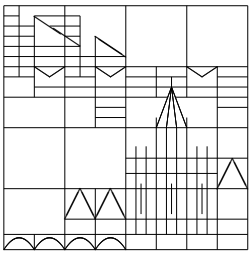
Skript zu

Numerik Partieller Differentialgleichungen II

Teil 1: Elliptische und parabolische Gleichungen

Sommersemester 2024

Johannes Schropp



Universität Konstanz

Fachbereich Mathematik und Statistik

Stand: 22. Mai 2024

Inhaltsverzeichnis

1 Finite Elemente für elliptische Differentialgleichungen 3

a) Gebräuchliche Finite Elemente 3

b) Finite Elemente für nichtlineare Probleme 14

2 Numerik parabolischer Differentialgleichungen 18

a) Finite Differenzenmodelle 18

b) Finite Elemente Methoden für parabolische Differentialgleichungen 24

c) Zeitintegratoren für gewöhnliche Differentialgleichungen 34

d) Zeitintegration für Liniensysteme parabolischer Anfangsrandwert-aufgaben 41

e) Nichtlineare parabolische Anfangsrandwertaufgaben 51

f) Finite Elemente für nichtlineare parabolische Probleme 58

1. Finite Elemente für elliptische Differentialgleichungen

a) Gebräuchliche Finite Elemente

Vorgelegt sei

$$\begin{aligned} -\Delta u &= g \text{ in } \Omega, \\ u &= \gamma \text{ auf } \partial\Omega. \end{aligned} \tag{1-1}$$

$\Omega \subset \mathbb{R}^2$ sei ein beschränktes Gebiet mit stückweise glattem Rand und strikter Kegeleigenschaft. Die Funktionen g und γ mögen stetig von Ω nach $\mathbb{R}$ abbilden.

Ist $u = u(x, y)$ eine klassische Lösung von (1-1) für $\gamma = 0$, so folgt für alle $\varphi \in C_0^\infty(\Omega)$

$$\begin{aligned} 0 = \int_\Omega -\Delta u\varphi - g\varphi \,\mathrm{d}(x,y) &= \int_\Omega \nabla u \cdot \nabla\varphi - g\varphi \,\mathrm{d}(x,y) - \int_{\partial\Omega} \frac{\partial u}{\partial n}\varphi \,\mathrm{d}S \\ &= \int_\Omega \nabla u \cdot \nabla\varphi - g\varphi \,\mathrm{d}(x,y). \end{aligned}$$

Wegen der Dichtheit von $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$ und der Stetigkeit des Funktionals $\langle \nabla u, \nabla \cdot \rangle - \langle g, \cdot \rangle$ auf $H_0^1(\Omega)$, lautet die zu (1-1) gehörige schwache Formulierung für $\gamma = 0$ somit

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}(x,y) = \int_\Omega g \cdot v \,\mathrm{d}(x,y) \quad \forall v \in V := H_0^1(\Omega) \tag{1-2}$$

bzw. mit

$$\begin{aligned} a : V \times V &\to \mathbb{R}, \\ a(u,v) &= \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}(x,y), \\ b : V &\to \mathbb{R}, \\ b(v) &= \int_\Omega g \cdot v \,\mathrm{d}(x,y) \end{aligned}$$

finden wir die sogenannte „Variationsgleichung“

$$a(u,v) = b(v) \quad \forall v \in V = H_0^1(\Omega). \tag{1-3}$$

1.1 Definition. $u \in V$ heißt eine „schwache Lösung“ von (1-1) mit $\gamma = 0$, wenn u (1-3) erfüllt.

© Johannes Schropp 22. Mai 2024

1.2 Bemerkung. Die Variationsgleichung (1-3) hat die gleichen Lösungen $u \in V$ wie die Aufgabe

$$F(v) = \frac{1}{2}a(v,v) - b(v) = \int_\Omega \frac{1}{2}|\nabla v|^2 - g \cdot v \,\mathrm{d}(x,y) \to \min_{v \in V},$$

wobei $F : V \to \mathbb{R}$. Letztere Minimierungsaufgabe heißt „Variationsproblem“.

Im allgemeinen Fall $\gamma \neq 0$ wird eine Funktion u mit $u - \gamma \in V = H_0^1(\Omega)$ gesucht. Dazu transformiert man die Aufgabe (1-1) auf homogene Randbedingungen. Ist $w = u - \gamma$ eine Lösung von

$$-\Delta w = g + \Delta\gamma \text{ in } \Omega, w = 0 \text{ auf } \partial\Omega,$$

so folgt

$$\begin{aligned} -\Delta u &= -\Delta(w+\gamma) = -\Delta w - \Delta\gamma = g, \\ u|_{\partial\Omega} &= w|_{\partial\Omega} + \gamma|_{\partial\Omega} = \gamma|_{\partial\Omega}. \end{aligned}$$

Ohne Einschränkung betrachten wir im Folgenden die Aufgabe (1-1) mit $\gamma = 0$.

Das Galerkin-Verfahren zur Aufgabe (1-1) lautet dann: Sei $V_h \subset V$ ein endlichdimensionaler Teilraum von V. Gesucht ist ein $u_h \in V_h$ mit

$$a(u_h, v) = b(v) \quad \forall v \in V_h.$$

Die Finite Elemente Methode ist dann ein Galerkin-Verfahren für einen Ansatzraum mit speziellen Eigenschaften.

Bei Ansätzen mit $V_h \subset V$ spricht man von konformen Finiten Elementen.

Sei nun $V_h \subset V$ mit $V_h = \operatorname{span}\{u_1, \ldots, u_m\}$ für gewisse Funktionen $u_i \in V$, $i = 1, \ldots, m$. Dann genügt es

$$a(u_h, v) = b(v)$$

für $v = u_i$, $i = 1, \ldots, m$, d.h. auf einer Basis von V_h, zu fordern.

Für unsere Gleichung (1-1) folgt mit $u_h = \sum_{i=1}^m c_i u_i$ sowie der Definition von a und b sofort

$$\int_\Omega \nabla \left(\sum_{j=1}^m c_j \cdot u_j \right) \cdot \nabla u_i - g \cdot u_i \,\mathrm{d}(x,y) = 0, \quad i = 1, \ldots, m$$

bzw.

$$\int_\Omega \sum_{j=1}^m c_j \nabla u_j \cdot \nabla u_i - g u_i \,\mathrm{d}(x,y) = 0, \quad i = 1, \ldots, m.$$

Wir erstellen also das lineare Gleichungssystem

$$Ac = r, \quad A \in \mathbb{R}^{m,m}, \quad c, r \in \mathbb{R}^m$$

© Johannes Schropp 22. Mai 2024

mit

$$\begin{aligned} A_{ij} &= \int_\Omega \nabla u_j \cdot \nabla u_i \,\mathrm{d}(x,y), \quad 1 \le i,j \le m, \\ r_i &= \int_\Omega g \cdot u_i \,\mathrm{d}(x,y), \quad 1 \le i \le m. \end{aligned}$$

1.3 Definition. A heißt „Steifigkeitsmatrix“ und r heißt „Ladevektor“.

Wahl der Ansatzfunktionen bei Finiten Elementen

Man unterteilt das Gebiet Ω in sogenannte Finite Elemente durch eine Triangulierung. Unsere vereinfachende Annahme sei dabei, dass Ω polygonal berandet ist, d.h. der Rand $\partial\Omega$ bestehe aus endlich vielen Geradestücken.

1.4 Definition. Eine Zerlegung $\Omega_{T_h} = \{e_1, \ldots, e_M\}$ von Ω in Dreieckelemente heißt „zulässige Triangulierung“, falls Folgendes gilt:

i) $\overline{\Omega} = \bigcup_{i=1}^{M} e_i$,

ii) Besteht $e_i \cap e_j$ aus genau einem Punkt, so ist dieser Eckpunkt sowohl von e_i als auch von e_j.

iii) Besteht $e_i \cap e_j$ für $i \neq j$ aus mehr als einem Punkt, so ist $e_i \cap e_j$ eine Kante sowohl von e_i als auch von e_j.

h ist dabei die maximale auftretende Kantenlänge.

1.5 Beispiel. Auf der nachstehenden Abbildung wird ein Beispiel einer Triangulierung gegeben.

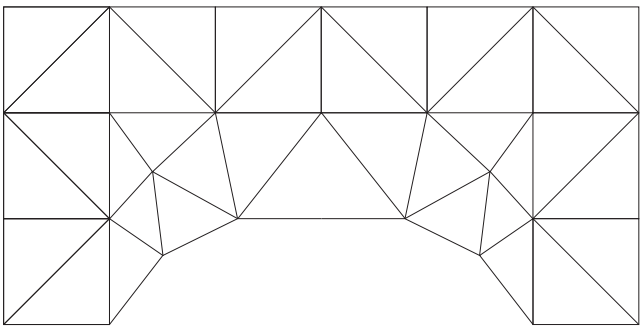


Abbildung 1: Triangulierung eines Gebietes $\Omega \subset \mathbb{R}^2$

Für die Lösung elliptischer Probleme zweiter Ordnung wählt man im Allgemeinen Finite Elemente in $H^1(\Omega)$.

© Johannes Schropp 22. Mai 2024

1.6 Satz. *Sei $k \geq 1$ und sei $\Omega \subset \mathbb{R}^2$ beschränktes Gebiet mit zulässiger Triangulierung $\Omega_{T_h} = \{e_1, \ldots, e_M\}$. Eine Funktion $v : \overline{\Omega} \to \mathbb{R}$ mit $v|_{e_i} \in C^\infty(e_i)$, $i = 1, \ldots, M$ gehört genau dann zu $H^k(\Omega)$, wenn $v \in C^{k-1}(\overline{\Omega})$ gilt.*

Beweis: Es genügt den Fall $k = 1$ zu zeigen. Für $k > 1$ folgt die Aussage sofort aus der rekursiven Anwendung auf die partiellen Ableitungen der Ordnung $k - 1$.

„⇒“ Sei $v \in C(\overline{\Omega})$. Wir setzen $w, z : \Omega \to \mathbb{R}$ stückweise, $w(x, y) = \frac{\partial}{\partial x} v(x, y)$, $z(x, y) = \frac{\partial}{\partial y} v(x, y)$, wobei auf jeder gemeinsamen Kante von zwei Dreiecken der Triangulierung einer der beiden Grenzwerte gewählt werden kann. Ferner sei $\varphi \in C_0^\infty(\Omega)$ beliebig.

Mit der Greenschen Formel folgt

$$\begin{aligned} \int_\Omega \varphi w \,\mathrm{d}(x, y) &= \sum_{j=1}^{M} \int_{e_j} \varphi \frac{\partial v}{\partial x} \,\mathrm{d}(x, y) \\ &= \sum_{j=1}^{M} \left\{ - \int_{e_j} \frac{\partial \varphi}{\partial x} v \,\mathrm{d}(x, y) + \int_{\partial e_j} \varphi v n_x \,\mathrm{d}S \right\}, \end{aligned}$$

wobei $n = (n_x, n_y)$.

Da v stetig ist, heben sich die Integrale über die inneren Kanten gegenseitig auf. Außerdem verschwindet φ auf $\partial\Omega$. Dies liefert

$$\begin{aligned} \int_\Omega \varphi w \,\mathrm{d}(x, y) &= - \sum_{j=1}^{M} \int_{e_j} \frac{\partial \varphi}{\partial x} v \,\mathrm{d}(x, y) \\ &= - \int_\Omega \frac{\partial \varphi}{\partial x} v \,\mathrm{d}(x, y). \end{aligned}$$

Analog folgt

$$\int_\Omega \varphi z \,\mathrm{d}(x, y) = - \int_\Omega \frac{\partial \varphi}{\partial y} v \,\mathrm{d}(x, y), \quad \varphi \in C_0^\infty(\Omega).$$

Dies stellt aber zusammengefasst die Definition der schwachen Differenzierbarkeit dar.

„⇐“ Sei jetzt $v \in H^1(\Omega)$. Betrachte v in der Umgebung einer Kante und drehe die Kante so um, dass sie auf der y-Achse liegt. Sie umfasse speziell das Intervall $[y_1 - \delta, y_2 + \delta]$ mit $y_1 < y_2$ und $\delta > 0$. Setze

$$\psi(x) = \int_{y_1}^{y_2} v(x, y) \, dy.$$

© Johannes Schropp 22. Mai 2024

Überdies sei nun $v \in C^\infty(\Omega)$ angenommen. Dann gilt

$$\psi(x_1) - \psi(x_2) \;=\; \int_{y_1}^{y_2} v(x_1, y) - v(x_2, y)\, dx \;=\; \int_{y_1}^{y_2} \int_{x_1}^{x_2} -\frac{\partial v}{\partial x}(x, y)\, dx\, dy,$$

und es folgt mit Cauchy-Schwarz

$$\begin{aligned} |\psi(x_1) - \psi(x_2)|^2 &= \left| \int_{y_1}^{y_2} \int_{x_1}^{x_2} -\frac{\partial v}{\partial x}(x, y)\ \mathrm{d}(x, y) \right|^2 \\ &\leq \left| \int_{y_1}^{y_2} \int_{x_1}^{x_2} \mathrm{d}(x, y) \right| \cdot \|v\|_{H^1}^2 \\ &\leq |x_1 - x_2| \cdot |y_1 - y_2| \cdot \|v\|_{H^1}^2. \end{aligned}$$

Wegen der Dichtheit von $C^\infty(\Omega)$ in $H^1(\Omega)$ gilt diese Aussage auch für $v \in H^1(\Omega)$. Also ist die Funktion $x \mapsto \psi(x)$ stetig und damit insbesondere stetig in Null.

Da y_1 und y_2 beliebig gewählt sind und der Bedingung $y_1 < y_2$ genügen, muss die stückweise stetige Funktion v auch auf der Kante stetig sein.

□

1.7 Bemerkung. Gilt für den Ansatzraum V_h die Inklusion $V_h \subset C^k(\Omega)$, $k = 0, 1, 2$, so spricht man von C^k-Finiten Elementen oder kurz C^k-Elementen.

Es sei e_μ das μ-te Element der Zerlegung $\Omega_{T_h} = \{e_1, \ldots, e_M\}$ des Grundgebietes Ω mit den Ecken $p^i = (x_i, y_i)$, $p^j = (x_j, y_j)$, $p^k = (x_k, y_k)$.

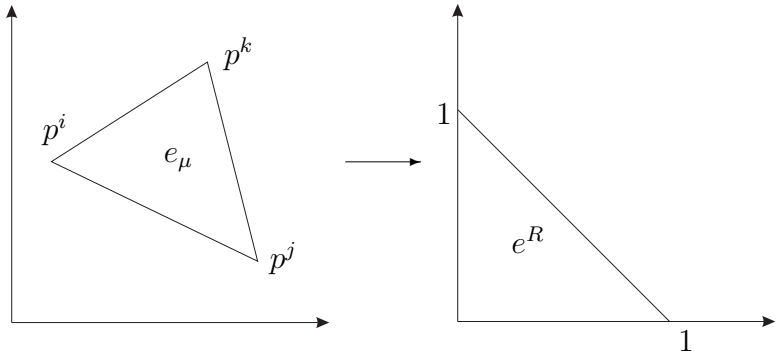


Abbildung 2: Transformation auf ein Referenzdreieck $e^R = \{(\xi, \eta) | 0 \leq \xi, \eta, \xi + \eta \leq 1\}$

Betrachte die Transformation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} x_j - x_i & x_k - x_i \\ y_j - y_i & y_k - y_i \end{pmatrix} \cdot \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} R_1(\xi, \eta) \\ R_2(\xi, \eta) \end{pmatrix} = R(\xi, \eta). \quad (1\text{-}4)$$

© Johannes Schropp 22. Mai 2024

Die Abbildung R ist affin linear und invertierbar. Offensichtlich gilt $R(0,0) = p^i$, $R(1,0) = p^j$, $R(0,1) = p^k$. Diese Transformation bildet das Referenzdreieck e^R auf das Dreieck e_μ ab.

Löst man (1-4) nach $\begin{pmatrix} \xi \\ \eta \end{pmatrix}$ auf, so hat man

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = g(x,y) = \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix}$$

mit affin linearen Funktionen g_1, g_2 in x und y.

Isoparametrisches Prinzip

Ist $p(\xi,\eta)$ eine Basisfunktion auf e^R, so ist $p(g_1(x,y), g_2(x,y)) = p(g(x,y))$ eine Basisfunktion auf e_μ.

Unser Ziel ist es, auf jedem Dreieck e_i, $i \in \{1,\dots,M\}$ geeignete glatte Funktionen so vorzugeben, dass global auf $\Omega_{T_h} = \bigcup_{i=1}^M e_i$ eine C^k-Funktion ($k = 0,1,2,\dots$) entsteht. Präziser: Man setze die Funktionen auf dem Referenzdreieck e^R mit den Eckpunkten $(0,0)$, $(0,1)$, $(1,0)$ an und benutze das ,,isoparametrische Prinzip“.

Konstruktion von C^0-Elementen mit Polynomen

Wählen den Ansatzraum

$$V_{T_h} = \{u \in C^0(\overline{\Omega}) \mid \begin{array}{l} u|_{e_i} \text{ ist ein Polynom mit} \\ \deg(u|_{e_i}) \le r, i = 1,\dots,M \text{ und } u|_{\partial\Omega} = 0\}. \end{array}$$

Sei $P_r(\Gamma) = \{u : \Gamma \to \mathbb{R} \mid u \text{ Polynom mit } \deg(u) = r\}$. Dann gilt für $\Gamma \subset \mathbb{R}^2$

$$\dim P_r(\Gamma) = \sum_{i=1}^{r+1} i = \frac{(r+1)(r+2)}{2}.$$

Somit erhalten wir 3, 6 bzw. 10 Freiheitsgrade für lineare, quadratische bzw. kubische C^0-Elemente. Man benötigt also ,,geeignet gesetzte“ $\frac{1}{2}(r+1)(r+2)$ Knoten im Referenzdreieck bei C^0-Elementen mit Polynomen vom Grade r.

1.8 Bemerkung. Sei $u : \Gamma \to \mathbb{R}$ ein Polynom mit $\deg(u) = r$. Wendet man eine affin lineare Transformation R an und drückt u in neuen Koordinaten aus, so erhält man wieder ein Polynom vom Grad r. $P_r(\Gamma)$ ist invariant unter R.

1.9 Lemma. *Es seien $\widetilde{e}$, $\hat{e}$ benachbarte Dreiecke einer Zerlegung mit einer gemeinsamen Kante K und die Polynome $\widetilde{u}$, $\hat{u}$ mit $\deg(\widetilde{u}) = \deg(\hat{u}) = r$ stimmen auf $(r+1)$-Punkten $p^1,\dots,p^{r+1} \in K$ überein, dann ist die Funktion*

$$u(x) = \begin{cases} \widetilde{u}(x), & x \in \widetilde{e}, \\ \hat{u}(x), & x \in \hat{e}\backslash K \end{cases} \tag{1-5}$$

© Johannes Schropp 22. Mai 2024

auf $\widetilde{e} \cup \hat{e}$ stetig.

Beweis: $\widetilde{u}$, $\hat{u}$ sind Polynome vom Grad r in 2 Variablen. Also sind $\widetilde{u}|_K$ und $\hat{u}|_K$ Polynome vom Grad r in einer Variablen. Betrachte $d = \widetilde{u}|_K - \hat{u}|_K$. Es folgt $d \equiv 0$, da $d(p^i) = 0$, $i = 1, \ldots, r+1$ gilt und d ein Polynom vom Grade r in einer Variablen auf K ist.

Dies motiviert eine Knotenverteilung wie in der Abbildung 3.

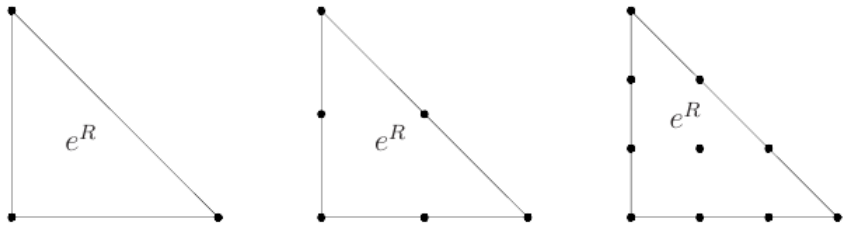


Abbildung 3: Knotierung für lineare, quadratische und kubische C^0-Elemente.

Man konstruiert nun zu diesen Knoten $p^1, \ldots, p^{N_r}$, $N_r = \frac{1}{2}(r+1)(r+2)$, eine Funktionenmenge $f_1, \ldots, f_{N_r}$ mit

$$f_i(p^j) = \delta_{ij}, \quad 1 \leq i, j \leq N_r \tag{1-6}$$

Man spricht von Lagrange-Elementen, wenn in (1-6) nur Funktionswerte vorgegeben sind. Man spricht von Hermite-Elementen, wenn statt (1-6) Funktions- und Ableitungsvorgaben gemacht werden.

$r = 1$: Es handelt sich hier um lineare Finite Elemente mit $N = 3$ und $e = e^R$ mit Knoten

$$p^1 = (1, 0), \quad p^2 = (0, 1), \quad p^3 = (0, 0)$$

und Funktionen

$$f_1(\xi, \eta) = \xi, f_2(\xi, \eta) = \eta, f_3(\xi, \eta) = 1 - \xi - \eta.$$

Dabei stellen f_1, f_2, f_3 eine nodale Basis auf e^R dar.

