Numerik partieller Differentialgleichungen

Norbert Köckler

SS 1998

Durchgesehen bis Kapitel 5 einschliesslich!

Inhaltsverzeichnis

1 Einführung **5**

1.1 Grundprobleme 5

1.1.1 Potentialgleichung 5

1.1.2 Wärmeleitungsgleichung 7

1.1.3 Wellengleichung 10

1.1.4 Qualität der Lösungen der Grundprobleme 11

1.2 PDEs 2. Ordnung; Typeneinteilung 12

1.3 Inkorrekt gestellte Probleme (Ill posed problems) 13

2 Variationsmethoden **15**

2.1 Variationsansatz für ein Dirichletsches RWP 15

2.2 Das Ritz-Verfahren von 1908 22

2.3 Regularität, Monotonie, 3. RWA 24

2.4 Die quasiharmonische 3. RWA im $\mathbb{R}^2$ 26

2.4.1 Die quasiharmonische PDE 26

2.4.2 Anwendung: Wärmeleitung 26

3 Die Methode der finiten Elemente (FEM) **31**

3.1 Geometrische Konstruktion und Ansatzfunktionen 31

3.2 Elementsteifigkeitsmatrix und -vektor 39

3.3 Aufbau der Gesamtsteifigkeitsmatrix 42

3.4 Weitere Elemente und Ansatzfunktionen 47

3.4.1 Bilineare Ansätze auf Vierecken 47

3.4.2 Hermitesche FEM-Ansätze 49

3.4.3 Krummlinig berandete Dreiecke als Elemente 50

3.4.4 Finite Elemente in Polarkoordinaten 54

3.4.5 Dreidimensionale finite Elemente . . . 57
3.5 "Optimale" Numerierung der Knotenpunkte . . . 59
3.5.1 Graphentheoretische Grundlagen . . . 60
3.5.2 Algorithmus von Cuthill/McKee . . . 61
3.5.3 Umgekehrter Cuthill/McKee-Algorithmus . . . 64
3.5.4 Strategien zur Auswahl des Startknotens . . . 65
3.6 Die FEM bei instationären Problemen . . . 66
3.7 Konvergenz und Fehler . . . 68
3.7.1 Lineare FEM im $\mathbb{R}^1$ . . . 70
3.7.2 Fehlerabschätzungen für lineare Dreieckselemente . . . 74
3.7.3 "Variational Crimes" . . . 77

4 Differenzenverfahren 79
4.1 Einführung . . . 79
4.2 Differenzenverfahren für elliptische PDE . . . 81
4.2.1 Eindimensionale Differenzenformeln . . . 81
4.2.2 Zweidimensionale Differenzensterne . . . 81
4.2.3 Neumann - RWA . . . 83
4.2.4 Gekrümmter Rand . . . 84
4.2.5 Allgemeine Forderungen . . . 85
4.3 Differenzenverfahren für parabolische PDE . . . 85
4.3.1 Eine Schar von Verfahren für (4.21) (im $\mathbb{R}^1$) . . . 86
4.3.2 Ein HOPSCOTCH - Verfahren . . . 93
4.4 Stabilitäts-Test bei Anfangswertaufgaben . . . 95
4.4.1 FOURIER-Stabilität (JOHN VON NEUMANN) . . . 95
4.4.2 Die Matrizen-Methode . . . 96
4.5 Inhomogene Anfangswertaufgaben . . . 98
4.6 Parabolische ARWA im $\mathbb{R}^n, n > 1$ . . . 99
4.6.1 Abstrakte Formulierung . . . 99
4.6.2 Konkrete Formulierung . . . 99
4.6.3 Stabilitäts-Test für (4.52) . . . 100

5 Grosse Gleichungssysteme 101
5.1 Einleitung . . . 102
5.2 Relaxationsverfahren (SOR) . . . 104
5.2.1 Erinnerung und Grundlagen . . . 104

5.2.2 Block-Relaxations-Verfahren . . . 106
5.3 Verfahrensvergleich für $-\Delta u = q$ . . . 108
5.4 Vorkonditionierte cg-Verfahren . . . 108
5.4.1 Das Verfahren der konjugierten Gradienten . . . 108
5.4.2 Konvergenzgeschwindigkeit . . . 111
5.4.3 Vorkonditionierung - die Idee . . . 112
5.4.4 PCG-Verfahren . . . 112
5.4.5 Zur Wahl von B . . . 113
5.5 Mehrgittermethoden – Multigrid-Methods (MGM) . . . 114
5.5.1 Die Idee der MGM . . . 114
5.5.2 Zweigittermethode als Grundlage der MGM . . . 116
5.5.3 Zwei einfache Beispiele . . . 117
5.5.4 Definition der Operatoren für das Modellproblem . . . 125
5.5.5 Vollständige Mehrgitterzyklen . . . 126
5.5.6 Mehrgittermethoden für FEM . . . 129

6 Gittererzeugung **131**
6.1 Möglichkeiten der Netzerzeugung . . . 131
6.2 Beispielgitter . . . 141
6.3 Strukturierte Gittererzeugung . . . 145
6.3.1 Transformationsmethoden (r-Typ) . . . 146
6.3.2 Parallelisierung durch Gebietszerlegung . . . 149
6.3.3 Anwendungen und Ergebnisse . . . 151
6.4 Unstruktierte Gittererzeugung . . . 153
6.4.1 Ein Quadtree-Beispiel . . . 153
6.5 Überblick . . . 158

A Tatsachen aus der Funktionalanalysis **175**
A.1 Banachräume . . . 175
A.2 Sobolev-Räume . . . 181

Literatur **187**

Kapitel 1

Einführung

1.1 Grundprobleme

Sei $u(x): \mathbb{R}^n \to \mathbb{R}, x = (x_1, \ldots, x_n)^T$.*

Sei $\Omega \subset \mathbb{R}^n$. Dann heißt die Beziehung

$$F(x, u(x), \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}, \frac{\partial^2 u}{\partial x_1^2}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \ldots) = 0 \qquad \forall x \in \Omega$$

eine partielle Differentialgleichung p-ter Ordnung für $u(x)$, wenn p die Ordnung der höchsten Ableitung ist (PDE). PDEs haben in der Regel einen physikalischen Hintergrund, sie beschreiben fundamentale Erhaltungssätze (Impuls, Ladung, Masse, oft: Energie, ...), und sie lassen sich in drei physikalische Kategorien einteilen:

- Gleichgewichtsprobleme, zeitunabhängig, stationär
- Ausbreitungs- (Fortpflanzungs-)probleme, zeitabhängig, instationär
- Eigenwertprobleme (Beschreibung des charakteristischen Zustands eines Systems, nicht einer speziellen Situation).

Die wesentlichen Bedingungen, Eigenschaften und Unterschiede von linearen PDEs (d.h. F ist linear in den Variablen u, u_{x_1}, ...) lassen sich an drei Grundproblemen 2. Ordnung schildern:

1.1.1 Potentialgleichung

$$\Delta u = f \text{ in } \Omega \subset \mathbb{R}^n \quad \text{(Poissongleichung}^\dagger\text{)} \tag{1.1}$$

mit dem Laplace-Operator‡ $\Delta := \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$.

*Wir befassen uns nicht mit Systemen von PDEs, wo also $u(x): \mathbb{R}^n \to \mathbb{R}^p$.

†Denis Poisson, 1781–1840

‡Pierre Simon Marquis de Laplace, 1749–1827

Physikalische Situation:

Gravitation:	u :	Gravitationspotential.
$(n = 3)$		
	f :	Massendichte.
Elektrostatik:	u :	elektrisches Potential
$(n = 3)$		
	f :	elektrische Ladung.

Die Lösung (für geeignete f) liefert das Newton Potential

$$u(x) = C \int_{\mathbf{R}^3} \frac{f(y)}{|x-y|} dy, x \in \mathbb{R}^3.$$

Dies ist nicht die einzige Lösung. Eindeutigkeit erzielt man nur durch Festlegung von RBen.

Lösungen von $\Delta u = 0$ (Laplacegleichung) nennt man harmonische Funktionen. Sie sind analytisch (C^∞) wie auch alle Lösungen von $\Delta u = f$ mit $f \in C^\infty$.

$\Delta u = f$ ist eine elliptische PDE (s.u.).

Beispiel 1.1.1 Die Wärmeverteilung in einem homogenen Körper Ω bei Kenntnis der Randtemperaturen g wird von der Laplace-Gleichung

$$\left.\begin{array}{rl} \Delta u = 0 & \text{in } \Omega \\ u = g & \text{auf } \partial\Omega \end{array}\right\} \tag{1.2}$$

beschrieben. Dabei ist $u(x)$ Temperatur.

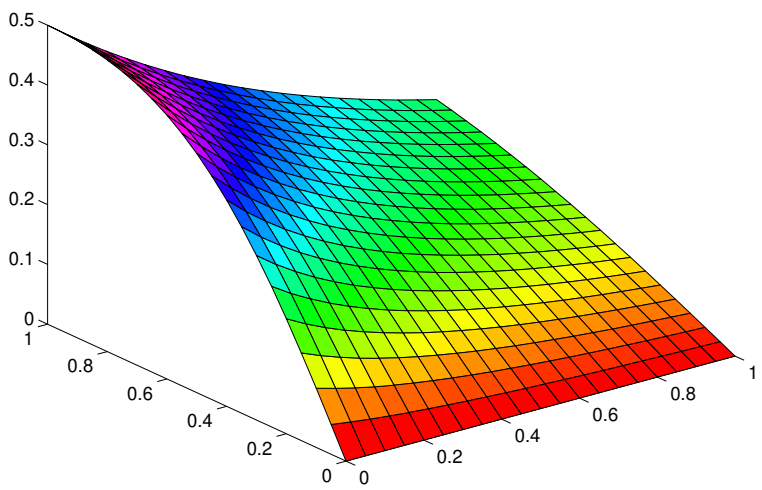
Ein konkretes Beispiel wollen wir exakt lösen. Dabei sei Ω das Einheitsquadrat. Also lösen wir

$$-\Delta u = 0 \text{ in } (0,1) \times (0,1).$$

Als Randbedingungen schreiben wir vor:

$$\begin{array}{lcll} u(x,0) & = & 0 & 0 \le x \le 1 \\ u(x,1) & = & \dfrac{1}{(1+x)^2+1} & 0 \le x \le 1 \\ u(0,y) & = & \dfrac{y}{1+y^2} & 0 \le y \le 1 \\ u(1,y) & = & \dfrac{y}{4+y^2} & 0 \le y \le 1 \\ \text{Exakte Lösung:} & & & \\ u(x,y) & = & \dfrac{y}{(1+x)^2+y^2}, & 0 \le x,y \le 1 \end{array}$$

Für diese Lösung ist $-\Delta u = 0$ leicht nachprüfbar (z.B. mit MAPLE).



Der Lösung ist gut anzusehen, daß Minimum und Maximum auf dem Rand liegen.

1.1.2 Wärmeleitungsgleichung

$$\Delta u(x,t) = \frac{1}{\kappa}\,\frac{\partial u(x,t)}{\partial t} \text{ für } (x,t) \in \Omega = \Omega' \times \mathbb{R}^+ \in \mathbb{R}^{n+1} \tag{1.3}$$

Physikalisch: Wärmeleitung:	u :	Temperatur
	κ :	Materialkonstante: Leitfähigkeit.
Diffusion:	u :	Konzentration (eines Stoffes, Gas, Staub)
	κ :	Diffusionskoeffizient.

Zusätzlich meist AB: $u(x,0) = \varphi(x) \quad \forall x \in \Omega'$

Für $\Omega' = \mathbb{R}^n$ bekommt man dann mit geeigneten Abklingbedingungen im Unendlichen

$$u(x,t) = \frac{1}{(4\pi t)^{n/2}} \int_{\mathbb{R}^n} \varphi(y) e^{-(x-y)^2/4t} dy \quad (x,t) \in \mathbb{R}^{n+1},\ t > 0! \tag{1.4}$$

Bemerkung 1.1.2 Ist $\varphi \geq 0$ und Träger (φ) kompakt,

(d.h. z.B. $\varphi > 0$ auf einem beschränkten Intervall, $\varphi = 0$ sonst)

$\Rightarrow u(x,t) > 0\ \forall x \in \mathbb{R}^n,\ t > 0.$

d.h. die Temperatur breitet sich mit unendlich großer Geschwindigkeit aus. Prozeß ist irreversibel (deshalb $t > 0$), deshalb (s.u.) ist das Umkehrproblem (Rückwärtswärmeleitung) inkorrekt gestellt. Ist $\Omega' \neq \mathbb{R}^n$, so benötigt man statt der Abkling- zusätzliche Randbedingungen, also liegt normalerweise eine ARWA vor:

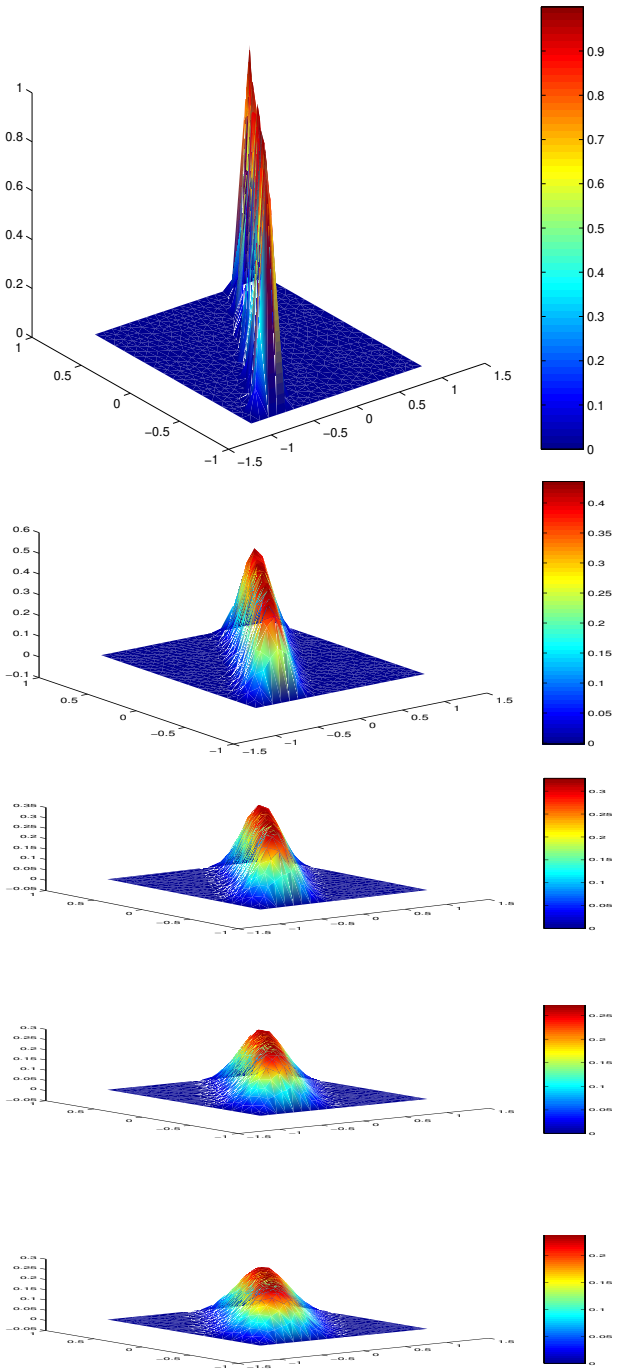
Wärmeleitungsgleichung ist parabolisch (s.u.)

$$\left.\begin{array}{lcll} \Delta u & = & \dfrac{\partial u}{\partial t} & \text{in } \Omega' \times (0,\infty) \\ u(x,0) & = & \varphi & \forall x \in \Omega' \\ u(x,t) & = & \psi & \forall x \in \partial\Omega, t \in (0,\infty) \end{array}\right\} \tag{1.5}$$

Beispiel 1.1.3 $n = 2$, $\Omega' = (-1,1) \times (-0.8, 0.8)$.

$$\begin{array}{rcl} u_t & = & (u_{xx} + u_{yy})/1000 \text{ in } \Omega = \Omega' \times (0,\infty) \\ u & = & 0 \text{ auf } \partial\Omega \\ u(x,y,0) & = & \exp(-100 * (x-y) .* (x-y)) \end{array}$$

Wir simulieren also einen diagonalen Heizstab, der in ein rechteckiges Gebiet gelegt und dann abgeschaltet wird, und beobachten die Wärmeverteilung danach. Typisch ist die schnelle Wärmeverteilung und -glättung. Wir sehen dies in einer Computersimulation (20 Sekunden) und an den folgenden Grafiken, die den Anfangszustand und die Verteilung nach 5, 10, 15 und 20 Sekunden zeigt. Ohne den Faktor 1000 wäre diese Entwicklung tausend mal so schnell!


1
0.8
0.6
0.4
0.2
0
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1
1.5
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.6
0.5
0.4
0.3
0.2
0.1
0
−0.1
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0

1.1.3 Wellengleichung

$$\Delta u = \frac{1}{c^2}\,\frac{\partial^2 u}{\partial t^2},\quad \Omega = \Omega' \times \mathbb{R} \subset \mathbb{R}^{n+1} \tag{1.6}$$

Physikalisch:

Akustik:	u :	Luftdichte
(n=3)	c :	Schallgeschwindigkeit
Elektrodynamik:	u :	elektromagnetische Potentiale (Feldstärke)
(n=3)	c :	Lichtgeschwindigkeit.
Elastizitätstheorie:	$n=1$:	schwingende Seite
	$n=2$:	schwingende Membran

Ist $c > 0$ konstant, so sind die Lösungen ebene Wellen:

$$u(x,t) = f(t - \xi \cdot x) \quad \forall (x,t) \in \mathbb{R}^{n+1}$$

mit $|\xi| = \frac{1}{c}, f : \mathbb{R} \to \mathbb{R} \in C^2$.

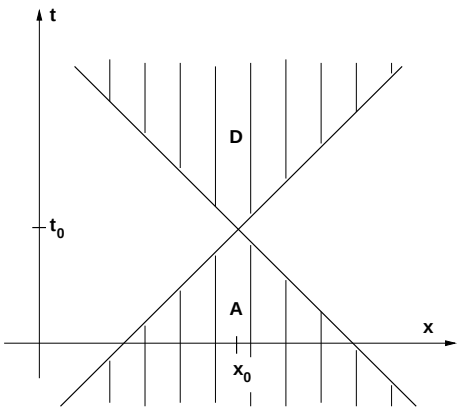
Wellen der Geschwindigkeit $c, u = const$ auf den Ebenen $t - \xi x = const$.

Ist $u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0$ für $|x| \le R$, so ist auch $u(x,t) = 0$ für

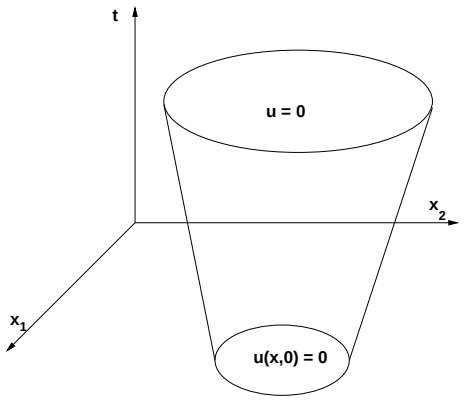
$$|x| \le R + ct,\ t > 0$$

Man bekommt also für jeden Punkt (x_0, t_0) ein Einfluß- und ein Abhängigkeitsgebiet:

Für $n = 1$:



Für $n = 2$:



Der Wert $u(x_0, t_0)$ kann nur Werte $u(x,t)$ mit $(x,t) \in D$ beeinflussen:

$$D(x_0, t_0) = \{(x,t) | t > t_0, |x - x_0| < c|t - t_0)\}$$

Umgekehrt: Abhängigkeitsgebiet

$$A(x_0, t_0) = \{(x,t) | t < t_0, |x - x_0| < c|t - t_0|\}$$

Normalerweise liegt die Wellengleichung als ARWP vor:

$$\Delta u = \frac{\partial^2 u}{\partial t^2} \text{ in } \Omega' \times \mathbb{R}$$

$$\left.\begin{array}{l} u(x,0) = \varphi_0 \\ u_t(x,0) = \varphi_1 \end{array}\right\} x \in \Omega' \qquad (1.7)$$
$$u(x,t) = \psi,\ x \in \partial\Omega,\ t \in \mathbb{R}$$

Beispiel 1.1.4 Wir wollen die Schwingungen einer quadratischen Membran berechnen, die an zwei gegenüberliegenden Seiten fest eingespannt ist, während sie an den beiden anderen frei schwingen kann. Das führt mit zusätzlicher Festlegung der Anfangsbedingungen zu folgendem Problem:

$$\Delta u = \frac{\partial^2 u}{\partial t^2} \text{ in } [-1,\,1]^2 \times \mathbb{R} \text{ mit den ARB:}$$

$$\left.\begin{array}{l} u(x,y,0) = \arctan\left(\cos(\pi/2\,x)\right) \\ u_t(x,y,0) = 3\,\sin(\pi\,x)\,\exp(\sin(\pi/2\,y)) \end{array}\right\} (x,y) \in [-1,\,1]^2$$
$$\left.\begin{array}{l} u(x,y,t) = 0, \text{ für } y = -1 \text{ und } y = 1, \\ \dfrac{\partial u}{\partial \mathfrak{n}} = 0, \text{ für } x = -1 \text{ und } x = 1, \end{array}\right\} t \in \mathbb{R}.$$

Diese Wellengleichung wurde mit MATLAB animiert.

1.1.4 Qualität der Lösungen der Grundprobleme

Potentialgleichung:	Stationäre Zustände Maximal mögliche Regularität.
Wärmeleitung:	Instationäre, irreversible Vorgänge mit unendlicher Ausbreitungsgeschwindigkeit. Lösungen werden mit wachsendem t glatter.
Wellengleichung:	Instationäre, reversible Prozesse mit endlicher Ausbreitungsgeschwindigkeit. Gleichbleibende Glattheit, u.U. sehr geringe Differenzierbarkeit.

Die Normalprobleme (RWP bzw. ARWP) sind korrekt gestellt, insbesondere hängen die Lösungen u stetig von den Daten $f, g, \varphi, \varphi_0, \varphi_1, \psi$ ab.

1.2 PDEs 2. Ordnung; Typeneinteilung

Die eben geschilderten Lösungsqualitäten lassen sich auf größere Klassen von PDEs übertragen. Dazu ist eine Typeneinteilung sinnvoll, die besonders anschaulich geometrisch für PDEs 2. Ordnung zu definieren ist. Sei $n = 2, x \equiv x_1, y \equiv x_2$:

Sei $\Omega \subset \mathbb{R}^2$ und

$a(x, y, u, u_x, u_y),\ b(x, y, u, u_x, u_y),\ c(x, y, u, u_x, u_y),\ f(x, y, u, u_x, u_y)$ reelwertig und stetig.

Gesucht ist $u(x, y) \in C^2(\Omega)$ mit

$$a u_{xx} + 2b u_{xy} + c u_{yy} + f = 0 \tag{1.8}$$

(1.8) ist eine quasilineare PDE

(d.h. in der höchsten Ordnung linear) mit dem Hauptteil

$$a u_{xx} + 2b u_{xy} + c u_{yy} \tag{1.9}$$

Die PDE heißt halblinear, wenn a, b, c nicht von u, u_x, u_y abhängen und

$$f = d(x, y, u) u_x + e(x, y, u) u_y + g(x, y, u) \tag{1.10}$$

und linear, wenn darüberhinaus

$$f = d(x, y) u_x + e(x, y) u_y + g(x, y) u + s(x, y) \tag{1.11}$$

Sei $D := \begin{vmatrix} a & b \\ b & c \end{vmatrix}$

Dann ist der Typ der PDE bez. einer Funktion u in einem Punkt $(x, y) \in \Omega$

hyperbolisch,	falls $D < 0$
parabolisch,	falls $D = 0$
elliptisch,	falls $D > 0$

(Normalerweise hyperbolisch, parabolisch, elliptisch $\forall (x, y) \in \Omega$, aber gemischte Typen möglich, z.B. bei transsonischen Strömungen um Flugzeuge).

Die PDE heißt gleichmäßig hyperbolisch bzw. gleichmäßig elliptisch, falls (unabhängig von einem speziellen u) $\forall (x, y) \in \Omega$ gilt:

$$\begin{aligned} D &\leq -\gamma < 0 \\ \text{bzw. } D &\geq \gamma > 0 \end{aligned} \qquad \gamma = const \tag{1.12}$$

Lineare PDEs mit konstanten Koeffizienten lassen sich immer durch lineare Trafos auf eine Form mit einem der folgenden drei Hauptteile transformieren

$$\begin{aligned} u_{xx} - u_{yy} &\quad \text{hyperbolische Normalform} \\ u_{xx} &\quad \text{parabolische Normalform} \\ u_{xx} + u_{yy} &\quad \text{elliptische Normalform} \end{aligned} \tag{1.13}$$

Die Typisierung ist eine geometrische, da für die quadratische Form

$$p(\xi, \eta) = a\xi^2 + 2b\xi\eta + c\eta^2$$

mit konstanten Koeffizienten a, b, c gerade eine Hyberbel, Parabel oder Ellipse in (ξ, η) beschrieben wird durch $p(\xi, \eta) = const$ in den entsprechenden Fällen.

1.3 Inkorrekt gestellte Probleme (Ill posed problems).

Definition 1.3.1 Hadamard[§] (1952)

Das Problem R, mit Hilfe eines gegebenen Elements $g \in Y$ ein Element $f \in X$ zu bestimmen, heißt korrekt gestellt, wenn

(i) zu jedem $g \in Y$ eine Lösung $f \in X$ existiert;

(ii) die Lösung f eindeutig bestimmt ist;

(iii) die Abhängigkeit der Lösung f von den "Daten" g stetig ist in den Topologien von X und Y (Stabilität).

Bemerkung 1.3.2 Viele schöne Beispiele zu korrekt oder inkorrekt gestellten Problemen bei PDE findet man in [MM78]. Bei der komplexen Vielfalt von PDEs und dem Einfluß von AB und RB auf die Korrektheit sind allgemeine Aussagen schwer.

Faustregel 1: Die Grundaufgaben aus 1.1 sind korrekt gestellt ((1.2), (1.5), (1.7)).

Die Korrektheit hängt auch vom Grundgebiet Ω ab.

Faustregel 2: Ω einfach zusammenhängend mit genügend glattem Rand.
(z.B. C^1 mit endlich vielen Ecken)

Zwei einfache Beispiele im $\mathbb{R}^2$ zeigen, wie ähnlich korrekt und inkorrekt gestellte Probleme sein können.

Beispiel 1.3.3

$$\begin{array}{lllll} \text{PDE:} & \dfrac{\partial u}{\partial y} & = & 0 & \forall x \in \mathbb{R}, y \in (0,T), T > 0. \\ \text{AB:} & u(x,0) & = & \varphi(x) & \forall x \in \mathbb{R}, \varphi \in C^1(\mathbb{R}). \end{array} \tag{1.14}$$

Existenz: $u(x,y) = \varphi(x)$.

Eindeutigkeit: Wegen $u_y = 0$ muß u bezüglich y konstant sein. $\varphi(x)$ hängt nicht von y ab, also auch u.

Stabilität: $\|u(x,y,\varphi) - u(x,y,\tilde{\varphi})\| = \|\varphi - \tilde{\varphi}\|$

$\Rightarrow$ Stetige Abhängigkeit u von Daten φ.

$\Rightarrow$ Korrekt gestellt.

Beispiel 1.3.4

$$\begin{array}{lllll} \text{PDE:} & u_x & = & 0 \text{ in } \mathbb{R} \times (0,T) \\ \text{AB:} & u(x,0) & = & \varphi(x), \varphi \in C^1(\mathbb{R}) \end{array} \tag{1.15}$$

[§]Jacques Hadamard, 1865–1963

In C^1 ist dieses Problem nur lösbar, wenn $\varphi(x) \equiv C$ (wegen $u_x = 0$ und C^1 mit $y \to 0$). Dann gibt es aber ∞ viele Lösungen

$$\begin{aligned} u(x,y) &= C + \psi(y) \\ \text{mit} & \\ \psi \in M &= \{\phi(y) | \phi \in C^1(0,T), \phi(0) = 0\} \end{aligned}$$

Also ist dieses Problem inkorrekt gestellt.

Bemerkung 1.3.5 Für die folgenden Abschnitte ist es hilfreich, einiges über Banach- und Hilbert-, speziell Sobolev-Räume zu wissen. Ich empfehle dazu die Lektüre von Anhang A.

Kapitel 2

Variationsmethoden für elliptische partielle Differentialgleichungen

2.1 Variationsansatz für ein Dirichletsches RWP

2.1.1 Problemklasse: Dirichletsches* RWP (DRWP)

Gegeben: $\Omega \subseteq \mathbb{R}^n$, nicht leer, offen, beschränkt, zusammenhängend, Normalgebiet, $a_{ij} \in C^1(\overline{\Omega}, \mathbb{R}), c, f \in C(\overline{\Omega}, \mathbb{R}), g \in C(\partial\Omega, \mathbb{R})$.

Gesucht: $u \in C^2(\Omega, \mathbb{R}) \cap C(\overline{\Omega}, \mathbb{R})$ mit:

$$\begin{aligned} &\text{PDE:} && Lu = f && \forall x \in \Omega, \\ &\text{RB:} && u = g && \forall x \in \partial\Omega. \end{aligned} \tag{2.1}$$

$$\begin{aligned} \text{mit} \quad Lu &:= -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}(a_{ij}\frac{\partial}{\partial x_j}u) + cu \\ &=: -\sum_{i,j=1}^{n} D_i(a_{ij}D_j u) + cu \\ &=: -\sum_{i,j=1}^{n} \partial_i(a_{ij}\partial_j u) + cu \end{aligned}$$

*Peter Dirichlet, 1805–1859

Anmerkungen:

- (2.1) heißt 1. RWA oder Dirichlet'sche RWA, die Randbedingung Dirichlet'sche RB.
- $\partial\Omega = \overline{\Omega} \setminus \Omega$ ist Rand von Ω.
- ein Normalgebiet ist ein Gebiet, für das die Integralformel von Gauß und damit auch die Greensche Formeln gelten.

Voraussetzung 2.1.2 (i) L sei gleichmäßig elliptisch, d.h. $\exists \lambda > 0$, sodaß

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \lambda \|\xi\|_2^2 \quad \forall x \in \Omega, \xi \in \mathbb{R}^n. \tag{2.2}$$

$\Rightarrow$ Die charakteristische Form

$$\sigma(L, x, \xi) := \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \tag{2.3}$$

von L ist auf Ω positiv definit.

(ii) Die Koeffizientenmatrix (a_{ij}) von L sei symmetrisch,
L heißt dann auch formal selbstadjungiert.

(iii) $c(x) \geq 0 \quad \forall x \in \Omega$

(iv) $g \in C(\partial\Omega, \mathbb{R})$ sei fortsetzbar zu $g \in C(\bar{\Omega}, \mathbb{R})$.

(v) Der Rand $\partial\Omega$ sei "genügend" glatt.

2.1.3 Operatorschreibweise des DRWP

$$\begin{aligned} \text{Seien } M &:= \{u \in C^2(\Omega) \cap C(\overline{\Omega})\} \\ \text{und } D_L &:= \{u \in M \mid u = g \text{ auf } \partial\Omega\}. \end{aligned} \tag{2.4}$$

u ist Lösung des DRWP nach (2.1) $\Leftrightarrow$ $u \in D_L \wedge Lu = f$.

Satz 2.1.4 Symmetrie von L

D_L ist ein reeller Vektorraum, für den wir ein Skalarprodukt $(\cdot, \cdot)$ mit zugehöriger Norm $\|\cdot\|_2$ definieren:

$$\begin{aligned} (v, w) &:= \int_\Omega v(x)w(x)\,\mathrm{d}x \quad \forall v, w \in D_L, \\ \|v\|_2 &:= \sqrt{(v, v)} \quad \forall v \in D_L. \end{aligned} \tag{2.5}$$

Dann gilt: $(v, Lw) = (Lv, w) \quad \forall v, w \in D_L.$ (2.6)

Definition 2.1.5 Sobolev-Skalarprodukt und Norm, Energie-Integral

(1) Wir können auf D_L ein Skalarprodukt $(\cdot\,,\cdot)_1$ mit zugehöriger Norm $\|\cdot\|_1$ definieren, das nach Sergej Lwowitsch Sobolev (1908-1989) benannt ist:

$$\begin{aligned}(v,w)_1 &:= (v,w) + \sum_{i=1}^{n}(D_i v, D_i w) \quad \forall v,w \in D_L,\\ \|v\|_1 &:= \sqrt{(v,v)_1} \quad \forall v \in D_L.\end{aligned} \tag{2.7}$$

(2) Weiter ordnen wir L eine quadratische Form zu, das sog. Energieintegral

$$B(u) = \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij} D_i u D_j u + c u^2 \right) dx \tag{2.8}$$

(Doppelte in Ω enthaltene Energie des physikalischen Systems, das Lu beschreibt).

(3) Entsprechend definieren wir eine symmetrische Bilinearform, das polarisierte Energieintegral.

$$\begin{aligned}[u,v] := B(u,v) &:= \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij} D_i u D_j v + c\, u\, v \right) dx \\ &= B(v,u)\end{aligned} \tag{2.9}$$

und die Energie-"Norm"

$$\|u\|_E := \sqrt{B(u)} = \sqrt{B(u,u)}$$

Satz 2.1.6 Vervollständigungen, Sobolev-Räume, Basen

(1) Die Skalarprodukträume $(M,(\cdot,\cdot)_1)$ und $(D_L,(\cdot,\cdot)_1)$ (siehe (2.4)) lassen sich nach Sätzen der Funktionalanalysis vervollständigen zu vollständigen Skalarprodukträumen H^1 bzw. $(D,(\cdot,\cdot)_1)$ (Hilberträume).

(2) $H^1 = H^1(\Omega)$ ist unter zusätzlichen Glattheitsvoraussetzungen an $\partial\Omega$ ein Sobolev-Raum, für den noch folgende Aussagen gelten:

(a) $D_1, D_2, (\cdot,\cdot), B(\cdot,\cdot)$ lassen sich stetig fortsetzen von D_L auf H^1.

(b) Die Identität (2.6) bleibt auch auf H^1 gültig:
$[v,w] = (Lv,w) \quad \forall v \in D_L\, \forall w \in H^1$.

(3) Für $g = 0$ schreibt man den Sobolev-Raum $D = H_0^1$.

(4) H^1 und H_0^1 sind separable Hilberträume, besitzen also Basen $\{u_i\}_{i\in\mathbb{N}}$, sodaß Elemente $u \in H^1$ bzw. $u \in H_0^1$ eindeutig bestimmte Fourierkoeffizienten $\{\alpha_i\}_{i=1}^{\infty}$ besitzen:

$$u = \sum_{i=1}^{\infty} \alpha_i u_i. \tag{2.10}$$

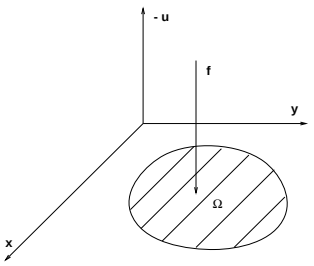
Der Reihengrenzwert ist im Sinne der Norm $\|\cdot\|_1$ gemeint. Aus dieser Definition folgt insbesondere, daß die u_i linear unabhängig sind.

Beispiel 2.1.7 Ideal elastische Membran, an den Rändern eingespannt, belastet.

Gleichgewichtslage:

$$-\Delta u = f \quad \text{in} \quad \Omega \subset \mathbb{R}^2$$

$$u = 0 \quad \text{auf} \quad \partial\Omega$$



Die Gleichgewichtslage kann auch als Variationsprinzip beschrieben werden:

Die gesamte potentielle Energie wird minimal, sie besteht aus 2 Anteilen:
der elastischen Energie

$$\frac{1}{2}B(u) = \int_\Omega (u_x^2 + u_y^2)dxdy$$

und der potentiellen Energie aus der äußeren Kraft mit der Dichte f:

$$\int_\Omega f\, u\, dxdy = (f,u)_0$$

Dabei kann die potentielle Energie nur unter möglichen Auslenkungsformen minimiert werden, d.h. die Randbedingung $u = 0$ auf $\partial\Omega$ muß erfüllt sein.

$\Rightarrow$

$$\min_{H_0^1}\{B(u) - 2(f,u)_0\}$$

Definition 2.1.8 (i) $u \in H^1(\Omega)$ heißt schwache Lösung der 1. RWA (2.1), wenn gilt

$$\begin{aligned} &B(u,\varphi) = (f,\varphi)_0 \quad \forall \varphi \in H_0^1(\Omega) \\ &u - g \in H_0^1(\Omega) \end{aligned} \tag{2.11}$$

(ii) Dirichlet-Integral

$$D_\Omega(u) := \int_\Omega \sum_{i=1}^n |D_i u|^2 dx \tag{2.12}$$

Lemma 2.1.9 Poincaré-Friedrichsche Ungleichung

Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt. Dann gibt es eine Konstante C, sodaß

$$\int_\Omega |v|^2 dx \le C^2\, D_\Omega(v) \qquad \forall v \in H_0^1(\Omega). \tag{2.13}$$

Lemma 2.1.10 Es gelten folgende Abschätzungen:

(i)
$$|B(u,v)| \leq C \|u\|_1 \|v\|_1 \qquad \forall u, v \in H^1(\overline{\Omega}) \tag{2.14}$$
wo C nur von $|a_{ij}|_{\overline{\Omega}}$ und $|c|_{\overline{\Omega}}$ abhängt.

(ii) Da L gleichmäßig elliptisch und $c \geq 0$, gilt
$$B(u) \geq \lambda D_\Omega(u) \qquad \forall u \in C^1(\overline{\Omega}) \tag{2.15}$$

(iii) $\exists m$, sodaß $\forall u \in H_0^1(\Omega)$ gilt:
$$B(u) \geq m \|u\|_1^2 \tag{2.16}$$

(iv) Die Normen $\|\cdot\|_1$ und $\|\cdot\|_E$ sind auf H_0^1 äquivalent:
$$c_1\|u\|_1 \leq \|u\|_E \leq c_2\|u\|_1 \tag{2.17}$$

Beweis: Ü. ■

Lemma 2.1.11 Umformulierung der schwachen 1. RWA: Sei g wieder fortgesetzt auf $\overline{\Omega}$ und
$$w = u - g.$$
Dann ist $w \in H_0^1(\Omega)$ und wir betrachten $\forall \varphi \in H_0^1(\Omega)$:
$$\begin{aligned} &B(w,\varphi) = l(\varphi) \\ &l(\varphi) := (f,\varphi)_0 - B(g,\varphi) \end{aligned} \tag{2.18}$$

Lemma 2.1.12 $l(\varphi)$ ist beschränkt.

Beweis:
$$|(f,\varphi)_0| \leq \|f\|_0 \|\varphi\|_0 \leq \|f\|_0 \|\varphi\|_1$$
und wegen (2.14)
$$\begin{aligned} |B(g,\varphi)| &\leq C\|g\|_1 \|\varphi\|_1 \\ \Rightarrow \qquad |l(\varphi)| &\leq K\|\varphi_1\| \text{ mit } K = \|f\|_0 + C\|g\|_1 \end{aligned}$$

■

Bemerkung 2.1.13 Da (2.18) und (2.11) völlig äquivalent sind, können wir in $H_0^1(\Omega)$ arbeiten, was viele Vorteile hat, z.B. ist die Energie-"Norm" eine Norm.

Wir haben jetzt also das abstrakte Hilbertraum-Problem:

Gegeben sind eine stetige Linearform $l(\varphi)$ und eine beschränkte, symmetrische und positiv definite Bilinearform $B(w,\varphi)$.

Gesucht ist ein $w \in H$ mit $B(w,\varphi) = l(\varphi) \quad \forall \varphi \in H$.

Satz 2.1.14 Darstellungssatz (Riesz†)

Gegeben: Linearform $l(u)$ und Bilinearform $B(u,v)$ auf Hilbertraum H mit:

$$\left.\begin{array}{rcl} |l(u)| & \leq & K\,\|u\| \text{ beschränkt} \\ B(u,v) & = & B(v,u) \text{ symmetrisch} \\ |B(u,v)| & \leq & M\,\|u\|\,\|v\| \text{ gleichmäßig beschränkt} \\ |B(u)| := |B(u,u)| & \geq & m\,\|u\|^2,\ m>0,\ \text{positiv definit.} \end{array}\right\} \tag{2.19}$$

Gesucht: Lösung $u \in H$ von

$$B(u,v) = l(v) \qquad \forall v \in H \tag{2.20}$$

Behauptung:

(i) (2.20) hat genau eine Lösung $u \in H$.

(ii) Die Lösung ist die eindeutig bestimmte Lösung des Variationsproblems

$$\min_{\varphi \in H} A(\varphi), \qquad A(\varphi) := B(\varphi) - 2l(\varphi) \tag{2.21}$$

(iii) Jede Minimalfolge $\{u_p\}$ des Variationsproblems (2.21) konvergiert gegen dessen Lösung:

$$\|u - u_p\| \longrightarrow_{p\to\infty} 0.$$

Dabei heißt $\{u_p\}$ Minimalfolge zu (2.21), wenn

$$\lim_{p\to\infty} A(u_p) = d = \inf_{\varphi \in H} A(\varphi)$$

Beweis:

(iii) Es gilt mit (2.19)

$$A(\varphi) \geq m\|\varphi\|^2 - 2K\|\varphi\| = \left[\sqrt{m}\|\varphi\| - \frac{K}{\sqrt{m}}\right]^2 - \frac{K^2}{m} \geq -\frac{K^2}{m} > -\infty$$

d.h. $A(\varphi)$ ist auf H nach unten beschränkt.

$\Rightarrow \exists\{u_p\}$ mit $\lim_{p\to\infty} A(u_p) = d$

Noch zu zeigen: Sei $\{u_p\}$ Minimalfolge, d.h. $\lim_{p\to\infty} A(u_p) = d$, dann $\|u_p - u\| \to 0$:

Parallelogrammsatz:

$$B(u_p - u) + B(u_p + u) = 2B(u_p) + 2B(u)$$

$\Rightarrow$

$$m\|u_p - u\|^2 \leq B(u_p - u) =$$

†Friedrich Riesz, 1880–1956

$$= 2B(u_p) + 2B(u) - 4B\big(\frac{u_p + u}{2}\big)$$

[Hierzu addieren wir $0 = -l\big(4u_p + 4u - 8\frac{u_p + u}{2}\big)$]

$$-4l(u_p) - 4l(u) + 8l\big(\frac{u_p + u}{2}\big)$$

$$= 2A(u_p) + 2A(u) - 4A\big(\frac{u_p + u}{2}\big)$$

also mit $d \le A(\frac{u_p + u}{2})$

$$m\|u_p - u\|^2 \le \underbrace{2A(u_p) + 2A(u)}_{\to 4d} - 4d \to 0 \text{ mit } p \to \infty.$$

Also ist $\{u_p\}$ Cauchyfolge, und da H vollständig ist, gibt es ein $u \in H$ mit $\lim\limits_{p\to\infty} \|u_p - u\| = 0$. Da $A(\varphi)$ stetig auf H ist, gilt

$$\lim A(u_p) = A(u) \quad \blacksquare$$

(ii) Wir zeigen jetzt mit einem typischen Variationsrechnungs-Ansatz, daß ein solches u auch (2.20) löst: Seien $\varphi \in H, \varepsilon \in \mathbb{R}$ beliebig. Dann ist

$$\begin{aligned} d &\le A(u + \varepsilon\varphi) = B(u + \varepsilon\varphi) - 2l(u + \varepsilon\varphi) \\ &= B(u) - 2l(u) + 2\varepsilon[B(u,\varphi) - l(\varphi)] + \varepsilon^2 B(\varphi) \\ &= A(u) + 2\varepsilon[B(u,\varphi) - l(\varphi)] + \varepsilon^2 B(\varphi) \\ &\Rightarrow \\ 0 &\le 2\varepsilon[B(u,\varphi) - l(\varphi)] + \varepsilon^2 B(\varphi) \end{aligned}$$

Sei zunächst $\varepsilon > 0$: multipliziert mit $\dfrac{1}{2\varepsilon}$ $\quad\to$

$$0 \le B(u,\varphi) - l(\varphi) + \frac{\varepsilon}{2} B(\varphi)$$

$\varepsilon \to 0$ ergibt dann

$$0 \le B(u,\varphi) - l(\varphi) \text{ für beliebiges } \varphi \in H$$

Sei jetzt $\varepsilon < 0$. Dann dreht sich bei Multiplikation mit $\dfrac{1}{2\varepsilon}$ (negativ) die Ungleichung um und man erhält mit $\varepsilon \to 0$

$$0 \ge B(u,\varphi) - l(\varphi) \text{ für beliebiges } \varphi \in H$$

$\Rightarrow$

$$B(u,\varphi) = l(\varphi) \qquad \forall \varphi \in H$$

(Es fehlt für (ii) noch die Eindeutigkeit). $\blacksquare$

(i) Eindeutigkeit der Lösung von (2.20) folgt aus der positiven Definitheit von B:
Sei $B(u_1,\varphi) = l(\varphi) = B(u_2,\varphi) \qquad \forall \varphi \in H$
$\Rightarrow B(u_1 - u_2, \varphi) = 0 \qquad \forall \varphi \in H$
Setze $\varphi := u_1 - u_2$
$\Rightarrow m\|u_1 - u_2\|^2 \leq B(u_1 - u_2) = 0, m > 0$
$\Rightarrow u_1 = u_2$ ■

(ii) Eindeutigkeit der Lösung von (2.21). Sei u eine Lösung, dann ist

$$\begin{aligned} A(u+\varphi) &= A(u) + 2\underbrace{[B(u,\varphi) - l(\varphi)]}_{=0} + B(\varphi) \\ &= d + B(\varphi) \end{aligned}$$

Nun ist $B(\varphi) > 0$ für $\varphi \neq 0$.
$\Rightarrow \quad A(u+\varphi) > d$
$\Rightarrow \quad u$ ist einzige Minimallösung ■

■

2.2 Das Ritz-Verfahren von 1908

2.2.1 Diskretisierung des DRWP[‡]

Wir wollen im folgenden o.B.d.A. $g = 0$ voraussetzen (siehe 2.1.11).

(1) Sei $\{u_n\}_{n\in\mathbb{N}}$ eine Basis von H_0^1. Definiere lineare Teilräume $S^{(n)}$ von H_0^1 als

$$S^{(n)} := \operatorname{span}(u_1, \ldots, u_n), \quad n \in \mathbb{N}.$$

(2) Bestimme für ein festes $n \in \mathbb{N}$ eine Minimalfunktion $u_S^{(n)}$ für A in $S^{(n)}$:

$$A(u_S^{(n)}) = \min_{v \in S^{(n)}} A(v).$$

Nach Satz 2.1.14 löst $u_S^{(n)}$ auch $B(u_S^{(n)}, v) = l(v) \qquad \forall v \in S^{(n)}$.

(3) $u_S^{(n)}$ ist die Lösung des LGS

$$S\beta = t, \tag{2.22}$$

wenn

$$S := ([u_i, u_j])_{i,j=1,\ldots,n}, \quad t := l(u_i)_{i=1,\ldots,n}. \tag{2.23}$$

Wir benutzen S als Symbol der Matrix, da S bei Finite-Elemente-Methoden Steifigkeitsmatrix heißt.

[‡]Walther Ritz, 1878–1909, Schweizer Mathematiker und Physiker, Zürich, Tübingen, Göttingen.

(4) Lösung ist dann $u_S^{(n)} = \sum_{i=1}^n \beta_i u_i$.

(5) Für den Fall $g \neq 0$ wird vorausgesetzt, dass g in H^1 auf ganz $\bar{\Omega}$ fortgesetzt werden kann. Dann wird in (2.23) $t := l(u_i)$ ersetzt durch $t := l(u_i) - B(g, u_i)$, wie in Lemma 2.1.11. Die Lösung $\bar{u}$ des Problems mit inhomogenen Randwerten ist dann

$$\bar{u} = u_S^{(n)} + g.$$

Lemma 2.2.2 $u_S^{(n)}$ ist Bestapproximation von u in $S^{(n)}$ bezüglich $[\cdot\,,\cdot]$:

$$[u_S^{(n)} - u, u_S^{(n)} - u] = \min_{v \in S^{(n)}} [v - u, v - u]. \tag{2.24}$$

Satz 2.2.3 Qualität des Ritz-Verfahrens

$u_S^{(n)}$ ist eine Minimalfolge, d.h. es gilt

$$\lim_{n \to \infty} A(u_S^{(n)}) = \inf_{v \in H^1} A(v) =: d.$$

Beweis: Gegeben sei ein beliebiges $\varepsilon > 0$. Wähle ein $w \in H^1$ mit $0 \leq A(w) - d \leq \varepsilon/2$. A ist stetig in w bezüglich $\| \cdot \|_1$ (Beweis!). Daher gilt:

$$\exists\, \delta > 0\ \forall v \in H^1 \wedge \|v - w\|_1 \leq \delta : |A(v) - A(w)| \leq \frac{\varepsilon}{2}.$$

Da $\{u_i\}$ eine Basis ist, besitzt w genau eine Darstellung

$$w = \sum_{i=1}^{\infty} \alpha_i u_i$$

im Sinne der $\| \cdot \|_1$-Konvergenz, woraus folgt:

$$\exists\, N_0 \in \mathbb{N}\, \forall N \in \mathbb{N} \wedge N \geq N_0 : \|w - \underbrace{\sum_{i=1}^{N} \alpha_i u_i}_{\in S^{(N)}} \|_1 \leq \delta$$

$$\Rightarrow \forall N \in \mathbb{N} \wedge N \geq N_0 :$$

$$0 \leq A(u_S^{(N)}) - d \leq A(\sum_{i=1}^{N} \alpha_i u_i) - d \qquad \text{(Definition von } u_S^{(N)})$$

$$\leq \left| A(\sum_{i=1}^{N} \alpha_i u_i) - A(w) \right| + |A(w) - d| \qquad \text{(Dreiecksungleichung)}$$

$$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \quad \blacksquare$$

Lemma 2.2.4 Die Matrix $S \in \mathbb{R}^{N,N}$ ist positiv definit.

Beweis: S ist symmetrisch, da $B(\cdot,\cdot)$ symmetrisch. Sei $x \neq 0, x \in \mathbb{R}^N$:

$$\begin{aligned} x^T S x &= x^T (B(u_i, u_j)) x \\ &= B\left(\sum \beta_i u_i, \sum \beta_i u_i \right), \text{ wenn } x = (\beta_1, \ldots \beta_N)^T \\ &= B(\varphi, \varphi) > 0 \text{ wenn } \varphi \neq 0 \Leftrightarrow x \neq 0. \quad \blacksquare \end{aligned}$$

Bemerkung 2.2.5 (i) Numerisch sind folgende Schritte zu erledigen:

(1) Umformulierung der 1. RWA in eine mit homogenen RBen, d.h. in $H_0^1(\Omega)$.
(2) Wahl einer geeigneten Basis $\{u_i\}$ in $H_0^1(\Omega)$.
(3) Aufstellen des Gleichungssystems (2.22)/(2.23) (Integrationen!).
(4) Lösen von (2.22).

Sowohl (3) als auch (4) sind meist sehr aufwendig, s.u.!

(ii) Bei unseren Voraussetzungen ist das Ritz-Verfahren mit dem Galerkin-Verfahren identisch.

(iii) Wir wollen noch einige ergänzende theoretische Ergebnisse ohne Beweis zur Abrundung anfügen.

2.3 Regularität, Monotonie, 3. RWA

Definition 2.3.1 Ein Gebiet $\Omega \subset \mathbb{R}^n$ erfüllt eine Kegelbedingung, wenn es zu jedem Punkt $x \in \overline{\Omega}$ einen offenen Kegel $K(x) \subset \Omega$ mit x als Spitze gibt, sodaß $K(x)$ zu einem festen Kegel K mit positivem Öffnungswinkel α kongruent ist.

Satz 2.3.2 Einbettungs-Lemma von Sobolev

Erfüllt Ω eine Kegelbedingung, so gilt

$$H^s(\Omega) \subset C^{s-[n/2]-1}(\overline{\Omega}) \tag{2.25}$$

und

$$|u|_{s-[n/2]-1} \leq const\, \|u\|_s \qquad \forall u \in H^s(\Omega) \tag{2.26}$$

mit $const = const(s, n, K)$ und $|u|_p := \sup\limits_{x \in \overline{\Omega}} \max\limits_{|\alpha| \leq p} |D^\alpha u|$

Satz 2.3.3 Regularitätssatz

Voraussetzung:

Sei $u \in H^1(\Omega)$ schwache Lösung der 1. RWA:

$$\begin{aligned} B(u, \varphi) &= (f, \varphi) \qquad \forall \varphi \in H_0^1(\Omega) \\ u - g &\in H_0^1(\Omega) \end{aligned}$$

Sei $f \in H^s(\Omega)$ und $g \in H^{s+2}(\Omega)$, $s \geq 0$.

Behauptung:

Dann ist $u \in H^{s+2}(\Omega)$ und

$$\|u\|_{s+2} \leq const\, \{\|u\|_0 + \|f\|_s + \|g\|_{s+2}\}, \tag{2.27}$$

mit $const = const(\Omega, L)$. Ist $s \geq [n/2] + 1$, so ist $u \in C^2(\overline{\Omega})$ und damit klassische Lösung von

$$\begin{aligned} Lu &= f \qquad \text{in } \Omega \\ u &= g \qquad \text{auf } \partial\Omega \end{aligned}$$

Satz 2.3.4 Maximum-Minimum-Prinzip

Sei L gleichmäßig elliptisch, $c \geq 0$. Dann gilt:

(i) Ist $u \in C^2(\Omega)$ eine Lösung der Ungleichung $Lu \leq f$ in Ω mit $f \leq 0$.
Dann hat u kein positives (relatives) Maximum in Ω.

(ii) Ist $u \in C^2(\Omega)$ eine Lösung der Ungleichung $Lu \geq f$ in Ω mit $f \geq 0$.
Dann hat u kein negatives (relatives) Minimum in Ω.

(iii) Sei $u \not\equiv const$ eine Lösung von

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \text{ mit } \underline{\text{entweder}}\ f \geq 0\ \underline{\text{oder}}\ f \leq 0 \\ u &= g \quad \text{auf } \partial\Omega \end{aligned}$$

Dann nimmt u negative Minima bzw. positive Maxima nur auf dem Rand $\partial\Omega$ an.

Lemma 2.3.5 Monotonieprinzip

(i) Aus $f(x) \geq 0$ in $\overline{\Omega}$ und $g(x) \geq 0$ auf $\partial\Omega$ folgt $u(x) \geq 0 \qquad \forall x \in \overline{\Omega}$

(ii) Seien $u_1, u_2 \in C^0(\overline{\Omega}) \cap C^2(\Omega)$ Lösungen von $Lu = f$ mit $u = g_1$ bzw. $u = g_2$. Dann gilt $|u_1 - u_2|_0 \leq |g_1 - g_2|_0$

(iii) $\exists K > 0$:

$$|u|_0 \leq |g|_{0,\partial\Omega} + K\,|f|_{0,\overline{\Omega}} \tag{2.28}$$

Bemerkung 2.3.6 Aus (iii) folgt wieder die Korrektheit der 1. RWA.

Definition 2.3.7 (i) Randbedingungen 2. Art oder Neumann'sche RB sind

$$\frac{\partial u}{\partial \mathfrak{n}} = g \text{ auf } \partial\Omega$$

Dabei ist $\mathfrak{n}$ der Vektor in Richtung der äußeren Normale. ($\frac{\partial u}{\partial \mathfrak{n}} = 0 \Leftrightarrow$ Isolation)

(ii) RB 3. Art oder gemischte RB oder Cauchy-RB sind

$$\frac{\partial u}{\partial \mathfrak{n}} + hu = g \text{ auf } \partial\Omega.$$

(iii) Entsprechend heißt

$$\begin{aligned} Lu &= f \quad \text{in} \quad \Omega \\ \frac{\partial u}{\partial \mathfrak{n}} + hu &= g \quad \text{auf} \quad \partial\Omega \end{aligned} \tag{2.29}$$

gemischte oder 3. RWA.

Satz 2.3.8 Die 3. RWA (2.29) ist eindeutig lösbar, korrekt gestellt und erfüllt ein Monotonieprinzip, falls gilt:

$$c + h > 0, \qquad c \geq 0,\ h \geq 0.$$

2.4 Die quasiharmonische 3. RWA im $\mathbb{R}^2$

Wir wollen ein spezielles Problem mit zahlreichen, ganz unterschiedlichen Anwendungen betrachten, bei denen man konkreter die Bedeutung des Variationsprinzips erkennen kann.

