omega \mapsto \mathbf{R}$. Für $f \equiv 0$ erhält man die *Laplace-* bzw. *Potentialgleichung*.

Es ist recht instruktiv, zunächst spezielle Lösungen der Laplace-Gleichung

$$-\Delta v = -\sum_{i=1}^{n} \frac{\partial^2 v}{\partial x_i^2} = 0 \tag{3.2}$$

der Form $v = v(r)$ mit $r := \|x\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$ zu suchen. Unter Beachtung von

$$\frac{\partial r}{\partial x_i} = \frac{x_i}{r}, \qquad \frac{\partial^2 r}{\partial x_i^2} = \frac{1}{r} - \frac{x_i^2}{r^3}, \qquad i = 1, ..., n$$

transformiert man die Laplace-Gleichung unter Beachtung der Kettenregel auf die gewöhnliche Differentialgleichung

$$\Delta v = v''(r) + \frac{n-1}{r} v'(r) = 0.$$

Mittels Trennung der Veränderlichen ergeben sich die partikulären Lösungen

$$v(r) = C_2 \ln \frac{1}{r}, \quad n = 2; \qquad v(r) = C_n \frac{1}{r^{n-2}}, \quad n \geq 3. \tag{3.3}$$

Diese Funktionen heißen *singuläre* bzw. *Fundamentallösungen* der Laplace-Gleichung. Sie zeigen, daß sich in dieser Gleichung codierte Information radialsymmetrisch ausbreitet (vgl. auch hierzu Beispiel 3.1). Ferner spielen sie eine zentrale Rolle in der sogenannten Potentialtheorie.

Zur Festlegung der Lösung von (3.1) benötigt man Zusatzbedingungen. Für den Fall eines beschränkten Gebietes Ω stellt man meist *Randbedingungen* auf dem Gebietsrand $\partial\Omega := \overline{\Omega} \setminus \Omega$. Zur weiteren Beschreibung sei $\vec{\nu} := (\nu_i)_{i=1}^n$ der äußere Normaleneinheitsvektor auf $\partial\Omega$, der zunächst überall auf dem Rand existieren möge. Dann bezeichnet $\frac{\partial(\cdot)}{\partial \vec{\nu}} := \text{grad}(\cdot) \cdot \vec{\nu}$ die äußere Normalenableitung.

Typische lineare *Randbedingungen* haben die Gestalt

- 1. Art (Dirichlet-Typ): $u(x) = g(x), \ x \in \partial\Omega$
- 2. Art (Neumann-Typ): $\frac{\partial u(x)}{\partial \vec{\nu}} = g(x), \ x \in \partial\Omega$
- 3. Art (Robin-Typ): $\frac{\partial u(x)}{\partial \vec{\nu}} + \alpha(x) u(x) = g(x), \ x \in \partial\Omega$

bei jeweils gegebenen Funktionen $g(\cdot)$ und $\alpha(\cdot)$. Bei Vorgabe verschiedener dieser Randbedingungen auf zueinander paarweise disjunkten Teilmengen des Randes spricht man von *gemischten* Randbedingungen, die in praktischen Anwendungen besonders wichtig sind.

Wir beschreiben kurz einige *physikalische Modelle* für Gleichung (3.1).

Beispiel 3.1. (i) *Stationäre Wärmeleitung*:
Für die (gesuchte) Temperatur $u(t, x)$ lautet die Fourier'sche Gleichung

$$c\rho \frac{\partial u}{\partial t} = \text{div}(k \ \text{grad} u) + f$$

bei gegebener spezifischer Wärmekapazität c, der Dichte ρ, Wämeleitfähgkeit k und Wärmequelle f. Wird im stationären Fall $\frac{\partial u}{\partial t} \equiv 0$ zusätzlich ein homogener Körper Ω (d.h. $k = \text{const.}$) angenommen, so entsteht die stationäre Wärmeleitungsgleichung

$$-k \ \text{div} \ (\text{grad} u) \equiv -k\Delta u = f,$$

die die Wärmeausbreitung durch Leitung bzw. Diffusion bei vorhandener Quelle f beschreibt.

(ii) *Stationäre Diffusionsgleichung*:
Für die (gesuchte) Konzentration $c(t, x)$ lautet die Nernst'sche Gleichung

$$\frac{\partial c}{\partial t} = \text{div}(D \ \text{grad} c) + f$$

bei gegebenem Diffusionskoeffizienten D und dem Quellterm f (z.B. Schadstoffeintrag). Wird wieder im stationären Fall $\frac{\partial c}{\partial t} \equiv 0$ ein homogenes Medium Ω (d.h. $D = \text{const.}$) angenommen, so entsteht die stationäre Diffusionsgleichung

$$-D \text{ div } (\text{grad} c) \equiv -D\Delta c = f.$$

(iii) *Potentialströmungen:*
Sei $\vec{v}(x) = (v_i(x))_{i=1}^n$ die Geschwindigkeit der stationären Strömung einer inkompressiblen Flüssigkeit, d.h. die Dichte ρ ist konstant. Dann existiert im Spezialfall einer wirbelfreien Strömung, d.h. rot $\vec{v} := \nabla \times \vec{v} \equiv 0$ ein Potential ϕ mit $\vec{v} = -\text{grad } \phi$. Ist die Strömung zusätzlich auch noch quellenfrei, d.h. div $\vec{v} := \sum_{i=1}^n \frac{\partial v_i}{\partial x_i} \equiv 0$, so erhält man die *Potentialgleichung*

$$-\text{div } (\text{grad } \phi) \equiv -\Delta\phi = 0.$$

Eine analoges Modell erhält man für das Potential elektrostatischer Felder. □

In einer offenen Umgebung $\mathcal{U}$ des Randes $\partial\Omega$ eines Gebietes sei eine Funktion $F \in C^1(\mathcal{U})$ gegeben, so daß der Rand implizit durch die Gleichung $F(x) = 0$ beschrieben wird. Für Punkte $x \in \mathcal{U} \cap \Omega$ bzw. $x \in \mathcal{U} \cap [\mathbf{R}^n \setminus \Omega]$ gelte $F(x) > 0$ bzw. $F(x) < 0$. Man spricht auch von einem C^1-Rand. Später werden wir die Voraussetzungen an F noch abschwächen bis zum Begriff der Lipschitz-Stetigkeit, um auch praktisch relevante Gebiete wie Polyeder usw. zu erfassen (vgl. Abschnitt 2.4).

Gilt $|\text{grad } F(x)| \neq 0$ in allen Punkten des Randes, so bezeichnet

$$\vec{\nu}_\Omega(x) = (\nu_1(x), \ldots, \nu_n(x))^T := \frac{\text{grad } F(x)}{|\text{grad } F(x)|}, \qquad x \in \partial\Omega$$

den äußeren *Normaleneinheitsvektor*. Dann gilt mit dem $(n-1)-$dimensionalen Oberflächenmaß $s = \sigma_{n-1}$ (vgl. z.B. [1], Kap. 1) der folgende für viele weitere Betrachtungen grundlegende Satz.

Lemma 1.1. (Integralsatz von Gauß)
Für jede vektorwertige Funktion $\vec{w} = (w_1, ..., w_n)^T$ mit $w_i \in C^1(\Omega) \cap C(\overline{\Omega})$ gilt auf einem beschränkten Gebiet Ω mit C^1-Rand

$$\int_\Omega \sum_{i=1}^n \frac{\partial w_i}{\partial x_i} dx = \int_{\partial\Omega} \sum_{i=1}^n w_i \nu_i ds. \tag{3.4}$$

Mit den oben eingeführten Bezeichnungen können wir die Gleichung (3.4) auch schreiben in der Form

$$\int_\Omega \text{div } \vec{w}(x)\ dx = \int_{\partial\Omega} \vec{w}(x) \cdot \vec{\nu}_\Omega(x)\ ds. \tag{3.5}$$

Diese Gleichung stellt eine Integralbilanz zwischen der Divergenz (grob gesagt Quellen und Senken) des Feldes $\vec{w}$ im Gebiet und dem *Fluß* $\vec{w} \cdot \vec{\nu}_\Omega$ des Feldes $\vec{w}$ durch den Rand dar. Man spricht auch von einem *Erhaltungssatz*. Wichtig ist, daß das Lemma 1.1 auch noch für Lipschitz-stetig berandete Gebiete gültig bleibt. Im Spezialfall $\vec{w} = (uv)\vec{e}_i$ erhalten wir die wichtige *Regel der partiellen Integration*

$$\int_\Omega \frac{\partial u}{\partial x_i} v\ dx = -\int_\Omega u \frac{\partial v}{\partial x_i}\ dx + \int_{\partial\Omega} uv \cdot \nu_i\ ds.$$

3.2 Einführendes Beispiel zur Modellbildung

Zur Motivation der weiteren Ausführungen betrachten wir das *Modell der Wärmeleitung*:

Sei $T = T(t, x)$ die Temperatur in einem Punkt x des Festkörpers $\Omega \subset \mathbf{R}^n$ zum Zeitpunkt t. Mit der spezifischen Wärme c und der Dichte ρ des Körpers berechnet sich die Gesamtenergie aus $e = c\rho T$. Weiter sei $G \subseteq \Omega$ ein beliebiges (!) Teilgebiet mit C^1-Rand ∂G und dem äußeren Normaleneinheitsvektor $\vec{\nu}_G$.

Ausgangspunkt der weiteren Betrachtung ist das physikalische *Axiom der Energieerhaltung*: Die zeitliche Änderung der Gesamtenergie (in G) ist gleich der Summe der Energie der einwirkenden inneren und

äußeren Kräfte und der Änderung der inneren Energie. Für die zeitliche Änderung der Gesamtenergie gilt

$$Q = \frac{d}{dt}\int_G e\ dx = \int_G \frac{\partial e}{\partial t}(t,x)\ dx.$$

(Innere) Wärmequellen bzw. -senken W im Gebiet G erzeugen die Wärmemenge $Q_1 = \int_G \rho W\ dx$. Äußere Kräfte wollen wir hier noch vernachlässigen.

Der durch *Leitung* bzw. *Diffusion* verursachte Vektor des Wärmestroms ist

$$\vec{q} = -\lambda \nabla T := -\lambda \left(\frac{\partial T}{\partial x_i}\right)_{i=1}^n$$

mit der Wärmeleitzahl $\lambda > 0$. Der Änderung der inneren Energie entspricht der Gesamtfluß des Wärmestroms durch die Oberfläche ∂G von G gemäß

$$Q_2 = -\int_{\partial G} (\vec{q}\cdot\vec{\nu}_G)\ ds = \int_{\partial G} \lambda(\nabla T\cdot\vec{\nu}_G)\ ds = \int_{\partial G} \lambda \frac{\partial T}{\partial \vec{\nu}_G}\ ds = \int_G \nabla\cdot(\lambda\nabla T)\ dx.$$

Im letzten Schritt wurde Lemma 1.1 mit $\vec{w} := \lambda\nabla T$ benutzt.

Nach dem Axiom der Energieerhaltung gilt in der Bilanz $Q = Q_1 + Q_2$, d.h.

$$\int_G \frac{\partial(c\rho T)}{\partial t}\ dx = \int_G \nabla\cdot(\lambda\nabla T)\ dx + \int_G \rho W\ dx$$

bzw.

$$\int_G \left(\frac{\partial(c\rho T)}{\partial t} - \nabla\cdot(\lambda\nabla T) - \rho W\right)\ dx = 0.$$

Da $G \subseteq \Omega$ beliebig ist, folgt dann punktweise die partielle Differentialgleichung

$$\frac{\partial(c\rho T)}{\partial t} - \nabla\cdot(\lambda\nabla T) = \rho W.$$

Im Fall eines homogenen Körpers, d.h. für konstante Stoffwerte c, ρ und λ, die *Wärmeleitungsgleichung* nach Fourier (1822)

$$\frac{\partial T}{\partial t} - a^2 \sum_{i=1}^n \frac{\partial^2 T}{\partial x_i^2} = \frac{W}{c}, \qquad a^2 := \frac{\lambda}{c\rho}. \tag{3.6}$$

Man beachte hier das Auftreten des Laplace-Operators $\Delta_x := \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}$.

Auf der Oberfläche $\partial\Omega$ wird der Wärmeübergang durch *Randbedingungen* modelliert. Das *Newton'sche Abkühlungsgesetz* lautet

$$\frac{\partial T}{\partial \vec{\nu}} + h(T - T_0) = 0, \qquad h := \frac{\alpha}{\lambda}. \tag{3.7}$$

bei Umgebungstemperatur T_0, Wärmeübergangszahl α und äußerer Normale $\vec{\nu}$ auf $\partial\Omega$. Man spricht auch von einer *Robin-Bedingung*. Grenzfälle sind für $\alpha = 0$ die *Neumann-Bedingung*

$$\frac{\partial T}{\partial \vec{\nu}} = 0$$

bzw. $\lambda \to \infty$ die *Dirichlet-Bedingung*

$$T = T_0.$$

Zum Anfangszeitpunkt $t = t_0$ gibt man als *Anfangsbedingung* eine Temperaturverteilung vor:

$$T(t_0, x) = \phi(x). \tag{3.8}$$

3.3 Separationsmethoden für die Poisson-Gleichung

Unser Ziel besteht in diesem Abschnitt darin, eine Lösung des 1. Randwertproblems der Poisson-Gleichung

$$-(\Delta u)(x) := -\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(x) = f(x), \quad x \in \Omega; \qquad u(x) = g(x), \quad x \in \partial\Omega \tag{3.9}$$

mittels sogenannter *Separationsmethoden* zu ermitteln. Bei sehr einfacher Geometrie des Gebiets Ω gelingt dabei die Überführung auf gewöhnliche Differentialgleichungen. Der Ansatz

$$u = U + W \tag{3.10}$$

führt wegen der Linearität des Problems (3.9) auf die Teilprobleme

$$-(\Delta W)(x) = 0, \qquad x \in \Omega; \qquad W(x) = g(x), \quad x \in \partial\Omega, \tag{3.11}$$

$$-(\Delta U)(x) = f(x), \quad x \in \Omega; \qquad U(x) = 0, \qquad x \in \partial\Omega. \tag{3.12}$$

Zur Motivation betrachten wir zunächst den eindimensionalen Fall (vgl. auch Übungsserie 1).

Beispiel 3.2. Sei $\Omega = (0, \pi) \subset \mathbf{R}$ sowie $g(0) = \alpha$, $g(\pi) = \beta$. Lösung von Problem (3.11) ist zum Beispiel

$$W(x) = \alpha + \frac{\beta - \alpha}{\pi} x.$$

Zur Lösung von (3.12) nutzt man das zugehörige Eigenwertproblem (EWP)

$$-\phi''(x) = \lambda\phi(x), \quad 0 < x < \pi; \qquad \phi(0) = \phi(\pi) = 0 \tag{3.13}$$

mit den Lösungen

$$\phi(x) = C_1 \cos(\sqrt{\lambda}x) + C_2 \sin(\sqrt{\lambda}x).$$

Die Auferlegung der homogenen Randbedingungen liefert $\phi(0) = C_1 = 0$ sowie die Eigenwertgleichung $\phi(\pi) = C_2 \sin(\sqrt{\lambda}\pi) = 0$. Letztere Gleichung hat die Eigenwerte $\lambda_k = k^2$, $k \in \mathbf{N}$ mit den nichttrivialen Eigenfunktionen $\phi_k(x) = c_k \sin(kx)$. Man kann nun zeigen, daß die normierten Eigenfunktionen $\phi_k(x) = \sqrt{\frac{2}{\pi}} \sin(kx)$ ein Orthonomalsystem bezüglich des Skalarproduktes $(v, w) := \int_0^\pi v(x)w(x)\ dx$ bilden.

Für Funktionen $f \in C[0, \pi]$, die dort auch stückweise stetig differenzierbar sind, gilt die folgende Fourier-Darstellung auf $(0, \pi)$ sogar punktweise:

$$f(x) = \sum_{k=1}^{\infty} f_k \phi_k(x), \qquad f_k := \int_0^\pi f(x)\phi_k(x)\ dx. \tag{3.14}$$

Der formale Ansatz

$$U(x) = \sum_{k=1}^{\infty} u_k \phi_k(x) \tag{3.15}$$

führt nach Einsetzen in die Differentialgleichung in (3.12) und unter Beachtung der Reihenentwicklung (3.14) für f sowie des EWP (3.13) auf

$$-\frac{d^2}{dx^2}\left(\sum_{k=1}^{\infty} u_k \phi_k\right) = -\sum_{k=1}^{\infty} u_k \phi_k'' = \sum_{k=1}^{\infty} u_k \lambda_k \phi_k = f = \sum_{k=1}^{\infty} f_k \phi_k.$$

Koeffizientenvergleich ergibt $\lambda_k u_k = f_k$ und somit die formale Lösungsdarstellung

$$U(x) = \sum_{k=1}^{\infty} \frac{f_k}{\lambda_k}\phi_k = \frac{2}{\pi} \sum_{k=1}^{\infty} \frac{1}{k^2} \left(\int_0^\pi f(x) \sin(kx)\ dx\right) \sin(kx). \tag{3.16}$$

Eine Rechtfertigung für diese formale Lösungsdarstellung gibt der nachfolgende

Satz 3.3. *Die Funktion $f \in C[0,\pi]$ sei stückweise stetig differenzierbar. Dann ist die Funktion U aus (3.16) klassische Lösung des RWP (3.12), d.h. es gilt $U \in C[0,\pi] \cap C^2(0,\pi)$.*

Beweis: Wir zeigen zunächst, daß $U \in C[0,\pi]$. Nach der sogenannten Besselschen Ungleichung für Fourier-Reihen gilt (bereits für Funktionen $f \in L^2(0,\pi)$) mit geeigneter Konstante C

$$\sum_{k=1}^{\infty} f_k^2 \le C \int_0^{\pi} [f(x)]^2 dx < \infty.$$

Folglich sind die Koeffizienten $|f_k|$ gleichmäßig beschränkt, d.h. $|f_k| \le K$ für alle $k \in \mathbf{N}$. Dann ist aber

$$|U(x)| \le \frac{2K}{\pi} \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty, \qquad \forall x \in [0,\pi].$$

Nach dem Majorantenkriterium konvergiert dann die Funktionenreihe für U gleichmäßig gegen eine stetige Grenzfunktion, die wir mit U identifizieren können.
Wir zeigen nun, daß sogar $U \in C^2(0,\pi)$ gilt. Bei formaler zweifacher Differentiation in (3.16) ergibt sich

$$-U''(x) = \frac{2}{\pi} \sum_{k=1}^{\infty} \left(\int_0^{\pi} f(x) \sin(kx)\ dx \right) \sin(kx).$$

Dies ist gerade die Fourier-Entwicklung von f. Unter den Voraussetzungen an f ist aus der Theorie der Fourier-Reihen bekannt, daß die Konvergenz der Reihe auch punktweise gilt. Wegen $f \in C[0,\pi]$ ist dann $U \in C^2(0,\pi)$, somit sind alle Operationen bei der formalen Herleitung der Lösungsdarstellung gerechtfertigt und U ist klassische Lösung von Problem (3.12). □

Wir betrachten nun den zweidimensionalen Fall des RWP (3.9) im Spezialfall $\Omega = (0,\pi) \times (0,\pi)$ und beginnen mit der Lösung von Problem (3.11). Dazu wählen wir den Ansatz

$$W = W_1 + W_2 \qquad \Delta W_i = 0, \quad i = 1,2 \tag{3.17}$$

mit den Randbedingungen

$$W_1|_{\partial\Omega} = \begin{cases} 0; & x_1 \in \{0,\pi\},\ 0 < x_2 < \pi, \\ \psi_1(x_1) := g(x_1,0); & 0 < x_1 < \pi \\ \psi_2(x_1) := g(x_1,\pi); & 0 < x_1 < \pi. \end{cases} \tag{3.18}$$

$$W_2|_{\partial\Omega} = \begin{cases} 0; & x_2 \in \{0,\pi\},\ 0 < x_1 < \pi, \\ \rho_1(x_2) := g(0,x_2); & 0 < x_2 < \pi \\ \rho_2(x_2) := g(\pi,x_2); & 0 < x_2 < \pi. \end{cases} \tag{3.19}$$

O.B.d.A. lösen wir das Problem für $W = W_1$. Dazu wählen wir den *Separationsansatz*

$$W(x_1,x_2) = X(x_1)\ Y(x_2). \tag{3.20}$$

Nach Einsetzen in die Differentialgleichung von (3.11) folgt

$$-\Delta W = -X''(x_1)Y(x_2) - X(x_1)Y''(x_2) = 0.$$

Nach Division durch XY erhalten wir formal über

$$-\frac{X''(x_1)}{X(x_1)} = \frac{Y''(x_2)}{Y(x_2)} = \lambda = \text{konst.}$$

das System entkoppelter eindimensionaler Probleme (vgl. Beispiel 3.2)

$$X''(x_1) + \lambda X(x_1) = 0, \quad 0 < x_1 < \pi; \qquad X(0) = X(\pi) = 0$$

mit den Eigenwerten $\lambda_k = k^2,\ k \in \mathbf{N}$ und den Eigenfunktionen $X_k(x_1) = \sin(kx_1)$ sowie

$$Y''(x_2) - \lambda_k Y(x_2) = 0, \quad 0 < x_2 < \pi$$

mit den Lösungen $Y_k(x_2) = a_k \sinh(kx_2) + b_k \sinh(k(\pi - x_2))$. Formale Summation (Superposition) liefert

$$W_1(x_1, x_2) = \sum_{k=1}^{\infty} \sin(kx_1) \left[a_k \sinh(kx_2) + b_k \sinh(k(\pi - x_2))\right].$$

Man entwickelt nun die Randfunktionen $\psi_i(x_2)$ in eine Fourier-Darstellung analog (3.14)

$$\psi_i(x_1) = \sum_{k=1}^{\infty} \psi_{ik} \sin(kx_1), \qquad \psi_{ik} := \frac{2}{\pi} \int_0^{\pi} \psi_i(x_1) \sin(kx_1)\ dx_1, \qquad i = 1, 2.$$

Dies ergibt

$$\begin{aligned} W_1(x_1, 0) &= \sum_{k=1}^{\infty} b_k \sinh(k\pi) \sin(kx_1) = \psi_1(x_1) \\ W_1(x_1, \pi) &= \sum_{k=1}^{\infty} a_k \sinh(k\pi) \sin(kx_1) = \psi_2(x_1). \end{aligned}$$

Nach Koeffizientenvergleich folgt

$$b_k = \frac{\psi_{1k}}{\sinh(k\pi)}, \qquad a_k = \frac{\psi_{2k}}{\sinh(k\pi)},$$

wir erhalten somit die formale Lösung

$$W_1(x_1, x_2) = \sum_{k=1}^{\infty} \frac{\sin(kx_1)}{\sinh(k\pi)} \left[\psi_{2k} \sinh(kx_2) + \psi_{1k} \sinh(k(\pi - x_2))\right]. \tag{3.21}$$

Analog bestimmt man formal die Funktion $W_2(x_1, x_2)$.

Wir kommen nun zur Bestimmung der Lösung des RWP (3.12). Die Idee besteht wieder in der Entwicklung der Lösung U nach Eigenfunktionen, d.h. Lösungen des zugehörigen EWP

$$-\Delta\phi(x_1, x_2) = \lambda\phi(x_1, x_2), \quad 0 < x_1, x_2 < \pi; \qquad \phi|_{\partial\Omega} = 0 \tag{3.22}$$

Mit dem Separationsansatz

$$\phi(x_1, x_2) = \prod_{i=1}^{2} \phi_i(x_i); \qquad \phi_i(0) = \phi_i(\pi) = 0, \quad i = 1, 2$$

werden die homogenen Randbedingungen erfüllt. Der Ansatz ergibt in der Differentialgleichung in (3.12)

$$-\Delta\phi = -\phi_1''(x_1)\phi_2(x_2) - \phi_1(x_1)\phi_2''(x_2) = \lambda\phi_1(x_1)\phi_2(x_2)$$

bzw. nach Division durch $\phi_1\phi_2$

$$-\left[\frac{\phi_1''(x_1)}{\phi_1(x_1)} + \frac{\phi_2''(x_1)}{\phi_2(x_2)}\right] = \lambda.$$

Dies wird erfüllt durch die entkoppelten gewöhnlichen Differentialgleichungen

$$-\frac{\phi_i''(x_i)}{\phi_i(x_i)} = \lambda^{(i)}, \qquad i = 1, 2$$

mit $\lambda = \lambda^{(1)} + \lambda^{(2)}$. Unter Beachtung der Randbedingungen erhält man eindimensionale EWP der Form

$$\phi_i''(x_i) + \lambda^{(i)}\phi_i(x_i) = 0, \quad 0 < x_i < \pi; \qquad \phi_i(0) = \phi_i(\pi) = 0$$

mit den Eigenwerten $\lambda_k^{(i)} = k_i^2$, $k_i \in \mathbf{N}$ und den Eigenfunktionen $\phi_{ik}(x_i) = c_k \sin(k_i x_i)$.
Mit dem Multiindex $\overline{k} = (k_1, k_2) \in \mathbf{N}^2$ erhalten wir somit die Eigenwerte $\lambda_{\overline{k}} = k_1^2 + k_2^2$ und die Eigenfunktionen $\phi_{\overline{k}}(x_1, x_2) = \sin(k_1 x_1)\sin(k_2 x_2)$. Man kann nun zeigen, daß die normierten Eigenfunktionen $\frac{2}{\pi}\phi_{\overline{k}}$ ein Orthonormalsystem bezüglich des Skalarproduktes $(v, w) := \int_\Omega v(x)w(x)\ dx$ bilden.

Der formale Ansatz

$$U(x_1, x_2) = \sum_{k_1,k_2=1}^{\infty} u_{\overline{k}}\phi_{\overline{k}}(x_1, x_2) \tag{3.23}$$

führt nach Einsetzen in die Differentialgleichung in (3.12) und unter Beachtung einer Reihenentwicklung

$$f(x_1, x_2) = \sum_{k_1,k_2=1}^{\infty} f_{\overline{k}}\phi_{\overline{k}}(x_1, x_2), \qquad f_{\overline{k}} := \int_\Omega f(x_1, x_2)\phi_{\overline{k}}(x_1, x_2)\ dx_1 dx_2 \tag{3.24}$$

von f analog zu (3.14) formal auf

$$-\Delta\left(\sum_{k_1,k_2=1}^{\infty} u_{\overline{k}}\phi_{\overline{k}}\right) = -\sum_{k_1,k_2=1}^{\infty} u_{\overline{k}}\Delta\phi_{\overline{k}} = \sum_{k_1,k_2=1}^{\infty} u_{\overline{k}}\lambda_{\overline{k}}\phi_{\overline{k}} \ = \ f \ = \sum_{k_1,k_2=1}^{\infty} f_{\overline{k}}\phi_{\overline{k}}.$$

Koeffizientenvergleich liefert wie in Beispiel 3.2 die Aussage

$$u_{\overline{k}} = \frac{f_{\overline{k}}}{\lambda_{\overline{k}}} = \frac{\int_\Omega f\phi_{\overline{k}}\ dx}{\lambda_{\overline{k}}}$$

bzw. als formale Lösung des RWP (3.12)

$$\begin{aligned} U(x_1, x_2) &= \sum_{k_1,k_2=1}^{\infty} \frac{\int_\Omega f\phi_{\overline{k}}\ dx}{\lambda_{\overline{k}}\int_\Omega(\phi_{\overline{k}})^2\ dx}\phi_{\overline{k}}(x) \qquad (3.25)\\ &= \frac{2}{\pi}\sum_{k_1,k_2=1}^{\infty} \frac{1}{k_1^2 + k_2^2}\left(\int_\Omega f(x_1, x_2)\sin(k_1 x_1)\sin(k_2 x_2)\ dx_1 dx_2\right)\sin(k_1 x_1)\sin(k_2 x_2). \end{aligned}$$

Die formale Herleitung der Lösungsdarstellungen von U und W läßt sich mit ähnlichen Argumenten wie in Beispiel 3.2 rechtfertigen. Somit ist dann $u = U + W$ klassische Lösung des 1. RWP der Poisson-Gleichung (3.9).

Bemerkung 3.4. Man kann die Vorgehensweise zur Lösung des 1. RWP der Poisson-Gleichung (3.9) für andere geometrisch einfache Gebiete Ω, zum Beispiel für allgemeine Quadergebiete $\Omega = \prod_{i=1}^n (a_i, b_i) \subset \mathbf{R}^n$, übertragen. Dies gilt auch im ebenen Fall $n = 2$ für Kreis-, Kreisring- oder Kreissektorgebiete sowie die entsprechenden Verallgemeinerungen im $\mathbf{R}^n$ mit $n \geq 3$. □

3.4 Finite-Differenzen-Methode für das Poisson-Problem

Wir wollen nun die numerische Lösung von Zweipunkt-RWP im eindimensionalen Fall auf mehrdimensionale Probleme erweitern. Vereinfachend betrachten wir auf dem Einheitsquadrat $\Omega = (0, 1) \times (0, 1)$ das Dirichletsche RWP der Poisson-Gleichung, d.h.

$$\begin{aligned} -(\Delta u)(x_1, x_2) := -\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) &= f(x_1, x_2), \qquad (x_1, x_2) \in \Omega \qquad (3.26)\\ u(x_1, x_2) &= g(x_1, x_2), \qquad (x_1, x_2) \in \partial\Omega. \qquad (3.27) \end{aligned}$$

Zur Definition des *klassischen Differenzen-Verfahrens* (FDM) definieren wir mit der (vereinfachend) äquidistanten Schrittweite $h = \frac{1}{N},\ N \in \mathbf{N}$ die Menge der Gitterpunkte

$$Z_h := \{(x_1, x_2) :\ \ x_1 = z_1 h, x_2 = z_2 h,\ \ z_1, z_2 \text{ ganz}\}.$$

Die Menge der inneren Gitterpunkte sei $\omega_h := Z_h \cap \Omega$, die Menge der Randgitterpunkte entsprechend $\gamma_h := Z_h \cap \partial\Omega$.

Wir approximieren die zweiten partiellen Ableitungen in x_1- bzw. x_2Richtung wie im eindimensionalen Fall durch den zentralen Differenzenquotienten 2. Ordnung, d.h.

$$\begin{aligned}(\Delta_h u)(x_1, x_2) \quad := \quad & \frac{1}{h^2}\left(\{u(x_1 + h, x_2) - 2u(x_1, x_2) + u(x_1 - h, x_2)\}\right. \\ & \left. + \{u(x_1, x_2 + h) - 2u(x_1, x_2) + u(x_1, x_2 - h)\}\right). \end{aligned} \tag{3.28}$$

Man spricht auch von einem sogenannten *Fünfpunkte-Stern.* Bezeichne wie im eindimensionalen Fall $R_h v$ die Einschränkung einer Funktion $v : \overline{\Omega} \to \mathbf{R}$ auf das Gitter $\omega_h \cup \gamma_h$. Ferner sei der Vektor $U = (U_i)_{i=1}^M$ die durch die FDM erzeugte Näherung an die Werte $R_h u$ der gesuchten stetigen Lösung auf dem Gitter. Dann lautet das dem 1. RWP der Poisson-Gleichung zugeordnete lineare Gleichungssystem

$$-\Delta_h U = R_h f \quad \text{in } \omega_h \tag{3.29}$$

$$U = R_h g \quad \text{in } \gamma_h. \tag{3.30}$$

Im Fall Dirichletscher Randbedingungen kann man (im Unterschied zum RWP 2. und 3. Art) die Randwerte $U = R_h g$ eliminieren. Die konkrete Gestalt des linearen Gleichungssystems hängt dann von der Numerierung der Gitterpunkte in ω_h ab. Der einfachste Fall entsteht bei *lexikographischer Anordnung* gemäß

$$\begin{array}{cccc} (h,h), & (2h,h), & ..., & (1-h,h) \\ (h,2h), & (2h,2h), & ..., & (1-h,2h) \\ \vdots, & \vdots, & \vdots, & \vdots, \\ (h,1-h), & (2h,1-h), & ..., & (1-h,1-h) \end{array}$$

und Numerierung der Unbekannten in den Gitterpunkten auf ω_h gemäß

$$U_1, ..., U_{N-1}, U_N, ..., U_{2N-2}, U_{2N-1}...., U_{3N-3}, ..., U_{(N-1)(N-1)}.$$

Mit der Tridagonal-Matrix

$$T = \text{tridiag}\,(-1,\ \ 4,\ \ -1) \in \ \mathbf{R}^{(N-1)\times(N-1)}$$

und der Einheitsmatrix $I \in \ \mathbf{R}^{(N-1)\times(N-1)}$ hat die entstehende Matrix des Gleichungssystems die Blocktridiagonal–Gestalt

$$A = \frac{1}{h^2}\text{tridiag}(-I,\ \ T,\ \ -I) \in \ \mathbf{R}^{(N-1)^2\times(N-1)^2} \tag{3.31}$$

(vgl. Übungsaufgabe).

Man charakterisiert Differenzenverfahren auf regelmäßigen Gittern oft durch *Differenzensterne* bezüglich eines Gitterpunktes (x_1, x_2). Im allgemeinen Fall entsteht als Approximation des Differentialoperators bei geeigneten Größen c_{ij} das Schema

$$\sum_{i,j} c_{ij} U(x_1 + ih, x_2 + jh).$$

Für den Fall $|i|, |j| \leq 1$ spricht man von *kompakten* Differenzen-Sternen. Der allgemeinste Fall ist dabei dann ein Neunpunkte-Stern. Der oben genannte Fünfpunkte-Stern ist ein Spezialfall.

Man kann die FDM auf allgemeineren Gebieten als dem hier betrachteten Einheitsquadrat erzeugen. Man überzieht den $\mathbf{R}^2$ erneut mit dem Gitter Z_h und verfährt in inneren Gitterpunkten $Z_h \cap \Omega$ wie oben beschrieben. Die Approximation in den randnahen Gitterpunkten erfordert jedoch eine gesonderte Behandlung. Hierzu sei - wie auch auf die Behandlung des 2. und 3. RWP - verwiesen auf die Darstellung in Großmann/Roos [10], Abschnitte 2.4.4 und 2.4.5.

Wir analysieren nun exemplarisch die gerade eingeführte klassische FDM für das 1. RWP des Poisson-Problems (3.26),(3.27) auf dem Einheitsquadrat. Dabei benutzen wir die in Abschnitt 1.4 eingeführten Grundbegriffe Konsistenz, Stabilität und Konvergenz wieder bezüglich der Maximum-Norm.

Lemma 3.5. *Die klassische Lösung des Problems (3.26),(3.27) liege in $C^4(\overline{\Omega})$. Dann gilt für den Konsistenzfehler der klassischen FDM (3.29),(3.30)*

$$\|AR_h u - R_h Lu\|_{\infty,\omega_h} \leq \frac{1}{6} h^2 \|u\|_{C^4(\overline{\Omega})}. \tag{3.32}$$

Beweis: vgl. Übungsaufgabe □

Lemma 3.6. *Die klassische FDM (3.29),(3.30) für das Problem (3.26),(3.27) ist bezüglich der Maximum-Norm stabil. Es gilt*

$$\|A^{-1}\|_\infty \leq C_S = \frac{1}{8}. \tag{3.33}$$

Beweis: Wir betrachten (ohne Beschränkung der Allgemeinheit) die bei lexikographischer Anordnung der inneren Gitterpunkte entstehende Matrix A aus (3.31). $A = (a_{ij})$ ist eine L_0-Matrix, denn es gilt $a_{ii} > 0$ sowie $a_{ij} < 0$ für $i \neq j$. Ferner prüft man sofort nach, daß die Matrix schwach diagonaldominant und irreduzibel ist. Damit ist A $M-$Matrix (vgl. Lemma 1.19), daher kann das $M-$Kriterium (vgl. Lemma 1.20) angewendet werden.

Wir nehmen vereinfachend an, daß der Punkt $(\frac{1}{2}, \frac{1}{2})$ zum Gitter ω_h gehört. Für das Polynom $e^*(x_1, x_2) := x_1(1 - x_1) + x_2(1 - x_2)$ gilt offenbar sowohl $e^* > 0$ als auch $-\Delta e^* = 4$. Für $e := R_h e^*$ gilt $-\Delta_h e = 4$, da quadratische Polynome durch den Fünfpunkte-Stern exakt diskretisiert werden. Wegen $\|e\|_{\infty,\omega_h} \leq \frac{1}{2}$ folgt nach dem $M-$Kriterium die gesuchte Aussage. □

Beide Lemmata ergeben dann die gewünschte Konvergenzaussage

Satz 3.7. *Die klassische FDM (3.29),(3.30) für das Problem (3.26),(3.27) ist unter der Regularitätsvoraussetzung $u \in C^4(\overline{\Omega})$ bezüglich der Maximum-Norm konvergent. Es gilt*

$$\|R_h u - U\|_{\infty,\omega_h} \leq \frac{1}{48} h^2 \|u\|_{C^4(\overline{\Omega})}. \tag{3.34}$$

Zur Illustration dieser Untersuchungen betrachten wir die folgenden Beispiele. Die Rechnungen wurden dazu mit einem in MATLAB erstellten Finite-Differenzen-Programm durchgeführt.

Beispiel 3.8. Gesucht wird die Lösung des Problems (3.26)- (3.27) mit $f(x_1, x_2) = 4 \sin 2\pi x_1 \sin \pi x_2$ und $g(x_1, x_2) = 0$. Die Lösung $u(x_1, x_2) = \sin 2\pi x_1 \sin \pi x_2$ entspricht damit gerade einer Eigenfunktion des Laplace-Operators mit homogenen Dirichlet-Bedingungen (vgl. Abschnitt 3.2). Die Abbildung 3.1 zeigt die Lösung und den Fehler des Finite-Differenzen-Schemas (vgl. Abschnitt 3.2) bei grober äquidistanter Schrittweite $h = 0.1$. Ferner wird der Fehler in der Supremum-Norm in zwei Diagrammen dokumentiert. In der halblogarithmischen Darstellung erkennt man sehr gut die in Satz 3.7 ermittelte quadratische Konvergenzordnung. □

Beispiel 3.9. Wir ermitteln die Lösung von Problem (3.26) - (3.27) mit $g(x_1, x_2) = 0$ und der unstetigen Quellfunktion f mit $f(x_1, x_2) = 1$ in $\Omega_0 := [0.6, 0.65] \times [0.6, 0.65]$ und $f(x_1, x_2) = 0$ in $\Omega \setminus \Omega_0$.

Die FDM-Lösung nach Abschnitt 3.2 mit der äquidistanten Schrittweite $h = 0.05$ ist in Abbildung 3.2 zu sehen. Trotz der relativ groben Diskretisierung wird die korrekte Lösung qualitativ richtig widergespiegelt. Man kann die Lösung u als Temperatur interpretieren. Insbesondere erkennt man die Rolle des Laplace-Operators, der die im Teilgebiet Ω_0 vorgegebene (unstetige) Wärmequelle diffusiv verteilt. Man vergleiche

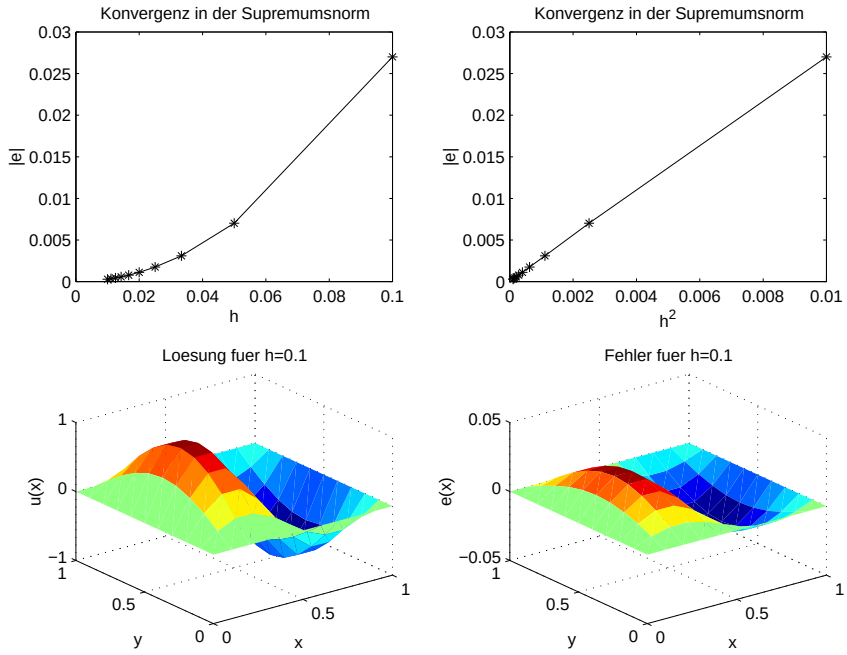


Abbildung 3.1: Lösungs- und Fehlerdarstellung zu Beispiel 3.8 für $h = 0.1$

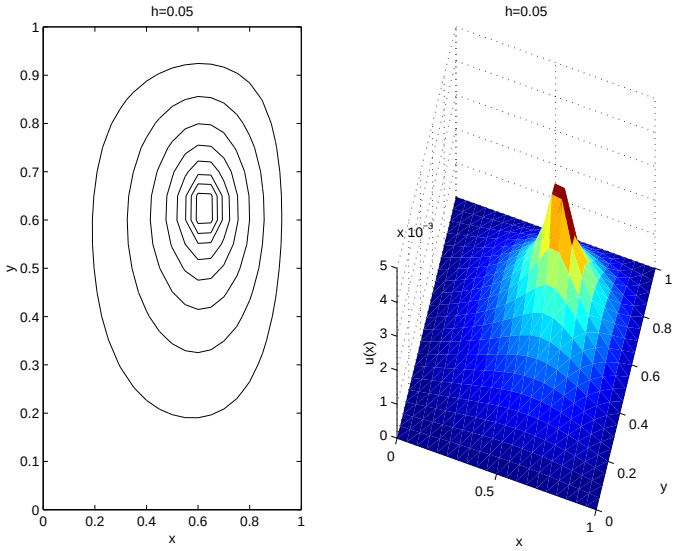


Abbildung 3.2: Lösungsdarstellung zu Beispiel 3.9 für $h = 0.05$

hierzu auch Beispiel 3.1 in Abschnitt 3.1. □

Bemerkung 3.10. Für die entsprechende Untersuchung der klassischen FDM auf allgemeineren Gebieten $\Omega \subset \mathbf{R}^2$ bzw. für das 2. und 3. RWP verweisen wir wieder auf die Darstellung bei Großmann/Roos [10], Abschnitte 2.4.4 bzw. 2.4.5. □

Kapitel 4

Klassische Lösungen elliptischer Randwertprobleme

Im vorliegenden Kapitel übertragen wir zunächst den klassischen Lösungsbegriff aus dem einführenden Kapitel zu Zweipunkt-Randwertaufgaben auf elliptische Randwertprobleme (RWP) 2. Ordnung. Eine geeignete Lösbarkeitstheorie kann in Hölder-Räumen (vgl. Abschnitt 4.1) formuliert werden, deren Ergebnisse wir hier weitgehend nach Gilbarg/ Trudinger [7] skizzieren (vgl. Abschnitt 4.2). Auf Grenzen des klassischen Lösungsbegriffs gehen wir dann in Abschnitt 4.3 ein.

4.1 Räume stetig differenzierbarer Funktionen

Seien $x = (x_1, ..., x_n)$ ein beliebiger Punkt im $\mathbf{R}^n$ und Ω ein beschränktes Gebiet im $\mathbf{R}^n$, d.h. eine offene und zusammenhängende Punktmenge. $\overline{\Omega}$ ist die abgeschlossene Hülle von Ω. Mit $\partial\Omega := \overline{\Omega} \setminus \Omega$ bezeichnen wir den *Rand des Gebietes.*

Für unsere weiteren Betrachtungen verwenden wir die in Kapitel 2 eingeführten *Räume stetig differenzierbarer Funktionen* $C^m(\Omega)$ bzw. $C^m(\overline{\Omega})$ mit $m \in \mathbf{N}_0$ (vgl. H.W. Alt: *Lineare Funktionalanalysis* [1] bzw. auch Kap. 5 der Vorlesung *Lineare Funktionalanalysis*, WS 2005/2006, kurz LFA 05/06).

Mit der punktweisen Addition und Skalarmultiplikation sind diese Räume linear. Weiterhin gilt folgende Charakterisierung.

Satz 4.1. *Sei $\overline{\Omega}$ kompakt, d.h. abgeschlossen und beschränkt. Dann ist $C^m(\overline{\Omega})$ ein Banach-Raum in Verbindung mit der Norm*

$$\|u\|_{C^m(\overline{\Omega})} := \max_{|\alpha| \le m} \sup_{x \in \Omega} |D^\alpha u(x)|, \quad u \in C^m(\overline{\Omega}). \tag{4.1}$$

Beweis: Der Nachweis der Normeigenschaften ist elementar. Beim Vollständigkeitsbeweis benutzt man insbesondere das Cauchy-Kriterium für die gleichmäßige Konvergenz von Funktionenfolgen, vgl. Satz 5.5 aus LFA 05/06. □

Es erweist sich, daß der Raum $C^2(\overline{\Omega})$ im mehrdimensionalen Fall $n \ge 2$ für eine geeignete Lösbarkeitstheorie von elliptischen Randwertproblemen 2. Ordnung nicht ausreichend ist. Hingegen erweisen sich Hölder-Räume als dafür geeignet.

Definition 4.2. *Seien $0 \le s \le 1$ und m eine nichtnegative ganze Zahl. Dann bezeichnet der* Hölder-Raum *$C^{m;s}(\overline{\Omega})$ die Menge der Funktionen aus $C^m(\overline{\Omega})$ mit*

$$\|u\|_{C^{m;s}(\overline{\Omega})} := \|u\|_{C^m(\overline{\Omega})} + \sum_{|\alpha|=m} \sup_{\substack{x,y \in \Omega \\ x \ne y}} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^s} < \infty. \tag{4.2}$$

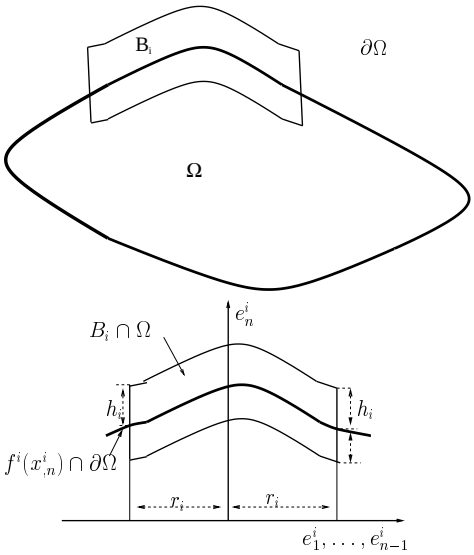


Abbildung 4.1: Prinzipskizze zur Definition 4.5 der Randglätte

Dann gilt folgende funktionalanalytische Charakterisierung.

Satz 4.3. *Sei $\overline{\Omega}$ kompakt, d.h. abgeschlossen und beschränkt. Dann ist $C^{m;s}(\overline{\Omega})$ in Verbindung mit der in Definition 4.2 definierten Norm Banach-Raum.*

Beweis: vgl. H.W. Alt [1] bzw. Übungsaufgabe in LFA 05/06 □

Es gilt insbesondere $C^m(\overline{\Omega}) = C^{m;0}(\overline{\Omega})$. Jedoch ist $C^{m;1}(\overline{\Omega}) \neq C^{m+1}(\overline{\Omega})$, man vergleiche hierzu Beispiel 4.4. Die Hölder-Räume $C^{m;s}(\overline{\Omega})$ "interpolieren" so in gewisser Weise zwischen $C^m(\overline{\Omega})$ und $C^{m+1}(\overline{\Omega})$.

Wir verwenden nun Hölder-Räume, um die *Glätte des Randes* $\partial\Omega$ eines beschränkten Gebietes zu beschreiben. Für das Verständnis der nachfolgenden Definition ist das folgende Beispiel nützlich.

Beispiel 4.4. Seien $n = 1$, $\Omega = (-1, 1)$ und $u(x) = |x|^s$. Dann gilt $u \in C^{0;s}[-1, 1]$. Für $s = 1$ ist jedoch $u \notin C^1[-1, 1]$. (Übungsaufgabe) □

Definition 4.5. *Ein beschränktes Gebiet $\Omega \subset \mathbf{R}^n$ gehört zur Klasse $C^{m;s}$ mit $m \in \mathbf{N}_0$ und $0 \leq s \leq 1$, wenn es endlich viele offene Gebiete $B_i, i = 1, ..., N$ gibt, so daß $\partial\Omega \cap B_i$ für jeweils $i = 1, ..., N$ der Graph einer $C^{m;s}$-Funktion ist und $\Omega \cap B_i$ auf jeweils einer Seite dieses Graphen liegt.*

Genauer gelte: Für $i = 1, ..., N$ gibt es ein euklidisches Koordinatensystem $(e_1^i, ..., e_n^i)$ im $\mathbf{R}^n$, Zahlen $r_i > 0$ und $h_i > 0$ sowie eine Funktion $f^i : \mathbf{R}^{n-1} \to \mathbf{R}$ aus der Klasse $C^{m;s}$, so daß mit den Bezeichnungen

$$x^i_{,n} := (x^i_1, ..., x^i_{n-1}), \qquad x = \sum_{j=1}^{n} x^i_j e^i_j$$

für $x \in \mathbf{R}^n$ mit $|x^i_{,n}| < r_i$ gilt

$$\begin{aligned} x^i_n = f^i(x^i_{,n}) \quad &\Rightarrow \quad x \in \partial\Omega \\ 0 < x^i_n - f^i(x^i_{,n}) < h_i \quad &\Rightarrow \quad x \in \Omega \\ 0 > x^i_n - f^i(x^i_{,n}) > -h_i \quad &\Rightarrow \quad x \notin \Omega. \end{aligned}$$

Die Gebiete

$$B_i := \{x \in \mathbf{R}^n |\ |x^i_{,n}| < r_i,\ |x^i_n - f^i(x^i_{,n})| < h_i\}, \quad i = 1, ..., N$$

bilden eine endlich offene Überdeckung des Randes $\partial\Omega$.

Speziell heißt ein zur Klasse $C^{0;1}$ gehörendes Gebiet Lipschitz-stetig.

Ein Gebiet besitzt ein Randstück $T \subset \partial\Omega$ der Klasse $C^{m;s}$, *falls in jedem Punkt $x_0 \in T$ eine Kugel $B = B(x_0)$ mit $B \cap \partial\Omega \subset T$ und den oben genannten Eigenschaften existiert.*

Für ein Gebiet Ω der Klasse $C^{m;s}$ besitzt jeder Punkt $x_0 \in \partial\Omega$ eine Umgebung, in der $\partial\Omega$ als Graph einer Funktion aus $C^{m;s}$ von $n-1$ der Variablen $x_1, ..., x_n$ darstellbar ist. Für $m \geq 1$ gilt auch die Umkehrung dieser Aussage.

Beispiel 4.6. (i) Kugeln $\Omega = \{x \in \mathbf{R}^n : \|x\| < r\}$ gehören zu $C^{m;s}$ für beliebige $m \in \mathbf{N}_0$, $s \in (0, 1]$.
(ii) Die Quadergebiete $\Omega = \{x = (x_1, ..., x_n) \in \mathbf{R}^n : -\infty < a_i < x_i < b_i < \infty, i = 1, ..., n\}$ sind Lipschitz-stetig. (Übungsaufgabe) □

Definition 4.7. *Sei $\Omega \subset \mathbf{R}^n$ ein beschränktes Gebiet der Klasse $C^{m;s}$. Eine Funktion $g : T \subset \partial\Omega \to \mathbf{R}$ gehört dann zur Menge $C^{m;s}(T)$, wenn $g \circ f_i^{-1} \in C^{m;s}(B \cap \partial\mathbf{R}^n_+)$ für jeden Punkt $x_0 \in T$ gilt. Dabei entspricht f_i der in Definition 4.5 eingeführten Funktion zur lokalen Beschreibung des Randstücks T in Umgebung von x_0. Ferner ist B eine geeignete Umgebung von x_0.*

Eine *wichtige Folgerung* bezieht sich auf die Fortsetzbarkeit von Funktionen: Für ein Gebiet Ω aus der Klasse $C^{m;s}$ mit $m \geq 1$ ist jede Funktion $g \in C^{m;s}(\partial\Omega)$ zu einer Funktion $\tilde{g} \in C^{m;s}(\Omega)$ fortsetzbar. Umgekehrt besitzt eine Funktion $\tilde{g} \in C^{m;s}(\Omega)$ Randwerte $g \in C^{m;s}(\partial\Omega)$ (vgl. Gilbarg/ Trudinger [7], Lemma 6.38).

Für Gebiete $\Omega \subset \mathbf{R}^n$ mit wenigstens Lipschitz-stetigem Rand $\partial\Omega$ definiert man in üblicher Weise ein $(n-1)-$dimensionales Lebesguesches Oberflächenmaß σ_{n-1} (vgl. z.B. H.W. Alt [1], Kap. 1). In fast allen Punkten $x \in \partial\Omega$ (bezüglich des $(n-1)-$dimensionalen Oberflächenmaßes) besitzt ein derartiges Gebiet einen eindeutig bestimmten äußeren Normaleneinheitsvektor $\nu = (\nu_1, ..., \nu_n)^*$. Dann gilt der folgende für viele weitere Betrachtungen grundlegende Satz.

Lemma 4.8. (Gaußscher Integralsatz)
Für jeden Vektor $(u_1, ..., u_n)^*$ von Funktionen $u_i \in C^1(\Omega) \cap C(\overline{\Omega})$ gilt auf einem beschränkten Gebiet Ω mit Lipschitz-stetigem Rand

$$\int_\Omega \sum_{i=1}^n \frac{\partial u_i}{\partial x_i} dx = \int_{\partial\Omega} \sum_{i=1}^n u_i \nu_i d\sigma_{n-1}. \tag{4.3}$$

4.2 Klassische Lösungen elliptischer RWP

Im Verlauf dieser Vorlesung untersuchen wir in einem beschränkten Gebiet $\Omega \subset \mathbf{R}^n$ Randwertprobleme für lineare partielle Differentialgleichungen 2. Ordnung

$$(Lu)(x) := -\sum_{i,j=1}^n a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j}(x) + \sum_{j=1}^n b_j(x) \frac{\partial u}{\partial x_j}(x) + c(x)u(x) = f(x), \quad x \in \Omega \tag{4.4}$$

bei gegebenen Funktionen $a_{ij} = a_{ji}, b_j, c, f : \Omega \to \mathbf{R}$, $i, j = 1, ..., n$. Einfachster und zugleich wichtiger Spezialfall von (4.4) ist die in Kapitel 3 behandelte Poisson-Gleichung mit $a_{ij} = \delta_{ij}$ sowie $b_j = c = 0$ für $i, j = 1, ..., n$.

Definition 4.9. *Der Differentialoperator L aus (4.4) heißt* gleichmäßig elliptisch *auf dem Gebiet $\Omega \subset \mathbf{R}^n$, falls eine Konstante $\lambda > 0$ existiert mit*

$$\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \geq \lambda \sum_{i=1}^n \xi_i^2, \quad \forall x \in \Omega, \ \forall \xi \in \mathbf{R}^n.$$

Definition 4.10. *Bei gegebener Funktion $g : \partial\Omega \to \mathbf{R}$ heißt das Problem*

$$(Lu)(x) = f(x), \quad x \in \Omega; \qquad u(x) = g(x), \quad x \in \partial\Omega \tag{4.5}$$

Dirichletsches Randwertproblem (oder 1. Randwertaufgabe).

Definition 4.11. *Für ein beschränktes Gebiet $\Omega \subset \mathbf{R}^n$ der Klasse $C^{2;s}$ mit $s \in (0,1]$ und hinreichend glatte Daten gemäß*

$$a_{ij} = a_{ji}, b_j, c, f \in C^{0;s}(\overline{\Omega}), \quad i,j = 1,...,n; \qquad \exists \tilde{g} \in C^{2;s}(\overline{\Omega}) : \tilde{g}|_{\partial\Omega} = g \tag{4.6}$$

heißt $u \in C^{2;s}(\overline{\Omega})$ klassische Lösung des Dirichletschen Randwertproblems (4.5) *genau dann, wenn die Gleichungen (4.5) punktweise auf Ω bzw. $\partial\Omega$ erfüllt sind.*

Es kann gezeigt werden, daß der (gegenüber Definition 4.11) *abgeschwächte* klassische Lösungsbegriff $u \in C^2(\Omega) \cap C(\overline{\Omega})$ nicht für eine geeignete Lösbarkeitstheorie für das Randwertproblem (4.5) ausreichend ist. Von Schauder stammt eine entsprechende Existenztheorie in Hölder-Räumen $C^{2;s}(\overline{\Omega})$ mit $s \in (0,1)$ (vgl. z.B. Gilbarg/ Trudinger [7], Kap. 6). Insbesondere gilt folgender *Alternativsatz,* den wir bereits für den eindimensionalen Fall in Kapitel 1 besprochen hatten.

Satz 4.12. *Unter den Voraussetzungen der Definition 4.11 gilt für die Lösbarkeit des Randwertproblems (4.5) bei gleichmäßig elliptischem Operator L die folgende Fredholm-Alternative:*

Es gilt genau einer der Fälle (i) oder (ii).

(i) Das homogene RWP

$$(Lu)(x) = 0 \quad in\ \Omega; \quad u(x) = 0 \quad auf\ \partial\Omega$$

hat nur die triviale Lösung. Dann besitzt das inhomogene RWP

$$(Lu)(x) = f(x) \quad in\ \Omega; \quad u(x) = g(x) \quad auf\ \partial\Omega$$

eine und nur eine klassische Lösung $u \in C^{2;s}(\overline{\Omega})$ für beliebige Daten f und g gemäß (4.6).

(ii) Das homogene Problem hat nichttriviale Lösungen, die einen endlichdimensionalen Teilraum von $C^{2;s}(\overline{\Omega})$ bilden.

Beweis: vgl. Gilbarg/ Trudinger [7], Theorem 6.15 □

Wir suchen nun (wie bereits im eindimensionalen Fall in Kapitel 1) nach *hinreichenden* Bedingungen für die Eindeutigkeit der Lösung des Randwertproblems (4.5) oder alternativ dafür, daß das entsprechende homogene Problem nur die triviale Lösung besitzt. Dazu kann man den folgenden Vergleichssatz benutzen, der aus dem Maximum-Prinzip folgt.

Satz 4.13. *Sei $\Omega \subset \mathbf{R}^n$ ein beschränktes Gebiet mit mindestens Lipschitz-stetigem Rand. Der Differentialoperator L aus (4.4) sei gleichmäßig elliptisch, ferner sei $c(x) \geq 0$. Für zwei Funktionen $U, V \in C^2(\Omega) \cap C(\overline{\Omega})$ gelte*

$$\begin{aligned} (LU)(x) &\leq (LV)(x) \qquad \forall x \in \Omega \\ U(x) &\leq V(x) \qquad \forall x \in \partial\Omega. \end{aligned}$$

Dann folgt $U(x) \leq V(x)$ für alle Punkte $x \in \overline{\Omega}$.

Beweis: Die Behauptung ist eine Folgerung aus dem nachstehenden *Maximum-Minimum Prinzip.*

Für die Daten des Operators L seien die Voraussetzungen von Satz 4.13 erfüllt. Für die Funktion $u \in C^2(\Omega) \cap C(\overline{\Omega})$ gelten dann folgende Aussagen:

(i) Aus $(Lu)(x) \leq 0$ folgt $u(x) \leq \max\{0; \max_{x\in\partial\Omega} u(x)\}$.

(ii) Aus $(Lu)(x) \geq 0$ folgt $u(x) \geq \min\{0; \min_{x\in\partial\Omega} u(x)\}$.

(Beweis: vgl. Übungsaufgabe - erfolgt analog zum eindimensionalen Fall) □

Als Folgerung ergibt sich die gesuchte Existenzaussage.

Satz 4.14. *Unter den Voraussetzungen der Definition 4.11 und des Satzes 4.13 gibt es eine und nur eine klassische Lösung des RWP (4.5).*

Nach einem Resultat von J.H. Michael (vgl. J. Diff. Equat. 23 (1977), 1-29) kann man das Resultat von Satz 4.14 abschwächen auf einen Lösungsbegriff in $C^{2;s}(\Omega) \cap C(\overline{\Omega})$ statt in $C^{2;s}(\overline{\Omega})$. Dabei benötigt man lediglich die Lipschitz-Stetigkeit des Randes des Gebietes sowie die Existenz einer Fortsetzung $\tilde{g} \in C^{2;s}(\Omega) \cap C(\overline{\Omega})$ für die Randwerte g.

Statt des Dirichletschen Randwertproblems kann man auch andere Randwertaufgaben für Gleichung (4.4) betrachten. Sei vereinfachend Ω ein Gebiet der Klasse $C^{2;s}$. Ein relativ *allgemeiner Fall* wird durch die Randbedingung

$$(Bu)(x) := \sum_{i=1}^{n} \beta_i(x) \frac{\partial u}{\partial x_i}(x) + \gamma(x) u(x) = g(x), \quad x \in \partial\Omega \tag{4.7}$$

beschrieben. Für $\gamma(x) = 1,\ \beta_i(x) = 0, i = 1, ..., n$ erhalten wir wieder die Dirichlet-Bedingung. Wir nehmen nun an, daß die Normalenkomponente des Vektors $\beta = (\beta_i)^*$ nicht verschwindet, d.h.

$$|(\beta \cdot \nu)(x)| := \left| \sum_{i=1}^{n} \beta_i(x) \nu_i(x) \right| \geq \gamma_0 > 0, \qquad x \in \partial\Omega. \tag{4.8}$$

Dann gilt analog zu Satz 4.14 der folgende Existenzsatz.

Satz 4.15. *Für ein beschränktes Gebiet $\Omega \subset \mathbf{R}^n$ der Klasse $C^{2;s}$ mit $s \in (0, 1]$ und hinreichend glatte Daten gemäß*

$$a_{ij}, b_j, c, f \in C^{0;s}(\overline{\Omega}); \qquad \exists \tilde{g} \in C^{2;s}(\overline{\Omega}) : \tilde{g}|_{\partial\Omega} = g$$

sowie

$$\gamma, \beta_i \in C^{1;s}(\partial\Omega)$$

sei der Operator L aus (4.4) gleichmäßig elliptisch auf Ω. Ferner gelte $c(x) \geq 0$, Ungleichung (4.8) sowie $\gamma(\beta \cdot \nu)(x) > 0$ auf $\partial\Omega$. Dann existiert eine und nur eine klassische Lösung $u \in C^{2;s}(\overline{\Omega})$ des Randwertproblems (4.4),(4.7).

Beweis: vgl. Gilbarg/Trudinger [7], Theorem 6.31 □

Für die Poisson-Gleichung gelte $\beta = \nu$. Dann heißen die speziellen Randwertprobleme mit

$$\frac{\partial u}{\partial \nu}(x) := \sum_{i=1}^{n} \nu_i(x) \frac{\partial u}{\partial x_i}(x) = g(x), \quad x \in \partial\Omega \tag{4.9}$$

bzw.

$$\frac{\partial u}{\partial \nu}(x) + \gamma(x) u(x) = g(x), \quad x \in \partial\Omega \tag{4.10}$$

Neumannsches Randwertproblem (oder 2. Randwertproblem) bzw. *Robinsches Randwertproblem* (oder 3. Randwertproblem) für die Poisson-Gleichung. Im Beispiel 6.19 in Kapitel 6 findet man eine gewisse anschauliche Interpretation von Randbedingungen 2. und 3. Art.

4.3 Grenzen des klassischen Lösungsbegriffs

Wir hatten in Abschnitt 4.2 gesehen, daß es für lineare elliptische RWP 2. Ordnung eine aus mathematischer Sicht durchaus befriedigende *klassische Lösbarkeitstheorie* in geeigneten Hölder-Räumen gibt. Für den Prototyp dieser Problemklasse, das Poisson-Problem, kann man ferner in Spezialfällen für viele Belange nützliche Lösungsformeln in Reihenform (vgl. Kap. 3) oder auch in Integralform (Potentialtheorie, vgl. zum Beispiel W. Hackbusch [11]) angeben.

Der klassische Lösungsbegriff für elliptische RWP ist aber sehr stark einschränkend. So ist er für das Dirichletsche RWP der Poisson-Gleichung

$$-(\Delta u)(x) = f(x), \quad x \in \Omega; \qquad u(x) = g(x), \quad x \in \partial\Omega$$

schon nicht mehr anwendbar, wenn die rechte Seite f oder die Randwerte g nicht stetig sind. Diese Daten des Problems stellen aber gerade bestimmte Quellterme des Diffusions-Modells dar, die in Anwendungen nicht mehr stetig sein müssen. Man vergleiche hierzu das Beispiel 3.9.

Oft ist in praxisrelevanten Aufgaben auch die Modellierung mit konstanten Koeffizienten nicht adäquat. In inhomogenen Medien betrachtet man so oft das verfeinerte Diffusionsmodell

$$-\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(a(x)\frac{\partial u(x)}{\partial x_i}\right) = f(x), \quad x \in \Omega; \qquad u(x) = g(x), \quad x \in \partial\Omega,$$

bei dem der Diffusionskoeffizient $a(\cdot)$ in der Regel nicht konstant und etwa durch unterschiedliche Materialeigenschaften nicht einmal stetig sein muß.

Aus derartigen Gründen ist die Suche nach einem *verallgemeinerten Lösungsbegriff* sinnvoll, der eine wesentliche Abschwächung der Voraussetzungen an die Daten ermöglicht. Zugleich soll dieser Begriff mit der physikalischen Modellbildung (zum Beispiel Extremalprinzipien und Erhaltungssätzen) verträglich sein. Dies führt auf die abgeschwächte Formulierung elliptischer RWP 2. Ordnung als *Variationsgleichungen*. Dies hatten wir bereits für den eindimensionalen Fall formal in Abschnitt 1.5 besprochen.

Ein weiterer wichtiger Aspekt bezieht sich auf die diskrete Approximation der Lösungen. Die in Abschnitt 3.3 betrachteten Finite-Differenzen-Verfahren (FDM) sind einerseits im Prinzip an eine sehr spezielle Gitterkonstruktion (sogenannte "kartesische Gitter") und an eine einfache Geometrie des betrachteten Lösungsgebietes gebunden. Jede Transformation auf kompliziertere Gebiete ist technisch sehr aufwendig. Andererseits sind bei der mathematischen Analyse dieser Verfahren sehr starke Forderungen an die klassische Differenzierbarkeit der Lösung zu stellen, um brauchbare Konvergenzabschätzungen zu erhalten.

Es wird sich zeigen, daß der zu entwickelnde verallgemeinerte Lösungsbegriff in natürlicher Weise auf diskrete Näherungsverfahren, sogenannte *Galerkin-Verfahren*, führt. Man hat zwar inzwischen auch für kompliziertere Probleme und Gebiete geeignete Anpassungen der FDM konstruiert, die auch erhebliche Abschwächungen des Lösungsbegriffes erlauben. Die dabei genutzten Ideen sind aber nicht so grundsätzlich von der in dieser Vorlesung verfolgten Idee der Galerkin-Verfahren entfernt.

Kapitel 5

Verallgemeinerte Lösungen elliptischer RWP 2. Ordnung

Der klassische Lösungsbegriff für RWP elliptischer Differentialgleichungen ist aus verschiedenen Gründen nicht ausreichend (vgl. Abschn. 4.3). Gegenstand dieses Abschnitts ist daher die Formulierung *verallgemeinerter Aufgabenstellungen* derartiger Probleme (vgl. Abschn. 5.2). Dazu stellen wir geeignete Funktionenräume bereit (vgl. Abschn. 5.1). Die Darstellung zu letzterem Abschnitt wird dabei bewußt knapp gehalten. Genauere Darstellungen findet man in dem Buch von H.W. Alt [1] sowie in den Abschnitten 6-8 bzw. 14 der Vorlesung *Lineare Funktionalanalysis* vom WS 05/06 (kurz LFA 05/06).

5.1 Angepaßte Funktionenräume

Wir stellen zunächst grundlegende Aussagen über angepaßte Funktionenräume zusammen. Dabei werden Grundkenntnisse der Maß- und Integrationstheorie, insbesondere des Lebesgue-Maßes, vorausgesetzt. Eine Zusammenfassung findet man zum Beispiel bei H.W. Alt [1] im Anhang A.1. Bei der Darstellung über angepaßte Funktionenräume wird weitestgehend auf Beweise verzichtet.

(i) L^p-Räume

Sei $\Omega \subset \mathbf{R}^n$ eine meßbare (nicht norwendig beschränkte) Punktmenge. Zueinander *äquivalente* integrierbare Funktionen, d.h. sie unterscheiden sich nur auf einer Menge vom Maß 0, faßt man jeweils zu einer *Äquivalenzklasse* zusammen.

Definition 5.1. *(i) Die Menge $L^p(\Omega)$ mit $1 \le p < \infty$ bezeichnet die Menge aller Äquivalenzklassen meßbarer Funktionen $u : \Omega \to \mathbf{R}$ mit*

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u(x)|^p dx \right)^{1/p} < \infty. \tag{5.1}$$

(ii) Die Menge von Äquivalenzklassen der auf Ω wesentlich beschränkten Funktionen *ist*

$$L^\infty(\Omega) := \{u : \Omega \to \mathbf{R} \ \ meßbar \mid \exists M < \infty : |u(x)| \le M \ \ f.ü. \ in \ \Omega\}$$

mit

$$\text{ess max}_{x\in\Omega}|u(x)| = \text{vrai max}_{x\in\Omega}|u(x)| := \inf M. \tag{5.2}$$

Bei beschränkter Punktmenge Ω ist der Raum $L^\infty(\Omega)$ offenbar Teilmenge aller Räume $L^p(\Omega)$ für beliebige Zahlen $p \in [1, \infty)$.

Für $p \in [1, \infty]$ ist die Menge $L^p(\Omega)$ ist ein linearer Raum. Dazu führt man Addition und Skalarmultiplikation jeweils für Repräsentanten der entsprechenden Äquivalenzklassen ein. Diese Operationen sind unabhängig

von der Auswahl der Repräsentanten. Das Nullelement in $L^p(\Omega)$ entspricht dann der Äquivalenzklasse der f.ü. auf Ω verschwindenden Funktionen.

Mit den Ausdrücken (5.1) bzw. (5.2) ist $L^p(\Omega)$ normierter Raum. Für verschiedene Abschätzungen sind folgende Ungleichungen oft nützlich, insbesondere ist Aussage (iii) die Dreiecksungleichung in $L^p(\Omega)$.