$r = 2$: Es ist $N = 6$. Wir sprechen dabei von quadratischen Elementen auf $e = e^R$ mit Knoten

$$\begin{aligned} p^1 &= (1, 0), \quad p^2 = (0, 1), \quad p^3 = (0, 0), \\ p^4 &= (1/2, 0), \quad p^5 = (0, 1/2), \quad p^6 = (1/2, 1/2) \end{aligned}$$

und Funktionen

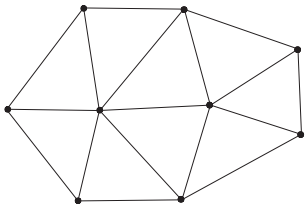
$$f_1(\xi, \eta) = \xi(2\xi - 1), \quad f_2(\xi, \eta) = \eta(2\eta - 1),$$

© Johannes Schropp 22. Mai 2024

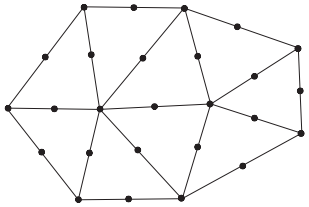
$$f_3(\xi,\eta) = (1-\xi-\eta)(1-2\xi-2\eta), \quad f_4(\xi,\eta) = 4\xi(1-\xi-\eta),$$
$$f_5(\xi,\eta) = 4\eta(1-\xi-\eta), \quad f_6(\xi,\eta) = 4\xi\eta.$$

Dabei gilt $\deg(f_i) = 2$, $i = 1,\dots,6$, $f_i(p^j) = \delta_{ij}$, $1 \le i,j \le 6$, d.h. $f_1,\dots,f_6$ ist eine nodale Basis auf e^R.

($r > 2$ entsprechend): Durch eine geeignete Festlegung der Knotenstellen ist sichergestellt, dass sich die gemäß des isoparametrischen Prinzips konstruierten Funktionen auf e_μ zu einer stetigen Funktion auf $\bigcup_{i=1}^{M} e_i = \Omega$ zusammensetzen lassen. Wie bei dem Referenzdreieck treten für $r = 1$ alle Eckpunkte der Triangulierung und für $r = 2$ alle Eckpunkte und Kantenmitten als Knoten auf.

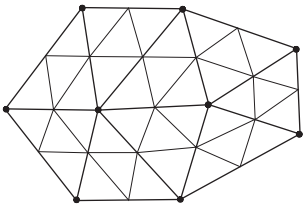


$r = 1$

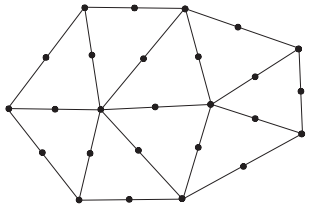


$r = 2$

Ein fairer Vergleich zwischen linearen und quadratischen Finiten Elementen benutzt aus diesem Grund die halbe Gitterweite.

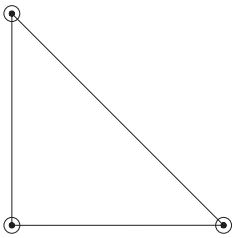


Lineare FE mit $h/2$

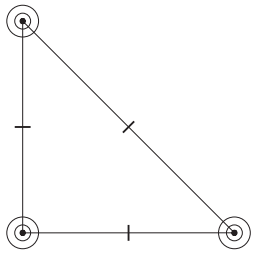


Quadratische FE mit h

Finite Elemente mit Hermite-Interpolation (wesentlich aufwändiger):



(a) Kubisches C^0-Element ($N = 9$)



(b) Quintisches C^1-Element (Argyris-Element) ($N = 21$)

Abbildung 3: Elemente mit Hermite-Interpolation

Dabei haben wir die folgende Bezeichnung verwendet:

© Johannes Schropp 22. Mai 2024

⊙ – Vorgabe der Funktionswerte und des Gradienten (d.h. 3 Vorgaben pro Punkt)

• – Vorgabe des Funktionswertes (1 Vorgabe)

◎ – Vorgabe des Funktionswertes, des Gradienten und der Hesse-Matrix (6 Vorgaben pro Punkt)

| – Vorgabe der Normalenableitung (1 Vorgabe)

□

1.10 Bemerkung (Argyris-Element)**.** Betrachtet man die Restriktion von $q \in P_5(e)$ auf eine Kante K_e von e, so sind in den Ecken die Werte bis zur zweiten Ableitung vorgegeben ($2 \times 3 = 6$ Vorgaben).

Somit ist die Hermitsche Interpolationsaufgabe in $P_5(K_e)$ eindeutig lösbar. Der Ansatz $h(t) = \sum_{i=0}^{5} \alpha_i t^i$ hat somit 6 Freiheitsgrade. Also hat man die Stetigkeit der Funktion auf K_e und die der tangentialen Ableitung. Ferner ist die Normalenableitung ein Polynom vierten Grades.

Restriktion auf die Kante

Restriktion auf die Ecken (5 Vorgaben)

Diese ist an den Ecken mitsamt der ersten Ableitungen und in den Seitenmitten gegeben. Da das zugeordnete eindimensionale Interpolationsproblem mit 5 Freiheitsgraden korrekt gestellt ist, hat man auch die Stetigkeit der Normalenableitung.

Wir skizzieren nun kurz den weiteren Weg bei C^0-Lagrange-Elementen.

Es sei $\Omega_{T_h} = \{e_1, \ldots, e_M\}$. Nach Konstruktion auf dem Referenzelement erhält man Knotenpunkte $p^1, \ldots, p^{M_f} \in \Omega_{T_h}$ und Funktionen $u_i : \Omega_{T_h} \to \mathbb{R}$ mit

- $u_i(p^j) = \delta_{ij}$, $1 \le i, j \le M_f$,
- $u_i \in C(\Omega_{T_h})$, $1 \le i \le M_f$,
- $u_i|_{e_j} \in P_r(e_j)$, $1 \le i, j \le M_f$.

mit Ansatzraum $V_{T_h} = \operatorname{span}\{u_1, \ldots, u_{M_f}\}$. Die Funktionenmenge $\{u_i\}_{i=1,\ldots,M_f}$ heißt „nodale Basis“ für V_{T_h}.

© Johannes Schropp 22. Mai 2024

1.11 Bemerkung. Gemäß Satz 1.6 mit $v \equiv u_l$ folgt sofort $u_l \in H^1(\Omega)$ für $l \in \{1, \ldots, M_f\}$, d.h. $V_{T_h} \subset V = H_0^1(\Omega)$, da $u_l|_{\partial\Omega_{T_h}} = 0$.

In unserem Fall erhält man das Gleichungssystem

$$Ac = r, \quad A \in \mathbb{R}^{M_f, M_f}, \quad c, r \in \mathbb{R}^{M_f}$$

mit

$$\begin{aligned} A_{ij} &= \int_\Omega \nabla u_j \cdot \nabla u_i \, \mathrm{d}(x,y), \quad 1 \le i, j \le M_f, \\ r_i &= \int_\Omega g \cdot u_i \, \mathrm{d}(x,y), \quad 1 \le i \le M_f. \end{aligned}$$

Mit $g_{T_h} = \sum_{j=1}^{M_f} g(p^j) u_j$ und $\Omega_{T_h} = \bigcup_{l=1}^{M} e_l$ bleiben damit die Integrale

$$\int_e \nabla u_i \cdot \nabla u_j \, \mathrm{d}(x,y), \quad \int_e u_i \cdot u_j \, \mathrm{d}(x,y), \quad 1 \le i, j \le M_f \tag{1-7}$$

auf einem Dreieck e zu bestimmen.

1.12 Beispiel ($r = 2$, $N_2 = 6$)**.** Es sind dann die Integrale

$$S_{jk}^e = \int_e \nabla u_{i_j} \cdot \nabla u_{i_k} \, \mathrm{d}(x,y), \quad 1 \le j, k \le N_r, \quad N_r = \frac{1}{2}(r+1)(r+2)$$

von Null verschieden.

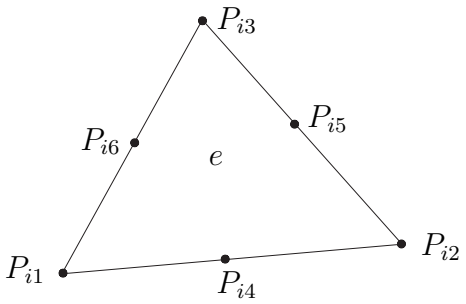


Abbildung 4: Knotennummerierung im Dreieck e

Die symmetrische Matrix $S^e = (S_{jk}^e)_{1 \le j,k \le N_r}$ ist beim Aufbau der Gesamtmatrix A auf die Untermatrix

$$\begin{pmatrix} A_{i_1,i_1} & A_{i_1,i_2} & \ldots & A_{i_1,i_{N_r}} \\ A_{i_2,i_1} & A_{i_2,i_2} & \ldots & A_{i_2,i_{N_r}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i_{N_r},i_1} & A_{i_{N_r},i_2} & \ldots & A_{i_{N_r},i_{N_r}} \end{pmatrix}$$

© Johannes Schropp 22. Mai 2024

aufzuaddieren.

Bei der Berechnung des zweiten Integrals in (1-7) tritt an die Stelle von $S^e \in \mathbb{R}^{N_r,N_r}$ die Matrix $M^e \in \mathbb{R}^{N_r,N_r}$, wobei

$$(M^e)_{jk} = \int_e u_{i_j} \cdot u_{i_k} \, \mathrm{d}(x,y), \quad 1 \le j,k \le N_r.$$

Erweiterung des Grundkonzepts auf Gebiete Ω mit glattem Rand

Sei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit glattem Rand. Bei linearen Finiten Ele-

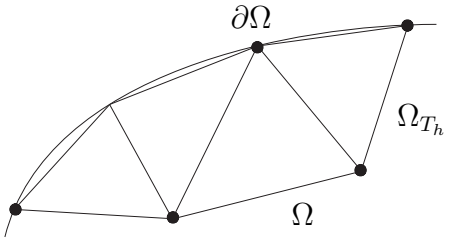


Abbildung 5: Polygonale Approximation Ω_{T_h} von Ω

menten ändert sich bei glattem Rand die Ordnung des Fehlers nicht, d.h. es gilt

$$\|u - u_h\|_{H^1} \le C \cdot h, \text{ falls } u \in H^2(\Omega) \cap H_0^1(\Omega).$$

Bei quadratischen Finiten Elementen gilt jedoch

$$\|u-u_h\|_{H^1} \le \begin{cases} C \cdot h^2, & \text{falls } u \in H^3(\Omega) \cap H_0^1(\Omega) \text{ und } \Omega \text{ polygonal berandet,} \\ C \cdot h^{3/2}, & \text{falls } u \in H^3(\Omega) \cap H_0^1(\Omega) \text{ und } \Omega \text{ beschränkt und glatt berandet.} \end{cases}$$

Die lineare Randapproximation führt also bei quadratischen Finiten Elementen zu einem Verlust in der Konvergenzordnung. Dieses Problem lässt sich aber über isoparametrische Elemente umgehen. Die Idee ist es, die Transformation R des Referenzdreiecks bei Randelementen geeignet zu modifizieren.

1.13 Beispiel (Quadratische Finite Elemente, $r = 2$, $N_2 = 6$)**.** Setze für quadratische Elemente an:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} + \begin{pmatrix} x_1 - x_3 \\ y_1 - y_3 \end{pmatrix} \xi + \begin{pmatrix} x_2 - x_3 \\ y_2 - y_3 \end{pmatrix} \eta + \begin{pmatrix} x_6 - (x_1+x_2)/2 \\ y_6 - (y_1+y_2)/2 \end{pmatrix} 4\xi\eta =: \hat{R}(\xi,\eta).$$

$\hat{R}$ ist in (ξ,η) ein Polynom vom Grade 2 mit

$$\hat{R}(0,0) = \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}, \ \hat{R}(0,1) = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ \hat{R}(1,0) = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ \hat{R}(1/2,1/2) = \begin{pmatrix} x_6 \\ y_6 \end{pmatrix}.$$

© Johannes Schropp 22. Mai 2024

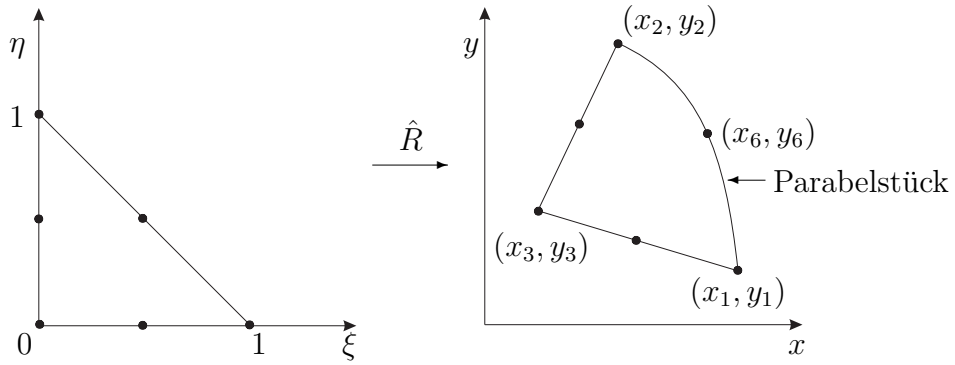


Abbildung 6: Transformation $\hat{R}$

Sind $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ und $\begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$ Randpunkte von Ω, so kann man natürlich den Rand durch eine Parabel besser approximieren als durch eine Gerade. Dies führt letzlich zu

$$\|u - u_h\|_{H^1} \leq C \cdot h^2,$$

falls $u \in H^3(\Omega) \cap H_0^1(\Omega)$ gilt und $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit glattem Rand ist, d.h. zur optimalen Ordnung quadratischer Finiter Elemente bei isoparametrischer Randapproximation.

b) Finite Elemente für nichtlineare Probleme

Vorgelegt sei das nichtlineare Poisson-Problem

$$\begin{aligned} -\Delta u &= f(u) \text{ in } \Omega, \\ u &= 0 \text{ auf } \partial\Omega \end{aligned} \tag{1-8}$$

mit $f \in C^1(\mathbb{R})$ für ein beschränktes Gebiet $\Omega \subset \mathbb{R}^2$. Ist $u = u(x, y)$ eine klassische Lösung von (1-8), so folgt für alle $\varphi \in C_0^\infty(\Omega)$

$$\begin{aligned} 0 = \int_\Omega -\Delta u\varphi - f(u)\varphi \,\mathrm{d}(x,y) &= \int_\Omega \nabla u \cdot \nabla\varphi - f(u)\varphi \,\mathrm{d}(x,y) - \int_{\partial\Omega} \frac{\partial u}{\partial n}\varphi \,\mathrm{d}S \\ &= \int_\Omega \nabla u \cdot \nabla\varphi - f(u)\varphi \,\mathrm{d}(x,y). \end{aligned}$$

Wegen der Dichtheit von $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$ und der Stetigkeit des Funktionals $\langle \nabla u, \nabla \cdot \rangle - \langle f(u), \cdot \rangle$ auf $H_0^1(\Omega)$, folgt die variationelle Formulierung

$$\int_\Omega \nabla u \cdot \nabla v - f(u)v \,\mathrm{d}(x,y) = 0, \quad \forall v \in H_0^1(\Omega).$$

© Johannes Schropp 22. Mai 2024

Das Galerkin-Verfahren hierzu lautet: Sei $V_h \subset V := H_0^1(\Omega)$. Finde ein $u_h \in V_h$ mit

$$\int_\Omega \nabla u_h \cdot \nabla v - f(u_h) v \,\mathrm{d}(x,y) = 0, \quad \forall v \in V_h.$$

Für $V_h = \operatorname{span}\{v_1, \ldots, v_M\}$, $u_h = \sum_{i=1}^M c_i v_i$ erhalten wir

$$\int_\Omega \sum_{j=1}^M c_j \nabla v_j \cdot \nabla v_i - f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}(x,y) = 0, \quad i = 1, \ldots, M,$$

was mit

$$\sum_{j=1}^M c_j \int_\Omega \nabla v_j \cdot \nabla v_i \,\mathrm{d}(x,y) = \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}(x,y), \quad i = 1, \ldots, M$$

äquivalent ist.

Mit

$$\begin{aligned} &a : V \times V \to \mathbb{R}, \quad a(v,w) = \int_\Omega \nabla v \cdot \nabla w \,\mathrm{d}(x,y), \\ &A = (a_{ij})_{1 \le i,j \le M} \in \mathbb{R}^{M,M}, \quad a_{ij} = a(v_i, v_j), \quad 1 \le i,j \le M \end{aligned}$$

und

$$\begin{aligned} &G : \mathbb{R}^M \to \mathbb{R}^M, \\ &G(c) := (G_i(c))_{1 \le i \le M}, \quad G_i(c) = \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}(x,y), \quad 1 \le i \le M \end{aligned}$$

erhalten wir das nichtlineare Problem

$$T(c) := Ac - G(c) = 0.$$

Dieses löst man mit dem Newton-Verfahren, d.h. man benötigt dazu $DT(c) = A - DG(c)$. Man findet

$$\begin{aligned} \frac{\partial G_i}{\partial c_k}(c) &= \frac{\partial}{\partial c_k} \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}(x,y) = \int_\Omega \frac{\partial}{\partial c_k} \Big(f\big(\sum_{j=1}^M c_j v_j\big) v_i \Big) \,\mathrm{d}(x,y) \\ &= \int_\Omega f'\big(\sum_{j=1}^M c_j v_j\big) \cdot v_k \cdot v_i \,\mathrm{d}(x,y), \quad 1 \le i,k \le M. \end{aligned}$$

1.14 Bemerkung. $DG(c)$ ist nun keine Diagonalmatrix wie bei den Differenzenverfahren. Eine eventuell vorhandene L_0- bzw. M-Struktur von A wird also im Allgemeinen durch $DG(c)$ zerstört.

© Johannes Schropp 22. Mai 2024

Dies ist unerwünscht und man geht deshalb in der Praxis auf andere Weise vor, indem man knotenorientierte Quadraturformeln zur Approximation des Vektorfeldes $G(c)$ benutzt.

Auf einer Triangulierung $\Omega_{T_h} = \{e_1, \ldots, e_s\}$ von Ω definiert man den Ansatzraum der Lagrangeschen Finiten Elemente

$$V_h = \{v \in C(\Omega_{T_h}) \,|\, v|_{e_k} \in P^r(e_k),\ k = 1, \ldots, s, \quad v|_{\partial\Omega_{T_h}} = 0\}.$$

Es sei ferner $V_h = \operatorname{span}\{v_1, \ldots, v_M\}$. Durch $v_1, \ldots, v_M$ und die Knoten $P_1, \ldots, P_M$ sei eine nodale Basis von V_h gegeben, d.h.

$$v_i(P_j) = \delta_{ij}, \quad 1 \le i, j \le M.$$

Wir benutzen eine Quadraturformel vom Typ

$$Q(g) = \sum_{i=1}^{M} w_i g(P_i)$$

für $g \in C(\overline{\Omega})$. Man ersetzt dann gemäß

$$\int_\Omega g\, d(x, y) = \sum_{i=1}^{M} w_i g(P_i) + R(g)$$

durch die Quadraturformel $Q(g)$ mit dem Quadraturfehler $R(g)$.

Bestimmung der Gewichte w_i, $i = 1, \ldots, M$:
Konstruiere eine derartige Formel durch

$$Q(g) := \int_\Omega I(g)\, \mathrm{d}(x, y) \text{ für } g \in C(\overline{\Omega})$$

mit dem Interpolationsoperator $I : C(\overline{\Omega}) \to V_h$, $I(g) := \sum_{j=1}^{M} g(P_j) v_j$. Man erhält dann also

$$\sum_{i=1}^{M} w_i f(P_i) = Q(f) = \int_\Omega I(f)\, \mathrm{d}(x, y) = \int_\Omega \sum_{i=1}^{M} f(P_i) v_i\, \mathrm{d}(x, y) = \sum_{i=1}^{M} f(P_i) \int_\Omega v_i\, \mathrm{d}(x, y),$$

d.h. die Gewichte lauten $w_i = \int_\Omega v_i\, \mathrm{d}(x, y)$, $i = 1, \ldots, M$. Konkret bekommt man z.B. für lineare Finite Elemente

$$w_i = \frac{1}{3} \sum_{e \in \Omega_{T_h}, P_i \in e} \mu(e) > 0, \quad i = 1, \ldots, M \text{ mit } \mu(e) = \int_e \mathrm{d}(x, y).$$

Das Vektorfeld G mit $G_i(c) = \int_\Omega f\big(\sum_{j=1}^{M} c_j v_j\big) v_i\, \mathrm{d}(x, y)$, $1 \le i \le M$ wird dann ersetzt durch

$$\sum_{k=1}^{M} w_k \left(f\big(\sum_{j=1}^{M} c_j v_j\big) v_i \right)(P_k) \overset{v_i(P_k)=\delta_{ik}}{=} \sum_{k=1}^{M} w_k f \left(\big(\sum_{j=1}^{M} c_j v_j\big)(P_k) \right) \cdot \delta_{ik}$$

© Johannes Schropp 22. Mai 2024

$$= w_i f\left(\big(\sum_{j=1}^{M} c_j v_j\big)(P_i) \right)$$
$$\overset{v_j(P_i)=\delta_{ji}}{=} w_i f\big(\sum_{j=1}^{M} c_j \delta_{ji}\big) = w_i f(c_i) =: \widetilde{G}_i(c).$$

Diese Technik nennt man Mass-Lumping.

$\widetilde{G} : \mathbb{R}^M \to \mathbb{R}^M$, $\widetilde{G}_i(c) = w_i f(c_i)$, $i = 1, \ldots, M$ ist ein Diagonalfeld. Die Linearisierung lautet

$$\frac{\partial(\widetilde{G}(c))_i}{\partial c_k} = \frac{\partial}{\partial c_k}(w_i f(c_i)) = w_i f'(c_i)\frac{\partial c_i}{\partial c_k} = w_i f'(c_i)\delta_{ik}, \quad 1 \le i, k \le M,$$

d.h. $D\widetilde{G}(c) = \operatorname{diag}(w_i f'(c_i) \,|\, i = 1, \ldots, M)$ ist eine Diagonalmatrix. Zu lösen ist also das Problem

$$\widetilde{T}(c) := Ac - \widetilde{G}(c) = 0$$

mit der Linearisierung $D\widetilde{T}(c) = A - D\widetilde{G}(c)$, welche nun mit dem Newton-Verfahren gelöst wird. Wie bei den Differenzenverfahren, bleibt eine L_0- oder M-Struktur von A erhalten.