Zu dieser Problemklasse ist in einer Diplom-Arbeit das Programmpaket QHARM entstanden, mit dem man die Probleme dieser Klasse auf einfache Weise lösen kann. Es soll in den Übungen eingesetzt werden.

2.4.1 Die quasiharmonische PDE

$$\frac{\partial}{\partial x}\left(K_1(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_2(x,y)\frac{\partial u}{\partial y}\right) + Q = 0 \tag{2.30}$$

(oder $-(K_1 u_x)_x - (K_2 u_y)_y = Q$) mit den Randbedingungen

$$u = g \text{ auf } \partial\Omega_1 =: C_1 \tag{2.31}$$

und/oder

$$K_1 u_x L_x + K_2 u_y L_y + q + \alpha(u - u_\alpha) = 0 \text{ auf } \partial\Omega_2 =: C_2 \tag{2.32}$$

mit

$$\partial\Omega_1 \cup \partial\Omega_2 = \partial\Omega,\ \partial\Omega_1 \cap \partial\Omega_2 = \emptyset$$

und

$$L_x = \cos(\mathfrak{n}, e_x), L_y = \cos(\mathfrak{n}, e_y)$$

wenn $\mathfrak{n}$ der äußere Normalenvektor ist.

2.4.2 Anwendung: Wärmeleitung

(1) Stationär, d.h. keine Zeitabhängigkeit. u = Temperatur.

(2) Bedeutung der Parameter:

K_1 : Wärmeleitfähigkeit des Materials in x-Richtung.
K_2 : Wärmeleitfähigkeit des Materials in y-Richtung.
Q : Wärmequelle im Gebiet Ω, wenn $Q > 0$
Wärmesenke im Gebiet Ω, wenn $Q < 0$.

(3) Für isotopes Material gilt $K_1 = K_2 = K$.

(z.B. Beton, aber nicht Fasermatten mit Faserrichtung). Dann wird (2.32) zu

$$K\frac{\partial u}{\partial \mathfrak{n}} + q + \alpha(u - u_\alpha) = 0.$$

Wir betrachten 3 Fälle:

(a) $q = \alpha = 0 : \dfrac{\partial u}{\partial \mathfrak{n}} = 0$
d.h. die entsprechenden Randteile sind perfekt isoliert. Die RB heißt mathematisch NEUMANN oder natürliche RB.

(b) $\alpha = 0 : K\dfrac{\partial u}{\partial n} = -q$
Hier fließt eine Wärmemenge q pro Einheitsfläche in das Gebiet Ω. Deshalb heißt sie auch Fluß-RB.

(c) $q = 0 : K\dfrac{\partial u}{\partial n} = -\alpha(u - u_\alpha)$
Hier ist der Wärmezufluß proportional zur Temperaturdifferenz zwischen äußerer und innerer Temperatur. Die Proportionalitätskonstante α heißt Wärmeübergangszahl. Diese RB heißt Konvektions-RB.

Satz 2.4.1 Seien K_1, K_2, c und Q in Ω gegebene C^1-Funktionen, g, q, α auf $\partial\Omega$ gegebene C^2-Funktionen, $u_\alpha \in \mathbb{R}$.

u löse das Variationsproblem $A(\varphi) \to \min\limits_{\varphi \in H^1}$ unter der Nebenbedingung $u = g$ auf $C_1 := \partial\Omega_1$ mit

$$\begin{aligned} A(\varphi) &= \int\limits_\Omega K_1(D_1\varphi)^2 + K_2(D_2\varphi)^2 + c\varphi^2 - 2Q\varphi \\ &+ \int\limits_{\partial\Omega} \alpha\varphi^2 + 2(q - \alpha u_a)\varphi \end{aligned}$$

Dann löst u die quasiharmonische 3. RWA (2.30) – (2.32).

Beweis: Wir gehen wie im Darstellungssatz, Beweis (ii), von einer Variation in der Lösung u aus, indem wir betrachten

$$\delta A := \left.\frac{A(u + \varepsilon\varphi) - A(u)}{2\varepsilon}\right|_{\varepsilon=0} = 0$$

Das ergibt

$$\begin{aligned} \delta A &= \int\limits_\Omega K_1 D_1 u D_1\varphi + K_2 D_2 u D_2\varphi + cu\varphi - Q\varphi \\ &+ \int\limits_{\partial\Omega} \alpha u\varphi + (q - \alpha u_\alpha)\varphi = 0 \end{aligned}$$

Das erste Integral formt man mit partieller Integration um:

$$\int\limits_\Omega v\, w_x = \int\limits_{\partial\Omega} v\, w\, L_x - \int\limits_\Omega v_x\, w$$

mit $v = K_1 D_1 u, w = \varphi$, also $w_x = D_1\varphi$, mit y entsprechend. Das ergibt

$$\begin{aligned} \delta A &= \int\limits_\Omega [-D_1(K_1 D_1 u) - D_2(K_2 D_2 u) + cu - Q]\,\varphi \\ &+ \int\limits_{\partial\Omega} [K_1 D_1 u\, L_x + K_2 D_2 u L_y + \alpha u + (q - \alpha u_\alpha)]\varphi \\ &= 0 \end{aligned}$$

Damit δA für alle $\varphi \in H^1$ verschwindet, müssen beide eckige Klammern verschwinden. Wegen der Nebenbedingung $u = g$ auf C_1 setzt man die Variation $\varphi = 0$ auf $C_1, u = g$ muß ja sowieso erfüllt sein, auf $C_2 := \partial\Omega \backslash C_1$ muß damit die 3. RB erfüllt sein und in Ω die PDE. ■

Bemerkung 2.4.2 (i) Bemerkenswert ist, daß die Cauchy'sche RB (Normalableitung = 0) nicht explizit im Variationsprinzip auftritt wie die Dirichlet'sche, sondern aus der Anwendung des Gauß'schen Satzes "automatisch" entsteht. Deswegen heißt das Verschwinden der Normalableitung auch natürliche RB, während die Dirichlet - RB auch Zwangsbedingung oder geometrische RB heißen.

(ii) Wir werden auf die quasiharmonische PDE in verschiedenen beispielhaften Anwendungen zurückkommen, wollen hier schon einmal die zahlreichen Anwendungen mit Bedeutung der eingebenden Größen tabellenartig aufzeigen.

Physikalische Interpretationsmöglichkeiten der quasiharmonischen Randwertaufgabe

Physikalisches Problem	gesuchte Funktion	K_1, K_2	Q	q, α
stationäre Wärmeleitung	Temperatur	Wärmeleitzahlen	Rate der intern erzeugten Wärme	q :Wärmefluß durch Rand α : Konvektionskoeffizient
Torsion prismatischer Balken	Wölbfunktion	Gleitmodule $K_1 \equiv G_1$ $K_2 \equiv G_2$	$Q = 0$	$q = G_2 \Theta x L_y - G_1 \Theta y L_x$ $\alpha = 0$
Torsion prismatischer Balken	Spannungsfunktion	$(\text{Gleitmodul})^{-1}$	$Q = 2\Theta$	$q = \alpha = 0$
drehungsfreie Strömung idealer Fluide	Strömungsfunktion	$K_1 = K_2 = 1$	$Q = 0$	q :Geschwindigkeit parallel zum Rand $\alpha = 0$
drehungsfreie Strömung idealer Fluide	Potentialfunktion	$K_1 = K_2 = 1$	$Q = 0$	q :Geschwindigkeit senkrecht zum Rand $\alpha = 0$
Sickerströmung in porösen Medien	Potentialfunktion	Materialdurchlässigkeitszahlen	$Q = S - \frac{dV}{dt}$	$q = 0$ $\alpha = 0$
Sickerströmung in porösen Medien	Strömungsfunktion	Materialdurchlässigkeitszahlen	$Q = 0$	$q = 0$ $\alpha = 0$
elektrisches Potential	Potentialfunktion	Dielektrizitätskonstante	elektrische Ladungsdichte	q : Randstrom $\alpha = 0$
magnetisches Potential	Potentialfunktion	sprezifischer magnetischer Widerstand	elektrische Stromdichte	q :angelegte Magnetfeldstärke $\alpha = 0$

Kapitel 3

Die Methode der finiten Elemente (FEM)

3.1 Geometrische Konstruktion und Ansatzfunktionen

3.1.1 Elemente-Geometrie

Wir wollen annehmen, daß das Gebiet Ω polygonal berandet ist, um das (wichtige!) Problem der Gebietsapproximation nicht betrachten zu müssen.

Das Gebiet Ω wird in endlich viele Teile, genannt Elemente, einer geometrischen Grundform zerlegt. In Frage kommen als Grundformen Dreiecke und Vierecke, geradlinig oder speziell krummlinig begrenzt. Viele Probleme kommen aus der technischen Mechanik, wo die Zerlegung oft durch das ingenieurtechnische Ausgangsproblem vorgegeben ist. Dann entsprechen die Elemente etwa Scheiben, Platten oder Fachwerken.

Wir beschränken uns auf geradlinig begrenzte Dreiecke als Grundformen. Dann nennt man die Zerlegung von Ω in solche finiten Elemente auch Triangulierung, die wir wie folgt formal definieren:

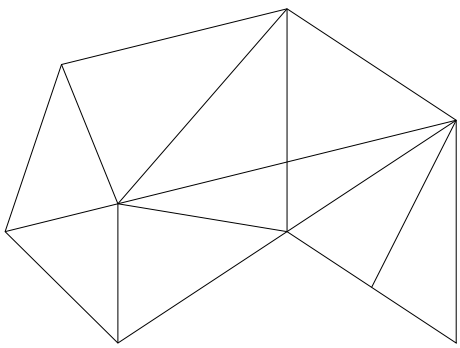
$$\begin{aligned} \tau &= \{e_1, e_2, \ldots, e_k\}, \\ \overline{\Omega} &= \bigcup_{i=1}^{k} e_i, \end{aligned} \tag{3.1}$$

wobei jedes Element e_i ein Dreieck mit nichtleerem Inneren und den Eckpunkten

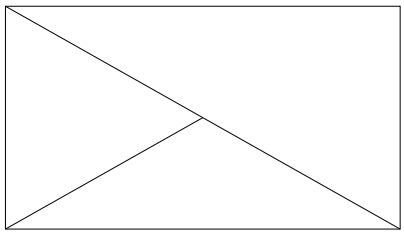
$$P_1 = (x_1, y_1),\ P_2 = (x_2, y_2),\ P_3 = (x_3, y_3)$$

ist. Dabei muß für zwei verschiedene Dreiecke $e_i,\ e_j,\ i \neq j$, folgendes gelten:

Entweder	sind sie disjunkt
oder	sie haben genau eine Kante gemeinsam
oder	sie haben genau eine Ecke gemeinsam.



Korrekte Triangulierung



Inkorrekt ('Briefumschlag')

3.1.2 Basisfunktionenansatz

Die FEM stellt ein spezielles Ritz-Verfahren 2.2.1 dar, indem sie spezielle Basisfunktionen verwendet. Dies sind Splinefunktionen auf der vorgegebenen Triangulierung τ des Gebietes Ω mit folgenden Eigenschaften:

(1) Eingeschränkt auf ein Dreieck ist die Basisfunktion ein Polynom vom Höchstgrad l.

(2) Jede Basisfunktion ist stetig, ihre Ableitung existiert fast überall und ist quadratintegrierbar, d.h. $u_i \in H^1$.

(3) In der Klasse der betrachteten Funktionen hat die Basisfunktion minimalen Träger, also eine Vereinigung möglichst weniger Dreiecke der Triangulierung.

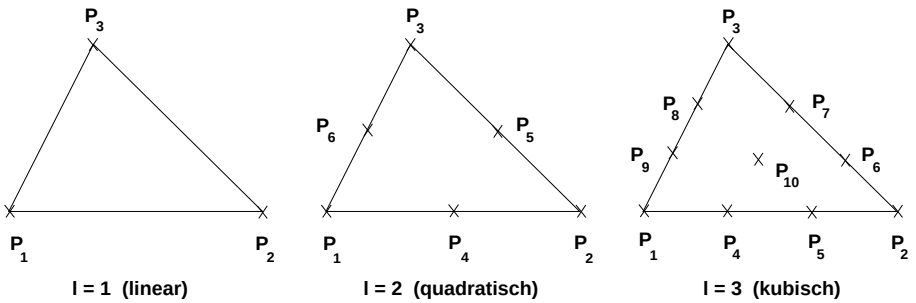
(4) Jede Basisfunktion nimmt in einem bestimmten Knotenpunkt (s.u.) den Wert 1 und in allen anderen Knotenpunkten den Wert 0 an.

Wir betrachten die Splinefunktionen aber nicht insgesamt, sondern meistens elementweise, also auf jedem Dreieck einzeln. Das hat den Vorteil, daß die Integrationen nur auf einem Modellelement formal ausgeführt werden müssen.

Auf einem beliebigen Element $e \in \tau$ ist die Splinefunktion ein Polynom in zwei Variablen. Abhängig von der Wahl von l erhalten wir p frei wählbare Koeffizienten:

$$
\begin{aligned}
l = 1 \text{ (linear)}: \quad & N(x,y) = \alpha_1 + \alpha_2 x + \alpha_3 y, && p = 3.\\
l = 2 \text{ (quadratisch)}: \quad & N(x,y) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2, && p = 6.\\
l = 3 \text{ (kubisch)}: \quad & N(x,y) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 && \\
& \qquad + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3, && p = 10.
\end{aligned}
$$

Wir wählen auf $e \in \tau$ jetzt Knotenpunkte $P_1^e, P_2^e, \ldots, P_p^e$ so, daß sich die Koeffizienten des Polynoms p aus vorgegebenen Werten in den Knotenpunkten durch Lagrange-Interpolation eindeutig ergeben.



Position und Reihenfolge der Knotenpunkte im Dreieck sind eindeutig festgelegt:

$l = 1$	Eckpunkte P_1, P_2, P_3, numeriert gegen den Uhrzeigersinn.
$l = 2$	Eckpunkte P_1, P_2, P_3 und Seitenmittelpunkte P_4, P_5, P_6; Position und Numerierung wie in der Zeichnung.
$l = 3$	Eckpunkte P_1, P_2, P_3, Seitendrittelpunkte P_4, P_5; P_6, P_7; P_8, P_9; Schwerpunkt P_{10}; Position und Numerierung wie in der Zeichnung.

Das ist wichtig für eine korrekte praktische Realisierung des Verfahrens.

Speziell betrachten wir die eindeutig bestimmten Polynomfunktionen $N_1^e, N_2^e, \ldots, N_p^e$ der vorgegebenen Form p mit:

$$N_i^e(P_j^e) = \delta_{ij} \quad \forall i, j = 1, \ldots, p. \tag{3.2}$$

Diese speziellen Polynomfunktionen N_i^e nennen wir lokale Formfunktionen. Wir fassen alle Knotenpunkte der Triangulierung τ, die nicht auf dem Rand $\partial\Omega$ liegen, in einer Menge κ zusammen. Auf dem Rand setzen wir zunächst $u = 0$ voraus. Randwerte ungleich Null werden aber später berücksichtigt, siehe 3.3.3. Für jeden (globalen) Knotenpunkt $P_K \in \kappa$ gibt es eine zugehörige Basisfunktion (oder: globale Formfunktion) N_K. Diese hat die Form:

$$N_K(x, y) = \begin{cases} N_i^e(x, y), & \text{falls} \quad P_i^e = P_K \text{ und } (x, y) \in e \\ 0 & \text{sonst.} \end{cases} \tag{3.3}$$

Die Linearkombinationen dieser Basisfunktionen bilden den Splineraum. Die Werte von Splinefunktionen in den Knotenpunkten heißen auch Knotenvariablen. Gibt man alle Knotenvariablen vor, ergibt sich daraus eine eindeutig bestimmte Splinefunktion.

Durch diese Konstruktion ist der Splineraum S eindeutig festgelegt, in dem wir das Ritz-Verfahren 2.2.1 durchführen. Jetzt geht es "nur" noch um die Fragen, wie man das zugehörige lineare Gleichungssystem $S\beta = t$ effizient aufstellt und löst.

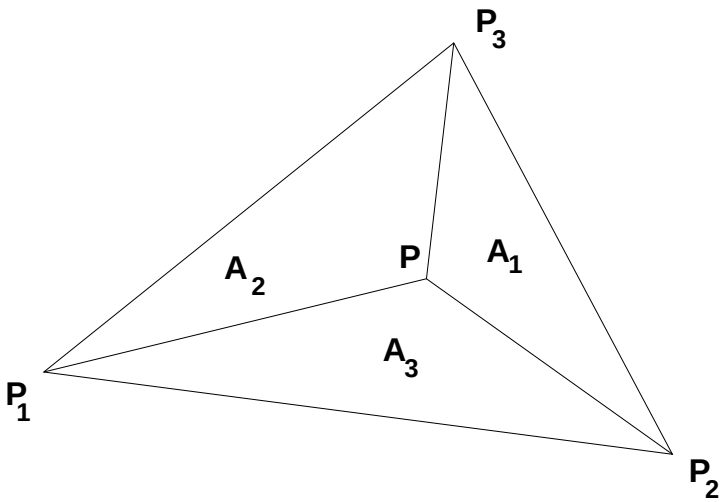
3.1.3 Natürliche Dreieckskoordinaten

Die Darstellung der lokalen Formfunktionen N_i^e wird besonders einfach, wenn man die kartesischen Koordinaten durch natürliche Dreieckskoordinaten ersetzt. Zunächst definieren wir vorzeichenbehaftete Flächeninhalte von Dreiecken. Dabei betrachtet man die Dreiecke als halbierte Parallelogramme, deren Inhalt man bekanntermaßen durch Determinanten bestimmen kann:

$$
\begin{aligned}
A(P_1,P_2,P_3) &:= \frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = \frac{1}{2}\begin{vmatrix} x_2-x_1 & y_2-y_1 \\ x_3-x_1 & y_3-y_1 \end{vmatrix} \\
&= \frac{1}{2}(x_1y_2 + x_2y_3 + x_3y_1 - x_1y_3 - x_2y_1 - x_3y_2) \qquad (3.4) \\
&\quad \forall P_1=(x_1,y_1), P_2=(x_2,y_2), P_3=(x_3,y_3) \in \mathbb{R}^2.
\end{aligned}
$$

Wir definieren die drei natürlichen Koordinaten eines beliebigen Punktes $P=(x,y)\in\mathbb{R}^2$ bezüglich eines gegebenen Dreiecks e mit den Eckpunkten $P_1,P_2,P_3\in\mathbb{R}^2$ jetzt als (siehe Zeichnung unten):

$$
\begin{aligned}
L_1(P) &:= \frac{A(P,P_2,P_3)}{A(P_1,P_2,P_3)} =: \frac{A_1}{A}, \\
L_2(P) &:= \frac{A(P_1,P,P_3)}{A(P_1,P_2,P_3)} =: \frac{A_2}{A}, \qquad (3.5) \\
L_3(P) &:= \frac{A(P_1,P_2,P)}{A(P_1,P_2,P_3)} =: \frac{A_3}{A}.
\end{aligned}
$$



Schon aus zwei der drei natürlichen Koordinaten kann man die kartesischen zurückgewinnen, mit allen drei natürlichen Koordinaten kann man aber einfacher rechnen. Wir listen einige Eigenschaften natürlicher Koordinaten auf. Es gilt $\forall P\in\mathbb{R}^2$:

(1) $L_1+L_2+L_3=1$.

(2) $x_1L_1+x_2L_2+x_3L_3=x$

(3) $y_1L_1+y_2L_2+y_3L_3=y$

(4) $L_1\geq 0\wedge L_2\geq 0\wedge L_3\geq 0 \Leftrightarrow P\in e$.

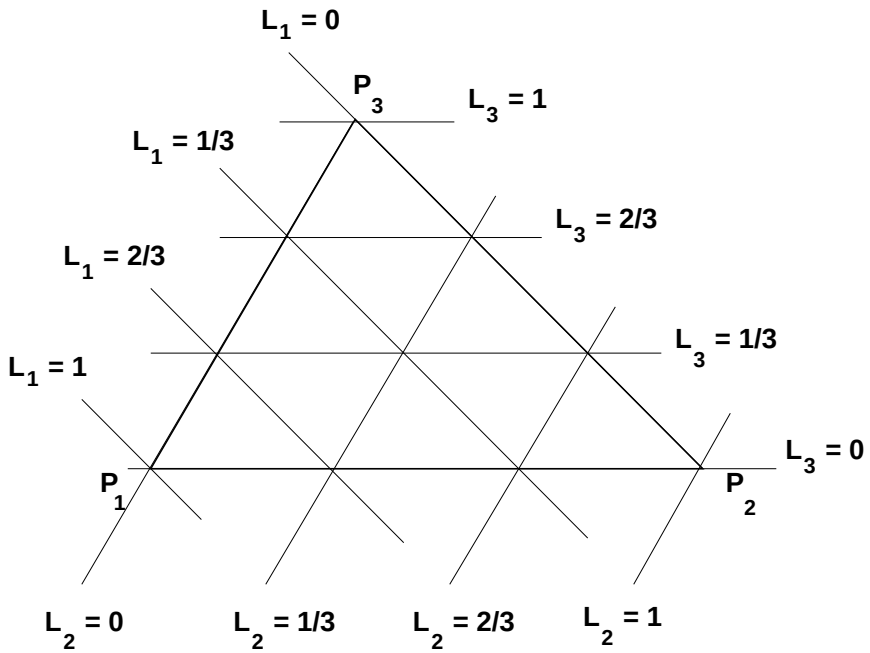
(5) $P=P_i \Rightarrow L_i=1\wedge L_j=0\wedge L_k=0$ mit i,j,k paarweise verschieden.

(6) $L_i=0\wedge L_j\geq 0\wedge L_k\geq 0 \Rightarrow P\in\overline{P_jP_k}$ mit i,j,k paarweise verschieden.

Wir beweisen exemplarisch (2):

$$
\begin{aligned}
L_1 &= \frac{1}{2A(P_1,P_2,P_3)}(x(y_2-y_3)+y(x_3-x_2)+x_2y_3-x_3y_2)\\
L_2 &= \frac{1}{2A(P_1,P_2,P_3)}(x(y_3-y_1)+y(x_1-x_3)+x_3y_1-x_1y_3)\\
L_3 &= \frac{1}{2A(P_1,P_2,P_3)}(x(y_1-y_2)+y(x_2-x_1)+x_1y_2-x_2y_1)\\
\Rightarrow\quad & x_1L_1+x_2L_2+x_3L_3\\
&= \frac{x}{2A(P_1,P_2,P_3)}(x_1(y_2-y_3)+x_2(y_3-y_1)+x_3(y_1-y_2))\\
&\quad +\frac{y}{2A(P_1,P_2,P_3)}(x_1(x_3-x_2)+x_2(x_1-x_3)+x_3(x_2-x_1))\\
&\quad +\frac{1}{2A(P_1,P_2,P_3)}(x_1x_2y_3-x_1x_3y_2+x_2x_3y_1-x_1x_2y_3+x_1x_3y_2-x_2x_3y_1)\\
&= x\frac{2A(P_1,P_2,P_3)}{2A(P_1,P_2,P_3)}+0+0=x.
\end{aligned}
$$

Die folgende Zeichnung die Verteilung der Werte der natürlichen Koordinaten verdeutlichen:



Satz 3.1.4 Integration von Monomen in natürlichen Koordinaten

Seien $a,b,c \in \mathbb{N}_0$ und L_1, L_2, L_3 die natürlichen Koordinaten in einem Dreieck e mit Eck-

punkten P_1, P_2, P_3. Dann gelten für die Dreiecksfläche $A = A(P_1, P_2, P_3)$ und jede Dreiecksseite $\overline{P_iP_j}$ die folgenden Integrationsformeln:

$$\int_e L_1^a L_2^b L_3^c \,\mathrm{d}x\mathrm{d}y = 2|A|\frac{a!\,b!\,c!}{(a+b+c+2)!} \tag{3.6}$$

$$\int_{\overline{P_iP_j}} L_i^a L_j^b \,\mathrm{d}s = \|P_i - P_j\|_2 \frac{a!\,b!}{(a+b+1)!} \tag{3.7}$$

Diese Integrale können auch nur positive Werte haben, da die natürlichen Koordinaten innerhalb des Dreiecks nichtnegativ sind.

Beweis: Die Rechnung ist elementar, aber sehr aufwendig und umfaßt etwa drei Druckseiten. Wir geben die Rechenschritte zum Beweis der ersten Gleichung an:

(1) Das Integral in kartesischen Koordinaten transformieren wir in ein Integral in den ersten beiden natürlichen Dreieckskoordinaten. Als transformiertes Integrationsgebiet erhalten wir dann ein Einheitsdreieck $d = \{(x, y) \in \mathbb{R}^2 \mid 0 \le x \le 1 \wedge 0 \le y \le 1 - x\}$. Das zu berechnende Integral hat dann den Wert
$2|A| \int_d \xi^a \eta^b (1 - \xi - \eta)^c \,\mathrm{d}\xi\mathrm{d}\eta$.

(2) Den letzten Ausdruck betrachten wir zunächst für $c = 0$, wir zeigen
$I_{ab} := \int_d \xi^a \eta^b \,\mathrm{d}\xi\mathrm{d}\eta = \frac{a!\,b!}{(a+b+2)!}$,
indem wir die Rekursionsformel $I_{ab} = q/(p+1) I_{a+1,b-1}$ durch partielle Integration herleiten und ein Integral ausrechnen als $I_{a+b,0} = 1/((p+q+1)(p+q+2))$.

(3) Das gewünschte Resultat bekommen wir nun, indem wir eine Induktion über $c \in \mathbb{N}_0$ durchführen. Der Induktionsanfang gilt nach Schritt 2, der Induktionsschritt wird dadurch realisiert, daß man das gegebene Integral mit c in eine Summe von drei Integralen mit $c - 1$ zerlegt, was offensichtlich möglich ist.

3.1.5 Lineare lokale Formfunktionen

Offenbar sind die natürlichen Dreieckskoordinaten L_1, L_2, L_3 lineare Polynome in zwei Variablen und erfüllen gerade die Anforderungen (3.2) an lokale Formfunktionen nach Basisfunktionenansatz 3.1.2. Folglich gilt:

$$N_i(x, y) = L_i \qquad \forall i = 1, 2, 3 \tag{3.8}$$

3.1.6 Quadratische lokale Formfunktionen

Auch die lokalen Formfunktionen für den quadratischen Fall kann man durch natürliche Koordinaten ausdrücken. Wir erhalten die gewünschte Darstellung von N_1 durch heuristische Überlegungen.

Wir brauchen insbesondere $N_1(P_1) = 1$ und $N_1(P_2) = N_1(P_3) = N_1(P_5) = 0$, was man durch einen Faktor L_1 in N_1 realisiert. Jetzt brauchen wir noch einen Faktor, der in den

Punkten P_4, P_6 null wird und in P_1 eins wird. Es gilt $L_1(P_4) = L_1(P_6) = 1/2$, was als Faktor $2L_1 - 1$ ergibt. Damit bekommen wir N_1, und durch zyklische Vertauschung erhalten wir auch N_2 und N_3:

$$\begin{aligned} N_1(x,y) &= L_1(2L_1 - 1) \\ N_2(x,y) &= L_2(2L_2 - 1) \\ N_3(x,y) &= L_3(2L_3 - 1) \end{aligned} \tag{3.9}$$

Damit haben wir drei von sechs lokalen Formfunktionen schon gefunden. Jetzt überlegen wir uns N_4: Alle Nullbedingungen sind erfüllt, wenn N_4 auf den Seiten $\overline{P_2P_3}$ und $\overline{P_3P_1}$ verschwindet. Das liefert der Faktor $L_1\,L_2$. Nun gilt noch $L_1(P_4) = L_2(P_4) = 1/2$. Damit bekommen wir N_4, und durch zyklische Vertauschung erhalten wir auch N_5 und N_6:

$$\begin{aligned} N_4(x,y) &= 4L_1L_2 \\ N_5(x,y) &= 4L_2L_3 \\ N_6(x,y) &= 4L_1L_3 \end{aligned} \tag{3.10}$$

3.1.7 Kubische lokale Formfunktionen

Wir wollen die lokalen Formfunktionen im kubischen Fall nur angeben und prüfen:

$$\begin{aligned} N_1 &= \frac{1}{2}L_1(3L_1 - 1)(3L_1 - 2) \\ N_2 &= \frac{1}{2}L_2(3L_2 - 1)(3L_2 - 2) \\ N_3 &= \frac{1}{2}L_3(3L_3 - 1)(3L_3 - 2) \\ N_4 &= \frac{9}{2}L_1L_2(3L_1 - 1) \\ N_5 &= \frac{9}{2}L_1L_2(3L_2 - 1) \\ N_6 &= \frac{9}{2}L_2L_3(3L_2 - 1) \\ N_7 &= \frac{9}{2}L_2L_3(3L_3 - 1) \\ N_8 &= \frac{9}{2}L_1L_3(3L_3 - 1) \\ N_9 &= \frac{9}{2}L_1L_3(3L_1 - 1) \\ N_{10} &= 27L_1L_2L_3 \end{aligned}$$

N_{10} verschwindet auf allen Seiten, im Schwerpunkt sind alle natürlichen Koordinaten 1/3, damit ist diese lokale Formfunktion korrekt. Die Prüfung der übrigen lokalen Formfunktionen geschieht leicht, wenn man die Zeichnungen auf den Seiten 33 und 35 zu Hilfe nimmt.

Bemerkung 3.1.8 Ansatz für glattere Splinefunktionen

Die globalen Splinefunktionen aus dem Basisfunktionenansatz 3.1.2 sind allesamt zwar stetig, aber in der Regel nicht überall differenzierbar. Wenn wir nicht nur die Funktionswerte, sondern auch die partiellen Ableitungen in einigen Knotenpunkten vorgeben, erhalten wir einmal stetig differenzierbare globale Splinefunktionen. Auch noch glattere Splinefunktionen sind realisierbar. Statt der Lagrange-Interpolation brauchen wir dann Hermite-Interpolation und erhalten mit den vorgegebenen Ableitungswerten zusätzliche Knotenvariablen. Darauf können wir hier aber nicht eingehen.

3.1.9 Partielle Ableitungen der Formfunktionen

Wir haben inzwischen gesehen, daß die natürlichen Dreieckskoordinaten lineare Polynome in den beiden kartesischen Koordinaten x, y sind. Wir führen lokal – d.h. im Dreieck e – die folgenden Abkürzungen für ihre Koeffizienten ein:

$$\begin{aligned} A &:= A(P_1, P_2, P_3), \\ a_1 &:= x_2y_3 - x_3y_2, \quad b_1 := y_2 - y_3, \quad c_1 := x_3 - x_2, \\ a_2 &:= x_3y_1 - x_1y_3, \quad b_2 := y_3 - y_1, \quad c_2 := x_1 - x_3, \\ a_3 &:= x_1y_2 - x_2y_1, \quad b_3 := y_1 - y_2, \quad c_3 := x_2 - x_1. \end{aligned} \tag{3.11}$$

Dann haben die natürlichen Koordinaten die Form:

$$\begin{aligned} L_1 &= \frac{1}{2A}(b_1x + c_1y + a_1), \\ L_2 &= \frac{1}{2A}(b_2x + c_2y + a_2), \\ L_3 &= \frac{1}{2A}(b_3x + c_3y + a_3). \end{aligned} \tag{3.12}$$

Mit der Kettenregel erhalten wir Darstellungen der partiellen Ableitungen der Formfunktionen:

$$\frac{\partial N(L_1(x,y), L_2(x,y), L_3(x,y))}{\partial x} = \sum_{i=1}^{3} \frac{\partial N}{\partial L_i} \frac{\partial L_i}{\partial x}. \tag{3.13}$$

Für partielle Ableitungen nach y erhält man eine entsprechende Formel. Hierbei sind die Komponenten $\partial N/\partial L_i$ wieder einfache Polynome in L_1, L_2, L_3, die Komponenten $\partial L_i/\partial x$ sind zwar element-abhängig, aber konstant:

$$\frac{\partial L_i}{\partial x} = \frac{b_i}{2A}, \quad \frac{\partial L_i}{\partial y} = \frac{c_i}{2A} \quad \forall i = 1, 2, 3. \tag{3.14}$$

(A) Im linearen Fall gilt gerade $N_i = L_i$:

$$\frac{\partial N_i}{\partial x} = \frac{b_i}{2A}, \quad \frac{\partial N_i}{\partial y} = \frac{c_i}{2A} \quad \forall i = 1,2,3. \tag{3.15}$$

(B) Der quadratische Fall:

i	$\partial N_i/\partial x$	$\partial N_i/\partial y$
$1,2,3$	$b_i(4L_i-1)/(2A)$	$c_i(4L_i-1)/(2A)$
4	$2(b_1L_2+b_2L_1)/A$	$2(c_1L_2+c_2L_1)/A$
5	$2(b_2L_3+b_3L_2)/A$	$2(c_2L_3+c_3L_2)/A$
6	$2(b_3L_1+b_1L_3)/A$	$2(c_3L_1+c_1L_3)/A$

(3.16)

Den kubischen Fall kann man analog erledigen, am besten nimmt man dazu ein Formelmanipulationssystem wie MAPLE.

3.2 Elementsteifigkeitsmatrix und -vektor

Erstens haben wir im Rahmen des Ritz-Verfahrens 2.2 den Splineraum S festgelegt, wobei die Basisfunktionen dieses Raumes globale Formfunktionen gemäß 3.1.2 sind. Zweitens sind diese globalen Formfunktionen elementweise entweder null oder lokale Formfunktionen, für die wir eine Darstellung in natürlichen Dreieckskoordinaten haben. Auch eine Darstellung der partiellen Ableitungen haben wir hergeleitet. Wir stellen jetzt Hilfsmittel zum Aufbau des linearen Gleichungssystems $S\beta = t$ bereit. Der einfacheren Darstellung wegen setzen wir in 2.1 die Koeffizienten $a_{ij} = \delta_{ij}$.

3.2.1 Allgemeine Definition der Elementmatrizen

Für jedes Element $e \in \tau$ berechnen wir lokale Energie-Skalarprodukte zweier lokaler Formfunktionen N_i^e, N_j^e. Durch diese Werte können wir die globalen Energie-Skalarprodukte $[N_I, N_J]$ für globale Formfunktionen N_I, N_J berechnen, die später die Elemente der Matrix S bilden.

$$\begin{aligned} K^e &:= (k_{ij}^e)_{i,j=1,\ldots,p} \in M(p,\mathbb{R}) \text{ mit} \\ k_{ij}^e &:= \int_e \left(\frac{\partial N_i^e}{\partial x}\frac{\partial N_j^e}{\partial x} + \frac{\partial N_i^e}{\partial y}\frac{\partial N_j^e}{\partial y} + c(x,y)\, N_i^e\, N_j^e \right) \mathrm{d}x\mathrm{d}y \end{aligned} \tag{3.17}$$

Außerdem brauchen wir zum Aufbau des Vektors y elementweise die folgenden Hilfsgrößen:

$$\begin{aligned} f^e &:= (f_i^e)_{i=1,\ldots,p} \in \mathbb{R}^p \text{ mit} \\ f_i^e &:= \int_e f(x,y) N_i^e(x,y)\, \mathrm{d}x\mathrm{d}y \end{aligned} \tag{3.18}$$

3.2.2 Elementmatrizen im linearen Fall

Zuerst berechnen wir die Elementmatrizen. Wir setzen (3.15) in (3.18) ein und erhalten nach der Vereinfachung $c(x,y) \equiv c^e = \text{const}$ und unter Anwendung von Satz 3.1.4:

$$\begin{aligned} k_{ij}^e &:= \int_e \left(\frac{b_i b_j}{4A^2} + \frac{c_i c_j}{4A^2} + c^e L_i L_j \right) \mathrm{d}x\mathrm{d}y \\ &= \frac{(b_i b_j + c_i c_j)|A|}{4A^2} + c^e\, 2|A| \frac{1!\,1!\,0!}{(1+1+0+2)!} \\ &= \frac{b_i b_j + c_i c_j}{4|A|} + \frac{c^e |A|}{12} \end{aligned} \tag{3.19}$$

Jetzt berechnen wir noch die Elementvektoren, wieder mit einer Vereinfachung $f(x,y) \equiv f^e = \text{const}$:

$$\begin{aligned} f_i^e &= \int_e f^e N_i^e \,\mathrm{d}x\mathrm{d}y \\ &= f^e\, 2|A| \frac{1!\,0!\,0!}{(1+0+0+2)!} \\ &= \frac{f^e |A|}{3} \end{aligned} \tag{3.20}$$

3.2.3 Elementmatrizen im quadratischen Fall

Hier wird die Auswertung sehr komplex. Wir rechnen daher nur exemplarisch k_{11}^e aus, wobei wir wieder $c(x,y) \equiv c^e = \text{const}$ setzen:

$$\begin{aligned} k_{11}^e &= \int_e \left[\left(\frac{b_1(4L_1 - 1)}{2A} \right)^2 + \left(\frac{c_1(4L_1 - 1)}{2A} \right)^2 + c^e L_1^2 (2L_1 - 1)^2 \right] \mathrm{d}x\mathrm{d}y \\ &= \frac{b_1^2 + c_1^2}{4A^2} \int_e (16L_1^2 - 8L_1 + 1)\,\mathrm{d}x\mathrm{d}y + c^e \int_e (4L_1^4 - 4L_1^3 + L_1^2)\,\mathrm{d}x\mathrm{d}y \\ &= \frac{b_1^2 + c_1^2}{4|A|} + \frac{c^e |A|}{30}. \end{aligned}$$

Mit Formelmanipulationssystemen wie MAPLE kann man die restlichen Formeln leicht gewinnen. Wir wollen diese hier in editorisch leicht aufgearbeiteter Fortran-Notation wiedergeben:

```
b1 = y2-y3
b2 = y3-y1
b3 = y1-y2
c1 = x3-x2
```

```
      c2 = x1-x3
      c3 = x2-x1
      a = (c3*b2-c2*b3)/2
      b1sqr=b1*b1
      b2sqr=b2*b2
      b3sqr=b3*b3
      c1sqr=c1*c1
      c2sqr=c2*c2
      c3sqr=c3*c3
      b1b2=b1*b2
      b1b3=b1*b3
      b2b3=b2*b3
      c1c2=c1*c2
      c1c3=c1*c3
      c2c3=c2*c3
c
      steif(1,1)=(b1sqr+c1sqr)/(4*a) + cquer*a/30
      steif(1,2)=(-b1b2-c1c2)/(12*a) - cquer*a/180
      steif(1,3)=(-b1b3-c1c3)/(12*a) - cquer*a/180
      steif(1,4)=(b1b2+c1c2)/(3*a)
      steif(1,5)= -cquer*a/45
      steif(1,6)=(b1b3+c1c3)/(3*a)
      steif(2,2)=(b2sqr+c2sqr)/(4*a)) + cquer*a/30
      steif(2,3)=(-b2b3-c2c3)/(12*a) - cquer*a/180
      steif(2,4)=(b1b2+c1c2)/(3*a)
      steif(2,5)=(b2b3+c2c3)/(3*a)
      steif(2,6)= -cquer*a/45
      steif(3,3)=(b3sqr+c3sqr)/(4*a) + cquer*a/30
      steif(3,4)= -cquer*a/45
      steif(3,5)=(b2b3+c2c3)/(3*a)
      steif(3,6)=(b1b3+c1c3)/(3*a)
      steif(4,4)=2*((b1sqr+b1b2+b2sqr)
     .           +(c1sqr+c1c2+c2sqr))/(3*a) + 8/45*cquer*a
      steif(4,5)=((b1b2+b2sqr+2*b1b3+b2b3)
     .           +(c1c2+c2sqr+2*c1c3+c2c3))/(3*a) + 8/45*cquer*a
      steif(4,6)=((b1sqr+b1b3+b1b2+2*b2b3)
     .           +(c1sqr+c1c3+c1c2+2*c2c3))/(3*a) + 8/45*cquer*a
      steif(5,5)=2*((b2sqr+b2b3+b3sqr)
     .           +(c2sqr+c2c3+c3sqr))/(3*a) + 8/45*cquer*a
      steif(5,6)=((2*b1b2+b2b3+b1b3+b3sqr)
     .           +(2*c1c2+c2c3+c1c3+c3sqr))/(3*a) + 8/45*cquer*a
      steif(6,6)=2*((b1sqr+b1b3+b3sqr)
     .           +(c1sqr+c1c3+c3sqr))/(3*a) + 8/45*cquer*a
c
c     Symmetrisch auffuellen: steif(i,j) = steif(j,i)
c
      do j=1,3
```

```
   rhs(j)=0
end do
do j=4,6
   rhs(j)=a*fquer/3
end do
```

Auf den kubischen Fall wollen wir aus verständlichen Gründen verzichten.

3.2.4 Numerische Integration

Die Integrale mit den Funktionen $c(x)$ und $f(x)$ können bisher nur durch Konstantsetzen dieser Funktionen in jedem Element analytisch bestimmt werden. Das hat unter Umständen einen zu starken Fehlereinfluß. Auf der rechten Seite hat man es bei höheren als quadratischen Ansätzen mit schwer manuell durchführbaren Integrationen zu tun. In beiden Fällen hilft eine numerische Integration.

Wir wollen nur eingehen auf GAUSS-Integration über Dreiecke:

$$\int_e f(x,y)\,\mathrm{d}x\mathrm{d}y \approx \sum_{i=1}^{n} w_i f(x_i, y_i), \tag{3.21}$$

wobei wir die Stützstellen (x_i, y_i) durch natürliche Koordinaten ausdrücken wollen:

Fall	i	L_1	L_2	L_3	Gewichte
$n = 3$	1	1/2	1/2	0	1/3
(im quadratischen Fall)	2	0	1/2	1/2	1/3
	3	1/2	0	1/2	1/3
$n = 7$	1	1/3	1/3	1/3	27/60
(im kubischen Fall)	2	1/2	1/2	0	8/60
	3	0	1/2	1/2	8/60
	4	1/2	0	1/2	8/60
	5	1	0	0	3/60
	6	0	1	0	3/60
	7	0	0	1	3/60
$n = 1$	1	1/3	1/3	1/3	1
(entspricht Konstantsetzen)					

(3.22)

3.3 Aufbau der Gesamtsteifigkeitsmatrix

Durch die Wahl der natürlichen Koordinaten haben wir jedes beliebige Element der Triangulation auf ein Modellelement transformiert. Das Ergebnis waren Integrationsformeln, in die nur noch die elementabhängigen Zahlen $b_i, c_i, i = 1, 2, 3$ und A eingesetzt werden mußten. Wir wollen Algorithmus-orientiert vorgehen, um das lineare Gleichungssystem $S\beta = t$ aufzubauen.

3.3.1 Datentabellen der Triangulation

Wir erinnern uns daran, daß wir Ω in Dreiecke $e_1, \ldots, e_k$ zerlegt haben und auf diesen Dreiecken Knotenpunkte angesiedelt haben. Wir numerieren erstens alle n globalen Knotenpunkte durch und stellen eine Tabelle der kartesischen Koordinaten der Eckpunkte auf:

Nummer i	x_i	y_i
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

(3.23)

Die Koordinaten der restlichen Punkte können vom Programm berechnet werden.

Die Elemente e_i sind bereits durchnumeriert. Wir stellen jetzt eine Tabelle der charakteristischen Werte aller Elemente auf. Wir listen für jedes Element e die Knotenpunkte $P_1^e, \ldots, P_p^e$ auf, indem wir ihre globalen Nummern aus der ersten Tabelle angeben. Wenn wir die Integrationen durch Konstantsetzen von c und f auf den Elementen erledigen, können wir zusätzlich noch die Funktionswerte c^e und f^e in die Tabelle mit aufnehmen:

Nummer von e	P_1^e	P_2^e	$\cdots$	P_p^e	c^e	f^e
1	Q_{11}	Q_{12}	$\cdots$	Q_{1p}	c^{e_1}	f^{e_1}
2	Q_{21}	Q_{22}	$\cdots$	Q_{2p}	c^{e_2}	f^{e_2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
k	Q_{k1}	Q_{k2}	$\cdots$	Q_{kp}	c^{e_K}	f^{e_K}

(3.24)

WICHTIG: Die Knotennummern müssen in der kanonischen Reihenfolge angegeben werden!

3.3.2 Algorithmus zur Berechnung von S und t

```
S := 0;
t := 0;
for m := 1 to k do (alle Elemente)
begin
  for i := 1 to p do
  begin
    for j := 1 to i do (alle Knotenkombinationen)
    begin
      l := Q_mi;
      r := Q_mj;
      S_lr := S_lr + k_ij^em;
      S_rl := S_lr;
    end;
  end;
  for i := 1 to p do (alle Knoten)
  begin
    l := Q_mi;
    t_l := t_l + f_i^em;
  end;
end.
```

(3.25)

Wir haben noch nicht berücksichtigt, daß wir nur aus den globalen Knotenpunkten globale Splinefuntionen machen dürfen, wenn diese Knotenpunkte nicht auf dem Rand $\partial\Omega$ liegen. Diesen Fehler korrigieren wir im nächsten Punkt. Bei der Berechnung der Matrix S oben haben wir auch ihre Symmetrie ausgenutzt.

3.3.3 Berücksichtigung der Randpunkte

Es gibt drei Möglichkeiten, Randpunkte und Randwerte zu berücksichtigen, wobei wir der größeren Allgemeinheit wegen hier auch Randwerte $\neq 0$ zulassen.*

Sei l die Nummer des betrachteten Randpunktes in der globalen Knotenpunktnumerierung, U_l der zugehörige Randwert.

(A) Elimination von Zeile und Spalte

(1) Aktualisiere die rechte Seite: setze $t_i = t_i - s_{il}U_l \quad \forall i = 1, \ldots, n$.

(2) Streiche die $l-te$ Spalte in S, dann die $l-te$ Zeile in S und t.

(3) Umnumerieren.

Nachteil dieser Methode: Das Umnumerieren (3) erfordert viel Programmier- und Rechenaufwand.

(B) Einheitszeilen setzen

(1) Aktualisiere die rechte Seite: setze $t_i = t_i - s_{il}U_l \quad \forall i = 1, \ldots, n$ wie in (A)
setze zusätzlich $\beta_l := U_l$.

(2) Ersetze die $l-te$ Zeile und Spalte durch eine Einheitszeile und -spalte: setze $S_{ll} := 1$, $S_{lj} := S_{jl} := 0, j = 1, \ldots, n \wedge j \neq l$.

(C) Randpunkte vorher berücksichtigen

Wir legen die Numerierung der Knotenpunkte aus 3.3.1 so an, daß die Indizes $1, \ldots, n_1$ zu inneren Punkten gehören und die Indizes $n_1 + 1, \ldots, n$ zu Randpunkten gehören.

(1) Aktualisiere die rechte Seite: setze $t_i = t_i - \sum_{l=n_1+1}^{n} s_{il}U_l \quad \forall i = 1, \ldots, n_1$.

(2) Streiche Zeilen und Spalten mit Indizes $n_1 + 1, \ldots, n$.

(3) Setze $n := n_1$.

Diese Methode hat den Nachteil, daß der Benutzer und nicht der Programmierer aufpassen muß, richtig zu numerieren. Für geometrisch nahe Punkte treten dann oft sehr unterschiedliche Punktnummern auf.

*Mathematisch bedeutet das eine Transformation des Problems, aber praktisch entspricht diese Transformation gerade den im folgenden angegeben Aktionen.

Bemerkung 3.3.4 Aufwand und Verfahrenswahl

(1) In 3.3.3 ist (B) die normale Methode. Die Aufwandserhöhung ist bei großen Systemen unerheblich.

(2) S ist symmetrisch und es sind links und rechts höchstens $n(\widehat{m}+1)$ Elemente $\neq$ Null, wenn $2\widehat{m}+1$ die Bandbreite von S ist:

$$\widehat{m} := \max_{l=1,\ldots,k} \max_{i,j=1,\ldots,p} |Q_{li} - Q_{lj}|.$$

Deshalb ist die Speicherung von S in einer $n \times (\widehat{m}+1)$-Matrix $\tilde{S}$ möglich. In einem Programm wird natürlich gleich nur mit $\tilde{S}$ statt S gerechnet, schon bei der Aufsummierung der Werte.

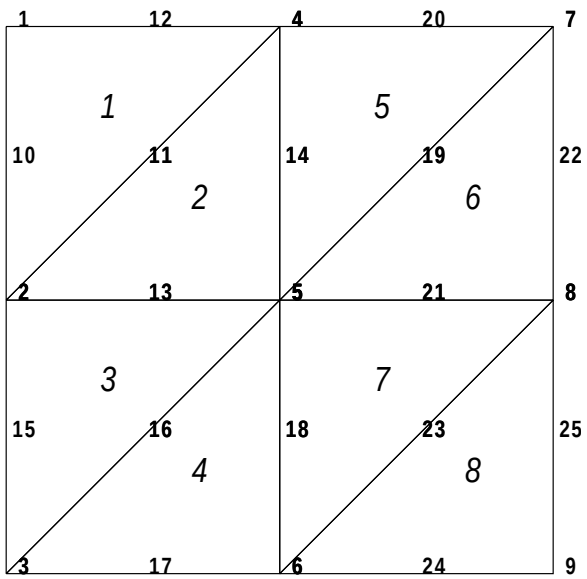
(3) Die Lösung von $S\beta = t$ erfolgt am einfachsten mit dem Cholesky-Verfahren, speziellere Methoden werden wir unten kennenlernen.

(4) Die Lösung im globalen Knotenpunkt P_k ist jetzt einfach gleich β_k. Zur Berechnung anderer Lösungswerte wird auf die Elementansätze zurückgegriffen.

Beispiel 3.3.5

$$\begin{aligned} -\Delta u &= 1 \text{ in } \Omega := (0,1) \times (0,1), \\ u &= 0 \text{ auf } \partial\Omega. \end{aligned}$$

Das bedeutet in unserem Grundproblem (2.1) $c \equiv 0$ und $f \equiv 1$. Wir triangulieren nach dem quadratischen Ansatz und der folgenden Grafik:



Die zugehörige Tabelle der globalen Eckpunkte lautet:

i	x_i	y_i
1	0	1
2	0	1/2
3	0	0
4	1/2	1
5	1/2	1/2
6	1/2	0
7	1	1
8	1	1/2
9	1	0

Die Koordinaten der Knotenpunkte, die keine Eckpunkte sind, werden automatisch berechnet.

Die zugehörige Tabelle der Elementwerte lautet, wobei hier nur die globalen Knotennummern aufgelistet werden:

Elementnummer	P_1	P_2	P_3	P_4	P_5	P_6	$\max \lvert Q_{mi} - Q_{mj} \rvert$
1	1	2	4	10	11	12	11
2	2	5	4	13	14	11	12
3	2	3	5	15	16	13	14
4	3	6	5	17	18	16	15
5	4	5	7	14	19	20	16
6	5	8	7	21	22	19	17
7	5	6	8	18	23	21	18
8	6	9	8	24	25	23	19

Damit ist die Halbbandbreite der 25×25-Steifigkeitsmatrix 19, d.h. die Matrix ist voll besetzt. Bei vernünftiger Numerierung wäre die Halbbandbreite 10. Da 16 Randpunkte vorliegen, bekommt man nach Elimination der zu Randpunkten gehörenden Zeilen und Spalten folgendes 9×9-System:

Zeile	Matrix									rechte Seite
(1= 5)	4	0	-4/3	-4/3	0	-4/3	0	-4/3	0	0
(2=11)	0	16/3	-4/3	-4/3	0	0	0	0	0	1/12
(3=13)	-4/3	-4/3	16/3	0	-4/3	0	0	0	0	1/12
(4=14)	-4/3	-4/3	0	16/3	0	0	-4/3	0	0	1/12
(5=16)	0	0	-4/3	0	16/3	-4/3	0	0	0	1/12
(6=18)	-4/3	0	0	0	-4/3	16/3	0	0	-4/3	1/12
(7=19)	0	0	0	-4/3	0	0	16/3	-4/3	0	1/12
(8=21)	-4/3	0	0	0	0	0	-4/3	16/3	-4/3	1/12
(9=23)	0	0	0	0	0	-4/3	0	-4/3	16/3	1/12

Durch die Symmetrie des Problems sind die Lösungswerte gleich in den Punkten 11, 16, 19, 23 und in 13, 14, 18, 21, sodaß zusammen mit dem Punkt 5 nur drei Punkte berücksichtigt werden müssen. Das liefert das 3×3-System

```
Zeile    Matrix                    rechte Seite
(1= 5)     4       0    -16/3          0
(2=11)     0     16/3    -8/3         1/12
(3=13)   -4/3    -8/3    16/3         1/12
```

Als Lösungen ergeben sich:

$$\begin{aligned} u_5 &= 0.075 \\ u_{11} &= 0.04375 \\ u_{13} &= 0.05625 \end{aligned}$$

3.4 Weitere Elemente und Ansatzfunktionen

Es sollen noch einige wichtige Elementetypen und die auf ihnen üblichen Ansatzfunktionen erwähnt werden:

- Die wichtigen Differenzenverfahren gehen meistens von einer Zerlegung des Gebietes in Rechtecke aus. Werden dementsprechend Rechtecke (spezieller Quadrate, allgemeiner Parallelogramme) als Elemente gewählt, so liefern **bilineare Funktionen** einen eindeutigen Interpolationsansatz.
- Die **Hermiteschen Elemente** sind finite Elemente, bei denen die Anzahl der Knotenvariablen größer ist als die Anzahl Knotenpunkte, bei denen also Ableitungen als Knotenvariablen auftreten. Dann wird statt der Lagrange-Interpolation zur Erfüllung der Freiheitsgrade die Hermite-Interpolation genommen.
- **Krummlinige Elemente** sind besser als mit Geraden berandete Elemente wie Dreiecke und Vierecke in der Lage, den Rand eines Gebietes zu approximieren. Dabei wählt man einen festen Funktionentyp zur Darstellung der Randstücke, die man parametrisch in Abhängigkeit vorzugebender Punkte darstellt. So kann man diese dann auf ein Einheitselement transformieren.
- **Krummlinige Koordiantensysteme** erfüllen etwa den gleichen Zweck, als Beispiel wollen wir Polarkoordinaten betrachten.
- Die gebräuchlichsten **3-dimensionalen Elemente** sind Tetraeder. Die ihnen zuzuordnenden Ansatzfunktionen sollen ebenso dargestellt werden wie zwei weitere mögliche 3-dimensionale Elemente.

Eine anschauliche Übersicht über gewöhnliche und ungewöhnliche Elemente im $\mathbb{R}^2$ und $\mathbb{R}^3$ findet man in [GR92].

3.4.1 Bilineare Ansätze auf Vierecken

Für Differenzenverfahren ist die günstigste Zerlegung eines Gebietes die in Quadrate, siehe Abb. 3.1 links. Im Einheitsquadrat (siehe Abb. 3.1 rechts) wird eine bilineare Funktion

$$u(\xi, \eta) = \alpha_1 + \alpha_2 \xi + \alpha_3 \eta + \alpha_4 \xi \eta \tag{3.26}$$

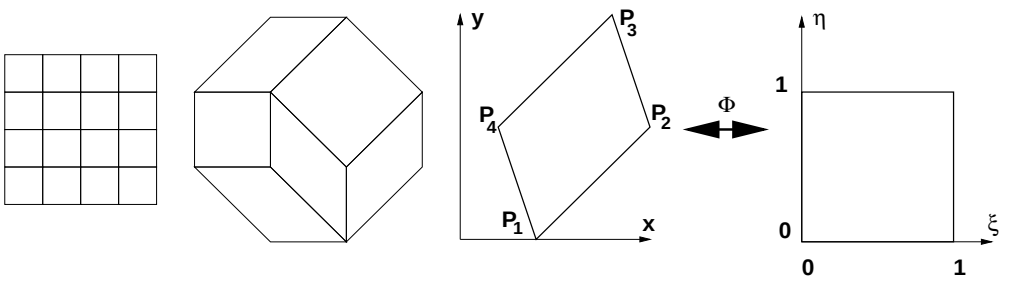


Abbildung 3.1: Quadrate und Parallelogramme

durch Vorgabe der Werte in den Eckpunkten eindeutig festgelegt. u ist auf den Seiten des Quadrats und auf allen Achsenparallelen linear, allerdings nicht auf den Seiten eines Parallelogramms. Deshalb wird das Einheitsquadrat auf ein beliebiges Parallelogramm abgebildet:

$$\Phi : (\xi, \eta) \to (x, y) = P_1 + \xi\,(P_2 - P_1) + \eta\,(P_4 - P_1). \tag{3.27}$$

Damit bekommt dort eine Ansatzfunktion

$$N(x, y) = u(\Phi^{-1}(x, y)). \tag{3.28}$$

Die Transformation muss nicht explizit ausgerechnet werden, da alle Integrationen auf dem Einheitsquadrat durchgeführt werden. Die Ansatzfunktion N ist linear auf den Seiten des Parallogramms. Das ist wichtig für einen gemischten Ansatz aus Dreiecken mit linearen Polynomen und Parallelogrammen, mit dem man ein Gebiet Ω besser ausfüllen kann als nur mit einem der beiden geometrischen Formtypen, siehe Abb. 3.2 links.