Lemma 5.2. *(i) Für $u \in L^p(\Omega)$ und $v \in L^q(\Omega)$ mit $1/p + 1/q = 1$ und $1 \le p, q \le \infty$ gilt die* Höldersche Ungleichung

$$\left| \int_\Omega u(x)v(x)dx \right| \le \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)}. \tag{5.3}$$

(ii) Für $u_i \in L^{p_i}(\Omega)$ mit $\sum_{i=1}^N 1/p_i = 1$ und $1 \le p_i \le \infty$ gilt die verallgemeinerte Höldersche Ungleichung

$$\left| \int_\Omega \prod_{i=1}^N u_i(x)dx \right| \le \prod_{i=1}^N \|u_i\|_{L^{p_i}(\Omega)}. \tag{5.4}$$

(iii) Für $u, v \in L^p(\Omega)$ mit $1 \le p \le \infty$ gilt die Minkowskische Ungleichung

$$\|u + v\|_{L^p(\Omega)} \le \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}. \tag{5.5}$$

Beweis: vgl. Vorlesung LFA 99/00, Lemmata 6.13 bzw. 6.14 □

Satz 5.3. *Die Menge $L^p(\Omega)$ der Lebesgue-integrierbaren Funktionen ist Banach-Raum bezüglich der Norm*

$$\|u\|_{L^p(\Omega)} := \begin{cases} \left(\int_\Omega |u(x)|^p dx\right)^{1/p}, & 1 \le p < \infty \\ \text{vrai } \max_{x\in\Omega} |u(x)|, & p = \infty. \end{cases} \tag{5.6}$$

Beweis: vgl. Vorlesung LFA 05/06, Satz 6.15 □

(ii) Mittelungsverfahren von Sobolev

Zur Einführung der für eine verallgemeinerte Lösungstheorie partieller Differentialgleichungen erforderlichen *Sobolev-Räume* benötigen wir dichte Teilmengen der Lebesgue-Räume reellwertiger Funktionen über meßbaren Mengen $\Omega \subset \mathbf{R}^n$.

Definition 5.4. *(i) Für eine Funktion $u : \Omega \mapsto \mathbf{R}$ bezeichnet man als* Träger von *u die Menge*

$$supp\ u := \overline{\{x | x \in \Omega :\ u(x) \neq 0\}}.$$

(ii) Eine Funktion heißt finit, *wenn ihr Träger beschränkt ist und im Gebiet Ω liegt.*
(iii) $C_0^\infty(\Omega)$ ist die Menge der bezüglich Ω finiten und unendlich oft stetig differenzierbaren Funktionen.

Beispiel 5.5. Im Raum $\mathbf{R}^n$ gehört die Funktion

$$\omega(x) = \begin{cases} C \exp\left(-\frac{1}{1-r^2}\right), & |r| < 1 \\ 0 & |r| \ge 1 \end{cases}, \quad r^2 := \sum_{i=1}^n x_i^2.$$

offenbar zu $C_0^\infty(\mathbf{R}^n)$ mit supp $\omega = \{x \in \mathbf{R}^n :\ |x| \le 1\}$ (vgl. Abb. 5.1). Die Konstante C wird so normiert, daß

$$\int_{\mathbf{R}^n} \omega(x)dx = \int_{\|x\|\le 1} \omega(x)dx = 1.$$

Ferner sei

$$\omega_h(x) := \frac{1}{h^n}\ \omega\left(\frac{x}{h}\right), \qquad h > 0.$$

Für hinreichend kleine Werte von $\|h\|$ gehört dann ω_h zu $C_0^\infty(\Omega)$. Ferner folgt nach kurzer Rechnung mit $y_j = x_j/h, j = 1, ..., n$, daß

$$\int_{\mathbf{R}^n} \omega_h(x)dx = 1.$$

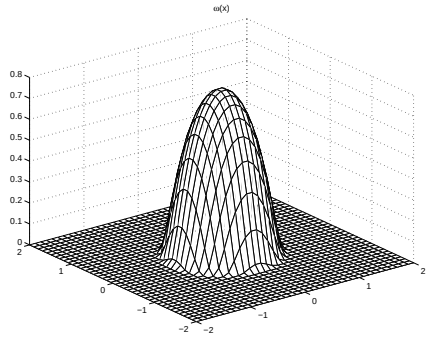


Abbildung 5.1: Sobolevsche Mittelungsfunktion $\omega(x)$ im $\mathbf{R}^2$

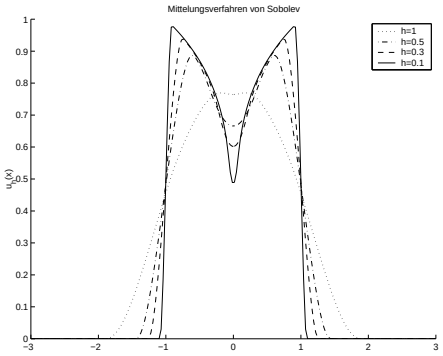


Abbildung 5.2: Sobolevsches Mittelungsverfahren für $u(x) = |x|^{\frac{1}{5}}, |x| \leq 1;\ u(x) = 0, |x| > 1$

Beispiel 5.6. (*Mittelungsverfahren von Sobolev*)
Die Funktion u gehöre zu $L^p(\Omega)$. Man setzt u außerhalb von Ω mit Null fort. Die so entstehende Funktion wird weiterhin mit u bezeichnet. Die *Sobolevsche Mittelungsfunktion* wird dann definiert durch

$$u_h(x) := \int_{\mathbf{R}^n} u(x - hy)\ \omega(y)\ dy = \int_{\|y\| \leq 1} u(x - hy)\ \omega(y)\ dy$$

bzw. nach Koordinatentransformation $z := x - hy$ durch

$$u_h(x) := \int_{\mathbf{R}^n} u(z)\ \omega\left(\frac{x - z}{h}\right) \frac{dz}{h^n} = \int_{\|x-z\| \leq h} \omega_h(x - z)\ u(z)\ dz.$$

Offenbar tragen zur Bildung von $u_h(x)$ nur die Werte von u mit $\|y - x\| \leq h$ bei.
Abbildung 5.2 zeigt die Sobolevsche Mittelungsfunktion für eine unstetige Funktion im eindimensionalen Fall bei verschiedenen Werten von h. Man erkennt die Konvergenz für $h \to 0$ gegen die Ausgangsfunktion, vgl. hierzu die Aussage von Lemma 5.7. □

Insbesondere gilt das folgende wichtige Resultat.

Lemma 5.7. *Sei $u \in L^p(\Omega)$ mit $1 \leq p < \infty$. Setzt man u außerhalb von Ω mit Null fort, so sind die Funktionen $u_h(x)$ mit $h > 0$ beliebig oft differenzierbar. Ferner ist $u_h \in L^p(\Omega)$ und es gilt*

$$\|u_h\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)}, \qquad \lim_{h \to 0} \|u - u_h\|_{L^p(\Omega)} = 0. \tag{5.7}$$

Beweis: vgl. Übungsaufgabe zur Vorlesung LFA 05/06 □

Für Beweisaussagen in Räumen verallgemeinert differenzierbarer Funktionen arbeitet man sehr oft mit den folgenden Resultaten. Insbesondere erlaubt die Aussage (ii), künftig verschiedentlich mit beliebig oft differenzierbaren Funktionen zu operieren und dann einen Dichteschluß in $L^p(\Omega)$ auszuführen.

Lemma 5.8. (*Dichte Teilmengen von $L^p(\Omega)$*)
(i) Die stetigen und finiten Funktionen liegen dicht in $L^p(\Omega)$ für $1 \le p < \infty$.
(ii) Die Menge $C_0^\infty(\Omega)$ ist dicht in $L^p(\Omega)$.

Beweis: vgl. Lemma 7.3 und Satz 7.8 der Vorlesung LFA 05/06 □

(iii) Verallgemeinerte Ableitungen

Definition 5.9. *Für meßbare Funktionen und beschränkte Gebiete Ω bezeichnet*

$$L^1_{loc}(\Omega) := \left\{ v : \Omega \to \mathbf{R} \ \ meßbar : \ \int_A |v(x)| \ dx < \infty \ \ \forall A \subset\subset \Omega \right\}$$

die Menge der lokal Lebesgue-integrierbaren Funktionen. $A \subset\subset B$ heißt dabei, daß A abgeschlossen ist und $A \subset B$ gilt.

Bemerkung 5.10. Für beschränkte Gebiete Ω gelten folgende Mengeninklusionen mit $k \in \mathbf{N}_0, \ p > 1$:

$$C_0^\infty(\Omega) \subset C^k(\overline{\Omega}) \subset L^\infty(\Omega) \subset L^p(\Omega) \subset L^1(\Omega) \subset L^1_{loc}(\Omega). \qquad \square$$

Für die weiteren Ausführungen benötigen wir die grundlegende *Regel der partiellen Integration* (vgl. Lemma 4.8) in folgender Form für $u \in C^1(\overline{\Omega})$ und beliebige *Testfunktionen* $v \in C_0^\infty(\Omega)$:

$$\int_\Omega \frac{\partial u}{\partial x_i} v \ dx = -\int_\Omega u \frac{\partial v}{\partial x_i} \ dx, \qquad i = 1, ..., n. \tag{5.8}$$

Nach der Hölderschen Ungleichung ergeben die Integrale in (5.8) noch Sinn für $u, \frac{\partial u}{\partial x_i} \in L^1_{loc}(\Omega)$.

Definition 5.11. *Gilt für eine Funktion $w_i \in L^1_{loc}(\Omega)$ die Aussage*

$$\int_\Omega w_i v \ dx = -\int_\Omega u \frac{\partial v}{\partial x_i} \ dx, \qquad \forall v \in C_0^\infty(\Omega), \tag{5.9}$$

so heißt sie verallgemeinerte erste Ableitung von $u \in L^1_{loc}(\Omega)$ in x_i-Richtung. *Man schreibt $w_i = \frac{\partial u}{\partial x_i}$.*

Man kann einfach zeigen, daß verallgemeinerte Ableitungen bis auf eine Menge vom Maß Null eindeutig bestimmt sind. Induktiv definiert man dann (unter Verwendung der Multiindexschreibweise) verallgemeinerte Ableitungen *höherer* Ordnung.

Definition 5.12. *Gilt für eine Funktion $w_\alpha \in L^1_{loc}(\Omega)$ die Aussage*

$$\int_\Omega w_\alpha v \ dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha v \ dx \qquad \forall v \in C_0^\infty(\Omega), \tag{5.10}$$

so heißt sie verallgemeinerte Ableitung $D^\alpha u$ von $u \in L^1_{loc}(\Omega)$. *Man schreibt $w_\alpha = D^\alpha u$.*

Beispiele 5.13. (i) Man überlegt sich leicht, daß für $u \in C^{|\alpha|}(\Omega)$ die "klassischen" (stetigen) und verallgemeinerten Ableitungen auf Ω übereinstimmen.
(ii) Sei $\Omega \subset \mathbf{R}^n$ beschränkt, $\partial\Omega \in C^{0,1}$ und gelte

$$\overline{\Omega} = \bigcup_{i=1}^{I} \overline{\Omega}_i; \quad \Omega_i \cap \Omega_j = \emptyset, \ i \neq j; \quad \partial\Omega_i \in C^{0,1}, \quad i = 1, ..., I.$$

Sei ferner $u \in C^k(\overline{\Omega})$ und $D^\alpha u$ stückweise stetig differenzierbar, so daß für $|\alpha| \le k+1$ gilt

$$D^\alpha u|_{\Omega_i} \in C(\Omega_i); \quad D^\alpha u \ \text{ stetig fortsetzbar auf } \overline{\Omega}_i, i = 1, ..., I.$$

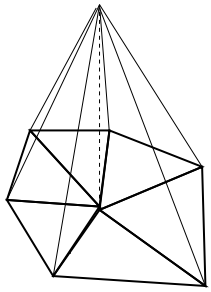


Abbildung 5.3: Finite-Elemente-Ansatzfunktionen über Dreiecksnetz im 2D-Fall

Dann ist v_α mit $v_\alpha|_{\Omega_i} = D^\alpha u$ mit $|\alpha| \leq k+1$ verallgemeinerte Ableitung von u.

Dieses Beispiel ist für die Näherungslösung elliptischer RWP mit Finite-Elemente-Verfahren von Bedeutung. Abbildung 5.3 zeigt die einfachsten Basisfunktion ("Hut-Funktionen") im zweidimensionalen Fall. Sie sind über jedem "finiten Element" Ω_i (hier: Dreieck) beliebig glatt, jedoch global (d.h. über der Vereinigung der finiten Elemente) stetig. Über die Elementkanten sind die Funktionen nicht klassisch differenzierbar, jedoch sind diese Funktionen verallgemeinert differenzierbar. □

Definition 5.14. *Für $1 \leq p \leq \infty$ ist der* Sobolev-Raum *der Funktionen mit verallgemeinerten und zur $p-$ten Potenz auf Ω integrierbaren Ableitungen bis zur Ordnung $k \in \mathbf{N}_0$ definiert durch*

$$W^{k,p}(\Omega) := \{v \in L^p(\Omega) : \ \exists D^\alpha v \in L^p(\Omega), \ \forall \alpha : |\alpha| \leq k\}. \tag{5.11}$$

Satz 5.15. *Sei Ω beschränktes Gebiet im $\mathbf{R}^n$. Dann ist der Raum $W^{k,p}(\Omega)$ Banach-Raum mit der Norm*

$$\|u\|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\alpha|\leq k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{1/p}, & 1 \leq p < \infty \\ \sum_{|\alpha|\leq k} \|D^\alpha u\|_{L^\infty(\Omega)}, & p = \infty. \end{cases} \tag{5.12}$$

Beweis: vgl. Satz 7.25 der Vorlesung LFA 05/06 □

Definition 5.16. *$W_0^{k,p}(\Omega)$ ist der Abschluß der Menge $C_0^\infty(\Omega)$ in der Norm $\|\cdot\|_{W^{k,p}(\Omega)}$.*

Bemerkung 5.17. Man kann zeigen, daß der Sobolev-Raum $W^{k,p}(\Omega)$ bei hinreichend glattem Rand $\partial\Omega$ gleich dem Abschluß der Menge $C^\infty(\Omega)$ der auf Ω unendlich oft stetig differenzierbaren Funktionen in der Norm von $W^{k,p}(\Omega)$ ist (vgl. z.B. E. Zeidler [25], Abschn. 21.4d). □

Satz 5.18. *Der Raum $W_0^{k,p}(\Omega)$ ist für $1 \leq p < \infty$ Banach–Raum mit der Norm $\|\cdot\|_{W^{k,p}(\Omega)}$.*

Beweis: vgl. Satz 7.28 der Vorlesung LFA 05/06 □

Für Anwendungen auf lineare partielle Differentialgleichungen spielt der Fall $p = 2$ eine wesentliche Rolle. Hier gilt sogar folgende Charakterisierung.

Satz 5.19. *Die Räume $W^{m,2}(\Omega)$ und $W_0^{m,2}(\Omega)$ sind Hilbert–Räume mit dem Skalarprodukt*

$$(u,v)_{W^{m,2}(\Omega)} := \sum_{0\leq|\alpha|\leq m} \int_\Omega D^\alpha u D^\alpha v \ dx.$$

Beweis: vgl. Satz 8.9 der Vorlesung LFA 05/06 □

5.2 Vertiefende Aussagen über Sobolev-Räume

In diesem Abschnitt stellen wir einige tieferliegende Aussagen über Sobolev–Räume zusammen, die wir an verschiedenen Stellen der Vorlesung zur mathematischen Formulierung und Analyse verallgemeinerter Aufgabenstellungen elliptischer RWP benötigen. Gegenstand und zeitlicher Rahmen der Vorlesung gestatten eine Darstellung der instruktiven Beweise verschiedener Aussagen leider nicht, so daß auch hier auf entsprechende Literatur verwiesen wird.

(i) Verallgemeinerte Randwerte

Die Formulierung verallgemeinerter Aufgabenstellungen für elliptische RWP erfordert eine geeignete Definition von *Randwerten* für Funktionen aus den Sobolev–Räumen $W^{k,p}(\Omega)$. Das Problem ist nicht trivial, da der Rand $\partial\Omega$ eines Lipschitz-stetigen Gebietes $\Omega \subset \mathbf{R}^n$ eine Menge vom Maß 0 bezüglich des $n-$dimensionalen Lebesgueschen Maßes ist. Werte von Funktionen aus $W^{k,p}(\Omega)$ sind aber nur bis auf eine Menge vom Maß 0 bestimmt.

Definition 5.20. *Für ein beschränktes Gebiet $\Omega \subset \mathbf{R}^n$ mit Lipschitz-stetigem Rand $\partial\Omega$ wird für $1 \le p < \infty$ definiert*

$$L^p(\partial\Omega) := \{v : \partial\Omega \to \mathbf{R} \text{ meßbar} \mid \|v\|_{L^p(\partial\Omega)} := \left(\int_{\partial\Omega} |u(x)|^p \, ds \right)^{1/p} < \infty \}. \tag{5.13}$$

Mit der Norm $\|\cdot\|_{L^p(\partial\Omega)}$ wird der Raum $L^p(\partial\Omega)$ zum Banach-Raum. Sinngemäß definieren wir auch den Raum $L^\infty(\partial\Omega)$, der mit der Norm

$$\|v\|_{L^\infty(\partial\Omega)} := \sup \operatorname{ess}_{x\in\partial\Omega} |v(x)| \tag{5.14}$$

ebenfalls Banach-Raum ist.

Der folgende Satz zeigt, daß Funktionen aus Sobolev-Räumen tatsächlich *verallgemeinerte Randwerte* zugeordnet werden können.

Satz 5.21. *Für ein beschränktes Gebiet $\Omega \subset \mathbf{R}^n$ mit Lipschitz–stetigem Rand $\partial\Omega$ sowie $1 \le p < \infty$ existiert eine Konstante $C > 0$, so daß für alle Funktionen $v \in W^{1,p}(\Omega)$ gilt*

$$\|v\|_{L^p(\partial\Omega)} \le C \|v\|_{W^{1,p}(\Omega)}. \tag{5.15}$$

Beweis: vgl. H.W. Alt: *Lineare Funktionalanalysis*, Satz A 5.7 □

Durch (5.15) wird eine lineare, stetige Abbildung

$$\gamma : \; W^{1,p}(\Omega) \to L^p(\partial\Omega), \tag{5.16}$$

die sogenannte *Spurabbildung*, motiviert. Die Menge der Werte γv von $v \in W^{1,p}(\Omega)$ heißt auch *Spur von* v, (5.15) wird als *Spurungleichung* bezeichnet.

Die angegebene Abbildung ist nicht surjektiv. Der Bildraum des Spurraums γ wird für $p = 2$ auch durch $H^{1/2}(\partial\Omega)$ bezeichnet. Hinsichtlich einer präzisen Einführung von Randräumen $H^s(\partial\Omega)$ mit $s \in \mathbf{R}$ im räumlich zweidimensionalen Fall sei auf die Monographie von R. Kreß [14], Kap. 8 verwiesen. Die Definition erfolgt hier über das (durch s definierte) Abklingverhalten von Fourier-Koeffizienten bei Reihenentwicklung der Randfunktion.

Für die Herleitung verallgemeinerter Aufgabenstellungen elliptischer RWP benötigen wir die folgende Verallgemeinerung der *Regel der partiellen Integration.*

Lemma 5.22. *Für ein beschränktes Gebiet $\Omega \subset \mathbf{R}^n$ mit Lipschitz-stetigem Rand $\partial\Omega$ und Funktionen $u \in W^{1,p}(\Omega)$, $v \in W^{1,q}(\Omega)$ mit $1 < p, q < \infty$ und $\frac{1}{p} + \frac{1}{q} = 1$ gilt für alle Multiindizes α mit $|\alpha| = 1$, daß*

$$\int_\Omega D^\alpha u \; v \; dx = -\int_\Omega u \; D^\alpha v \; dx + \int_{\partial\Omega} uv \; \nu_\alpha \; ds. \tag{5.17}$$

Dabei ist $\nu = (\nu_\alpha)$ der (fast überall existierende) äußere Normaleneinheitsvektor auf $\partial\Omega$.

Beweis: Dies ist eine Folgerung aus der klassischen Regel der partiellen Integration durch Grenzübergang in $W^{1,p}(\Omega)$ bzw. $W^{1,q}(\Omega)$, sofern der Rand $\partial\Omega$ aus C^1 ist. Die Erweiterung auf Lipschitz-stetige Gebiete findet sich z.B. bei H.W. Alt [1], Satz A 5.9. □

(ii) Äquivalente Normierungen von $W^{1,2}(\Omega)$

Definition 5.23. *Eine Abbildung $f : X \to \mathbf{R}$ heißt* stetige Linearform auf X, *falls*

$$f(\sum_{i=1}^{2} \alpha_i v_i) = \sum_{i=1}^{2} \alpha_i f(v_i), \qquad \forall v_i \in X,\ \forall \alpha_i \in \mathbf{R},\ i = 1, 2,$$

$$\exists M > 0 :\ |f(v)| \le M\|v\|_X, \qquad \forall v \in X.$$

Definition 5.24. *Eine Abbildung $a : X \times X \to \mathbf{R}$ heißt* stetige Bilinearform auf X, *falls*

$$a(\sum_{i=1}^{2} \alpha_i u_i, v) = \sum_{i=1}^{2} \alpha_i a(u_i, v), \qquad \forall u_i, v \in X,\ \forall \alpha_i \in \mathbf{R},\ i = 1, 2,$$

$$a(u, \sum_{i=1}^{2} \beta_i v_i) = \sum_{i=1}^{2} \beta_i a(u, v_i), \qquad \forall u, v_i \in X,\ \forall \beta_i \in \mathbf{R},\ i = 1, 2,$$

$$\exists K > 0 :\ |a(u, v)| \le K\|u\|_X\|v\|_X, \qquad \forall u, v \in X.$$

Für die Existenztheorie verallgemeinerter Lösungen verschiedener elliptischer Randwertprobleme 2. Ordnung benötigen wir Normen auf $W^{1,2}(\Omega)$, die zur Standardnorm äquivalent sind. Hilfreich ist dazu das folgende Resultat.

Lemma 5.25. *Sei $\Omega \subset \mathbf{R}^n$ ein beschränktes Gebiet mit Lipschitz-stetigem Rand $\partial\Omega$. Ferner seien*

$$A(\cdot, \cdot) : W^{1,2}(\Omega) \times W^{1,2}(\Omega) \to \mathbf{R} \tag{5.18}$$

eine stetige, symmetrische Bilinearform mit $A(v, v) \ge 0$ für alle $v \in W^{1,2}(\Omega)$ sowie $F(\cdot) : W^{1,2}(\Omega) \to \mathbf{R}$ eine stetige Linearform. Ferner gelte

$$A(1, 1) + |F(1)| > 0.$$

Dann ist die Norm $|||\cdot|||$ mit

$$|||v||| := |v|_{W^{1,2}(\Omega)} + \sqrt{A(v, v)} + |F(v)| \tag{5.19}$$

auf $W^{1,2}(\Omega)$ äquivalent zur Standardnorm

$$\|\cdot\|_{W^{1,2}(\Omega)} := \left(|\cdot|^2_{W^{1,2}(\Omega)} + \|\cdot\|^2_{L^2(\Omega)}\right)^{1/2}. \tag{5.20}$$

Beweis: vgl. H. Triebel [22], Satz 28.2. (A.a.O. findet man den Beweis für den Fall $\Omega \in C^{1;0}$. Der Fall Lipschitz-stetiger Fall kann durch einen technisch aufwendigen Grenzübergang, bei dem Ω durch eine geeignete Folge $C^{1;0}-$glatter Gebiete approximiert wird, bewiesen werden.) □

Insbesondere erhalten wir bei Anwendung von Lemma 5.25 folgenden

Satz 5.26. *Sei $\Omega \subset \mathbf{R}^n$ ein beschränktes Gebiet mit Lipschitz-stetigem Rand $\partial\Omega$. Dann sind die folgenden Normen äquivalent zur Standardnorm $\|\cdot\|_{W^{1,2}(\Omega)}$*

$$|||v||| := |v|_{W^{1,2}(\Omega)} + \left|\int_{\partial\Omega} v ds\right| \tag{5.21}$$

$$|||v||| := |v|_{W^{1,2}(\Omega)} + \left|\int_{\Omega} v\ dx\right| \tag{5.22}$$

$$|||v||| := |v|_{W^{1,2}(\Omega)} + \left(\int_{\Gamma_1} \eta|v|^2\ ds\right)^{1/2}, \tag{5.23}$$

falls $\Gamma_1 \subset \partial\Omega$, $meas_{n-1}(\Gamma_1) > 0$, $\eta \in L^\infty(\Gamma_1)$, $\eta > 0$ *auf* Γ_1 *sowie*

$$|||v||| := |v|_{W^{1,2}(\Omega)} + \left(\int_{\Omega_1} \eta |v|^2 \, dx \right)^{1/2}, \tag{5.24}$$

falls $\Omega_1 \subset \Omega$, $meas_n(\Omega_1) > 0$, $\eta \in L^\infty(\Omega_1)$, $\eta > 0$ *auf* Ω_1.

Beweis: Folgerung aus Lemma 5.25 □

5.3 Verallgemeinerte RWP der Poisson-Gleichung

(i) Dirichletsches RWP der Poisson-Gleichung

Wir betrachten zunächst das *homogene Dirichletsche RWP* für die Poisson-Gleichung

$$-(\Delta u)(x) \equiv -\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(x) = f(x), \quad x \in \Omega \tag{5.25}$$

$$u(x) = 0, \qquad x \in \partial\Omega \tag{5.26}$$

in einem Gebiet Ω im $\mathbf{R}^n$ mit dem Lipschitz-stetigen Rand $\partial\Omega$.

Nach Multiplikation der Differentialgleichung (5.25) mit einer beliebigen Testfunktion $v \in C_0^\infty(\Omega)$ und Integration über das Gebiet Ω erhalten wir

$$-\sum_{i=1}^{n} \int_\Omega \frac{\partial^2 u}{\partial x_i^2}(x) v(x) dx = \int_\Omega f(x) v(x) dx.$$

Nun wendet man auf der linken Seite die Regel der partiellen Integration an:

$$-\sum_{i=1}^{n} \int_\Omega \frac{\partial^2 u}{\partial x_i^2}(x) v(x) dx = \sum_{i=1}^{n} \left(\int_\Omega \frac{\partial u}{\partial x_i}(x) \frac{\partial v}{\partial x_i}(x) dx - \int_{\partial\Omega} \frac{\partial u}{\partial x_i} v \nu_i ds \right).$$

Da die Testfunktion auf dem Rand punktweise verschwindet, fällt das Randintegral weg. Damit folgt

$$\sum_{i=1}^{n} \int_\Omega \frac{\partial u}{\partial x_i}(x) \frac{\partial v}{\partial x_i}(x) dx = \int_\Omega f(x) v(x) dx, \quad \forall v \in C_0^\infty(\Omega). \tag{5.27}$$

Wir untersuchen, unter welchen Bedingungen Gleichung (5.27) noch sinnvoll bleibt. Per Definition ist der Raum $C_0^\infty(\Omega)$ dicht im Hilbert–Raum $X := W_0^{1,2}(\Omega)$. Wir definieren auf $X \times X$ die Bilinearform

$$a(u, v) := \int_\Omega \sum_{i=1}^{n} \frac{\partial u}{\partial x_i}(x) \frac{\partial v}{\partial x_i}(x) dx \tag{5.28}$$

sowie auf X die Linearform

$$f(v) := \int_\Omega f(x) v(x) dx. \tag{5.29}$$

Lemma 5.27. *Seien* $\Omega \subset \mathbf{R}^n$ *beschränktes Gebiet sowie* $f \in L^2(\Omega)$. *Dann sind durch (5.29) bzw. (5.28) eine beschränkte Linearform bzw. beschränkte Bilinearform auf* X *bzw.* $X \times X$ *definiert.*

Beweis: Die Linearität von f bzw. a folgen unmittelbar aus den Eigenschaften des Lebesgue-Integrals. Die Beschränktheit von f folgt mittels Hölderscher Ungleichung aus

$$|f(v)| \le \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \le \|f\|_{L^2(\Omega)} \|v\|_{W^{1,2}(\Omega)}.$$

Die Beschränktheit von a ergibt sich über die Höldersche Ungleichung sowie Satz 5.26, Formel (5.21) aus

$$\begin{aligned} |a(u,v)| &\leq \left(\sum_{i=1}^n \int_\Omega \left|\frac{\partial u}{\partial x_i}\right|^2 dx\right)^{1/2} \left(\sum_{i=1}^n \int_\Omega \left|\frac{\partial v}{\partial x_i}\right|^2 dx\right)^{1/2} \\ &\leq K\|u\|_{W^{1,2}(\Omega)}\|v\|_{W^{1,2}(\Omega)}. \qquad \square \end{aligned}$$

Man kann nun in der Formulierung (5.27) zu Elementen $u, v \in X$ übergehen. Seien dazu (u_n) und (v_n) Folgen in $C_0^\infty(\Omega)$ mit den Grenzwerten $u, v \in X$. Dann gilt nach Nullergänzung

$$\begin{aligned} |a(u_n,v_n) - a(u,v)| &= |a(u_n-u,v_n) + a(u,v_n-v)| \\ &\leq K\left(\|u_n-u\|_X\|v_n\|_X + \|u\|_X\|v_n-v\|_X\right) \;\to\; 0, \quad n\to\infty \\ |f(v_n)-f(v)| &\leq \|f\|_{L^2(\Omega)}\|v_n-v\|_X \to 0, \quad n\to\infty. \end{aligned}$$

Diese Vorbetrachtungen motivieren dann die nachfolgende

Definition 5.28. *Als* verallgemeinerte Aufgabenstellung *des homogenen Dirichletschen RWP der Poisson-Gleichung bezeichnet man*

$$\textit{Finde } u \in X: \quad a(u,v) = f(v), \quad \forall v \in X. \tag{5.30}$$

Die Lösung $u \in X$ heißt verallgemeinerte Lösung *des 1. RWP (5.25),(5.26).*

Für den Fall *inhomogener Dirichletscher Randbedingungen* $u(x) = g(x), x \in \partial\Omega$ sei $\tilde{g} \in W^{1,2}(\Omega)$ eine geeignete Fortsetzung, für die nach Satz 5.21 gilt $\gamma\tilde{g} = g$. Wir betrachten nun $U := u - \tilde{g} \in X = W_0^{1,2}(\Omega)$. Die entsprechende verallgemeinerte Aufgabenstellung für U lautet dann

$$\text{Finde } U \in W_0^{1,2}(\Omega): \quad a(U,v) = \tilde{f}(v)$$

mit $a(\cdot,\cdot)$ gemäß (5.28) und

$$\tilde{f}(v) := f(v) - a(\tilde{g},v).$$

Offenbar ist

$$|\tilde{f}(v)| \leq |f(v)| + |a(\tilde{g},v))| \leq \left(\|f\|_{L^2(\Omega)} + K\|\tilde{g}\|_{W^{1,2}(\Omega)}\right)\|v\|_{W^{1,2}(\Omega)}.$$

Die ursprüngliche Lösung ergibt sich dann aus $u = U + \tilde{g}$. Damit überträgt sich die Aussage von Lemma 5.27 auch auf den Fall inhomogener Dirichletscher Randwerte.

(ii) Neumannsches und Robinsches RWP der Poisson-Gleichung

Wir betrachten jetzt das folgende *Neumannsche* bzw. *Robinsche RWP*

$$-(\Delta u)(x) \equiv -\sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}(x) = f(x), \quad x \in \Omega \tag{5.31}$$

$$\sum_{i=1}^n \frac{\partial u}{\partial x_i}(x)\nu_i(x) + h(x)u(x) = g(x), \qquad x \in \partial\Omega \tag{5.32}$$

im Gebiet Ω im $\mathbf{R}^n$ mit dem Rand $\partial\Omega \in C^1$. Dabei ist $\nu(x) = (\nu_i)_{i=1}^n$ der äußere Normaleneinheitsvektor auf dem Rand. Für $h \equiv 0$ spricht man vom Neumann-Problem, für $h \geq 0$ mit $h \not\equiv 0$ vom Robin-Prblem.

Multiplikation der Gleichung (5.31) mit einer beliebigen Testfunktion $v \in C^\infty(\overline{\Omega})$, Integration über das Gebiet Ω sowie partielle Integration der linken Seite liefern

$$-\sum_{i=1}^n \int_\Omega \frac{\partial^2 u}{\partial x_i^2}(x)v(x)\,dx = \sum_{i=1}^n \left(\int_\Omega \frac{\partial u}{\partial x_i}(x)\frac{\partial v}{\partial x_i}(x)\,dx - \int_{\partial\Omega} \frac{\partial u}{\partial x_i}v\nu_i\,ds\right).$$

Im Randintegral auf der rechten Seite ergibt sich mittels Randbedingung (5.32)

$$-\sum_{i=1}^{n}\int_\Omega \frac{\partial^2 u}{\partial x_i^2}(x)v(x)\ dx = \sum_{i=1}^{n}\int_\Omega \frac{\partial u}{\partial x_i}(x)\frac{\partial v}{\partial x_i}(x)dx - \int_{\partial\Omega}(-hu+g)\,(x)v(x)\ ds.$$

Wir gelangen damit durch Umordnung zu

$$\sum_{i=1}^{n}\int_\Omega \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\ dx + \int_{\partial\Omega} huv\ ds = \int_\Omega fv\ dx + \int_{\partial\Omega} gv\ ds, \quad \forall v\in C^\infty(\overline{\Omega}). \tag{5.33}$$

Wir untersuchen, unter welchen Voraussetzungen die Formulierung (5.33) noch sinnvoll bleibt. Nach Bemerkung 5.17 ist der Raum $C^\infty(\overline{\Omega})$ dicht im Hilbert–Raum $X := W^{1,2}(\Omega)$.
Wir definieren nun auf $X\times X$ die Bilinearform

$$a(u,v) := \int_\Omega \sum_{i=1}^{n}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\ dx + \int_{\partial\Omega} huv\ ds \tag{5.34}$$

sowie auf X die Linearform

$$f(v) := \int_\Omega fv\ dx + \int_{\partial\Omega} gv\ ds. \tag{5.35}$$

Dann gilt

Lemma 5.29. *Seien $\Omega\subset\mathbf{R}^n$ beschränktes Gebiet mit Rand $\partial\Omega\subset C^1$ sowie $f\in L^2(\Omega)$, $h\in L^\infty(\partial\Omega)$, $g\in L^2(\partial\Omega)$. Dann sind durch (5.35) bzw. (5.34) eine beschränkte Linear- bzw. beschränkte Bilinearform auf X bzw. $X\times X$ definiert.*

Beweis: Die Linearität von f bzw. a folgen unmittelbar aus den Eigenschaften des Lebesgue–Integrals. Die Beschränktheit von f folgt mittels Hölderscher Ungleichung sowie Satz 5.26, Formel (5.21) aus

$$\begin{aligned} |f(v)| &\le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} \\ &\le \|f\|_{L^2(\Omega)}\|v\|_{W^{1,2}(\Omega)} + C\|g\|_{L^2(\partial\Omega)}\|v\|_{W^{1,2}(\Omega)}. \end{aligned}$$

Die Beschränktheit von a ergibt sich über die verallgemeinerte Höldersche Ungleichung sowie erneut Satz 5.26, Formel (5.21) aus der Abschätzung

$$\begin{aligned} |a(u,v)| &\le \left(\sum_{i=1}^{n}\int_\Omega\left|\frac{\partial u}{\partial x_i}\right|^2 dx\right)^{1/2}\left(\sum_{i=1}^{n}\int_\Omega\left|\frac{\partial v}{\partial x_i}\right|^2 dx\right)^{1/2} \\ &\quad +\|h\|_{L^\infty(\partial\Omega)}\|u\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} \\ &\le \|u\|_{W^{1,2}(\Omega)}\|v\|_{W^{1,2}(\Omega)} + C\|h\|_{L^\infty(\partial\Omega)}\|u\|_{W^{1,2}(\Omega)}\|v\|_{W^{1,2}(\Omega)}. \qquad \square \end{aligned}$$

Man kann nun in der Formulierung (5.33) durch Dichteschluß wieder zu Elementen $u,v\in X$ übergehen. Diese Vorbetrachtungen motivieren dann die nachfolgende

Definition 5.30. *Als* verallgemeinerte Aufgabenstellung *des Neumannschen bzw. Robinschen RWP der Poisson-Gleichung bezeichnet man*

$$\textit{Finde } u\in X: \quad a(u,v) = f(v), \quad \forall v\in X. \tag{5.36}$$

Die Lösung $u\in X$ heißt verallgemeinerte Lösung *des 2. bzw. 3. RWP.*

Kapitel 6

Existenz und Regularität verallgemeinerter Lösungen

Im vorliegenden Kapitel diskutieren wir Existenz, Eindeutigkeit und Regularität verallgemeinerter Lösungen strikt elliptischer (oder koerzitiver) Variationsgleichungen. Dazu nutzen wir einige wichtige Hilfsmittel der Funktionalanalysis, insbesondere den Darstellungssatz von Riesz. Wir beschränken uns hier auf die Aussagen der Lax-Milgram Theorie. Weitergehende Aussagen sind über den Fredholmschen Alternativsatz für kompakte Operatoren möglich (vgl. Vorlesung *Lineare Funktionalanalysis* WS 05/06).

6.1 Lax-Milgram Theorie

Zunächst stellen wir die Aussagen der *Lax-Milgram Theorie* in Hilbert-Räumen zusammen. Sei X Hilbert-Raum über $\mathbf{R}$ mit Skalarprodukt $(\cdot,\cdot)$ und der induzierten Norm $\|\cdot\|_X = \sqrt{(\cdot,\cdot)}$. Dann bezeichnet $X^* = \mathcal{L}(X,\mathbf{R})$ den zugehörigen Dualraum, d.h. die Menge der stetigen, linearen Funktionale auf X. Wir verwenden ferner die folgende Schreibweise für lineare Funktionale $f \in X^*$:

$$\langle f, v\rangle := f(v) \quad \forall v \in X.$$

Der Raum X^* ist sogar Banach-Raum mit der Norm $\|\cdot\|_{X^*}$ mit

$$\|f\|_{X^*} := \sup_{\|v\|_X=1} |f(v)|.$$

Grundlegend für unsere weiteren Betrachtungen ist der folgende *Darstellungssatz von Riesz* für Funktionale, der wegen der Symmetrie des Skalarproduktes in folgender Form formuliert werden kann.

Lemma 6.1. *Auf einem reellen Hilbert-Raum X existiert zu jedem Funktional $f \in X^*$ ein eindeutig bestimmtes Element $u \in X$, so daß*

$$f(v) = \langle f, v\rangle = (u, v) \quad \forall v \in X. \tag{6.1}$$

Der Rieszsche Darstellungsoperator $R : X^* \to X$ mit $R : f \mapsto u$ ist linear, bijektiv und isometrisch.

Beweis: 1) *Eindeutigkeit:* Wir nehmen an, es existieren zwei Elemente $u_1, u_2 \in X$ mit der gesuchten Eigenschaft. Aus

$$0 = (u_1, v) - (u_2, v) = (u_1 - u_2, v) \quad \forall v \in X$$

folgt mit $0 = u_1 - u_2$ über die Eigenschaften des Skalarprodukts die Eindeutigkeit.

2) *Konstruktion des Elements u:* Sei $f \neq 0$. Dann kann ein Element $w \in X$ mit $f(w) \neq 0$ gewählt werden. Wegen der Stetigkeit von f ist der Kern bzw. der Nullraum

$$N(f) := \{v \in X : \ f(v) = 0\}$$

abgeschlossener und damit vollständiger Unterraum des Hilbert-Raumes X.

Wir nutzen jetzt Aussagen über die Bestapproximation bezüglich vollständiger Unterräume von Hilbert-Räumen (vgl. Vorlesung LFA 05/06, Sätze 8.12 und 8.13). Danach existiert die Bestapproximation $\tilde{w} \in N(f)$ an w bezüglich $N(f)$ mit $w - \tilde{w} \perp N(f)$.
Wir setzen nun $\psi := w - \tilde{w}$. Wegen

$$f(f(\psi)v - f(v)\psi) = f(\psi)f(v) - f(v)f(\psi) = 0$$

ist

$$f(\psi)v - f(v)\psi \in N(f), \quad \forall v \in X,$$

also $(\psi, f(\psi)v - f(v)\psi) = 0$. Daraus errechnet man die gesuchte Darstellung

$$f(v) = \left(\frac{f(\psi)\psi}{\|\psi\|^2}, v\right), \quad \forall v \in X.$$

3) *Eigenschaften des Rieszschen Darstellungsoperators:*
Surjektivität: Für alle Elemente $u \in X$ definiert

$$f(v) = (u, v), \quad \forall v \in X$$

ein lineares Funktional mit $Rf = u$.

Beschränktheit und Isometrie: Die Beschränktheit folgt aus

$$|f(v)| \leq \|f\| \, \|v\|, \quad \forall v \in X.$$

Mit $u := \frac{v}{\|v\|}$ ergibt sich

$$\left| f\left(\frac{v}{\|v\|}\right)\right| = \left|\left(\frac{v}{\|v\|}, v\right)\right| = \|v\|$$

und daher wegen $\|f\| = \|v\|$ die Isometrie von R.

Linearität: Für beliebige Zahlen $\alpha, \beta \in \mathbf{R}$ und beliebige Elemente $f, g \in X^*$ gilt für alle $v \in X$

$$\begin{aligned}(R(\alpha f + \beta g), v) &= (\alpha f + \beta g)(v) = \alpha f(v) + \beta g(v) \\ &= \alpha(Rf, v) + \beta(Rg, v) = (\alpha Rf + \beta Rg, v).\end{aligned}$$

Daraus folgt $R(\alpha f + \beta g) = \alpha Rf + \beta Rg$. □

Wir erinnern an die Definitionen 5.23 und 5.24 stetiger Linearformen bzw. Bilinearformen aus Kapitel 5. Folgerung aus Lemma 6.1 ist der Darstellungssatz für stetige Bilinearformen.

Lemma 6.2. *Sei a stetige Bilinearform auf $X \times X$ nach Definition 5.24. Dann gibt es genau einen Operator $A \in \mathcal{L}(X, X^*)$ mit*

$$(i) \quad a(u, v) = \langle Au, v\rangle, \quad \forall u, v \in X, \qquad (ii) \quad \|A\|_{\mathcal{L}(X,X^*)} := \sup_{\|v\|_X = 1} \|Au\|_{X^*} \leq K.$$

Gegenstand der weiteren Untersuchungen ist nun die *Variationsgleichung*

$$\text{Finde } u \in X: \quad a(u, v) = f(v) \quad \forall v \in X. \tag{6.2}$$

Die Variationsgleichung (6.2) mit $f \in X^*$ kann nach Lemma 6.2 alternativ formuliert werden:

$$\langle Au - f, v\rangle = 0 \quad \forall v \in X \qquad \text{bzw.} \qquad Au = f \quad \text{in } X^*.$$

Andererseits ergibt der Rieszsche Darstellungssatz mit dem Riesz-Operator R

$$(RAu, v) = (Rf, v) \quad \forall v \in X \qquad \text{bzw.} \qquad RAu = Rf \quad \text{in } X.$$

Die Lösbarkeitsuntersuchung der Variationsgleichung (6.2) erfolgt mittels Fixpunktsatz von Banach nach Umformulierung in eine äquivalente *Fixpunktgleichung*

$$\text{Finde } u \in X: \quad u = T(u) := u - \rho(RAu - Rf) \tag{6.3}$$

im Hilbert-Raum X mit einem zunächst beliebigen Parameter $\rho > 0$ und dem Riesz-Operator R.
Der Banachsche Fixpunktsatz ist anwendbar, wenn gilt

$$(i)\ T: X \to X, \qquad (ii)\ T \text{ ist kontraktiv auf } X. \tag{6.4}$$

Die Eigenschaft (i) ist offenbar per Konstruktion erfüllt. Eigenschaft (ii) ist erfüllt, wenn die Lipschitz-Bedingung

$$(ii')\ \exists \tilde{L} \in [0,1): \quad \|Tv_1 - Tv_2\|_X \le \tilde{L}\|v_1 - v_2\|_X,\ \forall v_1, v_2 \in X$$

nachgewiesen wird. Dazu fordern wir eine zusätzliche Eigenschaft von a.

Definition 6.3. *Die Bilinearform $a: X \times X \to \mathbf{R}$ heißt $X-$*elliptisch (oder strikt koerzitiv auf X), *falls eine Konstante $\gamma > 0$ existiert mit*

$$a(v,v) \ge \gamma \|v\|_X^2, \quad \forall v \in X. \tag{6.5}$$

Für eine $X-$elliptische Bilinearform gilt dann

$$(RAv, v) = \ < Av, v > \ = a(v,v) \ge \gamma \|v\|_X^2.$$

Andererseits ist

$$\|RAv\|_X = \|Av\|_{X^*} \le \|A\| \|v\|_X \le K\|v\|_X.$$

Dann ergibt sich unter Beachtung der beiden letzten Beziehungen mit $v = v_1 - v_2$ die Lipschitz-Stetigkeit von T

$$\begin{aligned} \|Tv_1 - Tv_2\|_X^2 &= \|v - \rho RAv\|_X^2 = (v - \rho RAv, v - \rho RAv) \\ &= \|v\|_X^2 - 2\rho(RAv, v) + \rho^2 \|RAv\|_X^2 \\ &\le (1 - 2\rho\gamma + \rho^2 K^2)\|v\|_X^2 =: L(\rho)\|v_1 - v_2\|_X^2, \end{aligned}$$

d.h. (ii') wäre für $L(\rho) \in [0,1)$ erfüllt. Nun ist aber $L(0) = L(2\gamma/K^2) = 1$. Wegen der $X-$Elliptizität und Beschränktheit von a ist

$$\gamma \|v\|_X^2 \le (RAv, v) \le \|RAv\|_X \|v\|_X \le K\|v\|_X^2, \quad \forall v \in X, \tag{6.6}$$

d.h. $\gamma \le K$. Daraus folgt

$$L\left(\frac{\gamma}{K^2}\right) = \frac{K^2 - \gamma^2}{K^2} \ge 0.$$

Also liegt die Konstante $L(\rho)$ in $[0,1)$ genau für $0 < \rho < 2\gamma/K^2$.

Damit folgt die Existenz und Eindeutigkeit der Lösung $u \in X$ der Variationsgleichung. Ferner gilt nach Einsetzen von $v = u$ in (6.6) die folgende *a-priori Abschätzung* der Lösung:

$$\gamma \|u\|_X^2 \le (RAu, u) \le \|RAu\|_X \|u\|_X,$$

d.h.

$$\|u\|_X \le \frac{1}{\gamma}\|RAu\|_X \le \frac{1}{\gamma}\|Au\|_{X^*}.$$

Wir fassen die Ergebnisse zusammen im folgenden zentralen Resultat.

Satz 6.4. *(Lemma von Lax-Milgram)*
Auf dem Hilbert-Raum X seien $a: X \times X \to \mathbf{R}$ eine stetige, $X-$elliptische Bilinearform und $f: X \to \mathbf{R}$

eine stetige Linearform. Dann existiert eine und genau eine Lösung $u \in X$ der Variationsgleichung (6.2). Sie genügt der Abschätzung

$$\|u\|_X \leq \frac{1}{\gamma}\|f\|_{X^*}.$$

Auf der Fixpunktform (6.3) basiert auch das *Verfahren der sukzessiven Approximation* als konstruktives Lösungsverfahren: Sei $u^{(0)} \in X$ ein beliebiger Startwert des Verfahrens. Dann löse man für $n \in \mathbf{N}_0$

$$u^{(n+1)} := T(u^{(n)}) := u^{(n)} - \rho R\left(Au^{(n)} - f\right). \tag{6.7}$$

Satz 6.5. *Die Voraussetzungen von Satz 6.4 seien erfüllt. Ferner gelte $0 < \rho < 2\gamma/(K^2)$. Dann konvergiert die Lösungsfolge $(u^{(n)})$ der sukzessiven Approximation bei beliebigem Startwert $u^{(0)} \in X$ gegen die eindeutig bestimmte Lösung $u \in X$ der Variationsgleichung (6.2). Ferner gilt mit $L(\rho) := 1 - 2\rho\gamma + \rho^2 K^2$ die Fehlerabschätzung*

$$\|u - u^{(n)}\|_X \leq \frac{[L(\rho)]^{n/2}}{1 - [L(\rho)]^{1/2}}, \quad n \in \mathbf{N}_0.$$

Beweis: Folgerung aus dem Fixpunktsatz von Banach. □

Bemerkung 6.6. Das Verfahren der sukzessiven Approximation kann alternativ als *pseudo-instationäres* Lösungsverfahren

$$\frac{u^{(n+1)} - u^{(n)}}{\rho} = R(f - Au^n), \quad n \in \mathbf{N}_0$$

oder als *Defektkorrekturverfahren*

$$R^{-1}\left(u^{(n+1)} - u^{(n)}\right) = \rho\left[f - Au^n\right], \quad n \in \mathbf{N}_0 \tag{6.8}$$

interpretiert werden. Bei Kenntnis von R^{-1} kann jedes Problem (6.2) iterativ durch ein Problem vom Typ (6.8) gelöst werden. Man hofft, daß diese Operatorgleichungen einfacher als die ursprüngliche Variationsgleichung (6.2) zu lösen ist.

6.2 Anwendung auf elliptische RWP 2. Ordnung

(i) Dirichletsches RWP

Wir betrachten nun in einem beschränkten Gebiet $\Omega \subset \mathbf{R}^n$ das homogene Dirichletsche Randwertproblem für allgemeinere lineare partielle Differentialgleichungen 2. Ordnung in sogenannter *Divergenzform*

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)(x) + \sum_{j=1}^{n} b_j(x)\frac{\partial u}{\partial x_j}(x) + c(x)u(x) = f(x), \quad x \in \Omega \tag{6.9}$$

$$u(x) = 0, \quad x \in \partial\Omega \tag{6.10}$$

bei gegebenen und hinreichend glatten Funktionen $a_{ij}, b_j, c, f : \Omega \to \mathbf{R}$, $i, j = 1, ..., n$.

Bei formaler Herleitung einer verallgemeinerten Aufgabenstellung wie in Abschnitt 5.3 gelangen wir zu

$$\int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \left\{\sum_{j=1}^{n} b_j(x)\frac{\partial u}{\partial x_j} + c(x)u\right\} v\right) dx = \int_\Omega f(x)v dx. \tag{6.11}$$

Wir definieren

$$a(u,v) := \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \left\{\sum_{j=1}^{n} b_j\frac{\partial u}{\partial x_j} + cu\right\} v\right) dx, \tag{6.12}$$

$$f(v) := \int_\Omega fv\, dx. \tag{6.13}$$

Grenzübergang von $u, v \in C_0^\infty(\Omega)$ zu Elementen im Hilbert-Raum $X := W_0^{1,2}(\Omega)$ ergibt

Definition 6.7. *Als* verallgemeinerte Aufgabenstellung *des homogenen Dirichletschen Randwertproblems (6.9),(6.10) bezeichnet man*

$$\textit{Finde } u \in X: \quad a(u,v) = f(v), \quad \forall v \in X. \tag{6.14}$$

Die Lösung $u \in X$ heißt verallgemeinerte Lösung *von (6.9),(6.10).*

Wir untersuchen nun, ob die Voraussetzungen der Lax–Milgram Theorie für diese Variationsgleichung erfüllt sind.

Lemma 6.8. *Sei $\Omega \subset \mathbf{R}^n$ beschränktes Gebiet und gelte*

$$a_{ij}, b_j, c \in L^\infty(\Omega), \;\; i,j = 1,...,n; \quad f \in L^2(\Omega). \tag{6.15}$$

Ferner möge für die symmetrische Matrix $A(x) = [a_{ij}(x)]_{i,j=1}^n$ gleichmäßig auf Ω eine positive Konstante Γ existieren mit

$$\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \le \Gamma \sum_{i=1}^n \xi_i^2 \quad \textit{in } \Omega \textit{ f.ü., } \forall \xi = (\xi_i)_{i=1}^n \in \mathbf{R}^n.$$

Dann sind $f(\cdot)$ nach (6.13) bzw. $a(\cdot,\cdot)$ nach (6.12) beschränkte Linearform auf X bzw. beschränkte Bilinearform auf $X \times X$.

Beweis: Der Raum $X = W_0^{1,2}(\Omega)$ ist mit der Seminorm $|\cdot|_{W^{1,2}(\Omega)}$ nach Satz 5.26, Formel (21) ein Hilbert-Raum. Insbesondere ergibt sich die *Friedrichsche Ungleichung*

$$\|v\|_{L^2(\Omega)} \le C_F |v|_{W^{1,2}(\Omega)}.$$

(i) Linearität von $f(\cdot)$ bzw. Bilinearität von $a(\cdot,\cdot)$ sind offensichtlich.

(ii) Mittels der Ungleichungen von Cauchy-Schwarz und Friedrichs erhalten wir

$$|f(v)| \le \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \le C_F \|f\|_{L^2(\Omega)} |v|_{W^{1,2}(\Omega)}.$$

(iii) Die Beschränktheit von $a = a_1 + a_2$ ergibt sich in zwei Schritten. Zunächst erhalten wir über die verallgemeinerte Cauchysche Ungleichung

$$\left| \sum_{i,j=1}^n \int_\Omega a_{ij}\xi_j\eta_i \right| \le \left(\sum_{i,j=1}^n \int_\Omega |a_{ij}|\,|\xi_j||\xi_i| \right)^{\frac{1}{2}} \left(\sum_{i,j=1}^n \int_\Omega |a_{ij}|\,|\eta_j||\eta_i| \right)^{\frac{1}{2}}$$

und die Voraussetzung an die Matrix $A(x)$ die Abschätzung

$$\begin{aligned}
|a_1(u,v)| &\equiv \left| \sum_{i,j=1}^n \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} dx \right| \\
&\le \left(\sum_{i,j=1}^n \int_\Omega |a_{ij}| \left| \frac{\partial u}{\partial x_j} \right| \left| \frac{\partial u}{\partial x_i} \right| dx \right)^{\frac{1}{2}} \left(\sum_{i,j=1}^n \int_\Omega |a_{ij}| \left| \frac{\partial v}{\partial x_j} \right| \left| \frac{\partial v}{\partial x_i} \right| dx \right)^{\frac{1}{2}} \\
&\le \Gamma \left(\sum_{i=1}^n \left| \frac{\partial u}{\partial x_i} \right|^2 dx \right)^{\frac{1}{2}} \left(\sum_{i=1}^n \left| \frac{\partial v}{\partial x_i} \right|^2 dx \right)^{\frac{1}{2}} \\
&= K_1 |u|_{W^{1,2}(\Omega)} |v|_{W^{1,2}(\Omega)}.
\end{aligned}$$

Weiter ist nach verallgemeinerter Hölderscher und Friedrichscher Ungleichung

$$\begin{aligned}
|a_2(u,v)| &= \left| \sum_{j=1}^{n} \int_\Omega b_j \frac{\partial u}{\partial x_j} v dx + \int_\Omega cuvdx \right| \\
&\leq \sum_{j=1}^{n} \|b_j\|_{L^\infty(\Omega)} \left\| \frac{\partial u}{\partial x_j} \right\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|c\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\
&\leq \left(\left[\sum_{j=1}^{n} \|b_j\|^2_{L^\infty(\Omega)} \right]^{\frac{1}{2}} |u|_{W^{1,2}(\Omega)} + C_F \|c\|_{L^\infty(\Omega)} |u|_{W^{1,2}(\Omega)} \right) C_F |v|_{W^{1,2}(\Omega)} \\
&\leq K_2 |u|_{W^{1,2}(\Omega)} |v|_{W^{1,2}(\Omega)}.
\end{aligned}$$

Aus den beiden Abschätzungen folgt die Beschränktheit von a. □

Lemma 6.9. *Über die Voraussetzungen von Lemma 6.8 hinaus gelte*

$$\frac{\partial b_j}{\partial x_j} \in L^\infty(\Omega), \;\; j = 1, ..., n; \qquad c(x) - \frac{1}{2} \sum_{j=1}^{n} \frac{\partial b_j}{\partial x_j} \geq 0 \;\; f.ü. \;\; in \;\; \Omega.$$

Ferner existiere für die symmetrische Matrix $A(x) = [a_{ij}(x)]_{i,j=1}^n$ gleichmäßig auf Ω eine positive Konstante γ mit

$$\gamma \sum_{i=1}^{n} \xi_i^2 \leq \sum_{i,j=1}^{n} a_{ij}(x) \xi_i \xi_j \quad in \;\; \Omega \;\; f.ü., \;\; \forall \xi = (\xi_i)_{i=1}^n \in \mathbf{R}^n.$$

Dann ist die Bilinearform $a(\cdot,\cdot)$ nach (6.12) $X-$elliptisch.

Beweis: Zunächst gilt

$$a_1(v,v) \geq \gamma \int_\Omega \sum_{i=1}^{n} \left\| \frac{\partial v}{\partial x_i} \right\|^2_{L^2(\Omega)} dx \geq \gamma |v|^2_{W^{1,2}(\Omega)}.$$

Partielle Integration von $a_2(v,v)$ führt auf

$$a_2(v,v) = \int_\Omega \left(c - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \right) dx \geq 0.$$

Die Behauptung ergibt sich nach Satz 5.26 über die Äquivalenz der Seminorm $|\cdot|_{W^{1,2}(\Omega)}$ zur Standardnorm $\|\cdot\|_{W^{1,2}(\Omega)}$ im Falle von $X = W_0^{1,2}(\Omega)$. □

Aus den Lemmata 6.8 und 6.9 ergibt sich die Anwendbarkeit des Lemmas von Lax-Milgram.

Satz 6.10. *Unter den Voraussetzungen der Lemmata 6.8 und 6.9 existiert eine und genau eine verallgemeinerte Lösung $u \in X = W_0^{1,2}(\Omega)$ des RWP (6.9),(6.10).*

Bemerkung 6.11. Der Fall inhomogener Dirichletscher Randbedingungen kann nach der Vorgehensweise im Abschnitt 5.3 ebenfalls mittels Satz 6.10 erledigt werden, wenn zusätzlich die Lipschitz-Stetigkeit des Randes $\partial\Omega$ gilt. □

Beispiel 6.12. (*Transmissionsproblem*)
Sei $\Omega \subset \mathbf{R}^n$ ein beschränktes Gebiet mit Lipschitz-stetigem Rand $\partial\Omega$, so daß mit paarweise disjunkten und Lipschitz-stetigen Gebieten $\Omega_k, k = 1, ..., K$ gilt

$$\overline{\Omega} = \cup_{k=1}^{K} \overline{\Omega_k}.$$

Sei ferner

$$a_{ij}(x) := a(x)\delta_{ij}, \;\; i,j = 1, ..., n; \quad a(x)|_{\Omega_k} = a_k > 0, \;\; k = 1, ..., K.$$

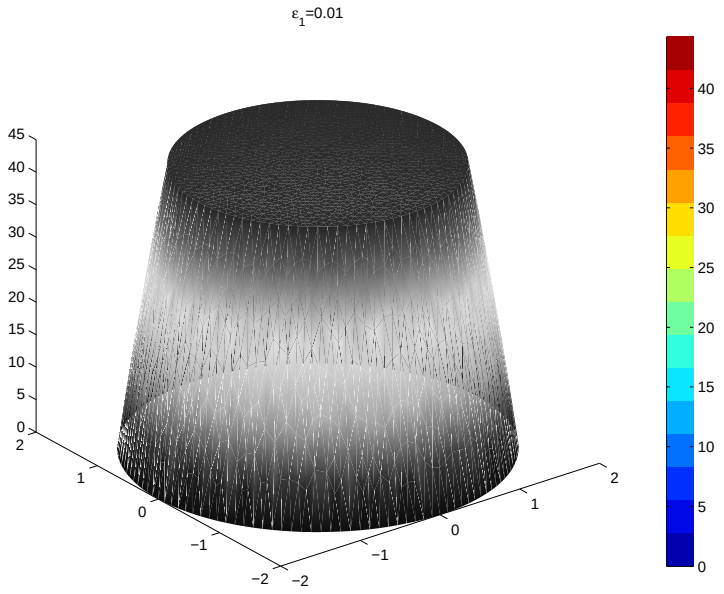


Abbildung 6.1: Lösung des Transmissionsproblems aus Beispiel 6.12

Dann hat das verallgemeinerte Problem

$$\int_\Omega \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}dx = \int_\Omega f(x)vdx.$$

eine und genau eine Lösung $u \in W_0^{1,2}(\Omega)$.

Derartige Transmissionsprobleme treten zum Beispiel bei der Wärmeleitung in einem Körper Ω auf, der aus verschiedenen Materialien mit unterschiedlicher Wärmeleitfähigkeit zusammengesetzt ist. Abbildung 6.1 zeigt die Lösung des Transmissionsproblems im Gebiet $\Omega = \{x \in \mathbf{R}^2 |\ \|x\| < 2\}$ sowie den Teilgebieten $\Omega_1 = \{x \in \mathbf{R}^2 |\ \|x\| < 1\}$ und $\Omega_2 = \Omega \setminus \Omega_1$ den Daten

$$a_1 = 0.01, \quad x \in \Omega_1; \quad a_2 = 1, \quad x \in \Omega_2; \qquad f(x) = 1, \qquad x \in \Omega.$$

Offenbar hat die Lösung keine klassischen zweiten partiellen Ableitungen auf dem Transmissionsrand $\|x\| = 1$. □

Bemerkung 6.13. (i) Die Lax-Milgram Theorie (vgl. Satz 6.4) liefert offenbar noch keine Lösbarkeitsaussage, wenn zum Beispiel die Voraussetzungen an die Terme 1. und 0. Ordnung nicht erfüllt sind. Ein wichtiger Spezialfall ist die sogenannte *Helmholtz-Gleichung*

$$-\left(\Delta u\right)(x) + c(x)u(x) = f(x), \quad x \in \Omega$$

ohne Vorzeichenbeschränkung an den Koeffizienten $c(\cdot)$. Für $c(x) = -\kappa^2 < 0$ beschreibt diese Gleichung zeitharmonische Schwingungsvorgänge. Für Lösbarkeitsaussagen derartiger Aufgaben kann man die Fredholm-Theorie kompakter Operatoren heranziehen (vgl. Kap. 19/20 der Vorlesung *LFA*, WS 05/06).

Abbildung 6.2 zeigt die (mit FEMLAB berechnete) Lösung des homogenen 1. RWP der Helmholtz-Gleichung auf dem Einheitsquadrat $\Omega = (0,1) \times (0,1)$ bei verschiedenen Werten von $c(x) = konst.$ Die Funktion $f(x) = \exp(-100\sqrt{((x_1 - 0.6)^2 + (x_2 - 0.6)^2)}$ simuliert eine "Punktquelle". Man beachte den unterschiedlichen Lösungscharakter für die Fälle $c \equiv 0$ und $c \equiv 100$ einserseits sowie $c \equiv -100$ andererseits (vgl. auch Übungsaufgabe in Serie 7).

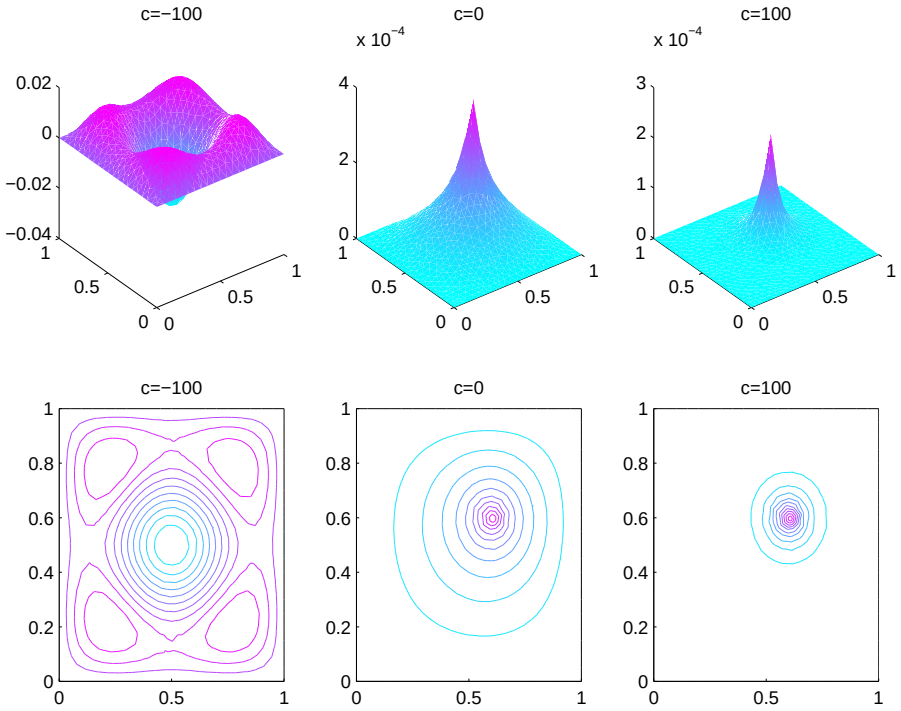


Abbildung 6.2: Lösungen des 1. RWP der Helmholtz-Gleichung für a) $c \equiv -100$, b)$c \equiv 0$, c) $c \equiv 100$

(ii) Für *singulär gestörte Probleme* (6.9)-(6.10) mit $a_{ij}(x) \equiv \epsilon\delta_{ij},\ i,j = 1,...,n$ und $0 < \epsilon \ll 1$ seien die Bedingungen von Lemma 6.9 erfüllt. Dann gilt für die Elliptizitätskonstante $\gamma \sim \epsilon$, d.h. im Grenzfall $\epsilon \to +0$ verliert man die Kontrolle über den Gradienten der Lösung. □

(ii) Neumannsches und Robinsches RWP

Auf einem beschränkten Gebiet $\Omega \subset \mathbf{R}^n$ betrachten wir das Neumannsche bzw. Robinsche RWP für eine (vereinfachend) symmetrische lineare partielle Differentialgleichungen 2. Ordnung in *Divergenzform*

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)(x) + c(x)u(x) = f(x), \quad x \in \Omega \tag{6.16}$$

$$\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial u}{\partial x_j}\nu_i(x) + h(x)u(x) = g(x), \qquad x \in \partial\Omega \tag{6.17}$$

bei gegebenen Funktionen

$$a_{ij}, c, f : \Omega \to \mathbf{R}, \quad i,j = 1,...,n, \quad a_{ij}, h, g : \partial\Omega \to \mathbf{R}, \quad i,j = 1,...,n.$$

Bei formaler Herleitung einer verallgemeinerten Aufgabenstellung wie in Abschnitt 5.3 gelangen wir zu

$$\int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + c(x)uv\right)\,dx + \int_{\partial\Omega} huv\,ds = \int_\Omega f(x)v\,dx + \int_{\partial\Omega} gv\,ds. \tag{6.18}$$

Mit

$$a(u,v) \quad := \quad \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + cuv\right)\,dx + \int_{\partial\Omega} huv\,ds \tag{6.19}$$

$$f(v) \quad := \quad \int_\Omega fv \; dx + \int_{\partial\Omega} gv \; ds \tag{6.20}$$

gelangt man durch Grenzübergang von $u, v \in C^\infty(\Omega)$ zu Elementen im Hilbert-Raum $X := W^{1,2}(\Omega)$ zur folgenden

Definition 6.14. *Als* verallgemeinerte Aufgabenstellung des *Neumannschen bzw. Robinschen RWP (6.16),(6.17) bezeichnet man*

$$\textit{Finde } u \in X: \quad a(u,v) = f(v), \quad \forall v \in X. \tag{6.21}$$

Die Lösung $u \in X$ heißt verallgemeinerte Lösung *von (6.16),(6.17).*

Wir untersuchen nun wieder, ob die Voraussetzungen der Lax–Milgram Theorie für diese Variationsgleichung erfüllt sind.

Lemma 6.15. *Sei $\Omega \subset \mathbf{R}^n$ beschränktes Gebiet mit Rand $\partial\Omega \subset C^1$ und gelte*

$$a_{ij}, c \in L^\infty(\Omega), \; i,j = 1,...,n; \quad f \in L^2(\Omega); \quad a_{ij}, h \in L^\infty(\partial\Omega), g \in L^2(\partial\Omega). \tag{6.22}$$

Ferner möge für die symmetrische Matrix $A(x) = [a_{ij}(x)]_{i,j=1}^n$ gleichmäßig auf Ω eine positive Konstante Γ existieren mit

$$\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \le \Gamma \sum_{i=1}^n \xi_i^2 \quad \textit{in } \Omega \textit{ f.ü.}, \; \forall \xi = (\xi_i)_{i=1}^n \in \mathbf{R}^n.$$

Dann ist $f(\cdot)$ nach (6.20) bzw. $a(\cdot,\cdot)$ nach (6.19) beschränkte Linearform auf X bzw. beschränkte Bilinearform auf $X \times X$.

Beweis: Der Beweis ist eine einfache Verallgemeinerung von Lemma 5.29. □

Lemma 6.16. *Über die Voraussetzungen von Lemma 6.15 hinaus gelte*

$$c(x) \ge 0, \quad x \in \Omega \textit{ f.ü.}, \quad h(x) \ge 0, \quad x \in \partial\Omega \textit{ f.ü.}$$

sowie einer der beiden Voraussetzungen

$$(i) \quad \exists \Omega_1 \subset \Omega, \; \textit{meas}_n(\Omega_1) > 0, \; c(x) \ge c_0 > 0, x \in \Omega_1$$

oder

$$(ii) \quad \exists \Gamma_1 \subset \partial\Omega, \; \textit{meas}_{n-1}(\Gamma_1) > 0, \; h(x) \ge h_0 > 0, x \in \Gamma_1.$$

Ferner möge für die symmetrische Matrix $A(x) = [a_{ij}(x)]_{i,j=1}^n$ gleichmäßig auf Ω eine positive Konstante γ existieren mit

$$\gamma \sum_{i=1}^n \xi_i^2 \le \sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \quad \textit{in } \Omega \textit{ f.ü.}, \; \forall \xi = (\xi_i)_{i=1}^n \in \mathbf{R}^n.$$

Dann ist die Bilinearform $a(\cdot,\cdot)$ nach (6.19) $X-$elliptisch.