© Johannes Schropp 22. Mai 2024

2. Numerik parabolischer Differentialgleichungen

a) Finite Differenzenmodelle

Vorgelegt sei die Anfangsrandwertaufgabe

$$\begin{aligned} u_t &= u_{xx} + f(u,x,t), \quad x \in \Omega = (0,1), \quad t \in (0,T), \\ u(x,0) &= u_0(x), \quad x \in \Omega, \\ u(0,t) &= \gamma_0(t), \quad u(1,t) = \gamma_1(t), \quad 0 \le t \le T. \end{aligned} \tag{2-1}$$

Wir diskretisieren das Problem (2-1) mit der Linienmethode, d.h. die Ableitungen in der Raumvariablen $x \in (0,1)$ werden durch entsprechende Differenzenquotienten ersetzt. Dazu wählen wir ein Ortsgitter $\Omega_{\Delta x} = \{j\Delta x \,|\, j = 0,\ldots,M\}$, $\Delta x = \frac{1}{M}$. Es sei ferner

$$v(t) = (u(\Delta x,t), u(2\Delta x,t), \ldots, u(1-\Delta x,t)) = (v_1, v_2, \ldots, v_{M-1})(t), \quad 0 \le t \le T.$$

Wir ersetzen den Ausdruck $u_{xx}(x,t) + f(u(x,t),x,t)$, $u(0,t) = \gamma_0(t)$, $u(1,t) = \gamma_1(t)$ für $t \in [0,T]$ durch

$$-C^{\Delta x} v(t) + r^{\Delta x}(t) + H^{\Delta x}(v(t))$$

mit

$$C^{\Delta x} = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix},$$

$$r^{\Delta x}(t) = \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0(t) \\ 0 \\ \vdots \\ 0 \\ \gamma_1(t) \end{pmatrix},$$

$$H^{\Delta x}(v(t)) = \begin{pmatrix} f(v_1(t), \Delta x, t) \\ f(v_j(t), j\Delta x, t), \quad j = 2,\ldots,M-2 \\ f(v_{M-1}(t), (M-1)\Delta x, t) \end{pmatrix}. \tag{2-2}$$

Präziser wird $A(t)u(\cdot,t) + f(u(\cdot,t),\cdot,t)$, $u(\cdot,t) \in D(A(t))$ durch

$$-C^{\Delta x} v(t) + r^{\Delta x}(t) + H^{\Delta x}(v(t))$$

© Johannes Schropp 22. Mai 2024

ersetzt, wobei der zeitabhängige Differentialoperator $A(t)$ mit dem Definitionsbereich

$$D(A(t)) := \{u \in C^2((0,1)) \cap C([0,1]) \,|\, u(0) = \gamma_0(t),\ u(1) = \gamma_1(t)\}$$

wie folgt definiert ist:

$$\begin{aligned} A(t) : D(A(t)) \subset C^2((0,1)) \cap C([0,1]) &\longrightarrow C((0,1)), \\ u &\longmapsto u_{xx} + f(u(\cdot,t),\cdot,t). \end{aligned}$$

Mit $v'(t) = (u_t(\Delta x, t), u_t(2\Delta x, t), \ldots, u_t(1-\Delta x, t))$ und (2-1) ergibt sich ein System gewöhnlicher Differentialgleichungen

$$\begin{aligned} v'(t) &= -C^{\Delta x} v(t) + H^{\Delta x}(v(t)) + r^{\Delta x}(t), \\ &= F_{\Delta x}(v(t), t), \quad 0 \le t \le T, \\ v(0) &= v^0 = (u_0(\Delta x), u_0(2\Delta x), \ldots, u_0(1-\Delta x)) \end{aligned} \tag{2-3}$$

als Ersatz für das Anfangsrandwertproblem (2-1).

2.1 Definition. Das Gleichungssystem (2-3) heißt Semidiskretisierung zu (2-1).

Wir lösen die Anfangswertaufgabe (2-3) mit dem ϑ-Verfahren. Sei $\Delta t = \frac{T}{N} > 0$. Der Wert v^j approximiere $v(j\Delta t)$, $j = 0, \ldots, N$. Das Verfahren lautet dann

$$\begin{aligned} v^{j+1} &= v^j + \Delta t \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j)\right], \quad j = 0, \ldots, N-1, \\ v^0 &= (u_0(\Delta x), \ldots, u_0((M-1)\Delta x)). \end{aligned}$$

Die Spezialfälle sind dabei:

$\vartheta = 0$: Euler-Cauchy Verfahren,

$\vartheta = \frac{1}{2}$: Crank-Nicholson Verfahren,

$\vartheta = 1$: implizites Euler-Cauchy Verfahren.

Fehleranalyse der Differenzenverfahren

Sei $\Gamma_h = \{(i\Delta x, j\Delta t) \,|\, i = 1, \ldots, M-1, j = 0, \ldots, N\}$, $h = (\Delta x, \Delta t)$. Wir schreiben das Differenzenverfahren in der Form

$$T^h(u) = 0, \quad T^h : \mathbb{R}^{\Gamma_h} \to \mathbb{R}^{\Gamma_h}, \quad u = (u_1^0, \ldots, u_{M-1}^0, \ldots, u_1^N, \ldots, u_{M-1}^N),$$

wobei

$$(T^h(u))_i^j = \begin{cases} u_i^0 - u_0(x_i), & j = 0,\ i = 1, \ldots, M-1 \\ \frac{1}{\Delta t}\left(u_i^j - u_i^{j-1}\right) & \\ \quad -\vartheta\left[\frac{1}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f(u_i^j, x_i, t_j)\right] & j = 1, \ldots, N,\ i = 1, \ldots, M-1 \\ -(1-\vartheta)\left[\frac{1}{\Delta x^2}\left(u_{i-1}^{j-1} - 2u_i^{j-1} + u_{i+1}^{j-1}\right) + f(u_i^{j-1}, x_i, t_{j-1})\right] & \end{cases} \tag{2-4}$$

© Johannes Schropp 22. Mai 2024

Seien $\overline{u}$ eine Lösung von (2-1) sowie $\overline{u}_h$ deren Restriktion auf das Gitter Γ_h.

2.2 Definition. $\|T^h(\overline{u}_h)\|$ heißt der Konsistenzfehler des Modells $T^h = 0$ bezüglich einer Norm $\|\cdot\|$.

Für das ϑ-Verfahren und $\|\cdot\| = \|\cdot\|_\infty$ gilt insbesondere

$$\|T^h(\overline{u}_h)\|_\infty = \begin{cases} O(\Delta t + \Delta x^2), & \vartheta \neq \frac{1}{2}, \\ O(\Delta t^2 + \Delta x^2), & \vartheta = \frac{1}{2} \end{cases}$$

an jeder Lösung $\overline{u} \in C^4([0,1] \times [0,T])$ von (2-1).

2.3 Definition. Das Modell $T^h = 0$ heißt stabil bezüglich einer Norm $\|\cdot\|$, falls es ein von h unabhängiges $C > 0$ derart gibt, dass

$$\|u - v\| \leq C \cdot \|T^h(u) - T^h(v)\|$$

für alle $u, v \in \mathbb{R}^{\Gamma_h}$ und $0 < h \leq h_0$ gilt.

Wir analysieren die Stabilität der linearen Wärmeleitungsgleichung, d.h. den Fall $f \equiv 0$ in (2-1), versehen mit den von t unabhängigen Randbedingungen $u(0,t) = \gamma_0$, $u(1,t) = \gamma_1$. Unter diesen vereinfachten Annahmen lautet das ϑ-Verfahren

$$\begin{pmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -B & A & 0 & \dots & 0 & 0 & 0 \\ 0 & -B & A & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A & 0 & 0 \\ 0 & 0 & 0 & \dots & -B & A & 0 \\ 0 & 0 & 0 & \dots & 0 & -B & A \end{pmatrix} \begin{pmatrix} v^0 \\ v^1 \\ v^2 \\ \vdots \\ v^{N-2} \\ v^{N-1} \\ v^N \end{pmatrix} - \begin{pmatrix} r^0 \\ r^1 \\ r^1 \\ \vdots \\ r^1 \\ r^1 \\ r^1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{2-5}$$

mit

$$r^0 = \begin{pmatrix} u_0(x_1) \\ u_0(x_2) \\ \vdots \\ u_0(x_{M-2}) \\ u_0(x_{M-1}) \end{pmatrix}, \quad r^1 = \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0 \\ 0 \\ \vdots \\ 0 \\ \gamma_1 \end{pmatrix},$$

$$A = \frac{1}{\Delta t} I + \vartheta C, \quad B = \frac{1}{\Delta t} I - (1-\vartheta) C,$$

$$C = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix}.$$

© Johannes Schropp 22. Mai 2024

Für die Stabilität von Blockmatrizen gilt allgemein:

2.4 Lemma. *Es sei auf $\mathbb{R}^m$ eine Norm $\|\cdot\|_*$ gegeben. Die Matrizen $A(\Delta t)$, $B(\Delta t)$ mögen von $\Delta t \in (0,T)$ abhängen. $A(\Delta t)$ sei invertierbar und erfülle*

$$\|A(\Delta t)^{-1}\|_* \le C_1 \Delta t \text{ für alle } t \in (0,T).$$

Ferner existiere ein $C_2 > 0$ mit

$$\|(A^{-1}B)^n(\Delta t)\|_* \le C_2 \text{ für alle } n \in \mathbb{N} \text{ mit } 0 \le n\Delta t \le T.$$

Dann gilt für die $(n+1)$-blockige Matrix

$$H(\Delta t) = \begin{pmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -B & A & 0 & \dots & 0 & 0 & 0 \\ 0 & -B & A & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A & 0 & 0 \\ 0 & 0 & 0 & \dots & -B & A & 0 \\ 0 & 0 & 0 & \dots & 0 & -B & A \end{pmatrix}$$

mit $0 \le n\Delta t \le T$ die Stabilitätsungleichung

$$\|v\|_{*,\infty} \le C_2(1 + C_1 T) \cdot \|H(\Delta t)v\|_{*,\infty}, \quad \forall v \in \mathbb{R}^{m(n+1)},$$

wobei

$$\|v\|_{*,\infty} = \|(v^0, v^1, \dots, v^n)\|_{*,\infty} = \max\{\|v^i\|_* \mid i = 0, \dots, n\}.$$

Daraus lässt sich die Gültigkeit des nachstehenden Lemmas folgern.

2.5 Lemma. *Unter der Bedingung*

$$\frac{\Delta t}{\Delta x^2} \le \frac{1}{2(1-\vartheta)}$$

sind die Voraussetzungen von Lemma 2.4 erfüllt und das ϑ-Verfahren ist für die Wärmeleitungsgleichung bezüglich der Norm $\|v\|_{\infty,\infty} = \|v\|_\infty$ auf $\mathbb{R}^{\Gamma_h}$ stabil.

Stabilität der Wärmeleitungsgleichung bezüglich $\|\cdot\|_{2,*}$ auf $\mathbb{R}^{\Gamma_h}$

Setzt man

$$\|v\|_* = \|v\|_2 = \left(\Delta x \sum_{i=1}^{M-1} v_i^2\right)^{1/2}, v \in \mathbb{R}^{M-1},$$

so stimmt $\|\cdot\|_2$ bis auf den Fehler der Ordnung $\sqrt{\Delta x}$ mit der euklidischen Norm überein, falls v Riemann-integrierbar ist. Im Grenzwert $\Delta x \to 0$ finden wir

$$\Delta x \sum_{i=1}^{M-1} v_i^2 \overset{\substack{\Delta x \to 0 \\ M(\Delta x) \to \infty}}{\longrightarrow} \int_0^1 v^2(x)\,\mathrm{d}x = \|v\|^2_{L^2(0,1)}$$

© Johannes Schropp 22. Mai 2024

aufgrund der Konvergenz der Riemannschen Summe gegen das Riemannsche Integral.

Ferner gilt für eine symmetrische Matrix $A \in \mathbb{R}^{M-1,M-1}$:

$$\begin{aligned} \|A\|_2 &= \sup\{\|Av\|_2 \mid v \in \mathbb{R}^{M-1}, \|v\|_2 = 1\} \\ &= \max\{|\lambda| \mid \lambda \in \sigma(A) \subset \mathbb{R}\}. \end{aligned}$$

2.6 Lemma. *Es sei*

$$C = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}, \quad \Delta x = \frac{1}{M} > 0.$$

C hat die Eigenwerte

$$\lambda_k = \frac{2}{\Delta x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right), \quad k = 1, \dots, M-1$$

mit den Eigenvektoren $v^k = (v_1^k, \dots, v_{M-1}^k) \in \mathbb{R}^{M-1}$,

$$v_i^k = \sin\left(\frac{ik\pi}{M}\right), \quad i = 1, \dots, M-1, \quad k = 1, \dots, M-1.$$

Beweis: Man rechne nach! □

Wir kehren nun zur Wärmeleitungsgleichung zurück und berechnen $\|A^{-1}(\Delta t)\|_2$ und $\|(A^{-1}B)(\Delta t)\|_2$ für

$$A(\Delta t) = \frac{1}{\Delta t} I + \vartheta C, \quad B(\Delta t) = \frac{1}{\Delta t} I - (1-\vartheta) C.$$

Gemäß Lemma 2.6 hat C die Eigenwerte $\lambda_k = \frac{2}{\Delta x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right)$, $k = 1, \dots, M-1$. Für λ_k gilt die Abschätzung

$$0 < \lambda_k < \frac{4}{\Delta x^2}, \quad k = 1, \dots, M-1.$$

Die Matrizen $A(\Delta t)$, $B(\Delta t)$, $A^{-1}(\Delta t)$ und C sind symmetrisch. Ferner gilt $(BA)(\Delta t) = (AB)(\Delta t)$ nach Definition von $A(\Delta t)$ und $B(\Delta t)$. Also folgt

$$(A^{-1}B)^T(\Delta t) = (BA^{-1})^T(\Delta t) = ((A^{-1})^T B^T)(\Delta t) = (A^{-1}B)(\Delta t),$$

© Johannes Schropp 22. Mai 2024

d.h. $(A^{-1}B)(\Delta t)$ ist symmetrisch.

$A(\Delta t)$ hat die Eigenwerte $\frac{1}{\Delta t} + \vartheta\lambda_k$. Somit sind

$$\frac{1}{\frac{1}{\Delta t} + \vartheta\lambda_k} = \frac{\Delta t}{1 + \Delta t\vartheta\lambda_k}, \quad k = 1, \dots, M-1$$

die Eigenwerte von $A^{-1}(\Delta t)$, und wir finden

$$\|A^{-1}(\Delta t)\|_2 = \max\left\{ \left| \frac{\Delta t}{1 + \Delta t\vartheta\lambda_k} \right| \mid k = 1, \dots, M-1 \right\} \leq \Delta t,$$

d.h. $C_1 = 1$ in Lemma 2.4.

Die Eigenwerte von $(A^{-1}B)(\Delta t) = \left(\frac{1}{\Delta t} I + \vartheta C\right)^{-1} \left(\frac{1}{\Delta t} I - (1-\vartheta)C\right)$ sind

$$\mu_k = \frac{\frac{1}{\Delta t} - (1-\vartheta)\lambda_k}{\frac{1}{\Delta t} + \vartheta\lambda_k} \leq 1, \quad k = 1, \dots, M-1.$$

Man beachte dazu, dass $A(\Delta t)$, $B(\Delta t)$, $A^{-1}(\Delta t)$, $(A^{-1}B)(\Delta t)$ dieselbe Basis $v^1, \dots, v^{M-1}$ aus Eigenvektoren wie C besitzen. Wir finden also

$$\|(A^{-1}B)(\Delta t)\|_2 \leq 1, \text{ falls } \mu_k \geq -1, \quad k = 1, \dots, M-1.$$

Dies ist äquivalent zu

$$\begin{aligned} \mu_k \geq -1 \quad &\Leftrightarrow \quad \frac{1}{\Delta t} - (1-\vartheta)\lambda_k \geq -\frac{1}{\Delta t} - \vartheta\lambda_k \\ &\Leftrightarrow \quad \frac{2}{\Delta t} \geq (1-2\vartheta)\lambda_k, \quad k = 1, \dots, M-1. \end{aligned}$$

Für $\vartheta \geq \frac{1}{2}$ ist dies stets erfüllt, während sich für $\vartheta < \frac{1}{2}$ die Bedingung

$$\lambda_k \leq \frac{2}{\Delta t(1-2\vartheta)}, \quad k = 1, \dots, M-1$$

ergibt. Dies ist insbesondere abgesichert, wenn

$$\lambda_k \leq \frac{4}{\Delta x^2} \leq \frac{2}{\Delta t(1-2\vartheta)} \Leftrightarrow \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\vartheta)}.$$

gilt. Das Lemma 2.4 ist also mit $C_1 = C_2 = 1$ anwendbar.

2.7 Satz. *Unter der Bedingung*

$$\frac{\Delta t}{\Delta x^2} \leq \begin{cases} \infty, & \text{für } \vartheta \geq \frac{1}{2} \\ \frac{1}{2(1-2\vartheta)}, & \text{für } \vartheta < \frac{1}{2} \end{cases}$$

© Johannes Schropp 22. Mai 2024

ist das ϑ-Verfahren für die Wärmeleitungsgleichung bezüglich der Norm

$$\|u\|_{2,\infty} = \max\left\{ \left(\Delta x \sum_{i=1}^{M-1} (u_i^j)^2 \right)^{1/2} \mid j = 0, \ldots, N \right\}, \quad u \in \mathbb{R}^{\Gamma_h},\ h = (\Delta x, \Delta t)$$

stabil. An jeder Lösung $\overline{u}$ der linearen Wärmeleitungsgleichung mit

$$\frac{\partial^\nu \overline{u}}{\partial t^\nu} \in C([0,1] \times [0,T]), \nu = 1,2, \quad \frac{\partial^\nu \overline{u}}{\partial x^\nu} \in C([0,1] \times [0,T]), \nu = 1,2,3,4$$

liegt die Konvergenz gemäß

$$\|\overline{u}_h - u^h\|_{2,\infty} = O(\Delta t + \Delta x^2), \quad T^h(u^h) = 0$$

vor. Für das Crank-Nicholson Verfahren gilt sogar

$$\|\overline{u}_h - u^h\|_{2,\infty} = O(\Delta t^2 + \Delta x^2),$$

falls zusätzlich $\frac{\partial^3 \overline{u}}{\partial t^3} \in C([0,1] \times [0,T])$.

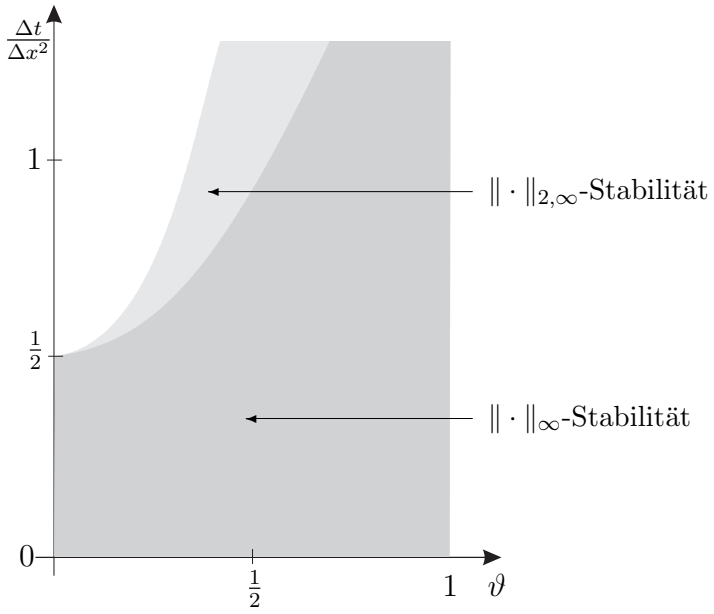


Abbildung 7: Stabilitätsbereiche

b) Finite Elemente Methoden für parabolische Differentialgleichungen

Vorgelegt sei die parabolische Anfangsrandwertaufgabe

$$\frac{\partial u}{\partial t} - \Delta u = f \text{ in } (0,T) \times \Omega, \tag{2-6}$$

© Johannes Schropp 22. Mai 2024

$$\begin{aligned} u &= 0 \text{ auf } (0,T)\times\partial\Omega, \\ u(0,\cdot) &= u_0 \text{ in } \Omega, \end{aligned}$$

wobei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet sei. Wir wenden auf (2-6) die Theorie schwacher Lösungen an.

Es sei $t \in (0,T)$ beliebig, aber fest. Ferner sei $v \in C_0^\infty(\Omega)$ willkürlich gewählt. Dann folgt

$$\int_\Omega \frac{\partial u}{\partial t}(t,x)v(x) - \Delta u(t,x)v(x)\,\mathrm{d}x = \int_\Omega f(t,x)v(x)\,\mathrm{d}x,$$

woraus sich mit der partiellen Integration

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(t,x)v(x)\,\mathrm{d}x + \int_\Omega \nabla u(t,x)\cdot\nabla v(x)\,\mathrm{d}x - \underbrace{\int_{\partial\Omega}\frac{\partial u(t,x)}{\partial n}v(x)\,\mathrm{d}S}_{=0} = \int_\Omega f(t,x)v(x)\,\mathrm{d}x$$

ergibt, wobei u eine klassische Lösung von (2-6) sei, d.h. $u \in C(\overline{\Omega}\times[0,T])$, $u_t, \Delta u \in C(\Omega\times(0,T))$. Also gilt

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(t,x)v(x)\,\mathrm{d}x + \int_\Omega \nabla u(t,x)\cdot\nabla v(x)\,\mathrm{d}x = \int_\Omega f(t,x)v(x)\,\mathrm{d}x$$

für alle $v \in C_0^\infty(\Omega)$ und $t \in (0,T)$.