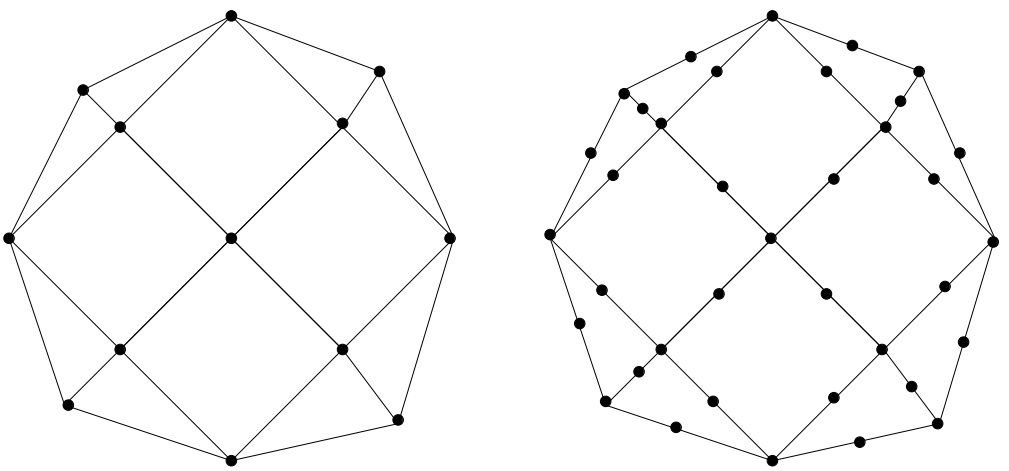


Abbildung 3.2: Gemischte Ansätze: bilinear/linear und quadratisch/Serendipity

Sowohl in [Sch90] als auch in [Hac86] wird noch kurz ein quadratischer Ansatz auf Parallelogrammen betrachtet. Dieser gehört zur sogenannten Serendipity-Klasse und passt zum quadratischen Ansatz auf Dreiecken bei einer Kombinationen von Dreiecken mit Parallelogrammen, siehe Abb. 3.2 rechts.

3.4.2 Hermitesche FEM-Ansätze

Hermitesche Interpolation berücksichtigt neben Funktions- auch Ableitungswerte, siehe Numerik I.

Als Ansatz bei Finite-Elemente-Methoden ist sie dann besonders wichtig, wenn die Differentialgleichung höherer als 2. Ordnung ist. Wichtigster Vertreter dieser Klasse ist die biharmonische Differentialgleichung (Plattengleichung):

$$\begin{aligned} u''''(x) &= f(x), \quad u(0) = u'(0) = u(1) = u'(1) = 0 \text{ im } \mathbb{R}^1 \\ \Delta^2 u &= f \text{ in } \Omega, \; u = \frac{\partial u}{\partial \mathfrak{n}} = 0 \text{ auf } \partial\Omega, \text{ im } \mathbb{R}^2 \\ \Delta^2 &= \frac{\partial^4}{\partial x^4} + 2\,\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4} \end{aligned} \tag{3.29}$$

Schwache Lösungen dieser Differentialgleichungen werden im Sobolev-Raum H_0^2 gesucht.

Eindimensionale kubische Hermite-Interpolation

Für eine Zerlegung $\{0 = x_0, x_1, x_2, \ldots, x_{n-1}, x_n, x_{n+1} = 1\}$ des Intervalls $[0,1]$ existieren zu jedem inneren Punkt x_i zwei Ansatzfunktionen N_{i1} und N_{i2}, die folgende Bedingungen erfüllen:

$$\begin{aligned} N_{i1}(x_i) = 1, \; N'_{i1}(x_i) = 0, \; N_{i1}(x_j) = 0, \; N'_{i1}(x_j) = 0, \; \forall j \neq i, \\ N_{i2}(x_i) = 0, \; N'_{i2}(x_i) = 1, \; N_{i2}(x_j) = 0, \; N'_{i2}(x_j) = 0, \; \forall j \neq i, \end{aligned}$$

Diese Basisfunktionen liegen in $C^1[0,1]$ und in $H_0^2[0,1]$.

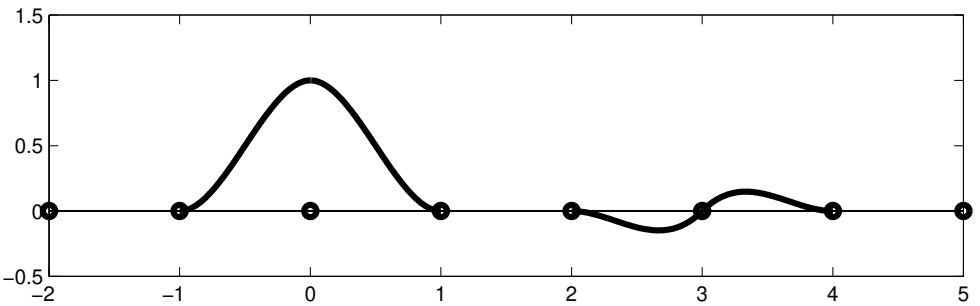


Abbildung 3.3: Zwei Grundformen kubischer Hermitescher B-Splines

Zweidimensionale Hermite-Ansätze

Die Übertragung des obigen eindimensionalen Ansatzes auf zwei Dimensionen gelingt mit einem Tensorproduktansatz

$$N(x,y) = \sum_{i,j=0}^{3} \alpha_{ij} x^i \, y^j \tag{3.30}$$

auf Rechtecken, siehe [Hac86] oder ausführlicher und allgemeiner in [MM78], siehe auch Abb. 3.4 rechts. Diese Interpolation hat 16 Freiheitsgrade und liefert stetig differenzierbare Ansatzfunktionen.

Ein – allerdings nur stetiger – kubischer Hermite-Ansatz auf Dreiecken ergibt sich bei Vorgabe der Funktionswerte in den Eckpunkten und im Schwerpunkt sowie Vorgabe der Ableitungswerte u_x und u_y in den Eckpunkten. Das ergibt den Freiheitsgrad 10 wie beim kubischen Lagrange-Ansatz mit 10 Knotenpunkten, siehe auch Abb. 3.4 links.

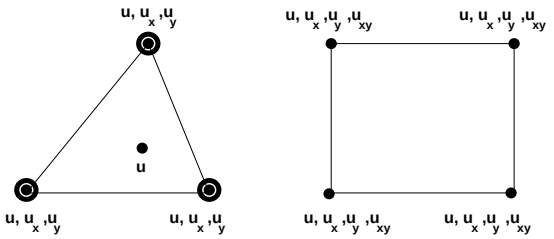


Abbildung 3.4: Zweidimensionale Hermitesche Elemente

3.4.3 Krummlinig berandete Dreiecke als Elemente

Besonders bei Ansätzen höheren Grades ist eine gute Randapproximation wichtig, um dort nicht einen Genauigkeitsverlust im Vergleich zur FEM-Lösung zu erleiden, siehe auch 3.7.3. Der Rand läßt sich natürlich besser mit krummlinig berandeten Elementen approximieren als mit Geradenstücken. Dabei muß allerdings auch bei krummlinig berandeten Elementen eine bestimmte Parameterdarstellung des Randes vorgegeben sein. Man spricht daher auch von **isoparametrischen** Elementen oder Transformationen.

Wir wollen hier nur Dreiecke mit quadratischen Kurven als Seiten betrachten und die Vorgehensweise an einem Beispiel klarmachen. Sei T_i ein krummliniges Dreieck mit den Punkten P_1, P_2 und P_3 als Eckpunkten und P_4, P_5 und P_6 als Punkten auf den drei Seiten in der bekannten Reihenfolge, siehe die Zeichnung unten.

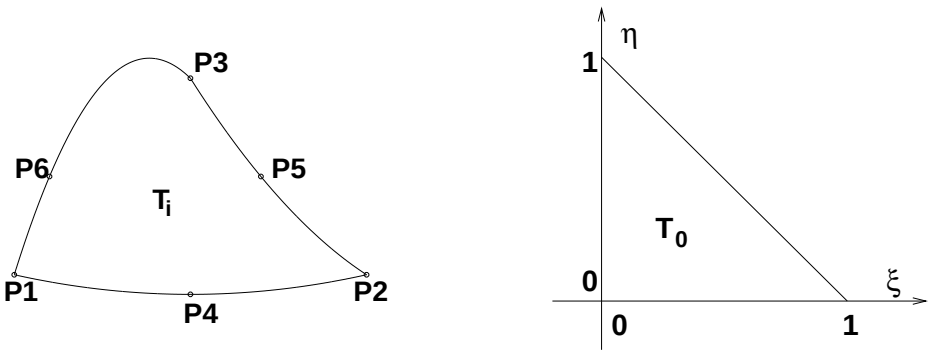
In der Zeichnung haben die Interpolationspunkte die Koordinaten:

```
x=[0.0, 1.0, 0.5,  0.5, 0.7, 0.1],
y=[0.0, 0.0, 1.0, -0.1, 0.5, 0.5].
```

Dann läßt sich das Dreieck mit zwei quadratischen Polynomen auf das Einheitsdreieck T_0 transformieren:

$$\begin{aligned} x &= \gamma_1 + \gamma_2\,\xi + \gamma_3\,\eta + \gamma_4\,\xi^2 + \gamma_5\,\xi\,\eta + \gamma_6\,\eta^2 \\ y &= \delta_1 + \delta_2\,\xi + \delta_3\,\eta + \delta_4\,\xi^2 + \delta_5\,\xi\,\eta + \delta_6\,\eta^2 \end{aligned} \tag{3.31}$$

Dabei sind (ξ, η) die Koordinaten in T_0.



Setzt man die Koordinaten der Eckpunkte des Einheitsdreiecks ein, so bekommt man

$$\begin{aligned} x_1 &= \gamma_1 \\ x_2 &= \gamma_1 + \gamma_2 + \gamma_4 \\ x_3 &= \gamma_1 + \gamma_3 + \gamma_6 \\ x_4 &= \gamma_1 + 1/2\,\gamma_2 + 1/4\,\gamma_4 \\ x_5 &= \gamma_1 + 1/2\,\gamma_2 + 1/2\,\gamma_3 + 1/4\,(\gamma_4 + \gamma_5 + \gamma_6) \\ x_6 &= \gamma_1 + 1/2\,\gamma_3 + 1/4\,\gamma_6 \end{aligned} \tag{3.32}$$

und dieselben Beziehungen zwischen y und δ.

Inversion der Abbildungsmatrix liefert

$$\begin{aligned} \gamma &= A\,x \\ \delta &= A\,y \\ \text{mit} & \\ A &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -3 & -1 & 0 & 4 & 0 & 0 \\ -3 & 0 & -1 & 0 & 0 & 4 \\ 2 & 2 & 0 & -4 & 0 & 0 \\ 4 & 0 & 0 & -4 & 4 & -4 \\ 2 & 0 & 2 & 0 & 0 & -4 \end{pmatrix} \end{aligned} \tag{3.33}$$

Die Transformationskoeffizienten γ und δ sind also durch die 6 Punkte und den quadratischen Ansatz eindeutig festgelegt. Dabei können durch Lage-Variation der Punkte P_4 bis P_6 stark unterschiedliche Randkurven erzeugt werden. Die Transformation ist nicht für alle möglichen Punktevektoren eineindeutig, z.B. ist sie es dann nicht, wenn P_1, P_2 und P_3 kollinear liegen. Es können sich auch die Kurven schneiden, was bei der gegebenen Problemstellung nicht sinnvoll ist. Mathematisch sind Bedingungen zur Verhinderung solcher Fälle schwer zu formulieren, geht man von einem sinnvollen Dreiecksnetz aus und legt nach ihm die Punkte fest, dann sind die Probleme aus praktischer Sicht nicht gravierend.

Die Eindeutigkeit der Abbildung kann mit Hilfe ihrer Jacobi-Determinante überprüft werden:

$$J = \left| \frac{\partial(x,y)}{\partial(\xi,\eta)} \right| = \begin{vmatrix} \gamma_2 + 2\gamma_4\xi + \gamma_5\eta & \delta_2 + 2\delta_4\xi + \delta_5\eta \\ \gamma_3 + \gamma_5\xi + 2\gamma_6\eta & \delta_3 + \delta_5\xi + 2\delta_6\eta \end{vmatrix}. \tag{3.34}$$

Berechnung der Steifigkeitsmatrix

Werden die Ansatzfunktionen vom selben Polynomtyp gewählt wie die Darstellung der Randkurven, dann kann die Variablentransformation mit Hilfe der Ansatzfunktionen dargestellt werden. Diese Methode soll hier am Beispiel quadratischer Polynome skizziert werden.

Im Einheitselement T_0 gilt (lokal):

$$u^{(e)}(\xi,\eta) = \sum_{i=1}^{p} c_i\, N_i^{(e)}(\xi,\eta), \quad (p=6). \tag{3.35}$$

Das lässt sich vektoriell schreiben. Sei dazu $u_e := (c_1,\ldots,c_p)^T$ der Vektor der Knotenvariablen und $N^{(e)}(\xi,\eta) := (N_1^{(e)}(\xi,\eta),\ldots,N_p^{(e)}(\xi,\eta))^T$ der Vektor der lokalen Ansatzfunktionen. Dann ist

$$u^{(e)}(\xi,\eta) = u_e^T\, N^{(e)}(\xi,\eta). \tag{3.36}$$

Sind $\hat{x} = (x_1,\ldots,x_p)^T$ und $\hat{y} = (y_1,\ldots,y_p)^T$ die Koordinatenvektoren der Knotenpunkte in T_i, dann lässt sich die Transformation

$$(x,y) \quad \leftrightarrow \quad (\xi,\eta)$$

mit den lokalen Formfunktionen darstellen als:

$$x = \hat{x}^T N^{(e)}(\xi,\eta), \quad y = \hat{y}^T N^{(e)}(\xi,\eta). \tag{3.37}$$

Damit erhält die Jacobimatrix die Form

$$J = \frac{\partial(x,y)}{\partial(\xi,\eta)} = \begin{pmatrix} x_\xi & y_\xi \\ x_\eta & y_\eta \end{pmatrix} = \begin{pmatrix} \hat{x}^T N_\xi^{(e)} & \hat{y}^T N_\xi^{(e)} \\ \hat{x}^T N_\eta^{(e)} & \hat{y}^T N_\eta^{(e)} \end{pmatrix}. \tag{3.38}$$

Diese Form kann ausgenutzt werden, um die Elementsteifigkeitsmatrix zu berechnen (hier wieder nur für die Bilinearform zum Laplace-Operator):

$$\begin{aligned} & \int_{T_i} \left(u_x^2 + u_y^2\right)\, dxdy \\ = & \int_{T_0} \left[\left(u_\xi \xi_x + u_\eta \eta_x\right)^2 + \left(u_\xi \xi_y + u_\eta \eta_y\right)^2\right] \det(J)\, d\xi d\eta \\ = & \int_{T_0} \left[\left(u_e^T N_\xi^{(e)} \xi_x + u_e^T N_\eta^{(e)} \eta_x\right)^2 + \left(u_e^T N_\xi^{(e)} \xi_y + u_e^T N_\eta^{(e)} \eta_y\right)^2\right] \det(J)\, d\xi d\eta. \end{aligned}$$

Aus diesem Integral kann der Vektor der Knotenvariablen als unabhängige Größe herausgezogen werden:

$$\begin{aligned} & \int_{T_i} \left(u_x^2 + u_y^2\right)\, dxdy \\ = & u_e^T \left\{ \int_{T_0} \left[\left(N_\xi^{(e)} \xi_x + N_\eta^{(e)} \eta_x\right)^2 + \left(N_\xi^{(e)} \xi_y + N_\eta^{(e)} \eta_y\right)^2\right] \det(J)\, d\xi d\eta \right\} u_e. \end{aligned}$$

Die Ausdrücke unter dem Integral von der Form (Vektor $\times$ Vektor) sind als Dyaden zu berechnen, es sind also Matrizen, und der Ausdruck in der geschweiften Klammer ist die

Elementsteifigkeitsmatrix. Zur Auswertung werden noch die Ableitungen ξ_x, η_x, ξ_y, η_y benötigt. Diese werden mit Hilfe der inversen Jacobimatrix und Cramers Regel berechnet:

$$\xi_x = \frac{y_\eta}{\det(J)},\ \xi_y = -\frac{x_\eta}{\det(J)},\ \eta_x = -\frac{y_\xi}{\det(J)},\ \eta_y = \frac{x_\xi}{\det(J)}.$$

Insgesamt entstehen also gebrochen rationale Ausdrücke, die numerisch integriert werden.

Die beiden anderen benötigten Integrale ergeben sich als

$$\int_{T_i} u^2\, dxdy = u_e^T\, M_e\, u_e \text{ mit } M_e = \int_{T_0} N^{(e)} N^{(e)} \det(J)\, d\xi d\eta \tag{3.39}$$

und

$$\int_{T_i} u\, dxdy = u_e^T\, b_e \text{ mit } b_e = \int_{T_0} N^{(e)} \det(J)\, d\xi d\eta. \tag{3.40}$$

Die entsprechende Rechnung für allgemeinere Differentialoperatoren lässt sich leicht übertragen.

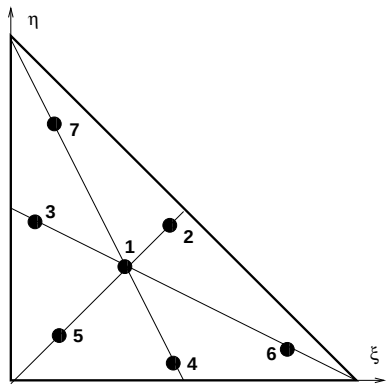
Numerische Integration im Einheitsdreieck

Wir wollen noch eine 7-Punkte-Gaußformel kennenlernen, die innere Punkte im Einheitsdreieck T_0 benutzt.

$$\int_{T_0} g(\xi, \eta)\, d\xi\, d\eta = \sum_{i=1}^{7} w_i\, g(\xi_i, \eta_i) \tag{3.41}$$

Diese Gaußformel integriert Polynome bis zum Höchstgrad 5 exakt. Die Koordinaten der Punkte und die Gewichte ergeben sich aus der folgenden Tabelle:

i	ξ_i	η_i	w_i
1	$1/3$	ξ_1	0.1125
2	$(6+\sqrt{15})/21$	ξ_2	$(115+\sqrt{15})/2400$
3	$(9-2\sqrt{15})/21$	ξ_2	$(115+\sqrt{15})/2400$
4	$(6+\sqrt{15})/21$	ξ_3	$(115+\sqrt{15})/2400$
5	$(6-\sqrt{15})/21$	ξ_5	$(115-\sqrt{15})/2400$
6	$(9+2\sqrt{15})/21$	ξ_5	$(115-\sqrt{15})/2400$
7	$(6-\sqrt{15})/21$	ξ_6	$(115-\sqrt{15})/2400$



Numerische Integration im Einheitsquadrat

Hier ergeben sich einfacher zweidimensionale Integrationsformeln durch Übertragung eindimensionaler Formeln:

Seien (x_i, w_i), $i = 1, \ldots, m$ Stützstellen und Gewichte einer eindimensionalen Gauß-Formel für das Intervall $[-1, 1]$. Dann liefern die transformierten Stützstellen $\sigma_i = 1/2 + 1/2\, x_i$ und Gewichte $w_i = 1/2$ die Formel

$$\int_{Q_0} g(\xi, \eta)\, d\xi d\eta = \frac{1}{4} \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j g(\sigma_i, \sigma_j). \tag{3.42}$$

Zusammenfassung der Schritte zur Erstellung der Steifigkeitsmatrix

(1) Berechne mit den Integrationspunkten $R_i := (\xi_i, \eta_i)$, $i = 1, \ldots, m$, die Vektoren $N(R_i)$, $N_\xi(R_i)$, $N_\eta(R_i)$ (in T_0 bzw. Q_0 mit $m^2 \to m$).

(2) Berechne mit den Koordinatenvektoren $\hat{x}$ und $\hat{y}$ eines Elementes e die Werte von $x_\xi(R_i)$, $x_\eta(R_i)$, $y_\xi(R_i)$, $y_\eta(R_i)$ und $\det(J)(R_i)$.

(3) Berechne für jedes Element e m Hilfsvektoren

$$\begin{aligned} h_1(R_i) &:= (y_\eta(R_i) N_\xi(R_i) - y_\xi(R_i) N_\eta(R_i)) / \sqrt{\det(J)(R_i)}, \\ h_2(R_i) &:= (-x_\eta(R_i) N_\xi(R_i) + x_\xi(R_i) N_\eta(R_i)) / \sqrt{\det(J)(R_i)}. \end{aligned}$$

(4)

$$\begin{aligned} \text{Berechne } S_e &= \sum_{i=1}^{m} w_i (h_1(R_i) h_1^T(R_i) + h_2(R_i) h_2^T(R_i)) \\ M_e &= \sum_{i=1}^{m} w_i \det(J)(R_i) N(R_i) N^T(R_i) \\ b_e &= \sum_{i=1}^{m} w_i \det(J)(R_i) N(R_i) \end{aligned}$$

3.4.4 Finite Elemente in Polarkoordinaten

Kreisbögen kommen bei realistischen Konstruktionen häufig als Randelemente vor. Die Parameter-Darstellung der Randkurven als Kreisbögen ist aber rechnerisch schwieriger als die mit polynomialen Kurven wie in 3.4.3. Unabhängig von der Finite-Elemente-Methode bietet sich außerdem bei kreisförmigen Gebieten die Darstellung und Lösung der PDE in Polarkoordinaten an. Das Problem wird also zunächst umformuliert und dann werden entsprechende Elemente definiert. Dabei werden hier die Rechnungen nur für den Laplace-Operator durchgeführt. Durch die Angabe der Transformationsformeln ist die Übertragung auf andere Operatoren einfach.

Für die Transformation kartesischer auf Polarkoordinaten gelten folgende Formeln:

$$\begin{aligned} x &= r \cos\varphi, \quad 0 \le r \le \infty \\ y &= r \sin\varphi, \quad 0 \le \varphi \le 2\pi \end{aligned}$$

$$\begin{aligned} J = \frac{\partial(x,y)}{\partial(r,\varphi)} &= \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} \\ \Longrightarrow \quad \det(J) &= r \\ J^{-1} = \frac{\partial(r,\varphi)}{\partial(x,y)} &= \begin{pmatrix} \cos\varphi & -\sin\varphi/r \\ \sin\varphi & \cos\varphi/r \end{pmatrix} \end{aligned} \tag{3.43}$$

$$\begin{aligned} \Longrightarrow \quad u_x = u_r\, r_x + u_\varphi\, \varphi_x &= u_r\, \cos\varphi - u_\varphi \frac{\sin\varphi}{r} \\ u_y = u_r\, r_y + u_\varphi\, \varphi_y &= u_r\, \sin\varphi + u_\varphi \frac{\cos\varphi}{r} \end{aligned} \tag{3.44}$$

Mit den Transformationsformeln (Differentialgeometrie, siehe [KS93] oder [TWM85])

$$\Delta = \frac{1}{\det(J)} \left(\frac{\partial}{\partial\, r}, \frac{\partial}{\partial\, \varphi} \right) \det(J)\ \left(J^T J\right)^{-1} \begin{pmatrix} \dfrac{\partial}{\partial\, r} \\ \dfrac{\partial}{\partial\, \varphi} \end{pmatrix} \tag{3.45}$$

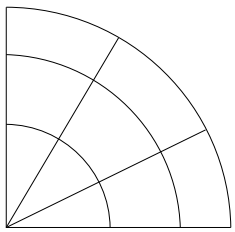
wird der Laplace-Operator in Polarkoordinaten zu

$$\Delta = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r}\frac{\partial}{\partial\varphi}\left(\frac{1}{r}\frac{\partial}{\partial\varphi}\right) = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\varphi^2}. \tag{3.46}$$

Werden jetzt formal Elemente definiert als

$$e = \{(r,\varphi) \,\Big|\, r_1 \le r \le r_2, \quad \varphi_1 \le \varphi \le \varphi_2\}, \tag{3.47}$$

dann ergeben sich Dreiecke und Vierecke:



Für die zu bildenden Integrale ergibt sich nach (3.44)

$$\begin{aligned} \int_e (u_x^2 + u_y^2) dx\, dy &= \int_e (u_r^2 + \frac{1}{r^2} u_\varphi^2)\, r\, dr\, d\varphi \\ \int_e u^2\, dx\, dy &= \int_e u^2\, r\, dr\, d\varphi \\ \int_e u dx\, dy &= \int_e u\, r\, dr\, d\varphi \end{aligned} \tag{3.48}$$

Wegen der Unstetigkeit der Polarkoordinaten im Nullpunkt, die sich auch im 'Schrumpfen' des Vierecks in ein Dreieck für $r_1 = 0$ zeigt, müssen Viereck- und Dreieckfall getrennt behandelt werden.

Das Viereck $\{0 < r_1 \le r \le r_2, \quad 0 \le \varphi_1 \le \varphi \le \varphi_2 \le 2\,\pi\}$

Für das erste Integral aus (3.48) wird die Variablentransformation

$$\begin{aligned} r &= r_1 + (r_2 - r_1)\,\xi \\ \varphi &= \varphi_1 + (\varphi_2 - \varphi_1)\,\eta \\ \frac{\partial \xi}{\partial r} &= \frac{1}{r_2 - r_1} \\ \frac{\partial \eta}{\partial \varphi} &= \frac{1}{\varphi_2 - \varphi_1} \end{aligned} \tag{3.49}$$

benutzt. Damit ergibt sich:

$$\begin{aligned} \int_e (u_r^2 + \frac{1}{r^2} u_\varphi^2)\, r\, dr\, d\varphi &= \frac{\varphi_2 - \varphi_1}{r_2 - r_1} \int_0^1 \int_0^1 (r_1 + (r_2 - r_1)\,\xi)\, u_\xi^2\, d\xi\, d\eta \\ &+ \frac{r_2 - r_1}{\varphi_2 - \varphi_1} \int_0^1 \int_0^1 \frac{u_\eta^2}{r_1 + (r_2 - r_1)\,\xi}\, d\xi\, d\eta \end{aligned} \tag{3.50}$$

Alle Integrale werden numerisch ausgewertet. (Das zweite Integral ist z.B. gebrochen rational in ξ.)

Entsprechend ergeben sich für die anderen Integrale aus (3.48):

$$\int_e u^\alpha\, r\, dr\, d\varphi = (r_2 - r_1)(\varphi_2 - \varphi_1) \int_0^1 \int_0^1 (r_1 + (r_2 - r_1)\,\xi)\, u^\alpha\, d\xi\, d\eta, \quad \alpha = 1, 2. \tag{3.51}$$

Ein bilinearer Ansatz

$$N(\xi, \eta) = \alpha_1 + \alpha_2\,\xi + \alpha_3\,\eta + \alpha_4\,\xi\,\eta \tag{3.52}$$

hat passend zu den Ecken den Freiheitsgrad 4.

Das Dreieck $\{0 = r_1 \le r \le r_2, \quad 0 \le \varphi_1 \le \varphi \le \varphi_2 \le 2\,\pi\}$

Durch Grenzübergang $r_1 \to 0$ ergibt sich

$$\int_e (u_r^2 + \frac{1}{r^2} u_\varphi^2)\, r\, dr\, d\varphi = (\varphi_2 - \varphi_1) \int_0^1 \int_0^1 \xi\, u_\xi^2\, d\xi\, d\eta + \frac{1}{\varphi_2 - \varphi_1} \int_0^1 \int_0^1 \frac{u_\eta^2}{\xi}\, d\xi\, d\eta \tag{3.53}$$

Der unangenehme Faktor $\frac{1}{\xi}$ kann durch einen Ansatz beseitigt werden, der ξ als Faktor in u_η enthält.

Die anderen beiden Integrale ergeben sich entsprechend.

Der Ansatz

$$N(\xi, \eta) = \alpha_1 + \alpha_2\,\xi + \alpha_3\,\xi\,\eta \tag{3.54}$$

hat passend zu den Ecken den Freiheitsgrad 3 und passt zu dem Ansatz (3.52). Ein Netz aus Dreiecken und Vierecken mit diesen Ansatzfunktionen liefert Stetigkeit der Basisfunktionen.

3.4.5 Dreidimensionale finite Elemente

Obwohl die Darstellung der Elemente und Ansatzfunktionen im $\mathbb{R}^3$ wesentlich komplexer und die rechnerische Durchführung viel aufwendiger wird, lassen sich einige der zweidimensionalen Ansätze auf drei Dimensionen übertragen.

Hier sollen im wesentlich drei geometrische Elementetypen grafisch gezeigt und die möglichen Ansätze auf ihnen erwähnt werden.

Tetraeder-Element

Der lineare Ansatz im Dreieck lässt sich direkt auf einen linearen Ansatz auf einem Tetraeder verallgemeinern:

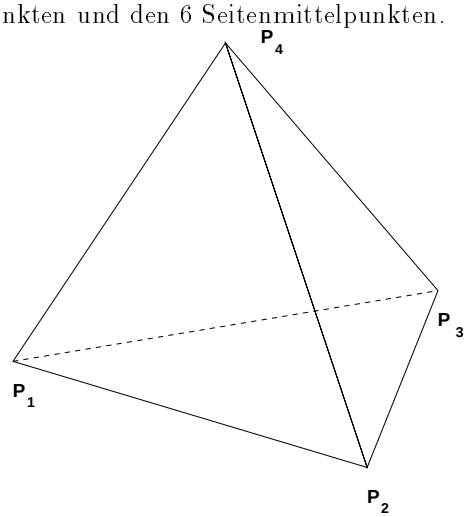
$$N(x,y,z) = \alpha_1 + \alpha_2\, x + \alpha_3\, y + \alpha_4\, z \tag{3.55}$$

Die vier Freiheitsgrade entsprechen den vier Eckpunkten des Tetraeders.

Auch der quadratische Ansatz im Dreieck lässt sich auf den Tetraeder übertragen. Die 10 Freiheitsgrade von

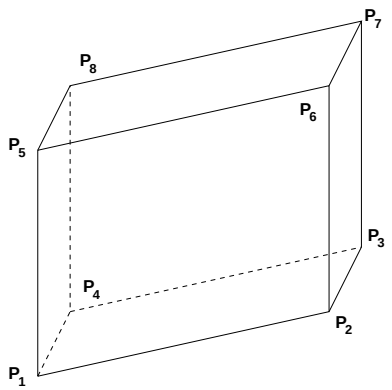
$$N(x,y,z) = \alpha_1 + \alpha_2\, x + \alpha_3\, y + \alpha_4\, z + \alpha_5\, x^2 + \alpha_6\, x\ y + \alpha_7\, x\, y + \alpha_8\, y^2 + \alpha_9\, y\, z + \alpha_{10}\, z^2 \tag{3.56}$$

entsprechen den 4 Eckpunkten und den 6 Seitenmittelpunkten.



Parallelepiped-Element

Auch die Parallelogramm-Ansätze des $\mathbb{R}^2$ lassen sich auf den $\mathbb{R}^3$ übertragen. Die im $\mathbb{R}^2$ notwendige Transformation vom Paralelogramm auf das Einheitsquadrat entspricht hier der entsprechenden Transformation des Parallelepipeds auf den Einheitswürfel.



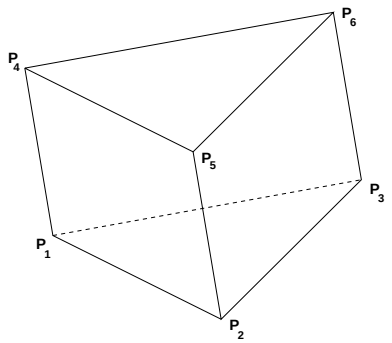
Die 8 Freiheitsgrade des trilinearen Ansatzes

$$N(x,y,z) = \alpha_1 + \alpha_2\, x + \alpha_3\, y + \alpha_4\, z + \alpha_5\, x\, y + \alpha_6\, x\, z + \alpha_7\, y\, z + \alpha_8\, x\, y\, z \tag{3.57}$$

entsprechen den 8 Eckpunkten des Parallelepipeds.

Prismen-Element

Prismen mit dreieckigerr Grundfläche lassen sich sowohl mit Tetraedern als auch mit Parallelepipeden kombinieren.



Die 6 Freiheitsgrade des Ansatzes

$$N(x,y,z) = \alpha_1 + \alpha_2\, x + \alpha_3\, y + \alpha_4\, z + \alpha_5\, x\, z + \alpha_6\, y\, z \tag{3.58}$$

stimmen mit den 6 Ecken überein und garantieren die Stetigkeit der Basisfunktionen, wenn dieser Prismen-Ansatz mit dem linearen Ansatz auf Tetraedern und dem trilinearen Ansatz auf Parallelepipeden kombiniert wird.

3.5 "Optimale" Numerierung der Knotenpunkte

Der Aufwand zur Lösung des Gleichungssystems $S\beta = t$ hängt sehr von der Numerierung der Knotenpunkte ab, deshalb gibt es verschiedene Verfahren, die diese Numeriering nach unterschiedlichen Gesichtspunkten und möglichst automatisch optimieren.

Die Ziele bei diesen Verfahren können u.a. sein:

- Minimierung des fill-in bei der Lösung des Gleichungssystems mit einem direkten Verfahren, also z.B. einer $L\,R$-Zerlegung mit dem Gauß- oder dem Cholesky-Verfahren.
- Minimierung der Bandbreite $2m+1$ (Halbbandbreite m) der Matrix. Die Minimierung der Bandbreite oder des Profils (siehe unten) ist auch für minimalen Speicheraufwand sehr wichtig.
- Minimierung des Profils, das wir weiter unten definieren werden.

Neben dem Ziel des Verfahrens unterscheidet man zwischen

- lokalen Verfahren, die eine (Um-)Numerierung während der Eliminationsschritte eines direkten Verfahrens jeweils für die Restmatrix vornehmen, und
- globalen Verfahren, die die Punkte vor der Lösung des linearen Gleichungssystems so gut wie möglich numerieren.

Rose und Tarjan haben 1978 für den unsymmetrischen und Yannakakis 1981 für den symmetrischen Fall gezeigt, daß das Problem der optimalen Numerierung NP-vollständig ist, wie etwa das berühmtere Problem des Handlungsreisenden (TSM-Problem).

Daher arbeiten die verschiedenen Algorithmen alle nur nach heuristischen Gesichtspunkten, deterministische Methoden mit beweisbar optimalem Ergebnis gibt es nicht.

Im Schwarz, [Sch90], findet man i.w. zwei Algorithmen, den von Cuthill/McKee, den wir unten vorstellen wollen und der ein Gebiet ohne vorgegebene Numerierung möglichst günstig numeriert, und den von Rosen, der zu einer gegebenen Numerierung, die schon nahe am Optimum sein sollte, eine Verbesserung liefert, für nicht nahe dem Optimum gelegene Numerierungen aber aufwendig und manchmal auch erfolglos ist. Eine ausführlichere Behandlung des Problems findet man im Duff/Erisman/Reid, [DER86].

3.5.1 Graphentheoretische Grundlagen

Graph G = Menge von n Knoten und Kanten.

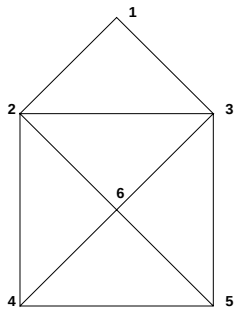
Knoten $x_i, i = 1, \ldots, n$.

Ungerichtete Kanten (x_i, x_j), keine Schleifen: (x_i, x_i).

Verknüpfungsmatrix $V = (v_{ij})$, $v_{ij} = \begin{cases} 1, \text{ falls } (x_i, x_i) \in G \\ 0, \text{ falls } (x_i, x_j) \notin G \end{cases}$

Beispiel 3.5.1

$$V = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ & 0 & 1 & 1 & 0 & 1 \\ & & 0 & 0 & 1 & 1 \\ & & & 0 & 1 & 1 \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix}$$



Benachbarte Knoten: $x_i, x_j \Leftrightarrow v_{ij} = 1$

Verbundene Knoten: $x_i, x_j \Leftrightarrow \exists i_1, i_2, \ldots, i_m : v_{i,i_1} = v_{i_1,i_2} = \cdots = v_{i_m,j} = 1$

Grad m eines Knoten x_i : Anzahl Kanten mit $(x_i, x_{j_l}) \in G, l = 1, \ldots, m$

Graph einer Matrix A: G (A)

Jeder symmetrischen Matrix A kann man einen Graphen $G(A)$ zuordnen, indem man
$a_{ij} \neq 0$ durch $v_{ij} := 1$ und
$a_{ij} = 0$ durch $v_{ij} := 0$ ersetzt,
$v_{ii} := 0$ setzt und
den der Verknüpfungsmatrix V zugeordneten Graphen nimmt. Grad eines Knotens x_i ist dann die Anzahl der Elemente $a_{ij} \neq 0$ außerhalb der Diagonale, im Beispiel oben also $\{2, 4, 4, 3, 3, 4\}$. Die Diagonalelemente spielen keine Rolle, weil es keine Schleifen gibt. Ein Baum ist ein Graph mit einem Knoten mehr als Kanten und ohne isolierten Knoten.

3.5.2 Algorithmus von Cuthill/McKee

Geg.: $S \in \mathbb{R}^{N,N}$ und $G(S)$.

Dazu ist eine Vornumerierung schon notwendig. Wie wir unten sehen, kann man aber auch die erste Numerierung der Knotenpunkte schon nach den Grundsätzen des Algorithmus vornehmen.

3.5.2 Algorithmus von CUTHILL/MCKEE

for $m = 1$ **to** n = Anzahl Knoten mit minimalem Grad **do**
begin
  Setze Level $l = 0$
  Level-Set[†] $S_{ml} = \{1\}$ = Nummer des Knoten minimalen Grades
  Nächste-Knoten-Nummer $k = 2$
  while Noch-Knoten-vorhanden
  begin
    Definiere die Menge der Nachbarn des Level-Sets l
    Sortiere diese Nachbarn nach aufsteigendem Grad und
    vergib die Nummern k bis r $(*)$
    Setze $l = l + 1$
    Level-Set $S_{ml} = \{k, \ldots, r\}$
    Nächste-Knoten-Nummer $k = r + 1$
  end
end
comment
  Es sind jetzt n Numerierungen mit unterschiedlichen Anzahlen
  von Level-Sets konstruiert. Davon muss eine ausgewählt werden.
  Wähle denjenigen Knoten minimalen Grades, der die größte Anzahl Level-Sets
  und die kleinste Maximalzahl von Knoten in einer Level-Set liefert,
  und numeriere dementsprechend.

Bemerkung 3.5.3 (i) Die kleinste Bandbreite innerhalb der durchgerechneten Mengen liefert genau die Numerierung nach dem letzten Kriterium. Es kann natürlich mehrere Knoten geben, die dieses Kriterium erfüllen, ebenso wie die Regel $(*)$ noch Entscheidungen offenläßt bei mehreren Knoten gleichen Grades. Es ist eine Aufwandsfrage, dies weiter mit Schleifen zu überprüfen.

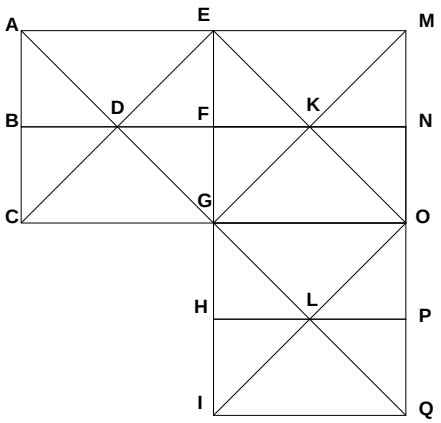
(ii) Es ist möglich, daß ein Knoten nicht minimalen Grades einen Knoten Nr. 1 mit kleinerer Bandbreite im Ergebnis abgibt. Das kann im Algorithmus berücksichtigt werden, siehe 3.5.4.

(iii) Knoten mit vielen Nachbarn sollen einen möglichst hohen Index erhalten, damit die Indexdifferenzen, die die Bandbreite bestimmen, im nächsten Schritt möglichst klein sind.

(iv) Der Algorithmus von CUTHILL/MCKEE liefert oft eine Knotennumerierung, die fast optimal ist. Sie ist daher eine gute Startnäherung für das Verfahren von ROSEN (siehe [Sch90]).

[†]Deutsch: Stufen-Menge!

Beispiel 3.5.4 Im Graphen dieses Beispiels gibt es viele Knoten mit minimalem Grad, zwei davon wollen wir vergleichend durchrechnen, das ist also nicht der vollständige Algorithmus von Cuthill/McKee.

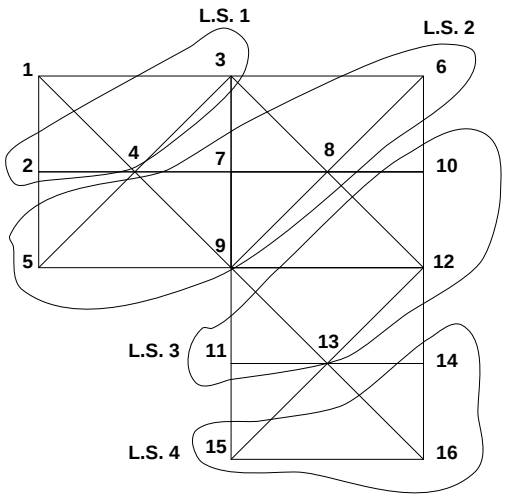


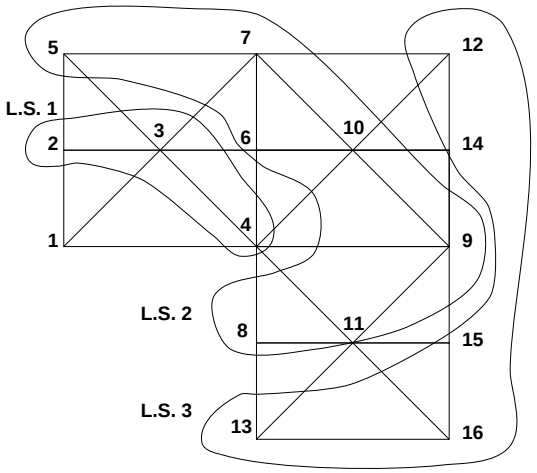
Minimaler Grad = 3

Knoten mit Grad 3:

A, B, C, H, I, M, N, P, Q

Wir betrachten als Startknoten die Punkte A und C:





Ablauf für Startknoten A

Schritt	Knoten	Nachbarknoten	Grad	Nummer
1	1 (A)	B	3	2
		D	6	4
		E	5	3
2	2 (B)	C	3	5
	3 (E)	F	4	7
		K	6	8
		M	3	6
	4 (D)	G	7	9
3	5 (C)	-	-	-
	6 (M)	N	3	10
	7 (F)	-	-	-
	8 (K)	O	5	12
	9 (G)	H	3	11
		L	6	13
4	10 (N)	-	-	-
	11 (O)	P	3	14
	12 (H)	I	3	15
	13 (L)	Q	3	16

Maximaler Grad ist 7, deshalb muß die Bandbreite $2m + 1 \geq 9$ sein (3 links, 4 rechts). In Wirklichkeit ergibt sich für:

Startknoten A: $2m + 1 = 11$ (Kanten (3,8), (4,9))

Startknoten C: $2m + 1 = 15$ (Kante (4,11))

Allgemein ist die Bandbreite meist kleiner, wenn die Stufenzahl größer ist. Deswegen kann man auch im Algorithmus oben ausschließlich nach der mit dem jeweiligen Startknoten erreichten Stufenzahl auswählen.

3.5.3 Umgekehrter Cuthill/McKee-Algorithmus

Wir betrachten zunächst statt der Bandbreite m das Profil p der Matrix S:

Definition 3.5.5 Sei

$$\begin{aligned} f_i(S) &= \min\{j \mid s_{ij} \neq 0,\ j \leq i\} \\ m_i(S) &= i - f_i(S) \quad \text{[linksseitige Bandbreite der } i.\text{Zeile]} \end{aligned}$$

Dann ist

$$m = \max_{1 \leq i \leq N} m_i(S)$$

die Halbbandbreite und

$$p := N + \sum_{i=1}^{N} m_i(S)$$

das Profil von S.

Das Profil erlaubt genauere Aussagen über den Aufwand gewisser Verfahren:

Kennt man die Werte $m_i(S)$, so kann man Speicherung und Rechnung auf die p Matrixelemente in der Hülle

$$\text{Env } (S) = \{(i,j) \mid f_i(S) \leq j \leq i, \quad 1 \leq i \leq N\}$$

beschränken mit

$$(p = |\text{Env } (S)|)$$

(z.B. beim Cholesky - Algorithmus).

Der umgekehrte CM-Algorithmus RCM entsteht durch Umnumerierung nach Durchführung des CM-Algorithmus:

Knoten $k \rightarrow$ Knoten $N - k + 1$.

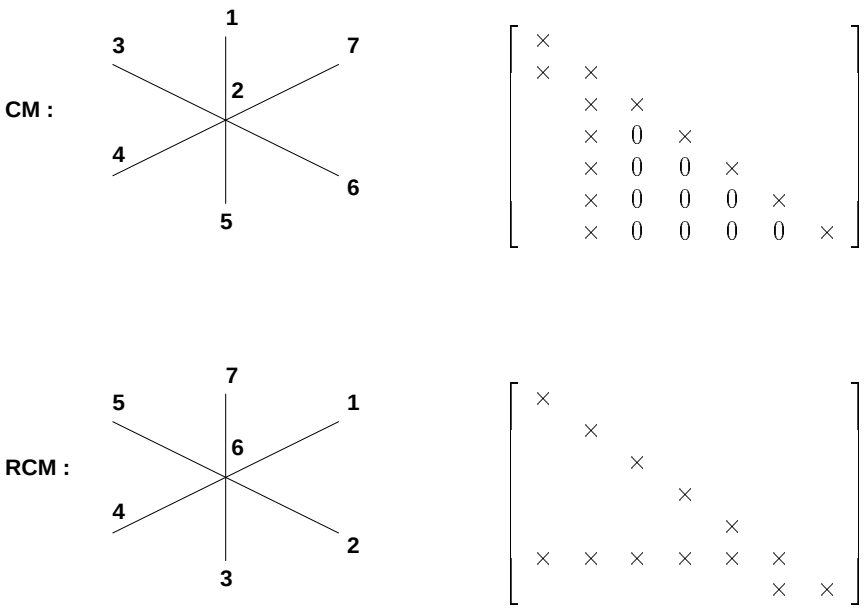
Man kann zeigen, daß für die Profile

$$p_{RCM} \leq p_{CM}$$

gilt. Die Bandbreite m ändert sich ohnehin nicht. In einer Reihe von Experimenten wurde gezeigt, daß oft

$$p_{RCM} < p_{CM}.$$

Beispiel 3.5.6 Ein weiterer Vorteil des RCM-Algorithmus ist, daß er oft weniger Speicherplatz zur Speicherung des Profils benötigt als der CM-Algorithmus. Dies sieht man sehr schön an einem einfachen Beispiel aus [DER86], das allerdings für unsere Anwendungen unrealistisch ist. Dabei sind alle zu speichernden Elemente angezeigt, $\times$ steht für Elemente ungleich Null. Man speichert also beim CM 10 überflüssige Nullen mehr als beim RCM.



3.5.4 Strategien zur Auswahl des Startknotens

Wie gut eine Numerierung ist, hängt stark davon ab, wieviele und welche Startknoten der Algorithmus ausprobiert. Da das Durchrechnen vieler Startknoten andererseits aber selbst einen hohen Aufwand darstellt, gibt es hierzu verschiedene strategische Vorschläge:

- Original CUTHILL/MCKEE: Probiere alle Startknoten kleinsten Grades durch.

- COLLINS: Probiere alle Knoten durch, brich aber den Algorithmus jeweils ab, wenn die Bandbreite größer als die bisher erreichte maximale Bandbreite ist.

- GIBBS, POOLE, STOCKMEYER:

  (1) Beginne mit einem beliebigen Knoten als Startknoten. Das liefert ein letztes Level-Set S_l.

  (2) Nimm alle Knoten aus dem Level-Set S_l als Startknoten. Liefert einer dieser Startknoten eine größere Zahl von Level-Sets, also ein letztes Level-Set $S_{\tilde{l}}$ mit $\tilde{l} > l$, so ist dies der neue Startknoten für den ersten Schritt.

  (3) Führe (1)/(2) durch, bis mit den Startknoten aus dem letzten Level-Set kein letztes Level-Set mit einem größeren l mehr entsteht.

Variationen dieser Möglichkeiten liegen auf der Hand.

3.6 Die FEM bei instationären Problemen

Wir wollen uns auf die instationäre Wärmeleitung beschränken.

$$\text{PDE:} \qquad \Delta u - f(x,y,t) - \kappa \frac{\partial u}{\partial t} = 0 \quad \text{in } \Omega. \tag{3.59}$$

$$\text{RB I:} \qquad u(x,y,t) = \varphi(x,y,t) \quad \forall (x,y) \in C_1,\ \forall t > 0 \tag{3.60}$$

$$\text{RB II:} \qquad \frac{\partial u}{\partial \mathfrak{n}} + \alpha(x,y)u(x,y,t) = \gamma(x,y,t) \quad \forall (x,y) \in C_2,\ \forall t > 0 \tag{3.61}$$
$$C_1 \cup C_2 = \partial\Omega,\ C_1 \cap C_2 = \phi$$

$$\text{AB:} \qquad u(x,y,0) = u_0(x,y) \quad \text{in } \Omega. \tag{3.62}$$

1. Schritt: Bestimme eine Funktion $\varphi_0(x,y,t)$, die die RB (3.60) und (3.61) erfüllt.

2. Schritt: Bestimme einen Basisfunktionenansatz $\{\varphi_i(x,y)\}_{i=1}^N$, mit $\varphi_i \in H_0^1$, also $\varphi_i = 0$ auf $\partial\Omega$.

Damit bekommt man den Lösungsansatz

$$u(x,y,t) = \varphi_0(x,y,t) + \sum_{i=1}^{N} c_i(t)\varphi_i(x,y) \tag{3.63}$$

Dies ist also ein Ansatz im Raum, dessen Koeffizienten noch zeitabhängig sind. Man nennt das auch Separationsansatz. Man bekommt für verschiedene feste Zeitpunkte t_i ein Variationsprinzip, eine FEM, und muß bezüglich t eine zusätzliche Lösungsmethode überlagern. Dieses Variationsprinzip führt wie bei der 3. RWA, nur mit von t abhängigen Koeffizienten, zu

$$\begin{aligned}\sum_{i=1}^{N} \frac{dc_i}{dt} \int_\Omega \kappa\, \varphi_i\varphi_j \quad &+ \quad \sum_{i=1}^{N} c_i(t) \left[\int_\Omega D_1\varphi_i\, D_1\varphi_j + D_2\varphi_i D_2\varphi_j + \int_{C_2} \alpha\varphi_i\varphi_j \right] \\ &+ \quad \int_\Omega (D_1\varphi_0 D_1\varphi_j + D_2\varphi_0 D_2\varphi_j + [f + \kappa \frac{\partial\varphi_0}{\partial t}]\,\varphi_j) \\ &+ \quad \int_{C_2} [\alpha\varphi_0 - \gamma]\varphi_j = 0 \quad j = 1,\ldots,N\end{aligned}$$

Dies ist ein DGL-System 1. Ordnung mit N Gleichungen.

Sei

$$A = \{a_{ij}\}, \quad a_{ij} = \int_\Omega D_1\varphi_i D_1\varphi_j + D_2\varphi_i D_2\varphi_j + \int_{C_2} \alpha\varphi_i\varphi_j$$

$$B = \{b_{ij}\}, \quad b_{ij} = \int_\Omega \kappa\, \varphi_i\varphi_j$$

und

$$c(t) = (c_1(t), \ldots, c_N(t))^T,$$

$$d = (d_1, \ldots, d_N)^T \text{ mit}$$

$$d_j = \int_\Omega (D_1\varphi_0 D_1\varphi_j + D_2\varphi_0 D_2\varphi_j + [f + \kappa \frac{\partial \varphi_0}{\partial t}]\varphi_j) + \int_{C_2} [\alpha\varphi_0 - \gamma]\varphi_j \,,$$

dann ist

$$B\dot{c} + Ac + d = 0 \tag{3.64}$$

A, B sind symmetrisch, B pos. def., A, B konstante Matrizen. Nur die Inhomogenität d kann noch von der Zeit abhängen. Zur Lösung der DGL (3.64) benötigt man eine AB: Es muß wegen (3.62)

$$\varphi_0(x, y, 0) + \sum_{i=1}^{N} c_i(0)\varphi_i(x, y) = u_0(x, y) \text{ in } \Omega$$

sein. Diese Gleichung kann nicht für alle Punkte in Ω erfüllt werden. Man kann sie approximativ mit einer Galerkin-ähnlichen Methode lösen:

$$\sum_{i=1}^{N} c_i(0) \underbrace{\int_\Omega \varphi_i\varphi_j}_{B \text{ (bis auf } \kappa)} + \int_\Omega [\varphi_0(x, y, 0) - u_0(x, y)]\varphi_j = 0 \quad j = 1, 2, \ldots, N$$

Dies ist ein lineares Gleichungssystem für $c_i(0)$, das zunächst gelöst werden muß.

Anschließend wird die DGL mit der Trapezregel gelöst. Das ist für ein lineares DGL-System genau genug. Dazu setzen wir $t_i := i\,\Delta t, i = 0, 1, 2, \ldots$, mit einer Schrittweite Δt und formen die DGL in die Standardform um:

$$\dot{c} = B^{-1}(-Ac - d)$$

$\Rightarrow$

$$\begin{aligned} c(t_{i+1}) &= c(t_i) + \underbrace{\int_{t_i}^{t_{i+1}} B^{-1}\left(-Ac(\tau) - d(\tau)\right) d\tau}_{\text{Trapezregel}} \\ &\approx c(t_i) + \frac{\Delta t}{2} B^{-1}[-Ac(t_{i+1}) - Ac(t_i) - d(t_{i+1}) - d(t_i)] \\ c^{(i+1)} &= c^{(i)} - \frac{1}{2}\Delta t B^{-1}[Ac^{(i)} + Ac^{(i+1)} + d^{(i)} + d^{(i+1)}] \end{aligned}$$

Das schreibt man als Gleichungssystem

$$Bc^{(i+1)} = Bc^{(i)} - \frac{1}{2}\Delta t[\cdots] \quad \text{oder}$$

$$(B + \frac{1}{2}\Delta t A)c^{(i+1)} = (B - \frac{1}{2}\Delta t A)c^{(i)} - \frac{1}{2}\Delta t(d^{(i)} + d^{(i+1)}) \tag{3.65}$$

In jedem Schritt muß dieses Gleichungssystem gelöst werden, aber mit immer derselben i.d.R. positiv-definiten Koeffizientenmatrix $B + \frac{1}{2}\Delta t A$, d.h. wir zerlegen die Matrix einmal mit dem Cholesky-Verfahren und lösen dann mit Vor- und Rücksubstitution. (Aufwand: Einmal n^3 bzw. n^2 bei dünn besetzten Matrizen, x-mal n^2 bzw. n).

Die Lösung u_N ergibt sich dann als

$$u_N(x, y, t^{(n)}) = \varphi_0(x, y, t^{(n)}) + \sum_{i=1}^{N} c_i^{(n)} \varphi_i(x, y) \text{ mit } t^{(n)} = n\,\Delta t \tag{3.66}$$

Wir werden auf parabolische Probleme bei Differenzenverfahren im Kapitel 4 zurückkommen.

3.7 Konvergenz und Fehler

Konvergenz und Fehler der Finite-Elemente-Lösungen sollen nur für spezielle FEM-Ansätze untersucht werden. Dabei werden zunächst Probleme im $\mathbb{R}^1$ betrachtet, um dann die Ergebnisse auf lineare Ansätze auf Dreiecken im $\mathbb{R}^2$ zu verallgemeinern. Vorher stellen wir aber noch einige Aussagen über das Variationsproblem und seine Lösung im $\mathbb{R}^n$ zur Verfügung.