Beweis: Wir setzen $v = u$ in (6.19). Fall (i) ergibt sich unter Beachtung von Satz 5.26, Formel (5.24)

$$a(v,v) \ge \gamma \|v\|^2_{W^{1,2}(\Omega)} + c_0 \|v\|^2_{L^2(\Omega_1)} \ge C \|v\|^2_{W^{1,2}(\Omega)}.$$

Im Fall (ii) ist sinngemäß nach Satz 5.26, Formel (5.23)

$$a(v,v) \ge \gamma \|v\|^2_{W^{1,2}(\Omega)} + h_0 \|v\|^2_{L^2(\Gamma_1)} \ge C \|v\|^2_{W^{1,2}(\Omega)}.$$

Daraus folgt die Behauptung. □

Aus den Lemmata 6.15 und 6.16 ergibt sich mit dem Lemma von Lax-Milgram der

Satz 6.17. *Unter den Voraussetzungen der Lemmata 6.15 und 6.16 existiert eine und genau eine verallgemeinerte Lösung $u \in X = W^{1,2}(\Omega)$ des Randwertproblems (6.16),(6.17).*

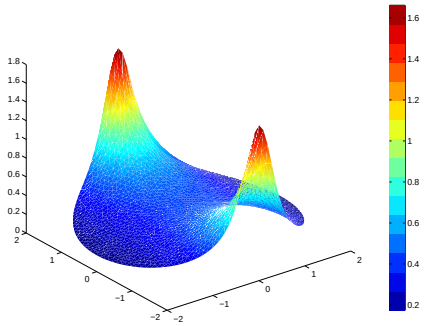


Abbildung 6.3: Lösung des 3. RWP aus Beispiel 6.19

Eine besondere Behandlung erfordert der Fall

$$c(x) = 0, x \in \Omega, \qquad h(x) = 0, x \in \partial\Omega.$$

Dies ist gerade das *Neumannsche Randwertproblem* der (verallgemeinerten) Poisson–Gleichung. Bei der Untersuchung der $X-$Elliptizität hat man nur

$$a(v, v) \geq \gamma |v|^2_{W^{1,2}(\Omega)}.$$

Wir verwenden als Ausweg die Seminorm $|\cdot|_{W^{1,2}(\Omega)}$ auf dem modifizierten Raum

$$\tilde{X} := \{v \in W^{1,2}(\Omega) : \int_\Omega v \, dx = 0\}, \tag{6.23}$$

da sie unter Beachtung von Satz 5.26, Formel (5.22) auf diesem Raum eine Norm ist. Wir erhalten dann den

Satz 6.18. *Unter den Voraussetzungen der Lemmata 6.15 und 6.16 existiert eine und genau eine verallgemeinerte Lösung $u \in \tilde{X}$ des Neumannschen Randwertproblems der (verallgemeinerten) Poisson–Gleichung (6.16),(6.17) mit $c(x) = 0, x \in \Omega$ und $h(x) = 0, x \in \partial\Omega$.*

Das folgende Beispiel veranschaulicht die unterschiedliche Wirkung der Randbedingungen 2. und 3. Art.

Beispiel 6.19. *(Randbedingungen 2. und 3. Art)*
Wir betrachten das 3. Randwertproblem der Wärmeleitungsgleichung für die Temperatur u mit

$$\begin{aligned} -\Delta u &= 0 \quad \text{in} \quad \Omega = \{x \in \mathbf{R}^2 | \ \|x\| < 2\} \\ \frac{\partial u}{\partial \nu} + h(u - g) &= 0 \quad \text{auf} \quad \partial\Omega. \end{aligned}$$

Die Randbedingung simuliert dabei den Wärmeübergang am Rand $\partial\Omega$ mit dem Wärmeübergangskoeffizienten $h \equiv 1$ und der Umgebungstemperatur g. Im Fall $g(x_1, x_2) := 5e^{-10x_1^2}$ werden hier zwei "Punktquelle" auf dem Rand simuliert (vgl. Abb. 6.3).

Ferner untersuchen wir das gemischte Randwertproblem der Wärmeleitungsgleichung für die Temperatur u mit "Wärmequelle" $f \equiv 1$ und

$$\begin{aligned} -\Delta u &= 1 \quad \text{in} \quad \Omega = \{x \in \mathbf{R}^2 | \ 1 < \|x\| < 2\}, \\ \\ u &= 1 \quad \text{auf} \quad \Gamma_1 = \{x \in \mathbf{R}^2 | \ \|x\| = 1\}, \\ \frac{\partial u}{\partial \nu} &= 0 \quad \text{auf} \quad \Gamma_2 = \{x \in \mathbf{R}^2 | \ \|x\| = 2\}. \end{aligned}$$

Auf dem Rand Γ_1 wird die Temperatur fixiert, d.h.man simuliert dort ebenfalls eine "Wärmequelle". Die homogene Neumann-Bedingung auf Γ_2 steht für den Fall der Wärmeisolierung (vgl. Abb. 6.4).

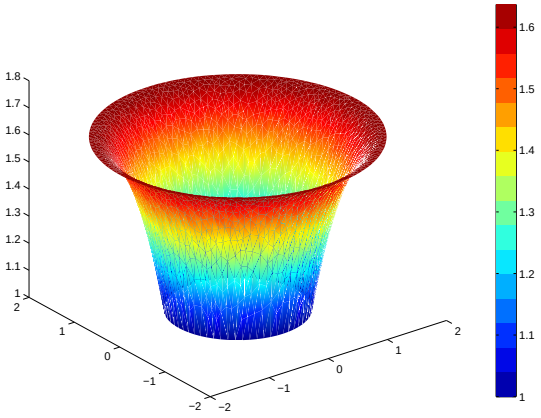


Abbildung 6.4: Lösung des gemischten RWP aus Beispiel 6.19

6.3 Regularität verallgemeinerter Lösungen

Regularitätsaussagen für verallgemeinerte Lösungen elliptischer Randwertprobleme sind später wesentlich für Konvergenzabschätzungen für numerische Lösungsverfahren. Entsprechende Resultate sind vorwiegend technischer Natur und werden wegen des Zieles der Vorlesung hier nur zitiert. Wir beschränken uns hier vereinfachend auf das homogene Dirichlet–Problem

$$Lu = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + \sum_{j=1}^{n} b_j(x)\frac{\partial u}{\partial x_j} + c(x)u \quad = \quad f(x),\ x \in \Omega \tag{6.24}$$

$$u(x) \quad = \quad 0,\ x \in \partial\Omega. \tag{6.25}$$

(i) Existenz verallgemeinerter zweiter Ableitungen

Wir nehmen an, das Problem (6.24),(6.25) besitzt eine verallgemeinerte Lösung $u \in W_0^{1,2}(\Omega)$. Ferner gelte für die Daten

$$a_{ij}, b_j, c \in L^\infty(\Omega), \quad i,j = 1,...,n; \quad f \in L^2(\Omega).$$

Formal erhält man

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) = f(x) - \sum_{j=1}^{n} b_j(x)\frac{\partial u}{\partial x_j} - c(x)u \in L^2(\Omega).$$

Das folgende Beispiel zeigt, daß die Lösung im allgemeinen Fall nicht in $W^{2,2}(\Omega)$ liegt.

Beispiel 6.20. Die Funktion

$$u(x_1, x_2) \equiv u(r, \phi) := r^\beta \sin(\beta\phi)$$

mit $\beta := \pi/\phi_0$ genügt einem inhomogenen Randwertproblem der Laplace–Gleichung im Kreissektor Ω mit $0 \le r < R, 0 < \phi < \phi_0$. Man rechnet nach (zur Übung empfohlen), daß $u \in W^{2,2}(\Omega)$ nur für $0 < \phi_0 \le \pi$ gilt. Man beachte, daß für $\pi < \phi_0 \le 2\pi$ das Gebiet Ω nicht konvex ist. □

Wir betrachten jetzt Kriterien für die Existenz zweiter verallgemeinerter Ableitungen.

Satz 6.21. *Zusätzlich zu den Voraussetzungen des Existenzsatzes 6.10 für verallgemeinerte Lösungen $u \in W_0^{1,2}(\Omega)$ von (6.24),(6.25) gelte $a_{ij} \in W^{1,\infty}(\Omega), i,j = 1,...,n.$*

(i) Für beliebige Teilgebiete $G \subset\subset \Omega$ gilt

$$u \in W^{2,2}(G) \qquad \textit{(innere Regularität).}$$

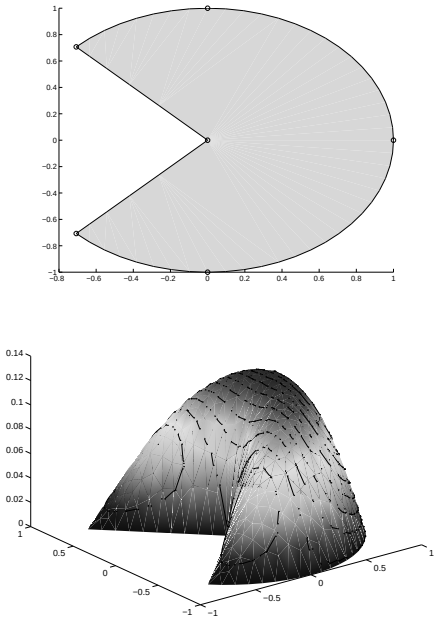


Abbildung 6.5: Lösung des 1. RWP der Poisson-Gleichung für $f \equiv 1$ in einem Kreissektor

(ii) Sei $\Sigma \subset \partial\Omega$ glatt im Sinne von $\Sigma \in C^2$. Dann gilt $u \in W^{2,2}(G)$ in beliebigen Teilgebieten

$$G = \{x \in \Omega : \ dist(x, \partial\Omega \setminus \Sigma) > \delta > 0\}.$$

Im Spezialfall $\partial\Omega \in C^2$ gilt dann sogar $u \in W^{2,2}(\Omega)$.

Beweis: vgl. O.A. Ladyshenskaja, N.N. Uralceva [15] oder z.T. in Gilbarg/ Trudinger [7], Theorem 8.8 □

Bemerkung 6.22. (i) Entscheidender Punkt im Beweis ist der Nachweis von

$$\|u\|_{W^{2,2}(\Omega)} \leq C \left(\|Lu\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} \right). \tag{6.26}$$

(ii) Satz 6.21 sagt aus, daß für beschränkte Gebiete Ω mit Lipschitz-stetigem, jedoch stückweisem C^2-Rand (z.B. Polyeder) Singularitäten (der Ableitungen) der Lösung von (6.24),(6.25) "lokale" Erscheinungen sind, d.h. z.B. auf eine Umgebung von Ecken und Kanten eines Polyeders beschränkt bleiben. Abbildung 6.5 zeigt die Lösung des 1. RWP der Poisson-Gleichung für $f \equiv 1$ in einem Kreissektor Ω mit einspringender Ecke. Man vergleiche hierzu auch das analytische Resultat aus Beispiel 6.20. □

Für konvexe Gebiete kann man die Aussage verschärfen.

Satz 6.23. *Zusätzlich zu den Voraussetzungen des Satzes 6.21 sei das Gebiet Ω konvex sowie $a_{ij} \in C^{0;1}(\overline{\Omega}), i, j = 1, ..., n$. Dann liegt die verallgemeinerte Lösung $u \in W_0^{1,2}(\Omega)$ sogar in $W^{2,2}(\Omega)$ und erfüllt die Ungleichung (6.26).*

Beweis: vgl. Kadlec [13]. □

(ii) Regularitätsaussagen über Einbettungsaussagen

Definition 6.24. *Für normierte Räume X, Y heißt die Einbettung $X \subset Y$* stetig, *falls eine Konstante $C_e > 0$ existiert mit $\|v\|_Y \leq C_e \|v\|_X$ für alle Funktionen $v \in X$. Die Einbettung heißt darüber hinaus* kompakt, *wenn der Einbettungsoperator $I \in \mathcal{L}(X, Y)$ mit $Iv = v$ für alle $v \in X$ kompakt ist.*

Wir zitieren zunächst den folgenden grundlegenden *Einbettungssatz von Sobolev.*

Satz 6.25. *Seien $\Omega \subset \mathbf{R}^n$ beschränktes Gebiet mit $\partial\Omega \in C^{0;1}$ sowie $p \geq 1$ und $k \in \mathbf{N}$. Dann gilt*

- *Falls $kp < n$, so gilt die stetige Einbettung $W^{k,p}(\Omega) \subset L^q(\Omega)$ für alle $q \in \left[1; \frac{np}{n-kp}\right]$, d.h. mit*

$$\|u\|_{L^q(\Omega)} \leq C\|u\|_{W^{k,p}(\Omega)}, \quad \forall u \in W^{k,p}(\Omega).$$

- *Falls $kp = n$, so gilt die stetige Einbettung $W^{k,p}(\Omega) \subset L^q(\Omega)$ für alle $q \in [1; \infty)$, d.h.*

$$\|u\|_{L^q(\Omega)} \leq C\|u\|_{W^{k,p}(\Omega)}, \quad \forall u \in W^{k,p}(\Omega).$$

- *Falls $kp > n$, so gilt die stetige Einbettung $W^{k,p}(\Omega) \subset C(\overline{\Omega})$, d.h. (ggf. nach Abänderung von u auf einer Menge vom Maß 0) ist*

$$\|u\|_{C(\overline{\Omega})} \leq C\|u\|_{W^{k,p}(\Omega)}, \quad \forall u \in W^{k,p}(\Omega).$$

Beweis: vgl. Alt [1], Satz 8.7 sowie Übungsaufgabe für den Fall $n = 1$ □

Wir wenden nun Satz 6.25 an, um exemplarisch Regularitätsaussagen für verallgemeinerte Lösungen $u \in W^{k,2}(\Omega)$ mit $k = 1$ bzw. $k = 2$ zu gewinnen.

Satz 6.26. *(i) Sei $u \in W^{1,2}(\Omega)$ verallgemeinerte Lösung von (6.24),(6.25). Dann gilt in Abhängigkeit von der Raumdimension n*

$$u \in \begin{cases} C(\overline{\Omega}) , & \text{falls} \quad n = 1 \\ L^q(\Omega), \ q \in [1, \infty) & \text{falls} \quad n = 2 \\ L^q(\Omega), \ q \in \left[1, \frac{2n}{n-2}\right] & \text{falls} \quad n \geq 3 \end{cases} .$$

(ii) Sei jetzt sogar $u \in W^{2,2}(\Omega)$. Dann gilt in Abhängigkeit von der Raumdimension n für alle ersten Ableitungen $D^\alpha u$ mit $|\alpha| = 1$

$$D^\alpha u \in \begin{cases} C(\overline{\Omega}) , & \text{falls} \quad n = 1 \\ L^q(\Omega), \ q \in [1, \infty) & \text{falls} \quad n = 2 \\ L^q(\Omega), \ q \in \left[1, \frac{2n}{n-2}\right] & \text{falls} \quad n \geq 3 \end{cases} .$$

Ferner ist $u \in C(\overline{\Omega})$ für $n \leq 3$.

Beweis: Anwendung von Satz 6.25 mit $kp = 2$ für (i),(ii) sowie $kp = 4$ für die letzte Aussage von (ii). □

Wir fügen an dieser Stelle noch Einbettungskriterien für Sobolev-Räume nach Rellich bzw. Sobolev zusammen, die wir später noch benutzen werden.

Satz 6.27 *Sei $\Omega \subset \mathbf{R}^n$ eine offene und beschränkte Punktmenge.*
(i) Dann ist die Einbettung $W_0^{k,2}(\Omega) \subset W_0^{l,2}(\Omega)$ für $k > l$, $k, l \in \mathbf{N_0}$ kompakt.

(ii) Ist außerdem der Rand Lipschitz-stetig, so ist auch die Einbettung $W^{k,2}(\Omega) \subset W^{l,2}(\Omega)$ für $k > l$, $k, l \in \mathbf{N_0}$ kompakt.

Beweis: vgl. H.W. Alt [1], Satz 8.7 □

(iii) Regularität in Sobolev-Slobodeckij-Räumen

Wir hatten gesehen, daß verallgemeinerte Lösungen $u \in W^{1,2}(\Omega)$ des RWP (6.24)-(6.25) nicht zwingend im Raum $W^{2,2}(\Omega)$ liegen. Man kann jedoch - ähnlich wie bei Hölder-Räumen (vgl. Kap. 4) - die Skala der Sobolev-Räume $W^{k,p}(\Omega)$ durch Übergang zu $s = k + \lambda \in \mathbf{R}^+$ verfeinern.

Sei $s = k + \lambda$ mit $k \in \mathbf{N}_0$ und $0 < \lambda < 1$. Ferner bezeichne $\| \cdot \|_{W^{k,p}(\Omega)}$ die in Kapitel 5 eingeführte Sobolev-Norm. Wir setzen für $1 \leq p < \infty$

$$\begin{aligned} \|v\|_{W^{s,p}(\Omega)} &:= \left(\|v\|^p_{W^{k,p}(\Omega)} + |v|^p_{W^{s,p}(\Omega)} \right)^{1/p} \\ |v|_{W^{s,p}(\Omega)} &:= \left(\sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{n+\lambda p}} \, dx \, dy \right)^{1/p} . \end{aligned}$$

Der Fall $p = \infty$ wird wie üblich modifiziert.

Definition 6.28. *Sei $\Omega \subset \mathbf{R}^n$ ein Gebiet, $1 \leq p < \infty$ sowie $s = k + \lambda$ mit $k \in \mathbf{N}_0$ und $0 < \lambda < 1$. Dann wird der* Sobolev-Slobodeckij-Raum $W^{s,p}(\Omega)$ *definiert durch*

$$W^{s,p}(\Omega) := \{v \in W^{k,p}(\Omega) : \ \|v\|_{W^{s,p}(\Omega)} < \infty\}.$$

Es gelten folgende Aussagen:

- Der Raum $W^{s,p}(\Omega)$ ist vollständig bezüglich der Norm $\|\cdot\|_{W^{s,p}(\Omega)}$.
- Es gilt (in Verallgemeinerung von Bemerkung 5.17): $C^\infty(\Omega) \cap W^{s,p}(\Omega)$ ist dicht in $W^{s,p}(\Omega)$.
- Ist Ω beschränktes Gebiet mit Lipschitz-stetigem Rand, so kann der Einbettungssatz 6.25 verallgemeinert werden: $W^{s,p}(\Omega)$ ist stetig in $C(\overline{\Omega})$ eingebettet, falls $s > \frac{n}{p}$.

Schließlich gilt noch folgendes Regularitätsresultat von Necas.

Satz 6.29. *Zusätzlich zu den Voraussetzungen des Existenzsatzes 6.10 gelte für die Daten des RWP (6.24)-(6.25), daß $a_{ij} \in C^{0;t}(\overline{\Omega})$, $i, j = 1, ...,$ für geeignetes $t \in (0, \frac{1}{2}]$. Dann liegt die Lösung des RWP in $W^{1+s,2}(\Omega)$ für $0 \leq s < t$.*

Teil II

Finite-Elemente-Methoden

Kapitel 7

Konforme Approximation elliptischer Variationsgleichungen

Teil II der Vorlesung ist der numerischen Approximation elliptischer Variationsgleichungen in endlich-dimensionalen Unterräumen gewidmet. Im vorliegenden Abschnitt behandeln wir *konforme* Näherungsverfahren zur approximativen Lösung elliptischer Variationsgleichungen in endlich-dimensionalen Unterräumen $X_n \subset X$ des Lösungsraumes X. Die Darstellung ist dabei zunächst *abstrakt* gehalten. In den nachfolgenden Kapiteln betrachten wir speziell *Finite-Elemente-Methoden* (FEM) für elliptische Randwertaufgaben 2. Ordnung.

7.1 Ritz-Galerkin Verfahren

Ausgangspunkt ist die Variationsgleichung

$$\text{Finde } u \in X: \quad a(u,v) = f(v) \quad \forall v \in X \tag{7.1}$$

im Hilbert-Raum X. Dabei verwenden wir die im Abschnitt 6.1 eingeführten Bezeichnungen und Voraussetzungen an die Bilinearform $a(\cdot,\cdot)$ sowie die Linearform $f(\cdot)$.

Gesucht ist nun eine Näherung u^n an die Lösung u von (7.1) im endlich–dimensionalen Teilraum $X_n \subset X$ mit $\dim X_n = n < \infty$. Dann ist $\{X_n; \|\cdot\|_X\}$ Banach–Raum.

Definition 7.1. *Die Aufgabe*

$$\textit{Finde } u^n \in X_n: \quad a(u^n,v) = f(v) \quad \forall v \in X_n \tag{7.2}$$

heißt Ritz–Galerkin Verfahren zur Variationsgleichung (7.1).

Wir zeigen zunächst, daß das Ritz-Galerkin Verfahren stets einem linearen Gleichungssystem entspricht: Sei $\{\phi_i\}_{i=1}^n$ eine Basis von X_n. Es bezeichne $P: \mathbf{R}^n \to X_n \subset X$ die durch

$$P\underline{v} = \sum_{i=1}^n v_i \phi_i, \qquad \underline{v} = (v_1, ..., v_n)^*$$

erklärte Abbildung. Offensichtlich ist P ein Isomorphismus zwischen $\mathbf{R}^n$ und X_n. Unter Beachtung der Basisdarstellung in $X_n = \text{span}\{\phi_1, ..., \phi_n\}$ erhält man das

Lemma 7.2. *Das Ritz-Galerkin Verfahren (7.2) ist äquivalent zu dem System der Gleichungen*

$$\text{Finde } u^n \in X_n: \quad a(u^n, \phi_i) = f(\phi_i) \quad i = 1, ..., n. \tag{7.3}$$

Mit den Bezeichnungen

$$\underline{u} = (u_1, ..., u_n)^* \in \mathbf{R}^n, \qquad u^n := P\underline{u};$$

$$A = (A_{ij}) \in \mathbf{R}^{n\times n},\ A_{ij} := a(\phi_j, \phi_i); \qquad \underline{f} = (f_1, ..., f_n)^* \in \mathbf{R}^n,\ f_i := f(\phi_i)$$

formulieren wir

Satz 7.3. *Das Ritz–Galerkin Verfahren (7.2) ist äquivalent zu dem linearen Gleichungssystem*

$$A\underline{u} = \underline{f}. \tag{7.4}$$

Beweis: Nach Lemma 7.2 sind (7.2) und (7.3) äquivalent. Die Behauptung folgt dann mit $u^n = P\underline{u} = \sum_{j=1}^n u_j\phi_j$ aus

$$a(u^n, \phi_i) = \sum_{j=1}^n u_j a(\phi_j, \phi_i) = \sum_{j=1}^n A_{ij}u_j = f(\phi_i) = f_i, \quad i = 1, ..., n \qquad \square$$

Bemerkungen 7.4. (i) Mit dem Skalarprodukt

$$\langle \underline{u}, \underline{v}\rangle := \sum_{i=1}^n u_i v_i$$

im $\mathbf{R}^n$ sowie $u = P\underline{u}$, $v = P\underline{v}$ gilt

$$a(u, v) = \langle A\underline{u}, \underline{v}\rangle, \quad f(v) = \langle \underline{f}, \underline{v}\rangle.$$

(ii) Das lineare Gleichungssystem (7.4) besitzt genau dann eine eindeutig bestimmte Lösung $u^n \in X_n$, wenn die Matrix A regulär ist. $\square$

Folgende Aufgaben sind nun zu lösen:

- Konstruktion geeigneter Unterräume X_n
- Generierung und Lösung des linearen Gleichungssytems
- Ableitung von Fehlerabschätzungen.

7.2 Lösbarkeit des Ritz-Galerkin Problems

Nachfolgend geben wir hinreichende Lösbarkeitsbedingungen für das Ritz-Galerkin Verfahren sowie eine a-priori Abschätzung der Lösung an.

Satz 7.5. *Seien $X_n \subset X$, dim $X_n = n < \infty$ und X Hilbert-Raum. Ferner sei $a(\cdot,\cdot) : X \times X \to \mathbf{R}$ strikt $X-$elliptische, stetige Bilinearform, d.h. gelte*

$$\exists \gamma > 0 : \ \ a(v, v) \geq \gamma \|v\|_X^2 \quad \forall v \in X \tag{7.5}$$

sowie

$$\exists M > 0 : \ \ |a(u, v)| \leq M\|u\|_X\|v\|_X \quad \forall u, v \in X. \tag{7.6}$$

Weiterhin sei $f : X \to \mathbf{R}$ linear und stetig, d.h.

$$\exists K > 0 : \ \ |f(v)| \leq K\|v\|_X \quad \forall v \in X. \tag{7.7}$$

Dann gilt

(i) Die Matrix $A = (a(\phi_j, \phi_i)) \in \mathbf{R}^{n\times n}$ ist regulär. (Daraus folgt die eindeutige Lösbarkeit von (7.4).)

(ii) Für die Lösung $u^n \in X_n$ des Ritz–Galerkin Verfahrens erhält man

$$\|u^n\|_X \leq \frac{K}{\gamma}. \tag{7.8}$$

Beweis: (i) Mit $\underline{u} \neq 0$ folgt wegen der $X-$Elliptizität von $a(\cdot,\cdot)$ die Aussage $P\underline{u} \neq 0$ sowie

$$\langle A\underline{u}, \underline{u}\rangle = a(P\underline{u}, P\underline{u}) \geq \gamma \|P\underline{u}\|_X^2 > 0,$$

d.h. $A\underline{u} \neq 0$.

(ii) Wegen (7.6) und (7.7) gilt mit $u^n = P\underline{u}$

$$\gamma\|P\underline{u}\|_X^2 \leq a(P\underline{u}, P\underline{u}) = f(P\underline{u}) \leq K\|P\underline{u}\|_X,$$

also (7.8). □

Eine gegenüber der Forderung der strikten $X-$Elliptizität abgeschwächte *hinreichende Lösbarkeitsbedingung* des diskreten Problems (7.2) gibt

Satz 7.6. *Sei $X_n \subset X$ Unterraum der Dimension $n < \infty$. Die stetige Bilinearform $a(\cdot,\cdot) : X \times X \to \mathbf{R}$ genüge der* diskreten Babuska-Bedingung

$$\inf_{u\in X_n\setminus\{0\}} \sup_{v\in X_n\setminus\{0\}} \frac{|a(u,v)|}{\|u\|_X\|v\|_X} = \gamma_n > 0. \tag{7.9}$$

Dann exisitiert eine und nur eine Lösung $u^n \in X_n$ des Galerkin-Problems (7.2). Ferner ist

$$\|u^n\|_X \leq \frac{1}{\gamma_n}\|f\|_{X_n^*} \leq \frac{1}{\gamma_n}\|f\|_{X^*}.$$

Beweis: Übungsaufgabe □

Die diskrete Babuska-Bedingung garantiert die Existenz von A^{-1}, nicht jedoch eine gute Kondition $\mathrm{cond}(A) := \|A\| \cdot \|A^{-1}\|$. Wesentlich ist dafür die Wahl des Unterraumes X_n. Ideal ist, wenn die Basis $\{\phi_1, ..., \phi_n\}$ ein Orthonormalsystem ist.

Beispiel 7.7. Für das Variationsproblem

$$\text{Finde } u \in X: \quad a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, dx = f(v) := \int_\Omega fv \, dx$$

zum homogenen Dirichlet-Problem der Poisson-Gleichung in $X := W_0^{1,2}(\Omega)$ besitzt das zugehörige, schwach formulierte Eigenwertproblem

$$a(u,v) = \lambda \int_\Omega uv \, dx$$

nur reelle Eigenwerte endlicher Vielfachheit mit

$$0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_k \leq ... \to \infty, \quad k \to \infty.$$

Die zugehörigen Eigenfunktionen ϕ_i bilden ein Orthonormalsystem in X. Man erhält dann mit der Wahl $X_n = \mathrm{span}\{\phi_1, ..., \phi_n\}$ die Eigenschaft $a(\phi_j, \phi_i) = \delta_{ij}$ und damit $\mathrm{cond}(A) = 1$.

Entscheidender Nachteil dieses Zugangs ist, daß die Bestimmung der Eigenwerte und Eigenfunktionen nur für spezielle Gebiete über die Methode des Separationsansatzes möglich ist (vgl. auch Kap. 3). Ferner ist der Träger der Eigenfunktionen in Ω in der Regel nicht klein. □

7.3 Fehlerabschätzungen in der $X-$Norm

Eine Abschätzung zwischen den Lösungen $u \in X$ der Variationsgleichung (7.1) und $u^n \in X_n$ des Ritz-Galerkin Verfahrens (7.2) liefert der

Satz 7.8. *Seien $X_n \subset X$, dim $X_n = n < \infty$, X Hilbert-Raum und sei die diskrete Babuska-Bedingung (7.9) für die stetige Bilinearform $a(\cdot,\cdot) : X \times X \to \mathbf{R}$ erfüllt. Dann folgt*

$$\|u - u^n\|_X \leq \left(1 + \frac{M}{\gamma_n}\right) \inf_{v \in X_n} \|u - v\|_X. \tag{7.10}$$

Beweis: Aus (7.1) und (7.2) folgern wir zunächst die sogenannte *Fehlergleichung* (oder Galerkin-Orthogonalität)

$$a(u - u^n, w) = a(u, w) - a(u^n, w) = 0 \quad \forall w \in X_n. \tag{7.11}$$

Damit ist

$$a(u^n - v, w) = a(u^n - u, w) + a(u - v, w) = a(u - v, w)$$

für alle $w, v \in X_n$. Unter Beachtung von (7.6) und (7.9) ergibt sich dann als Abschätzung nach oben

$$|a(u^n - v, w)| = |a(u - v, w)| \leq M\|u - v\|_X \ \|w\|_X$$

bzw. nach unten

$$\|u^n - v\|_X \leq \frac{1}{\gamma_n} \sup_{w \in X_n \setminus 0} \frac{|a(u^n - v, w)|}{\|w\|_X} \leq \frac{M}{\gamma_n}\|u - v\|_X.$$

Die Dreiecksungleichung liefert dann

$$\|u - u^n\|_X \leq \|u - v\|_X + \|u^n - v\|_X \leq \left(1 + \frac{M}{\gamma_n}\right) \|u - v\|_X$$

und wegen $v \in X_n$ beliebig die Behauptung. □

Bei $X-$Elliptizität der Bilinearform a läßt sich die Abschätzung (7.10) verschärfen.

Satz 7.9. *Unter den Voraussetzungen von Satz 7.8 sei zusätzlich die Bilinearform $a(\cdot,\cdot) : X \times X \to \mathbf{R}$ strikt $X-$elliptisch. Dann folgt*

$$\|u - u^n\|_X \leq \frac{M}{\gamma} \inf_{v \in X_n} \|u - v\|_X. \tag{7.12}$$

Beweis: Zur Übung empfohlen ! □

Mit den Sätzen 7.8 und 7.9 ist die Fehlerabschätzung auf eine Abschätzung des *Interpolationsfehlers* zurückgeführt. Auf Details dieser Interpolationstheorie in Sobolev-Räumen gehen wir in einem der nachfolgenden Kapitel ausführlich ein. Es gilt zumindest

Lemma 7.10. *Seien*

$$X_1 \subset ... \subset X_{n-1} \subset X_n \subset ... \subset X$$

sowie $X = \overline{\cup_{n=1}^{\infty} X_n}$. Dann ist

$$\lim_{n \to \infty} \inf_{w \in X_n} \|u - w\|_X = 0. \tag{7.13}$$

Beweis: Folgerung aus Dichtheit von $\cup_{n=1}^{\infty} X_n$ in X. □

7.4 Fehlerabschätzungen in der $H-$Norm

Im Fall einer *stetigen Einbettung* $X \subset H$ der Hilbert-Räume X und H, d.h. es gibt eine Konstante $C > 0$ mit

$$\|u\|_H \leq C\|u\|_X, \quad \forall u \in X,$$

kann man mittels eines *Dualitätsargumentes* auch eine Fehlerabschätzung in der $H-$Norm gewinnen. Dazu benötigen wir den Begriff der zu einer Bilinearform $a : X \times X \to \mathbf{R}$ adjungierten Bilinearform $a^* : X \times X \to \mathbf{R}$ gemäß

$$a^*(u, v) = a(v, u) \qquad \forall u.v \in X.$$

Man erwartet nun eine verbesserte Fehlerabschätzung gegenüber der Abschätzung in der $X-$Norm.

Satz 7.11. *Zusätzlich zu den Voraussetzungen von Satz 7.8 sei $X \subset H$ mit stetiger Einbettung der Hilbert-Räume X und H. Ferner besitze das adjungierte Variationsproblem*

$$a^*(w_g, \psi) = (g, \psi)_H \quad \forall \psi \in X \tag{7.14}$$

für beliebige $g \in H$ eine und nur eine Lösung $w_g \in X$. Dann gilt

$$\|u - u^n\|_H \leq M\|u - u^n\|_X \cdot \sup_{\substack{g \in H \\ \|g\|_H = 1}} \left(\inf_{\phi \in X_n} \|w_g - \phi\|_X \right). \tag{7.15}$$

Beweis: Die Schwarzsche Ungleichung liefert

$$\|u - u^n\|_H = \sup_{\|g\|_H = 1} |(u - u^n, g)_H|.$$

Ferner ist wegen $u - u^n \in X$ nach (7.14)

$$a^*(w_g, u^n - u) = (g, u^n - u)_H.$$

Zusammen mit der Fehlergleichung (7.11) ergibt sich dann

$$a(u^n - u, w_g - \phi) = a(u^n - u, w_g) - a(u^n - u, \phi) = a^*(w_g, u^n - u) = (g, u^n - u)_H$$

für alle $\phi \in X_n$. Über die Stetigkeit von a folgt dann die Behauptung weiter aus

$$|(u^n - u, g)_H| \leq M\|u^n - u\|_X \, \|w_g - \phi\|_X, \quad \forall \phi \in X_n. \qquad \square$$

7.5 Fall Gardingscher Formen

Zur Information für den an Erweiterungen interessierten Leser wollen wir hier kurz darstellen, wie man die Voraussetzung der strikten $X-$Elliptizität der Bilinearform $a(\cdot, \cdot)$ abschwächen kann:

Die Lösbarkeit der elliptischen Variationsgleichung (7.1) kann auch noch in der folgenden Situation bewiesen werden mittels des *Fredholmschen Alternativsatzes* für kompakte Operatoren (vgl. Kap. 19, 20 der Vorlesung LFA 05/06). Dabei heißt ein linearer Operator $K : X \to Y$ mit normierten Räumen $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y$ *kompakt*, wenn für jede beschränkte Folge $(u_n)_{n \in \mathbf{N}}$ in X die Bildfolge $(Ku_n)_{n \in \mathbf{N}}$ eine konvergente Teilfolge in Y enthält.

Seien $X \subset H$ Hilbert-Räume mit

(i) *stetiger Einbettung*, d.h. es gibt eine Konstante $C > 0$ mit $\|u\|_H \leq C\|u\|_X, \quad \forall u \in X$ sowie

(ii) *dichter Einbettung*, d.h. für alle $u \in H$ und jedes $\epsilon > 0$ gibt es ein Element $w \in X$ mit $\|u - w\|_H < \epsilon$.

Wir zitieren (vgl. Lemma 20.1 der o.a. Vorlesung vom WS 05/06)

Lemma 7.12. *Für Hilbert-Räume $X \subset H$ mit dichter und stetiger Einbettung ist auch die Einbettung*

$H^* \subset X^*$ der Dualräume dicht und stetig.

Nach dem Satz von Riesz können die Räume H und sein Dualraum H^* identifiziert werden. Man sagt, daß X und H ein *Evolutionstripel* $X \subset H \subset X^*$ bilden. Bei der Identifizierung von H mit H^* und von H^* mit einem Teilraum von X^* wird ein Element $v \in H$ mit dem Element $f_v \in X^*$ identifiziert, für das

$$(v,u)_H = \langle f_v, u\rangle, \quad u \in H$$

gilt. Dabei ist $\langle\cdot,\cdot\rangle$ das Dualitätsprodukt zwischen X^* und X. Wegen der Identifizierung führt es nicht zu Mißverständnissen, wenn dieses Dualitätsprodukt ebenso wie das Skalarprodukt auf H mit $(\cdot,\cdot)_H$ bezeichnet wird.

Definition 7.13. *Eine stetige Bilinearform $a(\cdot,\cdot) : X \times X \to \mathbf{R}$ ist eine* Gardingsche Form, *falls es Konstanten $\gamma > 0$ und $\delta \in \mathbf{R}$ gibt mit*

$$a(u,u) \geq \gamma\|u\|_X^2 - \delta\|u\|_H^2, \quad \forall u \in X. \tag{7.16}$$

Dann gilt

Satz 7.14. *Für das Evolutionstripel $X \subset H \subset X^*$ sei die Einbettung $X \subset H$ kompakt, d.h. der Einbettungsoperator $I_H \in \mathcal{L}(X,H)$ mit $I_H u = u$ für alle $u \in X$ ist kompakt. Ferner sei die stetige Bilinearform $a(\cdot,\cdot) : X \times X \to \mathbf{R}$ Gardingsche Form. Dann besitzt die Variationsgleichung (7.1) eine und genau eine Lösung $u \in X$, wenn $\lambda = 0$ kein Eigenwert des zu $a(\cdot,\cdot)$ gehörenden Operators $A \in \mathcal{L}(X,X^*)$ ist.*

Beweis: vgl. Folgerung 20.10 aus Vorlesung *LFA*, WS 05/06 □

Folgendes Beispiel zeigt, daß die Garding-Eigenschaft von $a(\cdot,\cdot)$ im allgemeinen Fall nicht die Lösbarkeit des Galerkin-Problems (7.2) sichert.

Beispiel 7.15. Die Variationsformulierung der Zweipunkt-RWA

$$-u''(x) - 10u(x) = f(x), \;\; 0 < x < 1; \quad u(0) = u(1) = 0$$

lautet mit $X := W_0^{1,2}(0,1)$ und $H := L^2(0,1)$

$$\text{Finde } u \in X: \quad a(u,v) = f(v), \quad \forall v \in X$$

mit $a(u,v) = \int_0^1 (u'v' - 10uv)\, dx$. Offenbar ist a Gardingsche Form mit $\gamma = 1$ und $\delta = 10$. Die Eigenwerte des homogenen Dirichlet-Problems

$$-\phi''(x) = \lambda\phi(x), \quad x \in (0,1), \qquad \phi(0) = \phi(1) = 0$$

sind $\lambda_k = k\pi, k \in \mathbf{N}$. Damit besitzt das obige Variationsproblem eine und nur eine Lösung.
Sei nun

$$X_1 = \text{span}\{\phi_1\} \subset X, \quad \phi_1(x) := x(1-x).$$

Dann gilt

$$A = a(\phi_1,\phi_1) = \int_0^1 \left[(1-2x)^2 - 10\, x^2(1-x)^2\right]\, dx = 0,$$

damit ist das Galerkin-Problem im Fall $f \not\equiv 0$ nicht lösbar. □

Man kann jedoch die Lösbarkeit des diskreten Problems bei hinreichend großer Dimension des Lösungsraumes X_n zeigen. Dies wird durch folgendes Resultat von A. Schatz präziser ausgedrückt

Satz 7.16 *Zusätzlich zu den Voraussetzungen von Satz 7.8 sei $\lambda = 0$ kein Eigenwert des zur Bilinearform $a(\cdot,\cdot)$ gehörenden Operators $A \in \mathcal{L}(X,X^*)$. Ferner sei $(X_n)_{n\in\mathbf{N}}$ eine Familie endlich-dimensionaler Unterräume von X, so daß für jede Lösung $u^n \in X_n$ von (7.2) eine Abschätzung der Form*

$$\|u - u^n\|_H \leq w(n)\|u - u^n\|_X, \qquad \lim_{n\to\infty} w(n) = 0$$

gilt. Dann gibt es eine Konstante $n_0 = n_0(u,\gamma,M,w)$, so daß für alle $n \geq n_0$ jeweils eine eindeutig bestimmte Lösung $u^n \in X_n$ des Problems (7.2) existiert. Für sie gilt die Abschätzung

$$\|u - u^n\|_X \leq \frac{2M}{\gamma} \inf_{v\in X_n} \|u - v\|_X.$$

Kapitel 8

Konforme Finite-Elemente-Räume für elliptische RWP

Die *Finite-Elemente-Methode* (FEM) ist eine spezielle Variante von Ritz-Galerkin-Verfahren (vgl. Kapitel 7), die sich besonders zur Behandlung elliptischer RWP in Gebieten mit komplizierterer Geometrie eignet. Sie zeichnet sich durch eine stückweise Definition der Basis- und Testfunktionen mit *kleinem Träger* aus. In der Regel wählt man stückweise polynomiale Funktionen. Man erhält dann bei Anwendung auf elliptische RWP lineare Gleichungssysteme mit *schwachbesetzten Matrizen*, für die angepaßte Speicher- und Lösungstechniken zur Verfügung stehen.

Genauer ist die FEM durch folgende *Merkmale* gekennzeichnet:

- Zerlegung des Lösungsgebietes in geometrisch einfache Teilgebiete,
- Lokale Definition von Ansatz- und Testfunktionen über Teilgebieten,
- Sicherung globaler Eigenschaften (z.B. Konformität der Methode) durch Einhaltung von Übergangsbedingungen bei den Ansatz- und Testfunktionen.

8.1 Zulässige Zerlegungen polyedrischer Gebiete

Vereinfachend sei $\Omega \subset \mathbf{R}^n$ mit $n \leq 3$ ein beschränktes polyedrisches Gebiet. Wir betrachten eine nichtüberlappende Zerlegung $\mathcal{T}_h = \{K_i\}_{i=1}^M$ des Gebietes in konvexe polyedrische Teilgebiete K_i mit

$$\overline{\Omega} = \cup_{j=1}^M \overline{K}_j, \quad K_i \cap K_j = \emptyset,\ i \neq j; \quad h_i := \mathrm{diam}(K_i), \quad h := \max_{i=1,\ldots,M} h_i. \tag{8.1}$$

Im eindimensionalen Fall $n = 1$ zerlegen wir das Intervall $\Omega = (a, b)$ mit Hilfe des Gitters

$$a = x_0 < x_1 < x_2 < ... < x_{M-1} < x_M = b \tag{8.2}$$

in Teilgebiete $K_i := (x_{i-1}, x_i),\ i = 1, ..., M$ mit $h_i := x_i - x_{i-1}$.

Für Gebiete $\Omega \subset \mathbf{R}^n$ mit $n = 2$ oder $n = 3$ ist eine Zerlegung in Teilgebiete nicht mehr offensichtlich. Später seien K_j für $n = 2$ Dreiecke oder konvexe Vierecke und für $n = 3$ Tetraeder oder konvexe Hyperquader. Wir fordern die *Zulässigkeit* der Zerlegung in Teilgebiete gemäß

Definition 8.1. *Eine Zerlegung $\mathcal{T}_h = \{K_i\}_{i=1}^M$ des Gebietes Ω heißt* zulässig, *falls jeweils zwei verschiedene abgeschlossene Teilgebiete $\overline{K}_j$ und $\overline{K}_i$ entweder*

- *genau eine vollständige gemeinsame Fläche (nur für $n = 3$),*
- *genau eine vollständige gemeinsame Kante (für $n \geq 2$),*

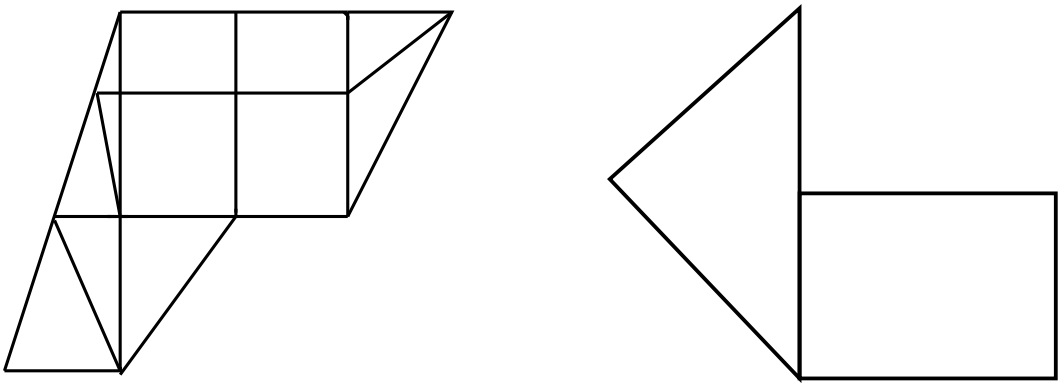


Abbildung 8.1: Zulässige und unzulässige Zerlegungen

- *genau einen Punkt (für $n \geq 1$) gemeinsam haben*
- *oder paarweise durchschnittsfremd sind.*

Beispiele einer zulässigen bzw. unzulässigen Zerlegung sind in der Abbildung 8.1 angegeben. Insbesondere sind per Definition sogenannte *"hängende Knoten"* nicht erlaubt. □

8.2 Finite Elemente

Finite Elemente lassen sich charakterisieren über die Geometrie der Teilgebiete $K \in \mathcal{T}_h$ sowie durch Anzahl, Lage und Art der Vorgabe der Ansatzfunktionen.

Definition 8.2. *Ein* finites Element *ist ein Tripel $(K, \mathcal{P}, \Sigma)$ mit den Eigenschaften*

- *$K \subset \mathbf{R}^n$ ist ein konvexes polyedrisches Gebiet im $\mathbf{R}^n$. Die Teile des Randes ∂K, die auf einer Hyperfläche liegen, heißen* Seiten.
- *Der* Raum der Formfunktionen *$\mathcal{P}$ ist ein auf K definierter endlichdimensionaler linearer Funktionenraum der Dimension d.*
- *Die* Menge der Freiheitsgrade *Σ besteht aus d linear unabhängigen Funktionalen über $\mathcal{P}$. Dabei ist jede Funktion $p \in \mathcal{P}$ durch die Werte der d Funktionale aus Σ eindeutig bestimmt.*

Benutzt man lediglich vorgegebene Funktionswerte als Freiheitsgrade, so heißen die zugehörigen finiten Elemente auch *Lagrange-Elemente.* Bei *Hermite-Elementen* verwendet man auch Ableitungswerte zur Bestimmung der Ansatzfunktionen.

Man kann auch folgende Charakterisierung vornehmen: Sei $\mathcal{P}'$ der algebraische Dualraum des linearen Raumes $\mathcal{P}$. Dann kann man eine Basis $\{N_1, ..., N_d\}$ von $\mathcal{P}'$, die Menge der *Knotenvariablen*, mit Σ identifizieren.

Definition 8.3. *Sei $(K, \mathcal{P}, \Sigma)$ finites Element. Eine Basis $\{\phi_1, ..., \phi_d\}$ von $\mathcal{P}$ mit $N_i(\phi_j) = \delta_{ij}$, $1 \leq i, j \leq d$ heißt* nodale Basis *von $\mathcal{P}$.*

Wir illustrieren die Definitionen am einfachsten Beispiel.

Beispiel 8.4. *(Eindimensionales lineares Lagrange-Element)*
Seien $K = (0, 1)$, $\mathcal{P}$ die Menge der linearen Polynome sowie $\Sigma = \{N_1, N_2\}$ mit $N_1(v) = v(0)$ sowie $N_2(v) = v(1)$ für alle $v \in \mathcal{P}$. Dann ist $(K, \mathcal{P}, \Sigma)$ finites Element mit der nodalen Basis $\{\phi_1, \phi_2\}$ mit $\phi_1(x) = 1 - x$ und $\phi_2(x) = x$. □

Zur Charakterisierung der Basis im Raum $\mathcal{P}'$ nutzt man oft folgende Aussagen.

Lemma 8.5. *Seien $\mathcal{P}$ ein d-dimensionaler linearer Raum von Funktionen auf $\mathbf{R}^n$ sowie $\{N_1, ..., N_d\}$ eine Teilmenge des Dualraumes $\mathcal{P}'$. Dann sind folgende Aussagen äquivalent:*

(i) $\{N_1, ..., N_d\}$ *ist Basis von* $\mathcal{P}'$.

(ii) *Gilt* $N_i(v) = 0$, $i = 1, ..., d$ *für* $v \in \mathcal{P}$, *so ist* $v \equiv 0$.

Beweis: *Übungsaufgabe !* □

Definition 8.6. *Man sagt "*Σ *bestimmt* $\mathcal{P}$*", falls für* $v \in \mathcal{P}$ *die Aussage* $N(v) = 0$ *für alle* $N \in \Sigma$ *die Aussage* $v \equiv 0$ *impliziert.*

Lemma 8.7. *Sei* P *ein Polynom vom Grad* $l \geq 1$, *das auf der Hyperebene* $\{x : L(x) = 0\}$ *bei nichtentarteter linearer Funktion* L *verschwindet. Dann kann man* P *faktorisieren gemäß* $P = LQ$, *wobei* Q *ein Polynom vom Grad* $(l-1)$ *ist.*

Beweis: Zur Übung empfohlen ! □

8.3 Lokale und globale Interpolation

Eine geeignete lokale Beschreibung von Ansatz- und Testfunktionen eines Finite-Elemente-Raums über einem Teilgebiet ist durch *lokale Interpolation* möglich. Sie spielt auch bei der Vorbereitung von FEM-Fehlerabschätzungen in Kapitel 10 eine wesentliche Rolle.

Definition 8.8. *Seien* $(K, \mathcal{P}, \Sigma)$ *ein finites Element und* $\{\phi_1, ...\phi_d\}$ *eine nodale Basis von* $\mathcal{P}$. *Sei* v *eine Funktion, für die alle* $N_i \in \Sigma$, $i = 1, ..., d$ *definiert sind. Als* lokalen Interpolant von v *bezeichnet man eine Funktion* $\Pi_K v \in \mathcal{P}$ *mit*

$$v \mapsto \Pi_K v := \sum_{i=1}^{d} N_i(v)\phi_i. \tag{8.3}$$

Offenbar gelten folgende Aussagen:

Lemma 8.9. *Für den nach Definition 8.8 erklärten lokalen Interpolant gilt:*

(i) *Die Abbildung* $v \mapsto \Pi_K v$ *ist linear.*

(ii) *Es gilt* $N_i(\Pi_K v) = N_i(v)$, $i = 1, .., d$.

(iii) Π_K *ist ein Projektor, d.h. es gilt* $\Pi_K(v) = v$ *für alle* $v \in \mathcal{P}$.

Beweis: *Übungsaufgabe !* □

Für die in den folgenden Abschnitten diskutierten Beispiele finiter Elemente auf simplizialen Gebieten oder Hyperquadern kann der lokale Interpolant in kanonischer Weise definiert werden.

Bei den hier zunächst betrachteten *konformen* FEM ist bei der stückweisen Definition der Ansatz- und Testfunktionen über den Teilgebieten K_i zu sichern, daß sie zum Lösungsraum X des elliptischen RWP paßfähig sind, d.h. es gilt $X_h \subset X$. Das nachfolgende Lemma gibt wesentliche Bedingungen zur Konstruktion passender Ansatzräume an.

Lemma 8.10. *Sei* $\overline{\Omega} = \cup_{j=1}^{M} \overline{K}_j$ *eine zulässige Zerlegung des Polyeders* $\Omega \subset \mathbf{R}^n$ *gemäß Definition 8.1. Sei* $k \geq 1$ *sowie* $v : \overline{\Omega} \to \mathbf{R}$ *eine Funktion mit* $v|_{K_j} \in C^k(\overline{K}_j)$, $j = 1, ..., M$. *Dann gilt* $v \in W^{k,2}(\Omega)$ *genau für* $v \in C^{k-1}(\overline{\Omega})$.

Beweis: Man beweist nur den Fall $k = 1$, für $k \geq 2$ schließt man induktiv.

(i) Sei $v \in C(\overline{\Omega})$. Die Existenz verallgemeinerter erster Ableitungen folgt nach Beispiel 5.13. Die ersten Ableitungen sind wesentlich beschränkt, gehören also zu $L^\infty(\Omega)$. Damit ist aber auch $v \in W^{1,2}(\Omega)$.

(ii) Beweis der Umkehrung vgl. Braess [5], Kap. II, Satz 5.2. □

Eine wichtige *Schlußfolgerung* für die Approximation elliptischer RWP 2. Ordnung in passenden Unterräumen $X_h \subset W^{1,2}(\Omega)$ ist, daß man lokal definierte Ansatzfunktionen lediglich stetig an den Seiten der Teilgebiete koppeln muß. Hingegen ist die Approximation elliptischer RWP der Ordnung $2m$ mit

$m \geq 2$ wesentlich aufwendiger. So muß man für $m = 2$ schon $X_h \subset C^1(\overline{\Omega})$ sichern.

Zur Diskretisierung von RWP wird angemerkt, daß stetige Fortsetzungen von Funktionen über dem Lösungsgebiet Ω auf den Rand $\partial\Omega$ Spuren im Sinne der Sobolev-Räume besitzen. Speziell gilt für RWP 2. Ordnung, daß

$$v \in C(\overline{\Omega}),\ v|_{K_j} \in C^1(\overline{K}_j),\ j = 1, ..., M; \qquad v|_{\partial\Omega} = 0 \quad \Longrightarrow \quad v \in W_0^{1,2}(\Omega).$$

Wir definieren noch eine geeignete *globale Interpolation* bei finiten Elementen unter Verwendung der lokalen Interpolation.

Definition 8.11. *Sei das beschränkte polyedrische Gebiet $\Omega \subset \mathbf{R}^n$ mittels einer zulässigen Zerlegung $\mathcal{T} = \{K_i\}_{i=1}^m$ exakt trianguliert. Zu jedem Gebiet K_i seien ein finites Element $(K_i, \mathcal{P}_i, \Sigma_i)$ und eine nodale Basis in $\mathcal{P}_i$ erklärt. Ferner sei $k \in \mathbf{N}_0$ die höchste in der nodalen Basis vorkommende Ordnung partieller Ableitungen. Für $v \in C^k(\overline{\Omega})$ wird ein* globaler Interpolant $\Pi_{\mathcal{T}} v$ *definiert durch*

$$\Pi_{\mathcal{T}} v|_{K_i} = \Pi_{K_i} v \qquad \forall K_i \in \mathcal{T}. \tag{8.4}$$

Man beachte, daß diese Definition noch keine Glattheitsaussagen über die globale Interpolationsfunktion impliziert. Für Abschätzungen des Interpolationsfehlers in $X_h \subset X$ (vgl. Kap. 7) ist jedoch zu sichern, daß $\Pi_{\mathcal{T}} v \in X_h$ gilt. Man sagt, daß

$$X_{\mathcal{T}} := \{\Pi_{\mathcal{T}} v :\ v \in C^m(\overline{\Omega})\}$$

ein *C^m-Finite-Elemente-Raum* mit $m \in \mathbf{N}_0$ ist, wenn für alle $v \in C^m(\overline{\Omega})$ auch $\Pi_{\mathcal{T}} v \in C^m(\overline{\Omega})$ folgt.

Für Lagrange-Elemente auf simplizialen Zerlegungen werden wir in Abschnitt 8.5 ein hinreichendes Kriterium für den Fall $m = 0$ angeben.

Die Auswahl bestimmter Elemente im Rahmen eines FEM-Programms ist eine wichtige Entscheidung vor der Implementierung. Folgende Klassifizierung ist eventuell hilfreich.

(i) $h-$Methoden: Elemente mit fixiertem *niedrigen Ansatzgrad* (insbesondere lineare bzw. multilineare Elemente) erlauben (wenigstens derzeit) eine größere Flexibilität hinsichtlich der Entwicklung geeigneter Lösungsverfahren für die diskreten Verfahren (z.B. Mehrgitter- und andere Multilevel-Verfahren) sowie hinsichtlich der Datenmanipulation auf unstrukturierten Gittern (adaptive Netzgenerierung, grafische Aufbereitung usw.). Die geringere Genauigkeit führt andererseits bei komplexen Problemen auf sehr große diskrete Probleme. Höhere Genauigkeit wird durch Verkleinerung der Elementgröße erreicht (*h-Methode*)

(ii) $p-$Methoden: Elemente mit fixiertem, jedoch *höherem Ansatzgrad* haben neben dem Vorteil höherer Genauigkeit vor allem den Nachteil schlechterer algebraischer Eigenschaften. Ferner sind sie im allgemeinen Fall zunächst schwerer handhabbar bei der Datenmanipulation, auf Gebieten mit einfacher Geometrie (z.B. Quadern) erzielt man aber durch Einsatz von numerischen Integrationsformeln erhebliche Vereinfachungen. Höhere Genauigkeit erreicht man durch Erhöhung des Ansatzgrades (*p-Methode).*

(iii) $hp-$Methoden: Die Kombination beider Ansätze führt auf die sogenannten *hp-Methoden*, die zunehmend Anwendung finden. Hier werden die Vorteile der beiden Methoden sinnvoll verbunden, deren Nachteile kommen weniger zum Tragen. Die Implementierung von $hp-$Methoden ist jedoch sehr aufwendig. Eine gute Übersicht zu $p-$ und $hp-$Methoden findet man in der Monographie [21] von C. Schwab.

8.4 Finite-Elemente-Räume im 1D-Fall

Das Gebiet $\Omega = (a, b)$ wird gemäß (8.2) in Teilgebiete $K_i := (x_{i-1}, x_i),\ i = 1, ..., M$ mit Durchmesser $h_i := x_i - x_{i-1}$ zerlegt. Später nutzen wir oft aus, daß jedes Element K_i mittels affin-linearer Transformation

$$x \mapsto \xi = F_i(x) := h_i^{-1}(x - x_{i-1})$$

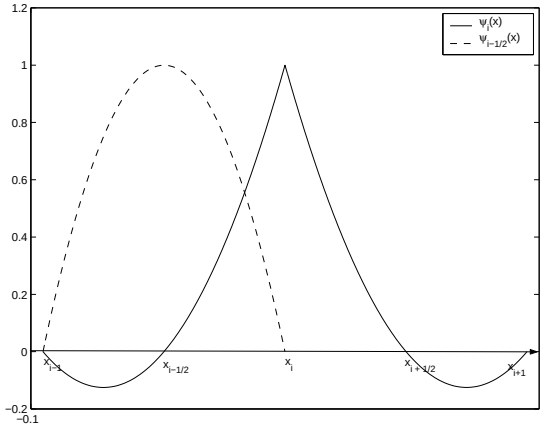


Abbildung 8.2: Quadratische Basisfunktionen

bijektiv auf das sogenannte *Referenzelement* $\tilde{K} = (0,1)$ überführt werden kann.

Wir beschreiben zunächst global stetige Ansatzräume $X_h \subset C^0(\overline{\Omega})$ mit *Lagrange-Elementen*. Lokal wählen wir $\mathcal{P}_i = P_{d-1}(K_i)$, wobei $P_l(K)$ die Menge der Polynome vom Grad $l \in \mathbf{N}$ über K ist. Die d linearen Funktionale aus Σ_i sind gerade die Funktionswerte in den d vorgegebenen Punkten $x_{i-1} + \frac{j}{d-1}h_i,\ j = 0, ..., d-1$ des Intervalls $\overline{K}_i$.

Beispiel 8.12. *Lagrange-Elemente vom P_1-Typ* (vgl. auch Beispiel 8.4)
Bei dem bereits im Abschnitt 1.5 untersuchten stückweise linearem Ansatz betrachten wir den endlichdimensionalen Raum $X_h := \text{span}\{\phi_i\}_{i=0}^M$ mit den stückweise linearen Lagrangeschen Basisfunktionen

$$\phi_i(x) := \begin{cases} \frac{x-x_{i-1}}{h_i}, & x \in K_i \\ \frac{x_{i+1}-x}{h_{i+1}}, & x \in K_{i+1} \\ 0, & \text{sonst} \end{cases} \quad , \quad i = 0, ..., M. \tag{8.5}$$

Jede Funktion $v_h \in X_h$ ist offenbar durch die Knotenwerte $v_i = v_h(x_i)$ eindeutig festgelegt und besitzt die Darstellung $v_h(x) = \sum_{j=0}^M v_j \phi_j(x)$. □

Beispiel 8.13. *Lagrange-Elemente vom P_2-Typ*
Für die Zerlegung (8.2) seien $x_{i-1/2} := \frac{1}{2}(x_{i-1} + x_i)$ die Mittelpunkte der Teilintervalle K_i. Bei stückweise quadratischem Ansatz betrachten wir Funktionen

$$v_h(x) = \sum_{i=0}^{M} v_i \psi_i(x) + \sum_{i=1}^{M} v_{i-1/2} \psi_{i-1/2}(x) \tag{8.6}$$

mit stetigen Funktionen $\psi_i,\ i = 0, ..., M$ und $\psi_{j-1/2},\ j = 1, ..., M$ mit

(i) $\psi_i|_{K_j}, \quad \psi_{i-1/2}|_{K_j} \in P_2; \quad j = 1, ..., M;$

(ii) $\psi_i(x_k) = \delta_{ik},\ \psi_i(x_{k-1/2}) = 0;\ \ \psi_{i-1/2}(x_k) = 0,\ \ \psi_{i-1/2}(x_{k-1/2}) = \delta_{ik}.$

Daraus ergibt sich die lokale Darstellung der Ansatzfunktionen unter Verwendung quadratischer Lagrangescher Funktionen (vgl. auch Abbildung 8.2)

$$\psi_i(x) \quad := \quad \begin{cases} \frac{2}{h_i^2}(x - x_{i-1})(x - x_{i-1/2}), & x \in K_i \\ \frac{2}{h_{i+1}^2}(x_{i+1} - x)(x_{i+1/2} - x), & x \in K_{i+1} \\ 0, & \text{sonst} \end{cases} \tag{8.7}$$

$$\psi_{i-1/2}(x) := \begin{cases} \frac{4}{h_i^2}(x-x_{i-1})(x_i-x), & x \in K_i \\ 0, & \text{sonst} \end{cases} . \tag{8.8}$$

Jede Funktion $v_h \in X_h$ ist durch die Knotenwerte $v_i = v_h(x_i)$ bzw. $v_{i-1/2} = v_h(x_{i-1/2})$ auf dem Gitter $x_0, x_{1/2}, x_1, ..., x_{M-1/2}, x_M$ eindeutig festgelegt. □

Bemerkung 8.14. Man kann auch die folgende Darstellung stückweise quadratischer Funktionen $v_h \in C(\overline{\Omega})$ angeben

$$v_h(x) = \sum_{i=0}^{M} v_i \phi_i(x) + \sum_{i=1}^{M} v_{i-1/2}\psi_{1-1/2}(x) \tag{8.9}$$

unter Verwendung der stückweise linearen Ansatzfunktionen ϕ_i aus Beispiel 8.12. Dies gilt wegen

$$\text{span}\{\psi_i\}_{i=0}^{M} \oplus \text{span}\{\psi_{i-1/2}\}_{i=1}^{M} = \text{span}\{\phi_i\}_{i=0}^{M} \oplus \text{span}\{\psi_{i-1/2}\}_{i=1}^{M}.$$

Die angegebene Darstellung entspricht einer speziellen *p−hierarchischen Basis*, bei der der Raum stückweise quadratischer Funktionen aus dem Raum der stückweise linearen Funktionen über dem Grundgitter durch Hinzunahme der zusätzlichen Basisfunktionen $\{\psi_{i-1/2}\}_{i=1}^{M}$ gewonnen wird. □

Die Konstruktion eines Ansatzraumes $X_h \subset C^1(\overline{\Omega})$, d.h. mit global stetig differenzierbaren Funktionen, ist aufwendiger. Die Menge Σ der stetigen Funktionale umfaßt dann neben Funktionswerten auch bestimmte Ableitungswerte.

Beispiel 8.15. *Kubische Hermite-Polynome*

Zur Erzeugung global stetig differenzierbarer Funktionen verwendet man *Hermite-Polynome*. Seien $\zeta_i,\ \eta_i \in C^1(\overline{\Omega}),\ i = 0, ..., M$ stückweise kubische Funktionen mit

$$\zeta_i(x_k) = \delta_{ik}, \quad \zeta_i'(x_k) = 0; \quad \eta_i(x_k) = 0, \quad \eta_i'(x_k) = \delta_{ik}, \quad k = 1, ..., M.$$

Dann erhält man die lokalen Darstellungen

$$\eta_i(x) := \begin{cases} \frac{1}{h_i^2}(x-x_i)(x-x_{i-1})^2, & x \in K_i \\ \frac{1}{h_{i+1}^2}(x-x_i)(x-x_{i+1})^2, & x \in K_{i+1} \\ 0, & \text{sonst} \end{cases} \tag{8.10}$$

bzw.

$$\zeta_i(x) := \begin{cases} \phi_i(x) - \frac{1}{h_i}\left[\eta_{i-1}(x) + \eta_i(x)\right], & x \in K_i \\ \phi_i(x) + \frac{1}{h_{i+1}}\left[\eta_i(x) + \eta_{i+1}(x)\right], & x \in K_{i+1} \\ 0, & \text{sonst} \end{cases} \tag{8.11}$$

unter Vewendung der stückweise linearen Ansatzfunktionen ϕ_i aus Beispiel 8.12.

In der Darstellung

$$v_h(x) = \sum_{i=0}^{M} v_i \zeta_i(x) + \sum_{i=0}^{M} w_i \eta_i(x) \tag{8.12}$$

bezeichnen die Werte v_i bzw. w_i die Funktions– bzw. Ableitungswerte der Funktion v_h in den zugehörigen Gitterpunkten x_i. Wegen $\zeta_i, \eta_i \in C^1(\overline{\Omega})$ ist

$$\text{span}\{\zeta_i\}_{i=0}^{M} \oplus \text{span}\{\eta_i\}_{i=0}^{M} \subset C^1(\overline{\Omega}). \quad \square$$

8.5 Finite Elemente im mehrdimensionalen Fall

Jetzt charakterisieren wir ausgewählte Realisierungen von finiten Elementen im mehrdimensionalen Fall durch Angabe der Teilgebiete sowie von *Formfunktionen* und deren globale Eigenschaften.

Sei $\mathcal{T}_h = \{K_i\}_{i=1}^M$ eine zulässige Zerlegung des Gebietes Ω in konvexe polyedrische Teilgebiete $K_i \in \mathcal{T}_h$. Das Gebiet Ω läßt sich dann über die Gesamtheit $\{p^j\},\ j = 1, ..., \overline{N}$ der Eckpunkte beschreiben. Jedes Teilgebiet $K \in \mathcal{T}_h$ kann als konvexe Hülle

$$\overline{K} = \ \mathrm{conv}\{p^j\}_{j\in J_K} := \left\{ x = \sum_{j\in J_K} \lambda_j p^j :\ \lambda_j \geq 0,\ \sum_{j\in J_K} \lambda_j = 1 \right\} \tag{8.13}$$

der zugehörigen Eckpunkte notiert werden. Dabei ist J_K die Menge aller Indizes der Eckpunkte von K. Über die Darstellung des Teilgebietes als Konvexkombination der Eckpunkte gewinnt man zugleich eine Standardparametrisierung.

Definition 8.16. *Die Koordinaten λ_j mit $j \in J_K$ in der Darstellung (8.13) des Teilgebietes K heißen* baryzentrische Koordinaten.

Spezielle finite Elemente werden gekennzeichnet durch die Lage und Art der verwendeten *Freiheitsgrade*, d.h. durch die Vorgabe des Funktionenraums $\mathcal{P}$ und der Funktionalmenge Σ. In den später folgenden Abbildungen werden im jeweiligen Gitterpunkt die folgenden Symbole benutzt:

- $\bullet$ - Verwendung des Funktionswertes als Freiheitsgrad
- $\circ$ - Verwendung aller ersten partiellen Ableitungen als Freiheitsgrade
- $\bigcirc$ - Verwendung aller zweiten partiellen Ableitungen als Freiheitsgrade.

Die Gesamtzahl der auf K verwendeten Freiheitsgrade sei d.

Nachfolgend betrachten wir Elemente auf *regulären Simplizes* in *allgemeiner Lage*. Bei der praktischen Berechnung (vgl. Kap. 9/10) greift man jedoch möglichst auf die Transformation auf *Referenzelemente* zurück.

Beispiel 8.17. *Finite Elemente über Dreiecken*

Sei $K \in \mathcal{T}_h$ ein nichtentartetes Dreieck mit den (in mathematisch positiven Drehsinn durchnumerierten) Eckpunkten p^1, p^2, p^3. Die zugehörigen baryzentrischen Koordinaten $\lambda_1, \lambda_2, \lambda_3$ werden den Punkten $x \in \overline{K}$ eineindeutig durch die Gleichungen

$$x = \sum_{i=1}^{3} \lambda_i p^i, \qquad \sum_{i=1}^{3} \lambda_i = 1 \tag{8.14}$$

zugeordnet. Mit $\tilde{\lambda} := (\lambda_1, \lambda_2)^* \in \mathbf{R}^2$ findet man stets zu (8.14) eine affine inverse Abbildung $\tilde{\lambda} = Bx + b$, die das Dreieck allgemeiner Lage in das Einheitsdreieck $\tilde{K} := \{\tilde{\lambda} :\ 0 < \lambda_1, \lambda_2 < 1;\ \lambda_1 + \lambda_2 < 1\}$ überführt. Diese Transformation werden wir insbesondere in Kapitel 10 benutzen.

Ansatzfunktionen über K lassen sich mittels der baryzentrischer Koordinaten angeben. So haben die über K affinen Funktionen $\phi_j,\ j = 1, 2, 3$ mit der Eigenschaft $\phi_j(p^k) = \delta_{jk}$ die Form

$$\phi_j(x) = \lambda_j(x), \quad j = 1, 2, 3.$$

Sie bilden die Menge $P_1(K)$ der *stückweise linearen Dreieckselemente*.

Man erhält allgemeiner *Dreieckselemente der Klasse $P_l(K),\ l \in \mathbf{N}$*, wenn neben den Eckpunkten p^j als weitere Interpolationspunkte

$$p^\alpha = \sum_{j=1}^{3} \frac{\alpha_j}{|\alpha|} p^j \tag{8.15}$$

benutzt werden. Dabei ist $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ ein Multiindex der Länge $|\alpha| = l$.

Abbildung 8.3 zeigt nacheinander

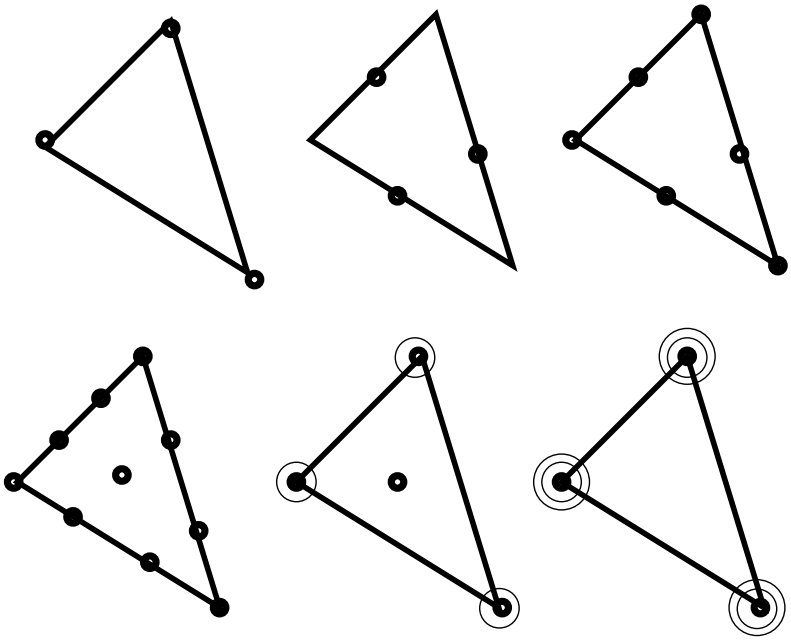


Abbildung 8.3: Auswahl von Dreieckselementen

- lineares C^0-Element (*"stetiges" P_1-Element*) mit $d = 3$
- "unstetiges" lineares Element (*Crouzeix-Raviart-Element*) mit $d = 3$
- quadratisches C^0-Element (*"stetiges" P_2-Element*) mit $d = 6$
- kubisches C^0-Element (*"stetiges" P_3-Element*) mit $d = 10$
- kubisches C^0-Element mit Hermite-Interpolation und $d = 10$
- reduziertes quintisches C^1-Element (*Bell-Element*) mit Hermite-Interpolation und $d = 18$.

Die "stetigen" P_l-Elemente mit $l = 1, 2, 3$ sind spezielle Lagrange-Elemente, die auch im Programmsystem FEMLAB Verwendung finden. Der nachfolgende Satz 8.19 klärt, daß man durch stetige Fortsetzung derartiger Elemente auf Nachbardreiecke einen konformen C^0-Finite-Elemente-Raum erhält.