Das funktionalanalytische Lösen der Aufgabe (2-6) erfordert eine besondere Behandlung der Variablen t und x:

a) Für jedes feste $t \in (0,T)$ handelt es sich bei der Abbildung

$$x \mapsto u(t,x), \text{ d.h. } u(t,\cdot)$$

um ein Element eines Sobolev-Raumes V, also $u(t,\cdot) \in V$ für $t \in (0,T)$. Hierfür schreibt man kurz $u(t) \in V$. Naheliegend ist die Wahl des Raumes $V = H_0^1(\Omega)$.

b) Variiert man nun $t \in (0,T)$, so entsteht eine Funktion $t \mapsto u(t)$ mit Werten im Banachraum V.

Für unsere weiteren Schritte benötigen wir den Begriff eines Bochner-Integrals.

Es seien (I,Σ,μ) ein Maßraum und H ein Banachraum. Das Bochner-Integral wird auf eine ähnliche Weise definiert wie ein Lebesgue-Integral.

2.8 Definition. Es sei s eine Stufenfunktion der Gestalt $s(t) = \sum_{i=1}^n \chi_{E_i}(t)h_i$, wobei $\{E_1,\ldots,E_n\} \subset \Sigma$ paarweise disjunkt sowie $\{h_1,\ldots,h_n\} \subset H$ paarweise verschieden sind und χ_E die charakteristische Funktion einer Menge $E \subset I$ bezeichnet.

© Johannes Schropp 22. Mai 2024

Das Bochner-Integral von s ist dann durch

$$\int_I s \, \mathrm{d}\mu = \sum_{i=1}^{n} \mu(E_i) h_i$$

gegeben.

2.9 Definition. Eine messbare Funktion $f : I \to H$ heißt Bochner-integrierbar, wenn es eine Folge $\{s_n\}_n$ von Stufenfunktionen gibt mit

$$\lim_{n\to\infty} \int_I \|f - s_n\|_H \, \mathrm{d}\mu = 0.$$

In diesem Fall wird das Bochner-Integral vermöge

$$\int_I f \, \mathrm{d}\mu = \lim_{n\to\infty} \int_I s_n \, \mathrm{d}\mu$$

definiert.

Nun können wir folgende funktionalanalytische Räume definieren.

2.10 Definition. Es seien (I, Σ, μ) ein Maßraum, H ein Banachraum und $1 \leq p \leq \infty$. Der Bochner-Raum $L^p(I, H)$ ist definiert als der Quotient (bezüglich der Gleichheit fast überall) des Raumes der messbaren Funktionen $u : I \to H$, deren zugehörige Norm $\|\cdot\|_{L^p(I,H)}$ endlich ist. Dabei ist

$$\begin{aligned} \|u\|_{L^p(I,H)} &:= \left(\int_I \|u(t)\|_H^p \, \mathrm{d}\mu(t) \right)^{1/p}, \quad 1 \leq p < \infty \\ \|u\|_{L^\infty(I,H)} &:= \operatorname{ess\,sup}_{t\in I} \|u(t)\|_H. \end{aligned}$$

2.11 Definition. Ist $p = 2$, so lässt sich $L^2(I, H)$ als ein Hilbertraum mit dem Skalarprodukt

$$\langle u, v \rangle_{L^2(I,H)} := \int_I \langle u(t), v(t) \rangle_H \, \mathrm{d}\mu(t)$$

auffassen.

2.12 Bemerkung. Ist $I = (0, T)$ ein offenes Intervall, Σ die Borel-σ-Algebra auf I sowie μ das dazu gehörige Lebesgue-Maß auf Σ, so gilt

$$L^p((0,T), H) = \{u : (0,T) \to H \mid \|u\|_{L^p((0,T),H)} < \infty\}.$$

© Johannes Schropp 22. Mai 2024

2.13 Definition. Sei $I \subset \mathbb{R}$ ein Intervall. Zu einem gegebenen Banachraum H wird der Raum der stetigen Funktionen definiert durch

$$C^0(I,H) = \{u : I \to H \,|\, u \text{ stetig in der } \|\cdot\|_H\text{-Norm}, \|u\|_{C^0(I,H)} < \infty\}$$

mit

$$\|g\|_{C^0(I,H)} := \sup_{t\in I} \|u(t)\|_H.$$

Es seien nun $f(t), u_0, z \in L^2(\Omega)$, $v, w \in H_0^1(\Omega)$. Wir definieren

$$\begin{aligned} \langle f(t), z\rangle_0 &:= \int_\Omega f(t,x)z(x)\,\mathrm{d}x, \\ a(v,w) &:= \int_\Omega \nabla v \cdot \nabla w\,\mathrm{d}x \end{aligned}$$

für alle $v, w \in H_0^1(\Omega)$. Damit finden wir unter Beachtung der Dichtheit der Einbettung $C_0^\infty(\Omega) \subset H_0^1(\Omega)$ sowie der Stetigkeit des Funktionals $\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), \cdot\rangle_0 + \langle \nabla u(t), \nabla\cdot\rangle_0 - \langle f(t), \cdot\rangle_0$ auf $H_0^1(\Omega)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), v\rangle_0 + a(u(t), v) = \langle f(t), v\rangle_0$$

für alle $v \in V$ und $t \in (0,T)$.

Lösungstheorie für parabolische Variationsgleichungen

Es seien $V = H_0^m(\Omega)$, $H = L^2(\Omega)$, $V' = \{f : V \to \mathbb{R} \,|\, f \text{ linear und stetig}\} = H^{-m}(\Omega)$ der Dualraum zu V. Da die Einbettungen $V \subset H \subset V'$ jeweils stetig und dicht sind, ist (V, H, V') ein Gelfand-Dreier.

2.14 Definition. Sei $u \in L^2((0,T), V)$. Dann heißt ein Element $w \in L^2((0,T), V')$ eine verallgemeinerte Ableitung, falls

$$\int_0^T \langle u(t), v\rangle_0 \varphi'(t)\,\mathrm{d}t = -\int_0^T \langle w(t), v\rangle_{V'\times V}\varphi(t)\,\mathrm{d}t$$

für alle $v \in V$ und $\varphi \in C_0^\infty((0,T))$ gilt. Man schreibt dazu wieder $w = u'$.

Somit erhalten wir

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), v\rangle_0 = \langle u'(t), v\rangle_{V'\times V}$$

für alle $v \in V$ und fast alle $t \in (0,T)$.

Dies motiviert folgende Definition.

2.15 Definition. Eine schwache Lösung der parabolischen Anfangsrandwertaufgabe (2-6) ist ein Element $u \in L^2((0,T), V)$, das eine schwache Ableitung $\frac{\mathrm{d}u}{\mathrm{d}t} = u' \in$

© Johannes Schropp 22. Mai 2024

$L^2((0,T),V')$ besitzt, und die Anfangsbedingung $u(0) = u_0$ sowie die Differentialgleichung

$$\langle u'(t), v\rangle_{V'\times V} + a(u(t), v) = \langle f(t), v\rangle_0 \quad \text{für alle } v \in V \text{ und fast alle } t \in (0,T)$$

erfüllt.

Man kann nun wieder das Skalarprodukt $\langle\cdot,\cdot\rangle_0$ auf $L^2(\Omega)$ zu einer stetigen Bilinearform $H^{-m} \times H_0^m \to \mathbb{R}$ fortsetzen und $\langle u'(t), v\rangle_0$ statt $\langle u'(t), v\rangle_{V'\times V}$ schreiben.

2.16 Bemerkung. Die Bedingungen $u \in L^2((0,T),V)$ und $u' \in L^2((0,T),V')$ implizieren $u \in C([0,T],H)$. Dieses Ergebnis ist als Interpolationssatz bekannt.

Im Folgenden sei $a : V \times V \to \mathbb{R}$ eine stetige und auf V-koerzive Bilinearform, d.h.

$$\begin{aligned} |a(u,v)| &\leq C \cdot \|u\|_V \cdot \|v\|_V, \quad \forall u, v \in V, \\ a(v,v) &\geq \alpha \|v\|_V^2, \quad \forall v \in V. \end{aligned}$$

Es gilt:

2.17 Lemma. *Es sei a eine V-elliptische, stetige Bilinearform. Sind $u_0 \in H$ und $f \in C([0,T],H)$, dann gilt für eine Lösung u des Problems*

$$\begin{aligned} \langle u'(t), v\rangle_0 + a(u(t), v) &= \langle f(t), v\rangle_0, \quad \forall v \in V, t \in (0,T), \\ u(0) &= u_0 \end{aligned} \tag{2-7}$$

die Abschätzung

$$\|u(t)\|_0 \leq \|u_0\|_0 \exp(-\alpha t) + \int_0^t \|f(s)\|_0 \exp(-\alpha(t-s))\,\mathrm{d}s$$

für alle $t \in (0,T)$.

Beweis: Wendet man (2-7) mit $v = u(t)$ an, so liefert die V-Elliptizität von a

$$\langle u'(t), u(t)\rangle_0 + \alpha\|u(t)\|_V^2 \leq \langle f(t), u(t)\rangle_0.$$

Ferner folgt mit

$$\langle u'(t), u(t)\rangle_0 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), u(t)\rangle_0 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0^2 = \|u(t)\|_0\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0$$

sofort

$$\|u(t)\|_0\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0 + \alpha\|u(t)\|_V^2 \leq \langle f(t), u(t)\rangle_0 \leq \|f(t)\|_0\|u(t)\|_0.$$

Mit $\|u(t)\|_0 \leq \|u(t)\|_V$ ergibt sich durch Division mit $\|u(t)\|_0$ dann

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0 + \alpha\|u(t)\|_0 \leq \|f(t)\|_0, \quad t \in (0,T).$$

© Johannes Schropp 22. Mai 2024

Man findet nun

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\exp(\alpha t)\cdot\|u(t)\|_0) &= \alpha\exp(\alpha t)\|u(t)\|_0+\exp(\alpha t)\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0\\ &= \exp(\alpha t)\left[\alpha\|u(t)\|_0+\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0\right]\\ &\le \exp(\alpha t)\|f(t)\|_0,\quad 0<t<T.\end{aligned}$$

Integration über $(0,t)$ liefert

$$\exp(\alpha t)\|u(t)\|_0-\|u(0)\|_0\le\int_0^t\exp(\alpha s)\|f(s)\|\,\mathrm{d}s,$$

was mit

$$\|u(t)\|_0-\exp(-\alpha t)\|u_0\|_0\le\int_0^t\|f(s)\|\exp(-\alpha(t-s))\,\mathrm{d}s$$

äquivalent ist. Wir bekommen schließlich

$$\|u(t)\|_0\le\|u_0\|_0\exp(-\alpha t)+\int_0^t\|f(s)\|_0\exp(-\alpha(t-s))\,\mathrm{d}s$$

für alle $t\in(0,T)$. □

2.18 Korollar. *Die Lösung u von (2-7) ist unter den Voraussetzungen des Lemmas 2.17 eindeutig.*

Beweis: Sind u_1,u_2 zwei Lösungen von (2-7), so löst $\hat{u}=u_1-u_2$ die Gleichung (2-7) mit $f\equiv 0$ und $u_0=0$. Lemma 2.17 liefert dann $\hat{u}=0$, also $u_1=u_2$. □

Semidiskretisierung im Raum

Wir wenden uns hier der numerischen Behandlung der folgenden variationellen Anfangsrandwertaufgabe zu: Gesucht ist ein $u\in L^2((0,T),V)$ mit $u'\in L^2((0,T),V')$ und

$$\begin{aligned}\langle u'(t),v\rangle_0+a(u(t),v) &= \langle f(t),v\rangle_0,\quad \forall v\in V, t\in(0,T),\\ u(0) &= u_0\in H.\end{aligned}\tag{2-8}$$

Wir wählen einen endlich dimensionalen Teilraum $V_h\subset V$ und bezeichnen mit u_{0h} eine Approximation von u_0 in V_h. Das Galerkin-Problem besteht darin, ein $u_h\in L^2((0,T),V_h)$ mit $u_h'\in L^2((0,T),V')$ mit

$$\begin{aligned}\langle u_h'(t),v\rangle_0+a(u_h(t),v) &= \langle f(t),v\rangle_0,\quad \forall v\in V_h,\quad t\in(0,T),\\ u_h(0) &= u_{0h}.\end{aligned}\tag{2-9}$$

zu finden. u_{0h} sei dabei als Näherung zu u_0 in V_h gegeben.

© Johannes Schropp 22. Mai 2024

Sei nun $V_h = \operatorname{span}\{v_1, \dots, v_M\}$, $u_h(t) = \sum_{i=1}^M c_i(t) v_i$ und $u_{0h} = \sum_{i=1}^M c_{i0} v_i$. Einsetzen in (2-9) liefert

$$\Big\langle \sum_{i=1}^M c_i'(t) v_i, v \Big\rangle_0 + a\Big(\sum_{i=1}^M c_i(t) v_i, v \Big) = \langle f(t), v \rangle_0, \quad \forall v \in V_h, \quad t \in (0,T).$$

Es genügt, dies auf der Basis $\{v_1, \dots, v_M\}$ zu sichern, d.h.

$$\sum_{i=1}^M c_i'(t) \langle v_i, v_j \rangle_0 + \sum_{i=1}^M c_i(t) a\big(v_i, v_j\big) = \langle f(t), v_j \rangle_0, \quad j = 1, \dots, M. \tag{2-10}$$

Das Galerkin-Verfahren (2-9) zu unserer Anfangsrandwertaufgabe (2-6) ist genau dann eindeutig lösbar, wenn es Funktionen $c_i \in C^1([0,T], \mathbb{R})$, $c_i(0) = c_{i0}$, $i = 1, \dots, M$ gibt, welche (2-10) erfüllen.

Mit der Steifigkeitsmatrix

$$A_h = (a_{ij})_{1 \le i,j \le M}, \quad a_{ij} = a(v_i, v_j)$$

und der Massenmatrix

$$B_h = (b_{ij})_{1 \le i,j \le M}, \quad b_{ij} = \langle v_i, v_j \rangle_0$$

sowie den Vektoren

$$\begin{aligned} r_h(t) &= (r_i(t))_{1 \le i \le M}, \quad r_i(t) = \langle f(t), v_i \rangle_0, \\ c_0 &= (c_{1,0}, c_{2,0}, \dots, c_{M,0}), \\ c(t) &= (c_1(t), c_2(t), \dots, c_M(t)) \end{aligned}$$

erhalten wir das System

$$\begin{aligned} B_h c'(t) + A_h c(t) &= r_h(t), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-11}$$

2.19 Definition. Das Problem (2-11) heißt semidiskrete Differentialgleichung des Anfangsrandwertproblems.

Ferner ist $B_h \in \mathbb{R}^{M,M}$ eine symmetrische, positiv definite Matrix. Die Abbildung $r_h : (0,T) \to \mathbb{R}^M$ ist stetig, da $f \in C([0,T], H)$ gilt und $\langle \cdot, \cdot \rangle_0$ stetig ist. Somit ist (2-11) äquivalent zur Anfangswertaufgabe

$$\begin{aligned} c'(t) &= B_h^{-1}(r(t) - A_h c(t)), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-12}$$

© Johannes Schropp 22. Mai 2024

Das Problem (2-12) ist eine lineare inhomogene Anfangswertaufgabe mit stetiger Inhomogenität $B_h^{-1} r_h(\cdot)$. Diese besitzt nach dem Existenz- und Eindeutigkeitssatz eine eindeutige Lösung $c : [0,T] \to \mathbb{R}^M$. Somit ist (2-12) eindeutig lösbar und damit auch (2-9).

Semidiskrete Fehlerabschätzung

Im Folgenden soll der Term $u(t) - u_h(t)$ abgeschätzt werden.

2.20 Definition. Die elliptische Projektion oder die Ritz-Projektion $R_h : V \to V_h$ ist für eine V-elliptische, stetige Bilinearform $a : V \times V \to \mathbb{R}$ durch

$$v \mapsto R_h(v) \iff (a(R_h v - v, w) = 0, \quad \forall w \in V_h)$$

definiert.

2.21 Bemerkung. Eine Ritz-Projektion besitzt die folgenden Eigenschaften:

i) R_h ist als Abbildung wohldefiniert.

ii) $R_h : V \to V_h$ ist linear und stetig,

iii) R_h liefert die quasioptimale Approximation, d.h.

$$\|v - R_h v\|_V \leq \frac{C}{\alpha} \inf_{w \in V_h} \|v - w\|_V.$$

iv) Es gilt

$$\|v - R_h v\|_0 \leq \widetilde{C} h^2 \|v\|_2, \quad \forall v \in H^2(\Omega).$$

Beweis: Siehe Übung. □

2.22 Satz. *Es sei a eine V-elliptische, stetige Bilinearform mit Konstanten C bzw. α. Ferner gelte $f \in C([0,T], H)$, $u_0 \in V$ sowie $u_{0h} \in V_h$. Dann folgt die Abschätzung*

$$\begin{aligned} \|u_h(t) - u(t)\|_0 \quad \leq \quad & \|u_{0h} - R_h u_0\|_0 \exp(-\alpha t) + \|(I - R_h)u(t)\|_0 \\ & + \int_0^t \|(I - R_h)u'(s)\|_0 \exp(-\alpha(t-s))\, \mathrm{d}s, \end{aligned}$$

falls $u \in C^1([0,T], H_0^1(\Omega))$.

Beweis: Es gilt

$$u_h(t) - u(t) = \underbrace{u_h(t) - R_h u(t)}_{=:\theta(t)} + \underbrace{R_h u(t) - u(t)}_{=:\rho(t)} = \theta(t) + \rho(t).$$

© Johannes Schropp 22. Mai 2024

Ferner sei $w \in V_h \subset V$. Dann gilt nach Definition von R_h

$$\langle u'(t), w\rangle_0 + a(u(t), w) = \langle u'(t), w\rangle_0 + a(R_h u(t), w) = \langle f(t), w\rangle_0,$$

wobei die Gültigkeit von $a(R_h v, w) = a(v, w)$ für alle $w \in V_h$ zu beachten ist. Weiter folgt

$$\langle u_h'(t), w\rangle_0 + a(u_h(t), w) = \langle f(t), w\rangle_0.$$

Subtraktion liefert

$$\begin{aligned} 0 &= \langle u_h'(t), w\rangle_0 - \langle u'(t), w\rangle_0 + a(\underbrace{u_h(t) - R_h u(t)}_{=\theta(t)}, w) \\ &= \langle \underbrace{u_h'(t) - u'(t)}_{=\theta'(t)+\rho'(t)}, w\rangle_0 + a(\theta(t), w), \end{aligned}$$

was mit

$$\langle \theta'(t), w\rangle_0 + a(\theta(t), w) = -\langle \rho'(t), w\rangle_0, \quad 0 < t < T$$

gleichbedeutend ist. Wendet man nun darauf das Lemma 2.17 an, so ergibt sich

$$\|\theta(t)\|_0 \le \|\theta(0)\|_0 \exp(-\alpha t) + \int_0^t \|\rho'(s)\|_0 \exp(-\alpha(t-s))\,\mathrm{d}s. \tag{2-13}$$

Weiter gilt $\rho(t) = (R_h - I)u(t)$, $0 < t < T$. Somit folgt

$$\rho'(t) = (R_h - I)u'(t), \quad 0 < t < T.$$

Die Abschätzung (2-13) impliziert nun mit der Dreiecksungleichung

$$\begin{aligned} \|u_h(t) - u(t)\|_0 \le\ & \|\underbrace{u_h(0)}_{=u_{0h}} - R_h \underbrace{u(0)}_{=u_0}\|_0 \exp(-\alpha t) + \|(I - R_h)u(t)\|_0 \\ & + \int_0^t \|(I - R_h)u'(s)\|_0 \exp(-\alpha(t-s))\,\mathrm{d}s. \end{aligned}$$

□

Die Abschätzung des Fehlerterms $\|u_h(t) - u(t)\|_0$ im Satz 2.22 erfolgt durch:

- den Anfangsfehler, welcher in der Zeit exponentiell abfällt und nur dann auftritt, wenn u_{0h} nicht mit $R_h u_0$ identisch ist,
- den Projektionsfehler in der Norm von H der exakten Lösung u,
- den durch die Integration über $(0, T)$ mit dem Faktor $\exp(-\alpha(t-s))$ gewichteten Projektionsfehler von $u'(t)$ in der L^2-Norm.

© Johannes Schropp 22. Mai 2024

2.23 Korollar. *Es gelten die Voraussetzungen von Satz 2.22. Gilt für die elliptische Projektion eine Fehlerabschätzung der Form*

$$\|(I - R_h)v\|_0 \leq C \cdot h^2 \|v\|_2, \quad \forall v \in H^2(\Omega),$$

so erhält man die Fehlerabschätzung

$$\begin{aligned} \|u_h(t) - u(t)\|_0 \quad \leq \quad & \|u_{0h} - R_h u_0\|_0 \exp(-\alpha t) \\ & + Ch^2 \left(\|(u(t)\|_2 + \int_0^t \|u'(s)\|_2 \exp(-\alpha(t-s))\, \mathrm{d}s \right), \\ & \text{falls } u \in L^2((0,T), H^3(\Omega) \cap H_0^1(\Omega)) \cap C^1([0,T], H_0^1(\Omega)). \end{aligned}$$

Man erhält also Konvergenz der Ordnung $O(h^2)$ bei regulären Finiten Elementen, falls $u_{0h} = R_h u_0$, $u \in L^2((0,T), H^3(\Omega) \cap H_0^1(\Omega)) \cap C^1([0,T], H_0^1(\Omega))$ und R_h die Ritz-Projektion ist.