Wir identifizieren folgende Bezeichnung: Sei $u^h := u_S := u_S^{(n)}$ die Lösung des diskretisierten Variationsproblems. Dabei symbolisiert der Parameter h etwa eine Diskretisierungsschrittweite im $\mathbb{R}^1$ oder einen Elementedurchmesser im $\mathbb{R}^n$. Sei weiter

$$e^h := u - u^h. \tag{3.67}$$

Als Folgerung aus den Sätzen von Riesz 2.1.14 und Ritz 2.2.3 und mit den abstrakten Bezeichnungen dort erhalten wir:

Satz 3.7.1 Sei $u \in H$ die Lösung von (2.20): $B(u, v) = l(v) \qquad \forall v \in H$.

Sei $S \subset H$ ein abgeschlossener, linearer Unterraum. Dann gibt es genau ein $u_S \in S$, die Ritz-Näherungslösung, das die folgenden äquivalenten Bedingungen erfüllt:

(i) $A(u_S) = B(u_S, u_S) - 2\,l(u_S) = \min\limits_{w \in S} A(w)$

(ii) $\|u_S - u\|_E \leq \|w - u\|_E \qquad \forall w \in S$

(iii) $B(u_S, w) = l(w) \qquad \forall w \in S$

(iv) $B(u - u_S, w) = 0 \qquad \forall w \in S$

Bemerkung 3.7.2 (i) definiert die Ritz-Approximation, siehe Satz 2.1.14.

(ii) zeigt, daß die Ritz-Approximation die Best-Approximation an u in der $\|\cdot\|_E$ ist, siehe Lemma 2.2.2.

(iii) zeigt, daß die "Ableitung" von $A|_S$ in u_S verschwindet.

(iv) sagt, daß man u_S erhält, wenn man von u auf S das Lot fällt (das Energielot).

Beweis: Zunächst gilt nach Riesz, daß $B(u_S, w) = l(w)$ eine eindeutig Lösung hat, die eindeutige Lösung des Minimalproblems ist, weil ja auch S Hilbertraum ist, also haben wir schon
(iii) $\Rightarrow$ (i). ■

Es gilt (zu zeigen wie am Schluß des Beweises von 2.1.14 (ii))

$$\begin{aligned} A(v) - A(w) &= 2B(w, v-w) - 2l(v-w) + B(v-w) \\ \Rightarrow \text{ mit } w = u & \\ A(v) - A(u) &= B(v-u) \qquad \forall v \in H \\ &= \|v-u\|_E^2 \end{aligned}$$

$(2(B(u, v-u) - l(v-u)) = 0$ wegen (2.20).)

$A(u)$ ist fester Wert, also wird die Differenz $A(v) - A(u)$ minimal, wenn $A(v)$ minimal wird, d.h. in $A(u_S)$. Also wird die linke und damit auch die rechte Seite minimal auf S für $v = u_S$, also sind (i) und (ii) äquivalent. ■

Wir verwenden jetzt wieder ein Variationsprinzip für $u_S \in S, \varepsilon \in \mathbb{R}, v \in S$:

$$\begin{aligned} A(u_S + \varepsilon v) &= B(u_S + \varepsilon v) - 2l(u_S + \varepsilon v) \\ &= B(u_S) + 2\varepsilon\, B(u_S, v) + \varepsilon^2 B(v) - 2l(u_S) - 2\varepsilon l(v) \\ &= A(u_S) + 2\varepsilon[B(u_S, v) - l(v)] + \varepsilon^2 B(v) \end{aligned}$$

u_S kann nur Minimum von $A|_S$ sein, wenn in S

$$\frac{d}{d\varepsilon} A(u_S + \varepsilon v)|_{\varepsilon=0} = 2[B(u_S, v) - l(v)] = 0 \quad \forall v \in S$$

Also gilt (i) $\Rightarrow$ (iii). Mit (iii) gilt

$$B(u_S, w) = l(w) = B(u, w)$$

$$\Leftrightarrow B(u_S - u, w) = 0 \quad \forall w \in S$$

Also sind (iii) und (iv) äquivalent. Damit sind die Äquivalenzen gezeigt. Existenz und Eindeutigkeit der Lösung u_S in S folgt aus dem Rieszschen Darstellungssatz. ■

Lemma 3.7.3 (i) Sei u^h das Minimum von A über S, dann gilt:

$$B(u - u^h, u - u^h) = \min_{v \in S} B(u - v, u - v).$$

(ii) $B(u - u^h, u^h) = 0$
oder in Äquivalenz dazu:
$B(u, u^h) = B(u^h, u^h) \;\Rightarrow\; B(u, u^h) - B(u^h, u^h) = 0.$

(iii) $B(e^h, e^h) = B(u, u) - B(u^h, u^h)$ (Pythagoras).

(iv) $B(u^h, u^h) \leq B(u, u)$.

Beweis: Übung!

3.7.1 Lineare FEM im $\mathbb{R}^1$

Wir wollen zunächst beispielhaft lineare FEM im $\mathbb{R}^1$ betrachten, das sind Hutfunktionen. Mit ihnen wollen wir das folgende RWP lösen:

$$\begin{aligned} -(a_{11}(x)\,u'(x))' + c(x)\,u(x) &= f(x), \quad x \in (a,b) \\ u(a) = u(b) &= 0 \end{aligned} \tag{3.68}$$

Dabei seien $a_{11} > 0$ und $c \geq 0$. Wir wählen ein äquidistantes Gitter $x_i = a + i\,h$, $i = 0, 1, \ldots, n$, also $h = (b-a)/n$. Diesem RWP können wir das Variationsfunktional A und die Bilinearform B zuordnen:

$$A(u) = \int_a^b [a_{11}u'^2 + cu^2 - 2fu]dx, \tag{3.69}$$

$$B(u,v) = \int_a^b a_{11}u'v' + cuv, \tag{3.70}$$

$$\text{in} \qquad \begin{aligned} S_h &= \left\{ \varphi \in C[a,b] \;\middle|\; \varphi|_{[x_i,x_{i+1}]} \in \Pi_1, \varphi(a) = \varphi(b) = 0 \right\}, \\ S_h &\in H_0^1[a,b]. \end{aligned} \tag{3.71}$$

Fragen:

(1) Approximationsgüte: $u - u^h$?

(2) Konvergenz $u^h \to u$ mit $h \to 0$?

(3) Fehlerschranken in Abhängigkeit von h?

Nach Lemma 3.7.3 erhält man eine Abschätzung von e^h in der Energienorm, wenn man die rechte Seite $B(u-v, u-v)$ für irgendein $v \in S^h$ berechnen kann. Das Problem der Fehlerabschätzung ist damit in ein Approximationsproblem überführt. Eine Funktion aus S^h ist die Interpolierende u_I:

$$u_I(x_i) = u(x_i) \quad i = 0, 1, \ldots, n$$

$u_I|_{[x_i,x_{i+1}]}$ ist linear.

u_I kann mit Hilfe der Hutfunktionen dargestellt werden: Wir wählen o.B.d.A.:

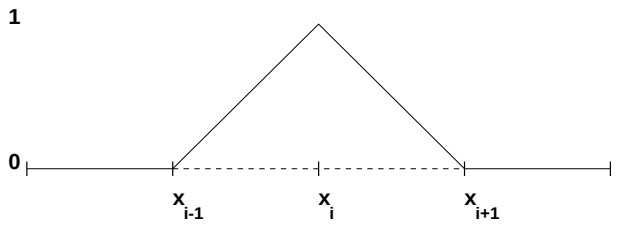
$$[a,b] := [0,1] \text{ und } x_i = i\,h, i = 0, 1, \ldots, n$$

also $h = \dfrac{1}{n}$. Sei noch $N := n - 1$. Dann ist für $i = 1, \ldots, N$:

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h} & x \in [x_{i-1}, x_i] \\ \dfrac{x_{i+1} - x}{h} & x \in [x_i, x_{i+1}] \\ 0 & \text{sonst.} \end{cases}$$

und es gilt

$$u_I(x) = \sum_{j=1}^{N} u(j\,h)\,\varphi_j(x) \tag{3.72}$$



Satz 3.7.4 Ist $u'' \in H^0$, so gilt

(i) $\|u - u_I\|_0 \le \dfrac{1}{\pi^2} h^2 \|u''\|_0$.

(ii) $\|u' - u_I'\|_0 \le \dfrac{h}{\pi} \|u''\|_0$.

(iii) $B(u - u_I, u - u_I) \le \left(\dfrac{h^2}{\pi^2} |a_{11}|_0 + \dfrac{h^4}{\pi^4} |c|_0 \right) \|u''\|_0^2$.
($|\cdot|_0$ ist die sup-Norm.)

Beweis: (i) In $0 \le x \le h$ sei

$$\delta(x) := u(x) - u_I(x), \text{ also } \delta(0) = \delta(h) = 0.$$

δ ist als Fourierreihe mit sin-Funktionen darstellbar:

$$\delta(x) = \sum_{1}^{\infty} a_j \sin \frac{j\pi x}{h} \Rightarrow \delta''(x) = -\sum a_j \frac{j^2\pi^2}{h^2} \sin \frac{j\pi x}{h}.$$

Beide Summen kann man exakt integrieren. Dabei beachtet man, daß

$$\int_0^h \sin \frac{i\pi x}{h} \sin \frac{j\pi x}{h} = \begin{cases} 0 & \text{für } i \ne j, \\ \dfrac{1}{2}\, x + \sin \dfrac{j\pi x}{h} \Big|_0^h = \dfrac{h}{2} & \text{für } i = j. \end{cases}$$

Dann folgt mit der Parsevalschen Gleichung

$$\int_0^h \delta^2(x)\,dx = \frac{h}{2} \sum_{1}^{\infty} a_j^2 = \|\delta\|_0^2$$

$$\int_0^h \delta''^2(x)\,dx = \frac{h}{2} \sum_{1}^{\infty} a_j^2 \frac{j^4\pi^4}{h^4} = \|\delta''\|_0^2$$

Summation über j liefert mit $a_j^2 \leq \left(\frac{h^4}{\pi^4}\right)\left(\frac{j\pi}{h}\right)^4 a_j^2$

$$\int_0^h \delta^2 \leq \frac{h^4}{\pi^4}\int_0^h \delta''^2 = \frac{h^4}{\pi^4}\int_0^h u''^2 dx \quad (\text{da } u_I \text{ in } (0,h) \text{ linear}).$$

Der Schluß gilt für jedes Intervall, Stetigkeit von u_I erlaubt die Summation zu:

$$\int_0^1 \delta^2 \leq \frac{h^4}{\pi^4}\int_0^1 u''^2 \Rightarrow (i)$$

Beweis von (ii), (iii) $\rightarrow$ Ü. ■

Aus den Sätzen 3.7.3 (i) und 3.7.4 (iii) folgt:

$$B(e^h, e^h) \leq c_1 h^2 \|u''\|_0^2 \tag{3.73}$$

Für den betrachteten Differentialoperator L gilt, daß die Lösung stetig von der rechten Seite f abhängt im Sinne von

$$\|u\|_2 \leq c\|f\|_0$$

(bei homogenen Randwerten).

Damit liefert (3.73)

$$B(e^h, e^h) \leq c_2 h^2 \|f\|_0^2. \tag{3.74}$$

Damit haben wir eine h^2-Schranke für das Quadrat der Energienorm des Fehlers. In den Konvergenzbeweisen wurde 2-malige Differenzierbarkeit von u vorausgesetzt. Das schließt $u \in H_1$, z.B. $\delta-$ Funktionen, aus. Man kann aber zeigen

Satz 3.7.5 Ist $u \in H_0^1$, so gilt

$$B(e^h, e^h) \rightarrow 0 \text{ mit } h \rightarrow 0.$$

Beweis:

$$u \in H_0^1 \Rightarrow \exists\{u_m\} \in C_0^\infty \text{ mit } \|u_m\|_E \rightarrow \|u\|_E.$$

Sei m fest und u_m^h die FEM-Approximation an u_m.

Dann gilt: $u_m^h \rightarrow u_m$ für $h \rightarrow 0$ in $\|\cdot\|_E$.

Also gibt es für jedes $u \in H_0^1$ ein N_0, so daß $u_{N_0} \in C_0^\infty$ und $\|u - u_{N_0}\|_E \leq \frac{\varepsilon}{2}$ und ein h_0, so daß $\|u_{N_0} - u_{N_0}^{h_0}\|_E \leq \frac{\varepsilon}{2}, u_{N_0}^{h_0} \in S^h$

$\Rightarrow \|u - u_{N_0}^{h_0}\|_E \leq \varepsilon$

Lemma 3.7.3 liefert $B(u-u^h, u-u^h) \leq B(u - u_{N_0}^{h_0}, u - u_{N_0}^{h_0})$ ■

Die FEM konvergiert also auch, wenn die rechte Seite f nicht stetig ist.

Fragen:

(i) Konvergenz in anderen Normen, z.B. der L^2-Norm?
(ii) Konvergenzordnung?

Eine L^2-Abschätzung ist schwierig, da die Best-Approximations-Eigenschaft nicht für die L^2-Norm gilt. Es gilt also nicht, daß u^h mindestens so dicht an u liegt wie u_I in der L^2-Norm.

Zur Überwindung dieser Schwierigkeiten gibt es zwei Möglichkeiten:

(i) Ausnutzung der speziellen Struktur des Beispiels:
(1-dimensionale, lineare Elemente)
Man berechnet den Abbruchfehler wie beim Differenzenverfahren und überträgt auf die FEM-Gleichungen.

Nachteil: Nur schwer übertragbar auf höhere Dimensionen und krummlinige Elemente.

(ii) Ausnutzung des Variationsprinzip (Nitsches Trick):

Satz 3.7.6 Sei u^h die lineare FEM-Lösung von $Lu = f$ und sei $u'' \in H^0$. Dann gilt:

$$\|u - u^h\|_0 \leq Ch^2\|u''\|_0 \leq \tilde{C}h^2\|f\|_0.$$

Beweis: Ist u Lösung von

$$B(u,v) = (f,v) \quad \forall v \in H_0^1$$

so gilt wegen der Stabilität (stetige Abhängigkeit von f, s.o.)

$$\|u\|_2 \leq C\|f\|_0 = C\|f\|_{L^2}$$

Sei S^h endlich-dimensionaler Teilraum von H_0^1.
Wir betrachten folgendes Problem:

Finde $z \in H_0^1$ mit $B(z,v) = (u - u^h, v) \quad \forall v \in H_0^1$.

Wähle speziell $v = u - u^h = e^h$:

$$\begin{aligned}
\|e^h\|_0^2 &= (e^h, e^h) = |B(z, e^h)| \\
&= |B(z - z^h, e^h)| \quad \forall z^h \in S^h, \\
&\quad [\text{da ja } B(z^h, e^h) = 0 \text{ nach Satz 3.7.1 (iv).}] \\
&= |B(z - z_I, e^h)| \quad (\text{da } z_I \in S^h) \\
&\leq \sqrt{B(z - z_I, z - z_I)}\sqrt{B(e^h, e^h)} \\
&\leq hC\|z''\|_0\sqrt{B(e^h, e^h)} \quad \text{nach Satz 3.7.4 (iii)} \\
&\leq hC\|z\|_2\sqrt{B(e^h, e^h)} \\
&\leq h\tilde{C}\|e^h\|_0\sqrt{B(e^h, e^h)} \quad (\text{Stabilität mit } e^h \text{ statt } f \text{ und } z \text{ statt } u) \\
\Rightarrow & \\
\|e^h\|_0 &\leq h\tilde{C}\sqrt{B(e^h, e^h)} \\
&\leq_{(3.73)} C_2h^2\|u''\|_0 \quad \blacksquare
\end{aligned}$$

Jetzt wollen wir noch betrachten, wie sich das Fehlerverhalten ändert, wenn statt der rechten Seite f deren Interpolierende f_I verwendet wird (numerische Integration).

Satz 3.7.7 Sei $f'' \in H^0$ und $\tilde{u}^h$ die FEM-Lösung zu $Lu = f_I$, dann gilt

$$B(u^h - \tilde{u}^h, u^h - \tilde{u}^h) \leq Kh^4 \|f''\|_0^2$$

Beweis:

$$\begin{aligned}
\|u - \tilde{u}\|_2 &\leq C\|f - f_I\|_0 \qquad \text{(Stabilität)} \\
\|f - f_I\|_0 &\leq \frac{h^2}{\pi^2}\|f''\|_0 \qquad \text{(Satz 3.7.4)} \\
\Rightarrow \|u - \tilde{u}\|_2 &\leq \tilde{C}h^2\|f''\|_0 \\
&\text{und} \\
B(u - \tilde{u}, u - \tilde{u}) &= (f - f_I, u - \tilde{u}) \\
&\leq \|f - f_I\|_0 \|u - \tilde{u}\|_0 \leq \|f - f_I\|_0 \|u - \tilde{u}\|_2 \\
&\leq \frac{h^2}{\pi^2}\|f''\|_0 \tilde{C}h^2\|f''\|_0 \\
&= K_2 h^4 \|f''\|_0^2 \\
\Rightarrow B(u^h - \tilde{u}^h, u^h - \tilde{u}^h) &\leq B(u - \tilde{u}, u - \tilde{u}) \quad \text{(wegen 3.7.3 (iv))} \\
&\leq \bar{C}h^4\|f''\|_0^2
\end{aligned}$$

u^h bzw. $\tilde{u}^h$ sind Ritz-Approximationen an u bzw. $\tilde{u}$. Wegen der Linearität der Probleme ist dann $u^h - \tilde{u}^h$ Ritz-Approximation an $u - \tilde{u}$. Deshalb kann man 3.7.3 (iv) anwenden. ■

3.7.2 Fehlerabschätzungen für lineare Dreieckselemente

Für lineare Ansätze auf einer zulässigen Triangulation τ von $\Omega \subset \mathbb{R}^2$ lassen sich einige der $\mathbb{R}^1$-Ergebnisse aus Abschnitt 3.7.1 übertragen. Zunächst werden drei Lemmata als Hilfsmittel bereitgestellt, die Beweise finden sich in [Hac86].

Lemma 3.7.8 Sei $T_0 = \{(\xi, \eta) : \xi, \eta \geq 0, \xi + \eta \leq 1\}$ das Einheitsdreieck. Für alle $u \in H^2(T_0)$ gilt

$$\|u\|_{H^2(T_0)}^2 \leq C\left[|u(0,0)|^2 + |u(0,1)|^2 + |u(1,0)|^2 + \sum_{|\alpha|=2} \|D^\alpha u\|_{L^2(T_0)}^2\right]. \tag{3.75}$$

Lemma 3.7.9 Sei $T_h := hT_0 := \{(x, y) : x, y \geq 0, x + y \leq h\}$. Für jedes $u \in H^2(T_h)$ gilt

$$\|D^\beta u\|_{L^2(T_h)} \leq$$

$$C\{h^{2-2|\beta|}[|u(0,0)|^2 + |u(0,h)|^2 + |u(h,0)|^2] + h^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|_{L^2(T_h)}^2\}, \tag{3.76}$$

$$|\beta| \leq 2,$$

mit C unabhängig von u, β und h.

Lemma 3.7.10 $\tilde{T}$ sei ein beliebiges Dreieck mit

$$\text{Seitenlängen } \leq h_{\max,} \qquad \text{Innenwinkel } \geq \alpha_0 > 0. \tag{3.77}$$

Für jedes $u \in H^2(\tilde{T})$ gilt

$$\|D^\beta u\|^2_{L^2(\tilde{T})} \leq C(\alpha_0)\left[h^{2-2|\beta|}_{\max} \sum_{x \text{ Ecke von } \tilde{T}} |u(x)|^2 + h^{4-2|\beta|}_{\max} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\tilde{T})}\right], |\beta| \leq 2, \tag{3.78}$$

wobei $C(\alpha_0)$ nur von α_0, nicht aber von u, β oder h abhängt.

Aus Lemma 3.7.10 ergibt sich folgendes Hauptresultat:

Satz 3.7.11 Sei $S := H^1_0(\Omega), S_N$ der FEM-Unterraum von S, $\alpha_0 > 0$ sei der kleinste Innenwinkel aller $T_i \in \tau$, während h die maximale Seitenlänge aller $T_i \in \tau$ sei. Dann ist

$$\inf_{v \in S_N} \|u - v\|_{H^k(\Omega)} \leq C'(\alpha_0) h^{2-k} \|u\|_{H^2(\Omega)} \text{ für } k = 0, 1 \text{ und alle } u \in H^2(\Omega) \cap S. \tag{3.79}$$

Beweis:

Zu $u \in H^2(\Omega)$ [bzw. $u \in H^2(\Omega) \cap S$] wählt man $v := \sum_i u(x^i) b_i \in S_N$, d.h. $v \in S_N$ mit $v(x^i) = u(x^i)$ an den [inneren] Knoten x^i. Da $w := u - v$ an den Ecken jedes $T_i \in \tau$ verschwindet und $D^\alpha w = D^\alpha u$ für $|\alpha| = 2$ (wegen $D^\alpha v = 0$ für alle $v \in S_N$), liefert (3.78) die Abschätzung

$$\|D^\beta w\|^2_{L^2(\Omega)} = \sum_{T_i \in \tau} \|D^\beta w\|^2_{L^2(T_i)} \leq \sum_{T_i \in \tau} [C(\alpha_0) h^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T_i)}]$$

$$= h^{4-2|\beta|} C(\alpha_0) \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\Omega)} \leq h^{4-2|\beta|} C(\alpha_0) \|u\|^2_{H^2(\Omega)}.$$

Summation über $|\beta| \leq k$ beweist die Ungleichung (3.79). ■

Im folgenden sei h stets die maximale Seitenlänge der Dreiecke aus τ. Da h der wesentliche Parameter ist, soll der Finite-Elemente-Raum durch S_h statt S_N bezeichnet werden. Eine Folge von Triangulationen τ_ν mit $h_\nu \to 0$ heißt quasiuniform, falls für alle ν die Innenwinkel $\geq \alpha_0 > 0$ bleiben.

Bei der Konstruktion der Triangulationen ist darauf zu achten, dass mit zunehmender Verfeinerung ($h_\nu \to 0$) nicht auch die Innenwinkel verkleinert werden. Methoden dazu werden in Kapitel 6 vorgestellt.

Der Fall $k = 1$ aus Satz 3.7.11 lässt sich nun wie folgt schreiben:

Satz 3.7.12 Sei τ_ν eine Folge quasiuniformer Triangulationen. Dann existiert eine Konstante C, so dass für alle $h = h_\nu$ und $S_h = S_{h_\nu}$ die folgende Abschätzung gilt:

$$\inf_{v \in S_h} \|u - v\|_1 \leq Ch\|u\|_2 \text{ für alle } u \in H^2(\Omega) \cap S. \tag{3.80}$$

Die Kombination dieses Satzes mit den Ritz-Eigenschaften der FEM (siehe Abschnitt 2.2) und entsprechenden Abschätzungen ergibt die beiden nächsten Sätze.

Satz 3.7.13 Sei τ_ν eine Folge quasiuniformer Triangulationen. Problem (2.19) habe eine Lösung $u \in H^2(\Omega) \cap S$. $u^h \in S_h$ sei die Finite-Elemente-Lösung. Dann gibt es eine von $u, h = h_\nu$ und ν unabhängige Konstante C, so dass

$$\|u - u^h\|_1 \leq Ch\|u\|_2. \tag{3.81}$$

Satz 3.7.14 Sei τ_ν eine Folge quasiuniformer Triangulationen mit $h_\nu \to 0$. Problem (2.19) habe eine eindeutige Lösung $u \in S$. Dann konvergiert die Finite-Elemente-Lösung $u^{h_\nu} \in S_{h_\nu}$ gegen u:

$$\|u - u^{h_\nu}\|_1 \to 0 \quad (\nu \to \infty). \tag{3.82}$$

Satz 3.7.14 beweist Konvergenz ohne einschränkende Voraussetzungen an u. Die Konvergenzordnung $O(h)$ benötigt dagegen die Annahme, dass die Lösung $u \in S$ in $H^2(\Omega)$ liegt. Diese Annahme ist allerdings nicht in jedem Fall erfüllt.

L^2- und H^s-Abschätzungen für lineare Elemente

Nach Satz 3.7.13 ist $O(h)$ die optimale Konvergenzordnung. Dies Ergebnis scheint im Widerspruch zur $O(h^2)-$Konvergenz der Fünfpunkteformel bei Differenzenverfahren zu stehen, denn das Finite-Elemente-Verfahren mit spezieller Triangulation ist fast identisch mit der Fünfpunkteformel (siehe Übungen). Der Grund hierfür ist, dass der Fehler $u - u^h$ in der $\|\cdot\|_1$-Norm gemessen wird. Die Abschätzung $\inf\{\|u - v\|_0 : v \in S_h\} \leq Ch^2\|u\|_2$ aus Satz 3.7.11 legt die Vermutung nahe, dass die $\|\cdot\|_0$-Norm des Fehlers von der Ordnung $O(h^2)$ ist: $\|u-u^h\|_0 \leq Ch^2\|u\|_2$. Diese Aussage ist jedoch ohne weitere Voraussetzungen falsch. Bisher wurde nur die Existenz einer schwachen Lösung $u \in S$ (z.B. $S = H_0^1(\Omega)$ oder $S = H^1(\Omega)$) garantiert. In Satz 3.7.13 benötigten wir jedoch die stärkere Annahme $u \in H^2(\Omega) \cap S$. Für $f \in L^2(\Omega) \subset S'$ wird $f(v)$ zu $(f, v)_{L^2(\Omega)}$. Die Regularitätsannahme lautet:

Für jedes $f \in L^2(\Omega)$ habe das Problem (2.19) eine Lösung

$$u \in H^2(\Omega) \cap S \text{ mit } \|u\|_2 \leq C_R\|f\|_0. \tag{3.83}$$

Diese Aussage trifft für hinreichend glatte Gebiete Ω zu.

Satz 3.7.15 Es gelte

$$\inf_{v \in S_h} \|u - v\|_1 \leq C_0 h\|u\|_2 \text{ für alle } u \in H^2(\Omega) \cap S. \tag{3.84}$$

Problem (2.19) habe eine Lösung $u \in S$. $u^h \in S_h \subset S$ sei die Finite-Elemente-Lösung. Dann gilt

$$\|u - u^h\|_0 \leq C_1 h\|u\|_1. \tag{3.85}$$

mit von u und h unabhängigem C_1. Falls die Lösung u auch zu $H^2(\Omega) \cap S$ gehört, gibt es eine von u und h unabhängige Konstante C_2, so dass

$$\|u - u^h\|_0 \leq C_2 h^2\|u\|_2. \tag{3.86}$$

Hinreichend für Bedingung (3.84) ist nach Satz 3.7.11, dass S_h der Finite-Elemente-Raum zu einer zulässigen und quasiuniformen Triangulation ist.

Folgerung 3.7.16 Die Ungleichungen (3.85) und (3.86) sind gleichbedeutend mit den folgenden Eigenschaften der Ritz-Projektion P_h:

$$\|I - P_h\|_{L^2(\Omega)\leftarrow S} \leq C_1 h, \tag{3.87}$$

$$\|I - P_h\|_{L^2(\Omega)\leftarrow H^2(\Omega)\cap S} \leq C_2 h^2. \tag{3.88}$$

3.7.3 "Variational Crimes"

In den bisherigen Fehlerbetrachtungen wurde angenommen:

Das Funktional

$$A(u) = B(u) - 2l(u) \text{ wird exakt berechnet (z.B. keine numerischen Integrale).} \tag{3.89}$$

$$\text{Basisfunktionen erfüllen wesentliche RB.} \tag{3.90}$$

$$\text{Basisfunktionen } \varphi^h \in H^m, \text{ konform.} \tag{3.91}$$

zu (3.89):

Sei $A^*(u) = B^*(u) - 2l^*(u)$ das bei der numerischen Behandlung verwendete Funktional. $\tilde{u}^h$ sei die A^* minimierende Funktion. Die wesentliche Bedingung, welche die verwendete Quadraturformel erfüllen muß, um Konvergenz $\tilde{u}^h \to u^h$ zu erhalten, ist:

Für Polynome P_m vom Grad m ist

$$B^*(P_m, v^h) = B(P_m, v^h) + O(h),$$

d.h. m-te Ableitungen der Basisfunktionen werden exakt integriert. Es gilt dann:

Satz 3.7.17 u^h bzw. $\tilde{u}^h$ minimieren die Funktionale A bzw. A^*. Weiter gelte:
$B^*(v^h) \geq \Theta \|v^h\|_m^2$ für alle $v^h \in H^m$ und $|(B^* - B)(u^h, v^h)| + |(l^* - l)v^h| \leq Ch^p \|v^h\|_m$.
Dann gilt

$$\|u^h - \tilde{u}^h\|_m \leq \Theta^{-1} C h^p$$

Beweis:

$$\begin{aligned}
& \Theta \|u^h - \tilde{u}^h\|_m^2 \leq B^*(u^h - \tilde{u}^h) \\
= \; & (B^* - B)(u^h, u^h - \tilde{u}^h) \\
& + B(u^h, u^h - \tilde{u}^h) - B^*(\tilde{u}^h, u^h - \tilde{u}^h) \\
= \; & |(B^* - B)(u^h, u^h - \tilde{u}^h) + (l - l^*)(u^h - \tilde{u}^h)| \\
\leq \; & Ch^p \|u^h - \tilde{u}^h\|_m \quad \blacksquare
\end{aligned}$$

Weiter kann man zeigen: Gilt für alle Polynome P_n vom Höchstgrad n,

$$\begin{aligned}
B(P_n, v^h) &= B^*(P_n, v^h), \text{ so folgt} \\
\|u^h - \tilde{u}^h\|_m &= O(h^{n-m+1})
\end{aligned}$$

zu (3.90):

Approximation des Gebietes und der wesentlichen RB.

Probleme:

(i) $\Omega \cong \Omega_h, \Omega_h$ hat polygonialen Rand oder stückweise polynomialen Rand.

(ii) RB selbst werden approximiert:

(W) $u = g$ auf $C_1 \qquad C_1 \cup C_2 = \partial\Omega$

(N) $\dfrac{\partial u}{\partial \mathfrak{n}} + \alpha u = h$ auf C_2

Natürliche RB (N) wird für $\alpha = h = 0$ automatisch erfüllt. Wesentliche RB (W) durch Ansatz im Raum

$$H_E^1 = \{v \in H^1 : v = g \text{ auf } \partial\Omega\}$$

Ist $v_g \in H_E^1$, so kann man auch schreiben

$$H_E^1 = H_0^1 + v_g$$

entsprechend

$$S^h = S_0^h + v_g^h$$

v_g^h Approximation von g.

Die Fehler bei (i) und/oder (ii) hängen stark vom speziellen Problem ab, allgemeine Aussage schwierig. In [SF73] (Strang/Fix) werden vier Beispiele diskutiert.

zu (3.91): Nicht-konforme Elemente

Ein Element-Typ kann praktisch nützlich sein, obwohl seine $(m-1)-$ten Ableitungen Unstetigkeitsstellen aufweisen, d.h.

$$S^h \not\subset H^m$$

Frage: Konvergenz?

Man minimiert

$$\begin{aligned} A_*(u) &= \sum_e B_e(u) - 2(f,u) \\ &:= B_*(u) - 2(f,u)_* \end{aligned}$$

Es liegt Konvergenz vor, wenn gilt:

$$B_*(P_m, \varphi_j) = B(P_m, \varphi_j) \;\; \forall P_m \text{ vom Hgrd. } m \text{ und } \forall \varphi_j \in S^h.$$

Ist die rechte Seite $B(P_m, \varphi_j)$ wohldefiniert?

Ja, siehe [SF73] S.176.

Kapitel 4

Differenzenverfahren

4.1 Einführung

Wir wollen hauptsächlich parabolische Probleme behandeln:

$$u_t = Lu, \tag{4.1}$$

wobei L ein elliptischer Operator in den Raumkoordinaten $x \in \mathbb{R}^n$ ist und

$$u = u(x,t) : \ \Omega \times \mathbb{R} \to \mathbb{R}, \ \ \Omega \subset \mathbb{R}^n$$

Hinzu kommen Anfangs- und Randbedingungen wie bei den Beispielen in Kapitel 1.

Faßt man L als formalen Operator auf (zunächst!), so kann man (4.1) als 1-dimensionales AWP behandeln. Hier kommen implizite und explizite Methoden in Frage (siehe Numerik II) wie z.B. das explizite Euler-Verfahren:

$$\begin{array}{rcl} U_{N+1}(x) & = & U_N(x) + \tau L U_N(x) \\ \text{wobei} \quad U_N(x) & \approx & u(x,t_N), \ t_N = N\tau. \end{array} \tag{4.2}$$

Die Ableitungen in L ersetzt man durch Differenzenquotienten, d.h. L durch L_h:

$$U_{N+1}(x_j) = U_N(x_j) + \tau L_h U_N(x_j), \ j = 1,2,\ldots,m. \tag{4.3}$$

Dabei ist m die Anzahl der Gitterpunkte, τ die Schrittweite in t.

Jetzt kann man den Fehler untersuchen. Wir betrachten zunächst den
globalen Diskretisierungsfehler

$$\varepsilon_{N,j} := u(x_j,t_N) - U_N(x_j). \tag{4.4}$$

Eine Taylor-Reihe in t liefert

$$\begin{array}{rcl} u(x_j,t_{N+1}) & = & u(x_j,t_N) + \tau u_t(x_j,t_N) + \tau\rho(x_j,t_N) \\ \text{mit} & & \end{array}$$

$$\begin{aligned} \rho(x_j,t_N) &= \frac{1}{2}\tau u_{tt}(x_j,t_N)+O(\tau^2). \\ &\Rightarrow \\ u(x_j,t_{N+1}) &= u(x_j,t_N)+\tau L_h u(x_j,t_N) \\ &+ \tau \underbrace{(L-L_h)u(x_j,t_N)}_{=:\alpha(x_j,t_N)} +\tau\rho(x_j,t_N) \end{aligned}$$

oder mit $\beta(x_j,t_N) := \alpha+\rho$

$$u(x_j,t_{N+1}) = (1+\tau L_h)u(x_j,t_N)+\tau\beta(x_j,t_N)$$

Zieht man auf beiden Seiten $U_{N+1}(x_j) = (1+\tau L_h)U_N(x_j)$ ab, so erhält man

$$\varepsilon_{N+1,j} = (1+\tau L_h)\varepsilon_{N,j}+\tau\beta(x_j,t_N) \tag{4.5}$$

Als lokalen Diskretisierungsfehler bezeichnet man den Fehler eines Schrittes unter der Voraussetzung, daß die bisherige Rechnung exakt war oder $\varepsilon_{N,j} := 0$. Division durch τ liefert gerade $\beta(x_j,t_N)$ (siehe Numerik II).

Es interessieren wieder die drei Begriffe:

Konsistenz = Konvergenz des lokalen Diskretisierungsfehlers mit $\tau \to 0$:

$$\lim_{\substack{t_N \text{ fest} \\ N\to\infty,\ \tau\to 0}} \tau\,\beta(x_j,t_N) = 0 \quad \forall\, x_j \tag{4.6}$$

Konvergenz des globalen Diskretisierungsfehlers:

$$\lim_{\substack{(x_j,t_N) \text{ fest} \\ \tau\to 0,\ h\to 0}} \varepsilon_{N,j} = 0 \tag{4.7}$$

Die Konvergenz folgt wieder aus der Stabilität, wenn wir Konsistenz voraussetzen. (Theoretisch ist das der Satz von Lax/Richtmeyer, siehe Meis/Marcowitz, [MM78].) Stabilität liegt vor, wenn die Fehlerrekursion (4.5) stabil ist, oder:

$$\text{Es soll } \|1+\tau L_h\| = 1+O(\tau) \text{ sein, besser noch } \|1+\tau L_h\| \le 1.$$

Die Stabilitätsfrage stellt also einen Zusammenhang zwischen Raumdiskretisierung h und Zeitdiskretisierung τ her. Hierfür gibt es Tests (von Neumann - Stabilität), auf die wir später zurückkommen werden. Es wird sich herausstellen, daß wie bei AWA bei gewöhnlichen DGL die impliziten Verfahren, z.B.

$$U_{N+1}(x_j) = U_N(x_j)+\tau L_h U_{N+1}(x_j) \tag{4.8}$$

stabiler sind als die expliziten. Dafür muß man in jedem Zeitschritt den Operator $(1-\tau L_h)$ invertieren, da ja dann

$$U_{N+1}(x_j) = (1-\tau L_h)^{-1}U_N(x_j). \tag{4.9}$$

Das bedeutet in der Regel die Lösung eines (linearen oder nichtlinearen) Gleichungssystems in jedem Zeitschritt. Dafür ist schon die recht ungenaue implizite Euler-Regel $\forall\tau > 0$ stabil, während bei Anwendung des expliziten Euler-Verfahrens Stabilitätsbedingungen an die Schrittweite τ zu stellen sind.

4.2 Differenzenverfahren für elliptische PDE

4.2.1 Eindimensionale Differenzenformeln

Sei $f(x)$ eine gegebene Funktion genügend hoher Differenzierbarkeit, sei weiter $f_i := f(x_i)$, $x_i := a + ih$. Dann ist

Vorwärtsdifferenz
$$\Delta f_i := \frac{f(x_{i+1}) - f(x_i)}{h} = f'(x_i) + \frac{h}{2} f''(\zeta)$$
Rückwärtsdifferenz
$$\nabla f_i := \frac{f(x_i) - f(x_{i-1})}{h} = f'(x_i) - \frac{h}{2} f''(\zeta) \tag{4.10}$$
Zentrale Differenz
$$\delta f_i := \frac{f(x_{i+1}) - f(x_{i-1})}{2h} = f'(x_i) + \frac{h^2}{3} f'''(\zeta)$$
Zweite Differenz
$$\delta^2 f_i := \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} = f''(x_i) + \frac{h^2}{12} f^{\mathrm{IV}}(\zeta)$$

Dabei ist ζ eine Zwischenstelle im entsprechenden Intervall.

δf_i und $\delta^2 f_i$ liefern also Näherungen 2. Ordnung für die erste bzw. zweite Ableitung von f am Punkt x_i.

Näherungen 4. Ordnung

$$f'(x_i) = \frac{f_{i-2} - 8f_{i-1} + 8f_{i+1} - f_{i+2}}{12h} + \frac{h^4}{30} f^{\mathrm{V}}(\zeta) \tag{4.11}$$
$$f''(x_i) = \frac{-f_{i-2} + 16f_{i-1} - 30f_i + 16f_{i+1} - f_{i+2}}{12h^2} + \frac{h^4}{90} f^{\mathrm{VI}}(\zeta) \tag{4.12}$$

Bemerkung 4.2.1 Genauso findet man Differenzen zweiter und vierter Ordnung für $f'''(x_i)$ und $f^{\mathrm{IV}}(x_i)$, man benötigt dann Werte bis $i \pm 2$ bzw. $i \pm 3$.

4.2.2 Zweidimensionale Differenzensterne

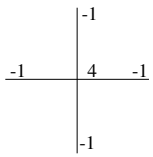
Für einen Differentialoperator im $\mathbb{R}^2$ setzt man die obigen Differenzenformeln für x und y zusammen. Dabei bekommt man Formeln mit Werten benachbarter Gitterpunkte, die man geometrisch-symbolisch als sogenannten Differenzenstern darstellen kann.

Beispiel 4.2.2 Laplace-Operator auf quadratischem Gitter (x_i, y_j) mit der Gitterweite h:

$$\begin{aligned} -\Delta u = & \frac{1}{h^2}(-u(x_{i-1}, y_j) - u(x_i, y_{j-1}) + 4u(x_i, y_j) - u(x_{i+1}, y_j) - u(x_i, y_{j+1})) \\ & - \frac{h^2}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right)(\xi, \eta) \end{aligned} \tag{4.13}$$

Teilweise Wiederholung aus Numerik I.

Der zugehörige Differenzenstern ist:



Sind die Schrittweiten in x- und y-Richtung unterschiedlich, so ergibt sich mit $x_i = a + ih, \quad y_j = c + jk$

$$\begin{aligned} -\Delta u(x_i, y_i) = & \quad \frac{-u_{i-1,j} + 2u_{ij} - u_{i+1,j}}{h^2} + \frac{-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}}{k^2} \\ & - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\zeta, \eta) - \frac{k^2}{12}\frac{\partial^4 u}{\partial y^4}(\tilde{\zeta}, \tilde{\eta}) \end{aligned} \tag{4.14}$$

Viele solche Sterne mit Fehlerglied findet man im Collatz, [7]. Wichtig sind noch verschiedene 9-Punkte-Sterne für Δu.

Im Collatz [7] findet man auch Sterne für triangulierte Gebiete.

Beispiel 4.2.3

$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad (0,1) \times (0,1) =: \Omega \\ u &= 0 \quad \text{auf} \quad \partial\Omega \end{aligned}$$

$x_i := i\,h$, $y_j := j\,h$, $h := 1/(N+1)$.

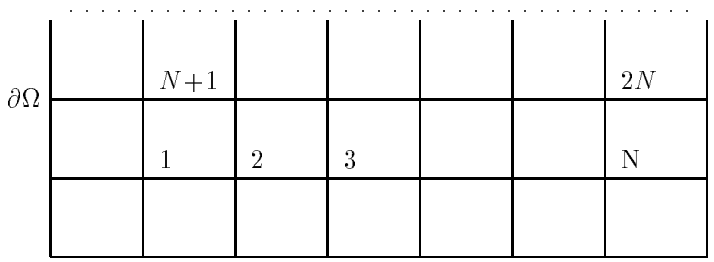
Diskretisierung mit (4.13): Mit $u_{ij} := u(x_i, y_j)$ bekommt man für alle inneren Punkte P_{ij} die Gleichungen

$$-u_{i-1,j} - u_{i+1,j} + 4u_{i,j} - u_{i,j-1} - u_{i,j+1} = h^2 f_{ij}$$

Ordnet man die Punkte in der Index-Reihenfolge

$$(1,1), (2,1), (3,1), \ldots, (N,1), (1,2), (2,2), (3,2), \ldots, (N,N),$$

also geometrisch zeilenweise von unten nach oben, so ergibt sich das folgende Gitter



und man bekommt man die $N^2 \times N^2$ - Matrix

$$A = \begin{pmatrix} B & -I & 0 & \cdots & 0 \\ -I & B & -I & 0 & \\ 0 & \ddots & \ddots & \ddots & \ddots \\ & \ddots & & & -I \\ 0 & \cdots & 0 & -I & B \end{pmatrix}$$

mit den Blöcken

$$B = \begin{pmatrix} 4 & -1 & 0 & \cdots & 0 \\ -1 & 4 & -1 & 0 & \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & -1 & 4 & -1 \\ 0 & \cdots & 0 & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N,N}, I \in \mathbb{R}^{NN}$$

Dabei wurde in den Randpunkten $u_{i,j} = 0$ eingesetzt.

Damit ergibt sich eine Halbbandbreite $m = N$, allerdings hat jede Zeile nur maximal fünf Elemente ungleich Null. Hier bieten sich dementsprechend zur Lösung spezielle Iterationsverfahren an, auf die wir unten näher eingehen werden.

4.2.3 Neumann - RWA

Ist in einem Randstück $C_1 \subset \partial\Omega$ die Randbedingung

$$\frac{\partial u}{\partial \mathfrak{n}} = g(x, y) \text{ auf } C_1 \tag{4.15}$$

vorgeschrieben, so muß diese zur Erhaltung der Ordnung mit einem zentralen Differenzenquotienten approximiert werden. Sei o.B.d.A. $C_1 = \{x = 0 | 0 \leq y \leq 1\}$, dann ergibt (4.15)

$$\frac{u_{-1,j} - u_{1,j}}{2h} = g_{0,j} = g(0, y_j) \tag{4.16}$$

Der Wert $u_{-1,j}$ für einen Punkt außerhalb des Gebietes Ω wird eliminiert durch Ansetzen der PDE im Randpunkt $(0, y_i)$. Lautet diese z.B. $-\Delta u = f$ wie in 4.2.3, so ergibt das für die "Punkte $i = 0$"

$$\begin{aligned} -u_{-1,j} - u_{1,j} + 4u_{0,j} - u_{0,j-1} - u_{0,j+1} &= h^2 f_{0,j} \\ \text{aber nach (4.16): } -u_{-1,j} &= -u_{1,j} - 2hg_{0,j} \\ \Rightarrow & \\ 4u_{0,j} - 2u_{1,j} - u_{0,j-1} - u_{0,j+1} &= h^2 f_{0,j} + 2hg_{0,j} \end{aligned} \tag{4.17}$$

Daß hier die Randpunktwerte als Unbekannte eingehen, ist vernünftig, da diese ja nicht vorgeschrieben sind. Im Fall gemischter Randbedingungen verfährt man entsprechend.

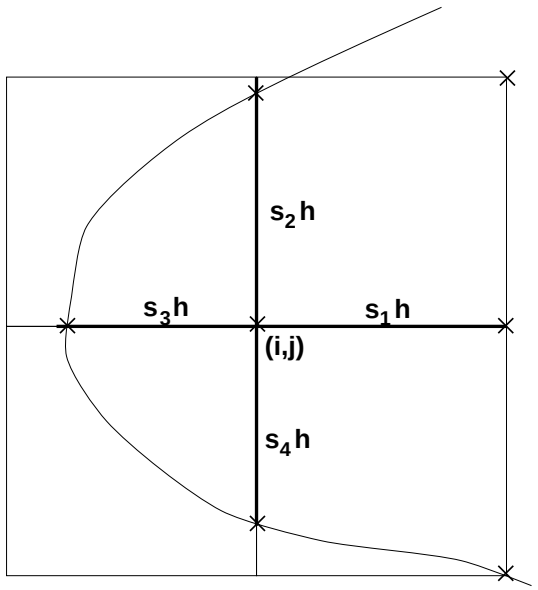
4.2.4 Gekrümmter Rand

Wir setzen voraus, daß wir ein quadratisches Netz über das Gebiet Ω gelegt haben. Bei gekrümmtem Rand kann man jetzt berücksichtigen, daß die Randwerte in kürzerem Abstand als h zu den inneren Punkten liegen. Durch Taylor-Reihen-Abgleich mit $s_i h$ ergibt sich

$$-\Delta u = \frac{1}{h^2}[\beta_1 u_{i+s_1,j} + \beta_2 u_{i,j+s_2} + \beta_3 u_{i-s_3,j} + \beta_4 u_{i,j-s_4} + \beta_0 u_{ij}] + O(h)$$

mit

$$\begin{aligned} \beta_1 &= \frac{-2}{(s_1+s_3)\,s_1}, \quad \beta_2 = \frac{-2}{(s_2+s_4)\,s_2} \\ \beta_3 &= \frac{-2}{(s_1+s_3)\,s_3}, \quad \beta_4 = \frac{-2}{(s_2+s_4)\,s_4} \\ \beta_0 &= -\sum_{i=1}^{4} \beta_i . \end{aligned} \tag{4.18}$$



Man beachte, daß der Fehlerterm hierbei eine Ordnung verliert, da das Glied dritter Ordnung sich aus den Taylor-Reihen nicht mehr heraushebt.

(Beweis: Ü.)

Will man nicht eine Fehlerordnung verlieren, so muß man am Rand kompliziertere Formeln einsetzen.

Auf gemischte RBen wollen wir nicht eingehen, hier ist zusätzliche Arbeit zu leisten.

4.2.5 Allgemeine Forderungen

Es gibt verschiedene Möglichkeiten, Diskretisierungsformeln zu entwickeln:
Taylor-Entwicklungen, Box-Integration etc., siehe Literatur. Ist L ein selbstadjungierter, elliptischer Operator, so sollte L_h drei Bedingungen erfüllen. Sei also

$$L_h u(P) \equiv \alpha_0 u(P) - \sum_{i=1}^{s} \alpha_i u(Q_i)$$

mit definierten Werten α_i und s Punkten Q_i, die P zugeordnet sind.

Bedingungen:

(i) Nicht-negative Koeffizienten:

$$\alpha_0 > 0 \text{ und } \alpha_i \geq 0,\ i = 1, 2, \ldots, s, \text{ für jeden Punkt } P. \tag{4.19}$$

(ii) Diagonal-Dominanz der Koeffizientenmatrix:

$$\begin{aligned} \sum_{i=1}^{s} |\alpha_i| &\leq \alpha_0 \quad \forall P, \\ \sum_{i=1}^{s} |\alpha_i| &< \alpha_0 \quad \text{für mindestens ein } P. \end{aligned} \tag{4.20}$$

(iii) Symmetrie der Koeffizientenmatrix

Durch (i) & (ii) übertragt sich das Maximumprinzip, das für L gilt, auf L_h.
(iii) entspricht der Selbstadjungiertheit von L und ist nur algorithmisch wichtig.

4.3 Differenzenverfahren für parabolische PDE

Wir betrachten nur das homogene selbstadjungierte Problem

$$\begin{aligned} \text{PDE:} \quad & u_t &&= \sum_{i=1}^{n} D_i(a_i(x) D_i u) \text{ in } \Omega \times \mathbb{R},\ \Omega \subset \mathbb{R}^n \\ \text{AB:} \quad & u(x,0) &&= u_0(x) \\ \text{RB:} \quad & u(x,t) &&= g(x) \quad \forall t \geq 0, \forall x \in \partial\Omega \\ & && \Omega \text{ beschränkt, zusammenhängend.} \end{aligned} \tag{4.21}$$

Oft betrachtet man das einfache Problem $u_t = \Delta u$ in einer Raumdimension ($n = 1$). Fehlerordnungen und Stabilitätsuntersuchungen lassen sich auf mehrere Raumdimensionen übertragen, nur die Struktur der linearen Gleichungssysteme ändert sich mit n.

4.3.1 Eine Schar von Verfahren für (4.21) (im $\mathbb{R}^1$)

Die PDE wird zu

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right), x \in [d,e], t \in [0,T] \tag{4.22}$$

Sei

$$\begin{aligned} a \in C^\infty(\mathbb{R}), \qquad & a' \in C^\infty(\mathbb{R}), \\ a(x) > 0 \qquad & \forall x \in [d,e], \end{aligned}$$

und sei

$$L_h u(x,t) := h^{-2}\left\{a\left(x+\frac{h}{2}\right)[u(x+h,t)-u(x,t)] - a\left(x-\frac{h}{2}\right)[u(x,t)-u(x-h,t)]\right\} \tag{4.23}$$

eine Differenzenapproximation zu $Lu = D(aDu)$.

Seien

$$\begin{aligned} t_j &= j\,\tau,\ \tau := \frac{T}{N},\ j = 0,1,\ldots,N, \\ x_i &= d + i\,h,\ h = \frac{e-d}{m},\ i = 0,1,\ldots,m, \text{ und} \\ U_i^j &\approx u(x_i,t_j). \end{aligned}$$

Dann hat die Verfahrensschar

$$U_i^{j+1} = U_i^j + \tau\left\{\alpha L_h U_i^j + (1-\alpha) L_h U_i^{j+1}\right\} \tag{4.24}$$

einen lokalen Diskretisierungsfehler

$$O(\tau + h^2) \text{ für } \tau, h \in \mathbb{R}, \alpha \in [0,1]. \tag{4.25}$$

Beweis: Folgt aus der Einleitung und dem $O(h^2)$ der zweiten Differenz in x.

Das Verfahren ist

explizit	für	$\alpha = 1$,
halbimplizit	für	$\alpha \in (0,1)$,
(voll) implizit	für	$\alpha = 0$.

Im Spezialfall $\alpha = 1/2$ bekommt man das günstige Crank-Nicholson-Verfahren, s.u. Verallgemeinerung auf höhere Dimensionen n ist einfach:

$$L_h u = h^{-2}\left(\sum_{i=1}^{n} a\left(x_{1,\ldots,}x_{i-1}, x_i + \frac{h}{2}, x_{i+1,\ldots}, x_1\right)\Delta_{x_i} u - a\left(\cdots, x_i - \frac{h}{2}, \cdots\right)\nabla_{x_i} u\right) \tag{4.26}$$

Das wird zu $\sum_{i=1}^{n} \delta_{x_i}^2 u$ im Fall $a \equiv 1$.

Beispiel 4.3.1

$$\begin{array}{lcll} u_t & = & u_{xx} & \text{für } x \in (0,1), t \in (0, 0.1] \\ u_0(x) & = & 0 & \text{für } t = 0 \\ u(0,t) & = & 1 & \text{(Sprungstelle!)} \\ u(1,t) & = & 0 & \end{array}$$

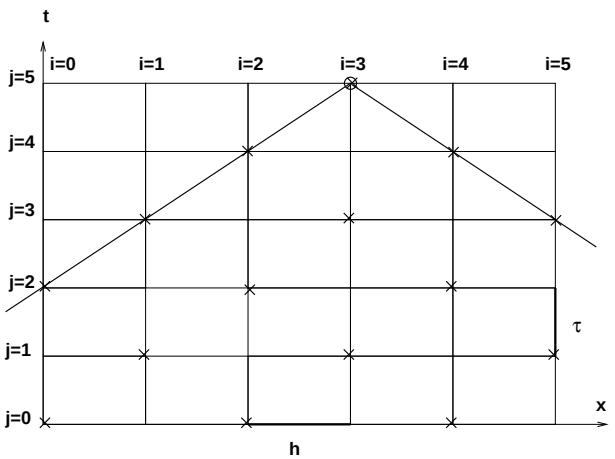
Wir setzen in (4.24) $\alpha = 1$ und $\tau = h^2/2$ und bekommen damit das explizite Verfahren

$$U_i^{j+1} = U_i^j + \frac{1}{2}(U_{i+1}^j - U_i^j - U_i^j + U_{i-1}^j)$$

oder

$$U_i^{j+1} = \frac{1}{2}(U_{i+1}^j + U_{i-1}^j),\ i = 1, \ldots, m-1,\ j = 0, 1, \ldots. \tag{4.27}$$

Dieses Verfahren hat den in der folgenden Zeichnung angedeuteten Einflußkegel. Der Punkt $(i = 3,\ j = 5)$ wird von den Werten $\{(i = 2,\ j = 4),\ (i = 4,\ j = 4)\}$ bestimmt und diese von den Werten $\{(i = 1,\ j = 3),\ (i = 3,\ j = 3),\ (i = 5,\ j = 3)\}$ usw.



Wir rechnen mit $h = 0.2$, also $\tau = 0.02$, und vergleichen die Ergebnisse mit der wahren Lösung

$$u(x,t) = 1 - x - \frac{2}{\pi}\sum_{i=1}^{\infty}\frac{1}{i}e^{-i^2\pi^2 t}\sin i\pi x \tag{4.28}$$

Ergebnisse nach (4.27) mit $h = 0.2,\ \tau = 0.02$						
t= / x=	0	0.2	0.4	0.6	0.8	1.0
0	1.0	0.0	0.0.	0.0	0.0	0.0
0.02	1.0	0.5	0.0	0.0	0.0	0.0
0.04	1.0	0.5	0.25	0.0	0.0	0.0
0.06	1.0	0.625	0.25	0.125	0.0	0.0
0.08	1.0	0.625	0.375	0.125	0.0625	0.0
0.10	1.0	0.6875	0.375	0.21875	0.0625	0.0
$(h = 0.1)$	1.0	0.664	0.383	0.187..	0.070...	0.0
$u(x, 0.1)$	1.0	0.655	0.371	0.178	0.066	0.0

Ergebnisse nach (4.29) (s.u.) mit $h = 0.2,\ \tau = 0.02$						
t= / x=	0	0.2	0.4	0.6	0.8	1.0
0	1.0	0.0	0.0.	0.0	0.0	0.0
0.02	1.0	0.375	0.0625	0.0	0.0	0.0
0.04	1.0	0.3125	0.25	0.0625	0.0	0.0
0.06	1.0	1.0	-0.0625	0.125	0.0625	0.0
0.08	1.0	-0.75	1.5	-0.1875	0.0	0.0
0.10	1.0	5.0	-4.0	2.0	-0.125	0.0

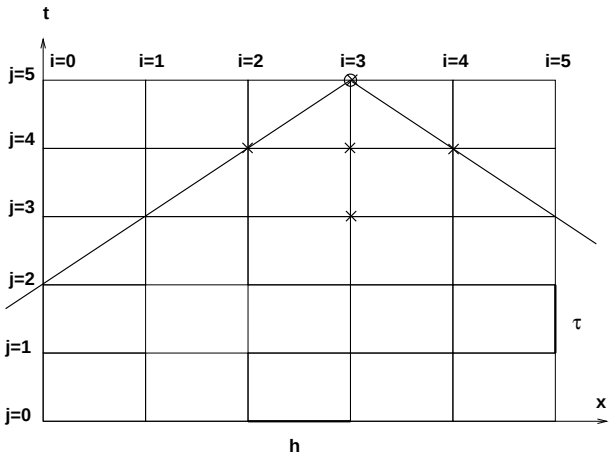
(4.29) ist offenbar für dieses Problem ein instabiles Verfahren.