Beim unstetigen P_1-Element verwendet man statt der Eckpunkte des Dreiecks die Seitenmittelpunkte als Interpolationspunkte. Das entstehende stückweise lineare Element ist das einfachste *nichtkonforme* Element. Offenbar ist eine stetige Fortsetzung einer Formfunktion auf ein benachbartes Element gleicher Art im allgemeinen Fall nicht möglich. Wir gehen auf nichtkonforme Elemente in Kapitel 11 ein.

Die beiden Hermite-Elemente erlauben auch den glatten Übergang in den ersten partiellen Ableitungen und somit die Konstruktion eines C^1-Finite-Elemente-Raums. Man beachte, daß beim Bell-Element die Zahl der Freiheitsgrade des originalen quintischen Elements deutlich reduziert wurde. Hierzu ist zu vermerken, daß die Zahl der Freiheisgrade pro Element wesentlich für den Besetzungsgrad der Matrix des entstehenden linearen Gleichungssystems ist. □

Beispiel 8.18. *Finite Elemente über regulären Simplizes*

Sei jetzt allgemeiner $K \in \mathcal{T}_h$ ein regulärer Simplex im $\mathbf{R}^n$ mit den Eckpunkten $p^1, ..., p^{n+1}$. Erneut ist K Konvexkombination aller Eckpunkte nach

$$\overline{K} = \ \mathrm{conv}\{p^j\}_{j\in J_K} := \left\{ x = \sum_{j\in J_K} \lambda_j p^j : \ \lambda_j \geq 0, \ \sum_{j\in J_K} \lambda_j = 1 \right\}.$$

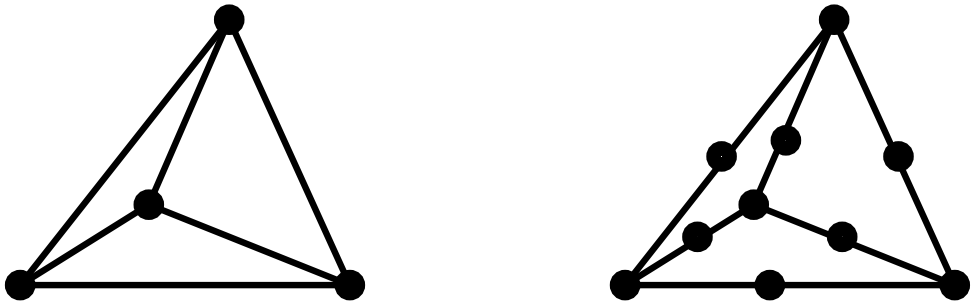


Abbildung 8.4: Auswahl von Tetraederelementen

Die Zuordnung zwischen den Punkten $x \in \overline{K}$ und den baryzentrischen Koordinaten $\lambda = (\lambda_1, ..., \lambda_{n+1})$ ist wieder eineindeutig.

Unter Verwendung des Multiindex $\alpha = (\alpha_1, ..., \alpha_{n+1})$ der Länge $l \in \mathbf{N}$ werden (in Verallgemeinerung von (8.15)) die Interpolationspunkte

$$p^\alpha = \sum_{j=1}^{n+1} \frac{\alpha_j}{|\alpha|} p^j$$

aus den Eckpunkten $p^1, ..., p^{n+1}$ des Teilgebietes K erzeugt.

Simpliziale Elemente der Klasse $P_l(K)$ mit $l \in \mathbf{N}$ erklärt man wie im vorhergehenden Beispiel. Die Abbildung 8.4 zeigt nacheinander für $n = 3$ die beiden Lagrange-Elemente niedrigster Ordnung, die auch im Programmsystem FEMLAB verwendet werden:

- lineares C^0-Element (*stetiges P_1-Element*) mit $d = 4$
- quadratisches C^0-Element (*stetiges P_2–Element*) mit $d = 10$. □

Der nachfolgende Satz klärt, daß über zulässigen simplizialen Zerlegungen mit Lagrange-Elementen ein C^0-Finite-Elemente-Raum erzeugt werden kann. Dies ist wesentlich für die numerische Lösung von elliptischen Randwertproblemen 2. Ordnung (etwa im System FEMLAB).

Satz 8.19. *Sei $\mathcal{T}$ eine zulässige Zerlegung des beschränkten, polyedrischen Gebietes $\Omega \subset \mathbf{R}^n$ in reguläre Simplizes. Dann wird durch die in Beispiel 8.18 erklärten Lagrange-Elemente der Klasse P_l, $l \in \mathbf{N}$ ein C^0-Finite-Elemente-Raum $X_\mathcal{T}$ gebildet.*

Beweis: vgl. S. Brenner/ R. Scott [4], Satz 3.3.17 für $n = 2$ und z.T. Übungsaufgabe. □

Wir gehen nun noch auf Zerlegungen in *Rechtecke* bzw. allgemeiner in *Quadergebiete* ein, die man sehr oft in Finite-Elemente-Paketen benutzt.

Beispiel 8.20. *Finite Elemente über Rechtecken*

Sei jetzt $K \subset \mathbf{R}^2$ ein Rechteck in allgemeiner Lage mit den (in mathematisch positiven Drehsinn durchnumerierten) Eckpunkten $p^1, ..., p^4$. Mit der folgenden affin-linearen Transformation unter Berücksichtigung der Eckpunkte p^1, p^2, p^4 kann man dieses Rechteck auf das Einheitsquadrat $\tilde{K} = (0,1) \times (0,1)$ als Referenzelement transformieren: Die Punkte p^1, p^2, p^4 werden wie in Beispiel 8.17 auf die Eckpunkte $(0,0)$, $(1,0)$ und $(0,1)$ abgebildet, der Punkt p^3 wird dabei auf den Eckpunkt $(1,1)$ transformiert.

Wir benutzen für das Referenzelement $\tilde{K}$ eine Parametrisierung der Form

$$\lambda_1 = (1-\xi)(1-\eta), \quad \lambda_2 = \xi(1-\eta), \quad \lambda_3 = \xi\eta, \quad \lambda_4 = (1-\xi)\eta \tag{8.16}$$

mit Parametern $\xi, \eta \in [0,1]$. Offenbar gelten die für baryzentrische Koordinaten gültigen Beziehungen

$$\lambda_i \geq 0, \ i = 1, ..., 4, \qquad \sum_{i=1}^{4} \lambda_i = 1, \qquad \forall \xi, \eta \in [0,1].$$

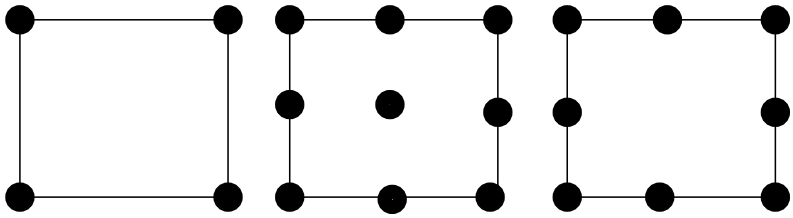


Abbildung 8.5: Auswahl von Rechteckselementen

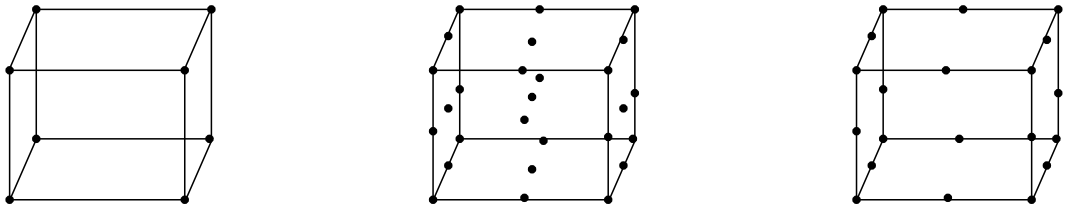


Abbildung 8.6: Auswahl von Quaderelementen

Sei jetzt $Q_l(\tilde{K})$ für $l \in \mathbf{N}$ die Menge aller Polynome auf $\tilde{K}$, die sich als Linearkombination von Produkten von Polynomen jeweils vom Grad l in jeder Richtung ξ bzw. η im Fall der Parametrisierung (8.16) darstellen lassen. Die Abbildung 8.5 zeigt nacheinander

- bilineares C^0-Element (*stetiges Q_1-Element*) mit $d = 4$
- biquadratisches C^0-Element (*stetiges Q_2-Element*) mit $d = 9$
- biquadratisches Serendipity-Element mit $d = 8$

Die Q_l-Elemente mit $l = 1, 2$ sind die üblichen Lagrange-Elemente. Bei dem biquadratischen Serendipity-Element wurde gegenüber dem Standard-Q_2-Element ein Knoten entfernt. □

Beispiel 8.21. *Quaderelemente*

Beispiel 8.20 läßt sich analog erweitern auf den Fall von Hyperquadern im $\mathbf{R}^n$. Für den dreidimensionalen Fall zeigt Abbildung 8.6 nacheinander die beiden Lagrange-Elemente niedrigster Ordnung sowie ein reduziertes Element:

- trilineares C^0-Element (*stetiges Q_1-Element*) mit $d = 8$
- triquadratisches C^0-Element (*stetiges Q_2-Element*) mit $d = 27$
- triquadratisches Serendipity-Element mit $d = 20$

Man beachte, daß beim reduzierten Q_2-Element die Zahl der lokalen Freiheitsgrade gegenüber dem üblichen Q_2-Element deutlich verringert ist. □

Bemerkung 8.22. Es ist möglich, simpliziale und Quader-Vernetzungen zu kombinieren. Wir betrachten eine deratige Variante in Kapitel 9. □

Kapitel 9

Praktische Aspekte der Finite-Elemente-Methode

Die Implementierung eines FEM-Programms zur Lösung partieller Differentialgleichungen ist eine sehr komplexe Aufgabenstellung. Im vorliegenden Kapitel werden die *Grundstruktur* eines FEM-Programms sowie einige Aspekte der praktischen Realisierung (vorwiegend an einem einfachen Beispiel nach Großmann/Roos [10], S. 175-180) behandelt.

9.1 Grundstruktur eines FEM-Programms

Ziel ist die Entwicklung eines FEM-Programms zur Lösung von Randwertproblemen partieller Differentialgleichungen in einem beschränkten Gebiet $\Omega \subset \mathbf{R}^n$. Wesentliche Bestandteile eines Finite-Elemente-Programms sind:

1. **Prä-Prozeß:** Eingabe und Beschreibung des (kontinuierlichen) Ausgangsproblems, Gittergenerierung, Generierung des endlichdimensionalen Problems
2. **Solver:** Lösung des entstehenden algebraischen Problems durch geeignete direkte oder iterative Lösungsverfahren
3. **Post-Prozeß:** Aufbereitung der Resultate (grafische Darstellung, Ermittlung abgeleiteter Größen), Bewertung der Resultate.

Oft wird dieser Zyklus iterativ abgearbeitet, da eine Bewertung der Resultate eine Verbesserung des Gitters und ggf. der Diskretisierung erfordert (vgl. Kapitel 12).

Die Komplexität eines FEM-Programms und die Vielzahl möglicher Modellprobleme erlauben keine Universallösungen für beliebige Klassen partieller Differentialgleichungen. Einerseits gibt es eine Vielzahl kommerzieller Lösungen für abgegrenzte Aufgabenstellungen (in der Regel für definierte Anwendungsfälle, z.B. ANSYS, NASTRAN, FLUENT oder FLOW3D in der Festkörper- oder Fluidmechanik). Andererseits gibt es jedoch inzwischen auch einige gut erprobte und ausbaufähige Programmsysteme für relativ große Aufgabenklassen. Dazu gehören die Systeme PLTMG (von R.E. Bank), ELLPACK (von J.R. Rice u.a.), KASKADE (von J. Leinen/ P. Deuflhard u.a.) und UG (von P. Bastian u.a.). Das im Rahmen dieser Vorlesung genutzte System FEMLAB gehört zu einer relativ neuartigen Klasse von Programmen mit multiphysikalischen Anwendungen.

Im vorliegenden Kapitel sollen einige Aspekte des *Prä-Prozesses* besprochen werden. Auf die Lösung der entstehenden diskreten Probleme gehen wir im Teil III dieser Vorlesung ein. Eines der weiteren Kapitel wird auch die Frage der a-posteriori Bewertung der diskreten Lösung als immer wichtiger werdender Bestandteil des Post-Prozesses berühren.

Die Komplexität eines FEM-Programms legt eine *modulare Struktur* der angestrebten Lösung nahe. Als

zunehmend wichtig erweist sich, derartige Programme einerseits erweiterungsfähig zu gestalten. Andererseits soll es möglich sein, einzelne Bestandteile (z.B. Prä- und Post-Prozessoren oder Löser für lineare Gleichungssyteme) auszuwechseln. Ebenso wichtig ist, daß ein FEM-Programm ohne großen Aufwand auf verschiedenen Hardware-Plattformen laufen kann. Aus Effektivitätsgründen muß ein FEM-Programm natürlich schnell abgearbeitet werden können.

9.2 Gebietsbeschreibung. Generierung eines Ausgangsgitters

Zunächst ist das Lösungsgebiet $\Omega \subset \mathbf{R}^n$ geeignet zu beschreiben. Wir diskutieren dies exemplarisch anhand des Programmsystems FEMLAB. Hiermit kann man einen Katalog bestimmter Grundgebiete benutzen, z.B. Quadrate, Rechtecke, Kreis- und Ellipsengebiete für $n = 2$ und Quader, Tetradeder, Zylinder- , Kegel- und Kugelgebiete für $n = 3$. Aus diesen Bausteinen kann man auch kompliziertere Gebiete mittels Boolescher Operationen wie Vereinigung und Schnitt konstruieren. Exemplarisch zeigt Abb. 9.1 die Konstruktion eines Beispiels jeweils für $n = 2$ und $n = 3$.

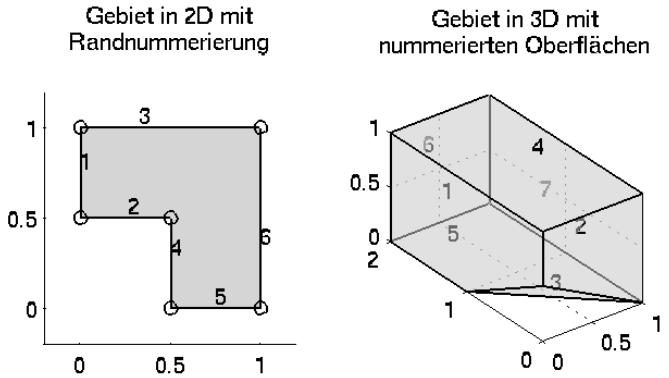


Abbildung 9.1: Beispielgebiete im zwei- bzw. dreidimensionalen Fall

Für die Eingabe von Randbedingungen ist eine geeignete Markierung entsprechender Teilgebiete des Randes $\partial\Omega$ geeignet, z.B. durch Nummerierung, erforderlich. Die ist in Abb. 9.1 für die Beispielgebiete ersichtlich. Natürlich unterstützt FEMLAB die grafische Ausgabe der so beschriebenen Gebiete.

Im nächsten Schritt erfolgt eine Zerlegung des Gebietes Ω. Ein wesentlicher Vorteil von FEM besteht in der Anpassungsfähigkeit der Zerlegung an die Gebietsgeometrie sowie an die Struktur der Lösung (z.B. Singularitäten). Wir behandeln hier nur die Erzeugung eines *Ausgangsgitters*, das die Geometrie beschreibt. Seine Qualität beeinflußt jedoch auch wesentlich Erfolg und Effizienz der gesamten Rechnung. Die Netzverbesserung auf der Basis von *a-posteriori Fehlerschätzern* behandeln wir in Kapitel 12.

Vereinfachend sei $\Omega \subset \mathbf{R}^n$ mit $n \leq 3$ ein beschränktes polyedrisches Gebiet. Wir betrachten eine nichtüberlappende, zulässige Zerlegung $\mathcal{T}_h = \{K_i\}_{i=1}^M$ des Gebietes in konvexe polyedrische Teilgebiete K_i mit

$$\overline{\Omega} = \cup_{j=1}^M \overline{K}_j, \quad K_i \cap K_j = \emptyset,\ i \neq j; \qquad h_i := \mathrm{diam}(K_i), \quad h := \max_{i=1,\ldots,M} h_i. \tag{9.1}$$

In FEMLAB sind die Teilgebiete K_i Simplizes, d.h. im Fall $n = 1$ Intervalle, für $n = 2$ Dreiecke bzw. und für $n = 3$ Tetraeder. (Eine andere gern benutzte Wahl sind konvexe Vierecke bzw. Hyperquader sowie die Kombination mit simplizalen Elementen, vgl. Abschnitt 9.3.)

Die Erzeugung eines Ausgangsgitters durch *Gittergeneratoren* ist sehr komplex und derzeit noch stark durch heuristische Prinzipien charakterisiert. In FEMLAB erfolgt ausgehend von der Randbeschreibung

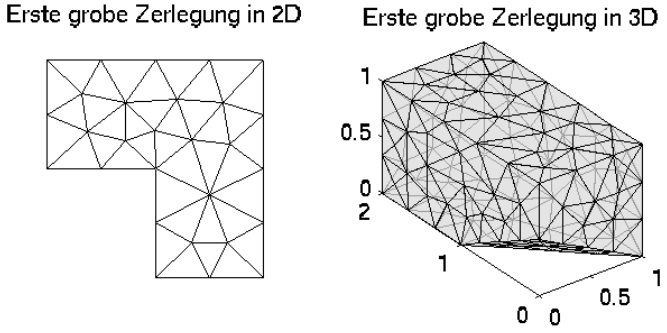


Abbildung 9.2: Ausgangszerlegung der Beispielgebiete im zwei- bzw. dreidimensionalen Fall

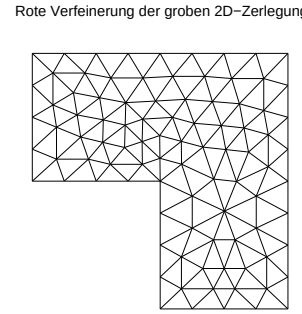


Abbildung 9.3: Rote Verfeinerung des Ausgangsgitters im zwei-dimensionalen Beispiel

und der Vorgabe eines maximalen Gitterparameters h_{max} zunächst eine interne Knotenverteilung auf $\partial\Omega$. Mit einer *advancing front*-Technik erfolgt dann die Gittererzeugung vom Rand in das Gebietsinnere. Das erzeugte Gitter ist stets zulässig.

Zur genaueren Beschreibung einer Zerlegung $\mathcal{T}$ definiert man für jedes Element K die Größe h_K als Durchmesser der kleinsten Kugel, in die K einbeschrieben werden kann. Ferner ist ρ_K der Durchmesser der größten in K einbeschriebenen Kugel. Jeder Zerlegung $\mathcal{T}$ ordnet man einen Index h gemäß (9.1) zu. Die so entstehende Familie $\{\mathcal{T}_h\}_h$ zulässiger Zerlegungen von Ω kann oft genauer charakterisiert werden.

Definition 9.1. *Eine Familie $\{\mathcal{T}_h\}_h$ heißt* isotrop, *falls gleichmäßig für $0 < h \leq h_0$ eine Beschränkung $\max_{K\in\mathcal{T}_h} h_K/\rho_K \leq c_0$ gibt. Sie heißt* quasi-uniform, *falls gleichmäßig für $0 < h \leq h_0$ Konstanten $0 < c_! < c_2$ existieren, so daß jedem Element K eine Kugel mit Durchmesser $c_1 h$ einbeschrieben und eine Kugel mit Durchmesser $c_2 h$ umbeschrieben werden kann.*

Die Isotropie schreibt geometrisch nichtentartete Elemente vor. So sind alle Innenwinkel von Dreiecken oder Tetraedern größer als ein gewisser Minimalwinkel. Quasi-uniforme Zerlegungen sind isotrop, aber zusätzlich sind alle Elemente von gleicher Größenordnung. Die mit FEMLAB erzeugten Ausgangsgitter sind automatisch isotrop. Dazu erfolgt intern eine geeignete Glättung des Gitters. Abb. 9.2 zeigt das Resultat so erzeugter Ausgangsgitter für unsere Beispielgebiete aus Abb. 9.1. Während das Ausgangsgitter die Geometrie des Gebietes beschreibt, ist oft für die Approximation der Lösung des Variationsproblems eine feinere Zerlegung erforderlich. FEMLAB erlaubt die automatische Verfeinerung des Ausgangsgitters. Bei der sogenannten *roten* Verfeinerung weden im Fall $n = 2$ Dreiecke in vier kongruente Teildreiecke zerlegt. Abb. 9.3 zeigt das Ergebnis dieser Operation für das zweidimensionale Beispiel aus Abb. 9.1 mit dem Ausgangsgitter in Abb. 9.2.

Bemerkung 9.2. Eine verbesserte Gitterqualität erreicht man oft mittels *Delaunay-Triangulation*. Gegebene innere Eckpunkte werden dabei so zu simplizialen Elementen verbunden, daß im Inneren des Umkreises eines jeden Elementes keine Eckpunkte der Zerlegung liegen. Man kann zeigen, daß eine derartige Triangulation zu kleine Innenwinkel der Dreieckselemente vermeidet. Eine weitere Verbesserung im Sinne dieses Kriteriums wird durch Verschieben innerer Eckpunkte erreicht. □

9.3 Datenstrukturen

Aus Komplexitätsgründen muß ein *Kompromiß* hinsichtlich des Umfangs der im Programm abzuspeichernden Daten gefunden werden. Nachfolgend stellen wir einige Informationen zusammen, die in einem FEM-Programm unbedingt gespeichert werden müssen:

- Liste der einzelnen Elemente durch geeignete Numerierung der Eckpunkte
- geometrische Lage der Gitterpunkte und der dort lokalisierten Freiheitsgrade oder Randbedingungen
- Information über die Approximation ggf. auftretender krummliniger Randkomponenten.

Hierzu hat sich eine *Listentechnik* als geeignet erwiesen (vgl. folgendes Beispiel), die auch in FEMLAB benutzt wird.

Beispiel 9.3. (*Listentechnik zur Beschreibung von Gebiet und FE-Zerlegung*)
Auf dem Dreiecksgebiet $\Omega := \{(x,y) \in \mathbf{R}^2 : x, y > 0, \quad 0 < x + y < 1\}$ wird eine Approximation an die Lösung $u \in W_0^{1,2}(\Omega)$ der Variationsgleichung

$$\int_\Omega \nabla u \cdot \nabla v \; dx = \int_\Omega f v \; dx, \quad \forall v \in W_0^{1,2}(\Omega) \tag{9.2}$$

zum homogenen Dirichlet-Problem der Poisson-Gleichung gesucht.

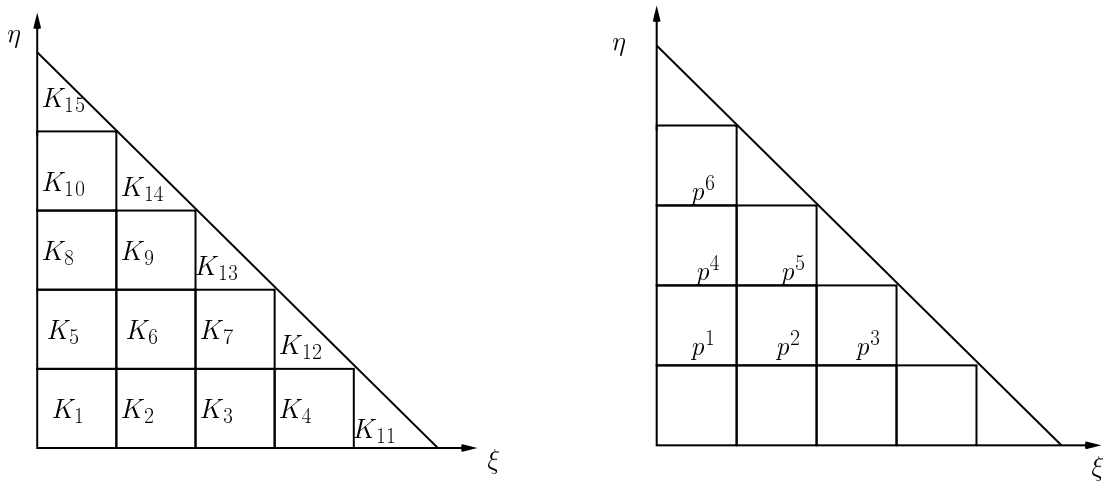


Abbildung 9.4: Zerlegung des Gebietes Ω und Knotennumerierung in ω_h

Bei Verwendung der FDM auf einem kartesischen Gitter (vgl. Kap. 3) hätte man Probleme mit der Diskretisierung an der nicht parallel zum Gitter liegenden Kante. Derartige Probleme enstehen bei der FEM nicht. Wir wählen eine gemischte Zerlegung des Gebietes Ω in Quadrate und Dreiecke $K_j, j = 1, ..., M$ (vgl. Abb. 9.4). Im betrachteten Fall ist $M = 15$. Wir numerieren die Quadrate K_i mit $i = 1, ..., 10$ und die Dreiecke K_i mit $i = 11, ..., 15$. Die Quadrate haben die Seitenlänge $h = 0.2$, entsprechend ist h auch die Länge der kurzen Dreiecksseiten.

Es bezeichne $p^j,\ j = 1, ..., N$ die Menge der inneren Gitterpunkte sowie $p^j,\ j = 1, ..., \overline{N}$ die Menge aller Gitterpunkte. Abkürzend sei

$$\omega_h := \{p^i\}_{i=1}^{N}, \qquad \overline{\omega}_h := \{p^i\}_{i=1}^{\overline{N}}.$$

Die folgende Liste der Koordinaten bezieht sich auf die inneren Gitterpunkte $p^i,\ i = 1, ..., 6$.

i	x_i	y_i
1	0.2	0.2
2	0.4	0.2
3	0.6	0.2
4	0.2	0.4
5	0.4	0.4
6	0.2	0.6

Tabelle 9.1. Liste der Koordinaten der inneren Gittepunkte

Ebenso stellt man eine Liste für die Randgitterpunkte $p^i,\ i = 7, ..., 21$ auf.

j					j				
1	7	8	1	13	6	1	2	5	4
2	8	9	2	1	7	2	3	16	8
3	9	10	3	2	8	15	4	6	17
4	10	11	14	3	9	4	5	18	6
5	13	1	4	15	10	17	6	20	19

Tabelle 9.2. Liste der Rechteckgebiete mit Eckenindizierung

Schließlich benötigt man eine Listenbeschreibung für die Teilgebiete $K_j,\ j = 1, ..., 15$ über die auftretenden Eckpunkte. Dabei ist zwischen Rechteck- und Dreieckselementen zu unterscheiden sowie ein einheitlicher Durchlaufsinn bei der Beschreibung der Eckpunkte zu beachten (vgl. Tabelle 9.2). Analog stellt man eine Liste der Dreieckselemente $K_j,\ j = 11, ..., 15$ auf. □

Bemerkungen 9.4. (i) Eine gesonderte Beschreibung der homogenen Dirichlet-Randbedingung ist im Beispiel 9.3 nicht erforderlich. Einerseits wurde bereits nach inneren und Randgitterpunkten unterschieden. Andererseits kann man die entsprechenden Freiheitsgrade auf dem Rand weglassen, indem man die Nullrandwerte im Gleichungssystem konsequent eliminiert und so dessen Dimension verringert. Es sei jedoch bereits auf Bemerkung 9.6 (i) verwiesen.

(ii) Oft ist es eine Aufbereitung weiterer Informationen im Programm wie etwa die Numerierung der Nachbarelemente in Listenform zweckmäßig, um einen schnellen Zugriff auf diese Daten zu erlauben. Dies spielt insbesondere bei Fragen der adaptiven Gitterverbesserung eine Rolle. □

9.4 Generierung des diskreten Problems

Wir befassen uns jetzt mit der Generierung der Matrix und der rechten Seite des entstehenden diskreten Problems. Ausgangspunkt ist (wie bereits in Abschnitt 7.1 beschrieben) eine elliptische Variationsgleichung der Form

$$\text{Finde } u \in X: \quad a(u, v) = f(v), \qquad \forall v \in X \tag{9.3}$$

im Hilbert-Raum X mit Bilinearform $a(\cdot, \cdot)$ und Linearform $f(\cdot)$. Zur Bestimmung einer Näherung u_h an die Lösung u von (9.3) im endlich–dimensionalen Teilraum $X_h \subset X$ mit $\dim X_h = \tilde{N} = \tilde{N}(h) < \infty$ über das Galerkin–Verfahren

$$\text{Finde } u_h \in X_h: \quad a(u_h, v) = f(v), \qquad \forall v \in X_h \tag{9.4}$$

geht man von einer Basis $\{\phi_i\}_{i=1}^{\tilde{N}}$ von X_h aus. Über die Basisdarstellung

$$u_h(x) = \sum_{i=1}^{\tilde{N}} u_i \phi_i(x) \tag{9.5}$$

erhält man für den Knotenvektor $\underline{u} = (u_1, ..., u_{\tilde{N}})^*$ das lineare Gleichungssystem

$$A_h \underline{u} = f_h. \tag{9.6}$$

mit der sogenannten *Steifigkeitsmatrix* A_h und dem *Lastvektor* f_h gemäß

$$A_h = (A_{ij})_{i,j=1}^{\tilde{N}}, \;\; A_{ij} := a(\phi_j, \phi_i), \qquad f_h = (f_1, ..., f_{\tilde{N}})^*, \;\; f_i := f(\phi_i).$$

Eine effiziente Generierung von (9.6), insbesondere der (*Assemblierung*) von A_h, erfolgt durch eine Schleife über alle Elemente K_j: Man nutzt, daß die FEM-Ansatzfunktionen ϕ_i einen kleinen Träger haben. Man berechnet die Anteile von A_h über den Teilgebieten K_j, die *Elementmatrizen*, gesondert und addiert sie dann in A_h auf. Dabei wird ausgenutzt, daß jeweils nur sehr wenige *Nichtnullelemente* zu berücksichtigen sind. Eine weitere Vereinfachung wird durch Transformation der Teilgebiete auf sogenannte *Referenzelemente* erreicht.

Wir beschreiben die Assemblierung der Matrix A_h am Beispiel 9.3. Zugleich sollen bei der Darstellung wesentliche Prinzipien bzw. Methoden verdeutlicht werden.

Beispiel 9.5. (*Assemblierung des linearen Gleichungssystems*)
Wir nutzen das Problem und die Bezeichnungen aus Beispiel 9.3. Im vorliegenden Fall stimmen die Zahl $\tilde{N}$ der Freiheitsgrade und $\overline{N}$ der inneren Knotenpunkte überein. Weiterhin ist $M = 15$ die Zahl der Elemente sowie $h = 0.2$.

Für die stückweise Definition der konformen Näherung u_h fordern wir neben $u_h \in X_h \subset C(\overline{\Omega})$, daß gilt

$$(i)\; u_h|_{K_i} \text{ bilinear auf } K_i, \; i = 1, ..., 10; \qquad (ii)\; u_h|_{K_i} \text{ linear auf } K_i, \; i = 11, ..., 15.$$

Zum Gitterpunkt p^j gehöre die bilineare bzw. lineare Funktion $\phi_j \in C(\overline{\Omega})$ mit

$$\phi_j(p^k) = \delta_{jk}, \quad j, k = 1, ..., \overline{N}.$$

In Umgebung der inneren Eckpunkte p^j mit $j = 1, 2, 4$ sind lediglich bilineare Ansatzfunktionen über Rechtecken wirksam. Sie haben die Form

$$\phi_j(x,y) := \begin{cases} \frac{1}{h^2}(h - |x - x_j|)(h - |y - y_j|), & \max\{|x - x_j|, |y - y_j|\} \le h \\ 0, & \text{sonst} \end{cases}.$$

In Umgebung der inneren Eckpunkte p^j mit $j = 3, 5, 6$ benötigt man sowohl bilineare Basisfunktionen über den Rechtecken als auch lineare Ansatzfunktionen über Dreiecken:

$$\phi_j(x,y) := \begin{cases} \frac{1}{h^2}(h - |x - x_j|)(h - |y - y_j|), & \max\{|x - x_j|, |y - y_j|\} \le h \\ & \text{und} \quad \min\{x - x_j, y - y_j\} \le 0 \\ \frac{1}{h}(h - (x - x_j) - (y - y_j)), & |x - x_j| + |y - y_j| \le h \\ & \text{und} \quad \min\{x - x_j, y - y_j\} \ge 0 \end{cases}.$$

Der Ansatz

$$u_h(x,y) := \sum_{j=1}^{\tilde{N}} u_j \phi_j(x,y) \tag{9.7}$$

erfüllt dann die oben genannten Anforderungen sowie die homogene Dirichlet-Randbedingung. Hier ist $\tilde{N} = 6$ (vgl. Abb. 9.3). Der Ansatz (9.7) führt in der Variationsgleichung (9.2) zum linearen Gleichungssystem

$$\sum_{j=1}^{6} \left(\int_\Omega \nabla\phi_j \cdot \nabla\phi_i \; dx \right) u_j = \int_\Omega f\phi_i \; dx, \quad i = 1, ..., 6.$$

Die Elementmatrizen zum Teilgebiet K_j haben die Gestalt

$$A_h^j = (A_{ik}^j)_{i,k\in I_j}, \quad A_{ik}^j := \int_{K_j} \nabla\phi_i \cdot \nabla\phi_k \; dx$$

mit der Indexmenge $I_j := \{i : \; \mathrm{supp}(\phi_i) \cap K_j \neq \emptyset\}$.

Sinngemäß erhalten wir elementweise für den Vektor der rechten Seite

$$f^j = (f_i^j)_{i\in I_j}.$$

Bei Durchlauf der Elemente der Zerlegung erhält man schließlich wegen der Linearität des Integrals durch Aufaddition der Elementbeiträge

$$A_h = (A_{ik})_{i,k=1}^{\tilde{N}}, \qquad A_{ik} = \sum_{j=1}^{m} \sum_{i,k\in I_j} a_{ik}^j; \qquad f_h = (f_i)_{i=1}^{\tilde{N}}, \qquad f_i = \sum_{j=1}^{m} \sum_{i\in I_j} f_i^j.$$

Bei den hier betrachteten stückweise bilinearen und linearen Ansatzfunktionen über Rechtecken bzw. Dreiecken sind die Indexmengen I_j gerade die in Beispiel 9.2 eingeführten Listen der inneren Eckpunkte der Elemente.

Wir berechnen jetzt exemplarisch die Elementmatrix für ein *Dreieck* $K = \overline{K}_j$. Die Eckpunkte seien o.B.d.A. mit den lokalen Indizes $l = 1, 2, 3$ versehen. Als zweckmäßig hat sich ferner die Transformation auf das *Referenzelement*

$$\tilde{K} = \{(\xi \; \eta)^* : \; \xi, \eta \geq 0, \; \xi + \eta \leq 1\}$$

erwiesen. Dies erfolgt mittels der affinen Abbildung

$$\begin{pmatrix} x \\ y \end{pmatrix} = F_j \begin{pmatrix} \xi \\ \eta \end{pmatrix} := \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} \tag{9.8}$$

mit $(\xi \; , \eta)^* \in \tilde{K}$. Die zugehörige Jacobische Funktionaldeteminante

$$|B_j| = \begin{vmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{vmatrix} \tag{9.9}$$

ist bei Numerierung der Eckknoten in mathematisch positivem Sinne positiv.

Zur Umrechnung der partiellen Ableitungen benutzt man

$$\begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \eta}{\partial x} \\ \frac{\partial \xi}{\partial y} & \frac{\partial \eta}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{pmatrix}.$$

Partielle Differentiation von (9.8) ergibt mit der Jacobischen Funktionalmatrix B_j

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = B_j \begin{pmatrix} \frac{\partial \xi}{\partial x} \\ \frac{\partial \eta}{\partial x} \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix} = B_j \begin{pmatrix} \frac{\partial \xi}{\partial y} \\ \frac{\partial \eta}{\partial y} \end{pmatrix}$$

und damit

$$\begin{pmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \eta}{\partial x} \\ \frac{\partial \xi}{\partial y} & \frac{\partial \eta}{\partial y} \end{pmatrix} = |B_j|^{-1} \, T_j; \qquad T_j := \begin{pmatrix} y_3 - y_1 & y_1 - y_2 \\ x_1 - x_3 & x_2 - x_1 \end{pmatrix}.$$

Damit erhalten wir zur Umrechnung des Gradientenoperators

$$\begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} = |B_j|^{-1} \, T_j \begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{pmatrix}.$$

Mit den über dem Referenzdreieck $\tilde{K}$ definierten linearen Ansatzfunktionen

$$\tilde{\phi}_1(\xi,\eta) = 1-\xi-\eta, \qquad \tilde{\phi}_2(\xi,\eta) = \xi, \qquad \tilde{\phi}_3(\xi,\eta) = \eta$$

folgt nach Integraltransformation

$$\begin{aligned}
A_h^j &= \left(A_{ik}^j\right)_{i,k\in\{1,2,3\}} = \left(\int_{K_j} \nabla\phi_k \cdot \nabla\phi_i \; dK\right)_{i,k\in\{1,2,3\}} \\
&= \left(\int_{\tilde{K}} \nabla\tilde{\phi}_k \cdot \tilde{\nabla}\phi_i \; |B_j| \; d\tilde{K}\right)_{i,k\in\{1,2,3\}} \\
&= |B_j| \; \mathrm{meas}(\tilde{K}) \left(\nabla\tilde{\phi}_k \cdot \nabla\tilde{\phi}_i\right)_{i,k\in\{1,2,3\}} \\
&= \frac{1}{2\,|B_j|} \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} T_j^* \; T_j \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \\
&= \frac{1}{2\,|B_j|} \begin{pmatrix} y_2-y_3 & x_3-x_2 \\ y_3-y_1 & x_1-x_3 \\ y_1-y_2 & x_2-x_1 \end{pmatrix} \begin{pmatrix} y_2-y_3 & y_3-y_1 & y_1-y_2 \\ x_3-x_2 & x_1-x_3 & x_2-x_1 \end{pmatrix}.
\end{aligned}$$

Für ein Dreieck K_j *in allgemeiner Lage* mit den Eckpunkten p^i, p^k, p^l, also der Indexmenge $I_j = \{i,k,l\}$, gilt

$$A_{ik}^j = \begin{cases} \frac{1}{2|D_j|}\left[(x_i-x_l)(x_l-x_k) + (y_i-y_l)(y_l-y_k)\right], & \text{falls } i \neq k \\ \\ \frac{1}{2|D_j|}\left[(x_k-x_l)^2 + (y_k-y_l)^2\right], & \text{falls } i = k \end{cases}.$$

Analog kann man die Elementmatrizen zu *bilinearen Ansatzfunktionen* über rechteckigen Elementen K_j ermitteln. Im hier vorliegenden *Fall achsparalleler Rechtecke*

$$K_j = \mathrm{conv}\{p^1, p^2, p^3, p^4\} = [x_1, x_2] \times [y_2, y_3]$$

vereinfacht sich die Rechnung. Mit den Bezeichnungen

$$\Delta x_j := x_2 - x_1, \qquad \Delta y_j := y_3 - y_2; \qquad \alpha_j := \frac{1}{6\Delta x_j \; \Delta y_j}$$

folgt nach längerer Zwischenrechnung

$$A_h^j = \alpha_j \begin{pmatrix} 2\Delta x_j^2 + 2\Delta y_j^2 & \Delta x_j - 2\Delta y_j^2 & -\Delta x_j^2 - \Delta y_j^2 & -2\Delta x_j^2 + \Delta y_j^2 \\ \Delta x_j - 2\Delta y_j^2 & 2\Delta x_j^2 + 2\Delta y_j^2 & -2\Delta x_j^2 + \Delta y_j^2 & -\Delta x_j^2 - \Delta y_j^2 \\ -\Delta x_j^2 - \Delta y_j^2 & -2\Delta x_j^2 + \Delta y_j^2 & 2\Delta x_j^2 + 2\Delta y_j^2 & \Delta x_j - 2\Delta y_j^2 \\ -2\Delta x_j^2 + \Delta y_j^2 & -\Delta x_j^2 - \Delta y_j^2 & \Delta x_j - 2\Delta y_j^2 & 2\Delta x_j^2 + 2\Delta y_j^2 \end{pmatrix}.$$

Für unser Beispiel erhält man nach Berechnung der entsprechenden Integrale die folgende Matrix

$$A_h = \left(\int_\Omega \nabla\phi_j \cdot \nabla\phi_i \; dx\right)_{i,j} = \frac{1}{3} \begin{pmatrix} 8 & -1 & 0 & -1 & -1 & 0 \\ -1 & 8 & -1 & -1 & -1 & 0 \\ 0 & -1 & 9 & 0 & -1 & 0 \\ -1 & -1 & 0 & 8 & -1 & -1 \\ -1 & -1 & -1 & -1 & 9 & -1 \\ 0 & 0 & 0 & -1 & -1 & 9 \end{pmatrix}.$$

Bei feinerer Zerlegung des Gebietes erhält man eine größere Dimension der Matrix, jedoch steigt die Zahl der *Nullelemente* stark an. Die entstehenden Matrizen bezeichnet man als *schwachbesetzt*. Für derartige Probleme stehen dann direkte und iterative Lösungsverfahren zur Verfügung, die diese spezielle Matrixstruktur ausnutzen.

Die Berechnung der diskreten rechten Seite f_h erfolgt analog zur elementweisen Ermittlung der Steifigkeitsmatrix. □

Bemerkungen 9.6. (i) Randbedingungen 1. Art werden in der Regel (als sogenannte *wesentliche* Randbedingungen) direkt im Lösungsansatz berücksichtigt. Eine übliche Variante ist, die entsprechenden Funktionswerte "hart"'vorzugeben und die zugehörigen Randvariablen in den diskreten Gleichungen zu eliminieren.

Im System FEMLAB werden Dirichlet-Bedingungen als Nebenbedingungen "schwach" eingearbeitet. Wir werden darauf im Rahmen der Übungen und eventuell zu einem späteren Zeitpunkt in der Vorlesung eingehen. Eine derartige Vorgehensweise erhöht zwar die Dimension des Gleichungssystems, ist aber methodisch besser angepaßt an die Behandlungen von Randbedingungen (zum Beispiel 2. und 3. Art), in denen Ableitungen der gesuchten Funktion auftreten. Derartige Randbedingungen werden bekanntlich in die Variationsformuierung eingearbeitet, die Freiheitsgrade auf entsprechenden Randstücken treten als Unbekannte im diskreten Problem auf.

(ii) Bei bestimmten Randwertproblemen mit konstanten Koeffizienten (z.B. beim Laplace-Operator), die auf regelmäßigen Zerlegungen von achsparallelen Rechecken erzeugt werden, erhält man diskrete Probleme, die mit einem klassischen Differenzen-Verfahren übereinstimmen.

(iii) Struktur und Kondition der Elementmatrizen werden wesentlich durch die konkrete Darstellung der Basisfunktionen beeinflußt. Bei unstrukturierten Gittern mit lokal variabler Feinheit ist es eventuell sinnvoll, nicht alle Basisfunktionen auf das feinste Gitter zu beziehen. Bei der Technik der *hierarchischen Basen* beginnt man mit der FEM–Basis über dem gröbsten Gitter und fügt auf dem nächstfeineren Gitter lediglich die Basisfunktionen über den neuen Gitterpunkten hinzu. Dies ist insbesondere bei adaptiver Netzverfeinerung und bei iterativer Lösung der Gleichungssysteme mit Mehrgitterverfahren sinnvoll. □

Kapitel 10

Fehlerabschätzungen für konforme FEM

Im Kapitel 7 hatten wir für *konforme* Approximationen von elliptischen Variationsgleichungen bereits Fehlerabschätzungen auf ein *Approximationsproblem* für die Lösung des kontinuierlichen Problems in den diskreten Unterräumen zurückgeführt. Gegenstand des vorliegenden Kapitels ist die Gewinnung von Interpolationsfehleraussagen und damit von Fehlerabschätzungen für konforme FEM. Wir beschränken uns dabei auf den Fall der Interpolation auf *simplizialen* Elementen.

Bei der Analyse werden hier vorerst praktisch wichtige Aspekte wie numerische Integration und Approximation krummliniger Randteile ignoriert. Diese Probleme behandeln wir im Kontext *nichtkonformer* Approximationen im Kapitel 11.

10.1 Transformation auf das Referenzelement

Das polyedrische Gebiet Ω sei mittels einer zulässigen Zerlegung $\mathcal{T} = \{K_j\}_{j=1}^M$ exakt in konvexe, polyedrische Teilgebiete K_j zerlegt. Für ein beliebiges Element K seien h_K bzw. ρ_K jeweils der Radius der kleinsten bzw. größten Kugel, in die K einbeschrieben bzw. die in K einbeschrieben werden kann. Weiter sei $h := \max_{K\in\mathcal{T}} h_K$ für fixiertes $\mathcal{T}$. Ordnet man so jeder Zerlegung einen Index $h > 0$ zu, d.h. $\mathcal{T} = \mathcal{T}_h$, erhält man eine Familie $\{\mathcal{T}_h\}_h$ von Zerlegungen des Lösungsgebietes Ω.

Interpolationsfehlerabschätzungen gewinnen wir durch Transformation auf ein geeignetes *Referenzelement*. Vereinfachend wird angenommen, daß die Familie $\{\mathcal{T}_h\}_h$ erzeugt wird durch Transformationen von einem einheitlichen Referenzelement $\tilde{K}$. Wir beschränken uns exemplarisch auf eine *Dreieckszerlegung* eines Gebietes $\Omega \subset \mathbf{R}^2$, alle Aussagen gelten analog für simpliziale Zerlegungen im $\mathbf{R}^n$.

Mit den Bezeichnungen aus Abschnitt 9.4 sei $\tilde{K} = \{(\xi\ \eta)^* :\ \xi, \eta > 0,\ \xi + \eta < 1\}$ ein Referenzelement für ein allgemeines Dreieck K_j. Wir wollen im Regelfall annehmen, daß die Abbildung $F_j :\ \tilde{K} \to K_j$ *affin linear*, d.h. von der Form

$$\begin{pmatrix} x \\ y \end{pmatrix} = F_j(p) = B_j p + b_j, \qquad p := \begin{pmatrix} \xi \\ \eta \end{pmatrix} \tag{10.1}$$

mit regulärer Matrix B_j und geeignetem Vektor b_j ist (vgl. auch Abb. 10.1). Vereinfachend wird nachfolgend der Index j des allgemeinen Dreiecks K_j mit den Eckpunkten $(x_i, y_i)^*,\ i = 1, 2, 3$ weggelassen. Dann gilt genauer

$$\begin{pmatrix} x \\ y \end{pmatrix} = F(p) = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} + \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}. \tag{10.2}$$

Bemerkung 10.1. Sofern nicht ausdrücklich auf den affin-linearen Fall verwiesen wird, gelten die Ausführungen in den Abschnitten 10.1-10.3 teilweise auch für nichtlineare Abbildungen $F : \tilde{K} \to K$.

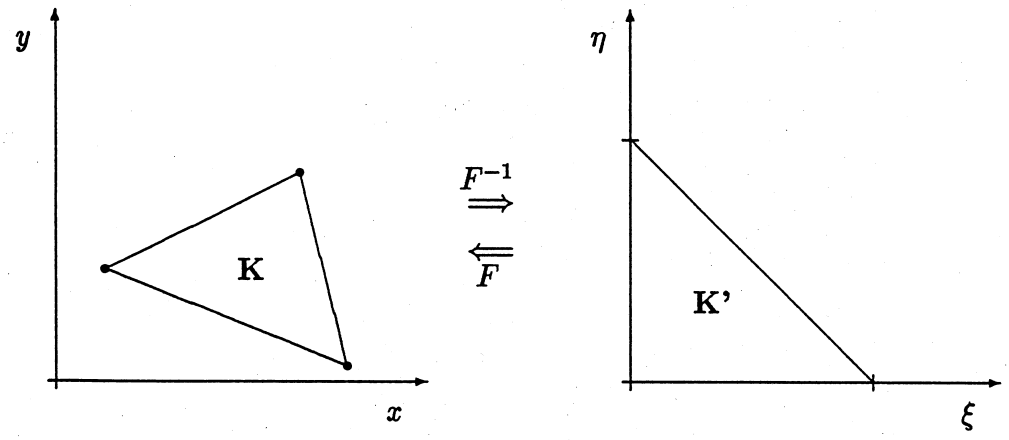


Abbildung 10.1: Transformation eines Dreiecks auf das Referenzelement

Darauf kommen wir in Kapitel 11 im Zusammenhang mit *isoparametrischen Elementen* zurück. □

Jeder Funktion $u : K \to \mathbf{R}$ wird eine Funktion $v : \tilde{K} \to \mathbf{R}$ über dem Referenzelement mit

$$v(p) = u(F(p)) \tag{10.3}$$

zugeordnet. Mittels Kettenregel folgt für differenzierbare Funktionen

$$\nabla_p v(p) = F'(p)\ \nabla_x u(F(p)). \tag{10.4}$$

Dabei sind ∇_p bzw. ∇_x die Gradientenoperatoren nach den $p-$ bzw. $x-$Koordinaten sowie $F'(\cdot)$ die Jacobische Funktionalmatrix. Für affin-lineare Abbildungen gilt nach (10.2) mit einer von h_K unabhängigen Konstanten C die Abschätzung

$$\|F'(p)\| \le C\ h_K, \quad \forall p \in \tilde{K}. \tag{10.5}$$

Nachfolgend betrachten wir über einem Gebiet $G \subset \Omega$ Sobolev-Räume $W^{l,2}(G),\ l \in \mathbf{N}_0$ mit der Standardnorm $\|\cdot\|_{l,G}$ bzw. Seminorm $|\cdot|_{l,G}$ gemäß

$$\|v\|_{l,G}^2 := \sum_{|\alpha|\le l} \|D^\alpha v\|_{0,G}^2, \qquad |v|_{l,G}^2 := \sum_{|\alpha|=l} \|D^\alpha v\|_{0,G}^2.$$

Zur Umrechnung von Normen über dem Element K bzw. über dem Referenzelement $\tilde{K}$ benötigt man die (o.B.d.A. als positiv angenommene) Funktionaldeterminante

$$d(p) = \det F'(p) > 0, \quad \forall p \in \tilde{K}.$$

Nach Gebietstransformation folgt über (10.3)

$$\int_K (u(x))^2\ dx = \int_{\tilde{K}} (v(p))^2 d(p)\ dp \tag{10.6}$$

und daraus für die Umrechnung der L^2-Normen

$$\left(\inf_{p\in\tilde{K}} d(p)\right)^{1/2} \|v\|_{0,\tilde{K}} \le \|u\|_{0,K} \le \left(\sup_{p\in\tilde{K}} d(p)\right)^{1/2} \|v\|_{0,\tilde{K}}. \tag{10.7}$$

Eine entsprechende Verallgemeinerung auf verallgemeinerte Ableitungen höherer Ordnung ist im Fall affin-linearer Abbildungen (10.1) gegeben durch

Lemma 10.2. *Ein Teilgebiet K und das Referenzelement $\tilde{K}$ seien durch die affin-lineare Abbildung*

$$x = F(p) = Bp + b, \quad p \in \tilde{K} \tag{10.8}$$

eineindeutig aufeinander abgebildet. Dann folgt für Transformationen nach (10.3)

(i) $u \in W^{l,2}(K) \iff v \in W^{l,2}(\tilde{K}), \quad l = 0, 1, ...$

(ii) Für die Seminormen gilt

$$|v|_{l,\tilde{K}} \leq C\|B\|^l \ |det(B)|^{-1/2} \ |u|_{l,K}, \qquad |u|_{l,K} \leq C\|B^{-1}\|^l \ |det(B)|^{1/2} \ |v|_{l,\tilde{K}}.$$

Beweis: (i) Die Aussage folgt unmittelbar aus (ii).

(ii) Wir beweisen die Aussage zunächst für hinreichend glatte Funktionen $u \in C^l(\overline{K})$. Mittels Kettenregel folgt aus (10.3) und (10.8)

$$\frac{\partial v}{\partial p_j} = \sum_{i=1}^{2} \frac{\partial u}{\partial x_i} \frac{\partial x_i}{\partial p_j},$$

damit

$$\left| \frac{\partial v}{\partial p_j} \right| \leq \|B\| \ \max_i \left| \frac{\partial u}{\partial x_i} \right|.$$

Rekursiv ergibt sich für einen Multiindex α, daß

$$|(D^\alpha v)(p)| \leq \|B\|^{|\alpha|} \max_{\substack{\beta \\ |\beta|=|\alpha|}} \left| (D^\beta u)(x(p)) \right|, \quad p \in \tilde{K}.$$

Wegen der Normäquivalenz auf endlich-dimensionalen Räumen ist

$$\sum_{|\alpha|=l} |(D^\alpha v)(p)|^2 \leq C\|B\|^{2l} \sum_{|\beta|=l} \left| (D^\beta u)(x(p)) \right|^2, \quad p \in \tilde{K}. \tag{10.9}$$

Aus (10.9) findet man durch Variablentransformation im Integral

$$\begin{aligned} |v|^2_{l,\tilde{K}} &= \int_{\tilde{K}} \sum_{|\alpha|=l} |(D^\alpha v)(p)|^2 \ dp \leq C\|B\|^{2l} \int_{\tilde{K}} \sum_{|\beta|=l} \left| (D^\beta u)(x(p)) \right|^2 \ dp \\ &\leq C\|B\|^{2l} \ |\det(B)|^{-1} \int_K \sum_{|\beta|=l} \left| (D^\beta u)(x) \right|^2 \ dx \\ &= C\|B\|^{2l} \ |\det(B)|^{-1} \ |u|^2_{l,K}. \end{aligned}$$

Diezweite Aussage von (ii) für hinreichend glatte Funktionen folgt analog unter Benutzung der Darstellung $p = B^{-1}x - B^{-1}b$. Wegen der Dichtheit von $C^l(\overline{K})$ in $W^{l,2}(K)$ erhalten wir die Behauptung (ii). □

Wir schätzen jetzt die in Lemma 10.2 auftretenden Größen $\|B\|$ und $\|B^{-1}\|$ ab.

Lemma 10.3. *Seien für ein Element K die Voraussetzungen von Lemma 10.2 erfüllt. Ferner sei $\tilde{K}$ ein festes und von der Zerlegung unabhängiges Referenzelement. Das Element K enthalte eine maximale Kugel mit Radius ρ_K und sei in eine minimale Kugel mit Radius h_K einbeschrieben. Dann gilt*

$$\|B\| \leq C \ h_K, \quad \|B^{-1}\| \leq C \ \rho_K^{-1}.$$

Beweis: Für das fixierte, nichtentartete Referenzelement $\tilde{K}$ existieren Kugeln mit den Radien $\tilde{\rho}$ und $\tilde{h}$, die in $\tilde{K}$ ein– bzw. um $\tilde{K}$ umbeschrieben sind. Dann existiert ein Punkt $p_0 \in \tilde{K}$ mit $p_0 + p \in \tilde{K}$ für beliebige p mit $\|p\|\infty = \tilde{\rho}$. Die mittels der affin–linearen Abbildung (10.8) zugeordneten Punkte $x_0 = Bp_0 + b$ sowie $x = B(p_0 + p) + b$ gehören zu K. Für sie ist dann $\|x - x_0\| \leq 2h_K$. Daraus folgern wir

$$\|B\| = \frac{1}{\tilde{\rho}} \sup_{\|p\|=\tilde{\rho}} \|Bp\| \leq \frac{1}{\tilde{\rho}} \|x - x_0\| \leq \frac{2h_K}{\tilde{\rho}}.$$

Die zweite Abschätzung folgt durch Vertauschung der Rolle von K und $\tilde{K}$. □

Folgerung 10.4. *Im Fall affin-linearer Transformationen eines Teilgebietes K auf das Referenzelement $\tilde{K}$ gibt es positive Konstanten $C_i, i = 0, 1, 2$ mit*

$$C_2 \left(\sup_{p\in\tilde{K}} d(p) \right)^{-\frac{1}{2}} \rho_K^r |u|_{r,K} \le |v|_{r,\tilde{K}} \le C_1 \left(\inf_{p\in\tilde{K}} d(p) \right)^{-\frac{1}{2}} h_K^r |u|_{r,K}, \tag{10.10}$$

$$0 < \frac{1}{C_0} \le \frac{\sup_{p\in\tilde{K}} d(p)}{\inf_{p\in\tilde{K}} d(p)} \le C_0 \tag{10.11}$$

für alle $u \in W^{r,2}(K)$.

Beweis: Die Funktionaldeterminante $d(P)$ ist bei affin-linearer Transformation konstant. Die Aussage folgt dann aus den Lemmata 10.2 und 10.3. □

Es sei angemerkt, daß eine Zerlegung $\mathcal{T}_h$ mit nichtlinearen Abbildungen $F : \tilde{K} \to K$ auch als *regulär* bezeichnet wird, wenn die Eigenschaften (10.11) und (10.10) bei Ersetzung von ρ_K und h_K durch h gleichmäßig erfüllt sind.

10.2 Lemma von Bramble-Hilbert

Jetzt werden wir den Fehler bei Polynominterpolation auf dem Referenzelement abschätzen. Wir benutzen dazu die Norm des zu $W^{l,2}(G)$ dualen Raumes

$$\|q\|_{-l,G} := \sup_{w\in W^{l,2}(G)} \frac{q(w)}{\|w\|_{l,G}}, \qquad q \in \left(W^{l,2}(G)\right)^* . \tag{10.12}$$

Das wesentliche Hilfsmittel für unsere Zwecke ist das folgende Resultat.

Lemma 10.5. *(Bramble-Hilbert Lemma)*
Sei $G \subset \mathbf{R}^n$ ein beschränktes Gebiet mit Lipschitz-stetigem Rand. Weiterhin sei q ein lineares stetiges Funktional auf $W^{k+1,2}(G)$ mit

$$q(w) = 0, \qquad \forall w \in P_k(G). \tag{10.13}$$

Dann findet man eine nur vom Gebiet G abhängige Konstante $C = C(G)$ mit

$$|q(v)| \le C\|q\|_{-(k+1),G}\ |v|_{k+1,G}, \qquad \forall v \in W^{k+1,2}(G). \tag{10.14}$$

Beweis: (i) Für beliebiges, jedoch fixiertes $v \in W^{k+1,2}(G)$ ermitteln wir ein Polynom $w \in P_k(G)$ mit

$$\int_G D^\alpha (v + w)\ dx = 0, \qquad \forall\ |\alpha| \le k. \tag{10.15}$$

Ausgangspunkt ist die Monom-Basisdarstellung

$$w(x) = \sum_{|\beta|\le k} c_\beta x^\beta, \quad x^\beta := \prod_{i=1}^n x_i^{\beta_i}, \quad c_\beta \in \mathbf{R}$$

mit dem Multiindex $\beta = (\beta_1, ..., \beta_n)$. Aus (10.15) folgt

$$\sum_{|\beta|\le k} c_\beta \int_G D^\alpha x^\beta\ dx = -\int_G D^\alpha v\ dx, \quad |\alpha| \le k, \tag{10.16}$$

d.h. ein lineares Gleichungssytem zur Bestimmung der Koeffizienten c_β, $|\beta| \le k$. Wegen der Eigenschaft $D^\alpha x^\beta = 0$ für alle Multiindizes α, β mit $\alpha_i > \beta_i$ für mindestens ein $i \in \{1, ..., n\}$ ist (10.16) sogar ein

gestaffeltes System. Es kann somit beginnend mit Indizes β mit $\beta_j = k$ für ein $j \in \{1, ..., n\}$ gelöst werden. Für diese gilt

$$c_\beta = \frac{-1}{k!\ \mathrm{meas}(G)} \int_G D^\alpha v\ dx.$$

Alle weiteren Koeffizienten erhält man rekursiv aus (10.16). Somit ist die Existenz des zu v gehörigen Polynoms $w \in P_k(G)$ bewiesen.

(ii) Wir benutzen jetzt die *Poincare–Ungleichung*

$$\|u\|^2_{k+1,G} \le C \left(|u|^2_{k+1,G} + \sum_{|\alpha|\le k} \left| \int_G D^\alpha u\ dx \right|^2 \right), \quad \forall u \in W^{k+1,2}(G),$$

die man analog zu Lemma 5.25 und Satz 5.26 beweisen kann. Dann gilt für das gemäß (i) konstruierte Element $w \in P_k(G)$, daß

$$\|v + w\|^2_{k+1,G} \le C|v + w|^2_{k+1,G} = C|v|^2_{k+1,G}.$$

Die Bedingung (10.13) und die Linearität von q ergeben

$$q(v) = q(v) + q(w) = q(v + w).$$

Damit folgern wir schließlich

$$|q(v)| \le \|q\|_{-(k+1),G}\ \|v + w\|_{k+1,G} \le C\|q\|_{-(k+1),G}\ |v|_{k+1,G}$$

mit einer geeigneten positiven Konstanten C. □

Folgerung 10.6. Durch Hintereinanderausführung des Beweisgedankens des Lemmas von Bramble-Hilbert folgt auch für stetige Bilinearformen $S : W^{k+1,2}(G) \times W^{r+1,2}(G) \to \mathbf{R}$ mit den Eigenschaften

$$(i) \qquad S(u, v) = 0 \quad \forall u \in W^{k+1,2}(G),\ v \in P_r(G)$$

$$(ii) \qquad S(u, v) = 0 \quad \forall u \in P_k(G), v \in W^{r+1,2}(G)$$

die Abschätzung

$$|S(u, v)| \le C\|S\|\ |u|_{k+1,G}\ |v|_{r+1,G} \quad \forall u \in W^{k+1,2}(G), v \in W^{r+1,2}(G). \tag{10.17}$$

Diese Ungleichung wird im Beweis des nachfolgenden Lemmas 10.7 benutzt. □

10.3 Interpolationsfehlerabschätzungen

Wir benutzen nun die Bramble-Hilbert Theorie, um Interpolationsfehlerabschätzungen zu gewinnen.

Lemma 10.7. *Für ein Gebiet $G \subset \mathbf{R}^n$ wie in Lemma 10.6 und natürliche Zahlen $k \ge r$ sei*

$$\Pi :\ W^{k+1,2}(G) \to P_k(G) \subset W^{r,2}(G)$$

ein linearer, stetiger Projektionsoperator. Dann existiert eine Konstante $C > 0$ mit

$$\|v - \Pi v\|_{r,G} \le C\|I - \Pi\|_{\mathcal{L}(W^{k+1,2}(G),W^{r,2}(G))}\ |v|_{k+1,G} \quad \forall v \in W^{k+1,2}(G).$$

Beweis: Wir definieren die in Folgerung 10.6 eingeführte Bilinearform S durch

$$S(u, v) := (u - \Pi u, v - \Pi v)_{r,G} \quad \forall u, v \in W^{k+1,2}(G)$$

mit dem Skalarprodukt $(\cdot, \cdot)_{r,G}$ auf dem Hilbert-Raum $W^{r,2}(G)$. Es sind die Voraussetzungen von Folgerung 10.6 zu prüfen. Die Ungleichung von Cauchy-Schwarz zeigt

$$|S(u, v)| \le \|u - \Pi u\|_{r,G}\ \|v - \Pi v\|_{r,G}.$$

Mit der Operatornorm

$$\|\cdot\| := \|\cdot\|_{\mathcal{L}(W^{k+1,2}(G),W^{r,2}(G))}$$

erhalten wir dann

$$|S(u,v)| \le \|I-\Pi\|^2 \; \|u\|_{k+1,G} \; \|v\|_{k+1,G} \quad \forall u,v \in W^{k+1,2}(G),$$

folglich ist $\|S\| \le \|I-\Pi\|^2$, d.h. S ist beschränkt.

Nach Definition des Projektors ist $(I-\Pi)v = 0$ für alle $v \in P_k$. Daraus folgt

$$S(u,v) = 0 \quad \forall u \in W^{k+1,2}(G), v \in P_k(G) \;\; \text{bzw.} \;\; \forall u \in P_k(G), v \in W^{k+1,2}(G).$$

Damit ist Folgerung 10.6 anwendbar und wir erhalten die Behauptung über

$$\|u-\Pi u\|_{r,G}^2 = |S(u,u)| \le C\|I-\Pi\|^2 \; |u|_{k+1,G}^2 \quad \forall u \in W^{k+1,2}(G). \qquad \square$$

Für die gesuchten Interpolationsfehlerabschätzungen definieren wir einen geeigneten *Projektionsoperator* in den Ansatzraum X_h: Wir nehmen an, daß X_h mittels eines einzigen Referenzelementes $(\tilde{K}, \tilde{\mathcal{P}}, \tilde{\Sigma})$ durch affin-lineare Abbildung erzeugt wird, d.h. für jedes finite Element $(K, \mathcal{P}, \Sigma)$ existiert eine nichtentartete affin-lineare Abbildung $F^{-1} : K \to \tilde{K}$, so daß

$$(i) \;\; F^{-1}(K) = \tilde{K}, \qquad (ii) \;\; F^*\tilde{\mathcal{P}} = \mathcal{P}, \qquad (iii) \;\; F_*\Sigma = \tilde{\Sigma} \tag{10.18}$$

mit

$$F^*\tilde{v} := \tilde{v} \circ F^{-1}, \qquad (F_*N)(\tilde{v}) := N(F^*(\tilde{v})) = N(\tilde{v} \circ F^{-1}).$$

In der letzten Formelzeile ist N ein Funktional aus Σ.

Somit kann man den in Abschnitt 8.3 eingeführten lokalen Interpolationsoperator Π auch auf $\tilde{K}$ erklären. Der globale Interpolationsoperator $\Pi_{\mathcal{T}}$ über der Zerlegung $\mathcal{T}_h$ wird stückweise durch Hintereinanderausführung der Interpolation Π auf $\tilde{K}$ und der zugehörigen Abbildung $F_j^* : \tilde{K} \to K_j$ bestimmt. Nach Lemma 8.9 ist $\Pi_{\mathcal{T}}$ ein (stückweise definierter) Projektor in den Raum X_h, d.h.

$$(I-\Pi_{K_j})v = 0 \quad \forall v|_{K_j} \in P_k(K_j) \qquad \forall j = 1,...,M.$$

Wir benutzen, daß bei affin-linearen Transformationen F_j Polynome auf dem Referenzelement $\tilde{K}$ auf Polynome über K_j übergehen.

Wir wenden nun Lemma 10.7 auf das Referenzelement $G = \tilde{K}$ an und berechnen dann hierfür die Norm des Operators $I - \Pi : \; W^{k+1,2}(G) \to W^{r,2}(G)$. Da $\tilde{K}$ unabhängig von $\mathcal{T}_h$ ist, kann $\|I-\Pi\|$ unabhängig von $\mathcal{T}_h$ ermittelt werden.

Satz 10.8. *Die zulässige Zerlegung $\mathcal{T}_h$ des Gebietes $\Omega \subset \mathbf{R}^2$ sei durch affin-lineare Transformation von einem Referenzelement $\tilde{K}$ erzeugt. Sei $\Pi_{\mathcal{T}} : \; W^{k+1,2}(\Omega) \to P_k^{\mathcal{T}}(\Omega) \subset W^{r,2}(\Omega)$ der über $\mathcal{T}_h$ definierte globale Projektor in die Menge $P_k^{\mathcal{T}}(\Omega)$ der stückweise polynomialen Funktionen vom Grad k. Ferner seien $r \in \mathbf{N}_0, \; k \in \mathbf{N}$ mit $r \le k$.*

(i) Dann existiert eine Konstante $C > 0$ mit der lokalen Interpolationsaussage

$$\|u - \Pi_{\mathcal{T}} u\|_{r,K} \le C \; h_K^{k+1} \; \rho_K^{-r} \; |u|_{k+1,K}. \tag{10.19}$$

(ii) Für eine isotrope Zerlegung gilt die globale Interpolationsaussage

$$\|u - \Pi_{\mathcal{T}} u\|_{r,\Omega} \le C \left(\sum_K h_K^{2(k+1-r)} \; |u|_{k+1,\Omega}^2 \right)^{1/2}.$$

(iii) Für eine quasi-uniforme Zerlegung gilt die globale Interpolationsaussage

$$\|u - \Pi_{\mathcal{T}} u\|_{r,\Omega} \le C \; h^{k+1-r} \; |u|_{k+1,\Omega}.$$

Beweis: Wir betrachten die Zerlegung $\mathcal{T}_h = \{K_j\}_{j=1}^M$ des Gebietes Ω. Wegen

$$\|u - \Pi_{\mathcal{T}} u\|_{r,\Omega}^2 = \sum_{j=1}^{M} \|u - \Pi_{K_j} u\|_{r,K_j}^2. \tag{10.20}$$

können wir den Fehler für jedes Teilgebiet $K = K_j$ abschätzen. Ferner sei $\tilde{K}$ das einheitliche Referenzelement der Zerlegung. Lemma 10.7 ergibt

$$\|u - \Pi u\|_{r,\tilde{K}} \leq C\|I - \Pi\| \; |u|_{k+1,\tilde{K}}.$$

Die Abschätzung (10.10) aus Folgerung 10.6 liefert

$$\|u - \Pi u\|_{r,\tilde{K}} \leq C\|I - \Pi\| \; h_K^{k+1} \left(\inf_{p \in \tilde{K}} |F'(p)| \right)^{-1/2} \; |u|_{k+1,K}.$$

Sei o.B.d.A. $\rho_K \leq 1$. Dann ist ggf. nach Mehrfachanwendung von (10.10)

$$\|u - \Pi u\|_{r,\tilde{K}} \geq C \left(\sup_{p \in \tilde{K}} |F'(p)| \right)^{-1/2} \rho_K^r \|u - \Pi u\|_{r,K}.$$

Aus beiden Ungleichungen folgt unter Beachtung von (10.11) für den lokalen Fehler

$$\|u - \Pi u\|_{r,K} \leq C\|I - \Pi\| \; h_K^{k+1} \; \rho_K^{-r} \; |u|_{k+1,K}.$$

Gleichung (10.20) ergibt die globalen Aussagen (ii) bzw. (iii) im isotropen bzw. quasi-uniformen Fall. □

Bemerkungen 10.9. (i) Die Aussage von Satz 10.8 gilt allgemeiner für simpliziale Zerlegungen auf beschränkten, polyedrischen Gebieten im $\mathbf{R}^n$.

(ii) Es ist ferner möglich, auch lokale Interpolationsabschätzungen in Sobolev-Räumen $W^{r,q}(K)$ zu gewinnen. So gilt

$$\|u - \Pi u\|_{W^{r,q}(K)} \leq C(\mathrm{meas}(K))^{1/q-1/p} h_K^{k+1} \rho_K^{-r} |u|_{W^{k+1,p}(K)},$$

sofern für Zahlen $m, k \in \mathbf{N}_0$ und $p, q \in [1, \infty]$ die Einbettung $W^{k+1,p}(\tilde{K}) \subseteq W^{r,q}(\tilde{K})$ auf dem Referenzelement $\tilde{K}$ stetig ist.