Volldiskretes Problem

Sei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit polygonalem Rand und $\Omega_{T_h} = \{e_1, \ldots, e_m\}$ sei eine Triangulierung. Sei $V_h = \operatorname{span}\{v_1, \ldots, v_M\} \subset H_0^1(\Omega)$. Typischerweise wählt man

$$V_h = V_{T_h}^{(1)} = \{v \in C(\overline{\Omega}) \mid \quad v|_{e_k} \in P^1(e_k) \text{ und } v = 0 \text{ auf } \partial\Omega_{T_h}\} \tag{2-14}$$

Diese Wahl wird oft als lineare Lagrangesche Finite Elemente bezeichnet. Mit dem Ansatz

$$u_h(t) = \sum_{i=1}^{M} c_i(t) v_i, \quad u_{0h} = \sum_{i=1}^{M} c_{i0} v_i$$

erhält man das Differentialgleichungssystem

$$\begin{aligned} B_h c'(t) + A_h c(t) &= r_h(t), \quad 0 < t < T, \\ c(0) &= c_0 \end{aligned} \tag{2-15}$$

mit

$$\begin{aligned} A_h &= (a_{ij})_{1 \leq i,j \leq M}, \quad a_{ij} = a(v_i, v_j), \\ B_h &= (b_{ij})_{1 \leq i,j \leq M}, \quad b_{ij} = \langle v_i, v_j \rangle_0, \\ r_h(t) &= (r_i)_{1 \leq i \leq M}, \quad r_i(t) = \langle f(t), v_i \rangle_0, \\ c_0 &= (c_{i0})_{1 \leq i \leq M}. \end{aligned}$$

Im Sinne von Satz 2.22 erscheint nun die Wahl $u_{0h} = R_h u_0$ optimal mit dem Ritz-Projektor $R_h : V \to V_h$ definiert durch

$$v \mapsto R_h v \Longleftrightarrow a(R_h v - v, w) = 0, \quad \forall w \in V_h.$$

© Johannes Schropp 22. Mai 2024

Die Anfangsrandwertaufgabe (2-15) kann dann wieder mit dem ϑ-Verfahren diskretisiert werden.

Sei $\Delta t = \frac{T}{N} > 0$, und sei $c^j \in \mathbb{R}^M$ die Approximation für $c(t_j)$, $t_j = j\Delta t$ und r^j bezeichne $r(t_j)$, $j = 0, \ldots, N$. Dann gilt

$$\begin{aligned} B_h \frac{c^{j+1} - c^j}{\Delta t} &= (1-\vartheta)(-A_h c^j + r^j) + \vartheta(-A_h c^{j+1} + r^{j+1}), \quad j = 0, \ldots, N-1, \\ c^0 &= c_0 \end{aligned}$$

mit $R_h u(0) = u_{0h} = \sum_{i=1}^M c_{i0} v_i$. Insgesamt erhält man die Approximation

$$u_h^j = \sum_{k=1}^M c_k^j \cdot v_k \in V_h$$

für $u(t_j)$. Es lässt sich zeigen:

$$\max\{\|u(t_j) - u_h^j\|_0 \,|\, j = 0, \ldots, N\} \leq C(u)(h^2 + \Delta t), \quad \frac{1}{2} < \vartheta \leq 1$$

bzw.

$$\max\{\|u(t_j) - u_h^j\|_0 \,|\, j = 0, \ldots, N\} \leq C(u)(h^2 + \Delta t^2), \quad \vartheta = \frac{1}{2},$$

falls die Triangulierungen $\{\Omega_{T_h}\}_{0<h<h_0}$ regulär sind und die Lösung $u \in C^2([0,T], H^2(\Omega) \cap H_0^1(\Omega))$, $\vartheta > 1/2$ bzw. $u \in C^3([0,T], H^2(\Omega) \cap H_0^1(\Omega))$ für $\vartheta = 1/2$.

c) Zeitintegratoren für gewöhnliche Differentialgleichungen

Einschrittverfahren

Wir betrachten eine allgemeine Anfangswertaufgabe

$$\begin{aligned} u'(t) &= f(t, u(t)) \text{ für } t \in [t_0, t_e], \\ u(t_0) &= \alpha \end{aligned} \tag{2-16}$$

für $f \in C^1([t_0, t_e] \times \mathbb{R}^N, \mathbb{R}^N)$. Wir nehmen an, dass (2-16) eine Lösung $\overline{u}(t)$ für $t \in [t_0, t_e]$ besitzt.

Zur numerischen Berechnung der Lösung gehen wir nun wie folgt vor. Wir wählen eine Schrittweite $h = \frac{t_e - t_0}{\sigma(h)} > 0$ und das Gitter

$$\Omega_h = \{t_j = t_0 + jh \mid j = 0, \ldots, \sigma(h)\}.$$

Ferner bezeichne u_j die numerische Approximation für $\overline{u}(t_j)$. Ein Einschrittverfahren zu (2-16) hat die allgemeine Form

$$\begin{aligned} \frac{1}{h}(u_{m+1} - u_m) &= V(h, t_m, u_m), \quad m = 0, \ldots, \sigma(h) - 1, \\ u_0 &= \alpha. \end{aligned} \tag{2-17}$$

© Johannes Schropp 22. Mai 2024

2.24 Definition. Die Funktion $V : (0, h_0] \times [t_0, t_e) \times \mathbb{R}^N \to \mathbb{R}^N$ in (2-17) heißt die Verfahrensfunktion des Einschrittverfahrens.

Zur Analyse schreibt man wieder (2-17) in die Form $T^h(u) = 0$ um mit $T^h : (\mathbb{R}^N)^{\Omega_h} \to (\mathbb{R}^N)^{\Omega_h}$ definiert durch

$$T_h(u) := (u_0 - \alpha, h^{-1}(u_{j+1} - u_j) - V(h, t_j, u_j), j = 0, \ldots, \sigma(h) - 1)$$

und $u = (u_0, u_1, \ldots, u_{\sigma(h)}) \in (\mathbb{R}^N)^{\Omega_h}$ versehen mit der Norm

$$\|u\|_\infty = \max\{|u_i(t_j) \mid i = 1, \ldots, N,\ j = 0, \ldots, \sigma(h)\}.$$

Man kann dann wieder die Begriffe der Konsistenz, Stabilität und Konvergenz definieren.

2.25 Definition. $\overline{u}_h$ bezeichne die Restriktion der wahren Lösung $\overline{u}(t)$, $t \in [t_0, t_e]$ auf das Gitter Ω_h. Dann heißt $\|T^h(\overline{u}_h)\|_\infty$ der Konsistenzfehler.

2.26 Definition. Ein numerisches Modell $T^h(u) = 0$ wird stabil genannt, falls es eine von h unabhängige Konstante $C > 0$ derart gibt, dass

$$\|u - v\|_\infty \leq C \cdot \|T^h(u) - T^h(v)\|_\infty$$

für alle $u, v \in (\mathbb{R}^N)^{\Omega_h}$ und $0 < h < h_0$ gibt.

2.27 Definition. Ist u^h die Lösung von $T^h(u) = 0$, so bezeichnet $u^h - \overline{u}_h \in (\mathbb{R}^N)^{\Omega_h}$ den Konvergenzfehler. Die Konvergenz der Ordnung p in der Maximumsnorm verlangt dann

$$\|u^h - \overline{u}_h\|_\infty = O(h^p).$$

Hinreichend hierfür ist die Stabilität des numerischen Modells $T^h(u) = 0$ und die Konsistenz der Ordnung p gemäß $\|T^h(\overline{u}_h)\|_\infty = O(h^p)$.

Die Runge-Kutta-Verfahren

Seien $h = \frac{t_e - t_0}{\sigma(h)} > 0$, $\Omega_h = \{t_j = t_0 + jh \mid j = 0, \ldots, \sigma(h)\}$. Ein s-stufiges Runge-Kutta-Verfahren für die Anfangswertaufgabe (2-16) ist gegeben durch

$$\begin{aligned} u_0 &= \alpha, \\ u_{m+1} &= u_m + h \sum_{i=1}^{s} b_i f(t_m + c_i h, U_i^m), \quad m = 0, \ldots, \sigma(h), \end{aligned} \tag{2-18}$$

wobei sich die sogenannten Stufenwerte $U^m = (U_1^m, U_2^m, \ldots, U_s^m)$ als Lösung des Gleichungssystems

$$U_i^m = u_m + h \sum_{j=1}^{s} a_{ij} f(t_m + c_j h, U_j^m), \quad i = 1, \ldots, s \tag{2-19}$$

© Johannes Schropp 22. Mai 2024

ergeben.

Ein Runge-Kutta-Verfahren wird durch die Vorgabe eines Runge-Kutta-Tableaus definiert:

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}$$

Hierfür schreibt man kurz: $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$, $c, b \in \mathbb{R}^s$, $A \in \mathbb{R}^{s,s}$.

2.28 Definition. Das Verfahren (2-18) ist explizit, falls $a_{ij} = 0$ für $j \geq i$ gilt, sonst ist es implizit.

Im Allgemeinen ist auf jedem Zeitlevel t_m das (sN)-dimensionale System (2-19) zu lösen.

Die freien Parameter c_i, b_i, $1 \leq i \leq s$, a_{ij}, $1 \leq i, j \leq s$ werden nun dazu benutzt, um dem Verfahren wünschenswerte Eigenschaften wie z.B. eine möglichst hohe Konsistenzordnung zu verleihen.

Die einfachsten Verfahren für $s = 1, 2$ sind folgende:

(i) Euler-Cauchy-Verfahren $(s = 1)$: $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$

$$\begin{aligned} u_0 &= \alpha, \\ u_{m+1} &= u_m + h \cdot 1 \cdot f(t_m, U_1^m) \\ &= u_m + hf(t_m, u_m), \quad m = 0, \dots, \sigma(h) - 1. \end{aligned}$$

(ii) implizites Euler-Cauchy-Verfahren $(s = 1)$: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$

(iii) ϑ-Verfahren für $\vartheta \in (0, 1)$ $(s = 2)$: $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\vartheta & \vartheta \\ \hline & 1-\vartheta & \vartheta \end{array}$

Für die Durchführbarkeit lässt sich der folgende Satz zeigen:

2.29 Satz. *f genüge einer Lipschitzbedingung*

$$\|f(t, v) - f(t, w)\|_\infty \leq L_f \cdot \|v - w\|_\infty \tag{2-20}$$

für alle $v, w \in \mathbb{R}^N$ und $t \in [t_0, t_e]$. Dann besitzt das Gleichungssystem

$$U_i^m = u_m + h \sum_{j=1}^{s} a_{ij} f(t_m + c_j h, U_j^m), \quad i = 1, \dots, s$$

© Johannes Schropp 22. Mai 2024

für $(t_m, u_m) \in [t_0, t_e] \times \mathbb{R}^N$ *und jede Schrittweite* $h \in (0, t_e - t_0)$ *mit* $q := hL_f \|A\|_\infty < 1$ *genau eine Lösung* $(U_1^m, \ldots, U_j^m) = U^m = U(t, t_m, u_m) \in \mathbb{R}^{sN}$.

Für die qualitative Untersuchung eines Runge-Kutta-Verfahrens sind die sogenannten Bedingungen von Butcher oft sehr nützlich:

$$B(p): \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, p,$$
$$C(q): \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, \ldots, s,\ k = 1, \ldots, q,$$
$$D(m): \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j (1 - c_j^k), \quad j = 1, \ldots, s,\ k = 1, \ldots, m.$$

Eines der zentralen Resultate lautet:

2.30 Satz. *Genügen die Koeffizienten* b, c, A *eines* s*-stufigen Runge-Kutta-Verfahrens den vereinfachten Bedingungen von Butcher* $B(p)$, $C(q)$ *und* $D(m)$ *mit* $p \le q+m+1$, $p \le 2q+2$, *so besitzt das Verfahren die Konsistenzordnung* p, *falls die Funktion* f p*-mal stetig differenzierbar in einer Umgebung der Lösung ist.*

Bezüglich der Stabilität lässt sich zeigen:

2.31 Satz. *Genügt* f *einer Lipschitz-Bedingung der Form (2-20), so ist das Runge-Kutta-Verfahren bzgl. der Maximumsnorm stabil.*

Man erhält dann also die Konvergenz der Ordnung p bzgl. $\|\cdot\|_\infty$, d.h.

$$\max\{\|u^h(t_j) - \overline{u}_h(t_j)\|_\infty \mid j = 0, \ldots, \sigma(h)\} = O(h^p),$$

falls $f \in C^p([t_0 t_e] \times \mathbb{R}^N, \mathbb{R}^N)$ global Lipschitz-stetig ist und die Butcher-Bedingungen $B(p)$, $C(q)$ und $D(m)$ mit $p \le q + m + 1$, $p \le 2q + 2$ erfüllt.

Lineare Mehrschrittverfahren

Vorgelegt sei die Anfangswertaufgabe

$$\begin{aligned} u'(t) &= f(t, u(t)), \quad t_0 \le t \le t_e, \\ u(t_0) &= \alpha \in \mathbb{R}^N \end{aligned}$$

mit $f \in C^1([t_0, t_e] \times \mathbb{R}^N, \mathbb{R}^N)$. In Erweiterung zu Einschrittverfahren machen Mehrschrittverfahren nicht nur von einem, sondern von mehreren vorangegangenen Näherungswerten Gebrauch.

Zu einer Schrittweite $h = \frac{t_e - t_0}{\sigma(h)}$ überziehe man das Intervall $[t_0, t_e]$ mit einem Gitter $\Omega_h = \{t_j = t_0 + jh \mid j = 0, \ldots, \sigma(h)\}$. Die allgemeine Form eines k-Schritt-Verfahrens

© Johannes Schropp 22. Mai 2024

mit Schrittweite h ist

$$\frac{1}{h}(a_0u_j + a_1u_{j+1} + \cdots + a_ku_{j+k}) = \Phi(h, t_j, \ldots, t_{j+k}, u_j, \ldots, u_{j+k}) \tag{2-21}$$

mit gegebenen Koeffizienten $a_i \in \mathbb{R}$, $i = 0, \ldots, k$, $a_k \neq 0$ und einer Verfahrensfunktion $\Phi : (0, h_0] \times [t_0, t_e]^{k+1} \times \mathbb{R}^{N(k+1)} \to \mathbb{R}^N$. Hierbei steht u_i für eine Approximation von $\overline{u}(t_i)$.

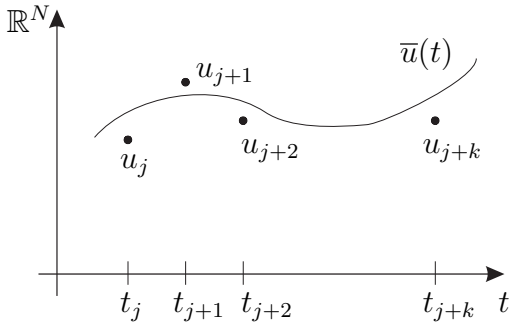


Abbildung 8: Ein Mehrschrittverfahren

2.32 Definition. Das Verfahren (2-21) heißt explizit, falls Φ nicht von u_{j+k} abhängt, sonst implizit.

Die gebräuchlisten Mehrschrittverfahren sind die linearen Mehrschrittverfahren der Gestalt

$$\begin{aligned}\Phi(h, t_j, \ldots, t_{j+k}, u_j, \ldots, u_{j+k}) &= b_0 f(t_j, u_j) + b_1 f(t_{j+1}, u_{j+1}) + \cdots + b_k f(t_{j+k}, u_{j+k}) \\ &= \sum_{i=0}^{k} b_i f(t_{j+i}, u_{j+i})\end{aligned}$$

mit Koeffizienten $b_i \in \mathbb{R}$, $i = 0, \ldots, k$. Das Verfahren lautet dann

$$\frac{1}{h}\sum_{i=0}^{k} a_i u_{j+i} = \sum_{i=0}^{k} b_i f(t_{j+i}, u_{j+i}), \quad j = 0, \ldots, \sigma(h) - k \tag{2-22}$$

zu gegebenen Startwerten $u_0, \ldots, u_{k-1}$. Die Parameter des Verfahrens lassen sich in Form einer Tabelle angeben:

a_0	a_1	$\ldots$	a_k
b_0	b_1	$\ldots$	b_k

Letztere wird als ein Mehrschrittverfahren-Tableau bezeichnet. Für $b_k = 0$ ist das Verfahren explizit und für $b_k \neq 0$ implizit.

© Johannes Schropp 22. Mai 2024

Wir können das lineare Mehrschrittverfahren (2-22) auf die Form $T^h(u) = 0$, $T^h : (\mathbb{R}^N)^{\Omega_h} \to (\mathbb{R}^N)^{\Omega_h}$, $u \in (\mathbb{R}^N)^{\Omega_h}$ bringen mit

$$T^h(u) = \Bigg(u_0 - \gamma_{0,h}, \ldots, u_{k-1} - \gamma_{k-1,h}, h^{-1} \sum_{i=0}^{k} a_i u_{j+i} - \sum_{i=0}^{k} b_i f(t_{j+i}, u_{j+i}), \\ j = 0, \ldots, \sigma(h) - k \Bigg). \quad (2\text{-}23)$$

Damit sind die Begriffe der Konsistenz, Stabilität und Konvergenz direkt übertragbar.

Für ein Mehrschrittverfahren mit k-Schritten benötigt man eine Anfangsrechnung zur Bestimmung der Approximationen $\gamma_{0,h}, \ldots, \gamma_{k-1,h}$ für $\overline{u}(t_0), \ldots, \overline{u}(t_{k-1})$. Dies kann z.B. durch $(k-1)$ Schritte eines Einschrittverfahrens geschehen.

Bei impliziten linearen Mehrschrittverfahren (d.h. im Falle $b_k \neq 0$) ist in jedem Zeitschritt das Gleichungssystem

$$u_{j+k} = \frac{hb_k}{a_k} f(t_{j+k}, u_{j+k}) + \frac{1}{a_k} \left[\sum_{i=0}^{k-1} hb_i f(t_{j+i}, u_{j+i}) - a_i u_{j+i} \right]$$

nach u_{j+k} aufzulösen.

Ein Fixpunktargument sichert die Auflösbarkeit unter der Bedingung

$$q := h \frac{|b_k|}{|a_k|} L_f < 1,$$

falls

$$\|f(t, v) - f(t, w)\|_\infty \leq L_f \|v - w\|_\infty$$

für alle $v, w \in \mathbb{R}^N$ und $t \in [t_0, t_e]$ gilt.

Man erhält ein Mehrschrittverfahren der Konsistenzordnung p, d.h. $\|T^h(\overline{u}_h)\|_\infty = O(h^p)$, falls für die Koeffizienten des linearen k-Schritt-Verfahrens

$$\sum_{i=0}^{k} a_i = 0, \quad (i^0 = 1,\ 0! = 1) \quad (2\text{-}24)$$

$$\sum_{i=0}^{k} a_i \frac{i^l}{l!} - b_i \frac{i^{l-1}}{(l-1)!} = 0 \text{ für } l = 1, \ldots, p$$

gilt und f p-mal stetig differenzierbar in einer Umgebung der Lösung ist, sowie $\|\gamma_{j,h} - \overline{u}(t_j)\|_\infty = O(h^p)$ für $j = 0, \ldots, k-1$ gilt. Ferner normiert man die Koeffizienten gemäß

$$\sum_{i=0}^{k} b_i = 1. \quad (2\text{-}25)$$

© Johannes Schropp 22. Mai 2024

Für die praktische Anwendung geben wir hier einige Formelsätze für Mehrschrittverfahren.

(i) Adams-Bashforth-Verfahren:
$a_i = 0$, $i = 0, \dots, k-2$, $b_k = 0$,
a_{k-1}, a_k, b_i, $i = 0, \dots, k-1$ bestimmt aus (2-24)— (2-25)

Das Verfahren ist explizit der Ordnung $p = k$.

Beispiel für $k = 2$: $u_{j+2} - u_{j+1} = h(3/2\, f(t_{j+1}, u_{j+1}) - 1/2\, f(t_j, u_j))$

(ii) BDF-Verfahren (BDF steht für „backward differentiation formulae“):
$b_0 = b_1 = \cdots = b_{k-1} = 0$,
$a_0, \dots, a_k, b_k$ bestimmt aus (2-24)— (2-25)

Das Verfahren ist implizit der Ordnung $p = k$

Beispiel für $k = 3$: $11/6\, u_{j+3} - 3\, u_{j+2} + 3/2\, u_{j+1} - 1/3\, u_j = h f(t_{j+3}, u_{j+3})$

Lineare Mehrschrittverfahren sind nicht mehr uneingeschränkt stabil. Zu einem linearen Mehrschrittverfahren (2-22) heißt

$$p(z) = \sum_{i=0}^{k} a_i z^i$$

das charakteristische Polynom des Verfahrens. Es gilt $\deg(p) = k$, da $a_k \neq 0$.

2.33 Satz (Dahlquist). *Das numerische Modell $T^h(u) = 0$ eines linearen Mehrschrittverfahrens mit T^h aus (2-23) erfüllt eine Stabilitätsungleichung*

$$\|u - v\|_\infty \leq C \cdot \|T^h(u) - T^h(v)\|_\infty, \quad u, v \in (\mathbb{R}^N)^{\Omega_h},\ 0 < h \leq h_0,$$

falls f einer globalen Lipschitz-Bedingung genügt, und die folgende Wurzelbedingung gilt:

Für jede Nullstelle $z \in \mathbb{C}$ des charakteristischen Polynoms gilt entweder $|z| < 1$ oder $|z| = 1$ und z ist eine einfache Wurzel. (2-26)

2.34 Definition. Ein Verfahren, dass die Bedingung (2-26) erfüllt, heißt nullstabil.

Als Konsequenz haben wir:

2.35 Korollar. *Ein lineares Mehrschrittverfahren, dessen charakteristisches Polynom die Wurzelbedingung (2-26) erfüllt, ist konvergent der Ordnung p, falls Konsistenz der Ordnung p vorliegt und f einer globalen Lipschitz-Bedingung genügt.*

© Johannes Schropp 22. Mai 2024

$z = 1$ ist immer eine Nullstelle von p, denn $p(1) = \sum_{i=0}^{k} a_i = 0$. Die $(k-1)$ anderen Nullstellen von p in $\mathbb{C}$ entscheiden also über die Stabilität.