Für $h = 0.1, \tau = 0.005$, müssen wir $9*20 = 180$ statt $4*5 = 20$ Werte berechnen, die Fehlerquadratsumme geht dabei von $3_{10}{-}3$ auf $4_{10} - 4$ zurück, aber einzelne Fehler werden auch größer. Deswegen versuchen wir eine andere explizite Regel (zentraler Differenzenquotient für u_t):

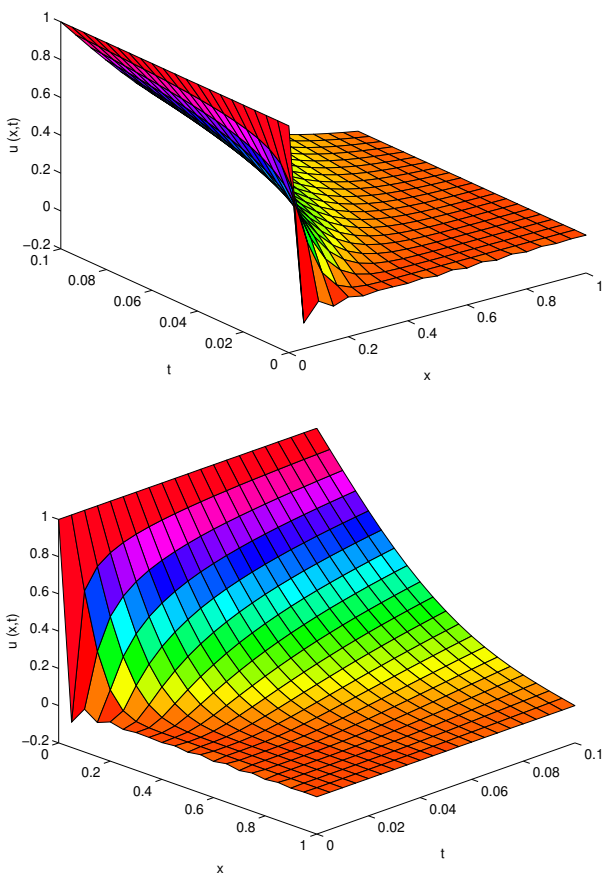
$$U_i^{j+1} = U_i^{j-1} + U_{i+1}^j - 2U_i^j + U_{i-1}^j \tag{4.29}$$

(wieder mit $\tau = \frac{1}{2}h^2$, also $2\tau = h^2$)

Der Einfluß in einem Punkt i, j erstreckt sich jetzt über zwei Zeitschichten (siehe die folgende Zeichnung), deswegen nimmt man für die erste Zeitschicht die alte Formel (4.27), allerdings mit $h = 0.1, \tau = 0.005$.



Die wahre Lösung aus zwei Sichten:



Explizites Euler-Verfahren

Satz 4.3.2 Sei für die PDE (4.22) $\lambda = \frac{\tau}{h^2}$ und

$$\left|\frac{\partial^{j+l} u(x,t)}{\partial x^j \partial t^l}\right| < L(x) \text{ mit } L(x) \in C^0 \cap L^2 \text{ für } 0 \le l \le 2, 0 \le j \le 4 \tag{4.30}$$

Dann ist das Verfahren (4.24) mit $\alpha = 1$ konsistent mit der Ordnung 1 und stabil, falls

$$0 < \lambda \max_{x \in \mathbf{R}} \{a(x)\} \le \frac{1}{2} \text{ (Stabilitätsbedingung)} \tag{4.31}$$

Für den Beweis benötigen wir noch:

Lemma 4.3.3 Sei $f(x) \in C^0(\mathbb{R}) \cap L^2(\mathbb{R})$ komplexwertig, $a \in C^0(\mathbb{R}), a(x) > 0, h \in \mathbb{R}$. Dann gilt

$$\begin{aligned} &\int_{-\infty}^{\infty} \overline{f(x)} \Big\{ a\big(x + \frac{h}{2}\big)[f(x+h) - f(x)] - a\big(x - \frac{h}{2}\big)[f(x) - f(x-h)] \Big\} dx \\ = \quad & -\int_{-\infty}^{\infty} a(x) |f\big(x + \frac{h}{2}\big) - f\big(x - \frac{h}{2}\big)|^2 dx \end{aligned} \tag{4.32}$$

und

$$\begin{aligned} &\int_{-\infty}^{\infty} |a\big(x + \frac{h}{2}\big)(f(x+h) - f(x)) - a\big(x - \frac{h}{2}\big)(f(x) - f(x-f))|^2 dx \\ \le \quad & 4 \int_{-\infty}^{\infty} (a(x))^2 |f\big(x + \frac{h}{2}\big) - f\big(x - \frac{h}{2}\big)|^2 dx \end{aligned} \tag{4.33}$$

Beweis: Übung
Schritte:

(i) Man benutzt $\int_{-\infty}^{\infty} g(x+h)dx = \int_{-\infty}^{\infty} g(x)dx$,

Der Rest ist Rechnen: 1. Produkt $f\, a\, [\cdots] - \frac{h}{2}$,

2. Produkt $f\, a\, [\cdots] + \frac{h}{2}$ im Argument

dann bekommt man $a(x)$ und der Faktor wird zu $|\cdots|^2$

(ii) Es ist $|\alpha + \beta|^2 \le 2(|\alpha|^2 + |\beta|^2)$. Das ergibt erste Umformung, dann wieder $\pm\frac{h}{2}$ im Argument wie oben. ■

Beweis von 4.3.2: Konsistenz klar, s.o., Stabilität:

Wir beweisen $\|1+\tau L_h\| \le 1$. Daraus folgt dann die Stabilität.

Sei $f \in C^0 \cap L^2$. Dann ist mit $C := 1+\tau L_h$

$$Cf = f(x) + \lambda\left\{a\left(x+\frac{h}{2}\right)[f(x+h)-f(x)] - a\left(x-\frac{h}{2}\right)[f(x)-f(x-h)]\right\}$$

$$\Longrightarrow \qquad \|Cf\|^2 = \|f\|^2 + 2\lambda \int_{-\infty}^{\infty} f(x)\{\cdots\}dx + \lambda^2 \int_{-\infty}^{\infty} \{\cdots\}^2$$

(reellwertig angesetzt, sonst ähnlich).

Jetzt liefert das Lemma 4.3.3:

$$2\lambda \int_{-\infty}^{\infty} f(x)\{\cdots\} =_{(4.32)} -2\lambda \int_{-\infty}^{\infty} a(x)|f\left(x+\frac{h}{2}\right) - f\left(x-\frac{h}{2}\right)|^2 dx$$

und

$$\lambda^2 \int_{-\infty}^{\infty} |\{\cdots\}|^2 dx \le_{(4.33)} 4\lambda^2 \int_{-\infty}^{\infty} (a(x))^2 |f(x+\frac{h}{2}) - f(x-\frac{h}{2})|^2 dx$$

$\Longrightarrow$

$$\|Cf\|^2 \le \|f\|^2 - 2\lambda \int_{-\infty}^{\infty} a(x)\underbrace{(1-2\lambda a(x))}_{\ge 0}\left|f(x+\frac{h}{2}) - f(x-\frac{h}{2})\right|^2 dx$$

(da $0 < \lambda \max\{a(x)\} \le \frac{1}{2}$)

$\Longrightarrow$

$$\|Cf\| \le \|f\| \quad \forall f$$

$\Longrightarrow$

$$\|C\| \le 1 \quad \blacksquare$$

Satz 4.3.4 PDE (4.22), $\lambda := \dfrac{\tau}{h^2}$, Verfahren (4.24). Dann bekommt man für $\alpha \in [\frac{1}{2}, 1]$ die Stabilitätsbedingung

$$(2\alpha - 1)\,\lambda \max\{a(x)\} \le \frac{1}{2}. \tag{4.34}$$

Für $\alpha \in [0, \frac{1}{2}]$ ist das Verfahren unbedingt stabil.

Beweis: Einen abstrakten Beweis für $\alpha = 0$ findet man im Meis-Marcowitz. Dort auch Stabilitätsausagen für einige weiterführende Verfahren. Ein konstruktiver Beweis $\forall\alpha$ (dort $\delta = 1-\alpha$) findet sich im Isaacson/Keller "Analyse numerischer Verfahren" von 1973. ■

Bemerkung 4.3.5 (i) Besonders günstig ist die Wahl $\alpha = \frac{1}{2}$. Man bekommt dann das berühmte CRANK-NICHOLSON-Verfahren

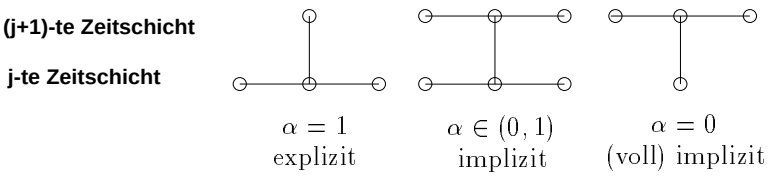
$$U_i^{j+1} = U_i^j + \frac{\tau}{2}\{L_h(U_i^j + U_i^{j+1})\} \tag{4.35}$$

oder

$$(1 - \frac{\tau}{2}L_h)U_i^{j+1} = (1 + \frac{\tau}{2}L_h)U_i^j, \tag{4.36}$$

das aufgrund seiner besonderen Symmetrie die Fehlerordnung $O(\tau^2 + h^2)$ hat.

(ii) Punkte-Sterne (Einflußbereich):



$\alpha = 1$ explizit | $\alpha \in (0,1)$ implizit | $\alpha = 0$ (voll) implizit

(iii) Maximale Fehlerordnung mit 3×2 Punkten erreicht das Verfahren von DOUGLAS, das für $a(x) \equiv 1$ die Form

$$\begin{aligned} U_i^{j+1} = U_i^j \quad &+\left(\lambda - \frac{1}{6}\right)[U_{i+1}^{j+1} - 2U_i^{j+1} + U_{i-1}^{j+1}] \\ &+\left(\lambda + \frac{1}{6}\right)[\underbrace{U_{i+1}^j - 2U_i^j + U_{i-1}^j}_{h^2 L_h U^j \ !!!}] \end{aligned} \tag{4.37}$$

hat und mit $\lambda = \frac{\tau}{h^2}$ die lokale Diskretisierungsfehler-Ordnung $O(\tau^2 + h^4)$ liefert.

Beispiel 4.3.6

$$\begin{aligned} u_t &= u_{xx} && \text{in } [0,\pi] \times [0,T] \\ u(x,0) &= \sin x && x \in [0,\pi] \\ u(0,t) &= u(\pi,t) = 0 && t \in [0,T] \end{aligned}$$

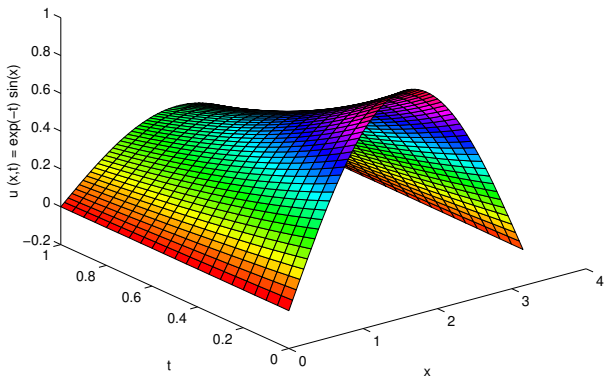
Exakte Lösung: $u(x,t) = e^{-t} \sin x$

$h = \frac{\pi}{20}, \lambda = \frac{1}{\sqrt{20}}$ (günstiger Wert für DOUGLAS)

$\Longrightarrow \tau = \frac{\pi^2}{400\sqrt{20}} = 0.0055\ldots$

Für $x = \frac{\pi}{2}, t = j\,\tau$ ergibt das folgende Fehler:

j	Exakte Lösung	Fehler $*10^6$	
		CN	DOUGLAS
1	0.9944979	11	– 0.000026
80	0.6431468	531	– 0.001331
800	0.0121088	100	– 0.000251



4.3.2 Ein HOPSCOTCH - Verfahren

Dieses explizit-implizite Schema basiert auf der kombinierten Anwendung der expliziten und der impliziten Euler-Regel (auch SPLITTING genannt). Man hat auf diese Weise ein Verfahren mit nicht soviel Aufwand wie beim voll-impliziten Euler und trotzdem guten Stabilitätseigenschaften. Wir betrachten es eindimensional und nehmen an

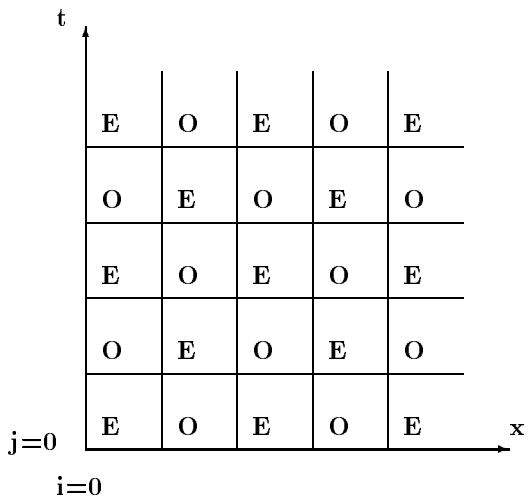
$$L_h U_i^j = f(U_{i-1}^j, U_i^j, U_{i+1}^j)$$

wie üblich. Jetzt unterscheiden wir zwischen Punkten

mit "$i+j$ gerade" (EVEN)
und "$i+j$ ungerade" (ODD)

Dann besteht das ODD-EVEN HOPSCOTCH-Verfahren (GORDON, GOURLAY) in

$$\left.\begin{array}{ll} U_i^{j+1} = U_i^j + \tau L_h U_i^j & \text{falls } (i+j) \text{ ungerade} \\ U_i^{j+1} = U_i^j + \tau L_h U_i^{j+1} & \text{falls } (i+j) \text{ gerade} \end{array}\right\} \tag{4.38}$$



Betrachtet man jetzt dieses Verfahren über zwei Zeitschritte, kann man es in ein explizites umformulieren. Allgemein kann man schreiben:

$$U_i^{j+1} = U_i^j + \tau\varphi_i^j L_h U_i^j + \tau\varphi_i^{j+1} L_h U_i^{j+1} \tag{4.39}$$

mit

$$\varphi_i^j = \begin{cases} 1 \text{ für } i+j \text{ ungerade} \\ 0 \text{ für } i+j \text{ gerade} \end{cases}$$

Damit lautet der nächste Schritt

$$U_i^{j+2} = U_i^{j+1} + \tau\varphi_i^{j+1} L_h U_i^{j+1} + \tau\varphi_i^{j+2} L_h U_i^{j+2} \tag{4.40}$$

(4.40)−(4.39) ergibt

$$U_i^{j+2} = 2U_i^{j+1} - U_i^j + \tau\varphi_i^{j+2} L_h U_i^{j+2} - \tau\varphi_i^j L_h U_i^j$$

Das führt wegen $\varphi_i^j = \varphi_i^{j+2}$ zu

$$\begin{cases} U_i^{j+2} = 2U_i^{j+1} - U_i^j & \text{für } (i+j) \text{ gerade} \\ U_i^{j+2} = U_i^j + \tau L_h U_i^j + \tau L_h U_i^{j+2}, & \text{falls } (i+j) \text{ ungerade} \end{cases} \tag{4.41}$$

Für $u_t = u_{xx}$ und die übliche Diskretisierung wird das mit $\lambda = \dfrac{\tau}{h^2}$ zu

$$U_i^{j+2} = U_i^j + \lambda(U_{i-1}^j - 2U_i^j + U_{i+1}^j) + \lambda(U_{i-1}^{j+2} - 2U_i^{j+2} + U_{i+1}^{j+2})$$

für $i+j$ ungerade, gerade Werte bekannt. Dies wird explizit zu:

$$\begin{aligned} U_i^{j+2} &= 2U_i^{j+1} - U_i^j \\ &\text{für } (i+j) \text{ gerade (reine Extrapolation)} \\ &\text{und} \\ U_i^{j+2} &= (1+2\lambda)^{-1}[U_i^j + \lambda(U_{i-1}^j - 2U_i^j + U_{i+1}^j) + \lambda(U_{i-1}^{j+2} + U_{i+1}^{j+2})] \\ &\text{für } (i+j) \text{ ungerade.} \end{aligned} \tag{4.42}$$

Solche zweistufigen Schemata nennt man auch LEAPFROG-Schema (Bockspringen) nach DUFORT/FRANKEL.

Da man jetzt über 2 Zeitschritte rechnet, braucht man eine Anlaufrechnung. Auf weitere Einzelheiten wollen wir aber nicht eingehen.

4.4 Stabilitäts-Test bei Anfangswertaufgaben

Spezielle PDE!

Festes Verfahren!

Ansatz für den Fehler auf der 0-ten Schicht $t = 0$ (statt der Anfangsbedingung)!

Sei $\overline{t} = \tau\, N$ ein fester Punkt!

Frage: Bleiben die Fehler beschränkt, wenn $\tau \to 0$?

Ist $C = C(\tau)$ der Differenzenoperator mit $U_i^{j+1} = CU_i^j$, so ist die Frage gleichbedeutend mit:

Ist $\|C^N\|$ gleichmäßig beschränkt in N?

4.4.1 FOURIER-Stabilität (JOHN VON NEUMANN)

Sei $i = \sqrt{-1}$. Wir machen einen Ansatz für den Fehler auf der 0-ten Schicht:

$$E(x) = \sum_j A_j e^{i\beta_j x} \tag{4.43}$$

mit beliebigen Koeffizienten $A_j, \beta_j,\ \beta_j$ verschieden. Wegen der linear unabhängigen Überlagerung der $e^{i\beta_j x}$-Terme genügt es, die Fehlerentwicklung eines Einzel-Terms $e^{i\beta x}$ zu untersuchen.

Lösungsstörung in $\overline{t}$, Ansatz

$$e^{\alpha \overline{t}} e^{i\beta x} = e^{\alpha N \tau} e^{i\beta x},\ \alpha = a(\beta) \in \mathbb{C}$$

Stabilitätsbedingung

$$|e^{\alpha\tau}| \le 1 \qquad \forall \beta\, x \tag{4.44}$$

(schwächer: $|e^{\alpha\tau}| = 1 + O(\tau)$), s.o.

Beispiel 4.4.1 Wir untersuchen die Stabilität des expliziten Verfahrens (4.24) mit $\alpha = 1$ und der üblichen zweiten Differenz für L_h, also

$$U_r^{s+1} = U_r^s + \lambda(U_{r-1}^s - 2U_r^s + U_{r+1}^s).$$

Wenden wir dieses Verfahren auf die PDE

$$u_t = u_{xx}$$

an und setzen als Anfangswerte den obigen Fehleransatz ein:

$$u(x,0) = E(x) = \sum_j A_j e^{i\beta_j x},$$

dann ergibt sich für den Fehler in (x_r, t_s) (mit $\xi := e^{\alpha\tau}$):

$$U_r^s = e^{\alpha s\tau} e^{i\beta rh} = \xi^s e^{i\beta rh}$$

$$\begin{aligned}
&\Longrightarrow \\
\xi^{s+1} e^{i\beta rh} &= \xi^s[(1-2\lambda)e^{i\beta rh} + \lambda(e^{i\beta(r+1)h} + e^{i\beta(r-1)h})] \\
&\Longrightarrow \\
\xi &= (1-2\lambda) + \lambda(e^{i\beta h} + e^{-i\beta h}) \\
e^{i\beta h} &= \cos\beta h + i\sin\beta h \\
&\Longrightarrow \\
\xi &= (1-2\lambda) + 2\lambda\cos\beta h \\
&= 1 - 2\lambda(1-\cos\beta h) \\
&= 1 - 4\lambda\sin^2\frac{\beta h}{2}
\end{aligned}$$

Die Bedingung (4.44) bedeuted $|\xi| \leq 1 \quad \forall \beta h$ oder hier mit $\lambda > 0$:

$$4\lambda \leq 2 \Leftrightarrow \lambda \leq \frac{1}{2}$$

Diese Bedingung ist identisch mit (4.31) in Satz 4.3.2.

Bemerkung 4.4.2 (i) Bei variablen Koeffizienten ist die Methode nur lokal anwendbar. Dann kann man z.B. MAPLE einsetzen.

(ii) RB werden ignoriert. Deshalb bekommt man nur notwendige Stabilitätsbedingung.

(iii) Im Fall $n > 1$ macht man einen entsprechenden Ansatz für den Fehler in $t = 0$: $E(x) = e^{i\beta_1 x_1} e^{i\beta_2 x_2} \cdots e^{i\beta_n x_n}$. Die Stabilitätsuntersuchungen verlaufen dann ganz entsprechend.

4.4.2 Die Matrizen-Methode

Fassen wir die Werte der j-ten Zeitschicht zu einem Vektor zusammen

$$U_j := (U_0^j, U_1^j, \ldots, U_m^j)^T,$$

dann läßt sich ein Zeitschritt schreiben als

$$\begin{aligned}
A_j U_{j+1} &= B_j U_j \text{ mit } A_j, B_j \in \mathbb{R}^{m+1,m+1} \\
&\text{oder} \\
U_{j+1} &= C_j U_j \qquad (4.45) \\
&\text{mit} \\
C_j &= A_j^{-1} B_j, \text{ falls } A_j \text{ regulär.}
\end{aligned}$$

Eine Anfangsstörung Z_0 ergibt die $(j+1)-$te Vektorstörung

$$Z_{j+1} = (C_j\, C_{j-1} \cdots C_0) Z_0\,.$$

Stabilitätsforderung

$$\| \prod_{k=0}^{j} C_k \| \leq K \quad \forall j. \tag{4.46}$$

Normalerweise sind die Matrizen konstant bezüglich j, also:

$$A_j = A, B_j = B \Longrightarrow C_j = C = A^{-1}B.$$

Dann bedeutet (4.46): Notwendig für die Stabilität ist

$$\rho(C) \leq 1, \tag{4.47}$$

hinreichend ist $\|C\| \leq 1$ für irgendeine submultiplikative Norm ($\rho(C) = \|C\|_2$, falls C symmetrisch), also ist (4.47) notwendig und hinreichend.

Beispiel 4.4.3 : Wir greifen Beispiel 4.4.1 wieder auf, es ist $A = I$

$$C = B = \begin{pmatrix} 1-2\lambda & \lambda & & \oslash \\ \lambda & \ddots & \ddots & \\ & \ddots & \ddots & \lambda \\ \oslash & & \lambda & 1-2\lambda \end{pmatrix}$$

Die Eigenwerte dieser Matrix berechnet man leicht als

$$\lambda_j = 1 - 4\lambda \sin^2 \frac{j\pi}{2m}$$

was wieder zu $\lambda \leq \frac{1}{2}$ führt. Also ist diese Bedingung für das explizite EULER-Verfahren auch hinreichend.

Beispiel 4.4.4 : CRANK-NICHOLSON für $u_t = u_{xx}$ mit Randwerten Null lautet

$$U_{i-1}^{j+1} - 2(1+\lambda^{-1})U_i^{j+1} + U_{i+1}^{j+1} = -(U_{i-1}^{j} - 2(1-\lambda^{-1})U_i^{j} + U_{i+1}^{j}), i = 1, \ldots, N-1.$$

Ist $P := \begin{pmatrix} -2 & 1 & & \oslash \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ \oslash & & 1 & -2 \end{pmatrix}$, so sind jetzt

$$A_n \equiv A = P - \frac{2}{\lambda} I, \qquad B_n \equiv B = -P - \frac{2}{\lambda} I$$

Die Eigenwerte von $C := A^{-1}B$ kann man ohne Berechnung der Inversen von A finden:

Aus

$$(A^{-1}B - \mu I)v = 0$$

folgt

$$(B - \mu A)v = 0$$

oder

$$\left(-P - \frac{2}{\lambda}I - \mu\left(P - \frac{2}{\lambda}I\right)\right)v = 0$$

oder

$$\left(-P(1+\mu) - \frac{2}{\lambda}(1-\mu)I\right)v = 0$$

oder

$$\left(P - \frac{2(\mu-1)}{\lambda(\mu+1)}I\right)v = 0$$

Man kennt die Eigenwerte von P:

$$(P - \gamma_n I)\, v = 0 \Longrightarrow \gamma_n = -4\sin^2\frac{n\pi h}{2},$$

wo h die räumliche Gitterweite ist. Damit folgt

$$\frac{2(\mu_n - 1)}{\lambda(\mu_n + 1)} = \gamma_n \Longrightarrow \mu_n = \frac{2 + \lambda\gamma_n}{2 - \lambda\gamma_n} = \frac{1 - 2\lambda\sin^2\frac{n\pi h}{2}}{1 + 2\lambda\sin^2\frac{n\pi h}{2}}$$

also $|\mu_n| < 1$ unabhängig von λ.

Daraus folgt, daß das CRANK-NICHOLSON-Verfahren unbedingt stabil ist.

4.5 Inhomogene Anfangswertaufgaben

Problem:

$$\begin{aligned} u_t &= Lu + q(t) \qquad t \in [0,T] \\ u(0) &= u_0 \\ u &\in D_L \qquad \text{(d.h. RB erfüllend)} \end{aligned} \tag{4.48}$$

Voraussetzungen wie oben.

Satz 4.5.1 Sei $q \in C^0[0,T]$, $U^{j+1} = CU^j$ ein konsistentes, stabiles Differenzenverfahren für $q \equiv 0, C = C(\tau), 0 < \tau \leq \tau_0$.
Sei $\{n_j, \tau_j\}$ eine Folge aus $\mathbb{N} \times \mathbb{R}_+$ mit $n_j\,\tau_j = \overline{t} \in [0,T]\ \forall\, j$ und $\tau_j \to 0$ mit $j \to \infty$.
Dann konvergiert die Lösung U^{n_j} der Differenzengleichung

$$\begin{aligned} U^{j+1} &= C(\tau_j)U^j + \tau_j\, q((j+1)\tau_j) \\ U^0 &= u_0|_{\text{Gitter}} \end{aligned}$$

mit $\tau_j \to 0$ gegen die verallgemeinerte Lösung von (4.48).

Beweis: und Einzelheiten siehe Meis/Marcowitz.

4.6 Parabolische ARWA im $\mathbb{R}^n, n > 1$

(kurz und beispielhaft)

4.6.1 Abstrakte Formulierung

$$\begin{aligned} u_t &= Lu && \text{in } \Omega \times \mathbb{R}_+, \Omega \subset \mathbb{R}^n \\ u(0) &= u_0 && \forall x \in \Omega \\ u(x,t) &= g(x) && \forall \in \geq 0, \forall x \in \partial\Omega \end{aligned} \tag{4.49}$$

Sei L gleichmäßig elliptisch, und Voraussetzung wie im elliptischen Fall an L, Ω erfüllt. Seien $C_\mu(\tau), \mu = 1, \ldots, n$, konsistente, stabile Differenzenverfahren im $\mathbb{R}^1$, also

$$\begin{aligned} \|C_\mu(\tau)\| &\leq 1 \\ \text{oder} &= 1 + O(\tau) \end{aligned}$$

Sei

$$C(\tau) := C_1(\tau) \circ C_2(\tau) \cdots \circ C_n(\tau). \tag{4.50}$$

Dann ist $C(\tau)$ auch konsistent und stabil, falls die $C_\mu(\tau)$ vertauschbar sind.

4.6.2 Konkrete Formulierung

$u_t = \Delta u$, $\Omega \subset \mathbb{R}^2$, $\lambda = \dfrac{\tau}{h^2}$, quadratisches x, y-Gitter.

Die einfachste zweidimensionale Taylor-Reihen-Approximation liefert:

$$U^{j+1} = \{1 + \lambda\delta_x^2 + \lambda\delta_y^2\}\, U^j \tag{4.51}$$

Mit der expliziten EULER-Methode ergibt sich für den Raumpunkt (x_r, y_s)

$$U_{r,s}^{j+1} = U_{r,s}^j + \lambda(U_{r-1,s}^j + U_{r+1,s}^j + U_{r,s-1}^j + U_{r,s+1}^j - 4U_{r,s}^j) \tag{4.52}$$

mit dem lokalen Diskretisierungsfehler $O(\tau + h^2)$. Aus (4.51) ergibt sich auch die zweidimensionale CRANK-NICHOLSON-Formel

$$\left(1 - \frac{1}{2}\lambda\delta_x^2 - \frac{1}{2}\lambda\delta_y^2\right) U^{j+1} = \left(1 + \frac{1}{2}\lambda\delta_x^2 + \frac{1}{2}\lambda\delta_y^2\right) U^j \tag{4.53}$$

Dies ist eine 5-Punkte-Formel.

Wird der gemischte Term $\frac{1}{4}\lambda^2\, \delta_x^2\, \delta_y^2$ auf jeder Seite hinzugenommen, so erhält man eine 9-Punkte-Formel und spricht auch von C/N:

$$\left(1 - \frac{1}{2}\lambda\delta_x^2\right)\left(1 - \frac{1}{2}\lambda\delta_y^2\right) U^{j+1} = \left(1 + \frac{1}{2}\lambda\delta_x^2\right)\left(1 + \frac{1}{2}\lambda\delta_y^2\right) U^j \tag{4.54}$$

Dieses Schema kann explizit gemacht werden durch Einführung einer Zwischenstufe

$$\begin{aligned} \left(1 - \frac{1}{2}\lambda\delta_x^2\right) U^{j+1/2} &= \left(1 + \frac{1}{2}\lambda\delta_y^2\right) U^j \\ \left(1 - \frac{1}{2}\lambda\delta_y^2\right) U^{j+1} &= \left(1 + \frac{1}{2}\lambda\delta_x^2\right) U^{j+1/2} \end{aligned} \tag{4.55}$$

Solche halbimpliziten Schemata wurden zuerst von Peaceman und Rachford 1955 eingeführt. Sie lösten damit Probleme der Modellierung von Ölreservoirs.

Aufwand

In einer Dimension hatte das Verfahren von Crank-Nicholson den Vorteil unbedingter Stabilität bei nur geringfügig höherem Aufwand. Die unbedingte Stabilität liegt für (4.53) auch bei zwei Raumdimensionen vor, allerdings wächst der Aufwand erheblich an. So ist z.B. für das Einheitsquadrat mit N^2 inneren Punkten ein lineares $N^2 \times N^2$-Gleichungssystem zu lösen. Das Verfahren (4.55) von Peaceman und Rachford benötigt hingegen für die bessere Approximation nur die Lösung von $2\,N$ Dreibandsystemen der Ordnung N. Unter Berücksichtigung der speziellen Struktur der Systeme ist das etwa nur ein Drittel des C/N-Aufwandes.

4.6.3 Stabilitäts-Test für (4.52)

Wir verallgemeinern die VON NEUMANN'sche FOURIER-Methode in der oben beschriebenen Weise:
Für $t = 0$ wird eine Störung $e^{i\beta x}e^{i\gamma y}$ angesetzt mit β, γ beliebig reell. Die Wirkung der Störung für $t > 0$ wird dann

$$e^{\alpha t}e^{i\beta x}e^{i\gamma y},\ \alpha = \alpha(\beta,\gamma) \in \mathbb{C},$$

und diese muß untersucht werden wie oben:

Forderung: $|e^{\alpha t}| \leq 1$.

Ganz analog zum Beispiel 4.4.1 erhält man:

$$\xi := e^{\alpha\tau} = 1 - 4\lambda\left(\sin^2\frac{\beta h}{2} + \sin^2\frac{\gamma h}{2}\right)$$

$\Longrightarrow$

$\lambda \leq \frac{1}{4}$ als Stabilitätsforderung.

Auch hier muß für variable Koeffizienten und/oder Einbeziehung der Randbedingungen die Matrix-Methode angewendet werden.

Kapitel 5

Lösung großer, dünn besetzter linearer Gleichungssysteme

Die numerische Lösung partieller Differentialgleichungen setzt sich aus vielen Anteilen zusammen. Bevor in diesem Kapitel die bei der Diskretisierung entstehenden großen Gleichungssysteme gelöst werden, sollen Rüdes Grundprinzipien zur effektiven Lösung von partiellen Differentialgleichungen einmal genannt werden, [Rüd93]:

- Gute Diskretisierungsmethoden (verschieden hoher Ordnung)
- Schnelle Lösungsmethoden für die entstehenden Gleichungssysteme
- Adaptivität
- Hochwertige Informatikanteile (Hardware, Algorithmen und Software)

Die wesentlichen Schritte vom kontinuierlichen zum diskreten Problem sind in den ersten vier Kapiteln vorgestellt worden. Dabei ist ein Problem angesprochen, aber ungelöst gelassen worden, nämlich das der Gittererzeugung. Mit ihm beschäftigen sich das nächste Kapitel und eine Spezialvorlesung, zu der sich auch ein Skriptum im Internet befindet. Was bleibt, ist die Lösung der bei der Diskretisierung entstehenden großen, dünn besetzten Gleichungssysteme. Die Adaptivität spielt bei beiden Problemklassen eine große Rolle, kann aber im Rahmen dieser Vorlesung nur am Rande behandelt werden. Dasselbe gilt für die Verwendung hochwertiger Hardware – wie Parallelrechner – und den Einsatz von Software, die strengen Anforderungen des Softwareengineering genügt, und die die Möglichkeiten spezieller Hardware algorithmisch nutzt.

Die letzten beiden Prinzipien sind nichtsdestowenigertrotz ausserordentlich wichtig für die effektive Lösung von realen Problemen. Leider ergänzen sich die Prinzipien nicht nur, sondern sie können sich auch behindern. So wird eine effektive parallele Lösung durch adaptive Ansätze erschwert. Und auch die mathematische Formulierung und Analyse der heute

Für dieses Kapitel setze ich die Kenntnis von "Stoer/Bulirsch: Numerische Mathematik 2" Kap. 8: 8.1 bis 8.4 voraus. Mindestens die Untermenge aus Numerik I/Köckler sollte bekannt sein.

schnellsten Methoden (Multilevel(Mehrgitter)-Methoden) wird schwieriger durch die Forderung der Adaptivität.

Der Versuch, alle wichtigen Kriterien für die effektive Lösung komplexer PDE-Probleme zu erfüllen, führt zur Gründung großer, interdisziplinärer Arbeitsgruppen, die Ingenieure, Informatiker und Mathematiker aus Hochschule und Industrie zusammenführt. An unserer Hochschule ist das z.B. der SFB 367 "Massive Parallelität".

Dieses Kapitel beginnt mit einem einführenden Abschnitt, der das Prinzip einer iterativen Approximation der Lösung und der Vorkonditionierung erklärt. Dieses Prinzip kann schon bei den bekannten Relaxationsverfahren gesehen werden. Wirklich wichtig und mit einer Vielzahl von Möglichkeiten in der Literatur beschrieben ist die Vorkonditionierung beim Verfahren der konjugierten Gradienten.

Für alle Methoden setzen wir die Symmetrie und positive Definitheit der Matrix des linearen Gleichungssystems voraus. Methoden für unsymmetrische Probleme werden in [Gre97] ausführlich beschrieben. Dort wird auch auf den Einfluss von Rundungsfehlern auf die Approximationsgüte eingegangen.

Höhepunkt und Abschluss dieses Kapitels bildet dann die Mehrgittermethode.

5.1 Einleitung

Es soll in diesem Kapitel das bei der Differenzen- oder der Finite-Elemente-Methode entstandene lineare Gleichungssystem gelöst werden:

$$Ax = b \quad \text{mit } A \in \mathbb{R}^{N,N}, \quad x, b \in \mathbb{R}^N. \tag{5.1}$$

Wenn das mit dem Cholesky-Verfahren ohne Berücksichtigung der dünnen Struktur geschehen würde, dann würden N^2 Speicherplätze benötigt und der Aufwand (die Komplexität) betrüge etwa $N^3/3$ Operationen. Bei Berücksichtigung einer Halbbandbreite m sind nur noch $(m+1)N$ Speicherplätze und etwa m^2N Operationen notwendig.

Wie wir oben gesehen haben, sind für große Probleme auch die Bandbreiten groß (mindestens $\sqrt{N}$) mit vielen Nullen innerhalb des Bandes. Die Nullen innerhalb des Bandes werden i.a. mit Elementen ungleich Null aufgefüllt (*fill-in*). Das führt von $O(N)$ Elementen ungleich Null zu $O(N\sqrt{N})$ (also z.B. von $5_{10}6$ zu $1_{10}9$).

Für Probleme dieser Größenordnung kommen daher zwangsläufig fast nur iterative Verfahren in Frage, die hauptsächlich aus Matrix×Vektor-Operationen bestehen. Bei diesen Operationen kann dann die dünne Struktur voll ausgenutzt werden, d.h. es wird keine Operation mit einem Element gleich Null ausgeführt.

Das Ziel bei diesen Algorithmen ist, Verfahren zu entwickeln, die für die behandelten Problemklassen *asymptotisch optimal* sind, d.h.

(1) Es werden nur $O(N)$ Operationen benötigt.

(2) Der Aufwand ist unabhängig von der Diskretisierungsgröße h.

Nur wenige Verfahren erfüllen diese Eigenschaft, und diese ist oft nur für Modellprobleme beweisbar, wird aber im Experiment für allgemeine Probleme bestätigt.

Liegt vor Beginn der Rechnung keine Näherung vor, dann liegt es nahe, eine Startnäherung

$$x^1 \in \text{ span}\{b\} \tag{5.2}$$

zu wählen. Iterative Verfahren, die die Minimierung des Residuums $b - Ax$ zum Ziel haben, berechnen dann in der Regel die erste Approximation

$$x^2 \in \text{ span}\{b,\, Ab\}. \tag{5.3}$$

Wird dieser Prozess fortgeführt, so werden Näherungslösungen im sogenannten *Krylov-Unterraum* berechnet:

$$x^k \in K^k := \text{ span}\{b,\, Ab, \ldots,\, A^{k-1}b\}, \quad k = 1, 2, \ldots \tag{5.4}$$

Natürlich stellen sich jetzt die Fragen:

(1) Wie gut ist die beste Approximation aus diesem Unterraum?

(2) Wie kann diese beste Approximation mit minimalem Aufwand und Speicher berechnet werden?

(3) Wie wird diese Berechnung durch Rundungsfehler beeinflusst?

Antworten auf diese Fragen sind nicht ermutigend. Deshalb lautet die nächste Frage:

(4) Kann das Problem so modifiziert werden, dass ein entsprechender Raum eine bessere Lösung enthält, die auch noch schneller berechnet werden kann?

Die positive Antwort auf diese Frage enthält das Wort *Vorkonditionierer (Preconditioner)*, und das modifizierte Problem ist entweder

$$M^{-1}Ax = M^{-1}b \tag{5.5}$$

oder im spd-Fall mit $M = L\,L^T$

$$L^{-1}AL^{-T}y = L^{-1}b, \quad x = L^{-1}y. \tag{5.6}$$

In beiden Fällen werden normalerweise keine Inversen ausgerechnet, sondern entsprechende Gleichungssysteme gelöst. M und L werden so konstruiert, dass der Aufwand dazu gering ist.

Algorithmen, die den Vektor x^k aus dem Krylov-Unterraum berechnen, der die kleinste euklidische Norm des Residuums hat

$$\|b - Ax^k\|_2 = \min_{x \in K^k} \|b - Ax\|_2 \tag{5.7}$$

heissen MINRES. Statt der 2-Norm des Residuums kann man die A-Norm des Fehlers

$$\|e_k\|_A := \|x^k - x\|_A = \|x^k - A^{-1}b\|_A = \sqrt{(Ax^k - b)^T A^{-1}(Ax^k - b)} \tag{5.8}$$

minimieren. Das Verfahren der *konjugierten Gradienten* erzeugt diese Lösung, siehe 5.4.1.

Für diese Algorithmen (MINRES und CG) besteht der Aufwand eines Iterationsschritte i.w. aus einer Matrix×Vektor-Operation. Die Frage, wie gut diese Algorithmen konvergieren, kann für spd-Matrizen mit Hilfe der Eigenwerte von A beantwortet werden. Für die MINRES-Algorithmen kann das Residuum $r_k = b - Ax^k$ wegen (5.4) geschrieben werden als

$$r_k = P_k(A)\, b, \tag{5.9}$$

wobei P_k ein gewisses Polynom k-ten Grades ist mit dem Wert 1 im Nullpunkt. Für jedes andere Polynom k-ten Grades p_k mit $p_k(0) = 1$ gilt

$$\|r_k\|_2 \leq \|p_k(A)b\|_2. \tag{5.10}$$

Sind Λ die Diagonalmatrix der Eigenwerte von A und Q die Modalmatrix, d.h. die Matrix der orthogonalen Eigenvektoren, dann gilt

$$A = Q\Lambda Q^T \tag{5.11}$$

und es folgt

$$\|r_k\|_2 \leq \|Q\, p_k(\Lambda)\, Q^T b\|_2 \leq \|p_k(\Lambda)\|_2\, \|b\|_2. \tag{5.12}$$

Da diese Ungleichung für alle Polynome p_k mit $p_k(0) = 1$ gilt, gilt auch

$$\frac{\|r_k\|_2}{\|b\|_2} \leq \min_{p_k} \max_{i=1,\ldots,n} |p_k(\lambda_i)|. \tag{5.13}$$

Für MINRES-Residuen ist diese Schranke scharf. Damit ist die Frage nach der Konvergenzgeschwindigkeit (Approximationsqualität) zurückgeführt worden auf ein Problem der Approximationstheorie: Wie gut kann ich die Null approximieren mit einem Polynom k-ten Grades auf der Menge der Eigenwerte von A, das im Nullpunkt den Wert 1 annimmt.

Will man diese Frage genau beantworten, bekommt man komplizierte Ausdrücke in allen Eigenwerten von A. Es gibt aber auch einfache Schranken mit Ausdrücken in wenigen Eigenwerten. Der wichtige Punkt ist, dass die Norm des Residuums ausschließlich von den Eigenwerten der Matrix A bestimmt wird, und dass es zumindest heuristische Ideen gibt über gute und schlechte Eigenwertverteilungen.

Mit derselben Argumentationskette kommt man bei den CG-Verfahren zu einer A-Norm-Abschätzung für $e_k = x^k - x = x^k - A^{-1}b$ in der A-Norm:

$$\frac{\|e_k\|_A}{\|A^{-1}b\|_A} \leq \min_{p_k} \max_{i=1,\ldots,n} \|p_k(\lambda_i)\|. \tag{5.14}$$

Auch diese Schranke ist scharf.

Die Konvergenztheorie für die MINRES- und CG-Verfahren ist also klar beschrieben. Diese Herleitungen setzen allerdings exakte Arithmetik voraus. Es ergibt sich ein abweichendes Konvergenzverhalten bei der Rechnung mit endlicher Arithmetik, also mit Rundungsfehlern.

5.2 Relaxationsverfahren (SOR)

5.2.1 Erinnerung und Grundlagen

Wir erinnern uns an folgende Begriffe:
irreduzibel, diagonal dominant, Spektralradius, Jacobi-Gesamtschritt-Verfahren (GSV),

Gauß-Seidel-Einzelschritt-Verfahren (ESV), konsistent geordnet.

Ein allgemeines Iterationsverfahren ist:

$$x^{i+1} = x^i + B^{-1}(b - Ax^i) = (I - B^{-1}A)x^i + B^{-1}b \tag{5.15}$$

mit $B = B(\omega)$ beliebig und

$$T(\omega) := (I - B^{-1}(\omega)A). \tag{5.16}$$

Hier kann also B als Vorkonditionierer aufgefasst werden. (5.15) konvergiert $\forall x^0 \in \mathbb{R}^N$, falls

$$K := \rho(I - B^{-1}A) = \rho(T(\omega)) < 1. \tag{5.17}$$

Dabei ist $\rho(\cdot)$ der Spektralradius der entsprechenden Matrix, der durch jede submultiplikative Norm nach oben abgeschätzt werden kann:

$$\rho(\cdot) \leq \|\cdot\|.$$

A sei aufgeteilt in Diagonal-, obere und untere Dreiecksmatrix:

$$A = D - E - F = \begin{pmatrix} \ddots & & -F \\ & D & \\ -E & & \ddots \end{pmatrix}.$$

Dann ist für die betrachtete Familie von Relaxationsverfahren

$$B := \frac{1}{\omega} D\,(I - \omega D^{-1}E).$$

Wähle $\omega \in (0, 2)$ so, daß $\rho(T(\omega))$ minimal wird.

Satz 5.2.1 (i) Es ist

$$\rho(T(\omega)) \geq |\omega - 1| \tag{5.18}$$

(ii) Ist A symmetrisch positiv-definit (spd), so ist

$$\rho(T(\omega)) < 1 \text{ für } 0 < \omega < 2. \tag{5.19}$$

(iii) A sei konsistent geordnet (siehe Stoer, [SB90]),
J sei die Iterationsmatrix des GSV:

$$J := D^{-1}E + D^{-1}F.$$

Die Eigenwerte von J seien reell und $\rho(J) < 1$.

Dann ist der optimale Relaxationsparameter ω_b gegeben als

$$\omega_b = \frac{2}{(1 + \sqrt{1 - \rho(J)^2})} \tag{5.20}$$

und es ist

$$\rho(T(\omega_b)) = \omega_b - 1 = \left(\frac{\rho(J)}{1 + \sqrt{1 - \rho(J)^2}}\right)^2 \tag{5.21}$$

5.2.2 Block-Relaxations-Verfahren

Alle in 5.2.1 kennengelernten Verfahren lassen sich in Blockform darstellen. Hierzu faßt man Teilmatrizen zu sogenannten Blöcken zusammen und ersetzt die Elemente bei der Verfahrensdefinition durch diese Blöcke, bei Division durch die Inverse. Sei also

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1K} \\ \vdots & \ddots & \vdots \\ A_{K1} & \cdots & A_{KK} \end{bmatrix}$$

mit quadratischen Matrizen A_{ii} und rechteckigen Matraizen A_{ij}, und sei $A = D - E - F$ mit $D := diag(A_{11}, \ldots, A_{KK})$

$$E := -\begin{bmatrix} 0 & & \ominus & \\ A_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ A_{K1} & \cdots & A_{K,K-1} & 0 \end{bmatrix}, F := -\begin{bmatrix} 0 & A_{12} & \cdots & A_{1K} \\ & \ddots & \ddots & \vdots \\ & & \ddots & A_{K-1,K} \\ & & & 0 \end{bmatrix},$$

$L := D^{-1}E$, $U := D^{-1}F$.

Dann lautet das Block-Jacobi-Verfahren

$$Dx^{i+1} = b + (E + F)x^i, i = 0, 1, \cdots$$

oder

$$A_{jj}x_j^{i+1} = b_j + \sum_{k \neq j} A_{jk}x_k^i, \qquad j = 1, \ldots, K, \quad i = 0, 1, \ldots.$$

Dabei sind x^j und b_j Vektorabschnitte der gleichen Dimension wie A_{jj}. Es müssen also pro Iterationsschritt K "kleine" Gleichungssysteme gelöst werden.

Bemerkung 5.2.2 (1) Sind die A_{jj} spd, so löst man diese mit dem Cholesky- Verfahren.

(2) Oft sind alle $A_{jj} \equiv M$, dann spart man Speicher und Rechenzeit und zerlegt nur einmal.

(3) Die Konvergenzgeschwindigkeit wird bestimmt vom Spektralradius

$$\rho(J) = \rho(L + U),$$

der i.a. kleiner ist als beim elementweisen Verfahren, da in B mehr Information über A steckt, oder: B ist ein besserer Vorkonditionierer.

Block-Gauß-Seidel

$$A_{jj}x_j^{i+1} = b_j - \sum_{k<j} A_{jk}x_k^{i+1} - \sum_{k>j} A_{jk}x_k^i \qquad (5.22)$$
$$j = 1, \ldots, K, \qquad i = 0, 1, \ldots$$

Block-SOR

$$\begin{aligned} A_{jj}x_j^{i+1} &= \omega(b_j - \sum_{k<j} A_{jk}x_k^{i+1} - \sum_{k>j} A_{jk}x_k^i) \\ &+ (1-\omega)x_j^i, \qquad j = 1,\ldots,K, \quad i = 0,1,\ldots. \end{aligned} \tag{5.23}$$

Zwischen den Gauß-Seidel- und den Block-SOR-Iterierten besteht folgende Beziehung:

$$x_{j,\mathrm{SOR}}^{(i+1)} = \omega\,(x_{j,\mathrm{ESV}}^{(i+1)} - x_{j,\mathrm{SOR}}^{(i)}) + x_{j,\mathrm{SOR}}^{(i)}, \qquad j = 1,\ldots,K.$$

Die Konvergenzgeschwindigkeit wird wieder bestimmt von

$$\rho(T(\omega)) \text{ mit } T(\omega) := (I - \omega L)^{-1}[(1-\omega)I + \omega U]$$

Dabei läßt sich die Theorie der konsistent geordneten Matrizen übertragen: A heißt jetzt konsistent geordnet, falls die Eigenwerte von

$$J(\alpha) := \alpha L + \alpha^{-1}U$$

unabhängig von α sind. Dann lassen sich optimale Relaxationsparameterwerte wie in Satz 5.2.1 bestimmen mit Hilfe von $\rho(J)$. Für ein Modellproblem ergibt sich:

$$\rho(J_{\mathrm{Block}}) = \frac{\cos\frac{\pi}{K+1}}{2 - \cos\frac{\pi}{K+1}} < \rho(J_{\mathrm{Element}})$$

und

$$\rho(T_{\mathrm{Block}}(\omega_b)) \approx \rho(T_{\mathrm{Element}}(\omega_b))^{\sqrt{2}},$$

d.h. die Zahl der Iterationen reduziert sich etwa um den Faktor $\sqrt{2}$.

Noch schnellere Verfahren ergeben sich, wenn man die Problemklasse noch stärker einschränkt. Das bekannteste Modellproblem ist die Poisson-Gleichung im Quadrat:

$$\left.\begin{aligned} -\Delta u = q \quad & \text{in } \Omega = (0,1)\times(0,1) \\ u = g \quad & \text{auf } \partial\Omega \end{aligned}\right\} \tag{5.24}$$

Verfahren, die sich (zunächst) auf dieses Problem beschränken, nennt man deshalb auch Fast Poisson Solvers oder allgemeiner Rapid Elliptic Solvers (RES).

5.3 Verfahrensvergleich für $-\Delta u = q$

Löst man Problem (5.24) auf einer IBM 370/158 mit $h = 1/256$, d.h. mit $N = 255^2 = 65025$ Unbekannten, so bekommt man etwa folgenden Aufwand:

Verfahren	asymptotischer Rechenaufwand	Rechenzeit [sec]
Gauß-Seidel	$\sim N^2$	ca. 100.000
SOR	$\sim N^{\frac{3}{2}}$	ca. 1.200
ADI $*$	$\sim NlogN$	ca. 130
Buneman $*$	$\sim NlogN$	15,5
FACR-Verfahren $*$	$\sim Nlog(logN)$	8,1
Mehrgitter-Verfahren	$\sim N$	7,5

Dabei sind die iterativen Verfahren auf eine relative Genauigkeit von $_{10} - 4$ festgelegt. Konvergenzgeschwindigkeit und Durchführung der mit $*$ gekennzeichneten Verfahren sind nur für das Modellproblem realisierbar.

5.4 Vorkonditionierte cg-Verfahren

5.4.1 Das Verfahren der konjugierten Gradienten

Zu lösen ist wieder das lineare Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{n,n}$ symmetrisch, positiv definit (spd). Die Lösung von $Ax = b$ erfüllt $\min\limits_{x \in \mathbf{R}^n} F(x) = 0$, wenn

$$F(x) := \frac{1}{2}(Ax - b)^T A^{-1}(Ax - b), \tag{5.25}$$

und diese Funktion wird vom cg-Verfahren formal minimiert.

Dazu führt das Verfahren im Schritt $x_k \to x_{k+1}$ eine $(k+1)-$dimensionale Minimierung durch:

$$\left.\begin{array}{ll} x_0 & \text{gegeben} \\ x_{k+1} \text{ aus} & F(x_{k+1}) = \min\limits_{u_0,\ldots,u_k} F(x_k + u_0 r_0 + \cdots + u_k r_k) \\ \text{mit} & r_i := b - Ax_i, \quad i \le k. \end{array}\right\} \tag{5.26}$$

Die durch diesen Algorithmus entstehenden Residuen r_k sind orthogonal, solange $r_k \neq 0$; d.h. man braucht höchstens n Schritte, bis $r_k = 0$ und $x_k = x$. Rechnerisch wird (5.26) folgendermaßen erreicht:

CG – Algorithmus (5.27)

Wähle $x_0 \in \mathbb{R}^n$ beliebig.

$p_0 := r_0 := b - Ax_0$.
Falls $r_0 = 0$: STOP: $Ax_k = b$.
sonst:
Für $k = 0, 1, 2, \ldots$

$$a_k := \frac{(r_k, r_k)}{(p_k, Ap_k)} \quad (a_k \in \mathbb{R})$$

$$x_{k+1} := x_k + a_k p_k$$

$$r_{k+1} := r_k - a_k Ap_k$$

Falls $r_{k+1} = 0$: STOP: $Ax_{k+1} = b$.

sonst:

$$b_k := \frac{(r_{k+1}, r_{k+1})}{(r_k, r_k)} \quad (b_k \in \mathbb{R})$$

$$p_{k+1} := r_{k+1} + b_k p_k$$

In einem Programm wird die Abfrage auf $r_{k+1} = 0$ durch eine auf $\|r_{k+1}\| < \varepsilon$ ersetzt.

Aufwand pro Schritt: 1 Matrix-Vektor-Multiplikation, 6 Skalarprodukte.

Satz 5.4.1 $A \in \mathbb{R}^{n,n}$ spd, $b \in \mathbb{R}^n, x_0 \in \mathbb{R}^n$ beliebig.

Dann gibt es eine kleinste Zahl $l, 0 \leq l \leq n$, mit $p_l = 0$.

Für $x_k, p_k, r_k, k \leq l$, nach (5.27) gilt:

$$\begin{array}{lll}
(a) & Ax_l = b & \\
(b) & r_i^T p_j = 0 & 0 \leq j < i \leq l \\
(c) & r_i^T p_i = r_i^T r_i & i \leq l \\
(d) & p_i^T Ap_j = 0 & 0 \leq i < j \leq l \\
 & p_j^T Ap_j > 0 & j < l \\
(e) & r_i^T r_j = 0 & 0 \leq i < j \leq l \\
 & r_j^T r_j > 0 & j < l \\
(f) & r_i = b - Ax_i & i \leq l
\end{array}$$

Der Satz besagt auch, daß das Verfahren wohldefiniert ist, weil für $p_k \neq 0$ auch $r_k^T r_k > 0$ und $p_k^T Ap_k > 0$ gilt.

Wegen (d) nennt man die Vektoren p_k A-konjugiert, was dem Verfahren seinen Namen gegeben hat.

Beweis: Folgende Eigenschaften werden durch vollständige Induktion nach $k \leq l$ gezeigt:

$$(A_k) \begin{cases} (1) & r_i^T p_j = 0 & \text{für } j < i \leq k & \Longrightarrow & \text{(b)} \\ (2) & r_i^T r_i > 0 & i < k & \Longrightarrow & \text{(e)(2)} \\ & r_i^T p_i = r_i^T r_i & i \leq k & \Longrightarrow & \text{(c)} \\ (3) & p_i^T A p_j = 0 & i < j \leq k & \Longrightarrow & \text{(d)(1)} \\ (4) & r_i^T r_j = 0 & i < j \leq k & \Longrightarrow & \text{(e)(1)} \\ (5) & r_i = b - A x_i & i \leq k & \Longrightarrow & \text{(f)} \end{cases}$$

(d)(2) ist trivial, (a) folgt.

Induktionsvor.: (A_0) ist trivial.

Induktionsann.: (A_k) sei richtig für ein $0 \leq k < l$.

Induktionsbew.: Zeige (A_{k+1}):

zu (1): $r_{k+1}^T p_k = (r_k - a_k A p_k)^T p_k = r_k^T r_k - \dfrac{r_k^T r_k}{p_k^T A p_k} p_k^T A p_k = 0$

Entsprechend für $j < k$:

$r_{k+1}^T p_j = (r_k - a_k A p_k)^T p_j = 0$ wegen (A_k), (1)/(3).