(iii) Satz 10.8 fordert, daß eine zu interpolierende Funktion u wenigstens in $W^{2,2}(\Omega)$ liegt. Ist $u \in W^{1,2}(\Omega)$ etwa verallgemeinerte Lösung eines elliptischen RWP 2. Ordnung, so ist diese Forderung oft nicht realistisch (vgl. Kapitel 6). Für eine global stetige Interpolation in X_h reicht jedoch nach dem Einbettungssatz von Sobolev (vgl. Satz 6.25) auch die Forderung $u \in W^{1,p}(\Omega),\ p > n$ zur Sicherung der Wohldefiniertheit von $\Pi_{\mathcal{T}}$ aus.

Gilt dies nicht, kann man regularisierende Interpolationsoperatoren nutzen. Man verliert jedoch den strikt elementweisen Charakter der Abschätzungen. Ein typisches Beispiel ist der Operator von *Clement*. Eine erste Information findet man bei D. Braess [5], Kap. II.6.9. □

10.4 Fehlerabschätzungen in der $X-$Norm

Die in Abschnitt 10.3 bereitgestellten Interpolationsabschätzungen werden jetzt für Fehlerabschätzungen bei konformen FEM benutzt. Sei $\Omega \subset \mathbf{R}^n$ ein polyedrisches Gebiet. Ausgangspunkt ist die elliptische Variationsgleichung

$$\text{Finde } \; u \in X: \quad a(u,v) = f(v) \quad \forall v \in X \tag{10.21}$$

mit stetigem linearen Funktional $f : X \to \mathbf{R}$ und stetiger, $X-$elliptischer Bilinearform $a : X \times X \to \mathbf{R}$. Dabei gelte die Einbettung $X \subset W^{m,2}(\Omega)$ mit $m \in \mathbf{N}$. Für RWP 2. Ordnung ist $m = 1$, jedoch gelten

die nachfolgenden Aussagen auch für elliptische RWP der Ordnung $2m \geq 2$.

Wir betrachten eine konforme FEM

$$\text{Finde } u_h \in X_h: \quad a(u_h, v_h) = f(v_h) \quad \forall v_h \in X_h \tag{10.22}$$

über einer quasi-uniformen Zerlegung $\mathcal{T}_h$ des Gebietes Ω mit einem einheitlichen Referenzelement $\tilde{K}$. Der diskrete Ansatzraum X_h sei so konstruiert, daß mit Hilfe des in Abschnitt 8.3 erklärten C^m-Finite-Elemente-Raumes $X_{\mathcal{T}}$ mit stückweise polynomialen Funktionen vom Grad $k \in \mathbf{N}$ gilt

$$X_h = X_{\mathcal{T}} \cap X \ (\subset X). \tag{10.23}$$

Die Bedingung $k \geq m$ ergibt sich aus dem Einbettungssatz von Sobolev (vgl. Satz 6.25). Dann gilt folgende Fehlerabschätzung:

Satz 10.10. *Die Lösung u der Variationsgleichung (10.21) sei regulär gemäß*

$$u \in X \cap W^{k+1,2}(\Omega), \qquad k \geq m. \tag{10.24}$$

Ferner seien die Voraussetzungen von Satz 10.8 erfüllt. Dann ist das diskrete Problem (10.22) eindeutig lösbar. Für den Fehler gilt

$$\|u - u_h\|_{m,\Omega} \leq C h^{k+1-m} \ |u|_{k+1,\Omega}. \tag{10.25}$$

Beweis: Wegen der Konformitätsbedingung $X_h \subset X$ übertragen sich Stetigkeit und Elliptizität der Bilinearform a auf X_h. Damit ist das Lemma von Lax-Milgram (vgl. Satz 7.12) anwendbar. Weiter liefert das Lemma von Cea (vgl. Satz 7.13)

$$\|u - u_h\|_X \leq C \inf_{v \in X_h} \|u - v\|_X \leq C \|u - \Pi_{\mathcal{T}} u\|_X$$

über Satz 10.9 und Voraussetzung (10.24) die Behauptung (10.25). □

Exemplarisch behandeln wir

Beispiel 10.11. *Dirichlet-Problem elliptischer Gleichungen 2. Ordnung*

Für $m = 1$ sei $X = W_0^{1,2}(\Omega)$. Das in Abschnitt 6.2.1 behandelte Problem (6.7),(6.8) ist unter den Voraussetzungen von Satz 6.12 eindeutig lösbar. Unter der zusätzlichen Regularitätsannahme $u \in W^{k+1,2}(\Omega)$ mit $k \geq 1$ gilt

$$\|u - u_h\|_{1,\Omega} \leq C h^k \ |u|_{k+1,\Omega}.$$

In Kapitel 6.3 hatten wir speziell hinreichende Bedingungen für die Existenz verallgemeinerter Ableitungen von u, d.h. für $k = 1$, hergeleitet. Insbesondere sind sie für das homogene Dirichlet-Problem der Poisson-Gleichung in einem konvexen polyedrischen Gebiet erfüllt. □

Bemerkungen 10.12. (i) Die obigen Fehlerabschätzungen bedürfen eines kritischen Kommentars. Einerseits ist die explizite Berechnung oder Abschätzung der auf der rechten Seite auftretenden Konstanten C schwierig. Andererseits tritt die Seminorm $|u|_{k+1,\Omega}$ der (unbekannten !) Lösung auf. Man kann versuchen, diese durch Problemdaten abzuschätzen.

(ii) Fehlerabschätzungen lassen sich für andere Randwertprobleme herleiten. Dabei werden Randbedingungen entweder (näherungsweise) in den diskreten Lösungsraum oder in die Variationsformulierung eingearbeitet. Man benötigt dann in der Regel auch Interpolationsabschätzungen auf dem Gebietsrand bzw. auf Teilmengen. □

Die Regularitätsvoraussetzung (10.24) ist in vielen Fällen nicht realistisch. Oft werden bereits durch Ecken eines polygonalen Gebietes Singularitäten der Ableitungen der Lösung verursacht. Wir werden daher versuchen, die Konvergenz des Diskretisierungsverfahrens ohne zusätzliche Regularitätsforderung zu beweisen.

Satz 10.13. *Für ein beschränktes polyedrisches Gebiet $\Omega \subset \mathbf{R}^n$ gelte $X \subset W^{1,2}(\Omega)$ und der $W^{2,2}(\Omega)$ sei dicht in X bezüglich der $\|\cdot\|_{1,\Omega}$–Norm eingebettet. Bei quasi-uniformer Zerlegung $\mathcal{T}$ des Gebietes Ω*

bestehe der diskrete Ansatzraum $X_h \subset X$ aus stückweise linearen Ansatzfunktionen. Für das Problem (10.21) seien die Voraussetzungen von Satz 10.8 erfüllt. Dann gilt für den Fehler des Verfahrens (10.22)

$$\lim_{h \to +0} \|u - u_h\|_{1,\Omega} = 0.$$

Beweis: Nach dem Lemma von Cea (vgl. Satz 7.13) und mittels Dreiecksungleichung ergibt sich mit zunächst freiem $w \in X$, daß

$$\|u - u_h\|_{1,\Omega} \leq C \inf_{v \in X_h} \|u - v\|_{1,\Omega} \leq C \left(\|u - w\|_{1,\Omega} + \|w - \Pi_{\mathcal{T}} w\|_{1,\Omega} \right).$$

Sei nun $\epsilon > 0$ beliebig. Wegen der Dichtheit von $W^{2,2}(\Omega)$ in X gibt es ein $w \in W^{2,2}(\Omega)$ mit

$$\|u - w\|_{1,\Omega} \leq \epsilon.$$

Nach Satz 10.8 gilt für das zu w gehörende Element $\Pi_{\mathcal{T}} w \in X_h$, daß

$$\|w - \Pi_{\mathcal{T}} w\|_{1,\Omega} \leq Ch|w|_{2,\Omega} \leq \epsilon$$

bei hinreichend kleinem $h > 0$. Damit ist aber $\|u - u_h\|_{1,\Omega} \leq 2C\epsilon$. Da $\epsilon > 0$ beliebig gewählt werden kann, folgt die Behauptung. □

10.5 Weitere Fehlerabschätzungen

(i) Fehlerabschätzungen in der $H-$Norm

Wir werden jetzt noch ausgehend vom Satz 7.11 Abschätzungen in der $H-$Norm angeben, falls für die Hilbert-Räume X und H die stetige Einbettung $X \subset H$ gilt. Die Fehlerabschätzung in der $W^{1,2}(\Omega)-$Norm für das homogene Dirichlet-Problem elliptischer Randwertprobleme 2. Ordnung (vgl. Beispiel 10.10) impliziert zwar bereits eine Abschätzung in der Norm von $H = L^2(\Omega)$. Diese ist jedoch offenbar unter Beachtung der Interpolationsabschätzung von Satz 10.7 nicht optimal. Der folgende Satz basiert auf einem *Dualitätsargument ("Aubin-Nitsche-Trick").*

Satz 10.14. *Gelte $X \subset W^{1,2}(\Omega)$ mit stetiger Einbettung. Die Lösung der Variationsgleichung (10.21) genüge der Regularitätsannahme (10.24) mit $m = 1$. Weiter seien die Voraussetzungen von Satz 10.10 erfüllt. Darüber hinaus besitze die zu (10.21) adjungierte Aufgabe*

$$a^*(w_g, v) := a(v, w_g) = (g, v)_H \quad \forall v \in X \tag{10.26}$$

für beliebige $g \in H := L^2(\Omega)$ eine eindeutig bestimmte Lösung $w_g \in X \cap W^{2,2}(\Omega)$ unter der zusätzlichen Forderung

$$|w_g|_{2,\Omega} \leq C\|g\|_{0,\Omega}. \tag{10.27}$$

Dann gilt für den Fehler des diskreten Problems (10.22) die Abschätzung

$$\|u - u_h\|_{0,\Omega} \leq Ch^{k+1}|u|_{k+1,\Omega}. \tag{10.28}$$

Beweis: Die Einbettung $X \subset H$ ist nach Voraussetzung stetig. Dann ist Satz 7.11 anwendbar und es gilt

$$\|u - u_h\|_H := \|u - u_h\|_{0,\Omega} \leq M\|u - u_h\|_X \sup_{g \in H} \frac{\inf_{\phi \in X_h} \|w_g - \phi\|_X}{\|g\|_H}. \tag{10.29}$$

Zur Abschätzung des letzten Terms in (10.29) betrachten wir jetzt das adjungierte Problem (10.26). Das ergibt in Verbindung mit Satz 10.10 und (10.27)

$$\inf_{\phi \in X_h} \|w_g - \phi\|_X \leq Ch|w_g|_{2,\Omega} \leq Ch\|g\|_{0,\Omega}.$$

In Verbindung mit (10.29) und Satz 10.10 folgt die Behauptung. □

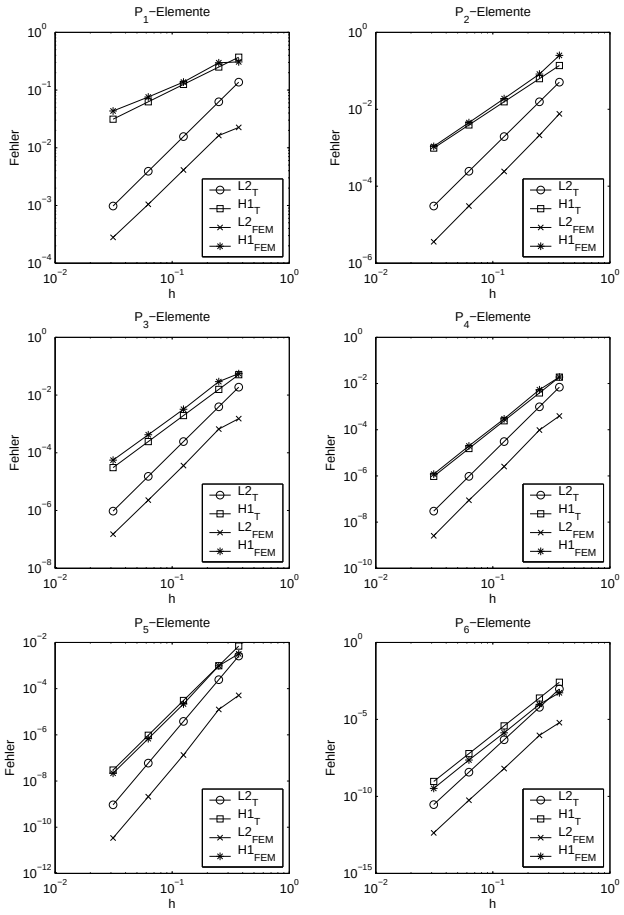


Abbildung 10.2: Fehlerdiagramme in der $W^{1,2}$- und L^2-Norm für Beispiel 10.15

Beispiel 10.15: Für das homogene 1. RWP der Poisson-Gleichung $-(\Delta u)(x) = f(x),\ x \in \Omega = (0,1) \times (0,1)$ sei die rechte Seite f so bestimmt, daß die Lösung gegeben ist durch $u(x) = \sin(\pi x_1)\sin(\pi x_2)e^{-x_1 x_2}$. Mit FEMLAB wurde die Lösung mit P_k-Elementen für $k \in \{1,\ldots,6\}$ auf einer Sequenz unstrukturierter quasi-uniformer Gitter von $h = \frac{1}{3}$ bis $h = \frac{1}{32}$ approximiert. Abb. 10.2 zeigt die Konvergenzdiagramme für die $W^{1,2}$- bzw. L^2-Norm. Die nach den Sätzen 10.8 bzw. 10.10 theoretisch erreichbaren Konvergenzordnungen h^k bzw. h^{k+1} werden praktisch erreicht. Ein Vergleich der Resultate für wachsendes k bei festem h zeigt den enormen Konvergenzgewinn bei Verwendung einer p-Methode im Fall glatter Lösungen. □

(ii) Fehlerabschätzungen in der L^∞-Norm

Wünschenswert sind auch für Finite-Elemente-Verfahren scharfe Abschätzungen in der L^∞-Norm. Deren Ableitung ist allerdings technisch recht kompliziert (vgl. dazu S. Brenner, R. Scott [4]). Wir zitieren lediglich folgendes Resultat für das homogene Dirichlet–Problem der Poisson-Gleichung.

Satz 10.16. *Sei $\mathcal{T}_h$ eine reguläre Triangulation des polyedrischen Gebietes $\Omega \subset \mathbf{R}^2$ durch Dreiecke. Bei Verwendung stückweise polynomialer Ansatzfunktionen vom Grad k gilt unter der Regularitätsannahme $u \in W_0^{1,2}(\Omega) \cap W^{k+1,\infty}(\Omega)$,* daß

$$\|u - u_h\|_{L^\infty(\Omega)} \leq \begin{cases} Ch^2|\log h|\ |u|_{W^{2,\infty}(\Omega)}, & \text{falls } k = 1 \\ Ch^{k+1}\ |u|_{W^{k+1,\infty}(\Omega)}, & \text{falls } k \geq 2 \end{cases}.$$

Kapitel 11

Nichtkonforme Finite-Elemente-Methoden

Im folgenden Kapitel wollen wir die Bedingung der Konformität $X_h \subset X$ an die Finite-Elemente-Räume abschwächen. Ferner soll erlaubt sein, daß die Variationsgleichung im diskreten Fall (z.B. bei numerischer Integration) abgeändert wird. Wir sprechen dann von *nichtkonformen Methoden.*

11.1 Begriffsbildung

Bisher haben wir für Variationsgleichungen

$$\text{Finde } \; u \in X: \quad a(u,v) = f(v) \quad \forall v \in X \tag{11.1}$$

lediglich Diskretisierungen in endlich–dimensionalen Unterräumen mit $X_h \subset X$ betrachtet, wobei das Funktional $f : X \to \mathbf{R}$ und die Bilinearform $a : X \times X \to \mathbf{R}$ nicht abgeändert wurden.

Oftmals ist es jedoch wünschenswert, von wenigstens einer dieser Annahmen abzuweichen. Dies ist z.B. sinnvoll, wenn a bzw. f durch numerische Integration (vgl. Abschn. 11.2) ausgewertet werden bzw. inhomogene wesentliche Randbedingungen oder ein krummliniger Rand $\partial\Omega$ keine exakte Erfüllung der Randbedingungen im diskreten Fall (vgl. Abschn. 11.3) erlauben. Oft findet man in der Literatur Arbeiten, in denen diese Betrachtungen unterbleiben (z. B. Annahme exakter Integration). Man hat dafür die etwas drastische Bezeichnung *"variational crimes"* geprägt.

Eine weitere Möglichkeit nichtkonformer Methoden entsteht, wenn (vor allem bei Randwertproblemen höherer Ordnung) die Einbettungsforderung $X_h \subset X$ zu sehr aufwendigen Elementen führen würde (vgl. Abschn. 11.4).

Wir werden fortan ein diskretes Problem

$$\text{Finde } \; u_h \in X_h: \quad a_h(u_h,v_h) = f_h(v_h) \quad \forall v_h \in X_h \tag{11.2}$$

betrachten. Dabei sei X_h Hilbert-Raum mit der Norm $\|\cdot\|_h$. Weiter seien $a_h : X_h \times X_h \to \mathbf{R}$ eine stetige Bilinearform, die gleichmäßig X_h–elliptisch ist, d.h. es gibt eine von der Diskretisierung unabhängige Konstante $\tilde{\gamma} > 0$ mit

$$a_h(v_h,v_h) \geq \tilde{\gamma}\|v_h\|_h^2 \quad \forall v_h \in X_h. \tag{11.3}$$

Außerdem sei $f_h : X_h \to \mathbf{R}$ ein stetiges lineares Funktional.

Im allgemeinen Fall sind a_h und f_h nicht auf $X \times X$ bzw. X erklärt. Wir führen daher normierte Räume Z mit der Norm $|||\cdot|||$ und Z_h mit den stetigen Einbettungen

$$X \subset Z, \qquad X_h \subset Z_h \subset Z$$

ein, so daß a_h und f_h nun auf $Z_h \times Z_h$ bzw. Z_h definiert und stetig sind. Dann gewinnen wir die folgende *abstrakte* Fehlerabschätzung, bei der die Wahl der Räume Z und Z_h zunächst offen bleibt.

Satz 11.1. *Seien die eingeführten Voraussetzungen erfüllt. Dann besitzt das diskrete Problem (11.2) eine eindeutige Lösung $u_h \in X_h$. Ferner gilt für den Diskretisierungsfehler des Verfahrens (11.2) die Abschätzung*

$$|||u - u_h||| \leq C \inf_{z_h \in X_h} (|||u - z_h||| + \|f_h - a_h(z_h, \cdot)\|_{*,h}) \tag{11.4}$$

mit der zu $\|\cdot\|_h$ gehörenden Dualnorm $\|\cdot\|_{*,h}$ gemäß

$$\|w\|_{*,h} := \sup_{v_h \in X_h \setminus \{0\}} \frac{|w(v_h)|}{\|v_h\|_h} \quad \forall w \in X_h^*.$$

Beweis: Unter den getroffenen Voraussetzungen ist auf das nichtkonforme Problem (11.2) das Lemma von Lax–Milgram betreffs Existenz und Eindeutigkeit der Lösung $u_h \in X_h$ anwendbar.

Zur Fehlerabschätzung zeigen wir eine Variante des Lemmas von Cea: Unter Beachtung von (11.2) gilt für beliebige $z_h \in X_h$ die Gleichung

$$a_h(u_h - z_h, v_h) = f_h(v_h) - a_h(z_h, v_h) \quad \forall v_h \in X_h. \tag{11.5}$$

Für fixiertes $z_h \in X_h$ ist $f_h - a_h(z_h, \cdot) \in X_h^*$. Damit erhalten wir unter Benutzung der X_h-Elliptizität

$$\tilde{\gamma}\|u_h - z_h\|_h^2 \leq \|f_h - a_h(z_h, \cdot)\|_{*,h} \, \|u_h - z_h\|_h,$$

also

$$\|u_h - z_h\|_h \leq \frac{1}{\tilde{\gamma}} \|f_h - a_h(z_h, \cdot)\|_{*,h}.$$

Wegen der stetigen Einbettung $X_h \subset Z$ liefert dann die Dreiecksungleichung

$$|||u - u_h||| \leq C \, (|||u - z_h||| + \|f_h - a_h(z_h, \cdot)\|_{*,h})$$

für beliebige $z_h \in X_h$. Das ist die Behauptung. □

11.2 Numerische Integration

Wir untersuchen jetzt den *Einfluß numerischer Integration* bei der Näherungsberechnung von a bzw. f, d.h. deren Ersetzung durch a_h bzw. f_h in (11.1). Vereinfachend gelte $X_h \subset X$. Daher gilt $Z := X$ und $|||\cdot||| := \|\cdot\|_X = \|\cdot\|_h$.

Eine geringe Modifizierung von Satz 11.1 durch genauere Auswertung der Approximationsgüte von f_h an f bzw. von a_h an a liefert der

Satz 11.2. *(1. Lemma von G. Strang)*
Gelte $X_h \subset X$. Ferner sei die Bilinearform $a_h : X_h \times X_h \to \mathbf{R}$ gleichmäßig X_h-elliptisch. Dann gilt für den Fehler des diskreten Problems (11.2)

$$\|u - u_h\|_X \leq C \inf_{z_h \in X_h} (\|u - z_h\|_X + \|a(z_h, \cdot) - a_h(z_h, \cdot)\|_{*,h} + \|f - f_h\|_{*,h}) \, . \tag{11.6}$$

Beweis: Zunächst folgt aus (11.1) und (11.2) für beliebige $z_h \in X_h$ die Fehlergleichung

$$a_h(u_h - z_h, v_h) = a(u, v_h) - a_h(z_h, v_h) + f_h(v_h) - f(v_h), \quad \forall v_h \in X_h.$$

Mit $v_h = u_h - z_h$ finden wir über Nullergänzung

$$\begin{aligned} \tilde{\gamma}\|u_h - z_h\|_X^2 \quad \leq \quad & M\|u - z_h\|_X \|u_h - z_h\|_X + \|f_h - f\|_{*,h} \|u_h - z_h\|_X \\ & + \|a(z_h, \cdot) - a_h(z_h, \cdot)\|_{*,h} \|u_h - z_h\|_X. \end{aligned} \tag{11.7}$$

Die Dreiecksungleichung ergibt dann die Behauptung. □

Der übliche Zugang der numerischen Integration ist die Anwendung von *Quadraturformeln* der Form

$$\int_\Omega g(x)\ dx = \sum_{j=1}^{M} \int_{K_j} g(x)\ dx = \sum_{j=1}^{M} \sum_{l=1}^{L} w_{l,j} g(z_{lj}) \tag{11.8}$$

zur Berechnung der in a bzw. f auftretenden Integrale mit geeigneten Gewichten w_{lj} und Integrationspunkten z_{lj}. Für das Modellproblem

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) = f(x) \quad x \in \Omega, \qquad u(x) = 0 \quad x \in \partial\Omega$$

mit $f \in W^{k,q}(\Omega),\ q \geq 2$ und $a_{ij} \in W^{k,\infty}(\Omega),\ k \geq n/q$ für $i,j = 1,\ldots,n$ gilt zum Beispiel folgendes Resultat.

Satz 11.3. *Der Raum X_h bestehe aus stückweise polynomialen Ansatzfunktionen vom Grad $k \in \mathbf{N}$. Ferner sei die numerische Integrationsformel (11.8) exakt für stückweise polynomiale Funktionen vom Grad $2k-2$. Dann gilt für den Diskretisierungsfehler des Galerkin-Verfahrens*

$$\|u - u_h\|_{1,\Omega} \leq Ch^k \left(\|u\|_{k+1,\Omega} + \|f\|_{k,\Omega}\right).$$

Beweis: vgl. Brenner/Scott [4]. □

Wir wählen hier für ein allgemeineres Problem einen *anderen Zugang*, der direkt auf dem 1. Lemma von Strang basiert und ein übersichtliches Beweiskonzept erlaubt. Sei $X := W_0^{1,2}(\Omega)$. Wir betrachten die Variationsformulierung eines homogenen Dirichletschen Randwertproblems 2. Ordnung (11.1) mit

$$a(u,v) := \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} v + cuv\right) dx \tag{11.9}$$

$$f(v) := \int_\Omega fv\ dx \tag{11.10}$$

sowie die nichtkonforme Approximation (11.2) mit

$$a_h(u,v) := \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}^h\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \sum_{i=1}^{n} b_i^h \frac{\partial u}{\partial x_i} v + c^h uv\right) dx \tag{11.11}$$

$$f_h(v) := \int_\Omega f^h v\ dx \tag{11.12}$$

mit

$$a_{ij}^h := \Pi_{k-1}^h a_{ij},\quad b_i^h := \Pi_{k-1}^h b_i,\quad c^h := \Pi_{k-1}^h c,\quad f^h := \Pi_{k-1}^h f,\quad i,j = 1,...,n. \tag{11.13}$$

Dabei ist Π_m^h bei $m \in \mathbf{N}$ der in Kapitel 10.3 eingeführte stückweise polynomiale Interpolationsoperator vom Grad m. Für $m = 0$ werden die Daten stückweise konstant auf den finiten Elementen approximiert, etwa durch den arithmetischen Mittelwert der Funktionswerte in den Eckpunkten des Elements.

Wir beweisen den

Satz 11.4. *Neben den Voraussetzungen des Existenzsatzes 6.10 gelte für die Daten $a_{ij}, b_i, c, \in W^{k,\infty}(\Omega)$, $f \in W^{k,2}(\Omega)$ mit $k \in \mathbf{N}$. Dann besitzt das diskrete Problem (11.2) mit (11.11)-(11.13) für hinreichend feine Vernetzung mit $0 < h \leq h_0$ eine und nur eine Lösung $u_h \in X_h \subset X$. Für den Diskretisierungsfehler gilt die Abschätzung*

$$\|u - u_h\|_{1,\Omega} \leq C\left(\inf_{z_h \in X_h} \|u - z_h\|_{1,\Omega} + Ch^k\right). \tag{11.14}$$

Beweis: Zunächst benutzen wir die Interpolationsabschätzungen für die Daten

$$\|a_{ij} - a_{ij}^h\|_{L^\infty(\Omega)} \le Ch^k \|a_{ij}\|_{W^{k,\infty}(\Omega)}; \quad \|b_i - b_i^h\|_{L^\infty(\Omega)} \le Ch^k \|b_i\|_{W^{k,\infty}(\Omega)};$$

$$\|c - c^h\|_{L^\infty(\Omega)} \le Ch^k \|c\|_{W^{k,\infty}(\Omega)}; \quad \|f - f^h\|_{L^2(\Omega)} \le Ch^k \|f\|_{W^{k,2}(\Omega)}.$$

Unter Verwendung der Hölderschen Ungleichung schätzen wir die Konsistenzfehler ab durch

$$|f(v) - f_h(v)| \le \|f - f^h\|_{0,\Omega} \|v\|_{0,\Omega} \le Ch^k \|f\|_{W^{k,2}(\Omega)} \|v\|_{0,\Omega} \tag{11.15}$$

bzw. nach kurzer Rechnung

$$\begin{aligned} & |a(u,v) - a_h(u,v)| \\ \le\ & Ch^k \left(\max_{i,j=1,\dots n} \|a_{ij}\|_{W^{k,\infty}(\Omega)} \|u\|_{1,\Omega} \|v\|_{1,\Omega} \right. \\ +\ & \left. \max_{i=1,\dots n} \|b_i\|_{W^{k,\infty}(\Omega)} \|u\|_{1,\Omega} \|v\|_{0,\Omega} + \|c\|_{W^{k,\infty}(\Omega)} \|u\|_{0,\Omega} \|v\|_{0,\Omega} \right) \\ \le\ & Ch^k \|u\|_{1,\Omega} \|v\|_{1,\Omega}. \end{aligned} \tag{11.16}$$

Dann gilt unter Beachtung der $X-$Elliptizität von a, daß

$$a_h(v,v) \ge a(v,v) - |a(v,v) - a_h(v,v)| \ge \gamma \|v\|_{1,\Omega}^2 - Ch^k \|v\|_{1,\Omega}^2$$

für beliebige $v \in X_h$.

Für hinreichend kleine Werte $h \le h_0$ ergibt sich die strikte $X-$Elliptizität von a_h. Das sichert nach dem Lemma von Lax-Milgram die eindeutige Lösbarkeit von (11.2). Die Konsistenzabschätzungen (11.15) und (11.16) liefern zusammen mit dem 1. Lemma von Strang die Behauptung. □

Bei stückweise polynomialen Ansatzfunktionen kann man nun zur Berechnung der im diskreten Problem (11.2) mit (11.11)-(11.13) entstehenden Integrale (mit polynomialen Integranden) exakte Quadraturformeln vom Typ (11.8) heranziehen. Oft verwendet man Formeln vom Gauß- bzw. Lobatto-Typ. Dazu verweisen wir auf A. Quarteroni/ A. Valli [18].

11.3 Approximation krummliniger Ränder

Eine Abänderung der Variationsformulierung (11.1) ist auch bei krummlinigem Rand des Gebietes Ω erforderlich, da dieser sich nicht exakt durch polyedrische finite Elemente beschreiben läßt.

Als erste Variante betrachten wir eine polyedrische Approximation des Gebietes. Exemplarisch untersuchen wir für das homogene Dirichlet-Problem einer elliptischen Gleichung 2. Ordnung den Fall, daß bei Zerlegung des Gebietes $\Omega \subset \mathbf{R}^2$ durch Dreiecke gilt $\overline{\Omega_h} := \cup_j \overline{K_j} \subseteq \overline{\Omega}$. Die Konformitätsbedingung $X_h \subseteq X := W_0^{1,2}(\Omega)$ erreicht man durch Nullfortsetzung der Testfunktionen v auf $\Omega \setminus \Omega_h$. Man kann dann noch folgende Interpolationsaussage beweisen:

Lemma 11.5 *Sei $\Omega_h \subseteq \Omega$ ein approximierendes Polygongebiet und gelte*

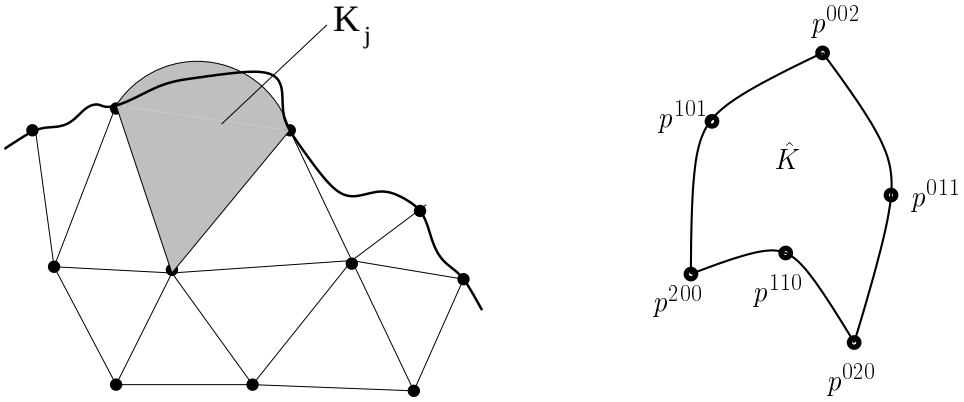
$$\Omega \setminus \Omega_h \subseteq S_\delta := \{x \in \Omega : \ dist(x, \partial\Omega) \le \delta\}$$

mit $\delta := max\{dist(x, \partial\Omega) : \ x \in \Omega \setminus \Omega_h\}$. Unter der Glättevoraussetzung $u \in W^{2,2}(\Omega) \cap W^{1,\infty}(\Omega)$ gilt dann

$$\inf_{v \in X_h} |u - v|_{1,\Omega} \le C \left(h|u|_{W^{2,2}(\Omega_h)} + \sqrt{\delta}|u|_{W^{1,\infty}(\Omega \setminus \Omega_h)} \right).$$

Beweis: Übungsaufgabe ! □

Ist Ω konvex und $\partial\Omega \in C^{1,1}$, so gilt $\delta = \mathcal{O}(h^2)$. Die Interpolationsaussage ist somit lediglich für stückweise lineare Ansatzfunktionen optimal. Diese erste Variante wird auch in der im Übungsbetrieb verwendeten



a) Approximation eines krummlinigen Randes b) Isoparametrisches Element

Abbildung 11.1: Randapproximation und isoparametrische Elemente

Version 2.1 des Programms FEMLAB, die bislang mit stückweise linearen Elementen arbeitet, benutzt. Für Finite-Elemente-Räume mit stückweise polynomialen Ansatzfunktionen höherer ordnung (wie ab Version 2.2 von FEMLAB vorgesehen) ist diese polyedrische Approximation des Gebietes also nicht sinnvoll, denn schon der Interpolationsfehler würde inakzetabel groß.

Eine zweite genauere Variante der Randapproximation besteht in der Einführung nichtlinear begrenzter Elemente K_j, vgl. Abbildung 11.1a. Exemplarisch betrachten wir den *Fall quadratischer Approximation* über einer Dreieckszerlegung (mit eventuell krummliniger Berandung) für ein Gebiet $\Omega \subset \mathbf{R}^2$.

Dazu führen wir mit dem Multiindex α mit $|\alpha| = 2$ geeignete Interpolationspunkte p^α auf dem Rand des krummlinig berandeten Dreieckselementes K_j ein. Seien insbesondere $p^{(0,0,2)}, p^{(2,0,0)}, p^{(0,2,0)}$ die in mathematischer Orientierung bezeichneten Eckpunkte eines beliebiges Dreiecks $K = K_j$ (vgl. Abb. 11.1b), denen die Eckpunkte $(0,0), (1,0), (0,1)$ des Referenzdreiecks

$$\tilde{K} = \{(\xi,\eta)^* : \ \xi,\eta > 0, \ \xi+\eta < 1\}$$

in der (ξ,η)-Ebene zugeordnet werden. Den Seitenmittelpunkten $p^{(1,0,1)}, p^{(1,1,0)}, p^{(0,1,1)}$ werden entsprechend die Seitenmittelpunkte von $\tilde{K}$ zugeordnet.

Wir definieren über die baryzentrischen Koordinaten $\lambda := (\lambda_1, \lambda_2, \lambda_3) \equiv (\xi, \eta, 1-\xi-\eta)$ die Formfunktionen $\Psi_\alpha = \Psi_\alpha(\lambda)$ mit der Interpolationsbedingung $\Psi_\alpha(\beta/2) = \delta_{\alpha,\beta}$. Dabei ist wie üblich $\delta_{\alpha,\beta} = 1$ genau für $\alpha = \beta$ und $\delta_{\alpha,\beta} = 0$ sonst.

Dann kann das Element $K = K_j$ näherungsweise beschrieben werden durch

$$x = \sum_{|\alpha|=2} \Psi_\alpha(\lambda) p^\alpha, \quad \sum_{i=1}^{3} \lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1,2,3.$$

Nach Elimination von λ_3 wird durch

$$F_j : \tilde{K} \to \hat{K}, \qquad x = \sum_{|\alpha|=2} \Psi_\alpha(\xi,\eta,1-\xi-\eta) p^\alpha \tag{11.17}$$

eine nichtlineare Abbildung F_j des Referenzdreiecks $\tilde{K}$ auf eine Approximation $\hat{K}$ von $K = K_j$, vgl. Abbildung 11.1b definiert.

Eine genauere Auswertung der Formfunktionen ergibt die Parameterdarstellungen

$$\begin{array}{lll} \Psi_{200}(\lambda) := \lambda_1(2\lambda_1 - 1), & \Psi_{020}(\lambda) := \lambda_2(2\lambda_2 - 1), & \Psi_{002}(\lambda) := \lambda_3(2\lambda_3 - 1), \\ \Psi_{110}(\lambda) := 4\lambda_1\lambda_2, & \Psi_{011}(\lambda) := 4\lambda_2\lambda_3, & \Psi_{101}(\lambda) := 4\lambda_1\lambda_3 \ . \end{array}$$

Insbesondere sind dadurch die Interpolationsbedingungen $\Psi_\alpha(\frac{1}{2}\beta) = \delta_{\alpha,\beta}$ erfüllt. Diese Transformationsfunktionen besitzen somit die gleiche Parametrisierung wie die üblichen (quadratischen) Formfunktionen über dem Referenzdreieck. Man nennt die darauf basierenden Diskretisierungen auch *isoparametrische finite Elemente*.

Mittels der gewählten Formfunktionen Ψ_α und der gemäß (11.17) definierten Abbildung $x := F_j(\xi, \eta)$ wird also über dem Referenzdreieck $\tilde{K}$ eine Näherung für das krummlinig berandete Element K definiert. Dabei wird jeweils eine Dreiecksseite von $\tilde{K}$ auf eine (im Parameter quadratische) interpolierende Kurve durch die zugehörigen drei Punkte der entsprechenden Seite von K abgebildet.

Durch diese Behandlung erhält man eine approximative Zerlegung von Ω in ggf. krummlinige Teilgebiete $\hat{K}_j$. Die über diesen Elementen definierten lokalen Formfunktionen gewinnt man durch die jeweilige Formfunktion über $\tilde{K}$ und die Abbildung F_j gemäß

$$u(x) = v(F_j^{-1}(x)), \quad x \in \hat{K}_j.$$

Dabei ist die Existenz von F_j^{-1} dann gesichert, wenn die in K_j zur Bildung von F_j verwendeten Referenzpunkte hinreichend nahe zur Lage in einem regulären Dreieck sind. Die in Kapitel 10.1-10.3 untersuchten Interpolationsabschätzungen zu affin–linearen Abbildungen (mit elementweise konstanter Transformationsmatrix) lassen sich entsprechend modifizieren. Es müssen dann der Regularitätsbedingungen (10.10) bzw. (10.11) in Folgerung 10.6 entsprechende Forderungen an die nichtkonstante (!) Transformationsmatrix gestellt werden.

Beispiel 11.6. Exemplarisch betrachten wir die Lösung des Poisson-Problems $-(\Delta u)(x) = f(x)$ im Gebiet $\Omega = \{x \in (0,1)^2 : \; x_1^2 + x_2^2 > \frac{1}{4}\}$. Die rechte Seite f und die Randbedingung $u = g$ sind gerade so gewählt, daß $u(x) = \sin(\pi x_1)\sin(\pi x_2)e^{x_1 x_2}$ die (glatte) Lösung des RWP ist.

Die Lösung wird mittels FEMLAB mit P_2-Elementen und isoparametrischer Randmodifikation approximiert. Abb. 11.2 zeigt links eine Isostufendarstellung der Lösung. Ferner werden rechts die theoretischen und praktisch erreichten Konvergenzraten für den $H^1(\Omega) = W^{1,2}(\Omega)$- sowie den $L^2(\Omega)$-Fehler gezeigt. Man erkennt, daß die theoretisch erwartete Ordnung h^2 bzw. h^3 tatsächlich erreicht wird. □

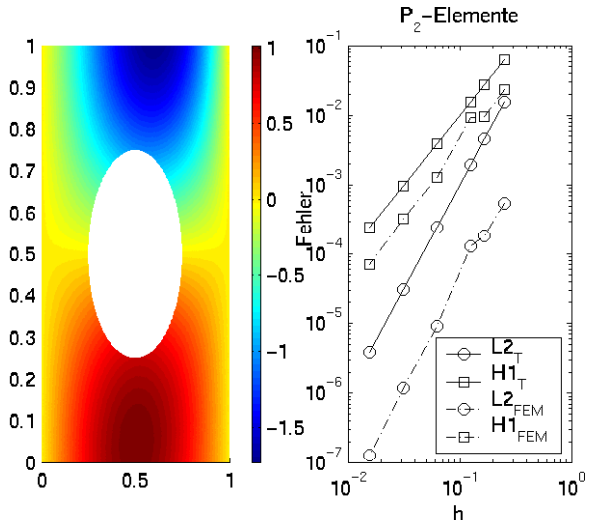


Abbildung 11.2: Isostufendarstellung der Lösung und Fehlerdiagramm zu Beispiel 11.6

Es sei vermerkt, daß man in FEMLAB auch mit P_1-Elementen und lokaler isotroper Verfeinerung durch Standard-Simplizes in Randnähe rechnen kann. Man kann über den Parameter `hcurve` des Befehls `meshinit` den Krümmungsradius des Randes analysieren und eine lokale Verfeinerung bewirken. □

11.4 Ansatzräume mit geringerer Glattheit

Wir betrachten jetzt den *nichtkonformen* Fall $X_h \not\subset X$ mit folgender Motivation: Betrachtet wird die FEM-Lösung eines elliptischen Randwertproblems der Ordnung $2m$ mit $m \in \mathbf{N}$. Im *konformen* Fall $X_h \subset X \subset W^{m,2}(\Omega)$ führt die stetige Einbettung in den Raum $C^l(\overline{\Omega})$ auf von der Raumdimension und der Ordnung des Randwertproblems abhängige Relationen zwischen den Zahlen $m, l \in \mathbf{N}$ (vgl. hierzu Einbettungssatz von Sobolev -s. Satz 6.25). Dies zieht Glattheitsforderungen an die Elemente in X_h nach sich, vgl. auch die Diskussion in Abschnitt 8.3 zum Fall $m = 2, l = 1$.

Um die ggf. erforderliche Konstruktion sehr aufwendiger finiter Elemente zu umgehen, lassen wir nun den Fall $X_h \not\subset X$ zu. Wir verwenden und präzisieren die in Abschnitt 11.1 eingeführte Notation: Sei

$$Z_h := X + X_h := \{z = v + v_h : \ v \in X, v_h \in X_h\}.$$

Ferner sei $|||\cdot|||_h$ eine auf Z_h definierte (eventuell gitterabhängige) Norm.

Mit $a_h : Z_h \times Z_h \to \mathbf{R}$ bezeichnen wir eine stetige, Z_h-elliptische Bilinearform, d.h. es gilt

$$|a_h(z_h, v_h)| \le M|||z_h|||_h|||v_h|||_h \quad \forall z_h, v_h \in Z_h,$$

$$a_h(z, z) \ge \gamma|||z|||_h^2 \quad \forall z \in Z_h.$$

Im Fall einer symmetrischen, positiv-semidefiniten Bilinearform a_h kann man zum Beispiel die induzierte Seminorm wählen

$$|||z|||_h := a_h(z, z)^{1/2} \quad \forall z \in Z_h.$$

Ferner setzen wir

$$\|z\|_h := |||z|||_h \quad \forall z \in X_h.$$

Wir erhalten die folgende Präzisierung von Satz 11.1.

Satz 11.7. *(2. Lemma von G. Strang)*
Unter den obigen Voraussetzungen gilt für den Diskretisierungsfehler des nichtkonformen diskreten Problems (11.2) an das Variationsproblem (11.1)

$$|||u - u_h|||_h \le C \left(\inf_{z_h \in X_h} |||u - z_h|||_h + \|f_h - a_h(u, \cdot)\|_{*,h} \right). \tag{11.18}$$

Beweis: Aus Gleichung (11.5) folgt

$$a_h(u_h - z_h, v_h) = a_h(u - z_h, v_h) + f_h(v_h) - a_h(u, v_h) \quad \forall z_h, v_h \in X_h.$$

Mit der Wahl $v_h := u_h - z_h$ haben wir wie im Beweis von Satz 11.1

$$|||u_h - z_h|||_h \le M|||u - z_h|||_h + \|f_h - a_h(u, \cdot)\|_{*,h}$$

und über die Dreiecksungleichung die Behauptung

$$|||u - u_h|||_h \le (1 + M) \inf_{z_h \in X_h} |||u - z_h|||_h + \|f_h - a_h(u, \cdot)\|_{*,h} \quad \forall z_h \in X_h. \quad \square$$

Als Anwendung betrachten wir einen sehr einfachen, jedoch unstetigen Ansatzraum $X_h \not\subset X$. Die hier exemplarisch vorgenommene Anwendung auf das Poisson-Problem ist eher atypisch, da man hier stetige Lagrange-Elemente zur Verfügung hat. Allerdings ist das betrachtete Element das einfachste finite Element, das bei der FEM-Approximation inkompressibler Strömungen eine Rolle spielt.

Beispiel 11.8. *(Crouzeix-Raviart Element)*
Sei $\Omega \subset \mathbf{R}^2$ polyedrisches Gebiet. Untersucht wird das Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \Omega. \tag{11.19}$$

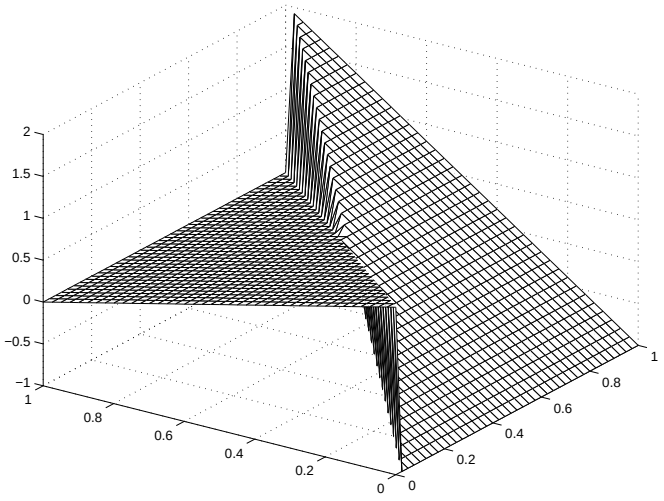


Abbildung 11.3: Crouzeix-Raviart-Elemente

Sei $\mathcal{T} = \{K_j\}_{j=1}^M$ eine zulässige und quasi-uniforme Dreieckszerlegung des Gebietes Ω. Der FE–Ansatzraum X_h werde gebildet durch stückweise lineare Funktionen über den Dreiecken, die in den Seitenmittelpunkten der Dreiecke stetig verheftet sind. Dieses sogenannte *Crouzeix-Raviart-Element* ist das einfachste Element mit $X_h \not\subset X = W_0^{1,2}(\Omega)$, vgl. Abbildung 11.3.

Seien jetzt genauer $p^j \in \Omega, j = 1, ..., N$ die inneren Gitterpunkte (Seitenmittelpunkte) und $p^j \in \partial\Omega, j = N+1, ..., \overline{N}$ die Randpunkte der Zerlegung. Wir wählen den Ansatzraum

$$\begin{aligned} X_h \quad := \quad & \{v_h \in L^2(\Omega) : v_h|_{K_j} \in P_1(K_j), \\ & v_h \text{ stetig in } p^i, i = 1, ..., N;\ v_h(p^j) = 0, p^j \in \partial\Omega\}. \end{aligned}$$

Wir definieren zu der zu (11.19) gehörenden Bilinearform

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \; dx, \quad \forall u, v \in X := W_0^{1,2}(\Omega)$$

auf $X_h \times X_h$ die Bilinearform

$$a_h(u_h, v_h) := \sum_{j=1}^M \int_{K_j} \nabla u_h \cdot \nabla v_h \; dx, \quad \forall u_h, v_h \in X_h.$$

Ferner sei

$$f_h(v) := f(v) := \int_\Omega fv \; dx, \quad \forall v \in X_h + X.$$

Satz 11.9. *Bei quasi-uniformer Dreieckszerlegung von Ω erhält man für die FEM–Lösung des Modellproblems (11.19) mit dem Crouzeix-Raviart Element unter der zusätzlichen Glättevoraussetzung $u \in W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega)$ an die Lösung von (11.19) die Fehlerabschätzung*

$$|||u - u_h|||_h \le Ch\|u\|_{2,\Omega}. \tag{11.20}$$

Beweis: Wir nutzen das 2. Lemma von Strang.

(i) Für den Interpolationsfehler in (11.18) hat man

$$\inf_{z_h \in X_h} |||u - z_h|||_h \equiv \inf_{z_h \in X_h} \|u - z_h\|_h \le Ch\|u\|_{2,\Omega},$$

da X_h speziell auch die stückweise linearen und in Ω stetigen Ansatzfunktionen enthält. Man kann also Satz 10.8 anwenden.

(ii) Wir untersuchen jetzt den Konsistenzfehler $\|f_h - a_h(u,\cdot)\|_{*,h}$. Wegen (11.19) gilt mit $v_j := v_h|_{K_j}$ bei partieller Integration

$$\begin{aligned} f_h(v_h) - a_h(u,v_h) &= \sum_{j=1}^{M} \left(\int_{K_j} f v_h \, dx - \int_{K_j} \nabla u \cdot \nabla v_h \, dx \right) \\ &= -\sum_{j=1}^{M} \left(\int_{K_j} \Delta u \; v_h \, dx + \int_{K_j} \nabla u \cdot \nabla v_h \, dx \right) \\ &= -\sum_{j=1}^{M} \int_{\partial K_j} \frac{\partial u}{\partial n_j} v_j(s) \, ds. \end{aligned} \tag{11.21}$$

Dabei ist n_j die äußere Normale auf ∂K_j. Bezeichne $[v_h|_e] := v_j|_e - v_i|_e$ den *Sprung* der Funktion v_h auf der Kante $e := \partial K_j \cap \partial K_i$ zwischen zwei benachbarten Dreiecken. Bei Kanten e auf dem Gebietsrand $\partial\Omega$ setzt man v_h mit Null außerhalb von Ω fort. Bei Summation über die Kanten e der Triangulation erhält man unter Beachtung von $n_i = -n_j$ für die Normalenvektoren auf einer Kante $e := \partial K_j \cap \partial K_i$ für den Konsistenzfehler

$$f_h(v_h) - a_h(u,v_h) = -\sum_{e \in \mathcal{T}_h} \int_e \frac{\partial u}{\partial n_e} [v_h|_e] \, ds. \tag{11.22}$$

Zur Abschätzung der Kantenterme in (11.22) betrachten wir die Vereinigung $G = T_1 \cup T_2$ zweier "Referenzdreiecke" mit $\operatorname{diam}(G) = 1$ und der gemeinsamen Kante $e = \overline{p^1 p^2}$ zwischen den Eckpunkten p^1 und p^2. Ferner sei m der Kantenmittelpunkt von e.

Seien nun $\zeta \in W^{1,2}(G)$ und $V := \{z|\ z|_{T_i} \in P_1(T_i),\ z \text{ stetig in } m\}$. Dann findet man eine endliche Konstante C, so daß

$$\left| \int_e \zeta \, (z|_{T_2} - z|_{T_1}) \; ds \right| \le C|\zeta|_{1,G} \, (|z|_{1,T_1} + |z|_{1,T_2}) \quad \forall z \in V. \tag{11.23}$$

Dies sieht man wie folgt: Mit $[z] := z|_{T_2} - z|_{T_1}$ ist $\int_e [z] \, ds = 0$. Mit beliebigen Konstanten c_1, c_2 folgt dann über die Cauchy-Schwarz Ungleichung sowie mittels Spurungleichung (vgl. Satz 5.21)

$$\begin{aligned} \left| \int_e \zeta[z] \, ds \right| &= \left| \int_e (\zeta - c_1)[z] \, ds \right| = \left| \int_e (\zeta - c_1)[z - c_2] \, ds \right| \\ &\le \|\zeta - c_1\|_{L^2(e)} \|[z - c_2]\|_{L^2(e)} \\ &\le C \|\zeta - c_1\|_{1,G} \, (\|z - c_2\|_{1,T_1} + \|z - c_2\|_{1,T_2}) \, . \end{aligned} \tag{11.24}$$

Dann folgt (11.23) aus (11.24) durch Infimumbildung über $c_1, c_2 \in \mathbf{R}$ aus der Interpolationsungleichung in Satz 10.8 mit $r = k = 0$ für den L^2-Anteil der auf der rechten Seite von (11.24) stehenden Terme.

Wir betrachten jetzt ein einzelnes Dreieck T mit der ausgewählten Kante $e = \overline{p^1 p^2}$ zwischen den Eckpunkten p^1 und p^2 und dem Kantenmittelpunkt m von e. Seien $\zeta \in W^{1,2}(T)$ und

$$V := \{z|\ z \in P_1(T),\ z(m) = 0\}.$$

Durch entsprechende Modifikation der Herleitung von (11.24) folgt die Existenz einer endlichen Konstante C mit

$$\left| \int_e \zeta[z] \, ds \right| \le C|\zeta|_{1,T}|z|_{1,T}. \tag{11.25}$$

Abschließend wenden wir die Aussagen (11.23) und (11.24) mit $\zeta = \frac{\partial u}{\partial n_e}$ an und benutzen ein Skalierungsargument, nach dem die Konstanten C in den genannten Ungleichungen durch Ch über einer quasi-uniformen Zerlegung $\mathcal{T}_h$ ersetzt werden können. Damit folgt nach Summation über alle Kanten e

$$|f_h(v_h) - a_h(u,v_h)| \le Ch|u|_{2,\Omega} |||v_h|||_h$$

und damit die Behauptung des Satzes. □

Bemerkung 11.10. Eine oft gewählte Alternative zu nichtkonformen FEM bei elliptischen Randwertaufgaben der Ordnung $2m$ stellen sogenannte *gemischte Methoden* dar. Dabei formt man zunächst das Ausgangsproblem in ein System von Differentialgleichungen niedriger Ordnung um und wendet dann passende Finite-Elemente-Ansätze mit wesentlich geringerer Zahl von Freiheitsgraden an. Man benötigt aber für die entstehenden diskreten Aufgaben vom *Sattelpunkt-Typ* eine spezielle Lösbarkeitstheorie.

Wir besprechen gemischte Probleme einführend im Teil IV der Vorlesung. Eine weitergehende Orientierung findet man bei S. Brenner/ R. Scott [4], Kap. 10, C. Großmann/H.G. Roos [10], Kap. 4.6. □

Kapitel 12

Fehlerschätzung und Adaptivität

Die bisher beschriebenen *a-priori Fehlerabschätzungen* der Form

$$|||u - u_h||| \leq Ch^p \|u\|_s \tag{12.1}$$

in einer geeigneten Norm $|||\cdot|||$ sind für praktische Rechnungen nur eingeschränkt sinnvoll. Zunächst gelten sie nur im asymptotischen Sinne $h \to +0$. Praktisch ist aber offen, ob ein gegebenes Gitter fein genug für diese Abschätzung ist. Andererseits sind im allgemeinen Fall weder realistische Schätzungen für die Konstanten C sowie der Norm $\|u\|_s$ der (unbekannten !) Lösung verfügbar. Ferner setzt die für stückweise Polynome vom Grad k geltende optimale Fehlerabschätzung die Glattheitsforderung $u \in W^{k+1,2}(\Omega)$ voraus, die im allgemeinen Fall nicht gilt.

Ziel adaptiver Verfahren ist, auf einem gegebenen Gitter aus einer lokalen *a–posteriori* Schätzung des Fehlers aus berechenbaren (!) Größen auf notwendige lokale Gitterverfeinerungen zu schließen.

12.1 Fehlerschätzer und -indikatoren

Die Struktur einer adaptiven Netzgenerierungsmethode bei gegebenem Fehlerschätzer ist:

1. Konstruktion eines Anfangsgitters $\mathcal{T}_0$ mit hinreichend guter Anpassung an die Problemgeometrie (und möglichst der Grobstruktur der Lösung). Setze $k := 0$.
2. Löse das diskrete Problem auf dem Gitter $\mathcal{T}_k$.
3. Berechne eine a–posteriori Fehlerschätzung für jedes Element $K \in \mathcal{T}_k$.
4. Ist der abgeschätzte globale Fehler hinreichend klein, so stop. Anderenfalls entscheide, welche Elemente K zu verfeinern oder zu vergröbern sind, und konstruiere das Gitter $\mathcal{T}_{k+1}$. Ersetze k durch $k + 1$ und gehe zu Schritt 2.

Sei jetzt η_K ein zum Element $K \in \mathcal{T}$ gehörender Schätzer.

Definition 12.1. *Die Größe η aus*

$$\eta^2 := \sum_{K \in \mathcal{T}} \eta_K^2 \tag{12.2}$$

heißt a-posteriori Schätzer zur Norm $|||\cdot|||$, *wenn sie nur aus den Problemdaten und der berechneten diskreten Lösung u_h ermittelt wird und falls es Konstanten d_u, d_l gibt mit*

$$d_l \eta \leq |||u - u_h||| \leq d_u \eta. \tag{12.3}$$

Zur Beschränkung des Fehlers der numerischen Lösung unter eine vorgegebene Toleranz ist lediglich ein *oberer Fehlerschätzer* (oder *Verfeinerungsindikator*) mit

$$|||u - u_h||| \leq d_u \eta \tag{12.4}$$

erforderlich. Dann besteht jedoch die Gefahr einer übermäßigen Verfeinerung. Zu deren Vermeidung sucht man auch untere Fehlerschranken der Form

$$d_l^* \eta_K \le |||u - u_h|||_{\omega_K}, \tag{12.5}$$

wobei ω_K eine möglichst kleine Umgebung des Elements K bezeichnet. Ist die Zahl der Elemente in ω_K unabhängig von K und h beschränkt, so kann auch eine untere Abschätzung wie in (12.3) abgeleitet werden.

Zur Bewertung eines a-posteriori Fehlerschätzers zieht man den *Effektivitätsindex*

$$\frac{\eta}{|||u - u_h|||}$$

heran. Ein Fehlerschätzer heißt *effizient*, wenn der Effektivitätsindex und seine Inverse für alle Gitterweiten beschränkt bleiben. Aus den Abschätzungen (12.3) folgt die Effizienz des Schätzers.

Ein Fehlerschätzer wird *asymptotisch exakt* genannt, falls gilt

$$\lim_{h \to +0} \frac{\eta}{|||u - u_h|||} = 1.$$

Es ist jedoch kritisch zu vermerken, daß die Forderung nach asymptotischer Exaktheit eines Schätzers unrealistisch und praktisch kaum realisiert ist.

Nachfolgend beschreiben wir einige typische Fehlerschätzer. Man vergleiche auch die Übersichtsarbeit von R. Verfürth [23].

12.2 Fehlerschätzer für die Poisson-Gleichung

Wir untersuchen wieder das Modellproblem

$$-\Delta u = f \quad \text{in} \quad \Omega \subset \mathbf{R}^2, \qquad u = 0 \quad \text{in} \quad \partial\Omega, \tag{12.6}$$

das mittels stückweise linearer konformer finite Elemente über einer quasiuniformen Dreieckszerlegung $\mathcal{T}$ näherungsweise gelöst werden soll. Dann sind $X := W_0^{1,2}(\Omega)$ und $X_h \subset X$. Wir wollen den Fehler in der Norm

$$|||\cdot||| := \|\nabla(\cdot)\|_{0,\Omega}$$

schätzen, die in natürlicher Weise auch bei a-priori Fehlerabschätzungen betrachtet wurde.

Für die verallgemeinerte Lösung $u \in X$ von (12.6) und die diskrete Lösung $u_h \in X_h$ gilt unter Verwendung des L^2-Skalarproduktes $(\cdot,\cdot)$ die Fehlergleichung

$$(\nabla(u - u_h), \nabla v_h) = (f, v_h) - (\nabla u_h, \nabla v_h), \quad \forall v_h \in X_h. \tag{12.7}$$

Gleichung (12.7) ist Ausgangspunkt für nachfolgende Untersuchungen von Fehlerschätzern.

(i) Ein residualer Fehlerschätzer

Aus der Friedrichschen Ungleichung (vgl. Beweis von Satz 5.26 (i)) folgt

$$\|v\|_{0,\Omega} \le C_F \|\nabla v\|_{0,\Omega}, \quad \forall v \in X,$$

damit

$$\frac{1}{\sqrt{1 + C_F^2}} \|v\|_{1,\Omega} \le \|\nabla v\|_{0,\Omega} = \sup_{\substack{w \in X \\ \|\nabla w\|_{0,\Omega} = 1}} (\nabla v, \nabla w) \le \|v\|_{1,\Omega}.$$

In Verbindung mit (12.7) finden wir

$$Q := \sup_{\substack{w \in X \\ \|\nabla w\|_{0,\Omega} = 1}} ((f, w) - (\nabla u_h, \nabla w)) = \|\nabla(u - u_h)\|_{0,\Omega}$$

und damit

$$\frac{1}{\sqrt{1+C_F^2}}\|u-u_h\|_{1,\Omega} \le Q \le \|u-u_h\|_{1,\Omega}. \tag{12.8}$$

Im nächsten Schritt ermitteln wir eine obere Schranke für Q. Die Projektionseigenschaft des Galerkin-Verfahrens (*Galerkin-Orthogonalität*) sichert

$$(\nabla(u-u_h),\nabla v_h) = 0, \quad \forall v_h \in X_h.$$

Mit einem zu präzisierenden Operator $P_h : X \to X_h$ folgt damit

$$Q^*(w) := (f,w) - (\nabla u_h, \nabla w) = (f, w - P_h w) - (\nabla u_h, \nabla(w - P_h w)), \quad \forall w \in X.$$

Partielle Integration über die einzelnen finiten Elemente $K \in \mathcal{T}$ ergibt

$$Q^*(w) = \sum_{K\in\mathcal{T}} (f + \Delta u_h, w - P_h w)_K - \sum_{K\in\mathcal{T}} \int_{\partial K} (n_K \cdot \nabla u_h)\,(w - P_h w)\; ds. \tag{12.9}$$

Dabei ist n_K die äußere Normale auf ∂K. Man beachte, daß sich im Fall stückweise linearer Elemente das Residuum $f + \Delta u_h$ auf f reduziert.

Sei $\mathcal{E}$ die Vereinigung aller Dreieckskanten der Zerlegung $\mathcal{T}$. Dann kann der zweite Term auf der rechten Seite von (12.9) als Summe über die Kanten $E \in \mathcal{E}$ unter Verwendung des Sprunges $[n_E \cdot \nabla u_h]_E$ der Normalenableitung von u_h über die Kante E (bei fester Wahl von n_E zu $E \in \mathcal{E}$) geschrieben werden:

$$Q^*(w) = \sum_{K\in\mathcal{T}} (f + \Delta u_h, w - P_h w)_K - \sum_{E\in\mathcal{E}} \int_E [n_E \cdot \nabla u_h]_E (w - P_h w)\; ds. \tag{12.10}$$

Da für $w \in W^{1,2}(\Omega)$ der übliche stückweise polynomiale Interpolationsoperator (vgl. Kap. 10) nicht definiert ist, betrachten wir nun einen auf Clement (1975) zurückgehenden Interpolationsoperator P_h: Seien ω_K bzw. ω_E die Vereinigung aller Dreiecke, die einen Eckpunkt mit dem Dreieck K bzw. der Kante E gemeinsam haben. Ferner bezeichne $\mathcal{U}(p^i)$ die Vereinigung der Dreiecke, die den Gitterpunkt p^i als Eckpunkt besitzen. Durch den lokalen L^2-Projektor

$$\tilde{\Pi}_i : \; L^2(\mathcal{U}(p^i)) \to X_h|_{\mathcal{U}(p^i)}$$

wird eine Mittelung einer Funktion auf $\mathcal{U}(p^i)$ erzeugt. Der Clement-Operator wird schließlich definiert durch

$$(P_h u)(x) := \sum_i (\tilde{\Pi}_i u)(p^i)\; \phi_i(x),$$

wobei $\{p^i\}_i$ die Menge aller Gitterpunkte in $\mathcal{T}$ ist. $\phi_i(\cdot)$ bezeichnet die zugehörige nodale Basisfunktion, d.h. $\phi_i(p^j) = \delta_{ij}$. Dabei gelten die folgenden Interpolationsaussagen (vgl. Ph. Clement, in: *RAIRO Anal. Numer.* 2 (1975), 77-84)

$$\begin{aligned} \|w - P_h w\|_{0,K} &\le C h_K \|w\|_{1,\omega_K} \\ \|w - P_h w\|_{0,E} &\le C h_E^{1/2} \|w\|_{1,\omega_E} \end{aligned} \tag{12.11}$$

Dabei sind h_K der Durchmesser des Elementes K und h_E die Kantenlänge von E. Einsetzen von (12.11) in (12.10) und die Ungleichung von Cauchy-Schwarz liefern unter Beachtung der Quasiuniformität des Netzes

$$\begin{aligned} Q^*(w) &\le \sum_{K\in\mathcal{T}} C h_K \|f + \Delta u_h\|_{0,K} \|w\|_{1,\omega_K} + \sum_{E\in\mathcal{E}} C h_E^{1/2} \|[n_E \cdot \nabla u_h]_E\|_{0,E} \|w\|_{1,\omega_E} \\ &\le C\|w\|_{1,\Omega} \left(\sum_{K\in\mathcal{T}} h_K^2 \|f + \Delta u_h\|_{0,K}^2 + \sum_{E\in\mathcal{E}} h_E \|[n_E \cdot \nabla u_h]_E\|_{0,E}^2 \right). \end{aligned} \tag{12.12}$$

Mit der Festlegung

$$\eta_{R,K}^2 := h_K^2 \|f + \Delta u_h\|_{0,K}^2 + \frac{1}{2} \sum_{E \in \partial K} h_E \|[n_E \cdot \nabla u_h]_E\|_{0,E}^2 \tag{12.13}$$

folgt aus den Abschätzungen (12.12) und (12.8), daß $\eta_{R,K}$ ein oberer Fehlerschätzer ist. Damit ergibt sich der folgende

Satz 12.2. *Durch Formel (12.13) ist bei Verwendung stückweise linearer Dreieckselemente zur Approximation des Modellproblems (12.6) ein oberer Fehlerschätzer gegeben.*

Bemerkungen 12.3. (i) Es kann mit ähnlichen Mitteln gezeigt werden, daß der residuale Fehlerschätzer auch ein unterer Schätzer für die Norm $\|\nabla(\cdot)\|_{0,\Omega}$ ist.

(ii) Eine Approximation dieses residualen Schätzers findet sich in der Version 2.3 des Programmsystems FEMLAB. Allerdings fehlt dort der zweite Anteil, der den Sprung der Normalenableitung über die Elementkanten mißt. Das ist bei stückweise linearen Elementen keine sehr günstige Lösung, da bereis der diskretisierte Laplace-Operator Δu_h elementweise verschwindet. Der Sprungterm wäre gerade dann wesentlich. Man vergleiche hierzu auch das Beispiel am Ende dieses Kapitels. □

(ii) Schätzer mit Lösung lokaler residualer Probleme

Eine weitere Gruppe von oberen Fehlerschätzern basiert auf der Lösung lokaler residualer Probleme auf einer Vereinigung finiter Elemente ω_K in Umgebung eines finiten Elementes K. Sei X_{hK} ein geeigeter (und noch festzulegender) Finite-Elemente Unterraum niedriger Dimension, der auf ω_K definiert ist. Allerdings sind die Basisfunktionen in $X_{h,K}$ von höherer Ordnung als im Finite-Elemente-Raum X_h.

Wir beginnen mit Schätzern, bei denen lokale Dirichlet-Probleme gelöst werden: Hierzu sei $v_K \in X_{hK}$ Lösung des lokalen residualen Problems

$$(\nabla v_K, \nabla w)_{\omega_K} = (f, w)_{\omega_K} - (\nabla u_h, \nabla w)_{\omega_K} \quad \forall w \in X_{hK}. \tag{12.14}$$

Wegen $X_{hK} \subset W_0^{1,2}(\omega_K)$ entspricht (12.14) einer Finite-Elemente-Approximation an $\phi = u_h + v_K$ des Dirichlet-Problems

$$-\Delta\phi = f, \quad \text{in } \omega_K; \qquad \phi = u_h, \quad \text{auf } \partial\omega_K.$$

Als lokalen Fehlerindikator wählt man

$$\eta_{L,K} := \|\nabla v_K\|_{0,\omega_K}. \tag{12.15}$$

Bei geeigneter Wahl von ω_K und X_{hK} approximiert die Größe $\eta_{L,K}$ den lokalen Fehler $\|u - u_h\|_{1,K}$.

Zur Charakterisierung einiger bekannter Varianten führen wir die folgende elementorientierte kubische *"Blasenfunktion" (bubble function)* auf K durch

$$b_K := \begin{cases} 27\lambda_{K1}\lambda_{K2}\lambda_{K3} & \text{auf } K \\ 0 & \text{in } \Omega \setminus K \end{cases}$$

ein, wobei $\lambda_{Ki}, i = 1, 2, 3$ die baryzentrischen Koordinaten auf K sind. Bei fixierter gemeinsamer Kante E von zwei Dreiecken K_1 und K_2 sollen die Eckpunkte der Dreiecke so numeriert werden, daß die Eckpunkte von E zuerst gezählt werden. Eine kantenorientierte quadratische Blasenfunktion b_E wird erklärt durch

$$b_E := \begin{cases} 4\lambda_{K_i 1}\lambda_{K_i 2} & \text{auf } K_i, i = 1, 2 \\ 0 & \text{in } \Omega \setminus (K_1 \cup K_2). \end{cases}$$

Eine Darstellung der element- bzw. kantenorientierten Blasenfunktionen findet man in Abbildung 12.1.

Wir nennen die folgenden oft verwendeten Varianten:

- *Babuska-Rheinboldt Schätzer:*
  Hier ist ω_K die Vereinigung der Dreiecke K', die einen Eckpunkt x mit K gemeinsam haben, sowie

$$X_{hK} := \text{span}\{b_{K'}, b_E : K' \subset \omega_K, E \cap x \neq \emptyset\} \subset W_0^{1,2}(\omega_K).$$

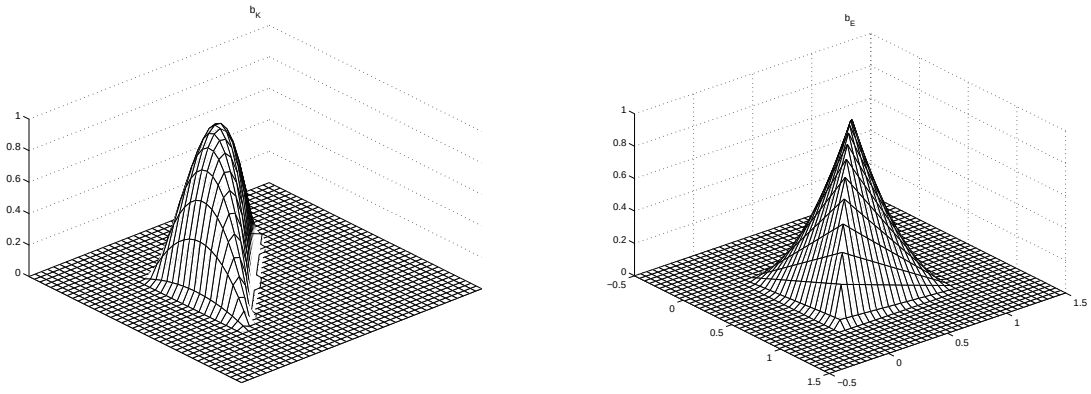


Abbildung 12.1: Darstellung der element- und kantenorientierten Blasenfunktionen

- *Verfürth Schätzer:*
  Hier besteht ω_K aus den Dreiecken K', die eine gemeinsame Kante mit K haben. Ferner ist

$$X_{hK} := \text{span}\{b_{K'}, b_E : K' \subset \omega_K, E \subset \partial K\} \subset W_0^{1,2}(\omega_K).$$

Schließlich betrachten wir noch einen Schätzer, bei dem lokale Neumann-Probleme gelöst werden:
Modifizierter Bank–Weiser Schätzer:
Seien $\omega_K = K$ sowie

$$X_{hK} := \text{span}\{b_K, b_E : E \subset \partial K\}.$$

Mit v_K wird die Lösung des Problems

$$(\nabla v_K, \nabla w)_K = (f + \Delta u_h, w)_K - \frac{1}{2} \sum_{E \in K} \int_E [n_E \cdot \nabla u_h]_E w \, ds, \quad \forall w \in X_{hK}$$

bezeichnet, das sich als diskretisierte Variationsgleichung des lokalen Neumann-Problems

$$-\Delta \phi = f + \Delta u_h \ \text{ in } K, \qquad \frac{\partial \phi}{\partial n} = -\frac{1}{2}[n_E \cdot \nabla u_h]_E \ \text{ auf } E \subset \partial K$$

ergibt.