2.36 Beispiel. a) Das charakteristische Polynom des Adams-Bashforth-Verfahrens $p(z) = z^k - z^{k-1} = z^{k-1}(z-1)$ erfüllt die Wurzelbedingung.

b) Die BDF-Verfahren erfüllen die Wurzelbedingung für $k = 1, 2, \ldots, 6$. Ab der Ordnung 7 sind die BDF-Verfahren instabil. Deshalb werden sie nur für $k = 1, 2, \ldots, 6$ benutzt.

c) Einschrittverfahren haben das charakteristische Polynom $p(z) = z - 1$, welches die Wurzelbedingung erfüllt.

d) Zeitintegration für Liniensysteme parabolischer Anfangsrandwertaufgaben

Betrachte

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta u &= 0 \text{ in } \Omega \times (0,T), \\ u &= 0 \text{ auf } \partial\Omega \times (0,T), \\ u(x,0) &= u_0(x) \text{ für } x \in \Omega, \end{aligned} \tag{2-27}$$

wobei $\Omega \subset \mathbb{R}^d$, $d = 1, 2$ ein beschränktes Gebiet ist.

Das mit klassischen Differenzenverfahren hergeleitete Liniensystem zu (2-27) erhält die Form

$$w' = \Gamma w, \quad w(0) = w^0.$$

i) Finite Differenzen: Hierbei bekommt man für $\Omega = (0,1)$

$$\Gamma = -\frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} = \Gamma_{\Delta x} \in \mathbb{R}^{M-1,M-1}, \quad \Delta x = \frac{1}{M},$$

$$w^0 = (u_0(\Delta x), \ldots, u_0((M-1)\Delta x))$$

mit den Eigenwerten

$$\lambda_k = -\frac{2}{\Delta x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right), \quad k = 1, \ldots, M-1.$$

© Johannes Schropp 22. Mai 2024

Im Fall $\Omega = (0,1)^2$ ergibt sich bei zeilenweiser Nummerierung der Gitterpunkte von links unten nach rechts oben

$$\Gamma = -\frac{1}{\Delta x^2}\begin{pmatrix} B & -C & 0 & \dots & 0 & 0 & 0 \\ -C & B & -C & \dots & 0 & 0 & 0 \\ 0 & -C & B & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & B & -C & 0 \\ 0 & 0 & 0 & \dots & -C & B & -C \\ 0 & 0 & 0 & \dots & 0 & -C & B \end{pmatrix} = \Gamma_{\Delta x} \in \mathbb{R}^{(M-1)^2,(M-1)^2},$$

wobei

$$\begin{aligned} B &= \begin{pmatrix} 4 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 4 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 4 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 4 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 4 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}, \\ C &= I_{M-1} \in \mathbb{R}^{M-1,M-1}, \quad w^0 = (u_0(i\Delta x, j\Delta x))_{1\le i,j\le M-1} \end{aligned}$$

mit den Eigenwerten

$$\lambda_{k,l} = -\frac{2}{\Delta x^2}\left(2 - \cos\left(\frac{k\pi}{M}\right) - \cos\left(\frac{l\pi}{M}\right)\right), \quad 1 \le k,l \le M-1.$$

ii) Finite Elemente: Für $\Omega \subset \mathbb{R}^2$ erhält man das System

$$\begin{aligned} Bw'(t) + Aw(t) &= 0, \quad 0 \le t \le T, \\ w(0) &= w^0 \end{aligned}$$

mit

$$A = (a_{ij})_{1\le i,j\le M} \in \mathbb{R}^{M,M}, \quad a_{ij} = a(v_i, v_j) = \int_\Omega \nabla v_i \cdot \nabla v_j \,\mathrm{d}x, \quad 1 \le i,j \le M,$$

$$B = (b_{ij})_{1\le i,j\le M}, \quad b_{ij} = \langle v_i, v_j\rangle_0 = \int_\Omega v_i v_j \,\mathrm{d}x, 1 \le i,j \le M,$$

$$R_h u_0 = \sum_{i=1}^{M} (w^0)_i v_i,$$

© Johannes Schropp 22. Mai 2024

wobei $R_h : V \to V_h := \text{span}\{v_1, \ldots, v_M\}$ der Ritz-Projektor sei. Ferner sind A, B symmetrisch und positiv definit. Wegen der Symmetrie und der positiven Definitheit von $B \in \mathbb{R}^{M-1,M-1}$ ist dies äquivalent zu

$$\begin{aligned} w' &= \Gamma w, \\ w(0) &= w^0 \end{aligned}$$

mit $\Gamma := -B^{-1}A$.

2.37 Bemerkung. i) Da mit B auch B^{-1} symmetrisch und positiv definit ist, existiert $B^{-1/2}$, und es gilt

$$\Gamma = -B^{-1}A = B^{-1/2}(B^{-1/2}(-A)(B^{-1/2})^T)B^{1/2},$$

d.h. $\sigma(\Gamma) = \sigma(B^{-1/2}(-A)(B^{-1/2})^T)$. Nach dem Sylvesterschen Trägheitssatz haben nun $(-A)$ und $B^{-1/2}(-A)(B^{-1/2})^T$ dieselbe Anzahl an positiven und negativen Eigenwerten. Somit sind $(-A)$ und $B^{-1/2}(-A)(B^{-1/2})^T$ negativ definit. Folglich hat Γ reelle, negative Eigenwerte.

ii) Im allgemeinen Fall, d.h. bei allgemeineren Randbedingungen (z.B. Neumann oder gemischte Randbedingungen) oder bei allgemeineren Gebieten Ω erhält man für die Liniensysteme, welche durch Semidiskretisierung im Raum durch Finite Differenzen oder Finite Elemente entstehen, wieder eine Anfangswertaufgabe der Form $w' = \Gamma w$, $w(0) = w^0$, i.a. aber mit einer nicht symmetrischen Matrix Γ, welche lediglich $Re(\lambda) < 0$ für $\lambda \in \sigma(\Gamma)$ erfüllt.

Wir betrachten nun ein beliebiges M-dimensionales System

$$w' = \Gamma w, \quad w(0) = w^0 \tag{2-28}$$

mit einer über $\mathbb{C}$ diagonalisierbaren Matrix $\Gamma \in \mathbb{R}^{M,M}$. Es existieren also linear unabhängige Vektoren $y^k \in \mathbb{C}^M$ derart, dass

$$\begin{aligned} \Gamma y^k &= \lambda_k y^k, \quad k = 1, \ldots, M, \\ Y^{-1}\Gamma Y &= \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_M), \quad Y = (y^1, \ldots, y^M). \end{aligned} \tag{2-29}$$

Die Lösung von (2-28) ergibt sich dann bekanntlich in der Form

$$w(t) = \sum_{k=1}^{M} c_k \exp(\lambda_k t) y^k, \quad t \in \mathbb{R}.$$

Dabei gilt $w(0) = w_0 = \sum_{k=1}^{M} c_k y^k$.

Wir analysieren nun Einschrittverfahren, welche angewandt mit der Zeitschrittweite Δt auf (2-28) eine Rekursion

$$w^m = g(\Delta t \Gamma) w^{m-1}, \quad m = 1, \ldots, N, \quad N\Delta t = T \tag{2-30}$$

mit einer rationalen Funktion $g(z) = \frac{p(z)}{q(z)}$, $z \in D \subset \mathbb{C}$ offen, für polynomiale p und q liefern.

© Johannes Schropp 22. Mai 2024

2.38 Definition. Zu einem Polynom $p : D \to \mathbb{C}$, $p(z) = \sum_{l=0}^{k} \alpha_l z^l$ und einer Matrix $B \in \mathbb{R}^{M,M}$ setze

$$p(B) := \sum_{l=0}^{k} \alpha_l B^l, \quad B^0 = I.$$

Ferner setzt man zu einer rationalen Funktion $g(z) = \frac{p(z)}{q(z)}$, $z \in D \subset \mathbb{C}$ offen, $q(z) \neq 0$:

$$g(B) := (q(B))^{-1} p(B),$$

falls $q(B)$ invertierbar ist.

2.39 Bemerkung. $q(B)$ hat die Eigenwerte $q(\lambda)$, $\lambda \in \sigma(B)$. Folglich ist $q(B)$ invertierbar, falls $\sigma(B) \cap \{\hat{z} \in D \,|\, q(\hat{z}) = 0\} = \varnothing$.

Eine Darstellung der Form (2-30) gilt allgemein für Runge-Kutta-Verfahren. Betrachte zunächst den Fall $M = 1$, d.h. die Anfangswertaufgabe

$$w' = \lambda w, \quad w(0) = w^0.$$

Sei $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ das Tableau eines s-stufigen Runge-Kutta-Verfahrens. Man findet

$$w^{m+1} = w^m + \Delta t \sum_{i=1}^{s} b_i F(t_m + c_i \Delta t, W_i^m) = w^m + \Delta t \sum_{i=1}^{s} \lambda b_i W_i^m = w^m + \Delta t \, \lambda b^T W^m$$

mit $W^m = (W_1^m, \ldots, W_s^m)$ gegeben als die Lösung von

$$\begin{aligned} W_i^m &= w^m + \Delta t \sum_{j=1}^{s} a_{ij} F(t_m + c_j \Delta t, W_j^m) \\ &= w^m + \Delta t \sum_{j=1}^{s} a_{ij} \lambda W_j^m, \quad i = 1, \ldots, s. \end{aligned}$$

Dies ist äquivalent zu

$$\begin{pmatrix} 1 - a_{11}\Delta t\lambda & -a_{12}\Delta t\lambda & \ldots & -a_{1s}\Delta t\lambda \\ -a_{21}\Delta t & 1 - a_{22}\Delta t\lambda & \ldots & a_{2s}\Delta t\lambda \\ \vdots & \vdots & \ddots & \vdots \\ -a_{s1}\Delta t\lambda & -a_{s2}\Delta t\lambda & \ldots & 1 - a_{ss}\Delta t\lambda \end{pmatrix} \begin{pmatrix} W_1^m \\ W_2^m \\ \vdots \\ W_s^m \end{pmatrix} = \begin{pmatrix} w^m \\ w^m \\ \vdots \\ w^m \end{pmatrix},$$

d.h.

$$(I - \Delta t \lambda A) W^m = \mathbb{I} w^m$$

mit $\mathbb{I} = (1, 1, \ldots, 1)^T \in \mathbb{R}^s$. Ist die Matrix $I - \Delta t \lambda A$ invertierbar, so finden wir

$$W^m = (I - \Delta t \lambda A)^{-1} \mathbb{I} w^m$$

© Johannes Schropp 22. Mai 2024

und daher

$$\begin{aligned} w^{m+1} &= w^m + \Delta t\lambda \cdot b^T(I - \Delta t\lambda A)^{-1}\mathbb{I}w^m \\ &= (1 + \Delta t\lambda b^T(I - \Delta t\lambda A)^{-1}\mathbb{I})w^m =: R(\Delta t\lambda)w^m \end{aligned}$$

mit

$$R(z) := 1 + zb^T(I - zA)^{-1}\mathbb{I} = g(z). \tag{2-31}$$

2.40 Definition. Die Funktion $R(z)$ in (2-31) heißt Stabilitätsfunktion des Runge-Kutta-Verfahrens.

Des Weiteren sichert die Cramersche Regel, dass $W_i^m = W_i^m(\Delta t\lambda)$, $i = 1, \ldots, s$ rationale Funktionen in $\lambda\Delta t$ sind, falls $I - \Delta t\lambda A$ invertierbar ist. Der Zählergrad und der Nennergrad überschreiten dabei nicht s. Somit hat die Funktion g aus (2-31) die Darstellung

$$g(z) = R(z) = \frac{p(z)}{q(z)},$$

wobei p und q Polynome mit $\deg(p) \leq s$ und $\deg(q) \leq s$ sind.

Wir kehren jetzt zum allgemeinen Fall (2-28) zurück. Wir nutzen die Diagonalisierbarkeit von Γ aus und setzen

$$u = Y^{-1}w$$

mit $Y = (y^1, \ldots, y^M) \in \mathbb{C}^{M,M}$ aus (2-29). Dann gilt

$$u' = Y^{-1}w' = Y^{-1}\Gamma w = Y^{-1}\Gamma Y u = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)u.$$

Einsetzen liefert dann mit kurzer Rechnung die Iteration

$$w^m = g(\Delta t\Gamma)w^{m-1}, \quad m = 1, \ldots, N. \tag{2-32}$$

Mit (2-32) und $Y^{-1}\Gamma Y = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)$ finden wir

$$\begin{aligned} w^m &= g(\Delta t\Gamma)w^{m-1} = g(\Delta t\Gamma)^m w^0 = g(\Delta t\Gamma)^m \underbrace{\left(\sum_{k=1}^{M} c_k y^k\right)}_{=w^0} \\ &= \sum_{k=1}^{M} c_k g(\Delta t\lambda_k)^m y^k, \quad m = 1, \ldots, N. \end{aligned}$$

Man beachte dabei, dass $g(\Delta t\Gamma)$ die Eigenwerte $g(\Delta t\lambda_1), \ldots, g(\Delta t\lambda_M)$ mit den Eigenvektoren $y^1, \ldots, y^M$ hat.

Vergleich der kontinuerlichen und der diskreten Lösung

© Johannes Schropp 22. Mai 2024

Ist $\overline{w}$ die wahre Lösung von (2-28), so folgt

$$\begin{aligned}\overline{w}(m\Delta t) - w^m &= \sum_{k=1}^{M} c_k \underbrace{\exp(m\Delta t\lambda_k)}_{=\exp(\Delta t\lambda_k)^m} y^k - \sum_{k=1}^{M} c_k g(\Delta t\lambda_k)^m y^k \\ &= \sum_{k=1}^{M} (\exp(\Delta t\lambda_k)^m - g(\Delta t\lambda_k)^m) c_k y^k.\end{aligned}$$

Zu vergleichen sind die sogenannten Amplikationsfaktoren $\exp(\lambda_k \Delta t)$ für die kontinuerliche Aufgabe und $g(\Delta t\lambda_k)$ für die diskrete Aufgabe.

Seien nun die Eigenwerte von $\Gamma \in \mathbb{R}^{M,M}$ gemäß

$$\mathrm{Re}(\lambda_M) \leq \ldots \mathrm{Re}(\lambda_2) \leq \mathrm{Re}(\lambda_1)$$

angeordnet. Damit $g(\Delta t\lambda_k)$ und $\exp(\Delta t\lambda_k)$ wenigstens qualitativ übereinstimmen, ist die Gültigkeit von

$$|g(\Delta t\lambda_k)| \begin{cases} < 1, & \text{falls } \mathrm{Re}(\lambda_k) < 0, \\ > 1, & \text{falls } \mathrm{Re}(\lambda_k) > 0 \end{cases}, \quad k = 1, \ldots, M \tag{2-33}$$

zu fordern. Der für die Anwendungen interessante Fall ist nun

$$\mathrm{Re}(\lambda_M) \leq \ldots \mathrm{Re}(\lambda_2) \leq \mathrm{Re}(\lambda_1) \leq 0. \tag{2-34}$$

Man vergleiche dazu beispielsweise Liniensysteme für parabolische Differentialgleichungen.

2.41 Definition. Sei i_0 der kleinste Index mit $Re(\lambda_{i_0}) < 0$. Das Verhältnis $\sigma = \frac{\mathrm{Re}(\lambda_M)}{\mathrm{Re}(\lambda_{i_0})}$ heißt im diesem Fall die Steifheit von Γ. Man sagt, $w'(t) = \Gamma w(t)$ ist eine steife Differentialgleichung, falls $\sigma = \sigma(\Gamma) \gg 1$.

2.42 Beispiel. Vorgelegt sei die Matrix

$$\Gamma = -\frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}$$

und es sei $M_0 \in \{1, \ldots, M-1\}$ fest. Dann folgt für die ersten M_0 Eigenwerte im Grenzwert $\Delta x \to 0$

$$\lambda_k = -\frac{2}{\Delta x^2}(1 - \cos(k\pi\Delta x)) = -\frac{2}{\Delta x^2}\left(1 - \left[1 - \frac{1}{2}k^2\pi^2\Delta x^2 + O(\Delta x^4)\right]\right)$$

© Johannes Schropp 22. Mai 2024

$$= \quad -k^2\pi^2 + O(\Delta x^2), \quad k = 1, \ldots, M_0.$$

Für die Steifheit der Differentialgleichung findet man

$$\begin{aligned} \sigma(\Gamma) &= \frac{\lambda_{M-1}}{\lambda_1} = \frac{1 - \cos((M-1)\pi\Delta x)}{1 - \cos(\pi\Delta x)} \\ &= \frac{2 + O(\Delta x^2)}{(1/2)\pi^2\Delta x^2 + O(\Delta x^4)} = O\left(\frac{1}{\Delta x^2}\right), \end{aligned}$$

d.h. Liniensysteme parabolischer Differentialgleichungen sind für $\Delta x \ll 1$ steife Systeme.

Im Falle von (2-34) erzwingt (2-33) die Bedingung

$$|g(\Delta t\lambda_k)| \leq 1, \quad k = 1, \ldots, M.$$

Dies motiviert:

2.43 Definition. Die Menge $S := \{z \in \mathbb{C} \mid |g(z)| \leq 1\}$ heißt der Bereich der absoluten Stabilität des Einschrittverfahrens mit der Funktion g.

In der Situation von (2-34) ist es wünschenswert, Verfahren mit einem möglichst großen Bereich der absoluten Stabilität zu haben, damit die Forderung $|g(\Delta t\lambda_k)| \leq 1$, $k = 1, \ldots, M$ nicht zu kleine Schrittweiten erzwingt. Im Idealfall soll gelten

$$\mathbb{C}_- := \{z \in \mathbb{C} \mid \operatorname{Re}(z) \leq 0\} \subset S. \tag{2-35}$$

2.44 Definition. Ein Verfahren mit der Eigenschaft (2-35) heißt absolut stabil oder kurz A-stabil.

2.45 Definition. Ein A-stabiles numerisches Verfahren heißt stark absolut stabil oder kurz L-stabil, wenn $g(z) \to 0$ für $\operatorname{Re}(z) \to -\infty$ gilt.

Die Koeffizienten $g(\Delta t\lambda_k)^j$ in der diskreten Lösung klingen dann umso stärker ab, je kleiner $\operatorname{Re}(\lambda_k)$ ist, genau wie dies die Koeffizienten $\exp(\Delta t\lambda_k)^j$ in der kontinuerlichen Lösung tun.

2.46 Beispiel (ϑ-Verfahren). Man findet $g(z) = g_\vartheta(z) = \frac{1+(1-\vartheta)z}{1-\vartheta z}$, $0 \leq \vartheta \leq 1$. Das ϑ-Verfahren ist A-stabil, d.h. $S \supset \mathbb{C}_-$, falls $1/2 \leq \vartheta \leq 1$. Ferner gilt

$$\lim_{\operatorname{Re}(z)\to-\infty} |g_\vartheta(z)| = \lim_{\operatorname{Re}(z)\to-\infty} \left|\frac{1 + (1-\vartheta)z}{1 - \vartheta z}\right| = \left|\frac{1-\vartheta}{\vartheta}\right|,$$

d.h. das ϑ-Verfahren ist L-stabil, nur falls $\vartheta = 1$.

© Johannes Schropp 22. Mai 2024

2.47 Beispiel (Explizite Runge-Kutta-Verfahren)**.** Ein explizites Runge-Kutta-Verfahren lautet

$$w^{m+1} = w^m + \Delta t \sum_{i=1}^{s} b_i \Gamma W_i^m, \quad m = 0, 1, \ldots, N-1$$

mit

$$\begin{aligned} W_1^m &= w^m, \\ W_i^m &= w^m + \Delta t \sum_{j=1}^{i-1} a_{ij} \Gamma W_j^m, \quad i = 2, \ldots, s, \end{aligned}$$

wobei $a_{ij} = 0$ für $j \geq i$ zu beachten ist. Man erhält, dass $g(z)$ ein Polynom vom höchstens s-ten Grade ist, d.h.

$$g(z) = \frac{p(z)}{1}, \quad \deg(p) \leq s.$$

Wegen des Satzes von Liouville, d.h. $|p(z)| \to \infty$ für $\mathrm{Re}(z) \to -\infty$ folgt, dass explizite Runge-Kutta-Verfahren niemals A-stabil sind.

2.48 Beispiel (Implizite Runge-Kutta-Verfahren)**.** Nicht alle impliziten Runge-Kutta-Verfahren sind A-stabil. Aber manche sind es. Ferner findet man A-stabile Runge-Kutta-Verfahren von beliebig hoher Ordnung. So sind z.B. die Verfahren vom Gauss-Typ oder vom Radau IIA Typ A-stabil. Die Radau IIA Verfahren sind sogar L-stabil.