$\Rightarrow$ (1).

zu (2): Annahme: $r_k^T r_k = 0 \Rightarrow r_k = 0$

$$\Rightarrow p_k = \begin{cases} r_0 & \text{falls } k = 0 \\ b_{k-1} p_{k-1} & \text{sonst} \end{cases}$$

Das ist ein Widerspruch im Fall $k = 0$. Für $0 < k < l$ führt $p_k = b_{k-1} p_{k-1}$ auch zum Widerspruch:

$$0 <_{(A \text{ spd})} p_k^T A p_k = b_{k-1} p_k^T A p_{k-1} =_{(A_k)(3)} 0 \Longrightarrow \text{ Widerspruch!}$$

Also ist $r_k^T r_k > 0$ und b_k, p_{k+1} sind wohldefiniert. Daraus folgt mit (5.27)

$$r_{k+1}^T p_{k+1} = r_{k+1}^T (r_{k+1} + b_k p_k) =_{(A_{k+1})(1)} r_{k+1}^T r_{k+1}.$$

$\Rightarrow$ (2)

zu (3): $r_k \neq 0 \Rightarrow a_k^{-1}$ existent. Daraus folgt mit (5.27) für $j \leq k$:

$$p_{k+1}^T A p_j =_{(5.27)} r_{k+1}^T A p_j + b_k p_k^T A p_j = a_j^{-1} r_{k+1}^T (r_j - r_{j+1}) + b_k p_k^T A p_j$$

$$[(5.27) : -A p_j = a_j^{-1} (r_{j+1} - r_j)]$$

$$\underbrace{=_{(5.27)}}_{j<k:} a_j^{-1} \underbrace{r_{k+1}^T (p_j - b_{j-1} p_{j-1} - p_{j+1} + b_j p_j)}_{\text{alles Null wg. (1)}} + \underbrace{b_k p_k^T A p_j}_{\text{Null wg. } (A_k)/(3)} .$$

Für $j = k$ setzen wir die Ausdrücke für a_k und b_k ein und beachten, daß aus der Klammer oben nur $p_{j+1} = p_{k+1}$ etwas ungleich Null ergeben kann:

$$p_{k+1}^T A p_k = -\left(\frac{r_k^T r_k}{p_k^T A p_k} \right)^{-1} \underbrace{r_{k+1}^T p_{k+1}}_{= r_{k+1}^T r_{k+1} \text{ wegen (2)}} + \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k} p_k^T A p_k = 0$$

Daraus folgt (3).

zu (4):

$$r_i^T r_{k+1} =_{(5.27)} (p_i - b_{i-1}p_{i-1})^T r_{k+1} =_{(1)} 0$$
$$\text{für } i \leq k. \text{ (Mit } p_{-1} := 0)$$

zu (5):

$$\begin{aligned} b - Ax_{k+1} &= b - A(x_k + a_k p_k) \\ &= \underbrace{b - Ax_k}_{r_k} - a_k A p_k =_{(5.27)} r_{k+1} \end{aligned}$$

$\Rightarrow (A_{k+1})$.

Also gilt (A_k) für $0 \leq k \leq l$. Wegen (A_l) (2) gilt $r_i \neq 0$ für $i < l$.

Mit (4) folgt: $\{r_i\}_{i=0}^{l-1}$ bilden ein Orthogonalsystem. Daraus folgt $l \leq n$.

Aus $p_l = 0$ folgt mit $(A_l)(2)$: $r_l^T r_l = r_l^T p_l = 0 \Rightarrow r_l = 0. \Rightarrow Ax_l = b.$ ■

Bemerkung 5.4.2 Man sieht jetzt leicht (Gradienten bilden, siehe auch Stoer), daß

$$\begin{aligned} S_k &:= \{u_0 r_0 + \cdots + u_k r_k \mid u_i \in \mathbb{R}\} \\ &= \{\rho_0 p_0 + \cdots + \rho_k p_k \mid \rho_i \in \mathbb{R}\} \end{aligned} \tag{5.28}$$

und damit

$$\min_{\rho_0,\ldots,\rho_k} F(x_k + \rho_0 p_0 + \cdots + \rho_k p_k) = \min_{u_0,\ldots,u_k} F(x_k + u_0 r_0 + \cdots + u_k r_k) = F(x_{k+1}). \tag{5.29}$$

5.4.2 Konvergenzgeschwindigkeit

Sei

$$\|u\|_A := \sqrt{u^T A u} \tag{5.30}$$

und

$$e_j := x_j - x, \qquad (Ax = b),$$

also

$$r_0 = -Ae_0.$$

Sei weiter

$$c = cond_2(A) = \frac{\lambda_1}{\lambda_n},$$

wenn $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ (A spd) die Eigenwerte von A sind. Dann gilt

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 2\left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^k, \tag{5.31}$$

d.h. das cg-Verfahren konvergiert um so schneller, je kleiner die Konditionszahl von A ist. Diese Beobachtung führt zu der Idee der Vorkonditionierung.

5.4.3 Vorkonditionierung - die Idee

Wegen (5.31) sucht man eine spd-Matrix B, sodaß $B^{-1}A \approx I$ und das Gleichungssystem $Bq = v$ leicht lösbar ist (wie bei der Erzeugung von Relaxationsverfahren).

Eine spd-Matrix B kann man immer symmetrisch zerlegen zu

$$B = B^{\frac{1}{2}} B^{\frac{1}{2}} \quad \text{bzw.} \quad B = L L^T.$$

Sei also

$$A' = B^{-\frac{1}{2}} A B^{-\frac{1}{2}} (= B^{\frac{1}{2}} (B^{-1} A) B^{-\frac{1}{2}}), \text{ bzw. } A' = L^{-1} A L^{-T} \quad \text{(vgl. (5.6))}. \tag{5.32}$$

Dann soll $cond(A') << cond(A)$. Das cg-Verfahren wird jetzt auf $A'x' = b'$ angewendet. Dazu berechnet man nicht A' explizit, sondern schreibt das cg-Verfahren entsprechend

$$r'_k = b' - A'x'_k = B^{-\frac{1}{2}} r_k, p'_k = B^{\frac{1}{2}} p_k \quad \text{um:}$$

5.4.4 PCG-Verfahren

(Preconditioned conjugate gradients method)

PCG – Algorithmus

$x_0 \in \mathbb{R}^n$ beliebig.

$r_0 := b - Ax_0$, $p_0 := q_0 := B^{-1} r_0$ (Gleichungssystem lösen)

Für $k = 0, 1, 2, \ldots$:

Falls $p_k = 0$: STOP: $Ax_k = b$.

sonst:

$$a_k := \frac{r_k^T q_k}{p_k^T A p_k}$$

$$x_{k+1} := x_k + a_k p_k$$

$$r_{k+1} := r_k - a_k A p_k$$

$q_{k+1} := B^{-1} r_{k+1}$ (Gleichungssystem lösen)

$$b_k := \frac{r_{k+1}^T q_{k+1}}{r_k^T q_k}$$

$$p_{k+1} := q_{k+1} + b_k p_k$$

In einem Programm wird die Abfrage auf $p_k = 0$ durch eine auf $\|p_k\| < \varepsilon$ ersetzt.

Zusätzlich zur Matrix-Vektor-Multiplikation und 6 Skalarprodukten muß man also jetzt pro Schritt 1 Gleichungssystem lösen, jedoch immer mit derselben und einfach strukturierten Matrix B.

5.4.5 Zur Wahl von B

(a) Symmetrisches SOR-Verfahren: SSOR

$$B = \frac{1}{2-\omega}\left(\frac{1}{\omega}D - E\right)\left(\frac{1}{\omega}D\right)^{-1}\left(\frac{1}{\omega}D - E^T\right)$$

mit $\omega \in (0,2)$, D, E w.o. bei den Relaxationsverfahren, $E^T = F$.

(Mehr über die SSOR im Hackbusch: *Iterative Lösung großer Gleichungssysteme*, [Hac91], der Aufwand ist ähnlich wie der der SOR)

Die Lösung $Bq = v$ entspricht also dem Aufwand eines SSOR - Schritts.

(b) Unvollständige Cholesky-Zerlegung

Voraussetzung: A $M-$Matrix, d.h. $a_{ij} \leq 0$ für $i \neq j$ und $A^{-1} \geq 0$ elementweise. Zerlege A nach Cholesky mit getrennter Speicherung der Diagonalelemente in D, aber setze gewisse Elemente in L gleich Null; z.B. akzeptiere nur eine Dreibandmatrix B.

Oder mit

$$B = LDL^T$$

Für $i = 1, 2, \ldots, n$:

$$d_i := a_{ii} - \sum_{k=1}^{i-1} d_k l_{ik}^2$$

Für $j = i+1, \ldots, n$:

$$d_i\, l_{ji} = \begin{cases} 0 \text{ falls } |i-j| > 1, \\ a_{ji} - \sum\limits_{k=1}^{i-1} d_k l_{jk} l_{ik}, \text{ sonst.} \end{cases}$$

Das ist das übliche Cholesky-Verfahren mit der Ausnahme, daß man an "verbotenen" Stellen die Cholesky-Matrix-Elemente Null setzt.

Bei beiden Methoden gelingt es, die Konditionszahl von A für das bekannte Modellproblem von $O(h^{-2})$ auf $O(h^{-1})$ zu reduzieren.

(c) Vorkonditionierung durch Gebietszerlegung:
Komplizierte Methode nach Bramble/Pasciak/Schatz, die in das Variationsprinzip für elliptische PDE eingreifen und dort die Bilinearform $A(\cdot,\cdot)$ durch $B(\cdot,\cdot)$ ersetzen, wobei in $B(\cdot,\cdot)$ ein vereinfachtes Variationsprinzip und die Gebietszerlegung einfließen. Darüber wird dann das cg-Verfahren gelegt.

5.5 Mehrgittermethoden – Multigrid-Methods (MGM)

Die MGM vereinen die Allgemeinheit der langsamen Verfahren wie SOR mit der Schnelligkeit der "Fast Poisson Solvers", die nur auf sehr spezielle Probleme anwendbar sind. Sie sind also das Ei des Kolumbus.

Die Grundidee geht zurück auf voneinander unabhängige Arbeiten von Federenko, 1962, Bakhalov, 1964, und Achi Brandt, 1970. Weiterführende grundlegende Arbeiten aus der Pionierzeit schrieben Wolfgang Hackbusch und Ulrich Trottenberg etwa ab 1975, siehe die Proceeding [HT82]. Eine schöne kleine Einführung ist das "Multigrid Tutorial" von Briggs, 1987, [Bri87]. Sehr viel ausführlicher, aber auch mit vielen schönen Beispielen und Zeichnungen ist das grundlegende Werk von Hackbusch, [Hac85]. Eine einführende Darstellung auf dem neuesten Stand findet man bei Wesseling, [Wes92].

Erweiterung und Alternative zu den MGM sind die Multilevel-Verfahren, die auch eine optimale Parallelisierung erlauben (also sozusagen das Doppelei des Kolumbus). Kurz gehaltene und praktisch orientierte Einführungen in diese Methoden sind die Bücher von McCormick, [McC89], und Ulrich Rüde ("Mathematical and Computational Techniques for Multilevel Adaptive Methods", nicht im Literaturverzeichnis).

5.5.1 Die Idee der MGM

Die MGM basiert auf folgenden Beobachtungen:

(i) Wenige Schritte eines Relaxationsverfahren (GSV, ESV, SOR) glätten den Fehler, d.h.:

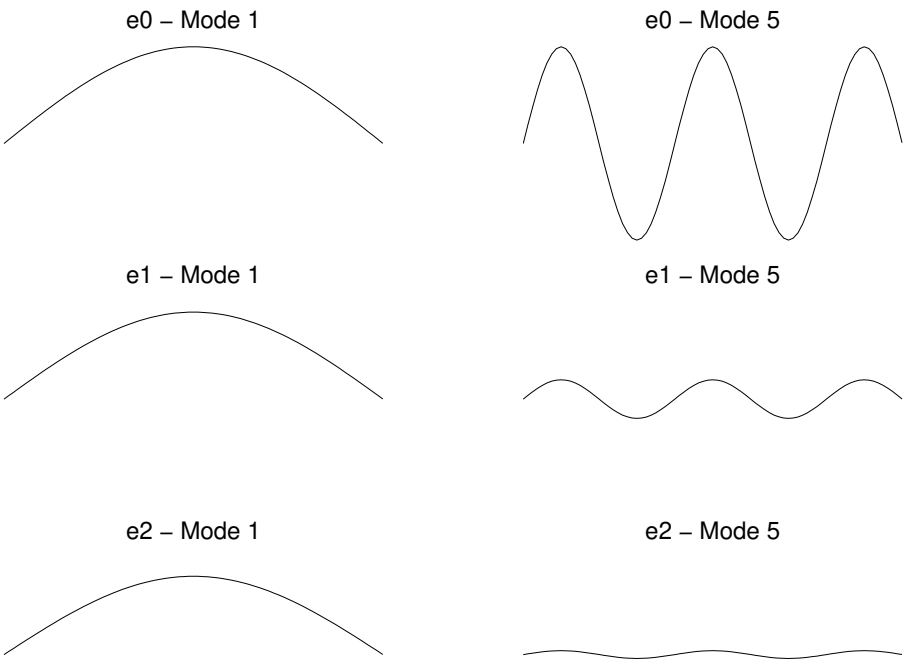
Sei $e_k := U - U^k$ der Fehler der k-ten Iterierten. Wir stellen uns eine Fourier-Entwicklung dieses Fehlers vor:

$$e_k(x) = \sum_{l=1}^{\infty} \alpha_l^{(k)} \exp(i\, l\, x), \quad i := \sqrt{-1}.$$

Bei Vektoren ($x = x_m, m = 1, \ldots, n$) entspricht die wachsende Frequenz der entsprechenden Zahl von Vorzeichenwechseln. Dann gilt für die Fourierkoeffizienten

$$\alpha_l^{(k)} << \alpha_j^{(k)}, \text{ wenn } l >> j.$$

Das heißt, daß niederfrequente Fehlerkomponenten schwächer gedämpft werden durch einen Schritt des Relaxationsverfahrens als hochfrequente Fehlerkomponenten:



Fehlerglättung durch Relaxation
(siehe auch Übungsserie 9)

(ii) Betrachtet man verschieden feine Gitter:

Maschenweite H: grobes Gitter,
Maschenweite h: feines Gitter,

dann kann man wie folgt vergleichen:

Grobes Gitter H	feines Gitter h
Wenige Punkte	Viele Punkte
Geringer Aufwand	Hoher Aufwand
Geringe Genauigkeit bei PDE-Lösung	Hohe Genauigkeit
Keine hochfrequenten Lösungsanteile	Nieder- und hochfrequente Lösungsanteile
Gute Approximation der niederfrequenten Lösungsanteile bei billiger exakter Lösung	gute Glättung der hochfrequenten Fehleranteile bei Relaxation

Diese Beobachtungen führen zu einer Verfahrenskombination mit grobem und feinem Gitter:

(1) Wenige Relaxationsschritte auf feinem Gitter h:
$\rightarrow$ Glättung des Fehlers.

(2) Restriktion der erhaltenen Näherungslösung auf Grobgitter H.

(3) Lösung der Defekt-Gleichung auf grobem Gitter. ($\sim$ Nachiteration mit Hilfe des Residuums!)

Bemerkung 5.5.1 (i) Praktisch arbeitet man nicht mit 2 Gittern, sondern mit einer Gitterfolge. Auf allen feineren Gittern wird relaxiert, nur auf dem gröbsten Gitter wird gelöst.

(ii) Es gibt jetzt verschiedene mögliche Verfahrenskombinationen mit folgenden algorithmischen Elementen:

Verfahrensteil	Gitter	Operator
Relaxation	h	S_h
Restriktion	$h \to H$	I_h^H
Lösung der Defektgleichung	H	L_H^{-1}
Interpolation	$H \to h$	I_H^h

Ziel der Kombination:

- Von h unabhängige Fehlerreduktion pro Gesamtschritt.
- Aufwand proportional zu N.

(iii) Relaxation ohne MGM liefert schlechte Konvergenz:

$$\begin{array}{ll} \text{Gauß-Seidel-ESV:} & \rho(T) = 1 - O(h^2) \\ \Longrightarrow & e_k \approx (1 - O(h^2))^k \, e_0 \\ \text{SOR } (\omega \text{ optimal}): & \rho(T) = 1 - O(h) \\ \Longrightarrow & e_k \approx (1 - O(h))^k \, e_0 \end{array}$$

5.5.2 Zweigittermethode als Grundlage der MGM

Es soll eine diskretisierte elliptische PDE

$$L_h u_h = f_h \tag{5.33}$$

iterativ gelöst werden mit den Näherungen $u_h^0, u_h^1, \cdots$, und es sei $H := 2h$.

Algorithmus

Schritt von u_h^j auf u_h^{j+1} :

(1) ν_1 Relaxationsschritte : $\bar{u}_h^j := S_h^{\nu_1} u_h^j$.

(2) Defektberechnung: $d_h := f_h - L_h \bar{u}_h^j$.

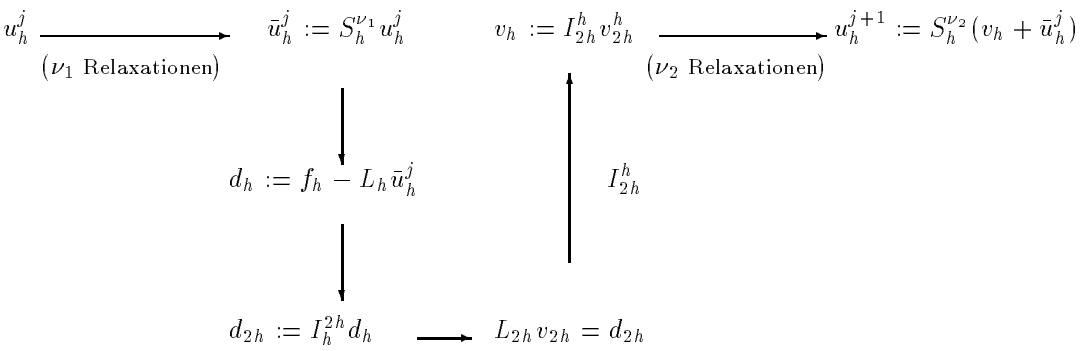
(3) Restriktion: $d_{2h} := I_h^{2h} d_h$.

(4) Lösung der Defektgleichung: $L_{2h} v_{2h} = d_{2h}$.

(5) Interpolation: $v_h := I_{2h}^h v_{2h}$.

(6) ν_2 Relaxationsschritte: $u_h^{j+1} := S_h^{\nu_2}(v_h + \bar{u}_h^j)$.

Schema eines Gesamtschrittes $u_h^j \to u_h^{j+1}$:



Das fassen wir zu einer Operatorgleichung zusammen:

$$u_h^{j+1} = S_h^{\nu_2}[\underbrace{S_h^{\nu_1} u_h^j}_{\bar{u}_h^j} + I_{2h}^h\, L_{2h}^{-1}\, I_h^{2h}(f_h - L_h \underbrace{S_h^{\nu_1} u_h^j}_{(1)})]. \tag{5.34}$$

(2), (3), (4), (5)

Der Iterationsoperator für dieses Verfahren ist also ($I_h :=$ Identität):

$$M_h^{2h} = S_h^{\nu_2} \underbrace{(I_h - I_{2h}^h L_{2h}^{-1} I_h^{2h} L_h)}_{=: K_h^{2h}} S_h^{\nu_1} \tag{5.35}$$

Jetzt müssen wir die Verfahrenskombination so wählen, daß $\rho(M_h^{2h})$ minimal wird. Beim Modellproblem ergibt sich etwa $0.01 \le \rho(M_h^{2h}) \le 0.1$.

Bemerkung: Übergang zum Mehrgitterverfahren:
Schneiden Sie den Algorithmus bei (4) auf und lösen Sie die Defektgleichung nicht genau, sondern durch ein Zweigitterverfahren, etc. $\cdots$.
Dabei entstehen zusätzliche Möglichkeiten für Varianten des Verfahrens, s.u.

5.5.3 Zwei einfache Beispiele

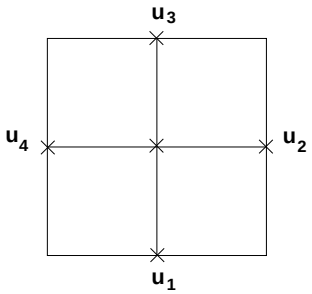
Beispiel 5.5.2 Ein einfaches und ungewöhnliches Mehrgittermodell (nach [Wei90])

$$\Delta u \;=\; 0 \text{ in } (0,1)^2$$

$$u = \begin{cases} x^2 & \text{für} \quad y = 0 \\ x^2 - 1 & \text{für} \quad y = 1 \\ -y^2 & \text{für} \quad x = 0 \\ 1 - y^2 & \text{für} \quad x = 1 \end{cases}$$

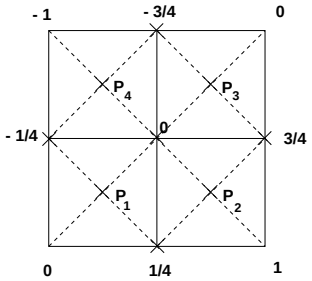
Lösung dieses Problems ist $u(x, y) = x^2 - y^2$.

1. Gitter:



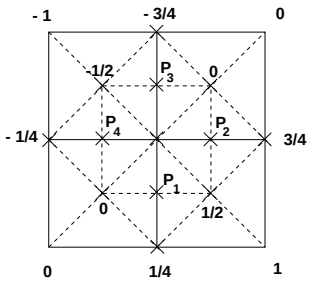
$$u^{(1)}(0.5, 0.5) = \frac{1}{4}(u_1 + u_2 + u_3 + u_4) = \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4} + \left(-\frac{3}{4}\right) - \frac{1}{4}\right) = 0$$

2. Gitter:



$$u^{(2)}(P_1) = 0, \;\; u^{(2)}(P_2) = \frac{1}{2}, \;\; u^{(2)}(P_3) = 0, \;\; u^{(2)}(P_4) = -\frac{1}{2}.$$

3. Gitter:



$$u^{(3)}(P_1) = \frac{3}{16}, \;\; u^{(3)}(P_2) = \frac{5}{16}, \;\; u^{(3)}(P_3) = -\frac{5}{16}, \;\; u^{(3)}(P_4) = -\frac{3}{16}$$

Alle bisher so berechneten Lösungswerte stimmen mit der analytischen Lösung überein.

Das liegt daran, daß der Diskretisierungsfehler hier $O(h^4)(u_{xxxx}(\tilde{x}, \tilde{y}) + u_{yyyy}(\tilde{\tilde{x}}, \tilde{\tilde{y}}))$ und die 4. Ableitungen verschwinden für alle Polynome vom Hgrd. 3!

Eine gute Startnäherung liefert ein solches Verfahren aber allemal.

Beispiel 5.5.3 Ein eindimensionales Beispiel (nach [Wes92])

Betrachte im $\mathbb{R}$ das Randwertproblem der gewöhnlichen DGL

$$\begin{aligned} -u'' &= f(x) \qquad \text{in } \Omega = (0,1) \\ u(0) &= u_x(1) = 0 \end{aligned}$$

Das (feine) Gitter sei

$$x_j = jh, j = 1, 2, \ldots, 2n, h = \frac{1}{2n}$$

Mit den üblichen Differenzen bekommen wir folgendes Gleichungssystem

$$\begin{aligned} 2u_1 - u_2 &= h^2 f_1 \\ -u_{j-1} + 2u_j - u_{j+1} &= h^2 f_j, \quad j = 2, 3, \ldots, 2n-1 \\ -u_{2n-1} + u_{2n} &= \frac{h^2}{2} f_{2n} \end{aligned} \tag{5.36}$$

(Die letzte Gleichung ist die Kombination der diskretisierten DGL im rechten Randpunkt und der rechten Randbedingung). (5.36) können wir in der Form

$$Au = f \tag{5.37}$$

schreiben. Das ESV mit einem Anfangsvektor u^0 lautet:

$$\begin{aligned} 2u_1^j &= u_2^{j-1} + h^2 f_1 \\ -u_{k-1}^j + 2u_k^j &= u_{k+1}^{j-1} + h^2 f_k, \quad k = 2, 3, \ldots, 2n-1 \\ -u_{2n-1}^j + u_{2n}^j &= \frac{h^2}{2} f_{2n} \end{aligned} \tag{5.38}$$

wobei j den Iterationsschritt kennzeichnet. Algorihmisch aufgelöst wird das zu:

$$\begin{aligned} u_1^j &= \frac{1}{2}(u_2^{j-1} + h^2 f_1) \\ u_k^j &= \frac{1}{2}(u_{k-1}^j + u_{k+1}^{j-1} + h^2 f_k), \quad k = 2, 3, \ldots, 2n-1 \\ u_{2n}^j &= u_{2n-1}^j + \frac{h^2}{2} f_{2n} \end{aligned} \tag{5.39}$$

ν Schritte dieses Verfahrens mit Anfangswert u^0 bezeichnen wir mit $S(u^0, A, f, \nu)$.

Fourier Analysis

(ganz entsprechend dem von Neumann'schen Stabilitätstest!)

Für die Fehleranalysis nehmen wir periodische Randbedingungen an

$$u(1) = u(0).$$

Dann ist auch der Fehler $e^j = u^j - u^\infty$ periodisch und erfüllt

$$\begin{aligned} -e_{k-1}^j + 2e_k^j &= e_{k+1}^{j-1}, \\ e_k^j &= e_{k+2n}^j. \end{aligned} \tag{5.40}$$

Fourier-Reihe für den Fehler

$$e_k^j = \sum_{\alpha=0}^{2n-1} c_\alpha^j e^{ik\Theta_\alpha}, \Theta_\alpha = \frac{\pi\alpha}{2n}, i = \sqrt{-1}$$

Da die $[e^{ik\Theta_\alpha}]$ linear unabhängig, ja sogar orthogonal sind, genügt es, einen Term in (5.40) einzusetzen:

$$\begin{aligned} e_k^{j-1} &= c_\alpha^{j-1} \exp(ik\Theta_\alpha) \\ e_k^j &= c_\alpha^j \exp(ik\Theta_\alpha) \\ e_{k-1}^j &= c_\alpha^j \exp(i(k-1)\Theta_\alpha) \end{aligned}$$

Damit wird

$$-e_{k-1}^j + 2e_k^j = e_{k+1}^{j-1}$$

zu

$$-c_\alpha^j e^{i(k-1)\Theta_\alpha} + 2c_\alpha^j e^{ik\Theta_\alpha} = c_\alpha^{j-1} e^{i(k+1)\Theta_\alpha}$$

und damit nach Auflösung und Kürzen von $e^{ik\Theta_\alpha}$

$$c_\alpha^j = \underbrace{\frac{\exp(i\,\Theta_\alpha)}{2 - \exp(-i\,\Theta_\alpha)}}_{=:g(\Theta_\alpha)} c_\alpha^{j-1}$$

$g(\Theta_\alpha)$ heißt Vergrößerungsfaktor. Er mißt den Anstieg oder Abstieg (Dämpfung) eines Fourier-Terms des Fehler bei einem Iterationsschritt.

Den Betrag von $g(\Theta_\alpha)$ simplifiziert MAPLE zu

$$|g(\Theta_\alpha)| = \frac{1}{\sqrt{5 - 4\cos\Theta_\alpha}}$$

Setzt man jetzt alle Werte Θ_α in diesen Ausdruck ein und bestimmt sein Maximum

$$\max\{|g(\Theta_\alpha)| \Big| \Theta_\alpha = \frac{\pi\alpha}{2n}, \alpha = 0, 1, \ldots, 2n-1\} = g(0) = 1,$$

dann sieht es so aus, als ob das ESV nicht konvergieren würde. Aber wegen der periodischen Randbedingung ist die Lösung ohnehin nur bis auf eine Konstante festgelegt, sodaß wir keinen Abfall des konstanten Teils $\Theta_\alpha = 0 (\alpha = 0)$ brauchen. Deshalb ist das korrekte Konvergenzmaß

$$\max\left\{ |g(\Theta_\alpha)| \;\middle|\; \Theta_\alpha = \frac{\pi\alpha}{2n}, \alpha = 1, 2, \ldots, 2n-1 \right\}$$

$$\begin{aligned} &= g(\Theta_1) = \frac{1}{\sqrt{5 - 4\cos\Theta_1}} \\ &= (5 - 4(1 - \Theta_1^2/2)^{-\frac{1}{2}} \\ &= (1 + 2\Theta_1^2 + O(\Theta_1^4))^{-\frac{1}{2}} = 1 - 2\pi^2 h^2 + O(h^4). \end{aligned}$$

$$\left[\text{ Wegen } \quad \Theta_1 = \frac{\pi}{2n} = \pi h \Rightarrow \Theta_1^2 = \pi^2 h^2. \right]$$

Das Mehrgitterprinzip

Wir sehen an $|g(\Theta_a)|$ wieder die Grundlage der MGM:
$|g(\Theta_\alpha)|$ wird kleiner mit wachsendem α.

Glättungsanalysis

Definition 5.5.4 Die Menge hochfrequenter Wellenzahlen sei

$$\Theta_r := \{\Theta_\alpha = \frac{\pi\alpha}{2n} \text{ mit } \alpha \geq c\,n\} \text{ mit } 0 < c \leq 1,$$

unabhängig von n, normalerweise $c = 1$, also

$$\Theta_r = \{\Theta_\alpha \text{ für } \alpha = n, n+1, \ldots, 2n-1\}.$$

(Das sind alle Frequenzen, die beim nächst größeren Gitter nicht vorkommen!)

Definition 5.5.5 Der Glättungsfaktor ρ ist definiert als

$$\rho = \max\{|g(\Theta_\alpha)| \,\big|\, \Theta_\alpha \in \Theta_r\}$$

Wenn $\rho < 1$ gleichmäßig in h (bzw. unabhängig von h), dann heißt die Methode smoother (Glätter!).
Für $c = 1$ haben wir beim Gauß-Seidel ESV

$$\rho = \left(5 - 4\cos\frac{\pi}{2}\right)^{-\frac{1}{2}} = 5^{-\frac{1}{2}}$$

Diese Fourier'sche Glättungsanalysis hat Achi Brandt schon 1977 so ausgeführt (hoffentlich korrekter als Wesseling!). Sie ändert sich nicht wesentlich bei anderen RBen, die periodischen haben nur das Rechnen erleichtert.

Zweigittermethode

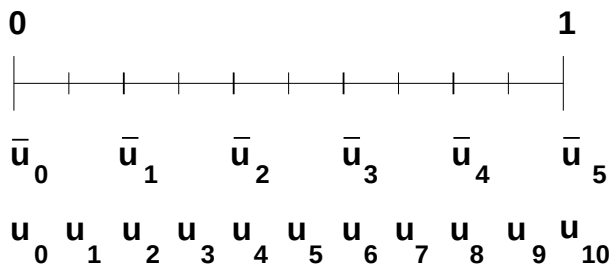
Das grobe Gitter wird durch verdoppelte Gitterweite definiert

$$\begin{aligned} \bar{G} &:= \{x \in \mathbb{R} \,\big|\, x = xj = j\bar{h}, j = 1, 2, \ldots, n\} \\ \bar{h} &= \frac{1}{n}. \end{aligned}$$

Das feine Gitter war gegeben durch

$$\begin{aligned} G &:= \{x \in \mathbb{R} \,\big|\, x = x_j = jh, j = 1, 2, \ldots, 2n\} \\ h &= \frac{1}{2n}. \end{aligned}$$

Funktionen $U : G \to \mathbb{R}$ und $\bar{U} : \bar{G} \to \mathbb{R}$ unterscheiden wir auch durch den Balken.



Den Übergang $\bar{U} \to U$ bezeichnen wir als Prolongation und benutzen lineare Interpolation:

$$\begin{aligned} P\bar{U} &= U \\ (P\bar{U})(x_j) &= \bar{u}_{\frac{j}{2}}, \text{ falls } x_j \in \bar{G}, j \text{ gerade} \\ (P\bar{U})(x_j) &= \frac{1}{2}(u_{\frac{j+1}{2}} + u_{\frac{j-1}{2}}), \text{ falls } x_j \notin \bar{G}, \text{ d.h. } j \text{ ungerade} \end{aligned}$$

Beim Übergang vom feinen aufs grobe Netz benutzen wir einen gewichteten Durchschnitt:

Restriktion

$$(RU)(x_j) = \frac{1}{4}u_{2j-1} + \frac{1}{2}u_{2j} + \frac{1}{4}u_{2j+1}$$

Dabei wird u_j Null gesetzt außerhalb G. Es gibt andere Möglichkeiten für Prolongation und Restriktion, hier gilt für die Matrizen P und R

$$R = \frac{1}{2}P^T$$

Auf dem groben Gitter gilt die Diskretisierung

$$\bar{A}\bar{u} = \bar{f}$$

Dieses Problem erhält man durch Diskretisierung der DGL auf dem groben Gitter. Es gibt andere Möglichkeiten!

Grobgitterkorrektur

Sei $\hat{u}$ eine Approximation der Lösung von $Au = f$ (also feines Gitter). Dann gilt für den Fehler $e = \hat{u} - u$:

$$A(-e) = r = f - A\hat{u}$$

Dieses Gleichungssystem lösen wir aber nur auf dem groben Gitter nach Restriktion von r:

$$\bar{A}\bar{u} = Rr$$

Die Korrektur $\bar{u}$ wird aufs feine Gitter interpoliert und zu $\hat{u}$ addiert:

$$\hat{\hat{u}} := \hat{u} + P\bar{u}$$

Zweigitteralgorithmus

Startwert u^0

Für

i	$=$	$1(1)nit$	(number of iterations)
$u^{\frac{1}{3}}$	$:=$	$S(u^0, A, f, \nu_1)$	(Vorglättung)
r	$:=$	$f - Au^{\frac{1}{3}}$	(Residuum)
$\bar{u}$	$:=$	$\bar{A}^{-1}Rr$	(Exakte Lösung der Defekt-Gleichung)
$u^{\frac{2}{3}}$	$:=$	$u^{\frac{1}{3}} + P\bar{u}$	
u^1	$:=$	$S(u^{\frac{2}{3}}, A, f, \nu_2)$	(Nachglättung)
u^0	$:=$	u^1	

Es werden also in diesem Algorithmus nit Zweigitterzyklen durchgerechnet mit
ν_1 Relaxationsschritten am Anfang und
ν_2 Relaxationsschritten am Ende jeden Zyklus.

Analysis des Zweigitterzyklus

Ziel dieser Analysis ist es, die Unabhängigkeit der Konvergenzrate von der Schrittweite h zu zeigen.

Wir analysieren den speziellen Fall $\nu_1 = 0$, d.h. keine Vorglättung.

Aus dem Algorithmus ersehen wir, daß der Fehler nach der Grobgitterkorrektur

$$e^{\frac{2}{3}} := u^{\frac{2}{3}} - u$$

$$e^{\frac{2}{3}} = Te^{\frac{1}{3}}$$

erfüllt mit der Iterations- oder Fehlerverstärkungsmatrix

$$T = I - P\bar{A}^{-1}RA$$

In unserem einfachen eindimensionalen Fall können wir $e^{\frac{2}{3}}$ explizit mit $e^{\frac{1}{3}}$ darstellen. Sei

$$e^{\frac{1}{3}} = d + P\bar{e} \text{ mit } \bar{e}_j := e^{\frac{1}{3}}_{2j},$$

d.h. wir definieren einen Vektor $\bar{e}$ auf $\bar{G}$, aus dem

$$P\bar{e} = (e^{\frac{1}{3}}_0, \frac{e^{\frac{1}{3}}_0 + e^{\frac{1}{3}}_2}{2}, e^{\frac{1}{3}}_2, \frac{e^{\frac{1}{3}}_2 + e^{\frac{1}{3}}_4}{2}, e^{\frac{1}{3}}_4, \ldots)$$

auf G wird. d ergänzt diesenVektor $P\bar{e}$ wieder zu $e^{\frac{1}{3}}$, wenn wir d definieren als

$$d_{2j} = 0, d_{2j+1} = -\frac{1}{2}e^{\frac{1}{3}}_{2j} + e^{\frac{1}{3}}_{2j+1} - \frac{1}{2}e^{\frac{1}{3}}_{2j+2}.$$

Diese Aufteilung führt gerade dazu, daß

$$Te^{\frac{1}{3}} = Td, \text{ d.h. } TP\bar{e} = 0$$

(was nicht leicht nachzurechnen ist, und nur in diesem speziellen Fall stimmt).

Daraus folgt, daß

$$e^{\frac{2}{3}} = Td$$

Wegen der speziellen Struktur von d ist aber auch $RAd = 0$ (auch eine Seite Rechnerei), also ist

$$e^{\frac{2}{3}} = d.$$

Jetzt betrachten wir den Effekt der Glättung am Ende des Algorithmus (post-smoothing) durch 1 ESV-Iteration ($\nu_1 = 0, \nu_2 = 1$).

Es ist $e^1 = u^1 - u$ und e_1 steht zu $e^{\frac{2}{3}}$ in folgender Beziehung, siehe (5.38):

$$\begin{aligned} 2e_1^1 &= e_2^{\frac{2}{3}} \\ -e_{j-1}^1 &+ 2e_j^1 = e_{j+1}^{\frac{2}{3}}, \quad j = 2, 3, \ldots, 2n-1 \\ -e_{2n-1}^1 &+ e_{2n}^1 = 0 \end{aligned}$$

Dies können wir mit den gerade hergeleiteten Beziehungen für $e^{\frac{1}{3}}, e^{\frac{2}{3}}, d$ umformen zu

$$\begin{aligned} e_1^1 &= 0 \\ e_{2j}^1 &= \frac{1}{2}d_{2j+1} + \frac{1}{4}e_{2j-2}^1, \quad j = 1, 2, \ldots, n-1, \\ e_{2j+1}^1 &= \frac{1}{2}e_{2j}^1 \\ e_{2n}^1 &= e_{2n-1}^1 \end{aligned}$$

Mit Induktion läßt sich jetzt leicht zeigen, daß

$$|e_{2j}^1| \leq \frac{2}{3}\|d\|_\infty$$

Damit haben wir eine Konvergenzrate unabhängig von h:

$$\|e^1\|_\infty \leq \frac{2}{3}\|e^0\|_\infty$$

Schlußfolgerung

Die von h unabhängige Konvergenzrate beruht i.w. auf der Gleichung

$$RAe^{\frac{2}{3}} = 0$$

R bewirkt einen gewichteten lokalen Durchschnitt mit positiven Gewichten, d.h. also, daß $Ae^{\frac{2}{3}}$ viele Vorzeichenwechsel hat, also hochfrequent ("rauh") ist. Da $e^{\frac{2}{3}}$ ein Residuums-Fehler ist, sehen wir, daß das Residuum nach einer Grobgitterkorrektur hochfrequent ist. Der Glätter kann deshalb diesen Fehler so effizient reduzieren, und das unabhängig von h.

5.5.4 Definition der Operatoren für das Modellproblem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0,1)\times(0,1) \\ u &= g \text{ auf } \partial\Omega \end{aligned} \tag{5.41}$$

Wir betrachten hier immer $H = 2h$. Die Operatoren L_h, I_h^{2h}, I_{2h}^h werden am einfachsten durch ihre "Sterne", d.h. ihre Wirkung auf einen inneren Punkt in Ω_h bzw. Ω_{2h} repräsentiert.

(i) Differenzenoperator:

$$L_h \hat{=} \frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_h \tag{5.42}$$

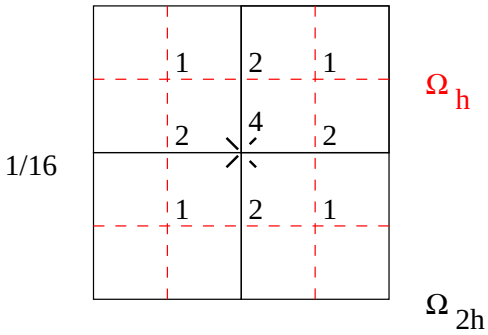
(ii) Restriktionsoperator:

$$I_h^{2h} : H(\Omega_h) \to H(\Omega_{2h})$$

z.B. im $\mathbb{R}^2$: $\Omega_{2h} = \{x = 2kh, k \in \mathbb{Z}^2, x, h \in \mathbb{R}^2\}$

Häufigstes Beispiel: Der full weighting FW-Operator:

$$I_h^{2h} \hat{=} \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}_h^{2h} \tag{5.43}$$



(iii) Interpolationsoperator:

$$I_{2h}^h : H(\Omega_{2h}) \to H(\Omega_h)$$

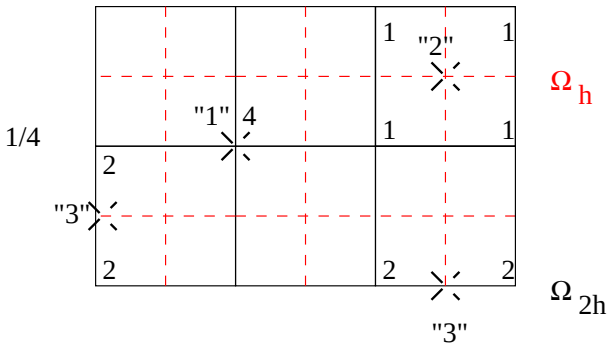
Der zum FW-Operator adjungierte Interpolationsoperator ist

$$I_{2h}^h \hat{=} \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}_{2h}^{h} \tag{5.44}$$

Her muß man drei Fälle unterscheiden:

- **Fall "1":**
  $x \in \Omega_h$ und $x \in \Omega_{2h}$, d.h. kein Nachbarpunkt von x im feinen Gitter liegt auf dem groben Gitter. Dann wird der Wert in x übernommen.

- **Fall "2":**
  $x \notin \Omega_{2h}$, 4 Nachbarn $\in \Omega_{2h}$.
  Dann bekommen diese den Faktor 1/4.
- **Fall "3":**
  $x \notin \Omega_{2h}$, 2 Nachbarn $\in \Omega_{2h}$.
  Dann bekommen diese den Faktor 1/2.



Bemerkung 5.5.6 Die Grobgitter-Korrektur CGC allein ohne Relaxation konvergiert nicht (CGC: Coarse Grid Correction). Man sieht leicht, daß

$$\rho(I_h - I_{2h}^h L_{2h}^{-1} I_h^{2h} L_h) \geq 1 \tag{5.45}$$

Beweis: Wähle $w_h \neq 0 \in H(\Omega_h)$ so, daß

$$L_h w_h \Big|_{\Omega_{2h}} = 0.$$

Dann ist

$$\underbrace{I_h w_h}_{w_h} - I_{2h}^h L_{2h}^{-1} \underbrace{I_h^{2h} L_h w_h}_{=0} = w_h$$

$\Rightarrow \rho(\cdot) \geq 1$. ■

Die Funktionen w_h, für die $L_h w_h = 0$ auf Ω_{2h} gilt, sind genau die stark oszillierenden Lösungs- bzw. Fehlerkomponenten, die nur auf dem feinen Gitter zu "sehen" sind, und deshalb nur mit der Relaxation auf diesem erhalten (Lösung) bzw. geglättet (Fehler) werden.

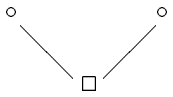
5.5.5 Vollständige Mehrgitterzyklen

Schematische Darstellung: Zeichenerklärung

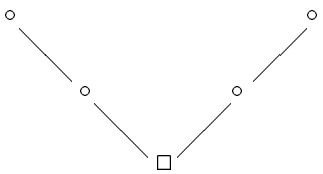
$\circ$	Relaxation
$\square$	Exakte Lösung der Defektgleichung
$\backslash$	Restriktion
$/$	Interpolation
γ	Wiederholungszahl der Lösungs/Relaxations-Stufen
$\gamma = 1$	V-Zyklus
$\gamma = 2$	W-Zyklus

Beispiele:

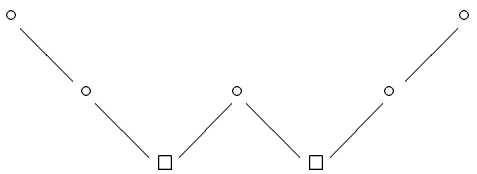
(1) Zweigitter-Methode (γ ohne Bedeutung)



(2) Dreigitter-Methode

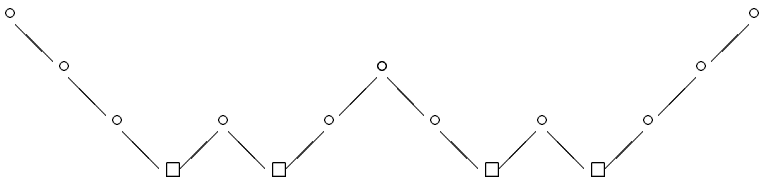


$\gamma = 1$



$\gamma = 2$

(3) Viergitter-Methode ($\gamma = 1$ trivial)



$\gamma = 2$

Weitere Zyklen wie Sägezahn- und F-Zyklus z.B. im Wesseling, [Wes92].

Die Gitter Ω_l mögen die Schrittweite h_l haben, wir ersetzen den Index h_l einfach durch l, also:

$$\begin{array}{lcll} L_l & : & H(\Omega_l) & \to & H(\Omega_l) \\ S_l & : & H(\Omega_l) & \to & H(\Omega_l) \\ I_l^{l-1} & : & H(\Omega_l) & \to & H(\Omega_{l-1}) \\ I_{l-1}^{l} & : & H(\Omega_{l-1}) & \to & H(\Omega_l) \end{array} \tag{5.46}$$

Wir setzen (natürlich) voraus, daß L_l^{-1} existiert $\forall l$.

Die Methode wird jetzt am besten rekursiv definiert wie in den beiliegenden Unterprogrammen aus [Wes92]. I.a. kann γ noch von der Stufe k abhängen. Wir wollen den Iterationsoperator $M_l(\gamma)$, l = Gitterzahl-1, darstellen und auf die von h unabhängige Konvergenz und den Aufwand eingehen.

Lemma 5.5.7 Löst man das Gleichungssystem $Aw = r$ iterativ mit dem Verfahren

$$w^{(j+1)} = Mw^{(j)} + s \tag{5.47}$$

mit dem Startwert $w^{(0)} = 0$ und $M = I - B^{-1}A, s = B^{-1}r$, dann gilt

$$w^{(j)} = (I - M^j)A^{-1}r \tag{5.48}$$

Beweis: Induktiv:
Induktionsanfang $(j = 0)$:
$w^{(0)} = (I - M^0)A^{-1}r = 0$

Induktionsvoraussetzung $(j = k)$:
$w^{(k)} = (I - M^k)A^{-1}r$ ist richtig!

Induktionsbeweis $(j = k + 1)$:

$$\begin{aligned} w^{(k+1)} &= Mw^{(k)} + s \\ &= M(I - M^k)A^{-1}r + s \quad (\text{ nach Induktionsvoraussetzung}) \\ &= (I - B^{-1}A - M^{k+1})A^{-1}r + s \quad (\text{ weil } M = I - B^{-1}A) \\ &= (I - M^{k+1})A^{-1}r \underbrace{-B^{-1}AA^{-1}r + s}_{=0} \blacksquare \end{aligned}$$

Beispiel: $l = 2, \gamma = 2$

Aus (5.35) folgt durch entsprechende Umnumerierung und Aufschneiden bei L_{2h}^{-1} offenbar

$$M_2 = S_2^{\nu_2}\,[I_2 - I_1^2\{\underbrace{S_1^{\nu_2}(I_1 - I_0^1 L_0^{-1} I_1^0 L_1)S_1^{\nu_1}}_{=:M_1}\}^2 I_2^1 L_2]\,S_2^{\nu_1} \tag{5.49}$$

Nach Lemma 5.5.7 (mit $M = M_1$ Iterationsoperator für das Zweigitterverfahren, $w^{(0)} = 0$ wegen Defektgleichung) kann man jetzt M_1^2 durch $(I_1 - M_1^2)L_1^{-1}$ ersetzen, also

$$M_2 = S_2^{\nu_2}\,[I_2 - I_1^2(I_1 - M_1^2)L_1^{-1} I_2^1 L_2]\,S_2^{\nu_1} \tag{5.50}$$

obwohl L_1^{-1} im Algorithmus nicht vorkommt. Rekursiv ausgenutzt bekommt man

Lemma 5.5.8 (i) Der Iterationsoperator $M_l := M_l(\gamma)$ ist durch folgende Rekursion gegeben

$$\begin{aligned} M_1 &= S_1^{\nu_2}(I_1 - I_0^1 L_0^{-1} I_1^0 L_1)S_1^{\nu_1} \\ M_{k+1} &= S_{k+1}^{\nu_2}\,[I_{k+1} - I_k^{k+1}(I_k - M_k^\gamma)L_k^{-1} I_{k+1}^k L_{k+1}]\,S_{k+1}^{\nu_1} \\ &\quad k = 1, \ldots, l-1 \end{aligned} \tag{5.51}$$

(ii) Seien $A_k^{k+1} := S_{k+1}^{\nu_2} I_k^{k+1}$ und $A_{k+1}^k := L_k^{-1} I_{k+1}^k L_{k+1} S_{k+1}^{\nu_1}$
Dann gilt für $k = 1, \ldots, l-1$

$$M_{k+1} = M_{k+1}^k + A_k^{k+1} M_k^\gamma A_{k+1}^k \tag{5.52}$$

wobei M_{k+1}^k der Zweigitteroperator zu h_{k+1}, h_k ist, also

$$M_{k+1}^k := S_{k+1}^{\nu_2}(I_{k+1} - I_k^{k+1} L_k^{-1} I_{k+1}^k L_{k+1})S_{k+1}^{\nu_1} \tag{5.53}$$

Bemerkung 5.5.9 Hier wird der Mehrgitteroperator als Störung der entsprechenden Zweigitteroperatoren aufgefaßt. Das führt zum leichteren Beweis der $h-$unabhängigen Konvergenz:

Satz 5.5.10 Sei für $k \leq l-1$:

$$\|M_{k+1}^k\| \leq \sigma^*, \quad \|A_k^{k+1}\| \, \|A_{k+1}^k\| \leq C \tag{5.54}$$

Dann ist $\|M_l\| \leq \eta_l$ mit rekursiv definiertem η_l:

$$\eta_1 := \sigma^*, \eta_{k+1} := \sigma^* + C\eta_k^\gamma \tag{5.55}$$

Beweis: folgt sofort aus (5.52). Aus diesem Satz folgt die $h-$unabhängige Konvergenz, wenn σ^* klein genug ist, d.h. die Zweigitter-Methode genügend gut konvergiert, ein Beispiel liefert das folgende Lemma. ■

Lemma 5.5.11 Sei $\gamma = 2$ und $4\,C\,\sigma^* \leq 1$.

Dann gilt

$$\|M_l\| \leq \eta := \frac{(1 - \sqrt{1 - 4C\sigma^*})}{2C} \leq 2\sigma^* \tag{5.56}$$

Bemerkung 5.5.12 Modellbeispiel $\sigma^* \leq 0.1, C = 1 \Rightarrow \eta \leq 0.113$.

Aufwand

Man kann jetzt auch die asymptotische Optimalität zeigen.

Für das Modellproblem (5.41) gilt für den Zweigitteroperator M_k^{k-1} und für ein Gitter mit N_k inneren Punkten

Aufwand $(M_k^{k-1}) = C\,N_k$ mit $C = 6\nu + 10$

$(\nu = \nu_1 + \nu_2)$.

Die für diesen geringen Aufwand entstehenden Konvergenzraten sind auch für allgemeinere Gebiete sehenswert, wie man an der beiliegenden Kopie sieht.

5.5.6 Mehrgittermethoden für FEM

- **1. Möglichkeit:**
  Fester Ansatz, feiner werdende Gitterfolge: Im Prinzip die gleiche Vorgehensweise wie bei den Differenzenverfahren, aber Restriktion und Interpolation müssen geeignet definiert werden.

- **2. Möglichkeit:**
  Verschiedene Ansätze mit identischen Gitterpunkten!

  Nur beschränkt anwendbar, ein Beispiel wäre die Kombination linear/quadratisch auf einem einmal regulär verfeinerten Gitter. Man erhält die gleiche Matrix-Größe und es sind Interpolation= Restriktion = Identität!

Kapitel 6

Gittererzeugung

Bevor eine PDE mit einer FEM oder einem Differenzenverfahren gelöst werden kann, muß auf dem Grundgebiet Ω ein Gitter vorhanden sein, das für die FEM die geometrischen Elemente, für ein Differenzenverfahren die inneren Punkte und deren Nachbarn definiert.

Für die Erzeugung eines sochen Gitters kommen mehrere Basismethoden in Frage:

(1) Erzeugung von Hand, d.h. Eingabe einer Koordinatenliste.

(2) Erzeugung einer groben Gitterstruktur per Hand, Verfeinerung automatisch uniform oder adaptiv mit entsprechenden Verfeinerungsalgorithmen ($\rightarrow$ mathematisches Grundpraktikum).

(3) Erzeugung eines **strukturierten** Gitters automatisch mit einem entsprechenden 'mathematischen' Algorithmus. Auch hier sind mehrere uniforme oder adaptive Verfeinerungsstufen möglich.

(4) Erzeugung eines **unstrukturierten** Gitters automatisch mit einem entsprechenden 'Informatik'-Algorithmus. Auch hier sind mehrere uniforme oder adaptive Verfeinerungsstufen möglich.

In allen Softwarepaketen zur Lösung partieller Differentialgleichungen sind heute Methoden zur Gittererzeugung und/oder Gitterverfeinerung enthalten.

Wir wollen zunächst graphisch orientiert einen Eindruck von den Möglichkeiten und Zielen der Gittererzeugung bekommen. Dazu habe ich Beispiele aus verschiedenen Vorträgen und Diplomarbeiten zusammengestellt.

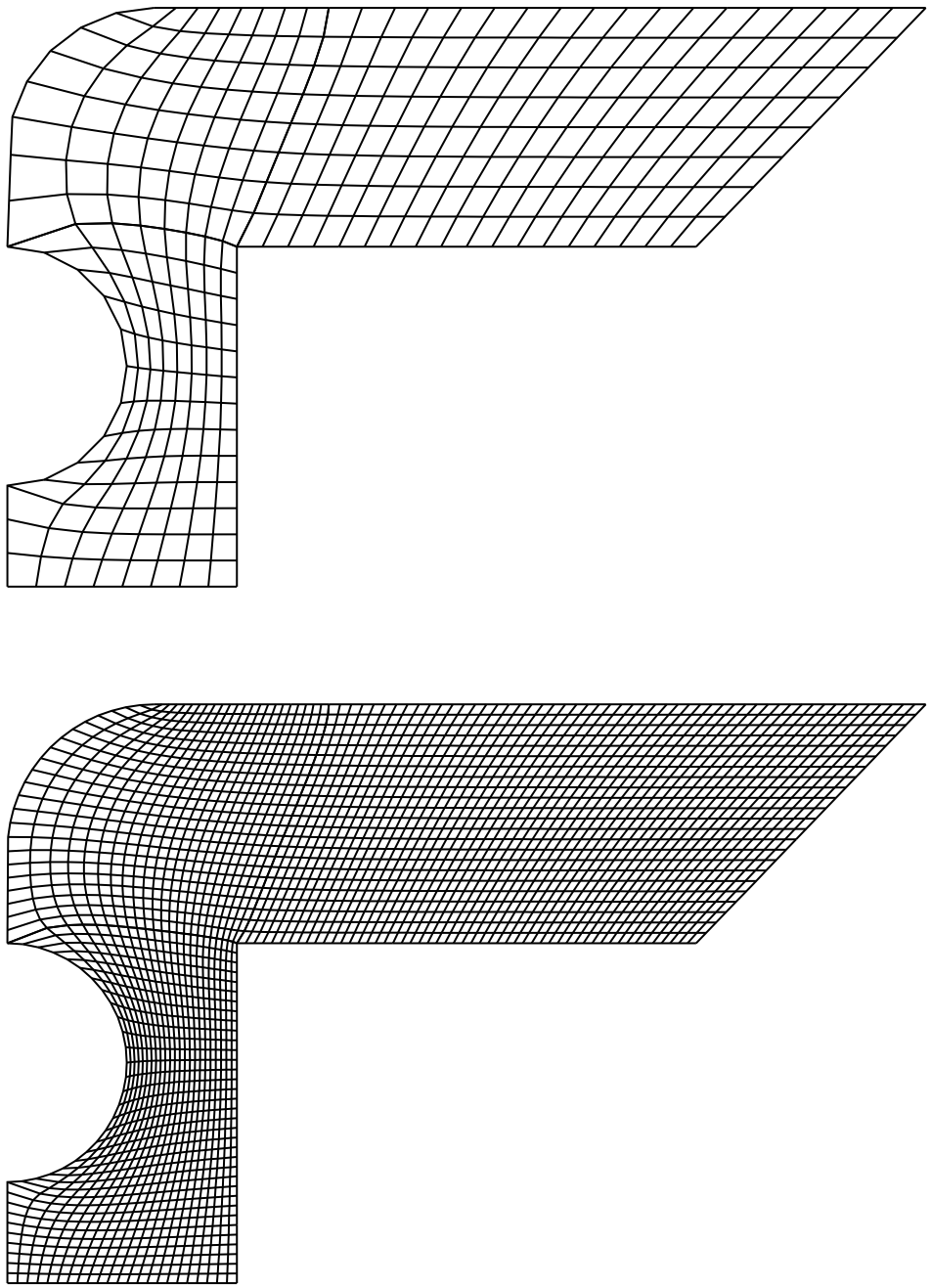
Anschließend wollen wir uns mit zwei mathematischen Methoden etwas näher beschäftigen, aber auch hauptsächlich am Beispiel orientiert.