Bemerkung 12. 4. Für elliptische Probleme mit dominierendem Anteil des elliptischen Hauptteils (d.h. den Termen mit zweiten partiellen Ableitungen der Lösung) ist die Wahl der $W^{1,2}-$Norm als Basis von Fehlerschätzungen einerseits natürlich. Dies ist aber nicht zwingend, denkbar sind auch $L^\infty-$ Schätzer. Bei Aufgaben mit Dominanz der Terme mit niedrigeren Ableitungen ist auch die Verwendung der L^2-Norm sinnvoll. □

Bemerkung 12. 5. Es gibt seit einiger Zeit wichtige Weiterentwicklungen adaptiver Fehlerschätzer. Neben den hier behandelten Verbesserungen bezüglich bestimmter "globaler" Normen $|||\cdot|||$ kann man auch als Zielfunktional gewisse Integralwerte über Teile des Randes $\partial\Omega$ oder sogar Punktfunktionale betrachten. Eine gute Übersicht zu dem Konzept der "dual gewichteten Residuen-Methoden (DWR) findet man in den Monographie [2] von Bangerth/Rannacher. □

12.3 Gitterverfeinerung

Wir wollen abschließend skizzieren, wie eine lokale Verfeinerungstechnik erreicht wird. Hierbei ist insbesondere zu beachten, daß bei der Verfeinerung nicht entartete Elemente (d.h. mit zu kleinen Innenwinkeln)

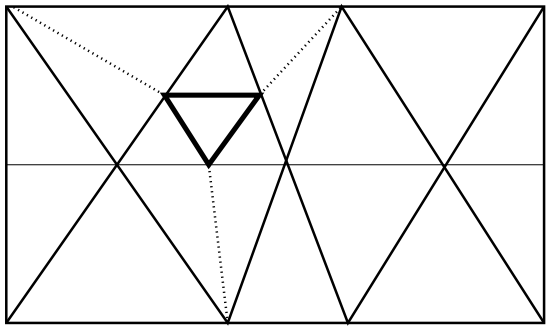


Abbildung 12.2: Rote und grüne Verfeinerung

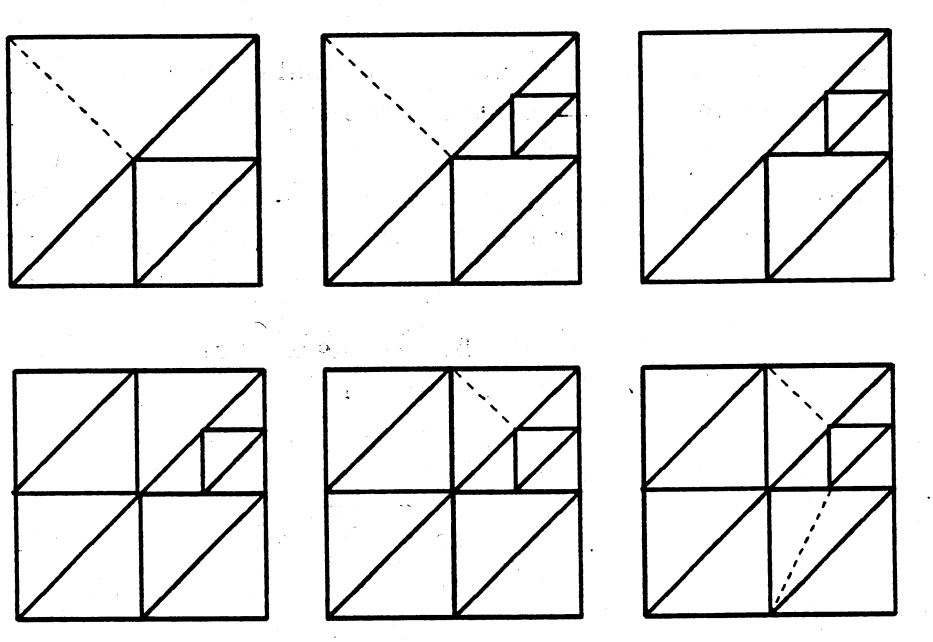


Abbildung 12.3: Rot-grüne Verfeinerung

entstehen. (Wir verweisen hierzu auf den Beweis von Satz 12.2. Hier wird implizit verlangt, daß die Anzahl der an ein Dreieck angrenzenden Elemente begrenzt bleibt.)

Im Programm PLTMG (*piecewise linear triangles multi grid*) von R. Bank und ähnlich in der Version 2.1 des Programmsystems FEMLAB werden auf Dreiecksnetzen zwei Varianten realisiert:

- Zerlegung eines Dreiecks durch Halbierung aller Seiten in vier kongruente Dreiecke (*reguläre* oder *rote Verfeinerung*),
- Zerlegung eines Dreiecks in zwei Dreiecke durch Halbierung einer Seite (*grüne Verfeinerung*).

In der Abbildung 12.2 wird die grüne Verfeinerung durch unterbrochene Linien markiert. Sie wird eingesetzt zur Sicherung der Zulässigkeit der Zerlegung bei roter Verfeinerung. Die Abbildung zeigt die Erzeugung benachbarter grüner Dreiecke bei einmaliger Zerlegung eines Elements in vier rote Dreiecke.

Eine weitere Zerlegung des erzeugten grünen "Zwillings" erfolgt nicht direkt: Ist eine derartige Zerlegung erforderlich, so werden die Teile zunächst wieder vereinigt und dann regulär zerlegt. Die folgende Bildsequenz (vgl. Abb. 12.3) zeigt exemplarisch eine derartige Prozedur. Die genannte Zerlegungsstrategie

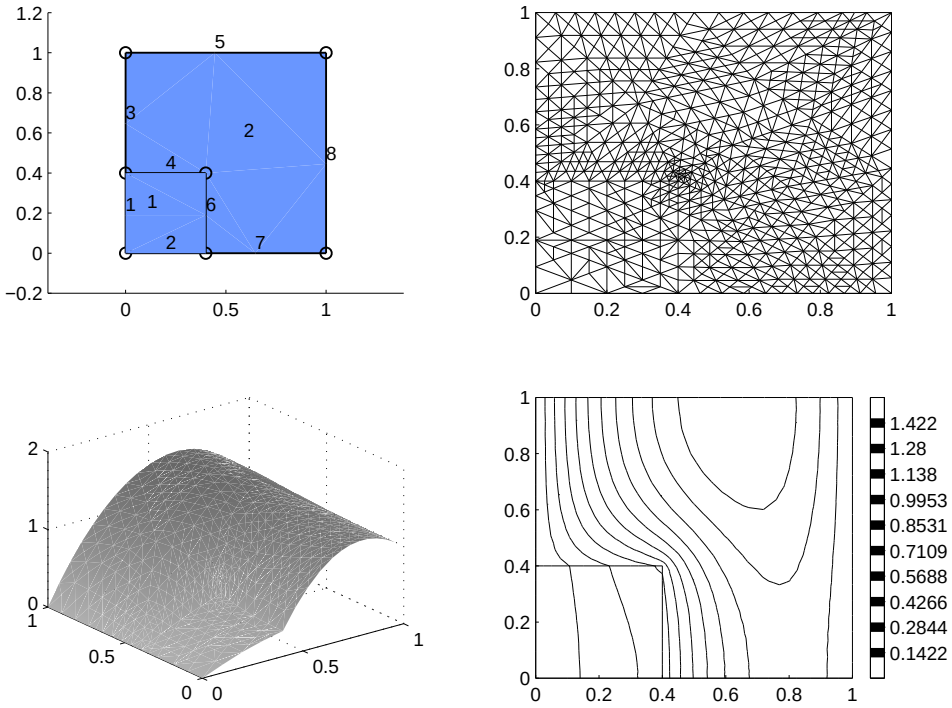


Abbildung 12.4: Netzadaption in Beispiel 12.7

sichert, daß alle Innenwinkel der im Verfeinerungsprozeß erzeugten Dreiecke gleichmäßig (d.h. unabhängig von der Tiefe der Zerlegung) nach unten beschränkt bleibt.

Durch die vorübergehende Bildung grüner Dreiecke und deren mögliche Auflösung sind die entstehenden Gitter nicht notwendig ineinander geschachtelt. Dies wäre insbesondere für Mehrgitterverfahren (vgl. eines der nachfolgenden Kapitel) sinnvoll. Durch zusätzliche Auswahlkriterien bei der Zerlegung kann dies jedoch erzwungen werden.

Es sei vermerkt, daß das Grundprinzip der aufgezeigten Verfeinerungsstrategie im Programmsystem KASKADE vom Zuse-Zentrum Berlin weiterentwickelt und mit der Anwendung hierarchischer Basen verbunden wurde.

Bemerkung 12.6. Für Probleme mit stark anisotropem Charakter, d.h. Aufgaben mit dominierenden Vorzugsrichtungen, wurde das PLTMG-Konzept von Kornhuber/ Roitzsch am Zuse-Zentrum Berlin 1990 um die *"blaue Verfeinerung"* erweitert. Dabei werden Dreiecke zunächst zu einem Viereck vereinigt und dieses dann entsprechend der Vorzugsrichtung in vier Dreiecke geteilt. □

Im Rahmen dieser Vorlesung können wir nicht weiter auf die Realisierung von Schritt 4 des zu Beginn des Kapitels dargestellten iterativen Zyklus der adaptiven Netzgenerierung eingehen. Der an Details interessierte Leser sei hier auf Teil IV der Monographie von R. Verfürth [23] verwiesen. Es sei hier nur vermerkt, daß bereits die Festlegung eines Abbruchkriteriums und die Details der Entscheidungsfindung über Verfeinerung oder Vergröberung von Elementen sehr viel praktische Erfahrung erfordern.

Wir wollen abschließend ein Anwendungsbeispiel des in FEMLAB realisierten residualen Fehlerindikators (vgl. Bemerkung 12.3 (ii)) angeben.

Beispiel 12.7. Im Einheitsquadrat $\Omega = (0,1) \times (0,1)$ betrachten wir das gemischte Randwertproblem

$$-\sum_{i=1}^{2} \frac{\partial}{\partial x_i} \left(a(x) \frac{\partial u}{\partial x_i} \right) = 0 \qquad \text{in } \Omega$$

$$\begin{aligned} u &= 0 \quad \text{auf } \Gamma_1 \\ u &= 1 \quad \text{auf } \Gamma_2 \\ \frac{\partial u}{\partial n} &= 0 \quad \text{auf } \Gamma_3. \end{aligned}$$

Dabei sind $\Gamma_1 = \{(x_1, x_2) \in \partial\Omega,\ x_1 = 0\}$, $\Gamma_1 = \{(x_1, x_2) \in \partial\Omega,\ x_1 = 1\}$ sowie $\Gamma_3 = \partial\Omega \setminus (\Gamma_1 \cap \Gamma_2))$. Der unstetige Diffusionskoeffizient ist

$$a(x_1, x_2) = 1, \quad \text{in } \Omega_1 = (0, 0.4) \times (0, 0.4); \qquad a(x_1, x_2) = 0.1, \quad \text{sonst in } \Omega.$$

Hier wurde der in der Version 2.3 von FEMLAB implementierte residuale Fehlerindikator verwendet. Es wurden vier Verfeinerungszyklen durchgeführt, bevor maximal 1.500 Dreiecke zur Diskretisierung herangezogen wurden.

Aus Abbildung 12.4 erkennt man, daß der Gradient der Lösung an den Linien $(x_1, x_2) \in (0, 0.4] \times \{0.4\}$ und $(x_1, x_2) \in \{0.4\} \times (0, 0.4]$ unstetig wird. Der Fehlerindikator erkennt offenbar diesen Sprung, allerdings erfolgt die Verfeinerung nicht sehr "lokal" entlang dieser Linien. Dies sollte an den in diesem Fehlerindikator fehlenden Kantensprung-Termen des residualen Fehlerschätzers aus Kapitel 12.2 (i) liegen, vgl. hierzu auch Bemerkung 12.3 (ii). □

Teil III

Algebraische Lösungsverfahren

Kapitel 13

Angepaßte direkte Lösungsverfahren

Im Teil III dieser Vorlesung befassen wir uns mit *Lösungsverfahren* für die bei der Diskretisierung elliptischer Randwertprobleme entstehenden großen linearen Gleichungssysteme. Direkte Lösungsverfahren spielen bei nicht zu großer Dimension der Probleme immer noch eine wichtige Rolle bei Anwendungen. Jedoch werden bei sehr großen Problemen (insbesondere im räumlich dreidimensionalen Fall) zunehmend iterative Lösungsverfahren in Verbindung mit geeigneter Vorkonditionierung unverzichtbar.

Im vorliegenden Kapitel betrachten wir zunächst Besonderheiten der bei der Diskretisierung elliptischer RWP entstehenden linearen Gleichungssysteme. Dann untersuchen wir Möglichkeiten und Grenzen der Anpassung *direkter Lösungsverfahren* an die spezielle Systemstruktur. In den Versionen 2.1 (und auch 2.2) des Systems FEMLAB spielen derartige Verfahren bei Problemen bis zu mittelgroßer Dimension eine wichtige Rolle.

Beweise werden hier nicht geführt, dazu wird auf die Vorlesung *Numerische Mathematik I* verwiesen.

13.1 Spezifik diskretisierter elliptischer RWP

Wir hatten in den vorhergehenden Kapiteln dargestellt, daß die Diskretisierung elliptischer Randwertprobleme auf die Lösung linearer Gleichungssysteme

$$\text{Finde } \; u_h \in \mathbf{R}^N : \quad A_h u_h = f_h \tag{13.1}$$

mit regulärer quadratischer Matrix $A_h \in \mathbf{R}^{N\times N}$ führt. Die Spezifik dieser Systeme kommt insbesondere im Falle von finiten Elementen niedriger Ordnung [1] zum Ausdruck durch

- die sehr große Dimension N mit $N = N(h) \to \infty,\ h \to 0$,
- eine sehr schwach besetzte Koeffizientenmatrix A_h,
- eine sehr schlechte Kondition der Matrix A_h.

Damit sind die Standardverfahren der linearen Algebra bei hinreichend großer Dimension des diskreten Problems entweder nicht effektiv oder aus Rechenzeit- und Speicherplatzbeschränkungen gar nicht anwendbar. Die (sehr) schlechte Kondition der Probleme (13.1) kann zu unakzeptablen Rechenfehlern führen. Wir betrachten ein einfaches Beispiel.

Beispiel 13.1. Für das Dirichlet-Problem der Poisson-Gleichung

$$-\Delta u = f \quad \text{in } \; \Omega = (0,1)^2; \qquad u = 0 \quad \text{auf } \; \partial\Omega \tag{13.2}$$

[1] Die ersten beiden Punkte sind natürlich zu relativieren, wenn man Elemente höherer Ordnung verwendet. Mit wachsendem Polynomgrad nimmt natürlich der Besetzungsgrad der Matrix zu. Andererseits kann man im Vergleich zu Elementen niedriger Ordnung die erforderliche Anzahl der Unbekannten (und damit die Dimension) deutlich reduzieren.

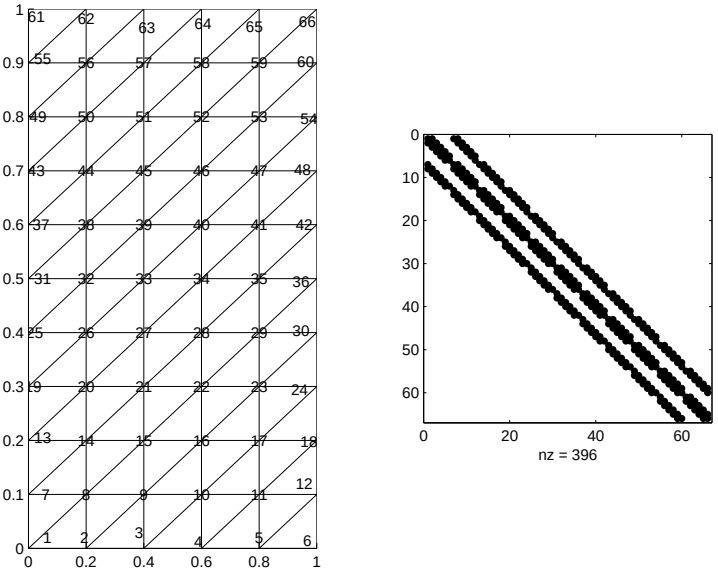


Abbildung 13.1: Strukturierte Zerlegung und zugehörige Matrixstruktur

wird eine FEM mit stückweise linearen Ansatzfunktionen auf einer isotropen Dreieckszerlegung verwendet. Das entstehende lineare Gleichungssystem erzeugt eine symmetrische, positiv definite Matrix A_h. Sind ferner alle Innenwinkel nicht stumpf, so genügt das diskrete Problem sogar dem diskreten Maximum-Prinzip.

Wir betrachten zunächst den Fall einer *regelmäßigen Zerlegung* (vgl. Abbildung 13.1). Dann entspricht die FEM gerade der Diskretisierung des Laplace-Operators auf einem Fünfpunkt-Stern (vgl. Kap. 4). Bei lexikographischer Numerierung der inneren Gitterpunkte hat die Matrix eine typische *Bandstruktur* (vgl. Abbildung 13.1) mit fünf Nichtnullelementen je Zeile.

Für den betragsmäßig größten bzw. kleinsten Eigenwert der Matrix A_h gilt

$$\lambda_{max}(A_h) = 0(1), \qquad \lambda_{min}(A_h) = 0(h^2) \tag{13.3}$$

(vgl. auch Beispiel 13.6 bzw. Übungsaufgabe), d.h. mit der Verfeinerung des Gitters verschlechtert sich die Kondition der Matrix gemäß

$$\text{cond}(A_h) = \frac{\lambda_{max}}{\lambda_{min}} = 0(h^{-2}). \tag{13.4}$$

(13.4) ist Konsequenz der folgenden *inversen Ungleichung* (vgl. Übungsaufgabe).

Lemma 13.2. *Sei $(\mathcal{T}_h)_h$ eine Familie von Zerlegungen, die mittels affin-linearer Abbildung aus einem einheitlichen Referenzelement erzeugt wird. Die Zerlegungen $\mathcal{T}_h$ seien isotrop, daß heißt: es gibt eine positive Zahl κ, so daß für alle Elemente $K \in \mathcal{T}_h$ die Abschätzung $h_K \leq \kappa \rho_K$ gilt. Der konforme Finite-Elemente-Raum X_h werde durch stückweise polynomiale Basisfunktionen vom Grad $k \in \mathbf{N}$ auf $\mathcal{T}_h$ erzeugt. Dann existieren Konstanten $C = C(k, t, \kappa)$ mit*

$$\|v_h\|_{t,K} \leq C h^{m-t} \|v_h\|_{m,K} \qquad \forall v_h \in X_h, \quad 0 \leq m \leq t. \tag{13.5}$$

Beweis: vgl. D. Braess [5], Lemma II.6.8. □

Ein wichtige Eigenschaft *strukturierter Gitter* ist, daß die Position der Nichtnullelemente sehr einfach bestimmt werden kann. Man speichert daher sehr oft diese Elemente gar nicht ab, sondern berechnet sie bei Bedarf einfach neu.

Eine andere Situation liegt bei Verwendung *unstrukturierter Gitter* vor. Die nachfolgende Abbildung 13.2

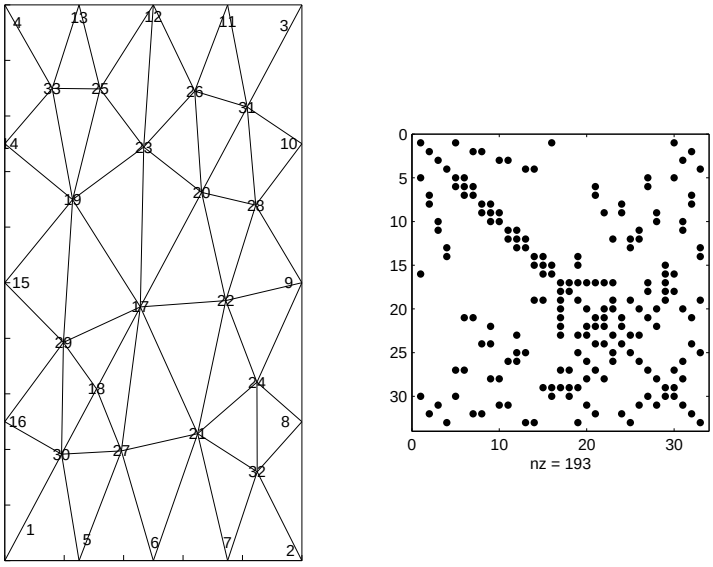


Abbildung 13.2: Unstrukturierte Zerlegung und zugehörige Matrixstruktur

zeigt eine derartige Situation und die resultierende Steifigkeitsmatrix mit unregelmäßigem Besetztheitsmuster. □

13.2 Angepaßte direkte Lösungsverfahren

Die Überlegungen aus dem letzten Abschnitt zeigen, daß eine Anpassung von Standard-Lösungsverfahren für lineare Gleichungsysteme an die spezielle Struktur dieser Systeme unumgänglich ist. Nachfolgend betrachten wir einige wichtige Varianten. Für die weiteren Darlegungen werden wir im linearen Gleichungssystem (13.1) jeweils den Index h weglassen.

(i) Gauß-Elimination für Bandmatrizen

Bei vollbesetzter Matrix des linearen Gleichungssystems

$$Au = f, \qquad A = (a_{ij}) \in \mathbf{R}^{N\times N} \tag{13.6}$$

erfordert die Anwendung der auf der Gauß-Elimination basierenden direkten Verfahren Speicherplatz für $\mathcal{O}(N^2)$ Nichtnullelemente und $\mathcal{O}(N^3)$ wesentliche Rechenoperationen. Das ist für den hier interessierenden Fall $N \gg 1$ natürlich nicht akzeptabel. Eine Anpassung an die schwachbesetzte Struktur der Matrix A is also unbedingt erforderlich.

Exemplarisch betrachten wir den Fall, daß die Matrix A eine *Bandstruktur* mit $m \ll N$ hat, d.h.

$$a_{ik} = 0, \quad |i-k| > m. \tag{13.7}$$

Durch geeignete Umordnung von Zeilen und/ oder Spalten der Matrix kann man eventuell die Bandbreite reduzieren (vgl. z.B. Y. Saad [20], Kapitel 3.3).

Die einfachste Methode der Aufwandsreduzierung beim Gauß-Algorithmus ist eine Anpassung an die Bandstruktur der Matrix. Vereinfachend sei die Matrix stark diagonaldominant, d.h.

$$|a_{ii}| > \sum_{k\neq i} |a_{ik}|, \quad i = 1, ..., N.$$

Dann kann auf die Matrix ohne vorherige Pivotisierung (d.h. ohne Umordnung von Zeilen oder Spalten) die $LU-$Zerlegung mit

$$A = LU \tag{13.8}$$

angewendet werden. Für die linke untere bzw. rechte obere Dreiecksmatrix $L = (l_{ij})$ bzw. $U = (u_{ij})$ gilt

$$l_{ij} = 0, \quad j > i \text{ oder } j < i - m \quad \text{bzw.} \quad u_{ij} = 0, \quad j < i \text{ oder } j > i + m. \tag{13.9}$$

Dies liefert vereinfachend

$$a_{ik} = \sum_{j=\max\{i,k\}-m}^{\min\{i,k\}} l_{ij}u_{jk}, \quad i,k = 1,...,N. \tag{13.10}$$

Mit der Festsetzung $l_{ii} = 1$ ergibt sich durch rekursive Auflösung von (13.10) für $\min(i,k) = 1,...,N$, daß

$$u_{ik} = a_{ik} - \sum_{j=\max\{1,k-m\}}^{i-1} l_{ij}u_{jk}, \quad k = i,...,i+m \tag{13.11}$$

und

$$l_{ik} = \frac{1}{u_{kk}}\left(a_{ik} - \sum_{j=\max\{1,i-m\}}^{k-1} l_{ij}u_{jk}\right), \quad i = k+1,...,k+m. \tag{13.12}$$

Bemerkung 13.3. (*Cholesky-Zerlegung*)
Für eine symmetrische und positiv definite Matrix A existiert eine eindeutig bestimmmte untere Dreiecksmatrix L derart, daß $A = LL^*$. (Für die Eindeutigkeit wird ohne Beschränkung der Allgemeinheit gefordert, daß die Hauptdiagonalelemente von L positiv sind.) □

Beispiel 13.4. Bei der Diskretisierung partieller Differentialgleichungen gelingt es nicht selten (z.B. beim *ADI-Verfahren – alternating direction iteration*), vereinfachte Probleme mit Tridiagonalmatrizen zu erzeugen. In diesem Fall vereinfacht sich die Rekursion (13.11),(13.12) mit $u_{11} = a_{11}$ und $u_{NN} = a_{NN} - l_{NN}u_{N-1,N}$ zu

$$u_{ii} = a_{ii} - l_{i,i-1}u_{i-1,i}, \quad u_{i,i+1} = a_{i,i+1}, \quad l_{i+1,i} = \frac{a_{i+1,i}}{u_{ii}}, \quad i = 2,...N-1.$$

Dies ist der sogenannte *Thomas-Algorithmus.* Bei dieser speziellen LU-Zerlegung kommt man mit $\mathcal{O}(N)$ wesentlichen Rechenoperationen aus. Das ist bei N Unbekannten ein optimales Ergebnis, ist aber an die sehr spezielle Struktur der Matrix gebunden. □

(ii) Frontlösungsverfahren:

Eine interessante Idee zur Reduktion des Datentransfers besteht darin, die Assemblierung der Steifigkeitsmatrix A bereits mit der LU-Zerlegung zu verbinden. Die Grundidee kann bereits am Fall von Bandmatrizen erläutert werden. Bei der LU-Zerlegung nach Formel (13.10) benötigt man nämlich nur die ersten m Zeilen und Spalten von A, um bereits mit der LU-Zerlegung zu beginnen. Die entsprechenden Elemente von L und U sind dann bereits bestimmbar. Dann setzt man dieses Verfahren sukzessiv fort. Man spricht von *Frontlösungsverfahren.* Eine elementare Beschreibung findet man bei [9], Abschnitt 3.1.3.

Die Effizienz dieser Technik basiert auf der Bestimmung der Elementmatrizen. Sind zu einem Gitterpunkt x_i bereits alle zugehörigen Elementmatrizen bekannt, so verfügt man damit über alle nichtverschwindenden Einträge der i-ten Zeile bzw. Spalte von A. Damit kann gleitend mit der Assemblierung auch die LU-Zerlegung durchgeführt werden. Insbesondere bietet sich hier eine effiziente Cache-Nutzung eines Rechners an.

Die Technik der Frontlösungsverfahren ist in den letzten Jahren systematisch weiterentwickelt worden. Insbesondere können sie auch bereits assemblierte Matrizen A angewendet werden. Im Programmpaket UMFPACK von T. Davis und I. Duff, das in den neuen Versionen von MATLAB angeboten wird, werden *uni-* und *multifrontale* Methoden geschickt kombiniert. Man konsultiere hierzu die Webseite

`http://www.cise.ufl.edu/research/sparse/umfpack/`.

Es sei unbedingt hervorgehoben, daß mit diesen Entwicklungen angepaßter direkter Lösungsverfahren eine deutliche Verschiebung ihrer Anwendungsgrenzen gegenüber iterativen Verfahren eingetreten ist. Insbesondere auch durch die zunehmende Verwendung finiter Elemente höherer Ordnung, die natürlich einen höheren Besetzungsgrad der Matrix A bewirken, sind diese Verfahren deutlich attraktiver geworden.

(iii) Approximative $LU-$Zerlegung (ILU)

Eine sehr negative Eigenschaft der $LU-$Zerlegung für schwachbesetzte Matrizen ist das sogenannte *"fill in"*. Im allgemeinen Fall werden dabei Nichtnullelemente auch an Positionen von L bzw. U erzeugt, an denen A Nullelemente besitzt. Es gibt Pivotstrategien, die das "fill in" reduzieren.

Eine wichtige Variante, die im Zusammenhang mit der *Vorkonditionierung* iterativer Verfahren auftritt, ist eine approximative $LU-$Zerlegung (*ILU – incomplete LU factorization*). Dabei wird das "fill in" außerhalb des Besetzungsmusters von A durch eine geeignete Löschmaske ignoriert. Man erhält die Zerlegung

$$A = \tilde{L}\tilde{U} - R,$$

wobei $\tilde{L}$ bzw. $\tilde{R}$ eine untere bzw. obere Dreiecksmatrix mit schwacher Besetztheit sind. Man hofft, daß der Einfluß der Restmatrix R gering ist.

Spezielle Varianten unterscheiden sich durch die Wahl der Löschmaske. Von dieser hängt auch ab, wieviel Information der Ausgangsmatrix A auf die approximative LU-Zerlegung übertragen wird. Beim $ILU(0)-$Verfahren wird die Löschmaske so gewählt, daß alle Nichtnullelemente außerhalb des originalen Besetzungsmusters von A ignoriert werden. Beim $ILU(p)-$Verfahren berücksichtigt man zusätzlich Nichtnullelemente in weiteren Nebendiagonalen.

Wir zitieren einen typischen Satz (vgl. Y. Saad [20]):

Satz 13.5. *Seien A eine $M-$Matrix und $\mathcal{P}$ eine geeignete Löschmaske, die jedoch keine Hauptdiagonalelemenete enthält. Dann ist die ILU-Zerlegung $A = \tilde{L}\tilde{U} - R$ ausführbar und regulär, d.h. $\tilde{L}\tilde{U}$ ist nichtsingulär und $(\tilde{L}\tilde{U})^{-1}$ sowie R sind nichtnegativ.*

Wir nehmen nun an, daß das Ausgangssystem (13.6) zum Beispiel von links mit der (als regulär vorausgesetzten) Matrix $(\tilde{L}\tilde{U})^{-1}$ multipliziert wird. Dann besteht die Hoffnung, daß bei geeignetem R die Matrix $\tilde{A} := (\tilde{L}\tilde{U})^{-1}LU$ nicht sehr stark von der Einheitsmatrix abweicht und eine wesentlich günstigere Kondition als A besitzt. Dies ist die Idee der *Vorkonditionierung* mittels $ILU-$Verfahren, die wir in den folgenden Kapiteln wieder aufgreifen.

(iv) Schnelle Fourier-Transformation für das diskrete Poisson-Problem

Wir betrachten das lineare Gleichungssystem

$$Au = f \tag{13.13}$$

mit regulärer, symmetrischer Matrix $A \in \mathbf{R}^{N\times N}$. Für spezielle Aufgaben, die eventuell bei Separationsansätzen auf speziellen Gebieten (z.B. Quader) entstehen, gelingt die Bestimmung eines vollständigen Orthogonalsystems $\{\phi^l\}_{l=1}^N$ von Eigenvektoren von A mit den zugehörigen Eigenwerten $\{\lambda_l\}_{l=1}^N$. Dann besitzt (13.13) die Lösungsdarstellung

$$u = \sum_{l=1}^{N} c_l\phi^l, \qquad c_l = \frac{1}{\lambda_l}\frac{(f,\phi^l)}{(\phi^l,\phi^l)}, \qquad l = 1,...,N \tag{13.14}$$

mit dem Euklidschen Skalarprodukt $(\cdot,\cdot)$ im $\mathbf{R}^N$. Bei vorliegenden Symmetrien im Problem ist es möglich, die in (13.14) auftretenden Summen effektiv zu berechnen. Die Idee wird am eindimensionalen Fall des Poisson-Problems dargestellt.

Beispiel 13.6. Bei Diskretisierung der Zweipunkt-Randwertaufgabe

$$-u''(x) = f(x) \quad \text{in} \quad \Omega = (0,1), \qquad u(0) = u(1) = 0$$

auf einem äquidistanten Gitter mit $h > 0$ entsteht das System

$$-u_{j-1} + 2u_j - u_{j+1} = h^2 f_j, \quad j = 1, ..., M-1; \qquad u_0 = u_M = 0.$$

Wir setzen dabei $N := M - 1$. Die Eigenvektoren $\phi^l = (\phi_j^l)_{j=1}^N$ der zugehörigen Matrix A sind Lösung des Systems

$$-\phi_{j-1}^l + 2\phi_j^l - \phi_{j+1}^l = \lambda_l \phi_j^l, \quad j = 1, ..., M-1.$$

Mit dem Ansatz $\phi_j^l := e^{i\rho_l j}$ gelangt man zu der Gleichung

$$2(1 - \cos\rho_l) e^{i\rho_l j} = \lambda_l e^{i\rho_l j}.$$

Die Randbedingungen $\phi_0^l = \phi_M^l = 0$ führen auf die Gleichung

$$\sin(\rho_l M) = 0, \quad l = 1, 2, ...$$

mit den Lösungen $\rho_l = \pm\frac{l\pi}{M},\ l = 1, 2, ...$ Für die Eigenwerte gilt damit

$$\lambda_l = 2\left(1 - \cos\frac{l\pi}{M}\right) = 4\sin^2\frac{l\pi}{2M}, \quad l = 1, ..., M-1.$$

Die entsprechenden reellwertigen Eigenvektoren $\phi^l \in \mathbf{R}^{M-1}$ haben die Komponenten

$$\phi_j^l = \sin\frac{l\pi j}{M}, \quad l, j = 1, ..., M-1.$$

Nachrechnen zeigt die Orthogonalität $(\phi^l, \phi^m) = 0, j \neq m$ sowie $(\phi^l, \phi^l) = \frac{M}{2}$. In (13.14) ergibt dies

$$c_l = \frac{1}{2M\sin^2\left(\frac{l\pi}{2M}\right)} \sum_{j=1}^{M-1} f_j \sin\left(\frac{l\pi j}{M}\right), \quad l = 1, ..., M-1.$$

Man benötigt dann eine effektive Berechnung von Summen der Form

$$c_l = \sum_{j=1}^{M-1} g_j \sin\left(\frac{l\pi j}{M}\right), \quad l = 1, ..., M-1.$$

Auch die Berechnung der Komponenten u_j der gesuchten diskreten Lösung $u_h = (u_j)_{j=1}^{M-1}$ führt auf derartige Summen.

Eine geeignete Lösung des Problems ist mit der *schnellen Fourier-Transformation* (FFT) möglich. Dabei nutzt man geschickt Symmetrien der Winkelfunktionen. Eine Darstellung findet man z.B. bei Hanke-Bourgeois [12], Abschnitt 53. Die FFT gehört heute zur Standardsoftware. □

Für einige Probleme kann die Methode auf den mehrdimensionalen Fall übertragen werden. Läßt sich das Ausgangsproblem mittels Separationsansatz auf eindimensionale Aufgaben reduzieren, so kann für diese jeweils die im Beispiel angegebene Vorgehensweise genutzt werden.

Kapitel 14

Klassische iterative Verfahren

Im vorliegenden Abschnitt betrachten wir grundlegende *iterative Verfahren* für lineare Gleichungssysteme, die bei elliptischen Randwertproblemen auftreten. Besonderer Wert ist auf die Möglichkeiten und Grenzen derartiger Verfahren zu legen. Beweise werden auch hier nicht geführt; dazu wird auf die Vorlesung *Numerische Mathematik I* verwiesen.

14.1 Grundstruktur iterativer Verfahren

Entscheidender Vorteil *iterativer Verfahren* zur Lösung linearer Gleichungssysteme gegenüber direkten Verfahren ist, daß das Besetzungsmuster der Matrix nicht verändert wird. Das zu lösende System

$$Au = f$$

mit regulärer Matrix $A \in \mathbf{R}^{N\times N}$ wird mit einer zu spezifizierenden regulären Matrix B in der folgenden äquivalenten Form

$$Bu = (B - A)u + f$$

geschrieben. Zur iterativen Berechnung der Lösung betrachten wir das Verfahren

$$Bu^{k+1} = (B - A)u^k + f \tag{14.1}$$

mit geeignetem Startvektor $u^0 \in \mathbf{R}^N$. Hieraus ergeben sich folgende *Anforderungen* an die Wahl von B

- effiziente Lösbarkeit der Systeme $Bv = g$,
- schnelle Berechenbarkeit von $g := (B - A)v + f$,
- möglichst kleine Kondition der Iterationsmatrix $B^{-1}(B - A)$.

Zur *Konvergenzuntersuchung* mit dem Fixpunktsatz von S. Banach führen wir eine geeignete Vektornorm $\|\cdot\|$ ein. Typische Varianten sind

$$\|v\|_p := \left(\sum_{j=1}^{N} |v_j|^p\right)^{\frac{1}{p}}, \quad 1 \le p < \infty, \tag{14.2}$$

die Maximumnorm

$$\|v\|_\infty := \max_{j=1,\dots,N} |v_j| \tag{14.3}$$

oder bei symmetrischer und positiv definiter Matrix A die *energetische Norm*

$$\|v\|_A := (v^* Av)^{1/2} . \tag{14.4}$$

Wir werden nachfolgend die Vektornorm $\|\cdot\|$ nicht indizieren, sofern es der Zusammenhang nicht erfordert. Ausdrücklich sei noch vermerkt: Alle Vektornormen sind zwar (über $\mathbf{R}^N$) äquivalent, jedoch hängen die entsprechenden Äquivalenzkonstanten von N bzw. h ab.

Weiter verwenden wir die durch

$$\|C\| := \sup_{v \in \mathbf{R}^N \setminus \{0\}} \frac{\|Cv\|}{\|v\|} \tag{14.5}$$

erklärte *Matrixnorm.*

Eine hinreichende Konvergenzbedingung des iterativen Verfahrens (14.1) folgt aus dem Fixpunktsatz von S. Banach für das Iterationsverfahren

$$u^{k+1} = Tu^k + t, \qquad k \in \mathbf{N}_0 \tag{14.6}$$

mit der Iterationsmatrix $T := B^{-1}(B - A)$ sowie $t := B^{-1}f$.

Satz 14.1. *Gelte $\|T\| < 1$. Dann hat die Fixpunktaufgabe $u = Tu + t$ eine eindeutige Lösung $u \in \mathbf{R}^N$. Für beliebige Startvektoren $u^0 \in \mathbf{R}^N$ konvergiert die Folge (u^k) aus (14.6) gegen u. Ferner gelten die a-priori Abschätzung*

$$\|u^{k+1} - u\| \leq \frac{\|T\|^k}{1 - \|T\|} \, \|u^1 - u^0\|, \quad k \in \mathbf{N}_0 \tag{14.7}$$

sowie die a-posteriori Abschätzung

$$\frac{1}{1 + \|T\|}\|u^{k+1} - u^k\| \leq \|u^k - u\| \leq \frac{1}{1 - \|T\|}\|u^{k+1} - u^k\|, \quad k \in \mathbf{N}_0. \tag{14.8}$$

Mit Hilfe des durch

$$\rho(T) := \max\{|\lambda_i| : \ \lambda_i \text{ Eigenwert von } T\} \tag{14.9}$$

erklärten *Spektralradius* von T hat man auch ein notwendiges und hinreichendes Konvergenzkriterium durch

Lemma 14.2. *Die durch das Verfahren (14.6) erzeugte Folge (u^k) konvergiert genau dann für jeden Vektor $t \in \mathbf{R}^N$ und jeden Startvektor $u_0 \in \mathbf{R}^N$, wenn $\rho(T) < 1$.*

14.2 Gesamt- und Einzelschrittverfahren

Wir zerlegen dazu die Matrix A in der folgenden Form

$$A = L + D + R \tag{14.10}$$

mit der Diagonalmatrix $D = \mathrm{diag}(a_{11}, ..., a_{NN})$ sowie der unteren bzw. oberen Dreiecksmatrix

$$L = \begin{pmatrix} 0 & & & & & \\ a_{21} & 0 & & & & \\ a_{31} & a_{32} & 0 & & & \\ \vdots & \vdots & & \ddots & & \\ a_{N1} & a_{N2} & \cdot & \cdot & a_{N,N-1} & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & a_{12} & \cdot & \cdot & \cdot & a_{1N} \\ & 0 & a_{23} & \cdot & \cdot & a_{2N} \\ & & \ddots & \cdot & \cdot & \cdot \\ & & & & 0 & a_{N-1,N} \\ & & & & & 0 \end{pmatrix}.$$

Für unsere Betrachtungen setzen wir voraus, daß (eventuell nach Vertauschung von Zeilen und Spalten der Matrix A) die inverse Matrix D^{-1} von D existiert. Dies ist äquivalent dazu, daß kein Element der Hauptdiagonale von A verschwindet.

Die beiden einfachsten Iterationsverfahren sind

- das *Gesamtschrittverfahren (Jacobi-Verfahren)* mit $B = D$ und

- das *Einzelschrittverfahren (Gauß-Seidel-Verfahren)* mit $B = L + D$.

Wir geben jetzt einige Konvergenzkriterien an, die auf der Abschätzung des Spektralradius $\rho(T)$ der Iterationsmatrix $T := B^{-1}(B - A)$ beruhen.

Lemma 14.3. *Die Matrix $A = (a_{ik})$ sei stark diagonaldominant, d.h.*

$$q_\infty := \max_{i=1,\ldots,N} q_i < 1, \quad q_i := \sum_{\substack{k=1 \\ k \neq i}}^{N} \left| \frac{a_{ik}}{a_{ii}} \right|. \tag{14.11}$$

Dann gilt für das Gesamt- bzw. Einzelschrittverfahren bezüglich der Maximumnorm

$$\|B^{-1}(B - A)\|_\infty \leq q_\infty < 1.$$

Bei der Diskretisierung elliptischer Randwertaufgaben hat man oft nur die Eigenschaft der *schwachen Diagonaldominanz* der Matrix A, d.h.

$$q_i \leq 1, \quad i = 1, ..., N: \qquad \exists j \in \{1, ..., N\}: \quad q_j < 1. \tag{14.12}$$

Ferner sei die Matrix unzerlegbar gemäß

Definition 14.4. *Eine Matrix $A = (a_{ik}) \in \mathbf{R}^{N \times N}$ heißt* unzerlegbar (oder irreduzibel), *falls sie nicht durch Umordnung von Zeilen und Spalten (d.h. durch Umnumerierung der Unbekannten) in der folgenden (entkoppelten) Form geschrieben werden kann:*

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}.$$

Beispiel 14.5. Tridiagonalmatrizen, bei denen die Elemente der Hauptdiagonale und der beiden Nebendiagonalen sämtlich nicht verschwinden, sind unzerlegbar.

Lemma 14.6. *Die Matrix $A = (a_{ik})$ sei unzerlegbar und schwach diagonaldominant. Dann konvergiert das Gesamtschrittverfahren.*

Man kann zeigen, daß diese Aussage auch für das Einzelschrittverfahren richtig bleibt (vgl. D.M. Young: *Iterative solution of large linear systems*, Academic Press, New York 1971, S. 108). Für das Einzelschrittverfahren gilt ferner noch

Lemma 14.7. *Die Matrix $A = (a_{ik})$ sei symmetrisch und positiv definit. Dann konvergiert das Einzelschrittverfahren.*

Die genannten Konvergenzkriterien erfordern einen *kritischen Kommentar*: Die Konvergenz der Verfahren ist umso schneller, je kleiner die Norm $\|T\| = \|B^{-1}(B - A)\|$ ist. Für das Gesamt– und Einzelschrittverfahren kann jedoch am Beispiel 13.1 gezeigt werden, daß gilt

$$\rho\left(B_h^{-1}(B_h - A_h)\right) = 1 - \mathcal{O}(h^2). \tag{14.13}$$

Daher sind auch diese beiden grundlegenden Iterationsverfahren bei sehr feiner Diskretisierung, d.h. für $N \gg 1$ nicht geeignet.

Bevor wir uns in den restlichen Abschnitten der Vorlesung mit besser konvergierenden Verfahren befassen, sollen noch einige Varianten diskutiert werden.

- Die Auflösung beim Einzelschrittverfahren hängt von der gewählten Numerierung der Variablen und somit der Gitterpunkte ab. Selbst bei symmetrischer Matrix A kann das zu Asymmetrien in der Lösungsfolge (u^k) führen. Eine Verbesserung erzielt man beim *symmetrischen Gauß–Seidel Verfahren* durch Verbindung einer Vorwärts- und einer Rückwärtsvariante in den Halbschritten

$$\begin{aligned} (L + D)u^{k+1/2} + Ru^k &= f \\ Lu^{k+1/2} + (D + R)u^{k+1} &= f. \end{aligned}$$

- Eine andere Variante der Symmetrisierung auf regelmäßigen Gittern besteht in der *Schachbrett-Iteration* (oder red–black Iteration).
- Bei Diskretisierung auf einem regelmäßigen Gitter (vgl. Beispiel 13.1) besitzt das Gleichungssystem eventuell eine *Blockstruktur*

$$\sum_{j=1}^{M} A_{ij}U_j = F_i, \quad i = 1, ..., M: \qquad F_i, U_i \in \mathbf{R}^{N_i}, \quad A_{ij} \in \mathbf{R}^{N_i \times N_i} \tag{14.14}$$

mit $\sum_{i=1}^{M} N_i = N$. Bei regulären Matrizen A_{ii} lassen sich die *Blockvarianten*

$$A_{ii}U_i^{k+1} + \sum_{j \neq i} A_{ij}U_j^k = F_i, \quad i = 1, ..., M$$

bzw.

$$\sum_{j=1}^{i} A_{ij}U_j^{k+1} + \sum_{j=i+1}^{M} A_{ij}U_j^k = F_i, \quad i = 1, ..., M$$

des Gesamt- bzw. Einzelschrittverfahrens angeben sowie entsprechende Konvergenzsätze formulieren.

14.3 Relaxations-Verfahren

Die einfachste Variante der *Konvergenzbeschleunigung* grundlegender iterativer Verfahren, wie des Gesamt- und Einzelschrittverfahrens, stellen *Relaxationsverfahren* dar. Die entsprechende Modifikation des Gesamtschrittverfahrens, d.h. für $B = D$, ist die *Jacobi-Relaxation*

$$u^{k+1} = (I - \omega D^{-1}A)u^k + \omega D^{-1}f. \tag{14.15}$$

Auskunft über die Konvergenz des Verfahrens und die geeignete Wahl des Relaxationsparameters ω gibt das folgende

Lemma 14.8. *Die zum Gesamtschrittverfahren gehörende Matrix $G := -D^{-1}(L + R)$ habe nur reelle Eigenwerte und einen Spektralradius $\rho(G) < 1$. Dann wird der Spektralradius der Matrix des Relaxationsverfahrens*

$$I - \omega D^{-1}A = (1 - \omega)I - \omega D^{-1}(L + R) \tag{14.16}$$

minimal bei

$$\omega_0 = \frac{2}{2 - \lambda_{min} - \lambda_{max}}. \tag{14.17}$$

Dabei sind λ_{min} bzw. λ_{max} der kleinste bzw. größte Eigenwert der Matrix G. Speziell weist bei $\lambda_{min} \neq -\lambda_{max}$ das Relaxationsverfahren eine schnellere Konvergenz als das Gesamtschrittverfahren auf.

Die entsprechende Relaxationsvariante für das Einzelschrittverfahren, d.h. für $B = D + L$, ist das *SOR-Verfahren („successive overrelaxation")*

$$u^{k+1} = u^k + \omega D^{-1}(f - Lu^{k+1} - (D + R)u^k), \quad k = 0, 1, 2, ... \tag{14.18}$$

Aus der Darstellung

$$(D + \omega L)u^{k+1} = \omega f + \{(1 - \omega)D - \omega R\}u^k$$

findet man die Gestalt der Iterationsmatrix zu

$$T(\omega) = (D + \omega L)^{-1}\{(1 - \omega)D - \omega R\}. \tag{14.19}$$

Offenbar hängt die Iterationsmatrix nicht linear ab vom Parameter ω, im Unterschied zum Gesamtschritt-Relaxationsverfahren. Daher bereitet die Wahl von ω auch größere Probleme.

Lemma 14.9. *Das SOR-Verfahren konvergiert höchstens im Parameterintervall $0 < \omega < 2$. Im Falle einer symmetrischen und positiv definiten Matrix A konvergiert das SOR-Verfahren für alle Werte aus diesem Intervall.*

Nachfolgend wird die Parameterwahl im Fall *konsistent geordneter Matrizen* behandelt. Dieser Fall schließt den wichtigen Fall nichtsingulärer Tridiagonalmatrizen ein.

Definition 14.10. *Eine Matrix $A = D + L + R$ heißt* konsistent geordnet, *falls die Eigenwerte der Matrix*

$$C(\alpha) := -\alpha D^{-1}L - \frac{1}{\alpha}D^{-1}R, \ \alpha \neq 0 \tag{14.20}$$

unabhängig vom Parameterwert α sind.

Lemma 14.11. *Tridiagonalmatrizen mit nichtsingulärer Diagonalmatrix sind konsistent geordnet.*

Das gesuchte Resultat für das SOR-Verfahren im Fall konsistent geordneter Matrizen ist

Lemma 14.12. *Die Matrix A sei konsistent geordnet. Die Eigenwerte der Iterationsmatrix des Jacobi-Verfahrens $-D^{-1}(L + R)$ seien sämtlich reell. Für deren Spektralradius gelte $\Lambda := \rho(-D^{-1}(L + R)) < 1$. Dann konvergiert das SOR-Verfahren für alle Parameterwerte $0 < \omega < 2$. Der Spektralradius der Iterationsmatrix*

$$T(\omega) = (D + \omega L)^{-1}\{(1 - \omega)D - \omega R\} \tag{14.21}$$

wird minimiert im Fall

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \Lambda^2}} > 1,$$

und es gilt

$$\rho(T(\omega_0)) = \frac{1 - \sqrt{1 - \Lambda^2}}{1 + \sqrt{1 - \Lambda^2}}.$$

Wir nehmen an, daß für das Gesamtschrittverfahren gilt $\Lambda = 1 - \mathcal{O}(h^2)$. Durch die Relaxation erhält man die verbesserte Abschätzung $\rho(T(\omega_0)) = 1 - \mathcal{O}(h)$.

Eine wichtige Variante des SOR-Verfahrens ist dessen symmetrische Version, das *SSOR-Verfahren*. Für die Iterationsmatrix erhält man

$$T^s(\omega) := (D + \omega R)^{-1}D(D + \omega L)^{-1} \cdot [(1 - \omega)D - \omega L]D^{-1}[(1 - \omega)D - \omega R]. \tag{14.22}$$

Im Fall einer symmetrischen, positiv definiten Matrix A gilt ebenfalls die Aussage von Lemma 14.9.

14.4 Kritische Wertung der Basisverfahren

Ein wichtiger *Vorteil* der grundlegenden Iterationsverfahren gegenüber direkten Eliminationsverfahren ist, daß das Besetzungsmuster (in der Regel schwache Besetztheit) nicht geändert wird. Insbesondere tritt das "fill in"-Problem nicht auf.

Ein grundlegender *Nachteil* der Basisiterationsverfahren ist aber, daß sie sehr langsam konvergieren. So erhält man bei der Diskretisierung elliptischer Randwertprobleme 2. Ordnung die folgende scharfe Abschätzung für die Iterationsmatrix

$$\|B_h^{-1}(B_h - A_h)\| = 1 - \mathcal{O}(h^\alpha)$$

mit $\alpha = 2$ im allgemeinen Fall bzw. $\alpha = 1$ bei Relaxation. Damit sind diese Verfahren für große Dimension $N = \mathcal{O}(h^{-d}) \gg 1$ des Gleichungssystems (bei Raumdimension d) völlig ungeeignet.

Die genannten Iterationsverfahren sind jedoch durchaus von großem Interesse bei der *Vorkonditionierung* anderer Lösungsverfahren. Die herausragende Eigenschaft der Basisverfahren ist die schnelle *Glättung hochfrequenter Fehleranteile*. So zeigt die Abbildung 14.1, wie beim Poisson-Problem der Fehler durch

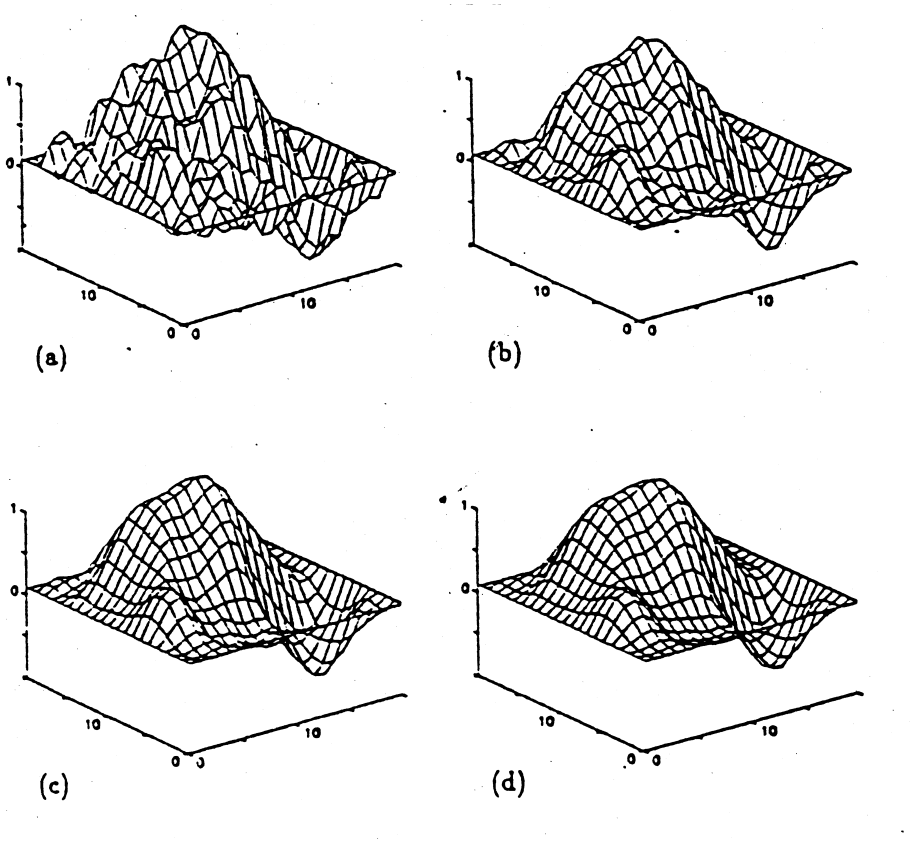


Abbildung 14.1: Glättungseigenschaft des Jacobi-Verfahrens

das Gesamtschrittverfahren tatsächlich schnell geglättet wird. Die niederfrequenten Anteile werden jedoch nur sehr langsam verbessert (vgl. Abschätzung der Iterationsmatrix).

Man nutzt diese *Glättungseigenschaft* der Basisverfahren im Rahmen von *Mehrgitterverfahren* (vgl. Kap. 16). Niederfrequente Fehleranteile werden dann auf gröberen Gittern mit entsprechend geringerem Aufwand reduziert.

Hinsichtlich der praktischen Erfahrungen mit Relaxationsverfahren ist mitzuteilen, daß ihre Konvergenzgeschwindigkeit sehr erheblich von einer optimierten Wahl des Relaxationsparameters abhängt. Es zeigt sich jedoch, daß diese starke Abhängigkeit deutlich abgemildert ist, wenn man ein Relaxationsverfahren zur Vorkonditionierung von zugkräftigeren Verfahren verwendet.

Kapitel 15

Krylov-Unterraum-Methoden

Die Ergebnisse aus dem vorhergehenden Kapitel zu den klassischen Iterationsverfahren legen nahe, bei iterativen Verfahren für lineare Gleichungssysteme

$$Au = b \tag{15.1}$$

mehr Informationen über die reguläre Koeffizientenmatrix A zu verarbeiten. Im vorliegenden Kapitel befassen wir uns mit *Krylov-Unterraum-Methoden*. Bei dieser Verfahrensklasse ist vor allem die effiziente Berechenbarkeit gewisser Matrix-Vektorprodukte (z.B. Au) wesentlich.

Ihren Ursprung haben diese Methoden im *Verfahren der konjugierten Gradienten (CG-Verfahren)* von Hestenes/ Stiefel (1952) für symmetrische, positiv definite Matrizen. Es gibt inzwischen zahlreiche *Verallgemeinerungen* auf Gleichungssyteme mit nichtsymmetrischer und/ oder indefiniter Matrix. Einen Überblick findet man in dem Lehrbuch von Y. Saad [20].

15.1 Krylov-Unterräume

Krylov-Methoden basieren auf der Konstruktion von an die Matrix A angepaßten Teilräumen des $\mathbf{R}^n$.

Definition 15.1. *Für gegebene Matrix $A \in \mathbf{R}^{n\times n}$ und Vektor $v \in \mathbf{R}^n \setminus \{0\}$ wird ein* Krylov-Unterraum *definiert durch*

$$\mathcal{K}_k(A, v) := span\{v, Av, \ldots, A^{k-1}v\} = \{p(A)v : \ p \in P_{k-1}\}. \tag{15.2}$$

Sind keine Mißverständnisse möglich, schreiben wir auch $\mathcal{K}_k := \mathcal{K}_k(A, v)$.

Wesentlich für Krylov-Methoden ist die effiziente Bestimmung einer Orthonormalbasis eines Krylov-Raumes (vgl. Abschnitt 15.2). Wir beschreiben nun die *Idee* von Krylov-Methoden:

Sei u_0 eine Näherung an die Lösung des Gleichungssystems (15.1). Als Vektor v wählt man das *Residuum* bzw. den *Defekt* $r_0 := b - Au_0$. In der Regel gilt $r_0 \neq 0$, anderenfalls wäre u_0 bereits Lösung. Man sucht nun eine Näherungslösung u_k im affinen Teilraum $u_0 + \mathcal{K}_k(A, r_0)$ durch geeignete Zusatzforderungen:

- Bei *Galerkin-Verfahren* wird gefordert, daß der Defekt $r_k := b - Au_k$ orthogonal zu $\mathcal{K}_k(A, r_0)$ oder einem anderen geeigneten Krylov-Unterraum ist (*Galerkin-Bedingung*). Exemplarisch besprechen wir in Abschnitt 15.3 das FOM-Verfahren.

- Bei *Minimierungs-Verfahren* minimiert man r_k in einer passenden Norm auf $\mathcal{K}_k(A, r_0)$ oder einem anderen geeigneten Krylov-Unterraum. In Abschnitt 15.4 behandeln wir als wichtiges Beispiel das GMRES-Verfahren.

Nach Bestimmung von u_k erhöht man dann entweder k oder startet das Verfahren mit $u_0 := u_k,\ r_0 := b - Au_0$ neu (*Restart*). Man hofft insbesondere, daß $k \ll n$ ist.

Wir definieren implizit im nächsten Lemma die Dimension eines Krylov-Unterraumes. Sei dazu

$$\deg(v) := \min\{l : \exists p \in P_l \setminus \{0\} \text{ mit } p(A)v = 0\}. \tag{15.3}$$

Nach dem Satz von Caley/Hamilton gilt für das charakteristische Polynom $p(\lambda) := \det(A - \lambda I)$ der Matrix $A \in \mathbf{R}^{n\times n}$ die Aussage $p(A) = 0$. Für beliebige Vektoren $v \in \mathbf{R}^n$ folgt somit $\deg(v) \leq n$.

Lemma 15.2. *Gegeben seien die Matrix $A \in \mathbf{R}^{n\times n}$ und der Vektor $v \in \mathbf{R}^n \setminus \{0\}$ mit $m := deg(v)$. Dann gelten folgende Aussagen:*

(i) *Es gilt $A(\mathcal{K}_m) \subset \mathcal{K}_m$, d.h. der Krylov-Unterraum $\mathcal{K}_m$ ist invariant unter A. Weiter gilt $\mathcal{K}_k = \mathcal{K}_m$ für alle $k \geq m$.*

(ii) *Es gilt $dim(\mathcal{K}_k) = k$ genau für $m \geq k$.*

(iii) *Es gilt $dim(\mathcal{K}_k) = \min(k, m)$.*

Beweis: (i) Für $u \in \mathcal{K}_m$ gilt per Konstruktion $u = \sum_{i=0}^{m-1} \alpha_i A^i v$. Ferner findet man Konstanten $\beta_0, \ldots, \beta_m$, die nicht alle gleichzeitig verschwinden, so daß

$$\sum_{i=0}^{m} \beta_i A^i v = 0.$$

Wegen $\deg(v) = m$ ist $\beta_m \neq 0$. Daraus folgt wegen

$$\begin{aligned} Au &= \sum_{i=1}^{m} \alpha_{i-1} A^i v - \frac{\alpha_{m-1}}{\beta_m} \sum_{i=0}^{m} \beta_i A^i v \\ &= -\frac{\alpha_{m-1}}{\beta_m}\beta_0 v + \sum_{i=1}^{m-1} \left(\alpha_{i-1} - \frac{\alpha_{m-1}}{\beta_m}\beta_i\right) A^i v \in \mathcal{K}_m, \end{aligned}$$

daß $A(\mathcal{K}_m) \subset \mathcal{K}_m$.

Für $k \geq m$ folgt per Konstruktion $\mathcal{K}_m \subset \mathcal{K}_k$. Seien nun $k > m$ und $u \in \mathcal{K}_k$. Dann gilt $u = \sum_{i=0}^{k-1} \alpha_i A^i v$. Außerdem findet man Konstanten $\beta_0, \ldots, \beta_m$ mit $\beta_m \neq 0$ und

$$\sum_{i=0}^{m} \beta_i A^i v = 0.$$

Dies impliziert

$$\begin{aligned} u &= \sum_{i=0}^{k-1} \alpha_i A^i v - \frac{\alpha_{k-1}}{\beta_m} A^{k-m-1} \left(\sum_{i=0}^{m} \beta_i A^i v\right) \\ &= \sum_{i=0}^{k-1} \alpha_i A^i v - \frac{\alpha_{k-1}}{\beta_m} \sum_{i=0}^{m} \beta_i A^{i+k-m-1} v \ \in \ \mathcal{K}_{k-1}. \end{aligned}$$

Dieser Schluß kann bis zur Aussage $u \in \mathcal{K}_m$ fortgeführt werden. Damit ist Aussage (i) bewiesen.

(ii) Die Vektoren $\{v, Av, \ldots, A^{k-1}v\}$ bilden genau dann eine Basis von $\mathcal{K}_k$, wenn für jede Menge $\{\gamma_0, \ldots, \gamma_{k-1}\}$ nicht gleichzeitig verschwindender Zahlen die Aussage

$$\sum_{i=0}^{k-1} \gamma_i A^i v \neq 0$$

folgt. Dies entspricht aber gerade der Bedingung, daß genau das Nullpolynom p in P_{k-1} der Bedingung $p(A)v = 0$ genügt. Dies ist äquivalent zu $m = \deg(v) \geq k$.

(iii) Aussage (ii) impliziert $\dim(\mathcal{K}_k) = k = \min(k, m)$, falls $m \geq k$. Im Fall $m < k$ liefert Teil (i) die Aussage $\mathcal{K}_k = \mathcal{K}_m$, somit ist $\dim(\mathcal{K}_k) = \dim(\mathcal{K}_m) = m$. Damit ist das Lemma bewiesen. □

15.2 Arnoldi-Verfahren

Die hier zu besprechenden Krylov-Verfahren erfordern die möglichst effiziente Konstruktion einer Orthonormalbasis für einen Krylov-Unterraum

$$\mathcal{K}_k := \text{span}\{v, Av, \ldots, A^{k-1}v\},$$

wobei wir $k \ll n$ annehmen wollen. Wir betrachten das folgende modifizierte Gram-Schmidt-Verfahren. Es heißt in der aktuellen Literatur auch

Arnoldi-Verfahren:

(1) Eingabegrößen sind $A \in \mathbf{R}^{n\times n}$, $v \in \mathbf{R}^n \setminus \{0\}$ sowie $k \in \mathbf{N}$.

(2) Berechne $q_1 := v/\|v\|_2$.

(3) Für $j = 1, \ldots, k$:

- $w := Aq_j$
- Für $i = 1, \ldots, j$:
  - $h_{ij} := q_i^* w$
  - $w := w - h_{ij} q_i$.
- $h_{j+1,j} := \|w\|_2$
- Falls $h_{j+1,j} = 0$, dann: STOP.
- $q_{j+1} := w/h_{j+1,j}$.

(4) Ausgabegrößen: Ohne vorherigen Abbruch erhält man die Matrizen

$$Q_k := (\ q_1\ \cdots\ q_k\ ) \in \mathbf{R}^{n\times k}, \tag{15.4}$$

und

$$\tilde{H}_k := \begin{pmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1k} \\ h_{21} & h_{22} & \ddots & & h_{2k} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & h_{k-1,k} \\ & & & h_{k,k-1} & h_{k,k} \\ & & & & h_{k+1,k} \end{pmatrix} \in \mathbf{R}^{(k+1)\times k}. \tag{15.5}$$

Mit $H_k \in \mathbf{R}^{k\times k}$ bezeichnen wir die Matrix, die aus $\tilde{H}_k$ durch Streichen der letzten Zeile entsteht. Ferner ermittelt man auch den Vektor $q_{k+1} \in \mathbf{R}^n$. Damit ist auch die Matrix $Q_{k+1} := (\ Q_k\ q_{k+1}\ )$ wohldefiniert.

Die Eigenschaften der im Verfahren erzeugten Matrizen fassen wir zusammen im

Lemma 15.3. *Das oben beschriebene Arnoldi-Verfahren breche nicht vorzeitig an. Dann gelten folgende Aussagen:*

(i) *Die Spalten $q_1, \ldots, q_k$ von Q_k bilden eine Orthonormalbasis von $\mathcal{K}_k$.*

(ii) *Es gilt $AQ_k = Q_{k+1}\tilde{H}_k$ sowie $Q_k^* A Q_k = H_k$.*

Beweis: (i) Mittels vollständiger Induktion nach j beweisen wir, daß $\{q_1, \ldots, q_j\}$ mit $j = 1, \ldots k+1$ ein Orthonormalsystem bildet. Der Induktionsanfang für $j = 1$ ist wegen $q_1 := v/\|v\|_2$ offenbar erfüllt.

Sei nun $\{q_1, \ldots, q_j\}$ ein Orthonormalsystem. Per Konstruktion ist $\|q_{j+1}\|_2 = 1$. Somit ist noch $q_l^* q_{j+1} = 0$ für $l = 1, \ldots j$ zu zeigen. Dazu notieren wir die Berechnungsvorschrift für q_{j+1} wie folgt

- $w^{(0)} := Aq_j$.
- Für $i = 1, \ldots, j$: $w^{(i)} := w^{(i-1)} - q_i^* w^{(i-1)} q_i$.
- $q_{j+1} := w^{(j)} / \|w^{(j)}\|_2$.

Hieraus folgt für $l = 1, \ldots, j$ mit der Induktionsvoraussetzung $q_l^* q_j = \delta_{lj}$, daß

$$q_l^* w^{(j)} = q_l^* w^{(j-1)} - q_j^* w^{(j-1)} q_l^* q_j = q_l^* w^{(j-1)} - q_j^* w^{(j-1)} \delta_{lj}.$$

Damit ist $q_l^* w^{(j)} = 0$ für $l = j$, ferner gilt $q_l^* w^{(j)} = q_l^* w^{(j-1)}$ für $l < j$.
Nun schließen wir analog weiter wegen

$$q_l^* w^{(j-1)} = q_l^* w^{(j-2)} - q_{j-1}^* w^{(j-2)} q_l^* q_{j-1} = q_l^* w^{(j-2)} - q_{j-1}^* w^{(j-2)} \delta_{l,j-1}.$$

Diese Prozedur kann weitergeführt werden. Man erhält, daß $w^{(j)}$ und damit q_{j+1} orthogonal zu $q_1, \ldots, q_j$ ist. Daher ist $\{q_1, \ldots, q_{j+1}\}$ ein Orthonormalsystem.

Wir zeigen, daß $\mathcal{K}_k = \text{span}\{q_1, \ldots, q_k\}$. Hierzu wird durch vollständige Induktion nach j bewiesen, daß mit geeignetem Polynom $p_{j-1} \in P_{j-1}$ gilt $q_j = p_{j-1}(A)v$. Der Induktionsanfang für $j = 1$ folgt wegen $q_1 = v/\|v\|_2$ mit $p_0(t) := 1/\|v\|_2$.
Für den Induktionsschritt sehen wir mit der Festsetzung des Polynoms $p_j \in P_j$ mittels

$$p_j(t) := \frac{1}{\|w\|_2} \left(t p_{j-1}(t) - \sum_{i=1}^{j} h_{ij} p_{i-1}(t) \right)$$

daß

$$\begin{aligned} q_{j+1} &= \frac{w}{\|w\|_2} = \frac{1}{\|w\|_2} \left(Aq_j - \sum_{i=1}^{j} h_{ij} q_i \right) \\ &= \frac{1}{\|w\|_2} \left(A p_{j-1}(A)v - \sum_{i=1}^{j} h_{ij} p_{i-1}(A)v \right) = p_j(A)v. \end{aligned}$$

Hieraus folgt $\text{span}\{q_1, \ldots, q_k\} \subset \mathcal{K}_k$. Per Konstruktion ist $\{q_1, \ldots, q_k\}$ Orthonormalbasis von $\mathcal{K}_k$.

(ii) Im Schritt (3) des Arnoldi-Verfahrens ergibt sich, daß

$$w_j = Aq_j - \sum_{i=1}^{j} h_{ij} q_i, \qquad h_{j+1,j} q_{j+1} = w_j.$$

Damit ist $Aq_j = \sum_{i=1}^{j+1} h_{ij} q_i$ und wir können schreiben

$$AQ_k e_j = Aq_j = \sum_{i=1}^{j+1} h_{ij} q_i = Q_{k+1} \tilde{H}_k e_j, \quad j = 1, \ldots, k.$$

Damit ist $AQ_k = Q_{k+1}\tilde{H}_k$, folglich auch $Q_k^* A Q_k = Q_k^* Q_{k+1} \tilde{H}_k$. Es bleibt zu zeigen, daß $Q_k^* Q_{k+1} \tilde{H}_k = H_k$ ist. Dies folgt aber wegen

$$Q_k^* Q_{k+1} \tilde{H}_k = Q_k^* \begin{pmatrix} Q_k & q_{k+1} \end{pmatrix} \begin{pmatrix} H_k \\ h_{k+1} e_k^* \end{pmatrix} = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} H_k \\ h_{k+1} e_k^* \end{pmatrix} = H_k.$$

Daraus ergibt sich die noch fehlende Aussage $Q_k^* A Q_k = H_k$. □

Notwendige und hinreichende Abbruchbedingungen beim Arnoldi-Verfahren gibt

Lemma 15.4. *Das Arnoldi-Verfahren bricht im Schritt j genau dann ab, wenn $\deg(v) = j$. Dann ist $\mathcal{K}_j$*

ein unter A invarianter Unterraum.