Absolute Stabilitätsbereiche für lineare Mehrschrittverfahren

Vorgelegt sei die Anfangswertaufgabe

$$w'(t) = \lambda w(t), \quad w(0) = w^0, \quad \lambda \in \mathbb{C}. \tag{2-36}$$

welche durch $\overline{w}(t) = \exp(\lambda t)$ gelöst wird. Wir diskretisieren das Problem (2-36) mit einem linearen Mehrschrittverfahren mit dem Tableau

a_0	a_1	$\ldots$	a_k
b_0	b_1	$\ldots$	b_k

und erhalten mit der Approximation w^l für $\overline{w}(t_l)$ das Schema

$$\sum_{i=0}^{k} a_i w^{j+i} = \Delta t \sum_{i=0}^{k} b_i \lambda w^{j+i}, \quad j = 0, \ldots, N-k$$

zu vorgegebenen $w^0, \ldots, w^{k-1}$. Die obige Iteration ist äquivalent zu

$$\sum_{i=0}^{k} (a_i - b_i \Delta t \lambda) w^{j+i} = 0, \quad j = 0, \ldots, N-k. \tag{2-37}$$

© Johannes Schropp 22. Mai 2024

(2-37) hat die Form

$$\sum_{i=0}^{k} g_i(\Delta t\lambda) w^{j+i} = 0$$

mit $g_i(z) = a_i - b_i z$, $i = 0, \dots, k$.

2.49 Definition. Zum Polynom

$$p(z, \xi) := \sum_{i=0}^{k} g_i(z)\xi^i \tag{2-38}$$

heißt die Menge

$$\begin{aligned} S := \{z \in \mathbb{C} \mid &\text{Für jede Nullstelle } \xi \text{ von } p(z,\cdot) \text{ gilt } |\xi| = 1, \\ &\text{und } \xi \text{ ist einfache Nullstelle oder } |\xi| < 1 \text{ sonst.}\} \end{aligned} \tag{2-39}$$

der absolute Stabilitätsbereich des linearen Mehrschrittverfahrens.

Zur Motivation von (2-38)—(2-39) betrachten wir die Lösung von (2-37) wobei wir der Einfachheit halber annehmen, dass das Polynom $p(\Delta t\lambda, \xi)$ nur einfache Nullstellen $\xi_j \in \mathbb{C}$, $j = 1, \dots, k$ hat. Wir bestimmen nun $\gamma_1, \gamma_2, \dots, \gamma_k$ eindeutig als Lösung von

$$\underbrace{\begin{pmatrix} 1 & 1 & \dots & 1 \\ \xi_1 & \xi_2 & \dots & \xi_k \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1^{k-1} & \xi_2^{k-1} & \dots & \xi_k^{k-1} \end{pmatrix}}_{=:V} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{pmatrix} = \begin{pmatrix} w^0 \\ w^1 \\ \vdots \\ w^{k-1} \end{pmatrix} \tag{2-40}$$

2.50 Bemerkung. Die Matrix $V \in \mathbb{R}^{k,k}$ ist die Vandermondesche Matrix, welche genau dann invertierbar ist, wenn $\xi_i \neq \xi_j$ für $i \neq j$.

Dann lautet die Lösung von (2-37)

$$w^j = \sum_{l=1}^{k} \gamma_l \xi_l^j, \quad j \in \mathbb{N}_0. \tag{2-41}$$

Sei $j \in \{0, \dots, k-1\}$. Nach Konstruktion von $\gamma_1, \dots, \gamma_k$ gilt

$$w^j = \sum_{l=1}^{k} \gamma_k \xi_l^j.$$

Für $j \geq k$ finden wir

$$\sum_{\kappa=0}^{k} g_\kappa(\Delta t\lambda) w^{j+\kappa} = \sum_{\kappa=0}^{k} g_\kappa(\Delta t\lambda) \sum_{l=1}^{k} \gamma_l \underbrace{\xi_l^{j+\kappa}}_{=\xi_l^j \cdot \xi_l^\kappa} = \sum_{l=1}^{k} \gamma_l \xi_l^j \underbrace{\sum_{\kappa=0}^{k} g_\kappa(\Delta t\lambda)\xi_l^\kappa}_{=p(\Delta t\lambda, \xi_l)=0} = 0.$$

© Johannes Schropp 22. Mai 2024

Anhand der Lösungsdarstellung (2-41) sehen wir, dass im Fall $\mathrm{Re}(\lambda) \leq 0$ wiederum $\Delta t\lambda \in S$ zu fordern ist, damit die numerische Lösung, genau wie die kontinuierliche Lösung, beschränkt bleibt.

2.51 Definition. Ein lineares Mehrschrittverfahren heißt absolut stabil oder A-stabil, falls $\mathbb{C}_- \subset S$ gilt. Es heißt sogar L-stabil, falls es A-stabil ist und falls zusätzlich für alle Nullstellen $\xi(z)$ von $p(z,\cdot)$

$$\lim_{\mathrm{Re}(z)\to-\infty} \xi(z) = 0$$

gilt.

Es gibt aber ein bekanntes Resultat von Dahlquist, nach dem jedes A-stabile lineare Mehrschrittverfahren implizit ist und höchstens die Konvergenzordnung 2 hat. Dazu gehören die ϑ-Verfahren für $1/2 \leq \vartheta \leq 1$ und das BDF-Verfahren der Stufe 2 mit dem Tableau $\begin{array}{ccc} 1/2 & -2 & 3/2 \\ \hline 0 & 0 & 1 \end{array}$.

Um also lineare Mehrschrittverfahren mit höherer Konvergenzordnung als 2 zu erhalten, muss man den Begriff der A-Stabilität etwas abschwächen. Dies führt dann zu folgender Definition.

2.52 Definition. Ein lineares Mehrschrittverfahren heißt A(α)-stabil mit $0 < \alpha < \frac{\pi}{2}$, falls

$$S \supset \mathbb{C}_{-,\alpha} = \{z \in \mathbb{C}_- \mid \frac{\mathrm{Im}(z)}{\mathrm{Re}(z)} < \tan(\alpha)\}.$$

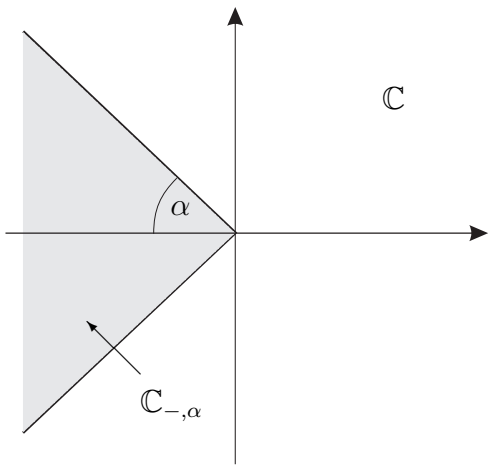


Abbildung 9: Die Menge $\mathbb{C}_{-,\alpha}$ für ein $\alpha > 0$

© Johannes Schropp 22. Mai 2024

Man erhält, dass die BDF-Formeln bis zur Ordnung 6 (nur bis zu dieser Ordnung sind sie überhaupt stabil) A(α)-stabil sind. Die BDF-Formeln erfüllen überdies

$$\lim_{\mathrm{Re}(z)\to-\infty} \xi(z) = 0$$

für alle Nullstellen $\xi(z)$ von $p(z,\cdot)$, d.h. sie sind L(α)-stabil.

2.53 Bemerkung. A(α)- bzw. L(α)-Stabilität, $\alpha > 0$, ist beispielsweise völlig hinreichend für Probleme der Gestalt $w' = \Gamma w$ mit lediglich reellen Eigenwerten λ. Man beachte dabei, dass $\Delta t\lambda \in \mathbb{R}$ für alle Eigenwerte $\lambda \in \sigma(\Gamma)$ gilt. Darunter fallen unsere Liniensysteme parabolischer Anfangsrandwertprobleme unabhängig davon, ob die Diskretisierung im Raum mit Finiten Differenzen oder Finiten Elementen gemacht wurde.

Software-Pakete für steife Differentialgleichungen

a) Radau 5:
Radau IIA Verfahren der Stufe $s = 3$ und Ordnung $p = 5$ (Runge-Kutta-Verfahren), E. Hairer (Genf).

b) DASSL:
BDF-Verfahren mit variabler Ordnung durch Benutzung der Stufen $k = 2, \ldots, 6$ (lineares Mehrschrittverfahren), L. Petzold (St. Barbara).

c) ode15s:
BDF-Verfahren für steife Differentialgleichungen in **Matlab**.

d) ode23tb:
BDF-Verfahren der Stufe 2 und Trapezregel.

e) Nichtlineare parabolische Anfangsrandwertaufgaben

Es sei $\Omega = (0,1)$. Betrachte die nichtlineare Anfangsrandwertaufgabe

$$\begin{aligned} \frac{\partial u}{\partial t} &= u_{xx} + f(u,x,t) \text{ in } \Omega \times (0,T), \\ u(x,0) &= u_0(x) \text{ in } \Omega, \\ u(0,t) &= \gamma_0(t), \quad u(1,t) = \gamma_1(t) \text{ in } (0,T). \end{aligned} \tag{2-42}$$

Es wird $f \in C^1(\mathbb{R} \times [0,1] \times [0,T], \mathbb{R})$ und

$$q \le \frac{\partial f}{\partial u}(u,x,t) \le \mu, \quad u \in \mathbb{R}, \quad 0 \le x \le 1, \quad 0 \le t \le T \tag{2-43}$$

© Johannes Schropp 22. Mai 2024

vorausgesetzt.

Es seien $\Delta x = \frac{1}{M} > 0$, $\Omega_{\Delta x} = \{x_j = j\Delta x \,|\, j = 1, \ldots, M-1\}$. Setze

$$v(t) = (u(x_1,t), \ldots, u(x_{M-1},t)) = (v_1(t), \ldots, v_{M-1}(t)), \quad 0 \le t \le T.$$

Das Liniensystem durch Diskretisierung mit dem klassischen Differenzenverfahren für den nichtlinearen Differentialoperator $u_{xx}(\cdot,t) + f(u(\cdot,t),\cdot,t)$, $u(0,t) = \gamma(0,t)$, $u(1,t) = \gamma_1(t)$, $t \in (0,T)$ lautet

$$\begin{aligned} v'(t) &= F_{\Delta x}(v(t),t), \quad 0 \le t \le T, \\ v(0) &= v^0 \end{aligned}$$

mit

$$\begin{aligned} F_{\Delta x}(v,t) &= -\Gamma v + E(v,t) + r^1(t), \\ \Gamma &= \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}, \\ r^1(t) &= \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0(t) \\ 0 \\ \vdots \\ 0 \\ \gamma_1(t) \end{pmatrix} \in \mathbb{R}^{M-1}, \\ E(v,t) &= \begin{pmatrix} f(v_1,x_1,t) \\ \vdots \\ f(v_i,x_i,t) \\ \vdots \\ f(v_{M-1},x_{M-1},t) \end{pmatrix}, \quad v^0 = \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_i) \\ \vdots \\ u_0(x_{M-1}) \end{pmatrix} \in \mathbb{R}^{M-1} \end{aligned} \tag{2-44}$$

Mit $v^j = (u_1^j, \ldots, u_{M-1}^j) \in \mathbb{R}^{M-1}$, $j = 0, \ldots, N$, $N\Delta t = T$, $t_j = j\Delta t$ lautet das ϑ-Verfahren

$$\begin{aligned} \frac{1}{\Delta t}(v^j - v^{j-1}) &= \vartheta F_{\Delta x}(v^j, t_j) + (1-\vartheta) F_{\Delta x}(v^{j-1}, t_{j-1}), \quad j = 1, \ldots, N, \\ v^0 &= r^0 = (u_0(x_1), \ldots, u_0(x_{M-1})), \end{aligned}$$

wobei $F_{\Delta x}(v,t) = -\Gamma v + E(v,t) + r^1(t)$. Mit den Abkürzungen

$$G_j(v) = \frac{1}{\Delta t} v - \vartheta F_{\Delta x}(v, t_j),$$

© Johannes Schropp 22. Mai 2024

$$H_j(v) \;=\; \frac{1}{\Delta t}v + (1-\vartheta)F_{\Delta x}(v,t_j)$$

schreiben wir das diskrete Problem in der Form $T^h(u) = 0$ mit $T^h : \mathbb{R}^{\Omega_h} \to \mathbb{R}^{\Omega_h}$ gegeben durch

$$T^h(v) = \begin{pmatrix} v^0 - r^0 \\ G_1(v^1) - H_0(v^0) \\ \vdots \\ G_N(v^N) - H_{N-1}(v^{N-1}) \end{pmatrix} = 0$$

wobei $\Omega_h = \{(x_i,t_j) \,|\, i = 1,\ldots,M-1,\ j = 0,\ldots,N\}$, $h = (\Delta x, \Delta t)$.

Gemäß Voraussetzung (2-43) gilt für $w, z \in \mathbb{R}^{M-1}$

$$\begin{aligned} F_{\Delta x}(w,t) \;-\; F_{\Delta x}(z,t) \;&=\; -\Gamma(w-z) + E(w,t) - E(z,t) \\ &=\; -\Gamma(w-z) + \begin{pmatrix} f(w_1,x_1,t) - f(z_1,x_1,t) \\ \vdots \\ f(w_{M-1},x_{M-1},t) - f(z_{M-1},x_{M-1},t) \end{pmatrix} \\ &=\; -\Gamma(w-z) + \begin{pmatrix} \frac{\partial f}{\partial u}(\xi_1^t,x_1,t)(w_1-z_1) \\ \vdots \\ \frac{\partial f}{\partial u}(\xi_{M-1}^t,x_{M-1},t)(w_{M-1}-z_{M-1}) \end{pmatrix} \\ &=\; -\Gamma(w-z) + \underbrace{\operatorname{diag}(D_{ii}(w_i,z_i,t), i = 1,\ldots,M-1)}_{=:D(w,z,t)}(w-z) \\ &=\; -\Gamma(w-z) + D(w,z,t)(w-z) \end{aligned}$$

mit

$$q \le D_{ii}(w_i,z_i,t) \le \mu, \quad i = 1,\ldots,M-1, \tag{2-45}$$

wobei

$$\begin{aligned} D_{ii}(w_i,z_i,t) \;&=\; \frac{\partial f}{\partial u}(\xi_i^t,x_i,t), \quad i = 1,\ldots,M-1, \\ \xi_i^t \;&=\; w_i + s^t(z_i - w_i) \text{ für ein } s^t \in (0,1),\ t \in]0,T[,\ i = 1,\ldots,M-1. \end{aligned}$$

Herleitung einer Stabilitätsungleichung

Wir wählen zwei Elemente $v = (v^0,\ldots,v^N)$, $z = (z^0,\ldots,z^N) \in \mathbb{R}^{(N+1)(M-1)}$ und setzen $g = (g^0,\ldots,g^N) := T^h(v) - T^h(z)$, d.h.

$$\begin{aligned} g^0 \;&=\; v^0 - z^0, \\ g^j \;&=\; G_j(v^j) - G_j(z^j) - (H_{j-1}(v^{j-1}) - H_{j-1}(z^{j-1})), \quad j = 1,\ldots,N. \end{aligned}$$

Diese Rekursion lässt sich ferner wie folgt schreiben:

$$g^j \;=\; \frac{1}{\Delta t}v^j - \vartheta F_{\Delta x}(v^j,t_j) - \frac{1}{\Delta t}z^j + \vartheta F_{\Delta x}(z^j,t_j)$$

© Johannes Schropp 22. Mai 2024

$$
\begin{aligned}
& -\left[\frac{1}{\Delta t}v^{j-1} + (1-\vartheta)F_{\Delta x}(v^{j-1}, t_{j-1}) - \frac{1}{\Delta t}z^{j-1} - (1-\vartheta)F_{\Delta x}(z^{j-1}, t_{j-1})\right] \\
= & \underbrace{\left(\frac{1}{\Delta t}I - \vartheta(-\Gamma + D(v^j, z^j, t_j))\right)}_{=:A_j}(v^j - z^j) \\
& - \underbrace{\left(\frac{1}{\Delta t}I + (1-\vartheta)(-\Gamma + D(v^{j-1}, z^{j-1}, t_{j-1}))\right)}_{=:B_{j-1}}(v^{j-1} - z^{j-1}) \\
= & A_j(v^j - z^j) - B_{j-1}(v^{j-1} - z^{j-1}), \quad j = 1, \ldots, N, \\
g^0 = & v^0 - z^0.
\end{aligned}
$$

Diskussion von A_j

A_j ist eine L_0-Matrix, denn $(A_j)_{lk} = \vartheta\Gamma_{lk} \leq 0$ für $l \neq k$, und es gilt

$$
A_j\mathbb{I} = \frac{1}{\Delta t}\mathbb{I} + \underbrace{\vartheta\Gamma\mathbb{I}}_{\geq 0} - \vartheta\underbrace{D(v^j, z^j, t_j)\mathbb{I}}_{\leq \mu\mathbb{I}} \geq \left(\frac{1}{\Delta t} - \vartheta\mu\right)\mathbb{I}.
$$

Setzt man nun $\Delta t\vartheta\mu \leq \kappa < 1$ voraus, so ist A_j sogar eine M-Matrix mit

$$
\mathbb{I} = A_j^{-1}A_j\mathbb{I} \geq \left(\frac{1}{\Delta t} - \vartheta\mu\right)A_j^{-1}\mathbb{I},
$$

d.h.

$$
A_j^{-1}\mathbb{I} \leq \frac{1}{\frac{1}{\Delta t} - \vartheta\mu}\mathbb{I} = \frac{\Delta t}{1 - \Delta t\vartheta\mu}\mathbb{I}. \tag{2-46}
$$

Wir erhalten ferner

$$
\|A_j^{-1}\|_\infty = \|A_j^{-1}\mathbb{I}\|_\infty \leq \left\|\frac{\Delta t}{1 - \Delta t\vartheta\mu}\mathbb{I}\right\|_\infty = \frac{\Delta t}{1 - \Delta t\vartheta\mu} \leq \frac{\Delta t}{1 - \kappa} =: C_1\Delta t. \tag{2-47}
$$

Durchführbarkeit des Verfahrens

Es ist zu zeigen, dass $G_j : \mathbb{R}^{M-1} \to \mathbb{R}^{M-1}$ für $j = 1, \ldots, N$ bijektiv ist. Betrachte hierzu die Gleichung $G(v) = \frac{1}{\Delta t}v + \vartheta\Gamma v - \vartheta E(v,t) = r$ für $r \in \mathbb{R}^{M-1}$ beliebig. Diese ist äquivalent zu

$$
C_{\Delta t}v = \vartheta E(v,t) + r \text{ mit } C_{\Delta t} = \frac{1}{\Delta t}I + \vartheta\Gamma.
$$

Man findet

$$
C_{\Delta t}\mathbb{I} = \frac{1}{\Delta t}\mathbb{I} + \vartheta\Gamma\mathbb{I} \geq \frac{1}{\Delta t}\mathbb{I} > 0,
$$

© Johannes Schropp 22. Mai 2024

d.h. $C_{\Delta t}$ ist eine M-Matrix. Analog zu (2-46), (2-47) folgt $C_{\Delta t}^{-1}\mathbb{I} \leq \Delta t\mathbb{I}$, d.h.

$$\|C_{\Delta t}^{-1}\|_\infty = \|C_{\Delta t}^{-1}\mathbb{I}\|_\infty \leq \Delta t.$$

Somit ist $G(v) = r$ äquivalent zur Fixpunktgleichung

$$v \;=\; C_{\Delta t}^{-1}\left(\vartheta E(v,t) + r\right) \;=\; T(v).$$

Wir zeigen nun, dass T eine Kontraktion ist, falls $\max(0,\mu)\vartheta\Delta t < 1$ ist. Dann folgt die Behauptung mit dem Fixpunktsatz von Banach. Es gilt

$$T(v) - T(z) \;=\; \vartheta C_{\Delta t}^{-1}(E(v,t) - E(z,t)) \;=\; \vartheta C_{\Delta t}^{-1} D(v,z,t)(v-z) \;\leq\; \vartheta C_{\Delta t}^{1}\mu(v-z)$$

und

$$\|T(v) - T(z)\|_\infty \;=\; \max(0,\mu)\vartheta\|C_{\Delta t}^{-1}\|_\infty\|(v-z)\|_\infty \;\leq\; \max(0,\mu)\vartheta\Delta t\|(v-z)\|_\infty.$$

Also ist das ϑ-Verfahren unter der Bedingung

$$\Delta t\vartheta\mu \leq \kappa < 1$$

durchführbar.

Diskussion von B_j

Es gilt

$$B_j = \frac{1}{\Delta t}I + (1-\vartheta)(-\Gamma + \underbrace{D(v^j, z^j, t_j)}_{\leq \mu I}) \leq \frac{1}{\Delta t}I + (1-\vartheta)(-\Gamma + \mu I) =: P. \quad (2\text{-}48)$$

Wir wollen nun die Ungleichung $|B_j| \leq P$, d.h.

$$|(B_j)_{kl}| \leq P_{kl}, \quad 1 \leq k,l \leq M-1$$

erfüllen. Dies ist äquivalent zu $B_j \leq P$ und $B_j \geq -P$. Die Formelzeile (2-48) sichert die Abschätzung $B_j \leq P$. Zu zeigen bleibt also, dass auch $B_j \geq -P$ gilt.

Für die Nebendiagonalelemente von B_j ergibt sich

$$(B_j)_{kl} = -(1-\vartheta)\Gamma_{k,l} \geq 0 \geq (1-\vartheta)\Gamma_{kl} = -P_{kl}.$$

Somit folgt $B_j \geq -P$, falls für die Hauptdiagonalelemente

$$\begin{aligned}(B_j)_{kk} \;&=\; \frac{1}{\Delta t} + (1-\vartheta)\Bigg(-\frac{2}{\Delta x^2} + \underbrace{(D(v^j, z^j, t_j))_{kk}}_{\geq q}\Bigg)\\ &\geq\; \frac{1}{\Delta t} + (1-\vartheta)\left(-\frac{2}{\Delta x^2} + q\right) \overset{!}{\geq} -\frac{1}{\Delta t} - (1-\vartheta)\left(-\frac{2}{\Delta x^2} + \mu\right) = -P_{kk}\end{aligned}$$

© Johannes Schropp 22. Mai 2024

gilt. Dies ist genau dann erfüllt, wenn

$$\frac{2}{\Delta t} \geq (1-\vartheta)\left(\frac{4}{\Delta x^2} - (q+\mu)\right),$$

was mit

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)} + \frac{q+\mu}{4}\Delta t \tag{2-49}$$

äquivalent ist.