Wenn danach noch Zeit bleibt, werde ich kurz auf die Parallelisierbarkeit der Algorithmen eingehen.

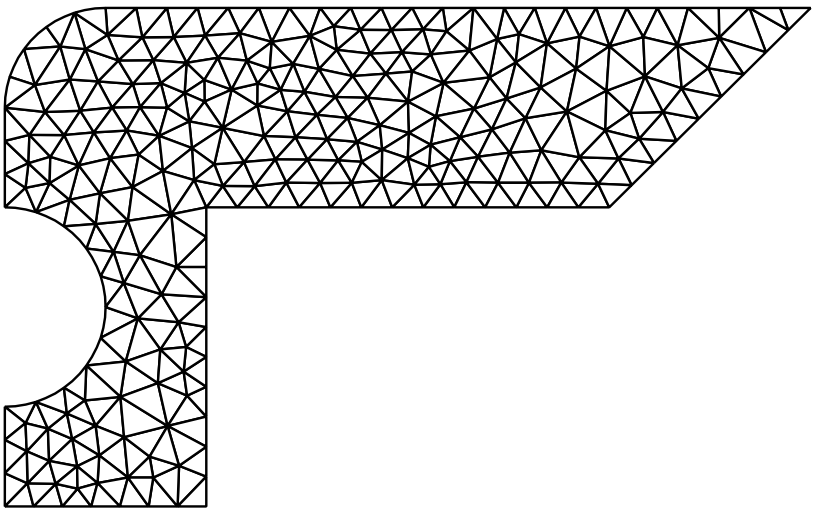
6.1 Möglichkeiten der Netzerzeugung

6.2 Beispielgitter

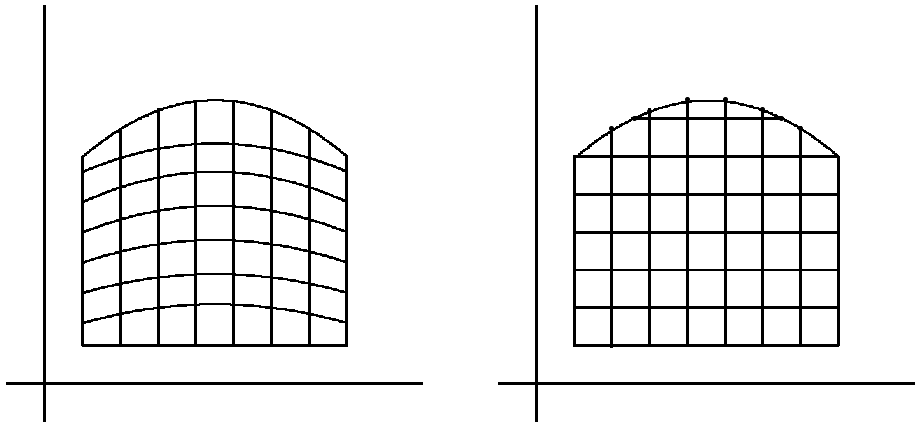
Zwei Transformationsgitter



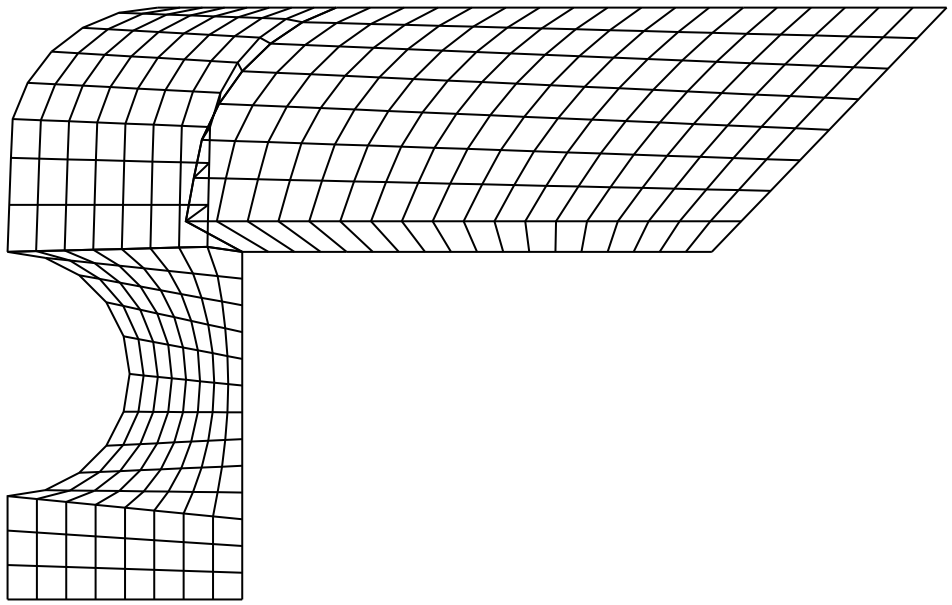
Ein Dreiecksnetz von PLTMG



Randkonform vs. Nicht-randkonform



Transfinite Interpolation: Gitter mit Falten



6.3 Strukturierte Gittererzeugung

Klassifizierung

- **r-Typ: Feste Anzahl Gitterpunkte wird 'optimal' verteilt.**
- **h-Typ: Adaptive Verfeinerung einer einfachen Anfangsverteilung.**
- **p-Typ: Verfeinerung der Elemente eines FEM-Ansatzes.**
  Erfüllung von Längen-, Volumen- und Winkelbedingungen.

Wünschenswerte Eigenschaften

- **Randkonform**
- **Automatisch**
- **Robust**
- **Sinnvoll**

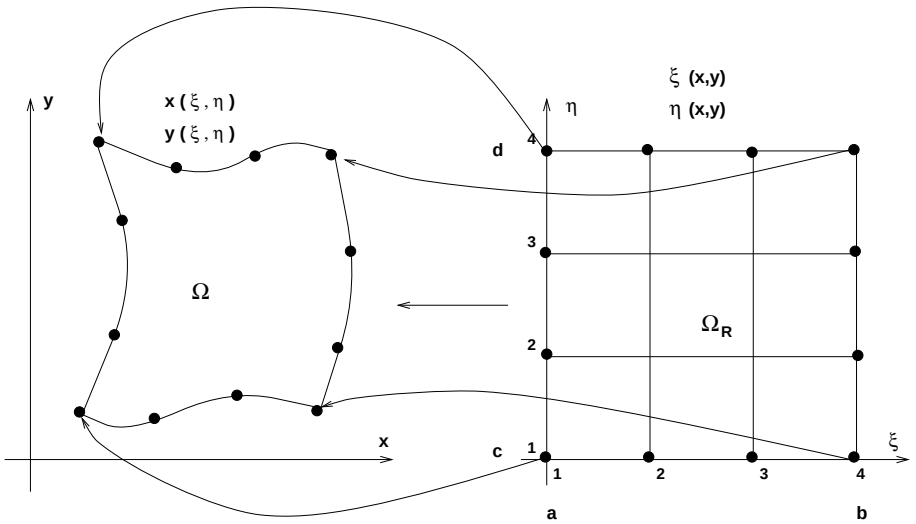
Methoden

- **Elementare 'Hand'-Methoden (der Computer hilft ein wenig)**
- **Algebraische Methoden, z.B. transfinite Interpolation**
- **PDE-Methoden (harmonische Abbildungen und Weiterentwicklungen davon)**
- **Variationsmethoden (teilweise ohne zugehörige PDE, dafür mit Einsatz geometrischer Formulierungen: Längen, Winkel, Volumen)**

Literatur

- **Winslow, 1967**
- **Amsden, Hirt, 1973**
- **Thompson, Thames, Mastin, 1974.**
- **Thompson, J.F.; Warsi, Z.U.A.; Mastin, C.W., 1985**
  Numerical Grid Generation: Foundations and Applications, [TWM85].
- **Castillo, J.E. (ed.), 1991**
  Mathematical Aspects of Numerical Grid Generation, [Cas91]
- **Parallelisierung: Aden/Köckler, 1995.**

6.3.1 Transformationsmethoden (r-Typ)



(x, y) : Kartesische Koordinaten in Ω (Physikalisches Gebiet):

$$(x(\xi,\eta), y(\xi,\eta))\,, \quad \mathfrak{x} = \begin{pmatrix} x \\ y \end{pmatrix}$$

(ξ, η) : Kartesische Koordinaten in Ω_R (Rechteck $[a, b] \times [c, d]$)

Die krummlinigen Koordinaten sind dann die Kurven-Familien

$$\begin{aligned} (x(s,\eta), y(s,\eta)), &\qquad a \le s \le b \\ (x(\xi,t), y(\xi,t)), &\qquad c \le t \le d \end{aligned}$$

Randerhaltungseigenschaften

$$\begin{aligned} (x(\xi,c), y(\xi,c)), &\qquad a \le \xi \le b \\ (x(\xi,d), y(\xi,d)), &\qquad a \le \xi \le b \\ (x(a,\eta), y(a,\eta)), &\qquad c \le \eta \le d \\ (x(b,\eta), y(b,\eta)), &\qquad c \le \eta \le d \end{aligned} \tag{6.1}$$

Jacobi-Matrix

$$M = \begin{pmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{pmatrix} \tag{6.2}$$

Jacobi-Determinante

$$J = x_\xi y_\eta - x_\eta y_\xi \,. \tag{6.3}$$

Ist $\tilde{J}$ die Jacobi-Determinante der inversen Transformation, dann ergibt die Kettenregel die folgenden Beziehungen:

$$J\tilde{J} = 1 \tag{6.4}$$

$$x_\xi = \eta_y/J, \;\; x_\eta = \xi_y/J, \;\; y_\xi = \eta_x/J, \;\; y_\eta = \xi_x/J \tag{6.5}$$

Transformationskoeffizienten und Differentialelemente

Kovariante Basisvektoren a_i sind die Tangenten an die krummlinigen Koordinatenlinien ξ bzw. η, also die Spalten der Funktionalmatrix

$$DX = \left(\frac{\partial(x,y)}{\partial(\xi,\eta)} \right), \;\; \text{also im } \mathbb{R}^2 \tag{6.6}$$

$$a_1 = \begin{pmatrix} x_\xi \\ y_\xi \end{pmatrix}, \quad a_2 = \begin{pmatrix} x_\eta \\ y_\eta \end{pmatrix} \tag{6.7}$$

Kontravariante Basisvektoren a^i: entsprechend.

Kontravariante Basisvektoren a^i sind die Normalen zu den Koordinatenlinien, auf denen ξ bzw. η konstant sind.

Transformationskoeffizienten sind die Elemente der ko- und kontravarianten Maßtensoren

$$g_{ij} := a_i \cdot a_j \quad \text{oder} \quad g := (DX)^T (DX), \quad \text{und damit } \; J = \sqrt{\det\, g} \tag{6.8}$$

g^{ij}, $\tilde{J}$ entsprechend.

Differentialelemente

- Differentialinkrement:

$$d\mathfrak{x} = d\xi a_1 + d\eta a_2 \tag{6.9}$$

- Inkrement der Bogenlänge s:

$$(ds)^2 = |d\mathfrak{x}|^2 = g_{11} d\xi d\xi + 2 g_{12} d\xi d\eta + g_{22} d\eta d\eta \tag{6.10}$$

- Inkrement der Bogenlänge auf der ξ-Koordinatenlinie

$$ds_\xi = \sqrt{g_{11}} d\xi, \tag{6.11}$$

  ds_η entsprechend.

- Inkrement einer Zellfläche

$$dS_\xi = \sqrt{g} d\xi d\eta, \tag{6.12}$$

Verfahrensanteile und -möglichkeiten

- Algebraische Methoden (Interpolation) $\rightarrow$ Startgitter
- Harmonische Transformation $\rightarrow$ Iterative Verbesserung
- Kontrollfunktionen $\rightarrow$ Strukturverbesserung
- Blockeinteilung $\rightarrow$ Strukturverbesserung und Parallelisierung

Erweiterungen und Spezialfälle

- Lokale Koordinatensysteme
- Gittererzeugung bei mehrfach zusammenhängenden Gebieten
- Adaptive Gittergenerierung
- Parabolische oder hyperbolische Systeme

Gittererzeugung mit Differentialgleichungen

$$\begin{aligned} \xi_{xx} + \xi_{yy} &= 0 && \text{in } \Omega, \\ \eta_{xx} + \eta_{yy} &= 0 && \\ \xi &= G && \text{auf } \partial\Omega. \\ \eta &= H && \end{aligned}$$

Eigenschaften dieser harmonischen Abbildung:

(1) Alle Punkte von Ω werden in Ω_R abgebildet.

(2) Die Gradienten $\nabla\xi$ und $\nabla\eta$ verschwinden nirgends in Ω.

(3) Die Jacobi-Determinante der Abbildung

$$J = \xi_x \eta_y - \xi_y \eta_x$$

verschwindet nirgends in Ω.

Wir benötigen aber die inverse Abbildung: $\Omega_R \rightarrow \Omega$. Sie führt zu dem nichtlinearen System

$$\begin{aligned} g_{22}\, \mathfrak{x}_{\xi\xi} + g_{11}\, \mathfrak{x}_{\eta\eta} - 2 g_{12}\, \mathfrak{x}_{\xi\eta} = 0 \qquad & \text{in} \quad \Omega_R \\ \mathfrak{x} = \begin{pmatrix} \tilde{G} \\ \tilde{H} \end{pmatrix} \qquad & \text{auf} \quad \partial\Omega_R \end{aligned}$$

g_{ij} sind die Transformationskoeffizienten, die im diskretisierten Fall für jeden Gitterpunkt berechnet werden:

$$g_{11} = x_\xi^2 + y_\xi^2, \quad g_{22} = x_\eta^2 + y_\eta^2, \quad g_{12} = x_\xi x_\eta + y_\xi y_\eta$$

6.3.2 Parallelisierung durch Gebietszerlegung

Ziele

- $$\frac{\#\text{ Gitterpunkte}}{\#\text{ Blockschnittpunkte}} \quad \text{groß}$$
- # Gitterpunkte/Block $\approx$ konstant
- Rechtecke $\approx$ Quadrate
- Keine Konvergenz-störende Blockaufteilung wie innere Blöcke
- Startblöcke günstig für Automatik-Blöcke

Gebietszerlegungs – Algorithmus

(1) **Anfangszerlegung** $\rightarrow$ Benutzer
Festlegung, ob innere Schnittlinien fest oder frei
Festlegung der Gitterpunktzahl, der Prozessorenzahl und der Architektur

Automatisch:

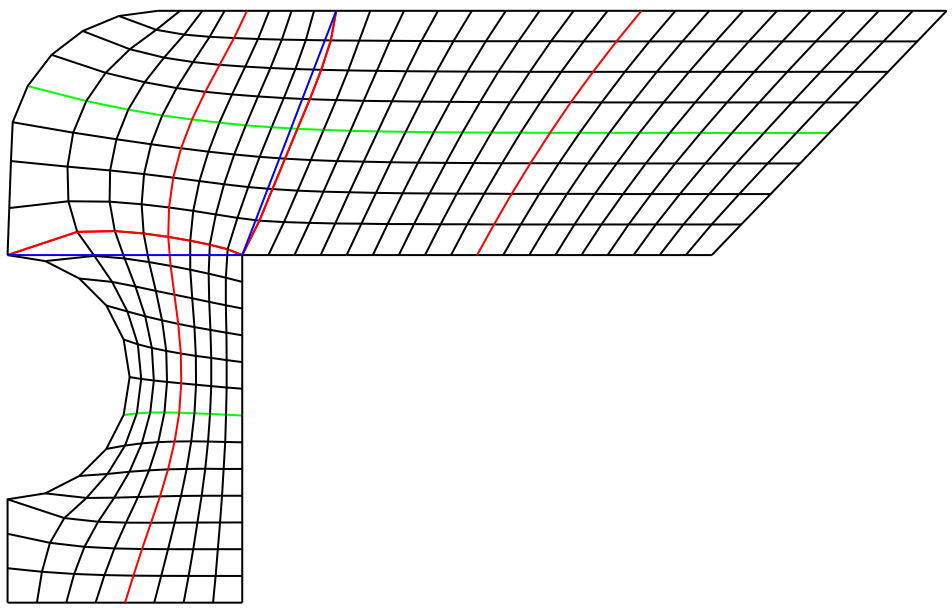
(2) Prüfung, ob für Zerlegung genügend Prozessoren vorhanden sind.

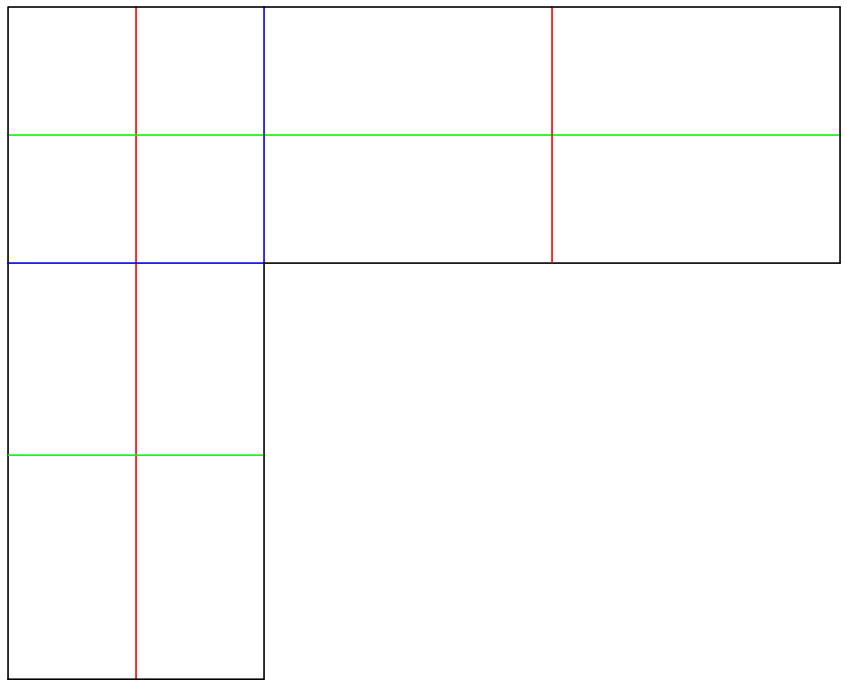
(3) Prüfung, ob vertikale oder horizontale Zerlegung günstiger.

(4) Prüfung, ob Zerlegung sinnvoll (bez. Lastausgleich, Speed-up).

(5) Zurück zu Punkt 2.

(6) Zerlegung durchführen, Variablen umsetzen, Prozessoren mit Daten beschicken.





6.3.3 Anwendungen und Ergebnisse

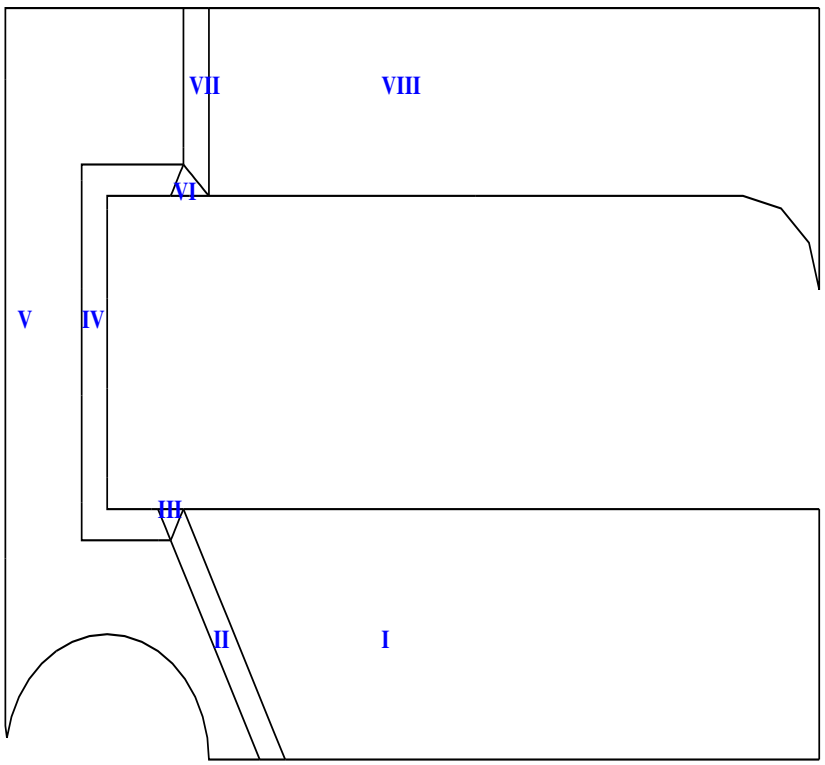
Stationäre Wärmeleitung mit unstetigen Koeffizienten

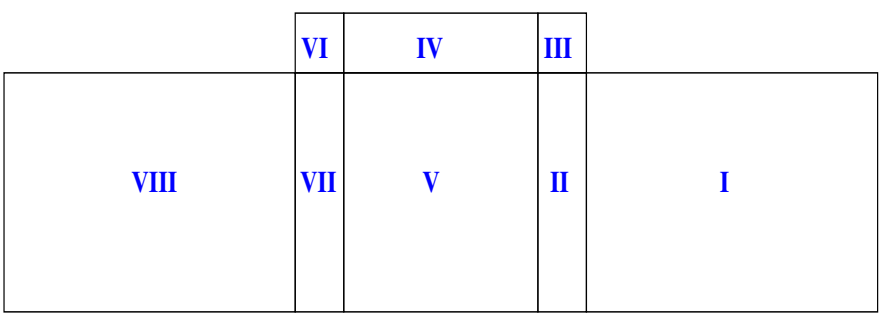
$$\begin{aligned} \frac{\partial}{\partial x}\left(K_1 \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_2 \frac{\partial u}{\partial y}\right) &= 0 \quad \text{in } \Omega \\ u &= g \quad \text{auf } \partial\Omega \end{aligned}$$

mit

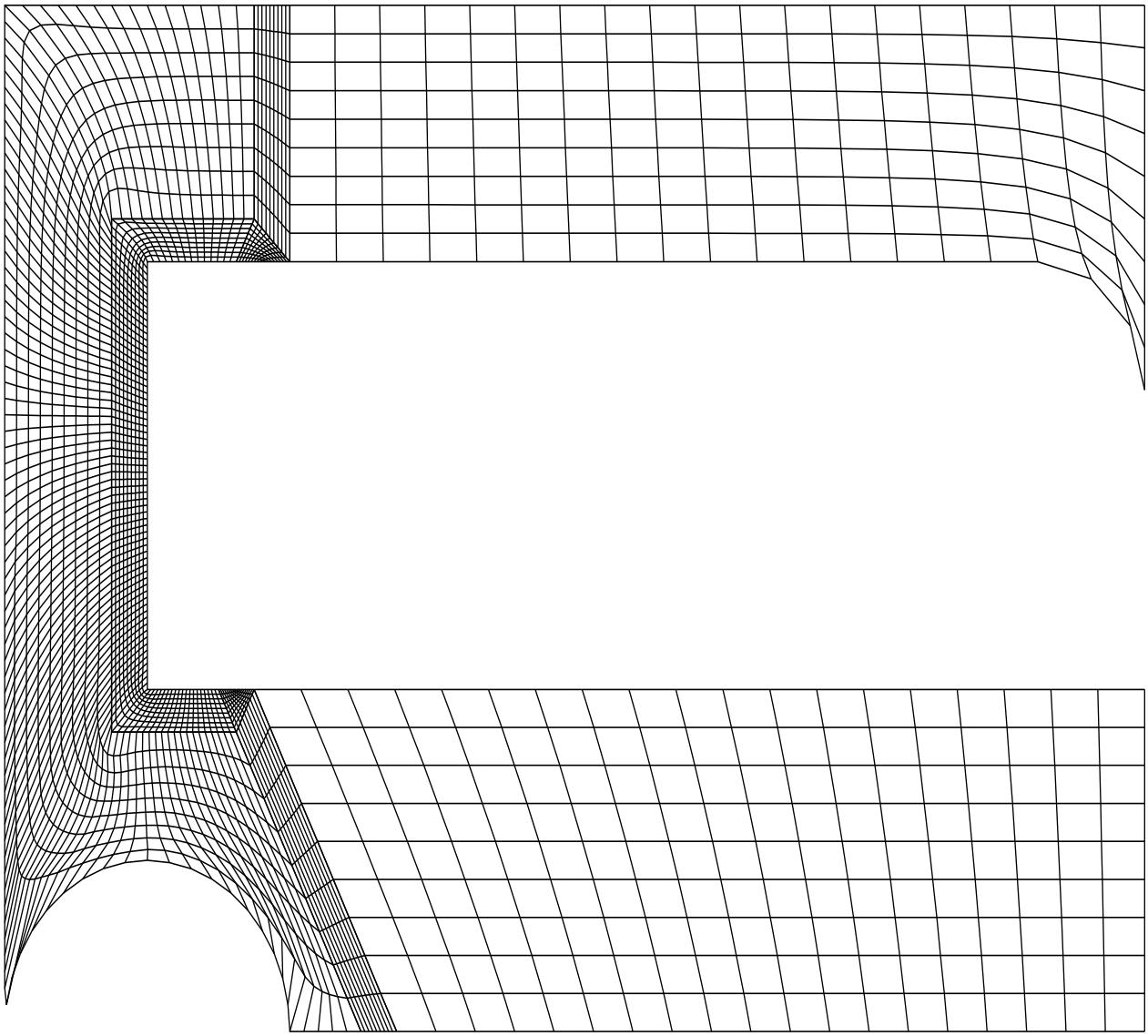
$K_1 =$ Wärmeleitzahl in x-Richtung
$K_2 =$ Wärmeleitzahl in y-Richtung

Hochofen mit Schornsteinanschluß – Blockaufteilung





Hochofen mit Schornsteinanschluß – Gitter



6.4 Unstruktierte Gittererzeugung

6.4.1 Ein Quadtree-Beispiel

Beispiel 6.4.1 Ein einfaches Beispiel

(1) Festlegung der Punkteanzahlen IMAX und JMAX

(2) Konfiguration des physikalischen Gebietes:
Zuordnung der Randpunkte

(3) Startnäherung durch transfinite Interpolation

(4) Iterativ:

(a) Berechnung der notwendigen Transformationskoeffizienten.

(b) Lösung der Trnsformations-PDEs

(1) **Festlegung der gewünschten Punktzahl**

Für das zweidimensionale Gebiet mit 4×4 Punkten wird IMAX $= 4$ und JMAX $= 4$ eingegeben. Die Numerierung von ξ und η läuft von $1 \ldots$ IMAX bzw. $1 \ldots$ JMAX.

(2) **Aufstellung einer Konfiguration für das transformierte Gebiet Ω**

Zunächst erfolgt die Eingabe der Randpunkte, in diesem Beispiel als Funktionen:

unterer/oberer Rand:

$$\begin{pmatrix} x(\xi,1) \\ y(\xi,1) \end{pmatrix} = \begin{pmatrix} K_1 \\ 0 \end{pmatrix} \text{ und } \begin{pmatrix} x(\xi,\ \text{JMAX}) \\ y(\xi,\ \text{JMAX}) \end{pmatrix} = \begin{pmatrix} K_1 \\ \frac{1}{2} + \frac{1}{2}K_1 \end{pmatrix} \text{ mit } K_1 = \frac{\xi - 1}{\text{IMAX } - 1}$$

linker/rechter Rand:

$$\begin{pmatrix} x(1,\eta) \\ y(1,\eta) \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2}K_2 \end{pmatrix} \text{ und } \begin{pmatrix} x(\ \text{IMAX},\eta) \\ y(\ \text{IMAX},\eta) \end{pmatrix} = \begin{pmatrix} 1 + 0,4K_2(1 - K_2) \\ K_2 \end{pmatrix} \text{ mit } K_2 = \frac{\eta - 1}{\text{JMAX} - 1}$$

In diesem Fall ergibt sich also $K_1 = \frac{1}{3}(\xi - 1), K_2 = \frac{1}{3}(\eta - 1)$.

Somit werden folgende Randpunkte vom Programm berechnet, dargestellt mit einer Vor- und fünf Nachkommastellen wie vom Generierungsprogramm ausgegeben:

unterer Rand:

(ξ,η)	$(1,1)$	$(2,1)$	$(3,1)$	$(4,1)$	
x	$0,0$	$3.33333E-01$	$6.66667E-01$	1.0	$E+00$
y	0.0	0.0	0.0	0.0	

oberer Rand:

(ξ,η)	$(1,4)$		$(2,4)$	$(3,4)$	$(4,4)$	
x	0.0		$3.33333E-01$	$6.66667E-01$	1.0	$E+00$
y	5.0	$E-01$	$6.66667E-01$	$8.33333E-01$	1.0	$E+00$

linker/rechter Rand:

(ξ,η)	$(1,2)$	$(1,3)$	$(4,2)$	$(4,3)$
x	0.0	0.0	$1.08889E-00$	$1.08889E-00$
y	$1.66667E-01$	$3.33333E-01$	$3.33333E-01$	$6.66667E-01$

Die Transformationskoeffizienten

	$(\xi=2,\eta=2)$	$(\xi=2,\eta=3)$	$(\xi=3,\eta=2)$	$(\xi=3,\eta=3)$
g_{11}	0.121	0.131	0.145	0.154
g_{22}	0.047	0.052	0.074	0.081
g_{12}	0.012	0.023	0.021	0.029

(3) **Generierung der Gitterpunkte**

Das Ziel ist es, mit den im Abschnitt "Generierungssysteme" beschriebenen Verfahren zu jedem Gitterpunkt $(\xi,\eta), \xi = 2,\ldots,$ IMAX $-1, \eta = 2,\ldots,$ JMAX-1 im Rechenraum Ω_R die Koordinaten des ihm entsprechenden Punkts im physikalischen Gebiet zu bestimmen. Die in Punkt 2 berechneten Randpunkte werden dabei (in diesem Beispiel) nicht mehr verändert.

Die Generierung soll mit einer elliptischen PDE ohne Kontrollfunktionen (Laplace-System) erfolgen, um eine gleichmäßige Verteilung der Gitterpunkte ohne Sonderbedingungen zu erreichen. Das Generierungssystem

$$\begin{aligned} \xi_{xx} + \xi_{yy} &= 0 \\ \eta_{xx} + \eta_{yy} &= 0 \end{aligned}$$

wird zunächst transformiert, um die Generierung im Rechenraum auf rechteckigem Gebiet Ω_R mit quadratischem Gitter und Einheitsschrittweite ausführen zu können:

$$g_{22}\Gamma_{\xi\xi} + g_{11}\gamma_{\eta\eta} - 2g_{12}\Gamma_{\xi\eta} = 0 \text{ mit } \Gamma = \begin{pmatrix} x \\ y \end{pmatrix} \tag{6.13}$$

(vgl. hierzu die Abschnitte 2.3 "Transformation" sowie 2.5 und 2.6 "Generierungssysteme").

Dann erfolgt die Berechnung der Startwerte für die inneren Punkte mit transfiniter linearer Lagrange-Interpolation, das heißt, einer vertikalen Interpolation schließt sich eine horizontale Korrektur an.

	Spalte $\xi = 2$		Spalte $\xi = 3$	
(ξ, η)	$(2,2)$	$(2,3)$	$(3,2)$	$(3,3)$
x	$3.62963E-01$	$3.62963E-01$	$7.25926E-01$	$7.25926E-01$
y	$2.22222E-01$	$4.44444E-01$	$2.77778E-01$	$5.55556E-01$

Interpolation

	Spalte $\xi = 2$		Spalte $\xi = 3$	
(ξ, η)	$(2,2)$	$(2,3)$	$(3,2)$	$(3,3)$
x	$3.38236E-01$	$6.89219E-01$	$3.39523E-01$	$6.92987E-01$
y	$2.12355E-01$	$2.66214E-01$	$4.31785E-01$	$5.43009E-01$

DGL

(4) **Abschließende Berechnung der Transformationskoeffizienten**

Die Transformatinskoeffizienten an jedem inneren Punkt können nun (abermals mit Differenzenausdrücken) aus den je 16 $x-$ und $y-$Werten endgültig bestimmt werden:

	Spalte $\xi = 2$		Spalte $\xi = 3$	
(ξ, η)	$(2,2)$	$(2,3)$	$(3,2)$	$(3,3)$
g_{11}	$1.21231E-01$	$1.31047E-01$	$1.44527E-01$	$1.54181E-01$
g_{22}	$4.66191E-02$	$5.16054E-02$	$7.38875E-02$	$8.05329E-02$
g_{12}	$1.1812\ E-02$	$2.29642E-02$	$2.13615E-02$	$2.90776E-02$

Hierbei handelt es sich um sämtliche in Frage kommenden ko- und kontravarianten Koeffizienten, von denen in der Regel nur einige wenige gebraucht werden. Sofern genügend Speicherplatz zur Verfügung steht, müssen die benönigten nur einmal berechnet werden.

Netzgenerierung

Definition 6.4.2 Definition 1 entfernt

Sei $x, y \in \delta P$

y liegt entfernt von x falls gilt:

x, y liegen auf verschiedenen Zusammenhangskomponenten von δP oder jeder Weg von x nach y auf δP enthält mindestens zwei Knoten.

Definition 6.4.3 Definition 2 nächster entfernter Nachbar

Sei $x, y \in \delta P$

y heißt nächster entfernter Nachbar von x falls gilt: y ist der nächste Punkt der entfernt von x liegt. (geodätische Distanz)

Definition 6.4.4 Definition 3 überfüllt

Eine Quadtreezelle b mit der Seitenlänge l heißt überfüllt falls eine der folgenden Bedingungen erfüllt ist:

(1) b enthält einen Punkt $x \in \delta P$ und der nächste entfernte Nachbar liegt näher als $2\sqrt{2l}$ am Punkt x

(2) b enthält einen Knotenpunkt $x \in X$ und einer der erweiterten Nachbarn von b wurde geteilt.

Definition 6.4.5 Definition 4 $q-$Knoten

Der Schnittpunkt des Polygons mit dem Rand einer Zelle heißt $q-$Knoten.

Definition 6.4.6 Definition 5 balanciert

Der Quadtree heißt balanciert falls gilt:

(1) Jede zu einer Quadtreezelle b benachbarte Zelle hat entweder die halbe, die gleiche oder die doppelte Seitenlänge wie b.

(2) Die Seitenlänge von zwei zu einer Quadtreezelle b benachbarten Zellen ist gleich oder sie unterscheidet sich nur um den Faktor 2.

Bedingung 1

Abschätzungsbedingung

Jede (Blatt) Zelle, die einen Knoten des Polygons enthält, muß von 24 Zellen gleicher Größe umgeben sein.

Der Algorithmus

(1) Berechne eine das Polygon umgebende Zelle.

(2) Solange eine überfüllte Zelle existiert, teile diese in vier gleichgroße Zellen auf.

(3) Blanciere den Quadtree.

(4) Verschiebe Quadtreeknoten auf die entsprechenden Polygonknoten.

(5) Verschiebe Quadtreeknoten auf die entsprechenden $q-$Knoten.

(6) Trianguliere die entstandenen Vierecke.

6.5 Überblick

Nach diesem Überblick konzentrieren wir uns auf die Möglichkeiten der numerischen Gittererzeugung. Dazu sehen wir uns zuerst die zur Lösung eines PDE-Problems notwendigen Schritte genauer an.

Lösung einer partiellen Differentialgleichung (PDE)

$$\begin{array}{lrl} \text{PDE:} & Lu = f & \text{in } \Omega \\ \text{BC:} & u = g & \text{in } \partial\Omega \end{array}$$

(Andere BCs sind möglich, wie später zu sehen sein wird.)

Lösungsschritte

(1) Geometrische Konfiguration:

Disketisierung, Transformation

Gittererzeugung

(2) Gittererzeugung

direkt oder mittels Transformation, per Hand oder automatisch.

manchmal iterativ:

(a) Erzeugung eines Startgitters (algebraische Methoden)

(b) Iterative Verbesserung.

(3) Transformation der PDE auf Ω_R

$$\left.\begin{array}{rlll} Lu = f & \text{in} & \Omega & \\ u = g & \text{auf} & \partial\Omega & \end{array}\right\} \rightarrow \left\{\begin{array}{rlll} \tilde{L}(u) = \tilde{f} & \text{in} & \Omega_R & \\ \tilde{u} = \tilde{g} & \text{auf} & \partial\Omega_R & \end{array}\right.$$

(Selbst wenn L linear war, kann und wird normalerweise $\tilde{L}$ nicht-linear sein, aber nur schwach nicht-linear, sozusagen quasi-linear.)

(4) Numerische Lösung der PDE (2) mit einer möglichst einfachen und schnellen Methode, die aber jetzt iterativ sein muß, weil PDE (2) nicht-linear ist.

2 Algebraische Gittererzeugung

Um die inneren Gitterpunkte

$(x_i, y_i) \in \Omega$ zu bestimmen, wenden wir Interpolationsmethoden und randabgleichende Ansätze an. Dazu stellen wir die Region

$$\Omega \subset \mathbb{R}^2$$

als Kurvenfamilien dar, die von den beiden Parametern ξ und η abhängen $(x(\xi,\eta), y(\xi,\eta))$:

$$\overline{\Omega}\{x(\xi,\eta), y(\xi,\eta) | 0 \leq \xi, \eta \leq 1\}$$

Dies ist äquivalent zu der Umformung:

1. Schritt: Zerlege $\partial\Omega$ (vier "Kanten" mit vier "Ecken"):

$$\partial\Omega = \underbrace{C_1 \cup}_{\text{"unten"}} \quad \underbrace{C_2 \cup}_{\text{"links"}} \quad \underbrace{C_3 \cup}_{\text{"oben"}} \quad \underbrace{C_4}_{\text{"rechts"}}$$

mit

$$\begin{aligned} C_1 \cap C_2 &= \big(x(0,0), y(0,0)\big) \\ C_2 \cap C_3 &= \big(x(0,1), y(0,1)\big) \\ C_3 \cap C_4 &= \big(x(1,1), y(1,1)\big) \\ C_4 \cap C_1 &= \big(x(1,0), y(1,0)\big), \end{aligned}$$

wobei die einzelnen Teile des Bereichs wie folgt gegeben sind:

entweder

als Gitter-Randpunkte $(x_i, y_i) \in \partial\Omega$ mit $(x_i, y_i) = \big(x(\xi_i, 0), y(\xi_i, 0)\big)$ z.B. für C_1 mit $0 \leq \xi_i \leq 1$.

oder

als Randfunktionen, ebenfalls gegeben als Funktionen der Parameter (ξ, η):

2. Schritt: Randdarstellung

$$\begin{aligned} C_1 = &\quad \{(x_1(\xi), y_1(\xi)) | 0 \leq \xi \leq 1\} \quad \eta = 0 \\ C_2 = &\quad \{(x_2(\eta), y_2(\eta)) | 0 \leq \eta \leq 1\} \quad \xi = 0 \\ C_3 = &\quad \{(x_3(\xi), y_3(\xi)) | 0 \leq \xi \leq 1\} \quad \eta = 1 \\ C_4 = &\quad \{(x_4(\eta), y_4(\eta)) | 0 \leq \eta \leq 1\} \quad \xi = 1 \end{aligned}$$

Zunächst werden wir uns drei Interpolationsmethoden, die direkte Anwendungen der Lagrange Interpolation sind, näher ansehen:

Vertikal-Interpolation zwischen den Gitterpunkten des unteren und des oberen Teils des Bereichs $C_1 - C_3$.

ZEICHNUNG

Die Gitterpunkte sind definiert durch äquidistante Punkte mit den Parametern ξ und η.

Horizontal-Interpolation entsprechend

ZEICHNUNG

Transfinite Interpolation, die aus einer Vertikal-Interpolation, gefolgt von einer Korrektur in der Horizontalen besteht.

Die Korrektur ist ein zusätzlicher Term, der aus der Differenz besteht zwischen den Differenzen der Ränder des physikalischen Bereichs und den Eckpunkten der interpolierten Grenzen (von links nach rechts). ???

Lagrange Interpolation, definiert durch

$$x(\xi) = \sum_{n=0}^{N} \Phi(\xi) x_n$$

mit $\Phi_n(\xi) = \prod_{i=0, i\neq n}^{N} \frac{\xi - \xi_i}{\xi_n - \xi_i}$, mit $\xi_i = ih, h = \frac{1}{n}$.

(Hierbei beschränken wir uns auf die $x - \xi$- Koordinaten mit offensichtlicher Übertragung auf y und η).

Normalerweise haben wir nun lediglich zwei Punkte ξ_0 auf C_1 und ξ_1 auf C_3, die zu einer Linearen Interpolation führen, so z.B. die einfache Formel

$$x(\xi) = x_0(\xi - \xi_0) + x_1(\xi - \xi_1)$$

Die Vertikal-Interpolation mit $\eta_0, \eta_1 = 1$ und den Bereichsfunktionen $\big(x_1(\xi), y_1(\xi)\big)$ auf C_1 und

$\big(x_3(\xi), y_3(\xi)\big)$ auf C_3

führt zu

$$(2.3) \begin{cases} x_v(\xi, \eta) = (1-\eta)x_1(\xi) + \eta x_3(\xi) \\ y_v(\xi, \eta) = (1-\eta)y_1(\xi) + \eta y_3(\xi) \end{cases}$$

Nun nehmen wir $k+1$ Punkte auf C_1 und C_3, beispielsweise

$$\begin{aligned} (x_{1i}, y_{1i}) &= \big(x_1(\xi_i), y_1(\xi_i)\big) \\ (x_{3i}, y_{3i}) &= \big(x_3(\xi_i), y_3(\xi_i)\big) \end{aligned}$$

mit $\xi_i = i \cdot h, \qquad h = \frac{1}{k}, \qquad , i = 0, \ldots, k$

und $m+1$ Werten $Y_j = j \cdot \tau, \qquad \tau = \frac{1}{m}$.

Somit erhalten wir $(m-1) \cdot (k-1)$ innere Gitterpunkte.

$$(2.4) \begin{cases} x(\xi_i, \eta_j) = (1-\eta_j)x_{1i} + \eta_j i x_{3i} \\ y(\xi_i, \eta_j) = (1-\eta_j)y_{1i} + \eta_j i x_{3i} \end{cases}$$

$i = 1, \ldots, k-1, \qquad j = 1, \ldots, m-1$

Horizontal-Interpolation ist entsprechend gegeben durch:

$$(2.5) \begin{cases} x_H(\xi, \eta) = (1-\xi)x_2(\eta) + \xi\ x_4(\eta) \\ y_H(\xi, \eta) = (1-\xi)y_2(\eta) + \xi\ y_4(\eta) \end{cases}$$

Die Transfinite Interpolation ist dann definiert als

$$(2.6)\quad \begin{cases} x(\xi,\eta) = x_\nu(\xi,\eta) + \\ \quad +(1-\xi)(x_2(\eta) - x_\nu(1,\eta) + \\ \quad +\xi(x_4(\eta) - x_\nu(1,\eta)) \\ y(\xi,\eta) \text{ entsprechend: ersetze alle Werte x durch y} \end{cases}$$

oder durch Zusammenschluß aller Terme

$$(2.7)\quad \begin{cases} x(\xi,\eta) = (1-\eta)x_1(\xi) + \eta x_3(\xi) + (1-\xi)x_2(\eta) + \xi x_4(\eta) \\ \quad -\{\xi\eta x_3(1) + \xi(1-\eta)x_1(1) + \\ \quad \eta(1-\xi)x_3(0) + (1-\xi)(1-\eta)x_1(0)\} \end{cases}$$

Im folgenden betrachten wir ein einfaches Beispiel mit nur einem zu berechnenden inneren Gitterpunkt.

$(0,6)$	$(4,6)$	$(8,6)$
$(-2,2)$	$\boxed{\left(x(\frac{1}{2},\frac{1}{2}), y(\frac{1}{2},\frac{1}{2})\right)}$	$(8,3)$
$(0,0)$	$(4,1)$	$(8,0)$

Dieses entspricht einem fast rechteckigen Gebiet $(0,8)$ x $(0,6)$ mit nur zwei krummlinigen Rändern C_1 und C_2 .

Vertikal-Interpolation (2.3) führt zu

$$\begin{aligned} x_\nu(1/2,1/2) &= 1/2\cdot 4 + 1/2\cdot 4 = 4 \\ y_\nu(1/2,1/2) &= 1/2\cdot 1 + 1/2\cdot 6 = 3.5 \end{aligned} \qquad (4, 3.5)$$

Horizontal-Interpolation (2.5) liefert die Werte

$$\begin{aligned} x_H(1/2,1/2) &= 1/2(-2) + 1/2\cdot 8 = 3 \\ x_H(1/2,1/2) &= 1/2\cdot 2 + 1/2\cdot 3 = 2.5 \end{aligned} \qquad (3, 2.5),$$

die sich von denen der Vertikal-Interpolation deutlich unterscheiden.

Um die Horizontal-Korrektur der Vertikal-Interpolation berechnen zu können, benötigt man zusätzlich die Werte der Vertikal-Interpolation für die Ränder C_2 und C_4:

$$x_\nu(0,1/2) = 1/2 \cdot 0 + 1/2 \cdot 0 = 0$$
$$y_\nu(0,1/2) = 1/2 \cdot 0 + 1/2 \cdot 6 = 3 \qquad (0,3)$$

$$x_\nu(1,1/2) = 1/2 \cdot 8 + 1/2 \cdot 8 = 8$$
$$y_\nu(1,1/2) = 1/2 \cdot 0 + 1/2 \cdot 6 = 3 \qquad (8,3)$$

Mit diesen Werten kann die transfinite Interpolation (2.6) wie folgt berechnet werden:

$$x(1/2,1/2) = 4 + 1/2(-2-0) + 1/2(8-8) = 3$$
$$y(1/2,1/2) = 3.5 + 1/2(2-3) + 1/2(3-3) = 3 \qquad (3,3)$$

Mit Hilfe von (2.7) hätten diese Werte auch direkt ermittelt werden können.

ZEICHNUNG

Spezialfälle:

(1) Rekonstruktion fehlender Randteile durch Interpolation.

(2) Zusätzliche Vorgabe innerer Punkte oder Gitterlinien (Blockaufteilung des Gebietes)

0: Eckpunkte

x: Vorgegebene Randpunkte

Im folgenden werden wir uns allgemeineren Methoden zur algebraischen Gittererzeugung zuwenden. Die sog. "Intrinsische Algebraische Gittererzeugung" wurde von P.M. Krupp in Castillo (Kap. 6) eingeführt.

Wie leicht zu erkennen ist, sind transfinite Interpolationskurven stetig und entsprechen den gegebenen Randfunktionen.

???? ????

Daher haben einige Mathematiker versucht, bessere Gitter mit Hilfe algebraischer Methoden zu erzeugen. Zwei dieser Gitter werden wir nun näher betrachten.

Intrinsische Gittererzeugung

ist als algebraische Gitter-Formel definiert, die die Grenzfunktionen mit den Ecken einschließt, aber keinen Gebrauch macht von ???Blending Function? so wie die Lagrange Interpolation $(\xi, \eta, 1-\xi, 1-\eta)$.

Gittererzeugung durch

Rand-erhaltende Ansätze AII (KNUPP)

Komplexe Randdarstellung ($i := \sqrt{-1}$)

$$\begin{aligned} C_1: &\quad z_1(\xi) = x_1(\xi) + iy_1(\xi) \\ C_2: &\quad z_2(\eta) = x_2(\eta) + iy_2(\eta) \\ C_3: &\quad z_3(\xi) = x_3(\xi) + iy_3(\xi) \\ C_4: &\quad z_4(\eta) = x_4(\eta) + iy_4(\eta) \end{aligned}$$

Gitter

$$z(\xi, \eta) = x(\xi, \eta) + iy(\xi, \eta)$$

Ansatz

$$z(\xi, \eta) = f(\eta) z_1(\xi) + g(\eta) z_3(\xi)$$

mit

$$f, g : [0, 1] \to \mathbb{C}$$

Bestimmung von f und g:

Lineares $2 \times 2-$ Gleichungssystem:

$f(\eta) z_1(0) + g(\eta) z_3(0) = z_2(\eta) \qquad \xi = 0$

$f(\eta) z_1(1) + g(\eta) z_3(1) = z_4(\eta) \qquad \xi = 1$

Determinate

$$D_{II} = z_1(0) z_3(1) - z_1(1) z_3(0)$$

Lösung:

$$f(\eta) = \begin{vmatrix} z_2(\eta) & z_3(0) \\ z_4(\eta) & z_3(1) \end{vmatrix} / D_{II}$$

$$g(\eta) = \begin{vmatrix} z_1(0) & z_2(\eta) \\ z_1(1) & z_4(\eta) \end{vmatrix} / D_{II}$$

Einsetzen:

$$\begin{aligned} z(\xi, \eta) = \frac{1}{D_{II}} \quad & \{ z_3(1) z_2(\eta) z_1(\xi) - \\ & - z_3(0) z_4(\eta) z_1(\xi) + \\ & + z_1(0) z_4(\eta) z_3(\xi) - \\ & - z_1(1) z_2(\eta) z_3(\xi) \} \end{aligned}$$

In unserem simplen Beispiel erhalten wir die folgenden Werte:

$$D_{II} = \underbrace{0}_{z_1(0)} - \underbrace{(8)}_{z_1(1)} \cdot \underbrace{6i}_{z_3(0)} = 48i$$

und mit (2.12) und (2.10):

$$
\begin{aligned}
-48i \cdot z(1/2,1/2) &= \begin{vmatrix} z_2(1/2) & z_3(0) \\ z_4(1/2) & z_3(1) \end{vmatrix} z_1(1/2) \\
&+ \begin{vmatrix} z_1(0) & z_2(1/2) \\ z_1(1) & z_4(1/2) \end{vmatrix} z_3(1/2)
\end{aligned}
$$

Mit dem Taschenrechner erhält man folgende Ergebnisse:

$-48i(x(1/2,1/2) + iy(1/2,1/2) = 164 - 154i$

$$\Rightarrow x(1/2,1/2) = \frac{-154}{-48} = 3.21$$

$$y(1/2,1/2) = \frac{164}{48} = 3.42$$

??? (Green cross on the slide)

Bei Verwendung von (2.13) gelangt man zu denselben Ergebnissen.

A III (KNUPP)

Diese Alternative beginnt mit einem Separationsansatz, die A II ähnelt und gewöhnliche Gitter erzeugt. Der Hauptgrund für die zusätzliche Einführung dieser Alternative ist der, daß sie leicht auf dreidimensionale Oberflächen ausgeweitet werden kann.

Separationsansatz

$$
(2.14)??? \begin{cases} x(\xi,\eta) = a_1(\eta)x_1(\xi) + b_1(\eta)y_1(\xi) + a_3(\eta)x_3(\xi) + b_3(\eta)y_3(\xi) \\ y(\xi,\eta) = a_1(\eta)y_1(\xi) + b_1(\eta)x_1(\xi) + a_3(\eta)y_3(\xi) + b_3(\eta)x_3(\xi) \end{cases}
$$

Dieser Ansatz unterscheidet sich an zwei Stellen von A II und zwar bezüglich der Vorzeichen von b_1 und b_3. Wie bei A II erhalten wir ein 4×4-lineares Gleichungssystem mit den Unbekannten a_1, b_1, a_3, b_3 und der Determinanten:

$(\xi = 0, \xi = 1, x, y)$

$$
D_{III} = \begin{vmatrix} x_1(0) & y_1(0) & x_3(0) & y_3(0) \\ y_1(0) & x_1(0) & y_3(0) & x_3(0) \\ x_1(1) & y_1(1) & x_3(1) & y_3(1) \\ y_1(1) & x_1(1) & y_3(1) & x_3(1) \end{vmatrix}
$$

Lösung:

$$
a_1(\eta) = \begin{vmatrix} x_2(\eta) & y_1(0) & x_3(0) & y_3(0) \\ y_2(\eta) & x_1(0) & y_3(0) & x_3(0) \\ x_4(\eta) & y_1(1) & x_3(1) & y_3(1) \\ y_4(\eta) & x_1(1) & y_3(1) & x_3(1) \end{vmatrix} D_{III}
$$

$b_1(\eta), a_3(\eta), b_3(\eta)$ entsprechend

(v.s. durch Spalten schieben!)

Für unser einfaches Beispiel erhalten wir die folgenden Werte:

$$D_{III} = \begin{vmatrix} 0 & 0 & 0 & 6 \\ 0 & 0 & 6 & 0 \\ 8 & 0 & 8 & 6 \\ 0 & 8 & 6 & 8 \end{vmatrix} = -2304$$

$$a_1(\frac{1}{2}) = \begin{vmatrix} -2 & 0 & 0 & 6 \\ 2 & 0 & 6 & 0 \\ 8 & 0 & 8 & 6 \\ 3 & 8 & 6 & 8 \end{vmatrix} / D_{III} = +8 \cdot (-264)/(-2304) = 0.92$$

$$b_1(\frac{1}{2}) = \begin{vmatrix} 0 & -2 & 0 & 6 \\ 0 & 2 & 6 & 0 \\ 8 & 8 & 8 & 6 \\ 0 & 3 & 6 & 8 \end{vmatrix} / D_{III} = 8 \cdot (-132)/(-2304) = 0.46$$

$$a_3(\frac{1}{2}) = \begin{vmatrix} 0 & 0 & -2 & 6 \\ 0 & 0 & 2 & 0 \\ 8 & 0 & 8 & 6 \\ 0 & 8 & 3 & 8 \end{vmatrix} / D_{III} = 8 \cdot (-96)/(-2304) = 0.33$$

$$b_3(\frac{1}{2}) = \begin{vmatrix} 0 & 0 & 0 & -2 \\ 0 & 0 & 6 & 2 \\ 8 & 0 & 8 & 8 \\ 0 & 8 & 6 & 3 \end{vmatrix} / D_{III} = 8 \cdot 96/(-2304) = -0.33$$

was mit (2.14) zu folgendem Ergebnis führt:

$$x(\frac{1}{2}, \frac{1}{2}) = 0.92 \cdot 4 + 0.46 \cdot 1 + 0.33 \cdot 4 + (-0.33) \cdot 6$$

$$y(\frac{1}{2}, \frac{1}{2}) = 0.92 \cdot 1 + 0.46 \cdot 4 + 0.33 \cdot 6 - 0.33 \cdot 4$$

daraus ergibt sich:

$$(\overline{x}, \overline{y}) = (3.48, 3.42)$$

was sich leicht von A II unterscheidet.

Im Knupp lassen sich zwei andere intrinsische Methoden finden, die auf einer gesunden mathematischen Basis miteinander verglichen werden.

Ein Nachteil von A II / A III ist der folgende:

Ein physikalisches Gebiet, das von einer realistischen Anwendung stammt, ist meistens unabhängig von einem Koordinatensystem. Daher sollte die numerische Gittererzeugung unabhängig sein, von den folgenden Operationen:

(1) Translation: $\tilde{z}(\xi, \eta) = z(\xi, \eta) + z_0$
(2) Rotation: $\tilde{\tilde{z}}(\xi, \eta) = e^{i\varphi} z(\xi, \eta)$
(3) Streckung: $\hat{z}(\xi, \eta) = r \cdot z(\xi, \eta)$

Alle Interpolationsmethoden führen zu geometrisch kongruenten Gittern, für Gebiete, die sich nur durch diese drei Bewegungen unterscheiden. A II dagegen ist unabhängig von

Drehung und Steckung, jedoch nicht von der Translation, A III ist weder unabhängig von der Translation noch von der Rotation.