Beweis: Gelte $\deg(v) = j$. Nach Lemma 15.2 hat man $\dim(\mathcal{K}_j) = j$, das Arnoldi-Verfahren kann also nicht vor dem Schritt j abgebrochen sein. Es bricht jedoch zwingend im Schrit j ab. Sonst könnte der normierte und zu $q_1, \ldots, q_j$ orthogonale Vektor q_{j+1} ermittelt werden. Dann wäre im Widerspruch zu Aussage (i) von Lemma 15.2 $\dim(\mathcal{K}_{j+1}) = j + 1$.

Wir nehmen nun an, daß das Arnoldi-Verfahren im Schritt j abbricht. Nach Definition des Grades wäre dann $\deg(v) \leq j$. Tatsächlich ist $\deg(v) = j$, denn sonst wäre der Algorithmus schon in einem früheren Schritt abgebrochen. □

15.3 FOM-Verfahren

Wir besprechen jetzt exemplarisch für die Klasse von Krylov-Verfahren vom *Galerkin-Typ* das sogenannte FOM-Verfahren. Ausgangspunkt ist der zum Residuum $r_0 := b - Au_0$ einer Startlösung $u_0 \in \mathbf{R}^n$ gehörige Krylov-Unterraum

$$\mathcal{K}_k := \mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}.$$

Beim FOM-Verfahren bestimmt man eine Näherung $u_k \in u_0 + \mathcal{K}_k$ so, daß $b - Au_k \perp \mathcal{K}_k$. Es basiert auf dem folgenden technischen Resultat.

Lemma 15.5. *Sei $\dim(\mathcal{K}_k) = k$. Mit dem Arnoldi-Verfahren seien die Matrix $Q_k = (q_1 \ \cdots \ q_k) \in \mathbf{R}^{n\times k}$ und die reduzierte obere Hessenberg-Matrix $H_k \in \mathbf{R}^{k\times k}$ mit*

$$Q_k^* Q_k = I, \qquad \mathcal{K}_k = span\{q_1, \ldots, q_k\}, \qquad Q_k^* A Q_k = H_k$$

ermittelt worden, insbesondere ist $q_1 = r_0/\|r_0\|_2$. Ferner sei H_k nichtsingulär. Dann gelten für den Vektor

$$u_k := u_0 + Q_k H_k^{-1}(\|r_0\|_2 e_1) \tag{15.6}$$

die Aussagen $u_k \in u_0 + \mathcal{K}_k$ und $b - Au_k \perp \mathcal{K}_k$.

Beweis: Die Spalten von Q_k bilden eine Basis des Krylov-Unterraums $\mathcal{K}_k$. Daher ist $u_k \in u_0 + \mathcal{K}_k$. Da $\{q_1, \ldots, q_k\}$ Basis von $\mathcal{K}_k$ ist, gilt $b - Au_k \perp \mathcal{K}_k$ genau bei $Q_k^*(b - Au_k) = 0$. Die letztere Beziehung gilt wegen

$$\begin{aligned} Q_k^*(b - Au_k) &= Q_k^* r_0 - \underbrace{Q_k^* A Q_k H_k^{-1}}_{=I}(\|r_0\|_2 e_1) = Q_k^* r_0 - \|r_0\|_2 e_1 \\ &= Q_k^*\,(r_0 - \|r_0\|_2 Q_k e_1) = Q_k^*\,(r_0 - \|r_0\|_2 q_1) = 0. \end{aligned}$$

Daraus folgt die Behauptung. □

Auf Basis des Arnoldi-Verfahrens erhält man dann das folgende Verfahren zur Lösung des linearen Gleichungssystems.

"Full Orthogonalization Method" (FOM) Arnoldi-Verfahren:

(1) Berechne für die Startlösung u_0 den Defekt $r_0 := b - Au_0$ sowie $q_1 := r_0/\|r_0\|_2$.
Initialisiere

$$\tilde{H}_k = (h_{ij})_{\substack{1\leq i\leq k+1\\ 1\leq j\leq k}} := 0.$$

(2) Für $j = 1, \ldots, k$:

- $w := Aq_j$
- Für $i = 1, \ldots, j$:
  - $h_{ij} := q_i^* w$
  - $w := w - h_{ij} q_i$.

- $h_{j+1,j} := \|w\|_2$
- Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zu Schritt (3).
- $q_{j+1} := w/h_{j+1,j}$.

(3) Setze $Q_k := (\ q_1\ \cdots\ q_k\ ) \in \mathbf{R}^{n\times k}$, $H_k := (h_{ij})_{1\le i,j\le k}$ und berechne

$$u_k := u_0 + Q_k H_k^{-1}(\|r_0\|_2 e_1).$$

Das im Fall $k \ll n$ im Vergleich zum Ausgangsproblems (15.1) niedrigdimensionale System

$$H_k y = \|r_0\|_2 e_1, \tag{15.7}$$

kann mittels Givens-Rotationen oder auch einem direkten Eliminationsverfahren effizient realisiert werden, vergleiche hierzu auch Abschnitt 15.4. Der wesentliche Aufwand des Verfahrens liegt jedoch bei $k \ll n$ im Schritt (2) bei der Berechnung der Matrix-Vektorprodukte Aq_j.

15.4 GMRES-Verfahren

Wir behandeln nun eine alternative Methode zur Lösung des Problems (15.1). Wir benutzen die Bezeichnungen und den Ansatz aus dem vorhergehenden Abschnitt. Im Unterschied zur FOM wird bei *Minimierungsverfahren* die neue Lösung $u_k \in u_0 + \mathcal{K}_k$ durch den Ansatz

$$\text{Minimiere}\ \ \|b - Au\|_2, \quad u \in u_0 + \mathcal{K}_k. \tag{15.8}$$

Mittels der Orthonormalbasis $\{\ q_1, \cdots, q_k\}$ von $\mathcal{K}_k$ bzw. der Matrix $Q_k = (\ q_1\ \cdots\ q_k\ )$ erhält man die äquivalente Aufgabe

$$\text{Minimiere}\ \ J(y) := \|b - A(u_0 + Q_k y)\|_2 = \|r_0 - AQ_k y\|_2, \quad y \in \mathbf{R}^k. \tag{15.9}$$

Nach Lemma 15.3 gilt $AQ_k = Q_{k+1}\tilde{H}_k$ mit der aus dem Arnoldi-Verfahren bestimmten Matrix $\tilde{H}_k \in \mathbf{R}^{(k+1)\times k}$. Für den ersten Spaltenvektor von Q_k bzw. Q_{k+1} gilt $q_1 = r_0/\|r_0\|_2$, damit gilt

$$r_0 - AQ_k y = Q_{k+1}\left(\|r_0\|_2 e_1 - \tilde{H}_k y\right).$$

Die Spalten der Matrix Q_{k+1} sind jedoch orthonormiert, somit ist das folgende lineare Ausgleichsproblem zu lösen:

$$\text{Minimiere}\ \ J(y) := \left\| \|r_0\|_2 e_1 - \tilde{H}_k y \right\|_2, \quad y \in \mathbf{R}^k. \tag{15.10}$$

Für die unreduzierte obere Hessenberg-Matrix $\tilde{H}_k$ ist $h_{j+1,j} \neq 0$ bei $j = 1, \ldots k$, somit hat H_k den Rang k. Dies impliziert die eindeutige Lösbarkeit des Ausgleichsproblems.

Damit ergibt sich das folgende Verfahren.

"Generalized Minimum Residual Method" (GMRES):

(1) Berechne für die Startlösung u_0 den Defekt $r_0 := b - Au_0$ sowie $q_1 := r_0/\|r_0\|_2$. Initialisiere

$$\tilde{H}_k = (h_{ij})_{\substack{1\le i\le k+1\\ 1\le j\le k}} := 0.$$

(2) Für $j = 1, \ldots, k$:

- $w := Aq_j$
- Für $i = 1, \ldots, j$:
  - $h_{ij} := q_i^* w$

  * $w := w - h_{ij} q_i$.
- $h_{j+1,j} := \|w\|_2$
- Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zu Schritt (3).
- $q_{j+1} := w/h_{j+1,j}$.

(3) Bestimme die Lösung y_k des linearen Ausgleichsproblems

$$\text{Minimiere } J(y) := \left\| \|r_0\|_2 e_1 - \tilde{H}_k y \right\|_2, \quad y \in \mathbf{R}^k.$$

Setze anschließend $u_k := u_0 + Q_k y_k$ mit $Q_k := (q_1 \ \cdots \ q_k)$.

Der Hauptaufwand des Verfahrens liegt im Fall $k \ll n$ wieder in Schritt (2) bei der Berechnung der Matrix-Vektorprodukte. Zur effizienten Lösung des linearen Ausgleichsproblems mit der bei $k \ll n$ niedrigdimensionalen Matrix $\tilde{H}_k$ bietet sich wegen der Struktur von $\tilde{H}_k$ wiederum die QR-Zerlegung von $\tilde{H}_k$ mittels Givens-Rotationen an:
Dabei multipliziert man die Matrix $(\tilde{H}_k \ \|r_0\|_2 e_1)$ sukzessive mit Givens-Rotationen $G_{j,j+1}$, $j = 1, ..., k$, um in der jeweils aktuellen Matrix das an der Position $(j+1, j)$ stehende Element für $j = 1, \cdots, k$ zu annulieren. Die Givens-Rotationsmatrizen lauten

$$G_{j,j+1} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & c_j & s_j & & & \\ & & & -s_j & c_j & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix}, \qquad j = 1, \ldots, k \tag{15.11}$$

mit $c_j = \cos\phi$, $s_j = \sin\phi$ bei geeignetem Winkel ϕ. Somit erhält man nach k Schritten die Matrix

$$(\tilde{R}_k \ \ \tilde{g}_k) := F_k(\tilde{H}_k \ \ \|r_0\|_2 e_1), \qquad F_k := G_{k,k+1} \cdots G_{12}.$$

Wir bezeichnen jetzt mit $R_k \in \mathbf{R}^{k \times k}$ die aus $\tilde{R}_k \in \mathbf{R}^{(k+1)\times k}$ durch Streichen der letzten (Null-)Zeile entstehende Matrix. Analog erhält man $g_k \in \mathbf{R}^k$ aus $\tilde{g}_k \in \mathbf{R}^{k+1}$ durch Weglassen der letzten Komponente. Es bezeichne γ_i die $i-$te Komponente von $\tilde{g}_k$.

Die Lösung des linearen Ausgleichsproblems ist dann offenbar gegeben durch $y_k = R_k^{-1} g_k$. Da $\tilde{H}_k$ den Rang k hat, ist tatsächlich R_k regulär. Wegen der Festsetzung von y_k ist

$$b - Au_k = Q_{k+1} \left(\|r_0\|_2 e_1 - \tilde{H}_k y_k \right) = Q_{k+1} F_k^* (\tilde{g}_k - \tilde{R}_k y_k) = Q_{k+1} F_k^* (\gamma_{k+1} e_{k+1}).$$

Aufgrund der Orthonormierung der Spalten von Q_{k+1} sowie der Orthogonalität von F_k ergibt sich damit

$$\|b - Au_k\|_2 = |\gamma_{k+1}|.$$

Diese Tatsache kann sehr günstig als Abbruchkriterium für das Verfahren verwendet werden, falls $|\gamma_{k+1}|$ hinreichend klein ist. Man kann den Vektor $\tilde{g}_k = (\gamma_i)_{1 \le i \le k+1}$ sehr einfach berechnen aus

$$\tilde{g}_k = F_k(\|r_0\|_2 e_1) = G_{k,k+1} \cdots G_{12}(\|r_0\|_2 e_1),$$

aus der Rekursion

- $\gamma_1 := \|r_0\|_2$.

- Für $j = 1, \ldots, k$:

$$\begin{pmatrix} \gamma_j \\ \gamma_{j+1} \end{pmatrix} := \begin{pmatrix} c_j & s_j \\ -s_j & c_j \end{pmatrix} \begin{pmatrix} \gamma_j \\ 0 \end{pmatrix}.$$

Insbesondere ist $\gamma_{j+1} = -s_j \gamma_j$. Daraus ergibt sich zugleich auch ein Abbruchkriterium für das GMRES-Verfahren.

Für die Realisierbarkeit des GMRES-Verfahrens benötigen wir

Lemma 15.6. *Bei regulärer Matrix $A \in \mathbf{R}^{n\times n}$ bricht das GMRES-Verfahren im $j-$ten Schritt wegen $h_{j+1,j} = 0$ genau dann ab, wenn u_j bereits Lösung des zu lösenden Gleichungssystems $Au = b$ ist.*

Beweis: Wir nehmen an, daß $h_{j+1,j} = 0$ ist. Im Verfahren wird dann $k := j$ gesetzt. Da das zu annulierende Element bereits verschwindet, ist die letzte Givens-Rotation die Identität, d.h. $s_k = 0$ und damit $\gamma_{k+1} = 0$. Also ist $Au_k = b$. Die Umkehrung wird analog gezeigt. □

Wir wollen uns nun mit Konvergenzeigenschaften des *GMRES*-Verfahrens befassen. Sei $\tilde{u}$ Lösung des Gleichungssystems. Per Konstruktion ist dann

$$\text{Minimiere } \|b - Au\|_2 = \|A(\tilde{u} - u)\|_2, \qquad u \in u_0 + \mathcal{K}_k, \tag{15.12}$$

also $\|b - Au_k\|_2 \leq \|b - Au_0\|_2$. Die Defektfolge verhält sich also monoton nichtwachsend. Für positiv definite, aber nicht notwendig symmetrische Matrizen A gilt sogar

Satz 15.7. *Sei $A \in \mathbf{R}^{n\times n}$ strikt positiv definit, d.h. $v^T Av \geq \alpha \|v\|_2^2$ für beliebige $v \in \mathbf{R}^n \setminus \{0\}$. Für die Näherungslösung u_k des GMRES(r)-Verfahrens mit Restart-Länge r und Startwert u_0 gilt*

$$\|b - Au_k\|_2 \leq \left(1 - \frac{\alpha^2}{\sigma^2}\right)^{k/2} \|b - Au_0\|_2, \quad k \in \mathbf{N}. \tag{15.13}$$

Dabei ist $\sigma := \|A\|_2$. Insbesondere konvergiert das Verfahren für $k \to \infty$ gegen die Lösung des Systems $Au = b$.

Beweis: Für beliebiges $\omega \in \mathbf{R}$ und $v \in \mathbf{R}^n$ gilt

$$\|(I - \omega A)v\|_2^2 \;=\; \|v\|_2^2 - 2\omega v^T Av + \omega^2 \|Av\|_2^2 \;\leq\; (1 - 2\omega\alpha + \omega^2 \|A\|_2^2)\|v\|_2^2.$$

Für $\omega = \omega_0 := \frac{\alpha}{\|A\|_2^2}$ folgt

$$\|(I - \omega_0 A)v\|_2 \leq q\|v\|_2, \qquad q := \left(1 - \frac{\alpha^2}{\|A\|_2^2}\right)^{\frac{1}{2}}.$$

Für $1 \leq k \leq r$ stimmen die Näherung u_k des GMRES(r)-Verfahrens und die des GMRES-Verfahrens überein. Wegen der Minimaleigenschaft der GMRES-Iterierten kann man das zugehörige Residuum vergleichen mit dem Residuum von

$$\tilde{u}_k = u_0 + \omega_0 \sum_{j=0}^{k-1} (I - \omega_0 A)^j r_0 \in u_0 + \mathcal{K}_k(A, r_0).$$

Wegen

$$\begin{aligned} b - A\tilde{u}_k &= r_0 - \omega_0 A \sum_{j=0}^{k-1} (I - \omega_0 A)^j r_0 \\ &= r_0 - \sum_{j=0}^{k-1} (I - \omega_0 A)^j r_0 + \sum_{j=0}^{k-1} (I - \omega_0 A)^{j+1} r_0 \\ &= r_0 - r_0 + (I - \omega_0 A)^k r_0 \;=\; (I - \omega_0 A)^k r_0 \end{aligned}$$

folgt

$$\|b - Au_k\|_2 \;\leq\; \|b - A\tilde{u}_k\|_2 \;=\; \|(I - \omega_0 A)^k r_0\|_2 \;\leq\; q^k \|r_0\|_2.$$

Nach dem ersten Restart, d.h. für $r < k \leq 2r$ gilt entsprechend

$$\|b - Au_k\|_2 \;\leq\; q^{k-r} \|b - Au_r\|_2 \;\leq\; q^{k-r} q^r \|r_0\|_2.$$

Analog gilt diese Abschätzung für alle $k \in \mathbf{N}$. Die Konvergenz des Verfahrens für $k \to \infty$ gegen die Lösung von $Au = b$ ergibt sich wegen $u - u_k = A^{-1}b - u_k = A^{-1}(b - Au_k)$ aus

$$\|u - u_k\|_2 \leq q^k \|A^{-1}\|_2 \|r_0\|_2, \quad k \in \mathbf{N}. \qquad \square$$

Bemerkung 15.8. (i) Die Konvergenzaussage von Satz 15.7 ist wenig hilfreich, wenn $\alpha \ll \sigma := \|A\|_2$ gilt. In vielen Fällen kann man jedoch die Situation durch geeignete Vorkonditionierung (vgl. folgender Abschnitt) erheblich verbessern.

(ii) Die Aussage von Satz 15.7 kann verallgemeinert werden auf den Fall diagonalisierbarer Matrizen A, d.h. man findet eine Matrix $X \in \mathbf{R}^{n \times n}$ mit $A = X \Lambda X^{-1}$ und $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Dabei sind $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A. $\square$

15.5 Vorkonditionierung von Krylov-Verfahren

Bemerkung 15.8 zeigt, daß die Konvergenz des GMRES-Verfahrens wesentlich vom *Spektrum* bzw. der Kondition der Matrix A abhängt. Wir wissen bereits aus Kapitel 13, daß sich die Kondition der aus der Diskretisierung elliptischer Randwertaufgaben resultierenden Matrizen mit der Verfeinerung des Gitters verschlechtert (vgl. Beispiel 13.1). Daher konvergiert das GMRES-Verfahren in der bisherigen Version in der Regel schlecht.

Ein Ausweg aus dieser Situation ergibt sich durch geeignete *Vorkonditionierung* des Problems mit einer regulären Matrix $M \in \mathbf{R}^{N \times N}$. Bei der *Linksvorkonditionierung* betrachtet man das zum Ausgangssystem (15.1) äquivalente Problem

$$M^{-1}Au = M^{-1}b. \tag{15.14}$$

Dabei soll M so gewählt werden, daß einerseits $M^{-1}A \approx I$ und damit die Kondition des geänderten Systems günstiger als die von A ist. Andererseits soll (15.14) "leicht(er)" lösbar sein als das Ausgangsproblem. Dieser Aspekt wird weiter unten präzisiert.

Bei der *Rechtsvorkonditionierung* gelangt man über die Transformation $u = M^{-1}x$ zum System $AM^{-1}x = b$. Man konstruiert M so, daß möglichst $AM^{-1} \approx I$ gilt. Man kann die Links- und Rechtvorkonditionierung auch kombinieren durch $u = M_2^{-1}x$ und $M_1^{-1}AM_2^{-1}x = M_1^{-1}b$.

Wir besprechen exemplarisch die Vorkonditionierung des GMRES-Verfahrens. Dabei spezifizieren wir die Vorkonditionierungsmatrizen noch nicht. Man beachte, daß gegenüber dem nichtvorkonditionierten Verfahren an verschiedenen Stellen der Algorithmen ein Gleichungssystem der Form

$$Mv = g$$

gelöst werden muß. Da dies ein System der Dimension n ist, muß dieses Hilfsproblem wesentlich effizienter als das Ausgangsproblem gelöst werden können.

Algorithmus: GMRES-Verfahren mit Linksvorkonditionierung

(1) Berechne für die Startlösung u_0 den vorkonditionierten Defekt $z_0 := M^{-1}(b - Au_0)$ sowie $q_1 := z_0/\|z_0\|_2$. Initialisiere

$$\tilde{H}_k = (h_{ij})_{\substack{1 \leq i \leq k+1 \\ 1 \leq j \leq k}} := 0.$$

(2) Für $j = 1, \cdots, k$:

- $w := M^{-1}Aq_j$
- Für $i = 1, \cdots, j$:
  - $h_{ij} := q_i^* w$
  - $w := w - h_{ij}q_i$.
- $h_{j+1,j} := \|w\|_2$
- Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zu Schritt (3).
- $q_{j+1} := w/h_{j+1,j}$.

(3) Bestimme die Lösung y_k des linearen Ausgleichsproblems

$$\text{Minimiere } J(y) := \left\| \|z_0\|_2 e_1 - \tilde{H}_k y \right\|_2, \quad y \in \mathbf{R}^k.$$

Setze anschließend $u_k := u_0 + Q_k y_k$ mit $Q_k := (\ q_1\ \cdots\ q_k\ )$.

Hier wird eine Orthonormalbasis zum modifizierten Krylov-Raum $\mathcal{K}_k(M^{-1}A, z_0)$ bestimmt. Man beachte, daß dabei der Defekt vorkonditioniert wird. Man hat jedoch nicht unmittelbar Zugriff auf den nicht vorkonditionierten Defekt. Dies gilt jedoch auch für den jetzt zu betrachtenden Fall der Rechtsvorkonditionierung, bei dem zunächst eine Orthonormalbasis für $\mathcal{K}_k(AM^{-1}, r_0)$ bestimmt wird.

Algorithmus: GMRES-Verfahren mit Rechtsvorkonditionierung

(1) Berechne für die Startlösung u_0 den Defekt $r_0 := b - Au_0$ sowie $q_1 := r_0/\|r_0\|_2$. Initialisiere

$$\tilde{H}_k = (h_{ij})_{\substack{1 \le i \le k+1 \\ 1 \le j \le k}} := 0.$$

(2) Für $j = 1, \cdots, k$:

- $w := AM^{-1}q_j$
- Für $i = 1, \cdots, j$:
  - $h_{ij} := q_i^* w$
  - $w := w - h_{ij}q_i$.
- $h_{j+1,j} := \|w\|_2$
- Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zu Schritt (3).
- $q_{j+1} := w/h_{j+1,j}$.

(3) Bestimme die Lösung y_k des linearen Ausgleichsproblems

$$\text{Minimiere } J(y) := \left\| \|r_0\|_2 e_1 - \tilde{H}_k y \right\|_2, \quad y \in \mathbf{R}^k.$$

Setze anschließend $u_k := u_0 + M^{-1}Q_k y_k$ mit $Q_k := (\ q_1\ \cdots\ q_k\ )$.

Der wesentliche Unterschied zwischen beiden Varianten der Vorkonditionierung soll im folgenden Lemma verdeutlicht werden.

Lemma 15.9. *Die Näherungslösung u_k ergibt sich im Fall des von links vorkonditionierten GMRES als Lösung von*

$$\textit{Minimiere } \|M^{-1}(b - Au)\|_2, \quad u \in u_0 + \mathcal{K}_k(M^{-1}A, z_0)$$

im Fall des von rechts vorkonditionierten GMRES als Lösung von

$$\textit{Minimiere } \|b - Au\|_2, \quad u \in u_0 + M^{-1}\mathcal{K}_k(AM^{-1}, r_0),$$

mit $r_0 := b - Ax_0$ *und* $z_0 := M^{-1}r_0$. *In beiden Varianten haben die (nicht zwingend gleichen) Lösungen* u_k *die Gestalt*

$$u_k = u_0 + s_{k-1}(M^{-1}A)z_0 = u_0 + M^{-1}s_{k-1}(AM^{-1})r_0, \quad s_{k-1} \in P_{k-1}.$$

Beweis: Die Aussage zur Linksvorkonditionierung folgt, da u_k bei Anwendung von GMRES auf das System $M^{-1}Au = M^{-1}b$ gebildet wird. Insbesondere findet man ein Polynom $s_{k-1} \in P_{k-1}$ mit

$$u_k = u_0 + s_{k-1}(M^{-1}A)z_0 = u_0 + s_{k-1}(M^{-1}A)M^{-1}r_0 = u_0 + M^{-1}s_{k-1}(AM^{-1})r_0.$$

Hierbei benutzt man die durch vollständige Induktion beweisbare Aussage

$$(M^{-1}A)^j M^{-1} = M^{-1}(AM^{-1})^j, \quad j = 0, \cdots k-1.$$

Im Fall der Rechtsvorkonditionierung ist $u_k = M^{-1}x_k$, wobei x_k Lösung der Minimierungsaufgabe

$$\text{Minimiere} \quad \|b - AM^{-1}x\|_2, \qquad x \in x_0 + \mathcal{K}_k(AM^{-1}, r_0)$$

mit $u_0 = M^{-1}x_0$ und $r_0 := b - Au_0$ ist. Die gesuchte Aussage erhält man mittels Transformation $u = M^{-1}x$. □

Bislang haben wir die Wahl der Vorkonditionierungsmatrix M nicht präzisiert. Das ist jedoch leider ein theoretisch noch nicht hinreichend gut gelöstes Problem, da offenbar diese Wahl sehr stark von Eigenschaften der Ausgangsmatrix A abhängt. Besonders kompliziert sind der Fall einer sehr feinen (unstrukturierten) Vernetzung bei elliptischen Randwertproblemen, insbesondere bei räumlich dreidimensionalen Aufgaben. Bei starker Nichtsymmetrie des Problems treten weitere erhebliche Probleme hinzu.

Bei geschickter Wahl von M erhält man durch Vorkonditionierung eine deutliche Beschleunigung gegenüber nichtvorkonditionierten Krylov-Methoden. Mitunter erreicht man auch erst dadurch Konvergenz der Iteration. Daher erfordert dieses Problem erhebliche praktische Erfahrung. In der Praxis verwendet man oft die folgenden Verfahren:

- Basis-Iterationsverfahren wie Gesamt- bzw. Einzelschrittverfahren bzw. dazugehörige Relaxationsverfahren (vgl. Kapitel 14),
- unvollständige $LU-$Zerlegungen wie $ILU(0)$ (vgl. Kapitel 13).

Auch Mehrgitter-Verfahren, die wir im folgenden Kapitel besprechen, spielen zunehmend eine Rolle bei der Vorkonditionierung.

Kapitel 16

Mehrgitterverfahren

Mehrgitterverfahren (MGV) zählen zu den schnellsten Iterationsverfahren zur Lösung von linearen Gleichungssystemen, die bei der Diskretisierung von Randwertproblemen entstehen. Wir hatten in Kapitel 14 gesehen, daß klassische Iterationsverfahren in der Regel in der Praxis nicht geeignet sind. Bei kleiner Diskretisierungsschrittweite h, d.h. bei Systemen großer Dimension, konvergieren sie sehr langsam.

Die MGV weisen dagegen unter gewissen Voraussetzungen eine von h unabhängige Konvergenzrate auf. Wir beschränken uns in dieser Darstellung auf den wichtigen Fall einer Variationsgleichung mit symmetrischer und positiv definiter Bilinearform.

16.1 Modellproblem. Vorbereitungen

Als *Modellaufgabe* betrachten wir ein symmetrisches elliptisches Randwertproblem 2. Ordnung. Sei $\Omega \subset \mathbf{R}^2$ ein konvexes, polygonal berandetes und beschränktes Gebiet. Mit der Bilinear- bzw. Linearform

$$a(u,v) := \int_\Omega [\alpha \nabla u \cdot \nabla v + \beta uv] \; dx, \qquad (f,v) := \int_\Omega fv \; dx \tag{16.1}$$

betrachten wir in Verallgemeinerung des Poisson-Problems die Variationsaufgabe

$$\text{Finde } \; u \in X := W_0^{1,2}(\Omega): \quad a(u,v) = (f,v) \;\; \forall v \in X. \tag{16.2}$$

Zur Näherungslösung von (16.1)-(16.2) sei $\{\mathcal{T}_l\}_{l=1}^k$ eine Gittersequenz, die durch dyadische Verfeinerung des quasi-uniformen Ausgangsgitters $\mathcal{T}_1$ entsteht, d.h.

$$h_l := \frac{1}{2} h_{l-1} \quad \text{mit} \;\; h_l := \max_{K \in \mathcal{T}_l} \text{diam}\,(h_K).$$

Bei konformer Approximation mit stückweise linearen Basisfunktionen ist

$$X_l := \{v_h \in C(\overline{\Omega}): \; v_h|_K \in P_1(K) \; \forall K \in \mathcal{T}_l, \; v|_{\partial\Omega} = 0\}$$

und

$$X_1 \; \subset \; X_2 \; \subset \; ... \subset \; X_{k-1} \; \subset \; X_k \; \subset \; X. \tag{16.3}$$

Wir vermerken nur, daß diese Inklusionskette die wesentliche Voraussetzung der weiteren Betrachtungen ist. Hierfür ist die dyadische ("rote") Verfeinerung eines Ausgangsgitters hinreichend, jedoch nicht notwendig. Man kann auch bei lokal verfeinerten Gittern in geeigneter Weise (16.3) erzwingen.

Eine Näherung $u_k \in X_k$ an die Lösung u von (16.2) suchen wir als Lösung von

$$\text{Finde } \; u_k \in X_k: \qquad a(u_k,v) = (f,v) \;\; \forall v \in X_k. \tag{16.4}$$

Bei hinreichend glatten Koeffizienten α und β mit $0 < \alpha_0 \leq \alpha(x) \leq \alpha_1$ und $0 \leq \beta(x) \leq \beta_1$ und für $f \in L^2(\Omega)$ erhält man nach den Ergebnissen aus den Kapiteln 6 bzw. 10 die Regularitätsaussage $u \in X \cap W^{2,2}(\Omega)$ sowie die Fehlerabschätzungen

$$\|u - u_k\|_{s,\Omega} \leq C h_k^{2-s} |u|_{2,\Omega}, \qquad s \in \{0,1\}, \quad k \in \mathbf{N}. \tag{16.5}$$

Ziel der weiteren Untersuchungen ist die Konstruktion eines *Mehrgitter-Verfahren* derart, daß

- eine Näherung $\hat{u}_k \in X_k$ mit $\mathcal{O}(n_k)$ wesentlichen algebraischen Operationen bei $n_k = \dim(X_k)$ bestimmt
- und dabei die Abschätzung

$$\|u_k - \hat{u}_k\|_{s,\Omega} \leq C h_k^{2-s} |u|_{2,\Omega}, \quad s = 0, 1,$$

  also auch (16.5), gesichert wird.

Wir stellen jetzt einige Begriffe und Hilfsaussagen zusammen, die für die Analyse des Mehrgitter-Verfahren nützlich sind. Zunächst benutzen wir das folgende *gitterabhängige* Skalarprodukt $(\cdot,\cdot)_k$ auf X_k mit

$$(v,w)_k := h_k^2 \sum_{j=1}^{n_k} v(p^j) w(p^j). \tag{16.6}$$

Dabei ist $\{p^j\}_{j=1}^{n_k}$ die Menge der inneren Gitterpunkte der Triangulation $\mathcal{T}_k$.

Dann wird durch die Beziehung

$$(A_k v, w)_k = a(v,w) \qquad \forall v, w \in X_k \tag{16.7}$$

sowie unter Beachtung des Darstellungssatzes von Riesz ein Operator $A_k : X_k \to X_k$ definiert. Das diskrete Problem (16.4) kann somit umformuliert werden als

$$A_k u_k = f_k \in X_k, \qquad (f_k, w)_k := (f, w) \ \forall w \in X_k. \tag{16.8}$$

Der Operator A_k ist offenbar symmetrisch und positiv definit bezüglich des Skalarproduktes $(\cdot,\cdot)_k$, d.h. durch

$$|||v|||_{s,k} := \sqrt{(A_k^s v, v)_k}, \quad s \in \mathbf{R} \tag{16.9}$$

wird eine *gitterabhängige Norm* $|||\cdot|||_{s,k}$ induziert.

Zur Erinnerung erklären wir den Ausdruck A_k^s: Seien $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_{n_k}$ die Eigenwerte des Operators A_k sowie $\phi_i,\ i = 1, ..., n_k$, die bezüglich des Skalarproduktes $(\cdot,\cdot)_k$ orthonormierten Eigenvektoren. Dann gilt

$$\mathbf{R}^{n_k} \ni v = \sum_{i=1}^{n_k} \alpha_i \phi_i \quad \mapsto \quad A_k^s v := \sum_{i=1}^{n_k} \lambda_i^s \alpha_i \phi_i.$$

Man rechnet dann sofort nach, daß

$$|||v|||_{s,k}^2 := (A_k^s v, v)_k = \left(\sum_{i=1}^{n_k} \lambda_i^s \alpha_i \phi_i, \sum_{j=1}^{n_k} \alpha_j \phi_j \right)_k = \sum_{i=1}^{n_k} \lambda_i^s \alpha_i^2. \tag{16.10}$$

Man sieht für $s = 1$ den Zusammenhang zur energetischen Norm $\|\cdot\|_E$ über

$$\|\cdot\|_E := \sqrt{a(\cdot,\cdot)} \ \equiv \ |||\cdot|||_{1,k}.$$

Das nächste Lemma zeigt die Äquivalenz der Norm $|||\cdot|||_{0,k}$ zur L^2-Norm.

Lemma 16.1. Es existieren positive Konstanten C_1, C_2, so daß

$$C_1\|v\|_{L^2(\Omega)} \le |||v|||_{0,k} \le C_2\|v\|_{L^2(\Omega)} \qquad \forall v \in X_k.$$

Beweis: Der Beweis wird über Quadraturformeln für Integranden aus $P_2(K)$ über jedem Dreieck K geführt. Für $v \in \mathbf{P}_1(K)$ gilt zunächst

$$\|v\|^2_{L^2(\Omega)} = \sum_{K\in\mathcal{T}_k} \int_K v^2\, dx = \sum_{K\in\mathcal{T}_h} \frac{\mathrm{meas}(K)}{3} \left(\sum_{i=1}^{3} v^2(m_i) \right).$$

Dabei sind $m_i, i = 1, 2, 3$, die Mittelpunkte der Kanten von K. Per Definition ist

$$|||v|||^2_{0,k} := (v, v)_k := h_k^2 \sum_{j=1}^{n_k} v^2(p^j).$$

Für eine quasiuniforme Zerlegung gilt $\mathrm{meas}(K) \sim h_k^2$. Schließlich hat man für Funktionen $v \in P_1(K)$ die Umrechnungsformel

$$\begin{pmatrix} v(m_1) \\ v(m_2) \\ v(m_3) \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} v(p^1) \\ v(p^2) \\ v(p^3) \end{pmatrix},$$

woraus sich die Behauptung ergibt. □

Nun schätzen wir den Spektralradius der Matrizen A_k ab.

Lemma 16.2. *Für den Spektralradius der Matrix A_k gilt $\rho(A_k) \le Ch_k^{-2}$.*

Beweis: Sei λ Eigenwert der Matrix A_k mit zugehörigem Eigenvektor ϕ. Dann ist

$$a(\phi, \phi) = (A_k\phi, \phi)_k = \lambda(\phi, \phi)_k = \lambda|||\phi|||^2_{0,k} \qquad \Longrightarrow \qquad \lambda = \frac{a(\phi, \phi)}{|||\phi|||^2_{0,k}}.$$

Die Behauptung erhält man dann mittels inverser Ungleichung (vgl. Lemma 13.2)

$$\|v\|^2_E = a(v, v) \le \alpha_1|v|^2_{1,\Omega} + \beta_1\|v\|^2_{0,\Omega} \le Ch_K^{-2}\|v\|^2_{0,\Omega} \quad \forall v \in X_k$$

sowie über die Normäquivalenz $|||\cdot|||_{0,k} \sim \|\cdot\|_{L^2(\Omega)}$ nach Lemma 16.1. □

Schließlich machen wir später wesentlich Gebrauch von der folgenden verallgemeinerten Ungleichung von Cauchy-Schwarz.

Lemma 16.3. *Für beliebige reelle Zahlen t gilt*

$$|a(v, w)| \le |||v|||_{1+t,k}\ |||w|||_{1-t,k} \qquad \forall v, w \in X_k.$$

Beweis: analog zum Beweis von Satz 16.8 (vgl. unten). □

16.2 Mehrgitter-Algorithmus

Das Mehrgitter-Verfahren ist für jedes k ein Iterationsverfahren zur Näherungslösung der Gleichung

$$A_k z = g \quad \text{in} \quad X_k. \tag{16.11}$$

Seine *Hauptkomponenten* sind

- *Glättung* hochfrequenter Fehleranteile auf dem "feinen" Gitter $\mathcal{T}_k$, d.h. in X_k,
- Approximation der niederfrequenten Fehleranteile von X_k auf dem "gröberen" Gitter $\mathcal{T}_{k-1}$, d.h. in X_{k-1}.

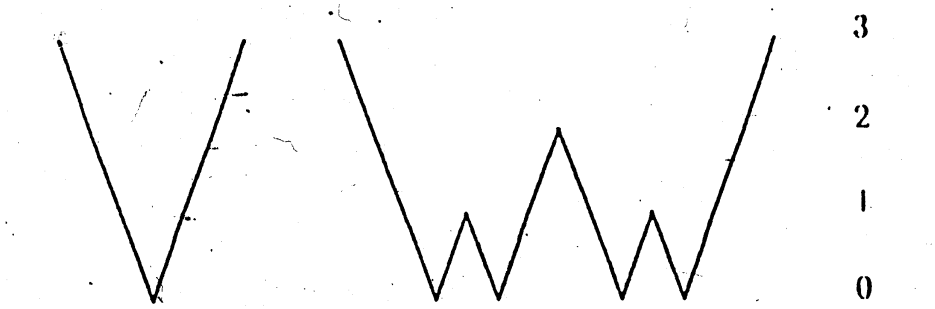


Abbildung 16.1: Schematische Darstellung des V- bzw. W-Zyklus

Wir benötigen dazu *Transfer-Operatoren* zwischen den Gittern:

- *Prolongation* (grob → fein) $I_{k-1}^{k} : X_{k-1} \to X_k$
- *Restriktion* (fein → grob) $I_{k}^{k-1} : X_k \to X_{k-1}$.

Diese Operatoren sind nicht zwingend festgelegt, jedoch wählt man sie als zueinander adjungiert. Hier gehen wir wie folgt vor: Als Prolongation verwenden wir die sogenannte natürliche Injektion

$$I_{k-1}^{k} v \;=\; v \quad \forall v \in X_{k-1}, \tag{16.12}$$

als Restriktion nutzen wir den dazu bezüglich $(\cdot,\cdot)_{k-1}$ bzw. $(\cdot,\cdot)_k$ adjungierten Operator

$$(I_{k}^{k-1} w, v)_{k-1} = (w, I_{k-1}^{k} v)_k = (w, v)_k \quad \forall v \in X_{k-1},\ \forall w \in X_k. \tag{16.13}$$

Wir beschreiben nun die *Iteration auf Level k* als *geschachteltes* Verfahren, d.h. das Verfahren ruft sich selbst während der Iteration wieder auf. Die Näherungslösung an die Lösung z von $A_k z = g$ auf dem Level k mit dem Startwert z_0 bezeichnen wir mit $\mathrm{MG}(k, z_0, g)$.

Mehrgitter-Verfahren: Iteration auf Level k

$\mathbf{k = 1}$: Durch ein direktes Verfahren löst man auf dem gröbsten Gitter exakt, d.h.

$$MG(1, z_0, g) = A_1^{-1} g.$$

$\mathbf{k \geq 2}$: Die Näherung $\mathrm{MG}(k, z_0, g)$ gewinnt man rekursiv aus folgenden Schritten:

1. **Vorglättung:**
$$z_l := z_{l-1} + \frac{1}{\Lambda_k}(g - A_k z_{l-1}), \qquad l = 1, ..., m_1.$$
Dabei ist $\Lambda_k \leq C h_k^{-2}$ eine obere Schranke für den Spektralradius $\rho(A_k)$ (vgl. Lemma 16.2).
2. **Grobgitterkorrektur:** Seien $q_0 = 0$ und $\overline{g} := I_k^{k-1}(g - A_k z_{m_1})$ die Restriktion des Vorglättungsresiduums. Nach Ermittlung einer Näherung
$$q_i := \mathrm{MG}(k-1, q_{i-1}, \overline{g}), \qquad i = 1, ..., p$$
erhält man die Korrektur
$$z_{m_1+1} := z_{m_1} + I_{k-1}^{k} q_p.$$
3. **Nachglättung:** Man berechnet
$$z_l := z_{l-1} + \frac{1}{\Lambda_k}(g - A_k z_{l-1}), \qquad l = m_1 + 2, ..., m_1 + m_2 + 1.$$

Das Ergebnis der Iteration auf dem Level k ist

$$\mathrm{MG}(k, z_0, g) := z_{m_1+m_2+1}.$$

Auch hier ist zu bemerken, daß die konkrete Wahl der Vor- und Nachglättung (hier als Richardson-Iteration) nicht zwingend ist.

Bei der Festlegung $p = 1$ bzw. $p = 2$ im Fehlerkorrekturschritt spricht man von einem *V-Zyklus* bzw. *W-Zyklus* (vgl. Abb. 16.1).

Praktisch geht man bei Lösung des linearen Gleichungssystems $A_k u_k = f_k$, vgl. (16.8), wie folgt vor: Sei $\hat{u}_{k-1}$ eine Näherungslösung der Gleichung $A_{k-1}u_{k-1} = f_{k-1}$. Dann wählt man $I_{k-1}^k \hat{u}_{k-1}$ als Startlösung zur Bestimmung von u_k. Dann wendet man die Iteration auf Level k mehrfach an ($r-$mal).

Kurz gefaßt läuft dann ein Mehrgitter-Zyklus zur Lösung der Gleichung $A_k u_k = f_k$ in vollständiger Form wie folgt ablaufen:

Vollständiger Mehrgitter-Algorithmus

$\mathbf{k = 1}$: Berechne $\hat{u}_1 = A_1^{-1} f_1$.

$\mathbf{k \geq 2}$: Bestimme die Näherungslösung $\hat{u}_k$ rekursiv aus

$$\begin{aligned} u_0^k &= I_{k-1}^k \hat{u}_{k-1} \\ u_l^k &= \mathrm{MG}(k, u_{l-1}^k, f_k), \quad l = 1, ..., r \\ \hat{u}_k &= u_r^k. \end{aligned}$$

16.3 Analyse des MGV auf Level k

Für die weitere Analyse bezeichnen wir durch $\mathbf{P}_k : X \to X_k$ die orthogonale Projektion bezüglich der Bilinearform $a(\cdot, \cdot)$, d.h.

$$\mathbf{P}_k v \in X_k : \quad a(v - \mathbf{P}_k v, w) = 0 \quad \forall w \in X_k.$$

Wir wollen zeigen, daß der Operator $\mathbf{P}_{k-1}$ den Vorglättungsfehler auf Level k in Relation zur exakten Lösung der Residuumsgleichung auf $\mathcal{T}_{k-1}$ setzt. Sei dazu $q \in X_{k-1}$ die Grobgitterkorrektur (vgl. Schritt (2)), d.h. Lösung von

$$A_{k-1} q = \overline{g} := I_k^{k-1}(g - A_k z_{m_1}).$$

Lemma 16.4. *Es gilt $q = \mathbf{P}_{k-1}(z - z_{m_1})$, d.h. bei exakter Lösung der Grobgittergleichung auf $\mathcal{T}_{k-1}$ ist*

$$z - (z_{m_1} + q) = (I - \mathbf{P}_{k-1})(z - z_{m_1})$$

der Fehler nach dem Korrekturschritt.

Beweis: Nach Definition von A_{k-1}, von q bzw. g und von I_k^{k-1} ist zunächst

$$a(q, w) = (A_{k-1} q, w)_{k-1} = (\overline{g}, w)_{k-1} = (I_k^{k-1}(g - A_k z_{m_1}), w)_{k-1}.$$

Nach Festlegung von I_k^{k-1}, von z sowie von A_k folgt schließlich die Behauptung über

$$a(q, w) = (g - A_k z_{m_1}, w)_k = (A_k(z - z_{m_1}), w)_k = a(z - z_{m_1}), w). \qquad \square$$

Die Aussage von Lemma 16.4 führt auf die Untersuchung des Operators $I - \mathbf{P}_{k-1}$.

Lemma 16.5. *Es existiert eine positive Konstante C mit*

$$|||(I - \mathbf{P}_{k-1})v|||_{0,k} \leq C h_k |||(I - \mathbf{P}_{k-1})v|||_{1,k} \qquad \forall v \in X_k.$$

Beweis: Nach dem Dualitätsargument aus Abschnitt 10.5 (vgl. Satz 10.16) gilt

$$\|(I-\mathbf{P}_{k-1})v\|_{L^2(\Omega)} \le Ch_k\|(I-\mathbf{P}_{k-1})v\|_E = Ch_k|||(I-\mathbf{P}_{k-1})v|||_{1,k}.$$

Die Behauptung folgt dann unter Beachtung von Lemma 16.1. □

Das gesuchte Ergebnis (übrigens ohne die Voraussetzung $v \in W^{2,2}(\Omega)$!) gibt

Satz 16.6. (*Approximationseigenschaft*)
Es existiert eine positive Konstante C mit

$$|||(I-\mathbf{P}_{k-1})v|||_{1,k} \le Ch_k|||v|||_{2,k} \qquad \forall v \in X_k.$$

Beweis: Unter Beachtung der Galerkin-Orthogonalität sowie der Lemmata 16.3 und 16.5 findet man

$$\begin{aligned} |||(I-\mathbf{P}_{k-1})v|||_{1,k}^2 &= \|v-\mathbf{P}_{k-1}v\|_E^2 \\ &= a(v-\mathbf{P}_{k-1}v, v-\mathbf{P}_{k-1}v) \\ &= a(v-\mathbf{P}_{k-1}v, v) \\ &\le |||v-\mathbf{P}_{k-1}v|||_{0,k}|||v|||_{2,k} \\ &\le Ch_k|||(I-\mathbf{P}_{k-1})v|||_{1,k}|||v|||_{2,k}. \end{aligned}$$

Daraus folgt die Behauptung. □

Wir zeigen nun, daß die Ausführung des W-Zyklus (d.h. mit $p=2$) auf Level k bei hinreichend großer Zahl von Glättungsschritten eine Kontraktion ist. Vereinfachend betrachten wir einen *einseitigen W-Zyklus*, d.h. es gelte $p=2$, $m_1 = m \ge 1$ sowie $m_2 = 0$.

Der Fehler der Iteration auf Level k sei $e_i := z - z_i$ für $i = 0, ..., m+1$. Als Fehler des Glättungsschrittes mit $l = 1, ..., m$ erhalten wir

$$\begin{aligned} e_l &:= z - z_l := z - z_{l-1} - \frac{1}{\Lambda_k}A_k(z - z_{l-1}) \\ &:= e_{l-1} - \frac{1}{\Lambda_k}A_k e_{l-1} = \left(I - \frac{1}{\Lambda_k}A_k\right)e_{l-1}. \end{aligned}$$

Nachfolgend bezeichnen wir den Ausdruck

$$R_k := I - \frac{1}{\Lambda_k}A_k : \ X_k \to X_k$$

als *Glättungs-* bzw. *Relaxationssoperator*. Der Fehler des Glättungsschrittes ist somit beschreibbar als

$$e_m = R_k^m e_0.$$

Für den Glättungsoperator zeigen wir die Abschätzung

$$|||R_k v|||_{s,k} \le |||v|||_{s,k} \quad \forall v \in X_k, \ \forall s \in \mathbf{R}. \tag{16.14}$$

Diese Aussage folgt mit der Basisdarstellung $v = \sum_{i=1}^{n_k} \alpha_i \phi_i$ aus der Darstellung (16.10) und wegen $0 < \lambda_i \le \Lambda_k$ aus

$$|||R_k v|||_{s,k}^2 \ = \ \sum_{i=1}^{n_k} \lambda_i^s \left(1 - \frac{\lambda_i}{\Lambda_k}\right)^2 \alpha_i^2 \ \le \ \sum_{i=1}^{n_k} \lambda_i^s \alpha_i^2 \ = \ |||v|||_{s,k}^2. \tag{16.15}$$

Wir wollen nun im ersten Schritt die Konvergenz der *Zweigitter-Methode* beweisen:
Sei dazu $q \in X_{k-1}$ Lösung von $A_{k-1}q = \overline{g}$. Das Ergebnis der Zweigitter-Iteration ist dann

$$\hat{z}_{m+1} = z_m + q.$$

Für den Fehler der Methode gilt somit unter Beachtung von Lemma 16.4:

$$\begin{aligned}\hat{e}_{m+1} &:= z - \hat{z}_{m+1} = z - z_m - q = e_m - q \\ &= e_m - \mathbf{P}_{k-1} e_m = (I - \mathbf{P}_{k-1}) R_k^m e_0.\end{aligned}$$

Neben der Approximationseigenschaft (vgl. Satz 16.6) benötigen wir also eine Abschätzung von R_k^m.

Satz 16.7. *(Glättungseigenschaft des W-Zyklus)*
Es gibt eine positive Konstante C mit

$$|||R_k^m v|||_{2,k} \leq C h_k^{-1} m^{-1/2} |||v|||_{1,k} \quad \forall v \in X_k.$$

Beweis: Analog zu (16.15) und unter Beachtung von $\lambda_i \leq \Lambda_k \leq C h_k^{-2}$ folgt dann nach kurzer Extremwertberechnung über

$$\begin{aligned}|||R_k^m v|||_{2,k}^2 &= \sum_{i=1}^{n_k} \left(1 - \frac{\lambda_i}{\Lambda_k}\right)^{2m} \lambda_i^2 \alpha_i^2 = \Lambda_k \left[\sum_{i=1}^{n_k} \left(1 - \frac{\lambda_i}{\Lambda_k}\right)^{2m} \left(\frac{\lambda_i}{\Lambda_k}\right) \lambda_i \alpha_i^2\right] \\ &\leq \Lambda_k \left[\sup_{0 \leq x \leq 1} (1-x)^{2m} x\right] \sum_{i=1}^{n_k} \lambda_i \alpha_i^2 \leq C h_k^{-2} \frac{1}{2m+1} |||v|||_{1,k}^2\end{aligned}$$

die Behauptung. □

Damit beweisen wir den

Satz 16.8. *(Konvergenz der Zweigitter-Methode)*
Es gibt eine von k unabhängige positive Konstante C mit

$$\|\hat{e}_{m+1}\|_E \leq \frac{C}{\sqrt{m}} \|e_0\|_E.$$

Beweis: Unter Beachtung der Approximations- und Glättungseigenschaft folgt

$$\begin{aligned}\|\hat{e}_{m+1}\|_E &= \|(I - \mathbf{P}_{k-1}) R_k^m e_0\|_E \\ &\leq C h_k |||R_k^m e_0|||_{2,k} \\ &\leq C h_k C h_k^{-1} m^{-1/2} |||e_0|||_{1,k} \\ &\leq \frac{C}{\sqrt{m}} \|e_0\|_E. \qquad \square\end{aligned}$$

Im zweiten Schritt können wir jetzt die Konvergenz der Iteration auf Level k beweisen:

Satz 16.9. (*Konvergenz der Iteration auf Level k*)
Für jede Konstante $\gamma \in (0,1)$ existiert eine (ggf. hinreichend große) Zahl $m \in \mathbf{N}$, so daß

$$\|z - \mathrm{MG}(k, z_0, g)\|_E \leq \gamma \|z - z_0\|_E, \quad k \in \mathbf{N}. \tag{16.16}$$

Beweis: Sei C^{AP} die Konstante aus Satz 16.8. Ferner wählen wir die Zahl m entsprechend

$$\mathbf{N}_0 \ni m > \left(\frac{C^{AP}}{\gamma(1-\gamma)}\right)^2.$$

Es bleibt zu zeigen, daß dann die Aussage (16.16) erfüllt ist. Der Nachweis erfolgt per Induktion:
Für $k = 1$ ist (16.16) erfüllt, denn die linke Seite verschwindet sogar. Für $k \geq 2$ gilt wegen $p = 2$ (beim W-Zyklus), daß

$$z - \mathrm{MG}(k, z_0, g) = z - z_{m+1} = z - z_m - q_2 = \hat{e}_{m+1} + q - q_2,$$

also unter Beachtung der Induktionsvoraussetzung sowie von Satz 16.8

$$\|z - \mathrm{MG}(k, z_0, g)\|_E \;\le\; \|\hat{e}_{m+1}\|_E + \|q - q_2\|_E \;\le\; \frac{C^{AP}}{\sqrt{m}}\|e_0\|_E + \gamma^2\|q\|_E.$$

Eine kurze Nebenrechnung ergibt unter Beachtung von Lemma 16.4, der Projektoreigenschaft von $\mathbf{P}_{k-1}$ sowie von (16.14)

$$\|q\|_E = \|\mathbf{P}_{k-1}(z - z_m)\|_E \le \|z - z_m\|_E = \|R_k^m e_0\|_E \le \|e_0\|_E.$$

Damit folgt wegen der Voraussetzung an die Zahl m schließlich die Behauptung mit

$$\|z - \mathrm{MG}(k, z_0, g)\|_E \le \left(\frac{C^{AP}}{\sqrt{m}} + \gamma^2\right)\|e_0\|_E \le \gamma\|e_0\|_E. \qquad \square$$

Das entsprechende *Ergebnis für den V-Zyklus* (d.h. mit $p = 1$ sowie $m = m_1 = m_2 \ge 1$) lautet bei verfeinerter Beweistechnik

Satz 16.9* (*Konvergenz der MG-Iteration auf Level k beim V−Zyklus)*
Sei m die Zahl der Glättungsschritte des V-Zyklus. Dann gilt unabhängig vom Level k die Aussage

$$\|z - \mathrm{MG}(k, z_0, g)\|_E \le \frac{C^{AP}}{m + C^{AP}}\|z - z_0\|_E.$$

Beweis: vgl. S. Brenner [4], Theorem 6.6.12 □

Aus dem Ergebnis und Beweis dieses Satzes kann man für den $W-$Zyklus folgern, daß dort auch lediglich ein Glättungsschritt (d.h. $m = 1$) für die Konvergenz ausreichend ist.

16.4 Konvergenz- und Aufwandsabschätzung

Wir analysieren schließlich die Konvergenz und Komplexität des vollständigen Mehrgitter-Verfahrens. Zunächst betrachten wir die Konvergenzfrage.

Satz 16.10. *(Konvergenz des vollständigen Mehrgitter-Verfahrens)*
Seien $u_k \in X_k$ bzw. $\hat{u}_k \in X_k$ die Lösung des diskreten Problems (16.4) bzw. des vollständigen Mehrgitter-Verfahrens. Die Mehrgitter-Iteration auf Level k sei eine Kontraktion mit einer Kontraktionszahl $\gamma \ne \gamma(k)$. Ferner sei die Zahl r der Mehrgitter-Zyklen hinreichend groß. Dann gibt es eine Konstante $C > 0$ mit

$$\|u_k - \hat{u}_k\|_E \le C h_k |u|_{W^{2,2}(\Omega)}.$$

Beweis: Sei $\hat{e}_k := u_k - \hat{u}_k$, also speziell $\hat{e}_1 = 0$. Unter Beachtung von Satz 16.9 und der bekannten Fehlerabschätzung in der $W^{1,2}$-Norm folgt dann

$$\begin{aligned}\|\hat{e}_k\|_E &\le \gamma^r\|u_k - \hat{u}_{k-1}\|_E \\ &\le \gamma^r\left(\|u_k - u_{k-1}\|_E + \|u_{k-1} - \hat{u}_{k-1}\|_E\right) \\ &\le C\gamma^r\left(h_k|u|_{W^{2,2}(\Omega)} + \|\hat{e}_{k-1}\|_E\right).\end{aligned}$$

Durch Iteration dieser Beziehung und Verwendung der geometrische Reihe mit $2C\gamma^r < 1$ sowie von $h_k = \frac{1}{2}h_{k-1}$ folgern wir über

$$\begin{aligned}\|\hat{e}_k\|_E &\le Ch_k\gamma^r|u|_{W^{2,2}(\Omega)} + C^2h_{k-1}\gamma^{2r}|u|_{W^{2,2}(\Omega)} + \ldots + C^k h_1\gamma^{kr}|u|_{W^{2,2}(\Omega)} \\ &\le \frac{C\gamma^r}{1 - 2C\gamma^r}h_k|u|_{W^{2,2}(\Omega)}\end{aligned}$$

die Behauptung. □

Wir wollen abschließend den Rechenaufwand des Mehrgitterverfahrens ermitteln. Dieser wird bestimmt durch den Aufwand für die Glättung, die Restriktion und Prolongation auf den Gittern l_{max} bis l_{min}. Hinzu kommt der Aufwand für die Berechnung der exakten Lösung des Gleichungssystems auf dem gröbsten Gitter.

Die Matrix auf dem Gitter Ω_l ist stets eine schwachbesetzte $n_l \times n_l-$Matrix. Daher ist die Summe der Rechenoperationen für Glättung, Restriktion und Prolongation proportional zur Anzahl der Gitterpunkte n_l dieses Gitters. C_S, C_R und C_P seien als Konstanten so gewählt, daß der Aufwand für die Vorglättung $\leq C_S n_l$, der Aufwand für die Berechnung des Defektes und dessen Restriktion $\leq C_R n_l$ sowie der Aufwand für die Prolongation $\leq C_P n_l$ für alle $l \geq 1$ beträgt.

Satz 16.11. *(Komplexität des vollständigen Mehrgitter-Verfahrens)*
Die Zahlen m_1 bzw. m_2 seien die Anzahl der Vor- bzw. Nachglättungsschritte. Durch $p = 1$ bzw. $p = 2$ werden der V-Zyklus bzw. der W-Zyklus gekennzeichnet. Weiter ist $n_{l+1} = 4n_l$. Dann gilt für den Aufwand W eines Mehrgitteriterationsschrittes

$$W \leq C_l(m_1 + m_2)n_l$$

mit

$$C_l(m) := \frac{mC_S + C_R + C_P}{1 - \frac{p}{4}} + \mathcal{O}\left(\left(\frac{p}{4}\right)^l\right).$$

Beweis: Der Rechenaufwand für einen Mehrgitterschritt sei $C_l n_l$. Aus der Iterationsvorschrift folgt

$$C_l n_l \leq (mC_S + C_R + C_P)\, n_l + pC_{l-1}n_{l-1},$$

also

$$C_l \leq (mC_S + C_R + C_P) + \zeta C_{l-1}$$

mit $\zeta := p/4$ und $n_{l-1}/n_l \leq 1/4$. Rekursiv folgt unter Benutzung der Darstellung der geometrischen Reihe

$$C_l \leq \frac{mC_S + C_R + C_P}{1 - \zeta} + \frac{p^l C_0}{4^l n_0}. \qquad \square$$

Wesentliches Resultat von Satz 16.11 ist, daß Mehrgitterverfahren den *optimalen Verfahrensaufwand* $\mathcal{O}(n_l)$ haben. Dies zeichnet das Verfahren vor anderen Lösungsverfahren für lineare Gleichungssysteme, die aus der Diskretisierung von Randwertproblemen entstehen, aus.

16.5 Erweiterungen. Ausblick

Die Ergebnisse dieses Kapitels zu Mehrgitter-Verfahren bei der Finite-Elemente Approximation von *symmetrischen* elliptischen Randwertproblemen zweiter Ordnung mit stückweise *linearen* Elementen erfordern einige Bemerkungen:

- Bei Verwendung von Ansatzfunktionen höherer Ordnung kann man die hier betrachtete Methode mit stückweise linearen Ansatzfunktionen als Vorkonditionierer verwenden.
- Die hier gewählte Darstellung ging vereinfachend von einer dyadischen Verfeinerung eines geeigneten Ausgangsgitters aus. Im Fall unstrukturierter Gitter, die eventuell adaptiv verfeinert werden, ist die Anwendung von Mehrgitter-Verfahren nach geeigneten Anpassungen möglich.
- Die hier vorgestellte numerische Analysis des Mehrgitter-Verfahrens basiert wesentlich auf der Annahme einer symmetrischen und positiv definiten Matrix. Direkte Erweiterungen auf den Fall *nichtsymmetrischer* und/ oder *indefiniter* Matrizen sind nicht direkt möglich und Gegenstand aktueller Forschung.

- Insgesamt haben Mehrgitter-Verfahren aufgrund verschiedener wählbarer Komponenten (Glättungsoperatoren, Gittertransfer-Operatoren usw.) eine komplexe Struktur. Es bedarf bei der Anwendung auf konkrete praktische Probleme doch erheblicher Erfahrungen. Insbesondere ist die oft zu hörende Meinung, daß Mehrgitter-Verfahren hinsichtlich der Konvergenzgeschwindigkeit und des Aufwandes ohne Konkurrenz seien, in Abhängigkeit vom konkreten Problem stark zu relativieren. In der Praxis verwendet man zunehmend Mehrgitter-Verfahren zur Vorkonditionierung von Krylov-Unterraum-Verfahren. Hierbei spielt natürlich die optimale Komplexität der MGV eine wesentliche Rolle.

Teil IV

Ausgewählte Erweiterungen

Kapitel 17

Probleme mit dominanter Konvektion

Bisher hatten wir elliptische Probleme der Form

$$Lu = f \qquad \text{in } \Omega$$

betrachtet mit einem elliptischen Operator

$$Lu := L_2u + L_1u := -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{j=1}^{n} b_j \frac{\partial u}{\partial x_j} + cu.$$

Wir ersetzen jetzt den Operator Lu durch den parameterabhängigen Operator $L_\epsilon u := \epsilon L_2 u + L_1 u$ mit $0 < \epsilon \leq 1$. Von besonderen Interesse ist in Anwendungen der Fall $0 < \epsilon \ll 1$. Natürlich entartet die bisherige Theorie im Grenzfall $\epsilon \to +0$. Ebenso müssen Diskretisierungsverfahren modifiziert werden.

17.1 Hyperbolische Gleichungen 1. Ordnung

Eine allgemeine skalare partielle Differentialgleichung 1. Ordnung der Form

$$F\left(x, u(x), \nabla u(x)\right) = 0 \tag{17.1}$$

für die Funktion $u = u(x)$ beschreibt physikalisch ein dichtes Teilchenfeld (zum Beispiel ein Strömungs- oder Windfeld) ohne Wechselwirkung. Man spricht bei (17.1) auch von einer *hyperbolischen Gleichung 1. Ordnung*. Hier beschränken wir uns auf den *linearen* Fall

$$\sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x). \tag{17.2}$$

Gleichung (17.2) besagt, daß sich der Wert der Lösung $u = u_0$ mit der Geschwindigkeit $f(x_0) - c(x_0)u(x_0)$ zu ändern beginnt, wenn sich der Punkt x beginnend in x_0 mit der Anfangsgeschwindigkeit $\vec{b}(x_0) = (b_1(x_0), ..., b_n(x_0))^T$ durch den Raum $\mathbf{R}^n$ bewegt. Hierdurch motiviert ergibt sich ein enger Zusammenhang zwischen Problem (17.2) und dem folgenden System gewöhnlicher Differentialgleichungen

$$\frac{d}{ds}\begin{pmatrix} x_1 \\ \vdots \\ x_n \\ u \end{pmatrix} = \begin{pmatrix} b_1(x) \\ \vdots \\ b_n(x) \\ f(x) - c(x)u(x) \end{pmatrix}, \tag{17.3}$$

dem *charakteristischen* System von (17.2). Die Lösungskurven $(x,u)^T(s)$ im $\mathbf{R}^n \times \mathbf{R}$ heißen *Charakteristiken* von (7.2). Die Tangentialrichtung $\frac{d}{ds}(x,\ u)^T(s)$ heißt charakteristische Richtung im Punkt $(x,u(x))^T$. Die Projektion $(x_1,...,x_n)^T(s)$ einer Charakteristik von (17.2) in den $(x_1,...,x_n)-$Raum heißt *Grundcharakteristik*. Insbesondere sind die Grundcharakteristiken unabhängig von der gesuchten Lösung u.

In physikalischer Interpretation beschreibt das charakteristische System (17.3) die Bewegung einzelner Teilchen eines durch (17.2) beschriebenen dichten Feldes. Die Integration des charakteristischen Systems (17.3) ist in der Regel nicht exakt möglich. Insbesondere kann dessen Lösung (deterministisch) chaotisch verlaufen (vgl. das Beispiel des Lorenz-Attraktors).

Wir können hier nicht näher auf den engen mathematischen Zusammenhang zwischen den Problemen (17.2) und (17.3) eingehen. Daher beschränken wir uns auf die Frage, wie man geeignete Zusatzbedingungen für Lösungen der Gleichung (17.2) findet. Die Vorgabe von Zusatzbedingungen richtet sich nach dem Verhalten der Grundcharakteristiken auf dem Rand von Ω. In geometrischer Interpretation werden Randwerte dort vorgegeben, wo die Grundcharakteristiken den Gebietsrand von außen nach innen schneiden.

Bei gegebenem Gebiet Ω mit Lipschitz-stetigem Rand $\partial\Omega$, d.h. der äußere Normaleneinheitsvektor $\vec{\nu} = (\nu_1,...,\nu_n)^T$ existiert fast überall auf $\partial\Omega$, gibt man Randwerte vor auf dem "Einströmrand"

$$\Sigma_- := \{x \in \partial\Omega :\ \exists \vec{\nu}(x),\ (\vec{b}\cdot\vec{\nu})(x) := \sum_{i=1}^{n} b_i(x)\nu_i(x) < 0\}. \tag{17.4}$$

Bemerkung 17.1. Man kann bei Problemen 1. Ordnung Zusatzbedingungen alternativ dort stellen, wo die Grundcharakteristiken den Gebietsrand von innen nach außen verlassen. Dies entspricht einer Umkehrung des Durchlaufsinns. Wir werden aber sehen, daß dies bei den in der Einleitung des Kapitels betrachteten Problemen mit $0 < \epsilon \ll 1$ nicht sachgemäß ist. □

17.2 Transportdominierte Konvektions-Diffusions Probleme

Nach unserem Plan betrachten wir jetzt elliptische Probleme 2. Ordnung

$$(L_\epsilon u)(x) := -\epsilon \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{j=1}^{n} b_j \frac{\partial u}{\partial x_j} + cu = f(x), \qquad x \in \Omega \tag{17.5}$$

mit einem (eventuell kleinen) Parameter $0 < \epsilon \leq 1$. Wir werten für deren Analyse unsere bisherigen theoretischen Resultate kritisch aus und geben notwendige Konsequenzen an. Dann untersuchen wir das Grenzverhalten solcher Modelle für $\epsilon \to +0$ und kombinieren die Untersuchung mit den Erkenntnissen zu hyperbolischen Problemen 1. Ordnung aus Abschnitt 17.1.

Bemerkung 17.2. In Anwendungen steht die gesuchte Größe $u = u(x)$ für ein skalares Feld (z.B. Temperatur, Dichte usw.). Wir hatten in Kapitel 4 gesehen, daß das Problem (17.5) als *Diffusion-Konvektion-Reaktion Modell* interpretiert werden kann. Die Änderung des Feldes durch Diffusion repräsentiert der elliptische Hauptteil. Die Änderung von u unter dem Einfluß eines konvektiven Feldes $\vec{b}$ (z.B. eines Strömungs- oder Windfeldes) sowie durch Reaktion (z.B. chemischer Art) beschreiben der konvektive Term (oder Transport-Term) $\vec{b}\cdot\nabla u$ bzw. der Reaktionsterm $f - cu$. Im Fall $\epsilon \to +0$ reduziert sich das Modell auf ein *Konvektion-Reaktion Modell*. □

Die Lösung des Modells (17.5) entartet in der Regel für $\epsilon \to +0$. Formal wird aus dem elliptischen Problem 2. Ordnung eine partielle Differentialgleichung 1. Ordnung, für die in der Regel eine Randbedingung nur noch auf einem Teil des Randes $\partial\Omega$, dem "Einströmrand", erfüllt werden kann

Beispiel 17.3. Zur Illustration betrachten wir das Zweipunkt-RWP

$$(L_\epsilon u)(x) := -\epsilon u''(x) - u'(x) = 0, \quad 0 < x < 1; \qquad u(0) = 0, \quad u(1) = 1 \tag{17.6}$$

mit der exakten Lösung $u(x) = \frac{1-\exp(-x/\epsilon)}{1-\exp(-1/\epsilon)}$. Für $\epsilon \to +0$ hat die Lösung den Grenzwert $\lim_{\epsilon \to +0} u(x) = 1$ für alle Punkte $x \in (0,1]$. Dies ist gerade die Lösung des Grenzproblems

$$-u'(x) = 0, \quad 0 \le x < 1; \qquad u(1) = 1, \tag{17.7}$$

das nach unseren Überlegungen in Abschnitt 7.1 nur die Randbedingung am "Einströmrand" $x = 1$ erfüllen kann. Am "Ausströmrand" $x = 0$ ändert sich die Lösung von (7.5) im Fall $0 < \epsilon \ll 1$ exponentiell schnell. Man spricht von *Grenzschichtverhalten.* □

Zur genaueren Analyse beschränken wir uns nachfolgend auf das Dirichletsche Randwertproblem

$$(L_\epsilon u)(x) = f(x), \quad x \in \Omega \tag{17.8}$$
$$u(x) = 0, \qquad x \in \partial\Omega. \tag{17.9}$$

Für dieses Problem können wir für jedes fest $\epsilon \in (0,1]$ die Theorie von Kapitel 4 anwenden. Insbesondere erhält man unter den Voraussetzungen von Lemma 4.1, d.h. u.a. unter der Bedingung

$$c(x) - \frac{1}{2}(\nabla \cdot \vec{b})(x) \ge 0, \qquad x \in \Omega,$$

die Existenz und Eindeutigkeit einer verallgemeinerten Lösung $u \in X := W_0^{1,2}(\Omega)$.

Für $\epsilon \to +0$ reduziert sich das RWP auf das Grenzproblem

$$(L_1 U)(x) := \sum_{i=1}^{n} b_i(x) \frac{\partial U}{\partial x_i} + c(x)U = f(x), \quad x \in \overline{\Omega} \setminus \Sigma_- \tag{17.10}$$
$$U(x) = 0, \qquad x \in \Sigma_-. \tag{17.11}$$

Zum Nachweis der *Existenz* einer verallgemeinerten Lösung kann man die (eindeutige) Lösung U von (17.10)-(17.11) als Grenzwert für $\epsilon \to +0$ des *singulär gestörten* elliptischen RWP 2. Ordnung (17.8)-(17.9) charakterisieren. Man spricht auch von einer *elliptischen Regularisierung* des RWP 1. Ordnung bzw. vom *"vanishing viscosity"*-Konzept. Die Lösbarbeit von (17.10)-(17.11) ergibt sich aus folgendem Resultat.

Satz 17.4. *Unter den Voraussetzungen von Satz 6.10 konvergiert die Lösungsfolge $\{u_\epsilon\}$ des singulär gestörten elliptischen Randwertproblems (17.8)-(17.9) schwach gemäß*

$$u_\epsilon \rightharpoonup U \quad \text{in } L^2(\Omega), \quad \epsilon \to +0 \tag{17.12}$$

gegen die eindeutig bestimmte verallgemeinerte Lösung U des RWP 1. Ordnung aus (17.10)-(17.11).

Wir vermerken noch einige *Regularitätsaussagen:*

- Seien $c(x) \ge \beta^2 > 0$ in $\overline{\Omega}$ sowie $b_1, ..., b_n \in C^1(\overline{\Omega})$ und $c, f \in C(\overline{\Omega})$. Dann ist U stetig in $\overline{\Omega}$ fast überall. Einfache Beispiele zeigen, daß i.a. Fall nicht $U \in C(\overline{\Omega})$ gilt.

- Differenzierbarkeitsaussagen an U gelten nur unter sehr einschneidenden geometrischen Voraussetzungen an das Randverhalten der Grundcharakteristiken, vgl. [17].

Wir hatten in Beispiel 17.3 gesehen, daß für $\epsilon \to +0$ die Lösung U des RWP 1. Ordnung (17.10)-(17.11) i.a. Fall nicht die Randbedingung $U = 0$ auf $\partial\Omega \setminus \Sigma_-$ erfüllt. Die Lösung $u \in W_0^{1,2}(\Omega)$ des singulär gestörten Problems (17.8)-(17.9) ändert sich in Umgebung von $\partial\Omega \setminus \Sigma_-$ eventuell sehr stark gegenüber U. Man spricht vom *Grenzschichtcharakter* von u für $0 < \epsilon \ll 1$. Ähnliche Probleme treten auch dort im Gebiet Ω auf, wo die Grenzlösung U nicht hinreichend glatt ist. Dort bilden sich sogenannte *innere* Grenzschichten aus. Genauere Aussagen hierzu findet man zum Beispiel in der Monographie [19].

17.3 Stabile Diskretisierung transportdominierter Probleme

In den letzten Jahren hat sich gezeigt, daß sich *stabilisierte* FEM zur numerischen Lösung von transportdominierten Problemen eignen. Nach unseren Vorbetrachtungen über hyperbolische Gleichungen 1. Ordnung scheint es sinnvoll, bei der Konstruktion solcher Verfahren die Informationsausbreitung entlang der Grundcharakteristiken auszunutzen. Es gibt eine Reihe von charakteristiken-basierten Verfahren (vgl. zum Beispiel [19], Kap. II.3), auf die wir hier aber nicht eingehen können. Die hier betrachteten Methoden berücksichtigen jedoch indirekt den Charakteristikenverlauf ebenfalls in geeigneter Weise.

Beispiel 17.5. Zur Motivation betrachten wir stabilisierte Diskretisierungen des eindimensionalen Problems

$$(L_\epsilon u) := -\epsilon u''(x) + bu'(x) = f(x), \quad 0 < x < 1; \qquad u(0) = u(1) = 0 \tag{17.13}$$

mit konstanten Koeffizienten $\epsilon > 0$ und (vereinfachend) $b > 0$. Die Galerkin-Diskretisierung mit P_1-Elementen auf einem äquidistanten Gitter mit Schrittweite $h = \frac{1}{N+1}$ führt auf das tridiagonale Gleichungssystem

$$-\epsilon\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + b\frac{U_{j+1} - U_{j-1}}{2h} = F_j \tag{17.14}$$

Dabei ist U_j eine Approximation an den gesuchten Wert $u(x_j) = u(jh),\ j = 1, \ldots, N$, analog für F_j. Dieses System erzeugt im Fall $\rho := \frac{h|b|}{2\epsilon} \gg 1$ eine unphysikalische oszillierende Lösung (bei linearer Interpolation), vgl. dazu Übungsaufgabe. Dabei ist ρ die sogenannte Gitter-Peclet-Zahl.

Als Ausweg ändert man das Schema ab in

$$-\epsilon\sigma\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + b\frac{U_{j+1} - U_{j-1}}{2h} = F_j \tag{17.15}$$

mit geeignetem Parameter $\sigma \geq 0$. Im Fall $\sigma = 1 + \rho$ erhält man das *einfache upwind*-Verfahren

$$-\epsilon\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + b\frac{U_j - U_{j-1}}{h} = F_j. \tag{17.16}$$

Das Verfahren berücksichtigt im konvektiven Term im Fall $\epsilon = 0$ die Richtung der Grundcharakteristik mit $\frac{dx}{ds} = b$. Es führt aufgrund der *künstlichen Viskosität* $\epsilon_h = \rho := \frac{h|b|}{2\epsilon}$ auf eine diagonal-dominante Matrix und damit zur Stabilität bezüglich der diskreten l_∞-Norm. Andererseits hat es wegen der geänderten Viskosität für festes $\epsilon \geq 0$ nur die Konvergenzordnung $\mathcal{O}(h)$.

Diese Betrachtung motiviert die Wahl von σ in Abhängigkeit von der Zahl $\rho := \frac{hb}{2\epsilon}$. Für $\sigma(\rho) > \rho$ bleibt die Stabilitätseigenschaft der zugehörigen Matrix erhalten. Eine genaue Analyse zeigt, daß das sogenannte *Iljin*-Schema mit

$$\sigma(\rho) = \rho\coth\rho, \qquad \rho := \frac{hb}{2\epsilon} \tag{17.17}$$

für $h \to +0$ in den Knotenpunkten gegen die exakte Lösung u konvergiert. Das Schema ist für $\epsilon \geq 0$ stabil und hat gleichmässig bezüglich $\epsilon \to +0$ die Konvergenzordnung $\mathcal{O}(h)$ in der Maximum-Norm. □

Die in Beispiel 17.5 betrachteten upwind-Verfahren haben vor allem den Nachteil einer niedrigen Konvergenzordnung. Hauptziel bei der Konstruktion geeigneter FEM ist daher neben der Sicherung guter Stabilitätseigenschaften die Gewinnung einer möglichst hohen Konvergenzordnung. Die Grundideen sind:

- Addition künstlicher Diffusion in Richtung der Grundcharakteristiken und
- Konsistenz der FEM zur Lösung des kontinuierlichen Problems.

Wir betrachten nun das stationäre Randwertproblem

$$-\epsilon\Delta u + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x), \quad x \in \overline{\Omega} \setminus \Sigma_- \tag{17.18}$$

$$u = 0, \quad x \in \Sigma_- \tag{17.19}$$

mit $\Sigma_- := \partial\Omega$ für $\epsilon > 0$ sowie $\Sigma_- := \{x \in \partial\Omega : \ \sum_{i=1}^n (b_i \nu_i)(x) < 0\}$ für $\epsilon = 0$. Es seien die Voraussetzungen aus Lemma 4.1 an die Daten des Problems erfüllt. Mit den Räumen

$$X := W_0^{1,2}(\Omega), \ \epsilon > 0; \qquad X := \{v \in L^2(\Omega) : \ \sum_{i=1}^n b_i(x) \frac{\partial v}{\partial x_i} \in L^2(\Omega), \ v|_{\Sigma_-} = 0\}, \ \epsilon = 0$$

erhält man in der üblichen Weise die Variationsformulierung

$$\text{Finde} \ \ u \in X : \ \ B_G(u,v) = f(v) \quad \forall v \in X \tag{17.20}$$

$$B_G(u,v) := \epsilon(\nabla u, \nabla v)_{L^2(\Omega)} + (\vec{b} \cdot \nabla u + cu, v)_{L^2(\Omega)}; \qquad f(v) := (f,v)_{L^2(\Omega)}.$$

Dabei sind $\vec{b} := (b_1, ..., b_n)^T$ sowie $(\cdot,\cdot)_{L^2(G)}$ das Skalarprodukt in $L^2(G)$ für meßbare Mengen $G \subseteq \Omega$.

Die a-priori Abschätzung

$$\epsilon \|\nabla u\|^2_{L^2(\Omega)} + \kappa \|u\|^2_{L^2(\Omega)} \leq C(\kappa, \epsilon) \|f\|^2_{L^2(\Omega)}$$

mit $\kappa = \min_{x \in \Omega}(c - \frac{1}{2} b)(x) \geq 0$ zeigt, daß für $\epsilon \to +0$ die Kontrolle über den Gradienten der Lösung verloren geht.

Zur FEM-Diskretisierung betrachten wir nun eine zulässige, quasi-uniforme Zerlegung $\mathcal{T}_h$ von Ω. K bezeichne ein beliebiges Element der Zerlegung. Sei ferner $X_h \subset X$ ein konformer FEM-Raum mit stückweise polynomialen Basisfunktionen vom Grad k, d.h.

$$X_h := \{v \in X : \ v|_K \in \mathcal{P}_k(T), \ \ \forall K \in \mathcal{T}_h\}.$$

In X_h gelten mit dem Interpolationsoperator $\Pi^k_{\mathcal{T}}$ nach Kapitel 6.1 die *lokale Interpolationsaussage*

$$|u - \Pi^k_{\mathcal{T}} u|_{W^{m,2}(K)} \leq C h^{k+1-m} |u|_{W^{k+1,2}(K)}, \quad m = 0, 1, 2. \tag{17.21}$$

Ferner gilt eine sogenannte *inverse* Ungleichung

$$\|\Delta v\|_{L^2(K)} \leq \mu_{inv} h_K^{-1} |v|_{W^{1,2}(K)} \quad \forall v \in X_h. \tag{17.22}$$

Für die Galerkin-FEM

$$\text{Finde} \ \ u \in X : \ \ B_G(u,v) = f(v) \quad \forall v \in X \tag{17.23}$$

übertragt sich die a-priori Abschätzung der Lösung, d.h. wegen

$$\epsilon \|\nabla u_h\|^2_{L^2(\Omega)} + \kappa \|\nabla u_h\|^2_{L^2(\Omega)} \leq C(\kappa, \epsilon) \|f\|^2_{L^2(\Omega)}$$

mit $\kappa \geq 0$ verliert man für $\epsilon \to +0$ die Kontrolle über den Gradienten der Lösung. Es tritt ein ähnlicher Destabilisierungeffekt wie beim zentralen Differenzen-Schema im räumlich eindimensionalen Fall auf. Man versucht daher, die Idee stabilisierter Verfahren (z.B. einfaches upwind- oder Iljin-Verfahren) aus Beispiel 17.5 zu übertragen. Im mehrdimensionalen Fall ist jedoch größere Sorgfalt bei der Wahl der Stabilisierung erforderlich.

Bei der *Stromlinien-Diffusion FEM* werden zur Galerkin-FEM gewichtete Residuen von Gleichung (17.18) addiert. Unter der Regularitätsannahme

$$-\epsilon \Delta u + \vec{b} \cdot \nabla u + cu = f \quad \text{in } L^2(K) \ \ \forall K \in \mathcal{T}_h \tag{17.24}$$

ist die Lösung von (17.18),(17.19) auch Lösung des abgeänderten kontinuierlichen Problems

$$B_{SG}(u,v) = L_{SG}(v) \quad \forall v \in X_h, \tag{17.25}$$

$$B_{SG}(u,v) := B_G(u,v) + \sum_{K \in \mathcal{T}_h} \delta_k (-\epsilon \Delta u + \vec{b} \cdot \nabla u + cu, \vec{b} \cdot \nabla v)_{L^2(K)} \tag{17.26}$$

$$L_{SG}(u,v) := f(v) + \sum_{K \in \mathcal{T}_h} (f, \vec{b} \cdot \nabla v)_{L^2(K)}. \tag{17.27}$$

Dabei sind $\delta_K \geq 0$ noch zu bestimmende Parameter. Damit lautet die *Stromlinien-Diffusions-Methode*

$$\text{Finde } U_h \in X_h: \quad B_{SG}(U_h, v) = L_{SG}(v) \quad \forall v_h \in X_h. \tag{17.28}$$

Unter der Regularitätsannahme (17.24) gilt dann die wichtige Relation der *Galerkin-Orthogonalität*

$$B_{SG}(U_h - u, v) = 0 \quad \forall v \in X_h. \tag{17.29}$$

Sie wird die gewünschte hohe Konsistenz- und Konvergenzordnung des Verfahrens sichern. Wir analysieren die Methode bezüglich der folgenden gitterabhängigen Norm

$$|||v|||_{SG} := \left(\epsilon \|\nabla v\|^2_{L^2(\Omega)} + c_0 \|v\|^2_{L^2(\Omega)} + \sum_K \delta_K \|\vec{b} \cdot \nabla v\|^2_{L^2(K)} \right)^{\frac{1}{2}}. \tag{17.30}$$

Dabei seien

$$c - \frac{1}{2} \nabla \cdot \vec{b} \geq c_0 > 0; \qquad c_K := \|c\|_{L^\infty(K)}. \tag{17.31}$$

Die Existenz und Stabilität der diskreten Lösung folgen aus

Satz 17.6. *Neben Glattheitsvoraussetzungen seien (17.31) sowie*

$$0 < \delta_K \leq \frac{1}{2} \min \left(\frac{c_0}{c_K^2}; \frac{h_K^2}{\epsilon \mu_{inv}^2} \right) \quad \forall K \in \mathcal{T}_h \tag{17.32}$$

erfüllt. Dann ist die Bilinearform $B_{SG}(\cdot, \cdot)$ X_h-elliptisch mit

$$B_{SG}(v, v) \geq \frac{1}{2} |||v|||^2_{SG} \quad \forall v \in X_h. \tag{17.33}$$

Beweis: Für alle Funktionen $v \in X_h$ gilt

$$B_{SG}(v, v) \geq \epsilon \|\nabla v\|^2_{L^2(\Omega)} + c_0 \|v\|^2_{L^2(\Omega)} + \sum_K \delta_K \|\vec{b} \cdot \nabla v\|^2_{L^2(K)} + \sum_K \delta_K (-\epsilon \Delta v + cv, \vec{b} \cdot \nabla v)_{L^2(K)}.$$

Mit der inversen Ungleichung (17.22) und Voraussetzung (17.32) ergibt sich die Behauptung wegen

$$\begin{aligned} & \sum_K \delta_K (-\epsilon \Delta v + cv, \vec{b} \cdot \nabla v)_{L^2(K)} \\ \leq \; & \sum_K \epsilon^2 \delta_K \|\Delta v\|^2_{L^2(K)} + \sum_K c_K^2 \delta_K \|v\|^2_{L^2(K)} + \frac{1}{2} \sum_K \delta_K \|\vec{b} \cdot \nabla v\|^2_{L^2(K)} \\ \leq \; & \frac{1}{2} \left(\epsilon \|\nabla v\|^2_{L^2(\Omega)} + c_0 \|v\|^2_{L^2(\Omega)} + \sum_K \delta_K \|\vec{b} \cdot \nabla v\|^2_{L^2(K)} \right). \end{aligned}$$

□

Bemerkung 17.7. Die Aussage von Satz 17.6 zeigt, daß sich (in Verallgemeinerung der Idee der *upwind*-Verfahren) durch die Stabilisierung des Galerkin-Verfahrens zusätzliche Kontrolle über die Ableitung der Lösung in charakteristischer Richtung $\vec{b}$, der *Stromlinien-Richtung*, ergibt. Dies kompensiert den Verlust jeglicher Kontrolle über Ableitungen der Lösung beim Galerkin-Verfahren ($\delta_K = 0$) für $\epsilon \to +0$. □

Eine *globale* Konvergenzaussage gibt der

Satz 17.8. *Unter den Voraussetzungen von Satz 17.6 und mit der Parameterwahl*

$$\delta_K \sim \frac{h_K}{b_K}, \qquad b_K := \|\vec{b}\|_{L^\infty(K)} \tag{17.34}$$

gilt unter der Regularitätsvoraussetzung $u \in W^{k+1,2}(\Omega)$ für $\epsilon \leq Ch$ mit $B := max_K b_K$ die folgende Fehlerabschätzung

$$|||u - U_h|||_{SG} \leq C \left(\sqrt{\epsilon} + \sqrt{Bh} + h \right) h^k |u|_{W^{k+1,2}(\Omega)}. \tag{17.35}$$

Das Programmsystem **FEMLAB** stellt für eine relativ große Problemklasse eine Routine zur Stabilisierung mittels Stromliniendiffusion (SUPG) zur Verfügung. Dies gilt auch für den hyperbolischen Grenzfall 1. Ordnung mit $\epsilon = 0$ und nichtlineare Probleme. Während in der Version 2.2 noch eine sehr aufwendige Berechnung der Parameter δ_K enthalten ist, liegt in der Version 2.3 eine stark vereinfachte Berechnung (ähnlich wie in Satz 17.8) vor.

Wir wollen noch zwei Beispiele mittels **FEMLAB** berechnen.

Beispiel 17.9. Wir betrachten das Problem

$$\begin{aligned} -\epsilon\Delta u + (-x_2, x_1)^T \cdot \nabla u &= 0, \qquad x \in \Omega \\ u(x) &= 0, \qquad x \in \partial\Omega \end{aligned}$$

im Gebiet $\Omega := (0,1)^2$ für den (extrem kleinen) Wert $\epsilon = 10^{-10}$. (Man behandelt quasi den hyperbolischen Grenzfall mit $\epsilon = 0$.) Die rechte Seite f wird so bestimmt, daß die (hinreichend glatte) exakte Lösung $u(x) = \sin(\pi x_1)\sin(\pi x_2)e^{x_1 x_2}$ lautet.

Das Galerkin-Verfahren (d.h. mit $\delta_K = 0$) mit P_k-Elementen für $k \in \{1, \ldots, 6\}$ (nicht gezeigt) liefert erst für hinreichend kleine Werte von h brauchbare Konvergenzresultate. Wir zeigen in den Abbildung 17.1 die Konvergenzdiagramme für das stabilisierte Galerkin-Verfahren für P_k-Elementen für $k \in \{1, \ldots, 6\}$. Man ersieht, daß für die $L^2(\Omega)$- und $W^{1,2}(\Omega)$-Norm jeweils die optimale Ordnung erreicht wird. □

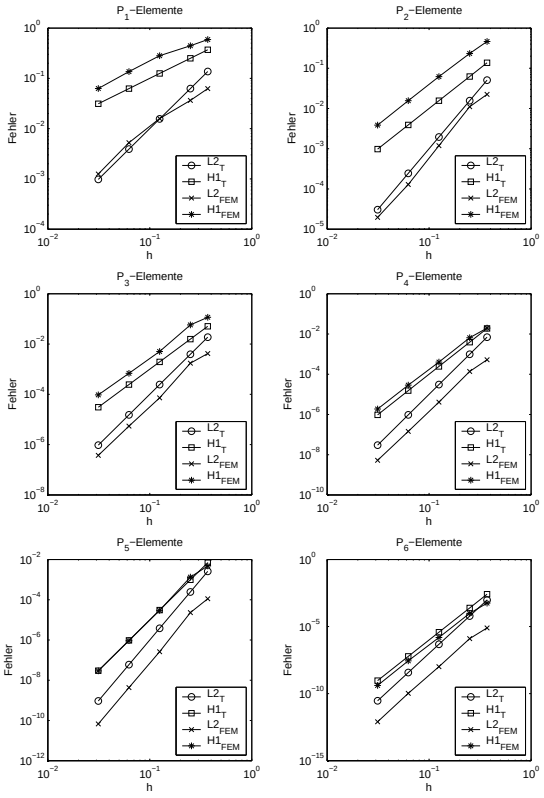


Abbildung 17.1: Konvergenzdiagramme für das stabilisisierte Galerkin-Verfahren in Beispiel 17.9

Beispiel 17.10. Wir greifen jetzt ein Beispiel mit nichtglatter Grenzlösung für $\epsilon \to +0$ auf:

$$\begin{aligned} L_\epsilon u := -\epsilon\Delta u + (1,1)^T \cdot \nabla u &= 0, \qquad x \in \Omega \\ u(x) &= 1, \qquad x \in \Gamma \\ u(x) &= 0, \qquad x \in \partial\Omega \setminus \Gamma \end{aligned}$$

im Gebiet $\Omega := (0,1)^2 \setminus \{x \in \mathbf{R}^2 : \ \|x - (\frac{1}{2}, \frac{1}{2})^T\| < \frac{1}{4}\}$. Es sei $\Gamma := \{(x_1, x_2) \in \partial\Omega \ : \ x_2 = 0\}$. Allerdings wurde hier das dazugehörige instationäre Problem $\frac{\partial u}{\partial t} + L_\epsilon u = f$ mit der Zeitschrittweite $\tau = 0.5$ gerechnet, vgl. hierzu Teil II der Vorlesung.

Die hier gezeigte Lösung zum Zeitpunkt $t = 10$ entspricht aber der Lösung des stationären Grenzfalls. Die Grenzlösung für $\epsilon = 0$ erfüllt in diesem Fall Randbedingungen dort, wo die Grundcharakteristiken $x_1 = x_2 + const.$ in das Gebiet Ω eintreten. Dies ist einerseits bei $x_1 = 0$ und $x_2 = 0$ sowie andererseits auf der der Strömungsrichtung abgewandten Seite des Innenkreises (den man als Hindernis in einer Strömung ansehen kann) der Fall. Durch die Wahl der Randbedingungen ist die Grenzlösung unstetig entlang der Grundcharakteristik $x_1 = x_2$ bis zum Erreichen des Hindernisses. Ferner tritt entlang der Grundcharakteristik $x_2 = x_1 - \frac{1}{3}$ nach Tangierung des Hindernisses bis zum Rand bei $x_1 = 1$ eine Unstetigkeit auf. Außerdem erfüllt die Grenzlösung nicht die Randbedingungen auf $\{1\} \times (0, \frac{2}{3})$ und auf der der Strömung zugewandten Seite des Hindernisses. Daher entstehen im Fall $0 < \epsilon \ll 1$ entlang der genannten eindimensionalen Mannigfaltigkeiten innere bzw. Randgrenzschichten der Lösung auf. Abbildung 17.2 zeigt

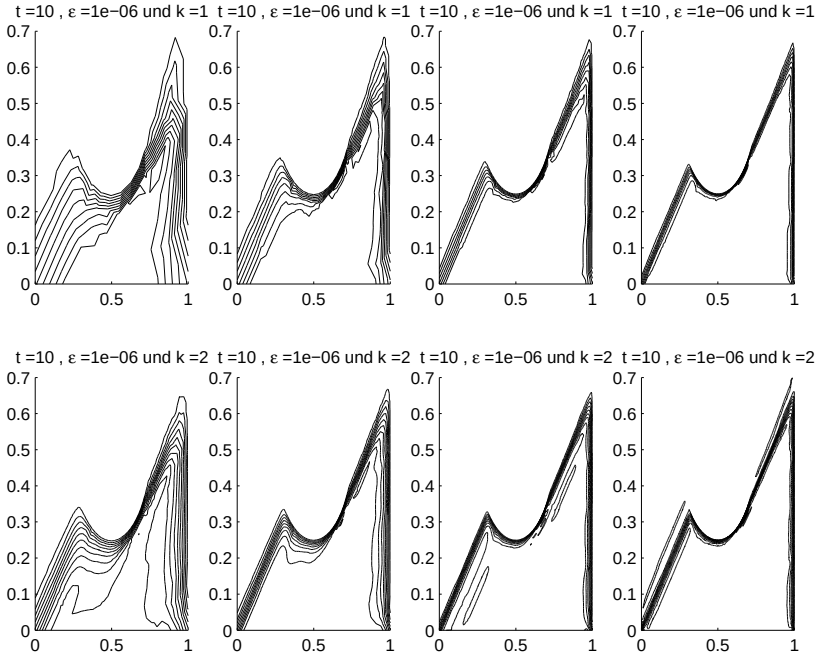


Abbildung 17.2: SUPG-Lösung von Beispiel 17.10 für $\epsilon = 10^{-6}$, $k = 1, 2$ und verschiedene Werte von h

die Isolinien der stabilisierten FEM-Lösungen im Fall $\epsilon = 10^{-6}$ mit P_k-Elementen mit $k \in \{1, 2\}$ jeweils auf einer Sequenz von quasi-uniformen Dreiecksnetzen mit $h \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$. Die Zahl der Unbekannten beträgt im Fall $k = 1$ jeweils $142,\ 520, 1984$ bzw. 7744, im Fall $k = 2$ sind es $520,\ 1984, 7744$ bzw. 16351.

Man erkennt, daß die inneren bzw. Randgrenzschichten mit wachsender Feinheit immer besser ermittelt werden. Auf groben Gittern werden die Grenzschichten durch P_1-Elemente stärker "verschmiert". Ein Vorteil der P_2-Elemente ist hier die schärfere Auslösung der Grenzschichten. Da die Lösung abseits der Grenzschichten im Prinzip konstant ist, kommt hier der Vorteil höherer Ordnung nicht zum Zug. Auf dem feinsten Gitter ist eigentlich die P_1-Lösung zu bevorzugen.

Man sieht aber auch in geringer Umgebung dieser Schichten noch Oszillationen der diskreten Lösung. Diese zeigen an, daß die gezielte Addition künstlicher Viskosität in Richtung der Charakteristiken (Stromlinien) allein noch nicht ausreicht. Mittels sogenannter "shock-capturing"-Stabilisierung quer zur Stromlinien-Richtung, in die sogenannte "crosswind"-Richtung, kann man auch die restlichen Instabilitäten der diskreten Lösung noch mildern. Allerdings führt die Sicherung der Konsistenz des Verfahrens leider auf eine nichtlineare Methode. In **FEMLAB** ist dieses Verfahren, das auch noch Gegenstand aktueller Forschung ist, noch nicht vorgesehen. □

Kapitel 18

Gemischte Probleme

In diesem Kapitel wollen wir die Theorie spezieller *Operatorgleichungen mit Nebenbedingungen*, die man als sogenannte *gemischte Probleme* formulieren kann, behandeln. Neben linearen gemischten Problemen im kontinuierlichen Fall wird auch deren konforme Finite-Elemente-Approximation besprochen. Einfache Anwendungen findet man bei der Behandlung inhomogener Dirichlet-Randbedingungen oder auch beim Neumann-Randwertproblem der Poisson-Gleichung.

Weitere wichtige Anwendungsbeispiele sind inkompressible Strömungmodelle, die hier leider aus Zeitgründen nicht weiter behandelt werden können. Als weiterführende Literaturquellen kann man z.B. die Monographie von V. Girault, P. A. Raviart [8] benutzen. Eine gründliche Behandlung gemischter Probleme erfolgt im kommenden Wintersemester im Rahmen eines Seminars.

18.1 Variationsprobleme mit Nebenbedingungen

Wir führen zunächst einige *Bezeichnungen* ein: V bzw. Q seien reelle Hilbert-Räume mit den Normen $\|\cdot\|_V$ bzw. $\|\cdot\|_Q$. V^* bzw. Q^* sind die zu V bzw. Q gehörenden dualen Räume. Die entsprechenden Normen sind $\|\cdot\|_{V^*}$ bzw. $\|\cdot\|_{Q^*}$. Mit $\langle\cdot,\cdot\rangle$ bezeichnen wir jeweils das Dualitätsprodukt zwischen V^* und V sowie zwischen Q^* und Q.

Weiter sind

$$a(\cdot,\cdot): \quad V\times V\to\mathbf{R} \qquad b(\cdot,\cdot): \quad V\times Q\to\mathbf{R}$$

stetige Bilinearformen mit den Normen

$$\|a\| := \sup_{u,v\in V\setminus\{0\}} \frac{a(u,v)}{\|u\|_V\|v\|_V}, \qquad \|b\| := \sup_{\substack{v\in V\setminus\{0\}\\ q\in Q\setminus\{0\}}} \frac{b(v,q)}{\|v\|_V\|q\|_Q}. \tag{18.1}$$

Gegenstand der Untersuchungen ist das folgende *gemischte Variationsproblem* **(Q)**:

$$\text{Finde } \ \hat u := (u,p)\in X := V\times Q, \ \text{ so daß } \ \forall \hat v := (v,q)\in X \tag{18.2}$$

$$\begin{aligned} a(u,v)+b(v,p) &= \langle f,v\rangle\\ b(u,q) &= \langle g,q\rangle. \end{aligned}$$

Wir wollen ein zu **(Q)** äquivalentes System von Operatorgleichungen ableiten. Nach dem Darstellungssatz für stetige Bilinearformen (vgl. Lemma 6.2) gilt

$$\exists\,!\quad A\in\mathcal{L}(V,V^*): \quad \langle Au,v\rangle := a(u,v) \quad \forall u,v\in V \tag{18.3}$$

$$\exists\,!\quad B\in\mathcal{L}(V,Q^*): \quad \langle Bv,q\rangle := b(v,q) \quad \forall v\in V, \ \ \forall q\in Q. \tag{18.4}$$

Ferner nutzen wir

Lemma 18.1 *(Dualer Operator)*
Für jeden Operator $C \in \mathcal{L}(X,Y)$ mit Banach-Räumen X und Y existiert eindeutig der duale Operator *$C^* \in \mathcal{L}(Y^*, X^*)$ mit*

$$\langle f, Cx\rangle = \langle C^* f, x\rangle \quad \forall f \in Y^*, x \in X.$$

Als Anwendung folgt sofort

$$\exists\,!\quad B^* \in \mathcal{L}(Q, V^*): \quad \langle B^* q, v\rangle = \langle Bv, q\rangle = b(v,q) \quad \forall v \in V, q \in Q. \tag{18.5}$$

Das gesuchte äquivalente System **(Q')** von Operatorgleichungen ist

$$\text{Finde } \hat{u} := (u,p) \in X := V \times Q, \text{ so daß} \tag{18.6}$$

$$\begin{aligned} Au + B^* p &= f \quad \text{in } V^* \\ Bu &= g \quad \text{in } Q^*. \end{aligned}$$

Wir benötigen noch einige Bezeichnungen: Für einen Operator $C : X \to Y$ sind

$$\mathcal{R}(C) := C(X) \quad \text{bzw.} \quad \mathcal{N}(C) := \{v \in X: \quad Cv = 0 \quad \in X^*\}$$

der *Bildbereich* bzw. der *Nullraum* (oder Kern) des Operators.

Die zweite Gleichung in **(Q')** bzw. **(Q)** kann auch als *Nebenbedingung* zur ersten Gleichung interpretiert werden. Dazu führen wir ein

$$W(g) := \{v \in V: \quad b(v,q) = \langle g, q\rangle \ \ \forall q \in Q\} \subseteq V. \tag{18.7}$$

Speziell ist die Menge

$$W := W(0) \equiv N(B) \tag{18.8}$$

wegen der Stetigkeit von B abgeschlossener Teilraum von V.

Das gemischte Problem **(Q)** lautet dann als *restringiertes Variationsproblem* **(P)** :

$$\text{Finde } \ u \in W(g): \quad a(u,v) = \langle f, v\rangle, \ \forall v \in W \equiv W(0). \tag{18.9}$$

Man beachte, daß in dieser Formulierung die Größe p nicht mehr explizit vorkommt. Offenbar ist für jede Lösung $(u,p) \in V \times Q$ von **(Q)** auch $u \in W(g)$ Lösung von **(P)**.

Zu klären ist, ob zu einer Lösung $u \in W(g)$ von **(P)** ein Element $p \in Q$ bestimmt werden kann, so daß (u,p) Lösung von **(Q)** ist. Stellt man bei bekannntem $u \in W(g)$ die erste Gleichung in **(Q')** um zu

$$B^* p = f - Au,$$

so gelangt man zur Frage der Invertierbarkeit des Operators B^*.

Wir betrachten folgendes Beispiel der schwachen Erfüllung inhomogener Dirichlet-Bedingungen, die (in modifizierter Form) auch im Programmsystem FEMLAB benutzt wird.

Beispiel 18.2. *(Inhomogenes Dirichlet-Problem)*
Gelöst werden soll vereinfachend das inhomogene Dirichlet-Randwertproblem der Poisson-Gleichung

$$-(\Delta u)(x) := -\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(x) = f(x) \quad \text{in } \ \Omega \subset \mathbf{R}^n; \qquad u(x) = g(x) \quad \text{auf } \ \partial\Omega. \tag{18.10}$$

Die Randbedingung soll in schwacher Form erfüllt werden. Dazu wählen wir den Lösungsraum $V = W^{1,2}(\Omega)$ und zunächst den Raum $Q = L^2(\partial\Omega)$. Mit den Bezeichnungen

$$\begin{aligned} a(u,v) &:= \int_\Omega \nabla u \cdot \nabla v \ dx, \qquad f(v) := \int_\Omega f v \ dx \\ b(v,w) &:= \int_{\partial\Omega} vw \ ds, \qquad g(w) := \int_{\partial\Omega} gw \ ds \end{aligned}$$

sowie

$$W(g) := \{v \in W^{1,2}(\Omega) : \quad \int_{\partial\Omega} vw \; ds = \int_{\partial\Omega} gw \; ds \quad \forall w \in L^2(\partial\Omega)\}$$

und $W := W_0^{1,2}(\Omega)$ erhalten wir eine erste schwache Formulierung des Problems (18.10).

Genauere Untersuchungen von Spurräumen zeigen, daß man die Bedingung $W(g) \neq \emptyset$ für glattere Randdaten $g \in H^{1/2}(\partial\Omega)$ erzwingen kann. Man wählt dann Q als Dualraum zu $H^{1/2}(\partial\Omega)$. Wir vermerken, daß das Arbeiten mit dem (schwer handhabbaren) Dualitätsprodukt $\langle \cdot, \cdot \rangle$ zwischen Q^* und Q im diskreten Fall durch ein geeignetes gewichtetes L^2-Skalarprodukt ersetzt werden kann. □

Im Rahmen der Übungen wird besprochen, wie auch die Lösung des inhomogenen Neumann-Problems der Poisson-Gleichung

$$-(\Delta u)(x) = f(x) \quad \text{in} \;\; \Omega \subset \mathbf{R}^n; \qquad \frac{\partial u}{\partial n}(x) = g(x) \quad \text{auf} \;\; \partial\Omega.$$

in diesem Kontext behandelt werden kann. Eindeutigkeit der Lösung erzwingt man im Lösungsraum

$$W(g) = \{v \in W^{1,2}(\Omega) : \quad \int_\Omega v \; dx = 0\}.$$

18.2 Lösbarkeit der kontinuierlichen Probleme

Ziel ist die Formulierung von Bedingungen für die *Invertierbarkeit* der Operators B^* und B. Wir benötigen dazu einige tieferliegende Aussagen der linearen Funktionalanalysis und eine Zusatzbedingung an b. Wesentlich ist eine geeignete Zerlegung des Raumes V gemäß

$$V = W \oplus W^\perp, \qquad W^\perp := \{w \in V : \; (w, v) = 0 \; \forall v \in W\}.$$

$W^\perp$ ist dabei das *orthogonale Komplement* von W in V bezüglich des Skalarproduktes $(\cdot, \cdot)$ in V.

Definition 18.3. *Für einen Unterraum $Z \subset X$ heißt*

$$Z^0 := \{g \in X^* : \quad \langle g, v \rangle = 0 \;\; \forall v \in Z\}$$

polare Menge (bzw. Annihilator) zu Z. Z^0 *besteht also aus den auf Z verschwindenden Funktionalen.*

Die benötigte Aussage aus der Theorie linearer Operatoren ist das *closed range theorem* von S. Banach.

Lemma 18.4. *(Satz vom abgeschlossenen Bildbereich)*
In reellen Banach-Räumen X und Y sei $C \in \mathcal{L}(X, Y)$. Dann sind folgende Aussagen äquivalent:

(i) *Der Bildbereich $\mathcal{R}(C)$ ist abgeschlossen.*

(ii) *Der Bildbereich $\mathcal{R}(C^*)$ ist abgeschlossen.*

(iii) *Es gilt $\mathcal{R}(C^*) = [\mathcal{N}(C)]^0 := \{x^* \in X^* : \quad \langle x^*, x \rangle = 0 \;\; \forall x \in \mathcal{N}(C)\}$.*

(iv) *Es gilt: $\mathcal{R}(C) = [\mathcal{N}(C^*)]^0 := \{y \in Y^* : \quad \langle y^*, y \rangle = 0 \;\; \forall y^* \in \mathcal{N}(C^*)\}$.*

Beweis: vgl. z.B. Yosida *Functional Analysis*, Springer-Verlag 1965. □

Mit dem Satz vom abgeschlossenen Bildbereich beweist man das für unsere Untersuchungen der Operatoren B^* und B entscheidende Lemma.

Lemma 18.5. *Folgende Eigenschaften sind äquivalent:*

(i) *Es gilt die* **Babuška-Brezzi-Bedingung:**

$$\exists \beta > 0 : \qquad \inf_{q \in Q \setminus \{0\}} \sup_{v \in V \setminus \{0\}} \frac{b(v, q)}{\|v\|_V \|q\|_Q} \geq \beta. \tag{18.11}$$

(ii) *Der Operator $B^* : Q \to W^0$ ist Isomorphismus. Ferner gilt*

$$\exists \beta > 0 : \quad \|B^* q\|_{V^*} \geq \beta \|q\|_Q \quad \forall q \in Q \tag{18.12}$$

(iii) *Der Operator $B : W^\perp \to Q^*$ ist Isomorphismus. Ferner gilt*

$$\exists \beta > 0 : \quad \|Bv\|_{Q^*} \geq \beta \|v\|_V \quad \forall v \in W^\perp. \tag{18.13}$$

Beweis: (1) Zunächst ersieht man, daß die Ungleichungen (18.11) und (18.12) offenbar äquivalent sind. Nun folgt aus dieser Äquivalenz bereits die Aussage $(ii) \Rightarrow (i)$.

(2) Wir zeigen: $(i) \Rightarrow (ii)$: Nach (18.12) ist $B^* : Q \to \mathcal{R}(B^*)$ bijektiv und $(B^*)^{-1}$ stetig. Damit ist $B^* : Q \to \mathcal{R}(B^*)$ Isomorphismus. Es verbleibt zu zeigen, daß $\mathcal{R}(B^*) = W^0$ ist. Wegen der Isomorphie-Eigenschaft von B^* ist $\mathcal{R}(B^*) \subseteq V^*$ abgeschlossener Unterraum. Nach Lemma 18.4 folgt dann $\mathcal{R}(B^*) = \mathcal{N}(B)^0 \equiv W^0$.

(3) Nachweis von $(ii) \Leftrightarrow (iii)$: Zur Übung empfohlen ! □

Das gesuchte Resultat über die Lösbarkeit der Probleme **(Q)** und **(P)** gibt

Satz 18.6. *(Lösbarkeit von* **(Q)** *bzw.* **(P)***)*
Unter den Bezeichnungen aus Abschnitt 18.1 sei die Bilinearform $a(\cdot,\cdot)$ $W-$elliptisch, d.h.

$$\exists \alpha > 0 : \quad a(v,v) \geq \alpha \|v\|_V^2 \quad \forall v \in W. \tag{18.14}$$

Ferner gelte für die Bilinearform $b(\cdot,\cdot)$ die Babuška-Brezzi-Bedingung (18.11). Dann hat man folgende Aussagen:

(i) *Es gibt eine und nur eine Lösung $u \in W(g)$ des Problems* **(P)**.

(ii) *Es gibt eine und nur eine Lösung $p \in Q$, so daß mit der Lösung $u \in W(g)$ nach (i) das Paar (u,p) Lösung des Problems* **(Q)** *ist.*

(iii) *Die Abbildung $(f,g) \mapsto (u,p)$ vermittelt einen Isomorphismus von $V^* \times Q^*$ auf $V \times Q$.*

Beweis: (1) *Existenz von u:* Wegen der Babuška-Brezzi-Bedingung (18.11) und Lemma 18.5 (iii) existiert die eindeutige Lösung $u_0 \in W^\perp$ von $Bu_0 = g$ mit

$$\|u_0\|_V \leq \frac{1}{\beta} \|g\|_{Q^*}.$$

Dann kann das Problem **(P)** homogenisiert werden zu

$$\text{Finde } \; w := u - u_0 \in W \; : \quad a(w,v) = \langle f, v \rangle - a(u_0, v) \qquad \forall v \in W. \qquad (\mathbf{P}')$$

Wegen der vorausgesetzten $W-$Elliptizität von $a(\cdot,\cdot)$ ist die Lax-Milgram Theorie anwendbar, d.h. es existiert die eindeutige Lösung $w \in W$ von **(P')** mit

$$\|w\|_V \leq \frac{1}{\alpha} \left(\|f\|_{V^*} + \|a\| \|u_0\|_V \right).$$

Somit ist $u = w + u_0 \in W(g)$ die eindeutige Lösung von **(P)**. Es gilt unter Beachtung der Abschätzungen von u_0 und w eine a-priori Abschätzung der Form

$$\|u\|_V \leq C_1 \left(\|f\|_{V^*} + \|g\|_{Q^*} \right), \quad C_1 := C_1(\alpha, \beta, \|a\|). \tag{18.15}$$

(2) *Existenz von p:* Wegen $f - Au \in W^0$ (vgl. **(P)**) und der Babuška-Brezzi-Bedingung (18.11) ist nun Lemma 18.5 (ii) anwendbar, d.h. es existiert die eindeutige Lösung $p \in Q$ des Problems

$$B^* p = f - Au.$$

Weiter gilt mit geeigneter Konstante $C_2 = C_2(\alpha, \beta, \|a\|)$ a-priori unter Verwendung von (18.15), daß

$$\|p\|_Q \leq \frac{1}{\beta}\|f - Au\|_{V^*} \leq C_2 \left(\|f\|_{V^*} + \|g\|_{Q^*}\right). \tag{18.16}$$

Das so bestimmte Paar (u, p) ist Lösung von **(Q)**.

(3) *Isomorphismus nach (iii):* Zur Übung empfohlen ! □

Bemerkung 18.7. Die Babuška-Brezzi Bedingung (18.11) ist nach dem Satz 18.6 *hinreichend* für die Isomorphieeigenschaft der Abbildung $(u, p) \mapsto (f, g)$ von $V \times Q$ auf $V^* \times Q^*$. Man kann zeigen, daß sie hierfür auch *notwendig* ist. (Zur Übung empfohlen !) □

18.3 Approximation durch Penalty-Regularisierung

Zur praktischen Lösung der Probleme **(Q)** bzw. **(P)** erweist sich oft eine *Regularisierung* als sinnvoll. Mit einem Parameter $\lambda > 0$ sei

$$c : Q \times Q \to \mathbf{R}$$

eine symmetrische und stetige Bilinearform mit der Norm

$$\|c\| := \sup_{p,q \in Q\setminus\{0\}} \frac{c(p,q)}{\|p\|_Q \|q\|_Q}.$$

Dann existiert eindeutig ein Operator $C \in \mathcal{L}(Q, Q^*)$ mit

$$\langle Cp, q\rangle = c(p,q) \quad \forall p, q \in Q.$$

Ferner sei $c(\cdot,\cdot)$ $Q-$elliptisch, d.h,

$$\exists \gamma > 0 : \quad c(q,q) \geq \gamma \|q\|_Q^2 \quad \forall q \in Q.$$

Das regularisierte Problem zu **(Q)** ist: Finde $\hat{u}^\lambda := (u^\lambda, p^\lambda) \in X := V \times Q$, so daß $(\mathbf{Q}^\lambda)$

$$a(u^\lambda, v) + b(v, p^\lambda) = \langle f, v\rangle, \quad \forall v \in V \tag{18.17}$$

$$-\lambda c(p^\lambda, q) + b(u^\lambda, q) = \langle g, q\rangle, \quad \forall q \in Q. \tag{18.18}$$

Wegen der $Q-$Elliptizität von $c(\cdot,\cdot)$ existiert der inverse Operator C^{-1}. Damit ist die regularisierte Nebenbedingung (18.18) äquivalent zu

$$p^\lambda = \frac{1}{\lambda} C^{-1}\left(Bu^\lambda - g\right).$$

Das Problem **(P)** wird somit abgeändert zum Problem $(\mathbf{P}^\lambda)$: Finde $u^\lambda \in V,\ \lambda > 0$:

$$a(u^\lambda, v) + \frac{1}{\lambda}\langle Bv, C^{-1}Bu^\lambda\rangle = \langle f, v\rangle + \frac{1}{\lambda}\langle Bv, C^{-1}g\rangle.\ \forall v \in V. \tag{18.19}$$

Satz 18.8. *(Lösbarkeit von $(\mathbf{Q}^\lambda)$)*
Für die stetige Bilinearform $b : V \times Q \to \mathbf{R}$ gelte die Babuška-Brezzi Bedingung (18.11). Die stetige, symmetrische Bilinearform $c : Q \times Q \to \mathbf{R}$ sei $Q-$elliptisch. Ferner existiere eine Konstante $\alpha > 0$ mit

$$a(v,v) + \langle Bv, C^{-1}Bv\rangle \geq \alpha \|v\|_V^2 \quad \forall v \in V. \tag{18.20}$$

Dann gelten folgende Aussagen:

(i) *Es existieren jeweils eine und nur eine Lösung $\hat{u} =: (u,p)$ bzw. $\hat{u}^\lambda := (u^\lambda, p^\lambda)$ der Probleme (**Q**) bzw. ($\mathbf{Q}^\lambda$), falls $0 < \lambda \leq 1$.*

(ii) *Für hinreichend kleine Werte* $0 < \lambda \ll 1$ *gilt die Abschätzung*

$$\|u^\lambda - u\|_V + \|p^\lambda - p\|_Q \le K\lambda \left(\|f\|_{V^*} + \|g\|_{Q^*}\right). \tag{18.21}$$

Beweis: *zu (i):*

(1) Wegen Voraussetzung (18.20) ist die Bilinearform in $(\mathbf{P}^\lambda)$ $V-$elliptisch für $0 < \lambda \le 1$, d.h. nach der Lax-Milgram Theorie existiert eine und nur eine Lösung u^λ von $(\mathbf{P}^\lambda)$.

(2) Durch die Beziehung

$$p^\lambda = \frac{1}{\lambda} C^{-1}(Bu^\lambda - g)$$

ist p^λ eindeutig festgelegt. Das Paar (u^λ, p^λ) ist damit die eindeutige Lösung von $(\mathbf{Q}^\lambda)$. Die Lösbarkeit von **(Q)** für $\lambda = 0$ folgt bereits aus Satz 17.6.

zu (ii):

(3) Differenzbildung von $(\mathbf{Q}^\lambda)$ und $(\mathbf{Q})$ ergibt die folgende Fehlergleichungen

$$\begin{aligned} a(u - u^\lambda, v) + b(v, p - p^\lambda) &= 0, \quad \forall v \in V & (18.22)\\ b(u - u^\lambda, q) \;+\; \lambda c(p^\lambda, q) &= 0, \quad \forall q \in Q. & (18.23) \end{aligned}$$

Mittels Babuška-Brezzi Bedingung, Fehlergleichung (18.22) und wegen der Stetigkeit von $a(\cdot,\cdot)$ folgt

$$\beta\|p - p^\lambda\|_Q \le \sup_{v\in V\setminus\{0\}} \frac{b(v, p - p^\lambda)}{\|v\|_V} \le \sup_{v\in V\setminus\{0\}} \frac{|a(u - u^\lambda, v)|}{\|v\|_V} \le \|a\|\|u - u^\lambda\|_V,$$

d.h.

$$\|p - p^\lambda\|_Q \le \frac{\|a\|}{\beta}\|u - u^\lambda\|_V. \tag{18.24}$$

(4) Mit der Festsetzung $v := u - u^\lambda$ bzw. $q := p - p^\lambda$ in den Fehlergleichungen (18.22), (18.23) erhält man unter Beachtung von Schritt (3), daß

$$\begin{aligned} a(u - u^\lambda, u - u^\lambda) &= -b(u - u^\lambda, p - p^\lambda) = \lambda c(p^\lambda, p - p^\lambda)\\ &= \lambda c(p, p - p^\lambda) - \underbrace{\lambda c(p - p^\lambda, p - p^\lambda)}_{\ge 0}\\ &\le \lambda c(p, p - p^\lambda)\\ &\le \lambda \frac{\|a\|\|c\|}{\beta}\|p\|_Q\|u - u^\lambda\|_V. \end{aligned} \tag{18.25}$$

(5) Weiterhin ist $B(u-u^\lambda) = g-Bu^\lambda = -\lambda Cp^\lambda$, d.h. wegen der Stetigkeit von $c(\cdot,\cdot)$ und unter Beachtung von Schritt (3)

$$\begin{aligned} \langle B(u - u^\lambda), C^{-1}B(u - u^\lambda)\rangle &= \lambda^2 c(p^\lambda, p^\lambda) \;\le\; \lambda^2\|c\|\|p^\lambda\|_Q^2 \;\le\; \lambda^2\|c\| \left(\|p\|_Q + \|p - p^\lambda\|_Q\right)^2\\ &\le \lambda^2\|c\| \sup\{1; \frac{\|a\|^2}{\beta^2}\} \left(\|p\|_Q + \|u - u^\lambda\|_V\right)^2 \end{aligned} \tag{18.26}$$

(6) Nach Voraussetzung (18.20) und (18.25), (18.26) ergibt sich schließlich

$$\begin{aligned} \alpha\|u - u^\lambda\|_V^2 &\le a(u - u^\lambda, u - u^\lambda) + \langle B(u - u^\lambda, C^{-1}B(u - u^\lambda)\rangle\\ &\le \underbrace{\lambda \frac{\|c\|\|a\|}{\beta}}_{=:K_2} \|p\|_Q\|u - u^\lambda\|_V + \underbrace{\lambda^2 \|c\| \sup\{1; \frac{\|a\|^2}{\beta^2}\}}_{=:K_1} \left(\|p\|_Q + \|u - u^\lambda\|_V\right)^2. \end{aligned}$$

Mit $Z := \|u - u^\lambda\|_V$ entsteht eine Ungleichung der Form

$$\alpha Z^2 \le \lambda^2 K_1 \left(\|p\|_Q + Z\right)^2 + \lambda K_2\|p\|_Q Z. \tag{18.27}$$

Für hinreichend kleine Werte von λ gelangt man aus (17.27) zu einer Ungleichung

$$\|u - u^\lambda\|_V \leq K_3 \lambda \|p\|_Q,$$

d.h. zusammen mit (18.24) auch

$$\|p - p^\lambda\|_Q \leq \frac{\|a\|}{\beta} \|u - u^\lambda\|_V \leq K_4 \lambda \|p\|_Q.$$

Das ist die gesuchte Fehlerabschätzung. □

Bemerkung 18.9. Die Aussage (i) des Satzes 18.8 gilt auch ohne die Babuška-Brezzi Bedingung (18.11), d.h. der Term $-\lambda c(p^\lambda, q)$ wirkt als *Regularisierungsterm* in $(\mathbf{Q}^\lambda)$. □

18.4 Iterationsverfahren

Wir untersuchen nun zwei typische iterative Verfahren zur entkoppelten Lösung des gemischten Problems **(Q)**. Diese sind im Prinzip geeignete Varianten des Regularisierungsverfahrens aus dem vorhergehenden Abschnitt. Diese Entkopplungsstrategie ist in Anwendungen auf die numerische Lösung (vor allem bei inkompressiblen Strömungsproblemen) von großer Bedeutung, da die Lösung des nichtentkoppelten gemischten Problems ggf. erhebliche Probleme hervorruft.

(i) *Uzawa-Verfahren*

Mit Parametern $\rho_m > 0$ lautet der Algorithmus $(\mathbf{Q}^\lambda_\mathbf{m})$ zur entkoppelten Lösung von $(\mathbf{Q})$ für $m \in \mathbf{N}_0$:

$$\text{Finde } (u^{(m+1)}, p^{(m+1)}) \in V \times Q, \text{ so daß } \forall \hat{v} = (v, q) \in V \times Q :$$

$$\begin{aligned} a(u^{(m+1)}, v) + b(v, p^{(m)}) &= \langle f, v \rangle \\ -\frac{c(p^{(m+1)}) - p^{(m)}), q)}{\rho_m} + b(u^{(m+1)}, q) &= \langle g, q \rangle. \end{aligned}$$

Falls die Bilinearformen $a(\cdot,\cdot)$ $W-$elliptisch sowie $c(\cdot,\cdot)$ $Q-$elliptisch sind, so ist auch die Lösungsfolge $(u^{(m+1)}, p^{(m+1)})_m$ von $(\mathbf{Q}^\lambda_\mathbf{m})$ eindeutig bestimmt.

Die Idee des Verfahrens ist, daß für $m \to \infty$ die Lösungsfolge $(p^{(m)})_m$ stationär wird, d.h. der regularisierende Term in der zweiten Gleichung verschwindet. Man kann diesen Term auch als zeitliche Diskretisierung eines instationären Terms der Form

$$-c(\frac{\partial p(t_m)}{\partial t}, q) \approx -\frac{c(p^{(m+1)}) - p^{(m)}), q)}{\rho_m}$$

mit der Zeitschrittweite ρ_m ansehen. Diese Idee findet man in zahlreichen Varianten sogenannter *Projektionsmethoden* oder bei *Stabilisierungstechniken* für inkompressible Strömungsprobleme wieder. Man bezeichnet sie auch als *Pseudokompressibilitäts–Methoden.* Die Wahl der Parameter ρ_m ist jedoch nicht trivial. Man vergleiche hierzu z.B. C. Großmann, H.G. Roos [10], Kap. 4.7.4.

(ii) *Augmented Lagrange Algorithmus*

Dieses Verfahren ist eine stabilisierte Variante des Uzawa-Verfahrens durch Addition eines Terms

$$r\langle Bv, C^{-1}(Bv - g)\rangle$$

zum Problem $(\mathbf{Q}^\lambda_\mathbf{m})$. Die Iterationsvorschrift $(\mathbf{Q}^{\lambda,\mathbf{r}}_\mathbf{m})$ lautet für $m \in \mathbf{N_0}$:

$$\text{Finde } (u^{(m+1)}, p^{(m+1)}) \in V \times Q, \text{ so daß } \forall \hat{v} = (v, q) \in V \times Q$$

$$\begin{aligned} a(u^{(m+1)},v) + r\langle Bv, C^{-1}Bu^{(m+1)}\rangle + b(v,p^{(m)}) &= \langle f,v\rangle + r\langle Bv, C^{-1}g\rangle \\ -\frac{c(p^{(m+1)}) - p^{(m)}),q)}{\rho_m} + b(u^{(m+1)},q) &= \langle g,q\rangle. \end{aligned}$$

Hinsichtlich der Wohldefiniertheit und Konvergenz des Verfahrens hat man folgende Aussage:

Satz 18.10. *(Konvergenz des Augmented Lagrange Algorithmus)*
Seien folgende Voraussetzung erfüllt:

- *Für die stetige Bilinearform $b : V \times Q \to \mathbf{R}$ gilt die Babuška–Brezzi Bedingung (17.11).*
- *Die stetige Bilinearform $a : V \times V \to \mathbf{R}$ ist nichtnegativ, d.h. $a(v,v) \geq 0, \qquad \forall v \in V$.*
- *Die symmetrische, stetige Bilinearform $c : Q \times Q \to \mathbf{R}$ ist $Q-$elliptisch, d.h.*

$$\exists \gamma > 0 : \quad c(q,q) \geq \gamma \|q\|_Q^2, \qquad \forall q \in Q.$$

- *Zusätzlich gelten folgende Abschätzungen*

$$\begin{aligned} \exists \alpha > 0 : \quad a(v,v) + \langle Bv, C^{-1}Bv\rangle &\geq \alpha \|v\|_V^2 \quad \forall v \in V \\ \exists \delta(r) > 0 : \quad a(v,v) + r\langle Bv, C^{-1}Bv\rangle &\geq \delta(r)\|Bv\|_{Q^*}^2 \quad \forall v \in V. \end{aligned}$$

Dann ist durch das Iterationsverfahren $(\mathbf{Q}_\mathbf{m}^{\lambda,\mathbf{r}})$ eindeutig eine Lösungsfolge $(u^{(m+1)}, p^{(m+1)})_m$ in $V \times Q$ bestimmt. Gilt außerdem

$$0 < \inf_m \rho_m \leq \sup_m \rho_m < 2\gamma\delta(r),$$

so konvergiert das Verfahren gemäß

$$\lim_{m\to\infty} \left(\|u - u^{(m)}\|_V + \|p - p^{(m)}\|_Q \right) = 0$$

gegen die Lösung des gemischten Problems **(Q)**.

18.5 Numerische Approximation

Wir betrachten jetzt endlichdimensionale Teilräume

$$V_h \subset V, \qquad Q_h \subset Q$$

sowie die entsprechenden Dualräume V_h^* bzw. Q_h^* mit den Normen $\|\cdot\|_{V^*}$ bzw. $\|\cdot\|_{Q^*}$ Das zum *kontinuierlichen* gemischten Problem gehörende *diskrete gemischte Variationsproblem* $(\mathbf{Q_h})$ ist:

$$\text{Finde } \hat{u}_h := (u_h, p_h) \in X_h := V_h \times Q_h, \text{ so daß}$$

$$\begin{aligned} a(u_h,v) + b(v,p_h) &= \langle f,v\rangle, \quad \forall v \in V \\ b(u_h,q) &= \langle g,q\rangle, \forall q \in Q. \end{aligned}$$

Wie im kontinuierlichen Fall findet man das zu **(Q)** äquivalente *System von Operatorgleichungen* $(\mathbf{Q}')$ mit Operatoren

$$A_h \in \mathcal{L}(V_h, V_h^*), \quad B_h \in \mathcal{L}(V_h, Q_h^*), \quad B_h^* \in \mathcal{L}(Q_h, V_h^*).$$

Wir wollen jetzt das zum kontinuierlichen Problem **(P)** analoge diskrete Problem formulieren. Dazu bezeichnen wir

$$\begin{aligned} W_h(g) &:= \{v_h \in V_h : \ b(v_h,q) = \langle g,q\rangle \ \ \forall q \in Q_h\}, \\ W_h &:= W_h(0) \equiv \mathcal{N}(B_h) \cap V_h. \end{aligned}$$

Das gemischte Problem $(\mathbf{Q_h})$ lautet dann als Problem mit eingearbeiteter Nebenbedingung $(\mathbf{P_h})$:

$$\text{Finde } u_h \in W_h(g) \text{ so daß}: \qquad a(u_h, v) = \langle f, v\rangle \quad \forall v \in W_h \equiv W_h(0).$$

Bemerkung 18.11: Da $Q_h \subset Q$ echter Unterraum ist, gilt im allgemeinen Fall

$$W_h(g) \not\subseteq W(g), \qquad W_h \not\subseteq W.$$

Daher kann der Existenzsatz 18.6 über das kontinuierliche gemischte Problem nicht einfach auf den diskreten gemischten Fall übertragen werden. □

Offenbar ist wieder für jede Lösung $(u_h, p_h) \in V_h \times Q_h$ von $(\mathbf{Q_h})$ auch $u_h \in W_h(g)$ Lösung von $(\mathbf{P_h})$. Zu klären ist, ob zu einer Lösung $u_h \in W_h(g)$ von $(\mathbf{P_h})$ ein Element $p_h \in Q_h$ bestimmt werden kann, so daß $(u_h, p_h) \in V_h \times Q_h$ Lösung von $(\mathbf{Q_h})$ ist.

Das entsprechende Lösbarkeitsresultat unter Einschluß einer Fehlerabschätzung gibt der folgende Satz.

Satz 18.12.a *Sei $W_h(g) \neq \emptyset$. Ferner sei die Bilinearform $a(\cdot,\cdot)$ W_h-elliptisch, d.h.*

$$\exists \alpha^* > 0: \quad a(v,v) \geq \alpha^* \|v\|_V^2 \quad \forall v \in W_h.$$

Dann gelten folgende Aussagen:

(i) *Es existiert eine und nur eine Lösung $u_h \in W_h(g)$ des Problems $(\mathbf{P_h})$.*

(ii) *Mit der Lösungen $\hat{u} = (u,p) \in V \times Q$ von $(\mathbf{Q})$ gilt die Fehlerabschätzung*

$$\|u - u_h\|_V \leq C_1(\alpha^*, \|a\|, \|b\|) \left(\inf_{v \in W_h(g)} \|u - v\|_V + \inf_{q \in Q_h} \|p - q\|_Q \right). \tag{18.28}$$

Dieses Resultat betrifft zunächst nur Problem $(\mathbf{P_h})$. Unbefriedigend sind ferner die Voraussetzung $W_h(g) \neq \emptyset$ und die Tatsache, daß die Interpolationsabschätzung auf der rechten Seite von (17.28) nicht in V_h sondern in der Menge $W_h(g)$ vorzunehmen ist. Letzteres ist praktisch kaum handhabbar.

Satz 18.12.b *Neben der W_h-Elliptizität von $a(\cdot,\cdot)$ gelte die* diskrete Babuška-Brezzi Bedingung, *d.h. es gibt eine von $X_h := V_h \times Q_h$ unabbhängige Konstante $\beta^* > 0$ so, daß*

$$\sup_{v \in V_h} \frac{b(v,q)}{\|v\|_V} \geq \beta^* \|q\|_Q, \quad \forall q \in Q_h. \tag{18.29}$$

Dann gelten folgende Aussagen:

(iii) *Es ist $W_h(g) \neq \emptyset$. (Damit gilt insbesondere die Aussage von Satz 18.12.a.)*

(iv) *Es gibt ein und nur ein Element $p_h \in Q_h$, so daß mit der Lösung u_h von $(\mathbf{P_h})$ das Paar (u_h, p_h) Lösung des diskreten gemischten Problems $(\mathbf{Q_h})$ ist.*

(v) *Es gilt die Fehlerabschätzung*

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq C_2(\alpha^*, \beta^* \|a\|, \|b\|) \left(\inf_{v \in V_h} \|u - v\|_V + \inf_{q \in Q_h} \|p - q\|_Q \right). \tag{18.30}$$

Beweis von Satz 18.12: (1) *Existenz und Eindeutigkeit von $(\mathbf{P_h})$:* $\Longrightarrow (i)$
Wegen $W_h(g) \neq \emptyset$ existiert ein Element $u_h^0 \in W_h(g)$. Dann hat das Hilfsproblem

$$\text{Finde } z_h \in W_h: \quad a(z_h, v_h) = \langle f, v_h\rangle - a(u_h^0, v_h) \qquad \forall v_h \in W_h$$

nach der Lax-Milgram Theorie eine und nur eine Lösung. Damit ist $u_h := z_h + u_h^0 \in W_h(g)$ die eindeutig bestimmte Lösung von $(\mathbf{P_h})$.

(2) *Fehlerabschätzung I:* $\Longrightarrow (ii)$
Für beliebiges $w_h \in W_h(g)$ ist $v_h := u_h - w_h \in W_h$. Dies führt wegen **(Q)** sowie wegen $v_h \in W_h$ auf

$$\begin{aligned} a(v_h, v_h) &= \langle f, v_h\rangle - a(w_h, v_h) \\ &= a(u, v_h) + b(v_h, p) - a(w_h, v_h) \\ &= a(u - w_h, v_h) + b(v_h, p - q_h) \quad \forall q_h \in Q_h. \end{aligned}$$

Die $V-$Elliptizität und Stetigkeit von $a(\cdot,\cdot)$ liefern

$$\alpha^* \|v_h\|_V^2 \le a(v_h, v_h) \le \|a\| \|u - w_h\|_V \|v_h\|_V + \|b\| \|v_h\|_V \|p - q_h\|_Q,$$

d.h.

$$\|v_h\| \equiv \|u_h - w_h\|_V \le \frac{1}{\alpha^*} \left(\|a\| \|u - w_h\|_V + \|b\| \|p - q_h\|_Q \right).$$

Die Dreiecksungleichung ergibt dann

$$\begin{aligned} \|u - u_h\|_V &\le \|u - w_h\|_V + \|u_h - w_h\|_V \\ &\le \left(1 + \frac{\|a\|}{\alpha^*}\right) \|u - w_h\|_V + \frac{\|b\|}{\alpha^*} \|p - q_h\|_Q \end{aligned}$$

und damit Aussage (ii).
(3) *Nachweis von (iii), (iv)*: Die Aussagen beweist man analog zum Beweis des Satzes 17.6 (vgl. auch Girault/ Raviart [8], Th. II.1.1).
(4) *Nebenrechnung zur Vorbereitung von (5):* Wir zeigen die Aussage

$$\inf_{w_h \in W_h(g)} \|u - w_h\|_V \le \left(1 + \frac{\|b\|}{\beta^*}\right) \inf_{w \in V_h} \|u - w\|_V. \tag{18.31}$$

Sei $v_h \in V_h$ beliebiges Element. Dann existiert ein und nur ein Element $z_h \in W_h^\perp$ mit $B_h z_h = B_h(u_h - v_h) \in Q_h^*$ bzw. von

$$b(z_h, q_h) = b(u - v_h, q_h) \qquad \forall q_h \in Q_h,$$

denn bei Modifizierung der Aussage von Lemma 18.5 (iii) ist $B_h :\ W_h^\perp \to Q_h^*$ Isomorphismus. Ferner gilt die Abschätzung

$$\beta^* \|z_h\|_V \le \|B_h(u - v_h)\|_{Q^*} \le \|b\| \|u - v_h\|_V.$$

Mit $w_h := z_h + v_h$ folgt

$$\begin{aligned} b(w_h, q_h) &= b(z_h, q_h) + b(v_h, q_h) \\ &= b(u - v_h, q_h) + b(v_h, q_h) \\ &= b(u, q_h) \ = \ \langle g, q_h\rangle \qquad \forall q_h \in Q_h, \end{aligned}$$

also ist $w_h \in W_h(g)$. Ferner ist nach Dreiecksungleichung

$$\|u - w_h\|_V \le \|u - v_h\|_V + \|z_h\|_V \le \left(1 + \frac{\|b\|}{\beta^*}\right) \|u - v_h\|_V.$$

Daraus folgt wegen $v_h \in V_h$ beliebig die gesuchte Aussage (18.31).
(5) *Fehlerabschätzung II:* $\Longrightarrow (v)$
Unter Beachtung der Probleme **(Q)** und $(\mathbf{Q_h})$ ist

$$\begin{aligned} b(v_h, p_h - q_h) &= b(v_h, p - q_h) + b(v_h, p_h - p) \qquad \forall q_h \in Q_h \\ &= b(v_h, p - q_h) + a(u - u_h, v_h), \qquad \forall v_h \in V_h. \end{aligned}$$

Mittels diskreter Babuška-Brezzi Bedingung sowie Stetigkeit von $a(\cdot,\cdot)$ bzw. $b(\cdot,\cdot)$ folgern wir

$$\begin{aligned}\beta^*\|p_h - q_h\|_Q &\leq \sup_{v_h \in V_h \setminus \{0\}} \frac{a(u-u_h, v_h) + b(v_h, p-q_h)}{\|v_h\|_V} \\ &\leq \|a\|\|u-u_h\|_V + \|b\|\|p-q_h\|_Q.\end{aligned}$$

Über die Dreiecksungleichung ergibt sich

$$\begin{aligned}\|p-p_h\|_Q &\leq \|p-q_h\|_Q + \|p_h - q_h\|_Q \\ &\leq \frac{\|a\|}{\beta^*}\|u-u_h\|_V + \left(1 + \frac{\|b\|}{\beta^*}\right)\|p-q_h\|_Q. \end{aligned} \tag{18.32}$$

Durch Zusammenfassung der bereits bewiesenen Aussage (ii) des Satzes sowie von (18.31) und (18.32) erhalten wir die gesuchte Aussage (v). □

Bemerkung 18.13. Die Erfüllung der diskreten Babuška-Brezzi Bedingung stellt eine Kompatabilitätsforderung zwischen den Räumen V_h und Q_h dar, die in der Regel sorgfältige Untersuchungen erfordert. Dies gilt insbesondere bei der Anwendung auf inkompressible Strömungsprobleme. Im Fall der Behandlung inhomogener Dirichlet-Bedingungen als Nebenbedingung in schwacher Form, vgl. dazu Beispiel 18.2, kann man jedoch folgendes zeigen:

Verwendet man bei exakter Triangulation des Gebiets Ω einen Finite-Elemente-Unterraum V_h mit global stetigen Lagrange-Elementen der Ordnung $k \in \mathbf{N}$ und wählt als Raum Q_h den Raum von Finite-Elemente-Funktionen, die bei Einschränkung von V_h auf den Rand $\partial\Omega$ entstehen, so ist die diskrete Babuška-Brezzi Bedingung erfüllt. □

Bemerkung 18.14. Die für kontinuierliche gemischte Probleme in den Abschnitten zuvor besprochenen Varianten der Regularisierung und iterativen Entkopplung sind auch insbesondere im diskreten Fall relevant. Sie wurden hier lediglich aus zeitlichen Gründen nicht gesondert besprochen. □

Literaturverzeichnis

[1] H. W. Alt: *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*, Springer-Lehrbuch. Berlin-Heidelberg-New York 1999.

[2] . W. Bangerth, R. Rannacher: *Adaptive finite element methods for differential equations.* Birkhäuser, Basel 2003

[3] C. Bardos: *Problemes aux limites pour les equations aux derivees partielles du premier ordre a coefficients reells; Theoremes d' approximation; Application a l' equation de transport*, Ann. Sci. Ec. Norm. Sup. 4 (1970) 3, 185-233)

[4] S. C. Brenner, L.R. Scott: *The Mathematical Theory of Finite Elements*, Springer-Verlag, Berlin - Heidelberg - New York 2002

[5] D. Braess: *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elasitizitätstheorie*, Springer-Verlag, Berlin - Heidelberg - New York 2003,

[6] L. C. Evans: *Partial Differential Equations*, Graduate Studies in Mathematics, Vol. 19, AMS, Providence, Rhode Island 1998.

[7] D. Gilbarg, N.S. Trudinger: *Elliptic partial differential equations of second order,* Springer-Verlag, Berlin - Heidelberg - New York 1998,

[8] V. Girault, P. A. Raviart: *Finite Element Methods for Navier-Stokes Equations*, Springer 1986

[9] H. Goering, , H.G. Roos, L. Tobiska: *Finite-Element-Methoden. Eine Einführung.* Akademie-Verlag, Berlin 1993

[10] Ch. Grossmann, H.G. Roos: *Numerische Behandlung partieller Differentialgleichungen,* Teubner, Stuttgart 2005.

[11] W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen.* Teubner, Stuttgart 1986

[12] M. Hanke-Bourgeois: Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens, Teubner 2006.

[13] Kadlec, Czech. Math. J. 14 (1964), 386–393.

[14] R. Kress: *Lineare Integralgleichungen*, Springer, Berlin-Heidelberg 1999

[15] O.A. Ladyshenskaja, N.N. Uralceva: *Linear and quasilinear differential equations of elliptic type*, Academic Press, New York 1968

[16] U. Nävert: *A finite element method for convection-diffusion problems*, Ph.D. Thesis, Chalmers Univ. Göteborg 1982

[17] O.A. Oleinik, R.U. Radkevic: *Second order equations with nonnegative characteristic form*, Amer. Math. Soc., New York 1973

[18] A. Quarteroni, A. Valli: *Numerical Approximation of Partial Differential Equations*, Springer-Verlag 1997

[19] H.G. Roos, M. Stynes, L. Tobiska: *Numerical Methods for Singularly Perturbed Differential Equations*, Springer 1996.

[20] Y. Saad: *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, 2003.

[21] C. Schwab: *p− and hp-methods*, Oxford Univ. Press 1998.

[22] H. Triebel: *Höhere Analysis*. VdW Berlin 1972,

[23] R. Verfürth: *A review of a-posteriori error estimation and adaptive mesh-refinement techniques*, Teubner-Wiley, Stuttgart 1996.

[24] W. S. Wladimirow: *Gleichungen der Mathematischen Physik*, Verlag der Wissenschaften, Berlin 1972.

[25] E. Zeidler: *Nonlinear Functional Analysis*, Bd. II A, Springer-Verlag, Berlin 1990