Invarianz - Eigenschaften

Abhängigkeit der Gittererzeugung von einer Bewegung des Gebietes im Koordinatensystem:

(1) Translation: $\tilde{z}(\xi,\eta) = z(\xi,\eta) + z_0$

??? (2) Rotation: $\tilde{\tilde{z}}(\xi,\eta) = e^{i\varphi} z(\xi,\eta)$

(3) Streckung: $\hat{z}(\xi,\eta) = r \cdot z(\xi,\eta)$

(4) Spiegelung: $z'(\xi,\eta) = \overline{z}(\xi,\eta)$

Alle betrachteten algebraischen Gittererzeugungsmethoden sind unabhängig von (3) und (4). Interpolationsmethoden sind von allen starren Bewegungen (1) - (4) unabhängig

Invarianz-Tabelle		
	(1)	(2)
Interpolation	invariant	invariant
AII	nicht i.	invariant
$AIII$	nicht i.	nicht i.
AII_∞	invariant	invariant

Dieses ist einerseits schlecht aber kann andererseits genutzt werden, um unterschiedliche Gitter durch folgende Verfahren zu erzeugen (hier nur für A II):

AII mit Verschiebung

(1) Gegeben sind z_1, z_2, z_3, z_4 sowie ein komplexer Parameter $\zeta \in \mathbb{C}$.
Ersetze die gegebenen Randfunktionen durch

$$\left.\begin{array}{ll} \tilde{z}_1(\xi) = & z_1(\xi) + \zeta \\ \tilde{z}_2(\eta) = & z_2(\eta) + \zeta \\ \tilde{z}_3(\xi) = & z_3(\xi) + \zeta \\ \tilde{z}_4(\eta) = & z_4(\eta) + \zeta \end{array}\right\} \zeta \in \mathbb{C} \text{ gegeben}$$

(2) Erzeuge mit den Randfunktionen $\tilde{z}_i, i = 1, \ldots, 4$, mit AII ein Gitter

$$\tilde{z}(\xi, \eta j \zeta)$$

(3) Schiebe zurück

$$z(\xi,\eta) = \tilde{z}(\zeta, \eta j \zeta) - \zeta$$

Interessant ist $(\zeta) \to \infty : AII_\infty$, denn dieses Gitter ist unabhängig von allen 3 o.g. Bewegungen.

Auf die mathematischen Eigenschaften werden wir nicht näher eingehen. Stattdessen werden wir abschließend einige Beispiele betrachten.

Verallgemeinerter Ansatz

Shift $\Rightarrow$ Längenminimierung

$$\begin{aligned} x(\xi,\eta) &= g(\eta)x_1(\xi) + f(\eta)x_3(\xi) \\ &+ k(\xi)x_2(\eta) + h(\xi)x_4(\eta) - e(\xi,\eta) \\ y(\xi,\eta) &= \hat{g}(\eta)y_1(\xi) + \hat{f}(\eta)y_3(\xi) \\ &+ \hat{k}(\xi)y_2(\eta) + \hat{h}(\xi)y_4(\eta) - \hat{e}(\xi,\eta) \end{aligned}$$

Für diesen Ansatz ergibt eine Verschiebung

$$\begin{array}{rcl} x & \rightarrow & x+p \\ y & \rightarrow & y+q \end{array} \quad \text{shift } (p,q)$$

Gitterfunktionen $x^R(\xi,\eta)$ und $y^R(\xi,\eta)$ mit

$$x^R(\xi,\eta) = x(\xi,\eta) - p[1 - f(\eta) - g(\eta)][1 - h(\xi) - k(\xi)]$$

$$y^R(\xi,\eta) = y(\xi,\eta) - q[1 - \hat{f}(\eta) - \hat{g}(\eta)][1 - \hat{h}(\xi) - \hat{k}(\xi)]$$

Diese Eigenschaft kann ausgenutzt werden, ein Funktional zu minimieren, z.B. die Längen

$$\begin{aligned} l_{ij}^2 &:= (x_{ij}^R - x_{i-1,j}^R)^2 + (y_{ij}^R - y_{i-1,j}^R)^2 \\ &\text{und} \\ \tilde{l}_{ij}^2 &:= (x_{ij}^R - x_{i,j-1}^R)^2 + (y_{ij}^R - y_{i,j-1}^R)^2 \end{aligned}$$

über alle zu erzeugenden inneren Punkte:

$$\min F(p,q) = \sum_{j=1}^{M} \sum_{i=2}^{N} (l_{ji}^2 + \tilde{l}_{ij}^2)$$

Man zeigt leicht, daß für die Differenzen gilt:

$$x_{ij}^R - x_{j-1,j}^R = (x_{ij} - x_{i-1,j}) + pA_{ij}$$

$$\cdots (A_{ij} = A_{ij}(f,g,h,k))$$

und bekommt so eine entkoppelte Lösung für p, q:

$$p = -(\sum\sum A_{ij}(x_{ij} - x_{i-1,j}) + \sum\sum C_{ij}(x_{ij} - x_{i,j-1}))/\Lambda_1$$

$$q = -(\sum\sum B_{ij}(y_{i,j} - y_{i-1,j}) + \sum\sum D_{ij}(y_{ij} - y_{i,j-1})/\Lambda_2$$

mit

$$\Lambda_1 = \sum\sum (A_{ij}^2 + C_{ij}^2)$$

$$\Lambda_2 = \sum\sum (B_{ij}^2 + D_{ij}^2)$$

$$
\begin{aligned}
y^R_{ii} - y^R_{i-1,j} &= (y_{ij} - y_{i-1,j}) + qB_{ij} \\
x^R_{ij} - x^R_{i,j-1} &= (x_{ij} - x_{i,j-1}) + pC_{ij} \\
y^R_{ij} - y^R_{i,j-1} &= (y_{ij} - y_{i,j-1}) + qD_{ij} \\
A_{ij} &= [1 - f(\eta j) - g(\eta j)]\cdot \\
&\quad [(h(\xi_i) - h(\xi_{i-1})) + (k(y_i) - k(\xi_{i-1}))] \\
B_{ij} &= [1 - \hat{f}(\eta_i) - \hat{g}(\eta_i)][(\hat{h}(\xi_i) - \hat{h}(\xi_{i-1})) + (\hat{k}(\xi_i) - \hat{k}(\xi_{i-1}))] \\
C_{ij} &= [1 - h(\xi_i) - k(\xi_i)][(f(y_i) - f(y_{j-1})) + (g(\eta_i) - g(y_{j-1})] \\
D_{ij} &= [1 - \hat{h}(\xi_i) - \hat{k}(\xi_i)][(\hat{f}(y_j) - \hat{f}(y_{j-1})) + (g(\eta_j) - \hat{g}(\eta_{j-1}))]
\end{aligned}
$$

Geometrische Möglichkeiten für randkonforme Koordinatensysteme

Ein einfaches Beispiel: Kreisring

Das Gebiet Ω ist ein kreisförmiges Gebiet, das in der untenstehenden Figur mit zylindrischen Koordinatenlinien überdeckt ist.

Dabei ist Ω gegeben durch

$$\Omega = \{(\nu, \varphi) | r_1 < r < r_2, 0 < \varphi \leq 2\pi\}$$

mit den beiden Koordinaten (ν, φ) und der bekannten Beziehung zwischen kartesischen Koordinaten:

$$\begin{cases} x(\nu, \varphi) = \nu \cos\ \varphi \\ y(\nu, \varphi) = \nu \sin\ \varphi \end{cases}$$ Transformationsgleichungen

$$\begin{cases} \nu(x, y) = \sqrt{x^2 + y^2} \\ \varphi(x, y) = \arctan y/x \end{cases}$$ inverse Trans.-gleichungen

Ω läßt sich im physikalischen (ν, φ)-Raum darstellen als Rechteck $[\nu_1, \nu_2]$ x $[0, 2\pi]$, wobei die Grenze $\partial\Omega$ gegeben ist durch zwei konstante Geraden ν.

Der nicht-triviale Teil dieses Beispiels läßt sich wie folgt beschreiben:

Im (ν, φ)-Raum wird das Gebiet Ω durch seine Grenze $\partial\Omega$ nicht vollständig eingeschlossen. Das liegt daran, daß es sich um ein mehrfach zusammenhängendes Gebiet handelt, bei der wir einen sog. branch-cut angewendet haben, um zu einem einfach verbundenen Rechteck zu gelangen.

Eine andere Transformation:

(ν_1, ν_2) x $(0, 2\pi] \longrightarrow (0,1)^2$:

mit $\xi := \frac{\varphi}{2\pi}$, $\eta := \frac{\nu - \nu_1}{\nu_2 - \nu_1}$

bekommen wir

$$\begin{cases} x(\xi,\eta) = [\nu_1 + (\nu_2 - \nu_1)\eta]\cos(2\pi\xi) \\ y((\xi,\eta) = [\nu_1 + (\nu_2 - \nu_1)\eta]\sin(2\pi\xi) \end{cases} \qquad \xi, \eta \in (0,1)$$

Verallgemeinerung der Transformation:

$$\begin{cases} x(\nu,\varphi) = \nu(\eta)\cdot\cos(\varphi(\xi)) \\ y(\nu,\varphi) = \nu(\eta)\cdot\sin(\varphi(\xi)) \end{cases}$$

ZEICHNUNG

Das Branch Cut Prinzip (Periodic BC)

Sehen wir uns zur Verdeutlichung des Branch-Cut Prinzips noch einmal die einfachste PDE mit der einfachsten ?discretisation ??? an:

Normalerweise -ohne Branch-Cut- löst man die PDE wie folgt:

$-\Delta u = f$ in Ω

$u = 0$ auf $\partial\Omega$

Es existiert nur ein innerer Punkt, der Fünf-Punkte-Stern liefert lediglich $(-1) * 0$, so daß man folgende lineare Gleichung erhält:

$$\begin{aligned} 4u_1 &= h^2 f(0.5, 0.5) \\ u_1 &= \frac{1}{16} f(0.5, 0.5) \end{aligned}$$

Mit Hilfe des Branch-Cut sieht diese einfache Situation vollkommen anders aus:

ZEICHNUNG

In diesem Fall gibt es zwei unbekannte Punkte, so daß man zwei lineare Gleichungen erhält:

$$\begin{aligned} 4u_1 - 2u_2 &= h^2 f(0.5, 0.5) \\ -2u_1 + 4u_2 &= h^2 f(1, 0.5) \end{aligned}$$

Dieses Beispiel führt zu den Grundsätzen:

(1) Krummlinige Koordinaten, so wie eine der Koordinaten des Beispiels, sind konstant zu machen ??? auf der Grenzkurve, während die anderen monoton entlang der Kurve variieren

(2) Die Möglichkeit von "Branch-Cuts", die geeignet sind für mehrfach zusammenhängende und für kompliziertere Gebiete.

Das vorliegende Konzept kann übertragen werden auf die Transformation von willkürlich komplexen Gebieten.

Kommen wir noch einmal zurück zu unserem Beispiel. Wenn wir die Transformationen verallgemeinern, so daß wir anstelle von (3) folgende Gleichungen erhalten:

$$(7) \qquad \begin{aligned} x(\nu,\varphi) &= \nu(\eta)\cdot\cos(\varphi(\xi)) \\ y(\nu,\varphi) &= \nu(\eta)\cdot\sin(\varphi(\xi)) \end{aligned} \quad ,$$

Dann können Punkte in $\nu-$ oder $\varphi-$ Richtung nicht äquidistant sein.

Diese eindimensional-funktionale Beziehung zwischen φ und ξ erfordert, daß die relative Verteilung von Randpunkten um die inneren und äußeren Kreise dieselbe ist.

Wenn wir diese Bedingung vernachlässigen, kommen wir zum allgemeinen Fall.

Der allgemeinste Fall

$$\begin{matrix} \nu(\xi,\eta) \\ \varphi(\xi,\eta) \end{matrix} \quad \text{bzw.} \quad \begin{matrix} x(\xi,\eta) \\ y(\xi,\eta) \end{matrix} \qquad \xi,\eta \in (0,1)$$

Aber für diesen allgemeinen Fall haben wir zwei Restriktionen einzuführen, die zwar offensichtlich sind, aber erfüllt sein müssen.

Notwendige Restriktionen

(1) Zwei krummlinige Koordinatenlinien derselben Koordinate dürfen sich nicht schneiden (Gitterfalten)

(2) Zwei Linien unterschiedlicher Koordinaten sollen sich genau einmal schneiden.

Wir sollten zur Kenntnis nehmen, daß wir zylindrische Koordinaten (ν,Θ) nur deshalb gewählt haben, weil wir diese gut kennen. Im allgemeinen wird man kein solches Koordinatensystem verwenden können.

Verallgemeinerung useres Beispiels

$$\begin{aligned} \Omega &:= \{(x,y)|r_1^2 < x^2+y^2 < r_2^2\} \\ \partial\Omega &= C_1 \cup C_2 : \\ C_1 &:= \{(x,y)|x^2+y^2 = r_1^2\} \\ C_2 &:= \{(x,y)|x^2+y^2 = r_2^2\} \end{aligned}$$

Bestimme (Finde) zwei Abbildungen $\xi(x,y), \eta(x,y)$, die für $(x,y) \in \overline{\Omega}$ definiert sind, sodaß

(1) $\eta = 0$ für $x^2 + y^2 = r_1^2$
$\eta = 1$ für $x^2 + y^2 = r_2^2$

(2) und daß ξ auf $[0,1]$ monoton variiert auf C_1 und C_2 mit demselben Richtungssinn

(3) und daß die Restriktionen erfüllt sind.

ABER

Wir benötigen die inverse Transformation:

Finde

$$x(\xi,\eta), y(\xi,\eta) \text{ für } (\xi,\eta) \in \Omega_R, \Omega_R = (0,1)^2$$

sodaß

z.B.	$x^2(\xi,0) + y^2(\xi,0) = r_1^2$	$\eta = 0$
	$x^2(\xi,1) + y^2(\xi,1) = r_2^2$	$\eta = 1$
Periodizität:	$x(1+\xi,\eta) = x(\xi,\eta)$	$\xi = 0, 1$
	$y(1+\xi,\eta) = y(\xi,\eta)$	
und	$(x(\xi,\eta), y(\xi,\eta)) \in \Omega$	für $(\xi,\eta) \in \Omega_R$

gilt, denn dann kann man ein Gitter in Ω erzeugen, das dem Quadratgitter in Ω_R entspricht.

So kommen wir zu den folgenden "Guidelines":

Zusammenfassung (Guidelines):

(1) Finde eine glatte, gleichmäßige Transformation

$$\Omega \to \Omega_R$$

die eine Reihe "vernünftige" Bedingungen erfüllt.

(2) Bestimme die Inverse dieser Transformation:

$$\Omega_R \to \Omega$$

(3) Erzeuge ein Gitter auf Ω durch Berechnung aller Punkte $(x_i, y_i) \in \Omega$, die Abbildung der inversen Transformation von Punkten $(\xi_i, \eta_i) \in \Omega_R$ eines quadratischen Gitters sind:

Mit $\begin{matrix} \tilde{\xi}_k := k \cdot h, k = 1, \ldots, n-1 \\ \tilde{\eta}_j := j \cdot h, j = 1, \ldots, n-1 \end{matrix}$, $n = \frac{1}{h}$

wird $(\xi_i, \eta_i) := (\tilde{\xi}_k, \tilde{\eta}_j)$ (alle Kombinationen) $i = 1, \ldots, (n-1)^2$ ein quadratisches Gitter.

Zwei einfache Beispiele ohne DGL

Abb. 2.3.4: linear krummlinige Koordinaten

Abb. 2.3.5: quadratisch krummlinige Koordinaten

Für das Beispiel in Abbildung 2.3.4 gilt:

$$\begin{aligned} x(\xi,\eta) &= \xi + \frac{1}{2}\eta - \frac{3}{2} \\ y(\xi,\eta) &= \frac{1}{2}\xi + \eta - \frac{3}{2} \end{aligned}$$

sowie

$$\begin{aligned} \xi(x,y) &= \frac{4}{3}x - \frac{2}{3}y + 1 \\ \eta(x,y) &= -\frac{2}{3}x + \frac{4}{3}y + 1 \end{aligned}$$

Somit ergeben sich:

$$\underline{a}_1 = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}, \underline{a}_2 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}, g_{ij} = \begin{pmatrix} \frac{5}{4} & 1 \\ 1 & \frac{5}{4} \end{pmatrix}, \sqrt{g} = \frac{3}{4}$$

$$\underline{a}^1 = \begin{pmatrix} \frac{4}{3} \\ -\frac{2}{3} \end{pmatrix}, \underline{a}^2 = \begin{pmatrix} -\frac{2}{3} \\ \frac{4}{3} \end{pmatrix}, g^{ij} = \begin{pmatrix} \frac{20}{9} & -\frac{16}{9} \\ -\frac{16}{9} & \frac{20}{9} \end{pmatrix}, \sqrt{det(g^{ij})} = \frac{4}{3}$$

Für die quadratischen krummlinigen Koordinaten (Abb. 2.3.5) gilt:

$$\begin{aligned} x(\xi,\eta) &= \xi - \frac{1}{16}\eta^2 + \frac{1}{8}\eta + \frac{15}{16} \\ y(\xi,\eta) &= -\frac{2}{25}\xi^2 + \eta + 1 \end{aligned}$$

Somit ergeben sich:

$$\underline{a}_1 = \begin{pmatrix} 1 \\ -\frac{4}{25}\xi \end{pmatrix}, \underline{a}_2 = \begin{pmatrix} -\frac{1}{8}\eta + \frac{1}{8} \\ 1 \end{pmatrix}, g_{ij} = \begin{pmatrix} \frac{16}{625}\xi^2 + 1 & -\frac{4}{25}\xi - \frac{1}{8}(\eta - 1) \\ -\frac{4}{25}\xi - \frac{1}{8}(\eta - 1) & \frac{1}{64}(\eta - 1)^2 + 1 \end{pmatrix}$$

$$\sqrt{g} = 1 - \tfrac{1}{50}\xi(\eta - 1)$$

$$\underline{a}^1 = \frac{1}{\sqrt{g}} \begin{pmatrix} 1 \\ \frac{1}{8}\eta - \frac{1}{8} \end{pmatrix}, \underline{a}^2 = \frac{1}{\sqrt{g}} \begin{pmatrix} \frac{4}{25}\xi \\ 1 \end{pmatrix},$$

$$g^{ij} = \frac{1}{g} \begin{pmatrix} \frac{1}{64}(\eta - 1)^2 + 1 & \frac{4}{25}\xi + \frac{1}{8}(\eta - 1) \\ \frac{4}{25}\xi + \frac{1}{8}(\eta - 1) & \frac{16}{625}\xi^2 + 1 \end{pmatrix}$$

Es sei jedoch an dieser Stelle nochmals darauf hingewiesen, daß die Transformation in der Praxis nicht analytisch gegeben ist, sondern alle Werte mit Differenzenausdrücken gebildet werden.

Continuum Variational Formulation???

Die glattestmöglichen Transformationen

$$\Omega \longrightarrow \Omega_R$$

können durch das Lösen von Variationsproblemen erreicht werden, beispielsweise durch Minimierung der potentiellen Energie der Transformationsfunktion:

$$\min \frac{1}{2} \int_\Omega \xi_x^2 + \xi_y^2 + \eta_x^2 + \eta_y^2 + BC$$

die zu den Laplace-Gleichungen

$$\left.\begin{array}{l} \Delta\xi = 0 \\ \Delta\eta = 0 \end{array}\right\} \text{ in } \Omega \quad + BC$$

Diese müssen nun invertiert werden mit differentialgeometrischen Koeffizienten, die auf C.F. Gauß zurückgehen:

$$\left.\begin{array}{l} g_{22}x_{\xi\xi} + g_{11}x_{\eta\eta} - 2g_{12}x_{\xi\eta} = 0 \\ g_{22}y_{\xi\xi} + g_{11}y_{\eta\eta} - 2g_{12}y_{\xi\eta} = 0 \end{array}\right\} \text{ in } \Omega_R$$

mit $g_{11} = x_\xi^2 + y_\xi^2$, $g_{22} = x_\eta^2 + y_\eta^2$, $g_{12} = x_\xi x_\eta + y_\xi y_\eta$

Offensichtlich handelt es sich hierbei um ein System nichtlinearer partieller Differentialgleichungen, das quasi-linear genannt wird, weil es nur in den Ableitungen höchster Ordnung linesr ist. Iterativ wird es wie folgt gelöst:

(1) Erzeuge ein Startgitter in Ω

(2) Berechne die Transformationskoeffizienten g_{11}, g_{22}, g_{12} in allen inneren Gitterpunkten durch Differenzenquotienten

$$G_{11}(i,j) = \left(\frac{x(\xi_{i+h}, \eta_i) - x(\xi_i, \eta_i)}{h}\right)^2 + \left(\frac{y(\xi_{i+h}, \eta_i) - y(\xi_i, \eta_i)}{h}\right)^2$$

usw.

(3) Löse die PDE (*) als lineare PDE, indem die Koeffizienten $g_{22}, g_{11}, -2g_{12}$ konstant gesetzt werden in jedem Gitterpunkt.

Interiere solange, bis $x(\xi_i, \eta_i)$ und $y(\xi_i, \eta_i)$ sich nur in ϵ- Grenzen unterscheiden, wobei ein $\epsilon > 0$ vorgegeben wird.

Discrete Variational Grid Generation ???

Eine andere Möglichkeit besteht darin, daß man eine Anzahl von Gitterpunkten als ???redistributed definiert und dann ein diskretes Variationsproblem löst. Dafür gibt es viele Möglichkeiten, von denen drei hier genannt werden:

(1) minimiere $F_s := \sum s_{i,j}^2$ Längen

(2) minimiere $F_A := \sum A_{ijhl}$ Gebiet

(3) minimiere $F_0 := \sum \vec{P}_i \cdot \vec{P}_j$ Produkt von Punkten oder

Man kann natürlich eine gewichtete Summe (Linearkombination) dieser Funktionen minimieren:

$$\text{minimiere} \quad F := aF_s + bF_A + cF_0 \qquad \text{mit } a, b, c \geq 0, \quad a + b + c = 1$$

Normalerweise hat man dann zusätzliche Bedingungen zu erfüllen, aber wir wollen hier nicht auf einzelheiten eingehen.

Erstaunlicherweise ist dieses diskrete Variationskonzept mathematisch sehr viel sensibler als die kontinuierliche Variationsformulierung. Wie leicht zu sehen ist, gibt es selbst bei einfachen Problemen keine oder mehrere Lösungen (siehe Kapitel 4 in Castillo).

Anhang A

Tatsachen aus der Funktionalanalysis

A.1 Banachräume

(nach Meis-Marcowitz, [MM78]) Wichtig für Stabilitäts- und Konvergenzbeweise nach der sog. Lax-Richtmeyer-Theorie, auch für Existenzsätze bei verallgemeinerten Aufgabenstellungen.

Definition A.1.1 Es sei B ein Vektorraum über dem Körper K ($K = \mathbb{R}$ oder $K = \mathbb{C}$). Dann heißt B reeller bzw. komplexer Banach-Raum, falls gilt:

(i) In B ist eine Abb. $\|\cdot\| : B \to [0,\infty)$, Norm, erklärt mit folgenden Eigenschaften:

$$\text{(S1)} \quad \begin{array}{ll} (a) & \|a\| = 0 \Longleftrightarrow a = 0 \qquad \forall a \in B \\ (b) & \|\lambda a\| = |\lambda|\,\|a\| \qquad \forall \lambda \in K, \forall a \in B \\ (c) & \|a+b\| \le \|a\| + \|b\| \quad \forall a, b \in B. \end{array}$$

(ii) B ist (bez. der Norm-Topologie) vollständig, d.h. jede Cauchy-Folge $\{a_n\}_{n\in\mathbb{N}}$, $a_n \in B$, konvergiert gegen ein Element $a \in B$.

$\{a_n\}$ heißt Cauchy-Folge, falls es zu jedem $\varepsilon > 0$ ein $n_0 \in \mathbb{N}$ gibt, so daß

$$\text{(S2)} \quad \|a_n - a_m\| < \varepsilon, \text{ falls } n, m > n_0.$$

$\{a_n\}$ heißt konvergent gegen $a \in B$, wenn $\|a_n - a\| \to 0$.

Banachräume sind also sowohl bez. der zugrundeliegenden Vektorräume (Elemente) als auch bez. der Norm zu unterscheiden.

Beispiel A.1.2 : Es sei $D \subset \mathbb{R}^m$ eine kompakte Menge. Der Vektorraum der stetigen komplexwertigen Vektor-Funktionen $C^0(D), f : D \to \mathbb{C}^n$ ([MM78]: $C^0(D,\ \mathbb{C}^n)$) ist vollständig bez. der ∞-Norm:

$$\|f\|_\infty := \max_j \max_{x\in D} |f_j(x)|$$

(Beweis: Wegen der Kompaktheit von D ist jede Cauchy-Folge stetiger Funktionen eine gleichmäßig konvergente Folge. Eine solche Folge hat aber eine stetige Funktion als Grenzwert.)

Beispiel A.1.3 Gegenbeispiel: Raum wie oben, aber mit der 2-Norm

$$\|f\|_2 := \left(\int_D \sum_{j=1}^n \overline{f_j(x)} f_j(x) dx \right)^{1/2}$$

Die Folge

$$f^{(i)}(x) = \begin{cases} x^i & \text{für} \quad x \in (0,1) \\ 1 & \text{für} \quad x \in [1,2] \end{cases}$$

konvergiert im Cauchy-Sinne. Der Grenzwert ist aber die unstetige Funktion,

$$f(x) = \begin{cases} 0 & \text{falls} \quad x \in (0,1) \\ 1 & \text{falls} \quad x \in [1,2] \end{cases} ,$$

also nicht im Raum.

Deshalb wird ab sofort immer angenommen, daß der Raum $C^0(...)$ mit der $\|\cdot\|_\infty$ versehen ist.

Beispiel A.1.4 $D \subset \mathbb{R}^m$.
$\bigwedge = \{f : D \to \mathbb{C}^n | f$ quadratintegrabel in $D\}$ mit der Norm

$$|||f|||_2 := \left(\int_D \sum_{j=1}^n (f_i(x)\overline{f_j(x)}) dx \right)^{1/2}$$

Dabei heißt eine Funktion quadratintegrabel, wenn das in der obigen Definition angegebene Integral im Lebesgueschen Sinne existiert.

Nun ist aber $|||\cdot|||_2$ keine Norm, da

$$|||f|||_2 = 0 \quad \forall f \in N := \{f \in \bigwedge | \{x \in D | f(x) \neq 0\} \text{ hat das L.-Maß } 0\}$$

Deshalb geht man zum Quotientenraum $\bigwedge / N$ über, d.h. man identifiziert Funktionen, die sich nur auf einer Menge vom Maß 0 unterscheiden.

Der Vektorraum $\bigwedge / N$ wird mit der Norm

$$\|\tilde{f}\|_2 := |||f|||_2 \quad \tilde{f} \in \frac{\bigwedge}{N}, f \in \tilde{f}$$

zum Banach-Raum $L^2(D)$.

Definition A.1.5 Sei B ein Banachraum.
D_1, D_2 Teilmengen von B, $D_1 \subset D_2$. Dann heißt D_1 dicht in D_2, wenn es zu jedem $a \in D_2$ und zu jedem $\varepsilon > 0$ ein $b \in D_1$ gibt mit $\|a - b\| < \varepsilon$.

Wichtig werden sein (auch später in Hilbert-Räumen) Raumschachtelungen, bei denen jeder Banachraum im nächstgrößeren dicht liegt. Diese dichten Schachtelungen liefern wichtige Approximationseigenschaften bzw. spiegeln diese wider.

Satz A.1.6 Approximationssatz von Weierstraß. Es sei $D \subset \mathbb{R}^m$ eine kompakte Menge. Dann gilt: Der Vektorraum der auf D definierten Polynome $\mathbb{P}(D)$ mit komplexen Koeffizienten liegt dicht in $C^0(D)$.

Das liefert z.B. eine Raumschachtelung

$$\mathbb{P} \subset C^\infty \subset C^k \subset C^0,$$

wobei jeder im nächsten dicht liegt. Es gilt sogar:

Satz A.1.7 (i) Der Vektorraum $V := \{f \in C^\infty([a,b]) |\ f^{(\nu)}(a) = f^{(\nu)}(b) = 0, \nu \in \mathbb{N}\}$ liegt dicht im $C^0([a,b])$.

(ii) Der Vektorraum der beschränkten Funktionen aus $C^\infty(\mathbb{R})$ liegt dicht im Banachraum der beschränkten Funktionen aus $C^0(\mathbb{R})$.

Bew.: [MM78]

Satz A.1.8 Es sei $D \subset \mathbb{R}^m$. $C_0^\infty(D)$ ist der Vektorraum der auf D definierten analytischen Funktionen mit kompaktem Träger.

(1) $C_0^\infty(D)$ liegt dicht in $L^2(D)$

(2) $\mathbb{P}(D)$ liegt dicht in $L^2(D)$

Definition A.1.9 B_1, B_2 Banachräume. D Unterraum von B_1. Eine Abbildung $A : D \to B_2$ heißt linearer Operator, falls gilt

$$A(\lambda a + \mu b) = \lambda Aa + \mu Ab \quad \forall \lambda, \mu \in K,\ \forall a, b \in D$$

Ein linearer Operator A heißt beschränkt, wenn es ein $\alpha > 0$ gibt mit $\|Aa\| \leq \alpha \|a\|\ \ \forall a \in D$. Die Größe

$$\|A\| := \inf\{\alpha \in \mathbb{R}_+ |\ \|Aa\| \leq \alpha \|a\|\ \forall a \in D\}$$

heißt dann Norm des Operators A.
Diese Klasse der linearen und beschränkten Operatoren von $D \to B_2$ bezeichnet man mit

$$L(D, B_2) := \{A : D \to B_2 | A \text{ linear und beschränkt}\}$$

Satz A.1.10 B_1, B_2 Banachräume, D dichter Unterraum in B_1, $A \in L(D, B_2)$. Dann gibt es eine eindeutig bestimmte normerhaltende Fortsetzung $\hat{A}$ von A in $L(B_1, B_2)$,

$$\hat{A} \in L(B_1, B_2), \quad \|\hat{A}\| = \|A\|.$$

Beweis: Es sei $a \in B_1$. Dann gibt es eine Folge $\{a_n\}, a_n \in D, a_n \to a$. Aa_n ist Cauchy-Folge in B_2, denn

$$\|Aa_j - Aa_k\| = \|A(a_j - a_k)\| \leq \|A\| \, \|a_j - a_k\|$$

$$\leq \underbrace{\|A\|}_{\text{beschränkt}} \quad (\underbrace{\|a - a_j\|}_{\to 0} + \underbrace{\|a - a_k\|}_{\to 0})$$

B_2 Banachraum $\Rightarrow Aa_j \to c$
c hängt nur von a ab, wie man sich durch Abschätzung einer zweiten Folge überzeugen kann. Definiere

$$\hat{A} : \hat{A}a := c$$

Dann ist $\hat{A}$ linear (für alle Werte $\hat{A}$ Grenzwerte A), ebenso zeigt man mit Hilfe der Grenzwerte

$$\|\hat{A}a\| = \|\lim \hat{A}a_j\| = \lim \|Aa_j\| \leq \|A\| \lim \|a_j\| = \|A\| \, \|\lim a_j\| = \|A\| \, \|a\|$$

die Beschränktheit und Normerhaltung von $\hat{A}$.

Eindeutigkeit: $\|\tilde{A}a - \hat{A}a\| \leq \|\tilde{A}a - Aa_j\| + \|\hat{A}a - Aa_j\| =$
$\|\tilde{A}a - \tilde{A}a_j\| + \|\hat{A}a - \hat{A}a_j\| \leq \|\tilde{A}\| \, \|a - a_j\| + \|\hat{A}\| \, \|a - a_j\| \to 0$

Der Satz von Hahn-Banach liefert, daß sich ein Operator $A : D \to B_2$ auch dann auf ganz B_1 fortsetzen läßt, wenn D nicht dicht in B_1 liegt. Dann ist allerdings die Fortsetzung nicht eindeutig.

Definition A.1.11 Sei M eine Menge von linearen beschränkten Operatoren $A : B_1 \to B_2$. M heißt glm. beschränkt, wenn $\{\|A\| \mid A \in M\}$ beschränkt ist, d.h. $\exists \alpha$:

$$\|Aa\| \leq \alpha \|a\| \quad \forall A \in M \quad \forall A \in B_1$$

Satz A.1.12 Prinzip der gleichmäßigen Beschränktheit (uniform boundedness).
B_1, B_2 Banachräume. $M \subset L(B_1, B_2)$. Es existiere eine Funktion $\beta : B_1 \to \mathbb{R}_+$ mit $\|Aa\| \leq \beta(a) \quad \forall a \in B_1, \forall A \in M$.
Dann ist M glm. beschränkt.

Definition A.1.13 Es sei B Banachraum, $I := [T_1, T_2] \subset \mathbb{R}$ ein reelles Intervall. Eine Abb. $u : [T_1, T_2] \to B$ heißt differenzierbar in $t_0 \in I$, falls es ein $a \in B$ gibt mit

$$\lim_{h \to 0, \; t_0 + h \in I} \frac{\|u(t_0 + h) - u(t_0) - h\, a\|}{|h|} = 0$$

a ist eindeutig bestimmt und heißt Ableitung:

$$a = u'(t_0) = \frac{du}{dt}(t_0)$$

Die Abb. u heißt differenzierbar, falls obiges für jedes $t_0 \in I$ gilt, und glm. diff.bar, falls zusätzlich

$$\frac{1}{|h|}\|u(t+h) - u(t) - hu'(t)\|$$

mit $h \to 0$ glm. gegen Null konvergiert $\forall t \in I$. u heißt stetig diff.bar, wenn $u'(t)$ in I stetig ist.

Folgerung A.1.14 (i) differenzierbar in $t_0 \Rightarrow$ stetig in t_0

(ii) Verallgem. MWS

$$\frac{1}{|h|}\|u(t+h) - u(t) - hu'(t)\| \leq \sup_{0<\nu<1} \|u'(t+\nu h) - u'(t)\|$$

(iii) stetig diff.bar $\Rightarrow$ glm. diff.bar

Integral im Banachraum

Definition A.1.15 B Banachraum, $I \subset \mathbb{R}, v \in C^0(I), v : I \to B$. $u : I \to B$ heißt Stammfunktion von v in I, falls u diff.bar und $u'(t) = v(t) \quad \forall v \in I$.

Lemma A.1.16 (1) u_1, u_2 Stammfunktionen von v in $I \Rightarrow u_1 - u_2$ konstant in I.

(2) Jede stetige Funktion v besitzt mindestens eine Stammfunktion u in I.

Definition A.1.17 B, I, v wie oben, u Stammfunktion von v in I. Dann heißt

$$\int_{T_1}^{T_2} v(t)dt := u(T_2) - u(T_2)$$

das Integral von v zwischen T_1 und T_2

Satz A.1.18 Eigenschaften von Integralen im Banachraum B:
Seien $I := [T_1, T_2] \subset \mathbb{R}$, $J := [T_3, T_4] \subset \mathbb{R}$, $c_1 \in \mathbb{R}$, $v \in C^0(I) \to B$.
Sei weiter $C \in L(B, B)$, f, g, $f_x(t,x) := \frac{\partial f}{\partial x}(t,x) \in C^0(I \times J) \to B$.
Dann gilt:

(i)
$$\int_{T_1}^{T_2} C(v(t))dt = c_1 \left(\int_{T_1}^{T_2} v(t)dt \right)$$

(ii) $\forall \varepsilon > 0\ \exists \delta > 0$, so daß für jede δ-Zerlegung von I:

$$\left\| \int_{T_1}^{T_2} v(t)dt - \sum_{k=0}^{n-1} v(t_k)(x_{k+1} - x_k) \right\| < \varepsilon$$

$(T_1 = x_0 \leq t_0 \leq x_1 \leq ... \leq x_k \leq t_k \leq x_{k+1} \leq ... \leq x_n = T_2,$
$x_{k+1} - x_k < \delta)$

(iii) $u(t) := \int_{T_1}^{t} v(\tau)d\tau$ ist stetig diff.bare Stammfunktion von v in I.

(iv) $w(x) := \int_{T_1}^{T_2} f(t,x)dt$ ist stetig diff.bar in J mit

$$\frac{d}{dx}w(x) = \int_{T_1}^{T_2} f_x(t,x)dt \qquad \forall x \in J$$

(v)

$$\int_{T_1}^{T_2}\left(\int_{T_3}^{T_4} g(t,x)dx\right)dt = \int_{T_3}^{T_4}\left(\int_{T_1}^{T_2} g(t,x)dt\right)dx$$

Bemerkung A.1.19 AWA werden wir oft in der Form

$$y' = Ty \qquad T : B \to B, \qquad B \text{ Banachraum}$$

schreiben. Bei Differenzenverfahren sind diese Abb. oft implizit gegeben, also $Ry' = Sy$, $R, S \in L(B,B)$.
Ist diese Gleichung $\forall x \in B$ eindeutig lösbar, so schreiben wir <u>formal</u> $T = R^{-1} \circ S$. Meist gilt auch konkret, daß $R^{-1} \in L(B,B)$ ist.

Satz A.1.20 Banach'scher Fixpunktsatz im Banachraum

$$D \in L(B,B), \qquad \|D\| < 1, \qquad R := I + D.$$

Dann gibt es $R^{-1} \in L(B,B)$ mit

$$R^{-1} \circ R = R \circ R^{-1} = I$$

<u>Beweis:</u> Neumann-Reihe von Op.

Bei impliziten Differenzenverfahren ergibt sich die Existenz von R^{-1} aus Stabilitätsaussagen zusammen mit dem Banachschen Fixpunktsatz.

Satz A.1.21 Es sei $R : [0,1] \to L(B,B)$ stetig, $R(0) = I$ und

$$\Lambda = \{\lambda \in [0,1] |\ \exists R^{-1}(\lambda) \in L(B,B) \text{ mit } R^{-1}(\lambda) \circ R(\lambda) = R(\lambda) \circ R^{-1}(\lambda) = I\}.$$

Wenn die Menge $\{R^{-1}(\lambda) | \lambda \in \Lambda\}$ glm. beschränkt ist, ist $\Lambda = [0,1]$.

<u>Bew.</u>: einfach, siehe [MM78]

A.2 Sobolev-Räume

Definition A.2.1 Unter einem Prae-Hilbert-Raum H (oder unvollständiger Hilbert-Raum) verstehen wir einen linearen Vektorraum über einem (reellen oder komplexen) Zahlenkörper K, auf dem eine bilineare Abbildung - genannt Skalarprodukt $(\cdot,\cdot) \to K$ -, definiert ist mit den Eigenschaften:

(i)
$$\begin{aligned}(\lambda f_1 + \mu f_2, g) &= \lambda(f_1, g) + \mu(f_2, g) \\ &\forall \lambda, \mu \in K, \forall f_1, f_2, g \in H\end{aligned}$$

(ii) $(f,g) = \overline{(g,f)}$

(iii) $(f,f) \geq 0 \qquad \forall f \in H, (f,f) = 0 \Leftrightarrow f = 0$

Folgerung A.2.2 (i) Jedem Element f aus H kann eine Länge zugeordnet werden, die eine Norm ist:
$$\|f\| := +\sqrt{(f,f)}$$

(ii) Es gilt die Schwarzsche Ungleichung
$$|(f,g)| \leq \|f\|\,\|g\| \qquad \forall f, g \in H$$

Definition A.2.3 Ein vollständiger Prae-Hilbert-Raum heißt Hilbert-Raum. (i.e. Banach-Raum mit Skalarprodukt)

Vervollständigung

Satz A.2.4 Jeder Prae-Hilbert-Raum H' kann in einen Hilbert-Raum H isometrisch und dicht eingebettet werden, d.h. $H' \subset H$ und auf H' stimmen Norm und Skalarprodukt von H und H' überein.

Definition A.2.5 (i) Eine Linearform (oder lineares Funktional) in H ist eine lineare Abbildung von $H \to K$:
$$l(\lambda f + \mu g) = \lambda l(f) + \mu l(g) \qquad \forall \lambda, \mu \in K$$

(ii) Eine Bilinearform $Q(f,g)$ ist eine Abb. $H \times H \to K$, die Linearform $\forall f, g \in H$ ist bei festem f bzw. g.

(iii) Eine Linearform heißt beschränkt, wenn
$$|l(f)| \leq \text{ const. } \|f\| \qquad \forall f \in H \text{ gilt .}$$

(iv) Der Raum $H' := L(H,K)$ der beschränkten linearen Funktionale heißt Dualraum zu H.

Folgerung A.2.6 (i) $l(f)$ ist stetig auf $H \Leftrightarrow l(f)$ ist beschränkt auf H.

(ii) Eine Bilinearform $Q(f,g)$ ist stetig $\Leftrightarrow$
$Q(f,g)$ ist beschränkt auf $H \times H$:

$$|Q(f,g)| \leq \text{ const. } \|f\| \, \|g\|$$

(iii) Riesz' Darstellungssatz: $F \in H' \Leftrightarrow \exists_1 v \in H : \forall u \in H : F(u) = (u,v)$ und $||F|| = \|v\|$
(Isomorphismus zwischen H und H')

Definition A.2.7 (i) Eine Teilmenge M eines Vektorraumes B heißt lineare Mannigfaltigkeit, wenn aus

$$f, g \in M \quad \Rightarrow \quad \lambda f + \mu g \in M \quad \forall \lambda, \mu \in K.$$

(ii) Eine Menge $M \in B$ heißt abgeschlossen, wenn aus

$$f_i \to f, f_i \in M \Rightarrow f \in M.$$

(iii) Eine abgeschlossene lineare Mannigfaltigkeit M in B heißt Unterraum von B.

Folgerung A.2.8 (i) Jeder Unterraum U eines Hilbertraumes H ist wieder ein Hilbert-Raum.

(ii) Jede endlichdimensionale lineare Mannigfaltigkeit ist abgeschlossen.

Einige Bezeichnungen:

Unter einem Gebiet $D \subset \mathbb{R}^n$ wollen wir - wenn nichts anderes gesagt ist - immer eine offene zusammenhängende Punktmenge verstehen. Zur Darstellung der Differentialoperatoren in allgemeiner Form verwenden wir sog. Multi-Indizes:

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \qquad \alpha_i \in \mathbf{N}_0$$

Ordnung von α ist

$$|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n$$

Es sei

$$\alpha! = \alpha_1! \alpha_2! \ldots \alpha_n!$$

Sei $x \in \mathbb{R}^n$, dann ist mit $x = (x_1, x_2, \ldots, x_n)$

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$$

Die partiellen Ableitungen wollen wir abkürzen mit

$$D_i := \frac{\partial}{\partial x_i}$$

Eine Ableitung p-ter Ordnung wird dann dargestellt als

$$D^\alpha = D_1^{\alpha_1} D_2^{\alpha_2} \ldots D_n^{\alpha_n} = \frac{\partial^p}{(\partial x_1)^{\alpha_1} \ldots (\partial x_n)^{\alpha_n}}$$

mit $|\alpha| = p$.

So ergibt sich z.B. als Potenzreihenentwicklung einer analytischen Funktion

$$f(x) = \sum_{|\alpha|=0}^{\infty} \frac{D^\alpha f(x_0)}{\alpha!}(x - x_0)^\alpha$$

dabei wird über alle möglichen Kombinationen für $|a| = p$ summiert.

Definition A.2.9 (i) f ist Hölder stetig auf D, falls

$$|f(x) - f(x')| \leq c|x - x'|^\mu \quad \forall x, x' \in D$$

mit Konstanten c und μ, $\quad 0 < \mu \leq 1$.

(ii) f heißt gleichmäßig H-stetig, falls c und μ für alle $x, x' \in D$ gleich sind.

(iii) f heißt Lipschitz-stetig, wenn $\mu = 1$.

(iv) Die Funktionen $f \in C^p(D)$, deren p-te Ableitung auf D H-stetig ist mit einem Exponenten μ bilden einen Vektorraum $C^{p+\mu}(D)$. Dieser ist ein Banach-Raum mit der Norm

$$|f|_{p+\mu} = |f|_p + \max_{|\alpha|=p}[D^\alpha f]_\mu$$

mit der Norm

$$|f|_p := \max_{|\alpha|\leq p} \sup_{x \in D} |D^\alpha f(x)|$$

und der Semi-Norm

$$[g]_\mu := \sup_{x, x' \in D} \frac{|g(x) - g(x')|}{|x - x'|^\mu}$$

Beispiel A.2.10 Räume

(a) $L^p(D)$ ist der Banachraum der auf D p-summierbaren Funktionen mit der Norm

$$\|f\|_p := \left(\int_D |f(x)|^p dx\right)^{\frac{1}{p}}$$

(Äquivalenzklassenbildung)

(b) Der Raum $L^2(D)$ ist mit dem Skalarprodukt

$$(f, g)_D := \int_D f(x)\,\overline{g(x)}dx$$

ein Hilbert-Raum.

(c) Wird der Raum $C^m(D)$ in der Norm

$$\|f\|_m := \left(\int_D \sum_{|\alpha|\leq m} |D^\alpha f|^2\right)^{\frac{1}{2}}$$

vervollständigt, so entsteht der Hilbert-Raum, den man Sobolev-Raum nennt,

$$H^m(D).$$

Das zugehörige Skalarprodukt ist

$$(f,g)_m := \int_D \Big(\sum_{|\alpha| \le m} D^\alpha f \, D^\alpha g \Big) dx$$

Jeder Funktion $f \in H^m$ kann man eindeutig eine Ableitung zuordnen, indem man eine Folge $\{f_i\}, f_i \in C^m$, nimmt mit

$$\lim_{i \to \infty} \|f - f_i\|_m = 0$$

Dann ist jede Folge $D^\alpha f_i$ mit $0 \le |\alpha| \le m$ eine Cauchy-Folge, die einen Grenzwert in $H_{m-|\alpha|}$ besitzt, den wir mit $D^\alpha f$ bezeichnen.

Folgerung A.2.11 Kalkül in Sobolev-Räumen

(a) Es gilt offenbar

$$\ldots \subset H^m(D) \subset H^{m-1}(D) \subset \ldots \subset H^0(D)$$

wegen

$$\|f\|_0 \le \|f\|_1 \le \ldots \le \|f\|_{m-1} \le \|f\|_m$$

(b)

$$H^m(D) \subset H^m(\tilde{D}) \text{ falls } \tilde{D} \subset D$$

(c) Ist

$$\eta \in C^m(\overline{D}), f \in H^m(D)$$

$$\{f_p\} \text{ eine Folge, } f_p \in C^m(\overline{D})$$

mit

$$f_p \to f \text{ in } H^m,$$

dann ist auch $\{\eta f_p\}$ eine Cauchy-Folge mit Grenzwert ηf in H^m.

(d) Ist $a \in C^0(\overline{D}), D$ kompakt, also $\|a\|_\infty \le M$ und sind $f, g \in C^m(\overline{D})$, so gilt für beliebige Multiindizes α, β mit $|\alpha|, |\beta| \le m$

$$(*) \left| \int_D a D^\alpha f D^\beta g dx \right| \le M \, \|f\|_m \, \|g\|_m,$$

so daß nach dem Fortsetzungssatz in Banachraum, der auch für lineare Funktionale gilt,

$$\int_D a D^\alpha f D^\beta g dx$$

auch für $f, g \in H^m(D)$ definiert und $(*)$ gültig ist.

Beispiel A.2.12 Sei $C_0^m(D)$ die Klasse der in D m-fach stetig differenzierbaren Funktionen mit kompaktem, ganz in D gelegenen Träger (d.h. Randwerte 0 bis zur m-1-ten Ordnung.) Dann ist der Sobolev-Raum $H_0^m(D)$ als die Vervollständigung von $C_0^m(D)$ in der m-Norm definiert. (oder als die von $C_0^\infty(D)$, da C_0^∞ dicht in C_0^m).

Satz A.2.13 (i) Sei $f \in H_0^m(D) \cap C^{m-1}(\overline{D})$.

Sei ∂D von der Klasse C^1.

Dann gilt $D^\alpha f = 0$ für $|\alpha| \leq m-1$ auf ∂D.

(ii) Umgekehrt: Sei $f \in C^m(\overline{D})$ und

$$D^\alpha f = 0 \text{ für } |\alpha| \leq m-1 \text{ auf } \partial D.$$

Dann ist

$$f \in H_0^m(D).$$

Definition A.2.14 $f, g \in H^m(D)$. Dann sagen wir

$$D^\alpha f = D^\alpha g \text{ für } |\alpha| \leq m-1 \text{ auf } \partial D,$$

falls

$$f - g \in H_0^m(D).$$

Bemerkung A.2.15 (1) Die "Kalkül-Eigenschaften" gelten auch in den Räumen H_0^m.

(2) Partielle Integration etc. lässt sich in H_0^m bzw. H^m analog zu C^m durchführen. Beweis immer durch Grenzübergang aus $C^m \to H^m$.

Noch ein Nachtrag zum Hilbert-Raum:

Definition A.2.16 (i) Zwei Funktionen $f, g \in H, H$ H.R., heißen <u>orthogonal</u>, falls

$$(f, g) = 0$$

(ii) Ein linearer Operator $T : H \to U \subset H$, heißt Projektion $\Leftrightarrow Tg = g$
$\forall g \in U, Tf = 0, f \in V \quad H = U + V$

(iii) Eine Projektion heißt orthogonale Projektion $\Leftrightarrow (f - Tf, g) = 0 \; \forall f \in H$

Satz A.2.17 <u>Projektionssatz</u>
Sei $f \in H, U$ Unterraum von H.

(i) Dann gibt es genau ein Element $Pf = u \in U$, so daß $(g, f-u) = 0 \quad \forall g \in U$, wobei P orthogonale Projektion auf U ist.

(ii) Das Element u wird eindeutig charakterisiert durch das Variationsproblem

$$\|f - u\| = \min_{\varphi \in U} \|f - \varphi\|$$

(iii) H sei ein H.R., U ein Unterraum von H

$$U^{\perp} = \{g \in H | (g, f) = 0 \quad \forall f \in U\}$$

Dann ist $U^{\perp}$ auch ein Hilbertraum.

(iv) Ist T eine orthogonale Projektion auf U, dann ist $I - T$ orth. Proj. auf $U^{\perp}$ und es gilt:

$$\begin{aligned} H &= U \oplus U^{\perp} \\ x &= u + v \end{aligned}$$

Definition A.2.18 (i) Ein linearer Operator T heißt positiv definit, falls

$$(Tf, f) \geq \gamma \|f\|^2, \gamma > 0, \forall f \in H.$$

(ii) Der adjungierte Operator T^* zu T ist definiert durch

$$(Tf, g) = (f, T^* g) \qquad \forall f, g \in H.$$

(iii) Ein Operator heißt selbstadjungiert, wenn

$$(Tf, g) = (f, Tg) \qquad \forall f, g \in H.$$

Definition A.2.19 (i) Eine Menge linear unabhängiger Elemente

$$S = \{f_1, f_2, \ldots, f_n\}$$

oder

$$S = \{f_1, f_2, \ldots\}$$

heißt Basis des H.R. H, falls es $\forall f \in H$ Zahlen $\beta_1, \beta_2, \ldots \in K$ gibt, so daß

$$f = \sum \beta_i f_i$$

gilt.

(ii) Ein Hilbert-Raum heißt separabel, wenn er eine abzählbare Basis hat.

Bemerkung A.2.20 Wir werden nur separable H.R. benutzen.

Literaturverzeichnis

[And95] J.D. Anderson. *Computational Fluid Dynamics.* McGraw-Hill, New York, 1995. TLO 3108.

[Ans78] R. Ansorge. *Differenzenapproximation partieller Anfangswertaufgaben.* Teubner, Stuttgart, 1978. TLB 1313.

[Ban94] R.E. Bank. *PLTMG Users' Guide – Edition 7.0.* SIAM, Philadelphia, 1994. TLB 2424.

[Bri87] W.L. Briggs. *A Multigrid Tutorial.* SIAM, Philadelphia, 1987. TKX 4040.

[Bri95] E.O. Brigham. *FFT: Schnelle Fourier-Transformation, 6. Auflage.* Oldenbourg, München, 1995. TIR 1400.

[Cas91] J.E. Castillo, editor. *Mathematical Aspects of Numerical Grid Generation.* SIAM, Philadelphia, 1991. 40 TAY 881089.

[CH68] R. Courant and D. Hilbert. *Methoden der Mathematischen Physik I, II, 3. bzw. 2. Auflage (1. Auflage 1924/1937).* Springer, Berlin, 1968. TLL 1101.

[Cia78] Ph.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam, 1978. TIK 2875.

[Col66] L. Collatz. *The Numerical Treatment of Differential Equations, 3. Aufl.* Springer, Berlin, 1966. TLA 1122.

[Col68] L. Collatz. *Funktionalanalysis und numerische Mathematik.* Springer, Berlin, 1968. TOQ 1247.

[DER86] I.S. Duff, A.M. Erisman, and J.K. Reid. *Direct Methods for Sparse Matrices.* Oxford University Press, Oxford, 1986. TKZ 1733.

[Geo91] P. L. George. *Automatic Mesh Generation. Applications to Finite Element Methods.* Wiley, New York, 1991. TLB 2246.

[GF63] I.M. Gelfand and S.V. Fomin. *Calculus of Variations.* Prentice-Hall, Englewood Cliffs, 1963. TIQ 1324.

[GR81] C. Greenough and K. Robinson. *Finite Element Library, Rutherford Appleton Laboratory.* zu beziehen über: NAG Ltd., Oxford, 1981.

[GR92] Ch. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen.* Teubner, Stuttgart, 1992. TLB 2212.

[Gre97] A. Greenbaum. *Iterative Methods for Solving Linear Systems.* SIAM, Philadelphia, 1997. TKZ 2268.

[GVL89] G.H. Golub and Ch.F. Van Loan. *Matrix Computations, 2. Auflage.* John Hopkins Press, Baltimore, 1989. TDQ 3588.

[Hac85] W. Hackbusch. *Multi-Grid Methods and Applications.* Springer, Berlin, 1985. TLB 1729.

[Hac86] W. Hackbusch. *Theorie und Numerik elliptischer DGL.* Teubner Studienbücher, 1986. TLB 1876.

[Hac91] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* Teubner Studienbücher, Stuttgart, 1991. TKZ 1929.

[HT82] W. Hackbusch and U. Trottenberg, editors. *Multigrid Methods.* Springer, Lecture Notes in Mathematics 960, Berlin, 1982. TAY 810402.

[KS93] P. M. Knupp and S. Steinberg. *The Fundamentals of Grid Generation.* CRC Press, Boca Raton, 1993. TLB 2474.

[KS95] J.J.I.M. van Kan and A. Segal. *Numerik partieller Differentialgleichungen für Ingenieure.* Teubner, Stuttgart, 1995. TLB 2571.

[LH95] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems.* SIAM, Philadelphia, 1995. TKZ 1212.

[Mar76] D. Marsal. *Die numerische Lösung partieller Differentialgleichungen.* Springer, Mannheim, 1976. TLB 1216.

[McC87] St.F. McCormick, editor. *Multigrid Methods.* SIAM, Philadelphia, 1987. TLB 1923.

[McC89] St.F. McCormick. *Multilevel Adaptive Methods for Partial Differential Equations.* SIAM, Philadelphia, 1989. TLB 2181.

[MM78] Th. Meis and U. Marcowitz. *Numerische Behandlung partieller Differentialgleichungen.* Springer, Berlin, 1978. TLB 1389.

[MM94] K.W. Morton and D.F. Mayers. *Numerical Solution of Partial Differential Equations.* Cambridge University Press, Cambridge, 1994. TLB 2513.

[Pic86] M. Pickering. *An Introduction to Fast Fourier Transform Methods for PDE, with Applications.* Wiley, New York, 1986. TIE 1227.

[Pre75] P.M. Prenter. *Splines and Variational Methods.* Wiley, New York, 1975. TLB 1224.

[Rüd93] U. Rüde. *Mathematical and Computational Techniques for Multilevel Adaptive Methods.* SIAM, Philadelphia, 1993. TKY 1869.

[SB90] J. Stoer and R. Bulirsch. *Einführung in die Numerische Mathematik II, 3. Auflage.* Springer, Berlin, 1990. TKX 1767.

[Sch73] M.H. Schultz. *Spline Analysis.* Prentice Hall, Englewood Cliffs, 1973. TZD 1165.

[Sch90] H.R. Schwarz. *Methode der finiten Elemente, 2. Auflage.* Teubner, Stuttgart, 1990. WCG 1531.

[SF73] G. Strang and G.J. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood Cliffs, 1973. TKX 1733.

[TWM85] J.F. Thompson, Z.U.A. Warsi, and C.W. Mastin. *Numerical Grid Generation – Foundations and Applications.* North Holland, Amsterdam, 1985. TLB 1779.

[Vve93] D. Vvedensky. *Partial Differential Equations with* MATHEMATICA. Addison-Wesley, Wokingham, 1993. TIK 6146.

[Wei90] J. Weissinger. *Spärlich besetzte Gleichungssysteme. Eine Einführung mit BASIC- und PASCAL-Programmen.* BI-Wissenschaftsverlag, Mannheim, 1990. TKZ 1937.

[Wes92] P. Wesseling. *An Introduction to Multigrid Methods.* Wiley, Chichester, 1992. TLB 2220.

[Zie77] O.C. Zienkiewicz. *The Finite Element Method.* McGraw-Hill, New York, 1977. WCA 1379.