Wir setzen nun (2-49) voraus und leiten eine Stabilitätsungleichung her. Aus

$$g^j = A_j(v^j - z^j) - B_{j-1}(v^{j-1} - z^{j-1})$$

folgt

$$\begin{aligned} \|v^j - z^j\|_\infty &= \|A_j^{-1}B_{j-1}(v^{j-1} - z^{j-1}) + A_j^{-1}g^j\|_\infty \\ &\leq \|A_j^{-1}B_{j-1}\|_\infty \cdot \|v^{j-1} - z^{j-1}\|_\infty + \|A_j^{-1}\|_\infty \cdot \|g^j\|_\infty. \end{aligned} \tag{2-50}$$

Nun gilt mit $|\cdot| : \mathbb{R}^{M-1,M-1} \to \mathbb{R}^{M-1,M-1}$ definiert durch

$$C \mapsto (|C_{ij}|)_{1\leq i,j\leq M-1}$$

sofort

$$\pm B_{j-1} \leq |B_{j-1}| \leq P \;\Rightarrow\; \pm A_j^{-1}B_{j-1} \leq A_j^{-1}|B_{j-1}| \leq A_j^{-1}P$$

und somit

$$\|A_j^{-1}B_{j-1}\|_\infty \leq \|A_j^{-1}P\|_\infty.$$

Einsetzen in (2-50) liefert

$$\|v^j - z^j\|_\infty \leq \|A_j^{-1}P\|_\infty \cdot \|v^{j-1} - z^{j-1}\|_\infty + \|A_j^{-1}\|_\infty \cdot \|g^j\|_\infty.$$

Wir benötigen jetzt eine Abschätzung für $\|A^{-1}P\|_\infty$:

$$\begin{aligned} (A_j - P)\mathbb{I} &= \left[\left(\frac{1}{\Delta t}I - \vartheta(-\Gamma + D(v^j, z^j, t_j))\right) - \left(\frac{1}{\Delta t}I + (1-\vartheta)(-\Gamma + \mu I)\right)\right]\mathbb{I} \\ &= (\Gamma - \vartheta \underbrace{D(v^j, z^j, t_j)}_{\leq \mu I} - (1-\vartheta)\mu I)\mathbb{I} \\ &\geq (\Gamma - \mu I)\mathbb{I} = \underbrace{\Gamma\mathbb{I}}_{\geq 0} - \mu\mathbb{I} \geq -\mu\mathbb{I}. \end{aligned}$$

Da A_j eine M-Matrix ist, folgt

$$(I - A_j^{-1}P)\mathbb{I} = A_j^{-1}(A_j - P)\mathbb{I} \geq -\mu A_j^{-1}\mathbb{I},$$

was mit

$$A_j^{-1}P\mathbb{I} \leq (I + \mu A_j^{-1})\mathbb{I}$$

© Johannes Schropp 22. Mai 2024

gleichbedeutend ist. Dies liefert

$$\begin{aligned} A_j^{-1}P\mathbb{I} &\leq \mathbb{I} + \mu A_j^{-1}\mathbb{I} = \mathbb{I} + \max\{\mu,0\} \underbrace{A_j^{-1}\mathbb{I}}_{\leq \frac{\Delta t}{1-\Delta t\vartheta\mu}\mathbb{I}} \\ &\overset{(2-46)}{\leq} \mathbb{I} + \underbrace{\max\{\mu,0\}\cdot\frac{\Delta t}{1-\Delta t\vartheta\mu}}_{\leq: C_2\Delta t}\mathbb{I} \leq (1+C_2\Delta t)\mathbb{I} \end{aligned}$$

mit $C_2 := \max\{\mu,0\}\frac{1}{1-\kappa}$.

Wegen $A_j^{-1}P \geq 0$ folgt ferner

$$\|A_j^{-1}P\|_\infty = \|A_j^{-1}P\mathbb{I}\|_\infty \leq \|(1+C_2\Delta t)\mathbb{I}\|_\infty = 1 + C_2\Delta t. \tag{2-51}$$

Benutzen wir nun (2-47) und (2-51), so folgt

$$\begin{aligned} \|v^j - z^j\|_\infty &\leq \|A_j^{-1}P\|_\infty \cdot \|v^{j-1} - z^{j-1}\|_\infty + \|A_j^{-1}\|_\infty \cdot \|g^j\|_\infty \\ &\leq (1+C_2\Delta t)\|v^{j-1} - z^{j-1}\|_\infty + C_1\Delta t \cdot \|g^j\|_\infty, \quad j = 1,\ldots,N. \end{aligned}$$

Durch Induktion erhalten wir dann

$$\|v^j - z^j\|_\infty \leq (1+C_2\Delta t)^j\|g^0\|_\infty + \sum_{k=1}^{j}(1+C_2\Delta t)^{j-k}\cdot C_1\Delta t\cdot\|g^k\|_\infty. \tag{2-52}$$

$j = 0$: Offensichtlich gilt

$$\|v^0 - z^0\|_\infty \leq \|g^0\|_\infty = \|v^0 - z^0\|_\infty.$$

$j-1 \to j$: Es folgt

$$\begin{aligned} \|v^j - z^j\|_\infty &\leq (1+C_2\Delta t)\cdot\|v^{j-1} - z^{j-1}\|_\infty + C_1\Delta t\|g^j\|_\infty \\ &\leq (1+C_2\Delta t)\big[(1+C_2\Delta t)^{j-1}\|g^0\|_\infty \\ &\quad + \sum_{k=1}^{j-1}(1+C_2\Delta t)^{j-1-k}\cdot C_1\Delta t\cdot\|g^k\|_\infty\big] + C_1\Delta t\|g^j\|_\infty \\ &= (1+C_2\Delta t)^j\|g^0\|_\infty + \sum_{k=1}^{j-1}(1+C_2\Delta t)^{j-k}\cdot C_1\Delta t\cdot\|g^k\|_\infty + C_1\Delta t\cdot\|g^j\|_\infty \\ &= (1+C_2\Delta t)^j\|g^0\|_\infty + \sum_{k=1}^{j}(1+C_2\Delta t)^{j-k}\cdot C_1\Delta t\|g^k\|_\infty. \end{aligned}$$

Mit (2-52) und $1 + C_2\Delta t \leq \exp(C_2\Delta t)$ ergibt sich dann

$$\|v^j - z^j\|_\infty \leq \exp(C_2\cdot j\Delta t)\cdot\|g^0\|_\infty + \sum_{k=1}^{j}\underbrace{(1+C_2\Delta t)^{j-k}}_{\leq(1+C_2\Delta t)^j\leq\exp(C_2 j\Delta t)}\cdot C_1\Delta t\cdot\|g^k\|_\infty$$

© Johannes Schropp 22. Mai 2024

$$\begin{aligned}
&\leq \exp(C_2 \cdot \underbrace{j\Delta t}_{\leq T}) \cdot \|g^0\|_\infty \\
&\quad + C_1 \underbrace{j\Delta t}_{\leq T} \cdot \exp(C_2 \underbrace{j\Delta t}_{\leq T}) \cdot \max\{\|g^j\|_\infty \,|\, j = 1, \ldots, N\} \\
&\leq \exp(C_2 T)(1 + C_1 T)\|g\|_\infty \\
&= \exp(C_2 T)(1 + C_1 T)\|T^h(v) - T^h(z)\|_\infty, \quad j = 0, \ldots, N.
\end{aligned}$$

Dies liefert die Behauptung mit der Stabilitätskonstanten $C := \exp(C_2 T)(1 + C_1 T)$.

Fasst man nun alles zusammen, so ergibt sich der nachstehende Satz.

2.54 Satz. *Unter der Voraussetzung $f \in C^1(\mathbb{R} \times [0,1] \times [0,T], \mathbb{R})$, $q \leq \frac{\partial f}{\partial u}(u,x,t) \leq \mu$, $u \in \mathbb{R}$, $0 \leq x \leq 1$, $0 \leq t \leq T$ sowie der Schrittweitenbedingung $\Delta t \vartheta \mu \leq \kappa < 1$ ist das ϑ-Verfahren zur Anfangsrandwertaufgabe (2-42) durchführbar. Mit der zusätzlichen Bedingung*

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)} + \frac{q+\mu}{4}\Delta t$$

ist es auch bezüglich $\|\cdot\|_\infty$ stabil. Es gilt dann die Stabilitätsungleichung

$$\|u - w\|_\infty \leq \exp(C_2 T)(1 + C_1 T)\|T^h(u) - T^h(w)\|_\infty, \quad \forall u, w \in \mathbb{R}^{\Omega_h}, \quad h = (\Delta x, \Delta t)$$

mit $C_1 = \frac{1}{1-\kappa}$, $C_2 = \max\{\mu, 0\} \cdot C_1$.

An jeder klassischen Lösung $\overline{u}$ mit $\frac{\partial^\nu}{\partial t^\nu}\overline{u} \in C([0,1]\times[0,T])$, $\nu = 1,2$, $\frac{\partial^\nu}{\partial x^\nu}\overline{u} \in C([0,1]\times[0,T])$, $\nu = 1,2,3,4$ liegt die Konvergenz der Ordnung $O(\Delta t + \Delta x^2)$ bezüglich $\|\cdot\|_\infty$ vor. Für das Crank-Nicholson-Verfahren erhält man sogar die Konvergenz der Ordnung $O(\Delta t^2 + \Delta x^2)$ bezgl. $\|\cdot\|_\infty$, falls zusätzlich $\frac{\partial^3}{\partial t^3}\overline{u} \in C([0,1] \times [0,T])$.

f) Finite Elemente für nichtlineare parabolische Probleme

Nun wenden wir uns dem Lösen des nichtlinearen parabolischen Problems

$$\begin{aligned}
u_t &= \Delta u + f(u) \text{ in } \Omega \times (0,T), \\
u(\cdot,t) &= 0 \text{ auf } \partial\Omega \times (0,T), \\
u(\cdot,0) &= u_0 \text{ in } \Omega
\end{aligned} \tag{2-53}$$

zu, wobei $\Omega \subset \mathbb{R}^2$ eine beschränktes Gebiet ist. Die Nichtlinearität $f \in C^1(\mathbb{R})$ sei global Lipschitz-stetig. Für die Anfangsdaten gelte $u_0 \in L^2(\Omega)$.

Analog zum linearen Fall erhält man

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(x,t)v(x)\,\mathrm{d}x + \int_\Omega \nabla u(x,t) \cdot \nabla v(x)\,\mathrm{d}x = \int_\Omega f(u(x,t))v(x)\,\mathrm{d}x$$

© Johannes Schropp 22. Mai 2024

für alle $v \in C_0^\infty(\Omega)$, $t \in (0,T)$. Wir interpretieren wieder für jedes feste t die Abbildung

$$x \mapsto u(t,x), \text{ d.h. } u(\cdot,t)$$

als ein Element eines Sobolev-Raumes V (hier: $V = H_0^1(\Omega)$) und schreiben kurz $u(t) \in V$. Wir setzen

$$\begin{aligned} \langle f(u(t)), v\rangle_0 &= \int_\Omega f(u(x,t))v(x)\,\mathrm{d}x, \quad 0 < t < T, \quad v \in H_0^1(\Omega), \\ a(v,w) &= \int_\Omega \nabla u(x) \cdot \nabla v(x)\,\mathrm{d}x, \quad v, w \in H_0^1(\Omega) \end{aligned}$$

und finden mit der Definition der verallgemeinerten Ableitung das folgende schwache Problem: Gesucht ist ein Element $u \in L^2((0,T),V)$, $V = H_0^1(\Omega)$, das eine schwache Ableitung $u' \in L^2((0,T),V')$ besitzt, mit $u(0) = u_0$ und

$$\langle u'(t), v\rangle_0 + a(u(t),v) = \langle f(u(t)), v\rangle_0, \quad \forall v \in V, \quad t \in (0,T). \tag{2-54}$$

2.55 Bemerkung. Es ist zu beachten, dass sich aus $u \in L^2((0,T),V)$, $u' \in L^2((0,T),V')$ mit dem Interpolationssatz $u \in C([0,T],H)$, $H = L^2(\Omega)$ ergibt. Mit einer Lipschitz-stetigen Abbildung $f : \mathbb{R} \to \mathbb{R}$ folgt dann, dass durch

$$\Omega \ni x \mapsto f(u(t))(x) = f(u(t,x))$$

eine Abbildung $f(u(\cdot)) \in C([0,T],H)$ definiert wird.

Das Galerkin-Verfahren zu (2-54) lautet: Sei $V_h \subset V$ endlich-dimensional und $u_{0h} \in V_h$ eine Approximation von u_0. Finde ein $u_h \in L^2((0,T),V_h)$ mit $u_h' \in L^2((0,T),V')$ und $u(0) = u_{0h}$ sowie

$$\langle u_h'(t), v\rangle_0 + a(u_h(t),v) = \langle f(u_h(t)), v\rangle_0, \quad \forall v \in V_h, \quad t \in (0,T). \tag{2-55}$$

Typischerweise wählt man $u_{0h} = R_h(u_0)$ mit der Ritz-Projektion $R_h : V \to V_h$.

Unsere nächste Aufgabe ist es, die Lösbarkeit des Problems (2-55) zu diskutieren. Der numerische Ansatz zu $V_h = \operatorname{span}\{v_1,\dots,v_M\}$ lautet

$$\begin{aligned} u_h(t) &= \sum_{i=1}^M c_i(t)v_i, \\ u_{0h} &= \sum_{i=1}^M c_{i0}v_i \in V_h. \end{aligned}$$

Analog zum linearen Fall erhält man

$$\sum_{i=1}^M c_i'(t)\langle v_i, v_j\rangle_0 + \sum_{i=1}^M c_i(t)a(v_i,v_j) = \Big\langle f\Big(\sum_{k=1}^M c_k(t)v_k\Big), v_j\Big\rangle_0, \quad j = 1,\dots,M,\ 0 < t < T,$$

© Johannes Schropp 22. Mai 2024

$$c(0) = c_0.$$

Mit der Steifigkeitsmatrix

$$A = (a_{ij})_{1\le i,j\le M}, \quad a_{ij} = a(v_i, v_j), \quad 1 \le i, j \le M,$$

der Massenmatrix

$$B = (b_{ij})_{1\le i,j\le M}, \quad b_{ij} = \langle v_i, v_j\rangle_0, \quad 1 \le i, j \le M$$

und dem Vektorfeld $G : \mathbb{R}^M \to \mathbb{R}^M$

$$G(a) := (G_j(a)), \quad G_j(a) = \int_\Omega f\Big(\sum_{k=1}^M a_k v_k\Big) v_j \,\mathrm{d}x, \quad 1 \le j \le M$$

erhalten wir das System

$$\begin{aligned} Bc'(t) + Ac(t) &= G(c(t)), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-56}$$

Dabei ist $B \in \mathbb{R}^{M,M}$ symmetrisch und positiv definit, und die Abbildung $G : \mathbb{R}^M \to \mathbb{R}^M$ ist global Lipschitz-stetig, da $f : \mathbb{R} \to \mathbb{R}$ Lipschitz-stetig ist.

Das System (2-56) ist äquivalent zu

$$\begin{aligned} c'(t) &= B^{-1}(G(c(t)) - Ac(t)), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-57}$$

Da die rechte Seite von (2-57) Lipschitz-stetig in c und stetig in t ist, existiert nach dem Satz von Picard und Lindelöf die Lösung c auf $[0, T]$. Somit ist das Galerkin-Problem (2-55) eindeutig lösbar.

Volldiskretes Problem

Wir diskretisieren nun (2-56) mit dem ϑ-Verfahren. Sei $\Delta t = \frac{T}{N}$, $\Omega_{T_h} = \{t_j = j\Delta t \,|\, j = 0, \dots, N\}$ und sei c^j die Approximation für $c(t_j)$, $j = 0, \dots, N$. Dies liefert

$$\begin{aligned} B\frac{c^{j+1} - c^j}{\Delta t} &= \vartheta(-Ac^{j+1} + G(c^{j+1})) + (1-\vartheta)(-Ac^j + G(c^j)), \quad j = 0, \dots, N-1, \\ c^0 &\in \mathbb{R}^M \text{ gegeben.} \end{aligned}$$

Letztere Iteration ist gleichbedeutend mit

$$\begin{aligned} (B + \Delta t\vartheta A)c^{j+1} - \Delta t\vartheta G(c^{j+1}) &= Bc^j + \Delta t(1-\vartheta)(-Ac^j + G(c^j)), \\ &\quad j = 0, \dots, N-1, \\ c^0 &\in \mathbb{R}^M \text{ gegeben.} \end{aligned} \tag{2-58}$$

© Johannes Schropp 22. Mai 2024

(2-58) hat die Operatorform

$$T(c^{j+1}) = r^j, \quad j = 0, \ldots, N-1 \tag{2-59}$$

mit

$$\begin{aligned} T(a) &= (B + \Delta t\vartheta A)a - \Delta t\vartheta G(a), \\ r^j &= Bc^j + \Delta t(1-\vartheta)(-Ac^j + G(c^j)). \end{aligned}$$

Löst man (2-59) mit dem Newton-Verfahren, so wird die Linearisierung

$$DT(a) = (B + \Delta t\vartheta A) - \Delta t\vartheta DG(a)$$

benötigt. Wie im stationären Fall gilt dabei

$$\frac{\partial G_i}{\partial a_k}(a) = \int_\Omega f'\big(\sum_{j=1}^M a_j v_j\big) v_k \cdot v_i \,\mathrm{d}x, \quad 1 \le i,k \le M.$$

Somit zerstören die nichtdiagonalen Matrizen B und $DG(a)$ eine eventuell vorhandene Struktur von A. Eine Abhilfe verspricht jetzt die schon aus dem stationären Fall bekannte Technik des Mass-Lumping.

Auf einer Triangulierung $\Omega_{T_h} = \{e_1, \ldots, e_s\}$ von Ω definieren wir wieder den Ansatzraum der Lagrangeschen Finiten Elemente

$$V_h = \{v \in C(\Omega_{T_h}) \,|\, v|_{e_k} \in P^r(e_k),\ k = 1, \ldots, s, \quad v|_{\partial\Omega_{T_h}} = 0\}.$$

Es sei ferner $V_h = \mathrm{span}\{v_1, \ldots, v_M\}$. Durch $v_1, \ldots, v_M$ und die Knoten $P_1, \ldots, P_M$ sei eine nodale Basis von V_h gegeben, d.h.

$$v_i(P_j) = \delta_{ij}, \quad 1 \le i,j \le M.$$

Wir benutzen nun wieder die knotenorientierte Quadraturformel aus dem stationären Fall. Das Vektorfeld G mit $G_i(a) = \int_\Omega f\big(\sum_{j=1}^M a_j v_j\big) v_i \,\mathrm{d}x$, $1 \le i \le M$ wird dann durch

$$\sum_{k=1}^M w_k \Big(f\big(\sum_{j=1}^M a_j v_j\big) v_i\Big)(P_k) = w_i f(a_i) =: \widetilde{G}_i(a), \quad i = 1, \ldots, M$$

ersetzt. Ferner wird die Matrix $B = (b_{ij})_{1 \le i,j \le M}$, $b_{ij} := \int_\Omega v_i v_j \,\mathrm{d}x$, $1 \le i,j \le M$ ersetzt durch

$$\sum_{k=1}^M w_k (v_j v_i)(P_k) \overset{v_i(P_k)=\delta_{ik}}{=} w_i v_j(P_i) = w_i \delta_{ji},$$

d.h. B wird durch $\widetilde{B} = \mathrm{diag}(w_i, i = 1, \ldots, M)$ ersetzt.

© Johannes Schropp 22. Mai 2024

Des Weiteren werden nur solche knotenorientierte Quadraturformeln verwendet, die $w_i > 0$, $i = 1, \ldots, M$ sicherstellen, was im Falle linearer und quadratischer Finiter Elemente immer erfüllt ist. Benutzt man den Mass-Lumping-Prozess, so lautet das ϑ-Verfahren

$$\begin{aligned}(\widetilde{B} + \Delta t\vartheta A)c^{j+1} - \Delta t\vartheta\widetilde{G}(c^{j+1}) &= \widetilde{B}c^j + \Delta t(1-\vartheta)(-Ac^j + \widetilde{G}(c^j)),\\ &\quad j = 0, \ldots, N-1,\\ c^0 &\in \mathbb{R}^M \text{ gegeben.}\end{aligned} \tag{2-60}$$

Die Formelzeile (2-60) hat die Gestalt

$$\widetilde{T}(c^{j+1}) = \widetilde{r}^j, \quad j = 0, \ldots, N-1 \tag{2-61}$$

mit

$$\begin{aligned}\widetilde{T}(a) &= (\widetilde{B} + \Delta t\vartheta A)a - \Delta t\vartheta\widetilde{G}(a),\\ \widetilde{r}^j &= \widetilde{B}c^j + \Delta t(1-\vartheta)(-Ac^j + \widetilde{G}(c^j)).\end{aligned}$$

Die Linearisierung lautet

$$D\widetilde{T}(a) = (\widetilde{B} + \Delta t\vartheta A) - \Delta t\vartheta D\widetilde{G}(a)$$

mit den Diagonalmatrizen

$$\begin{aligned}\widetilde{B} &= \operatorname{diag}(w_i, i = 1, \ldots, M),\\ D\widetilde{G}(a) &= \operatorname{diag}(w_i f_i(a), i = 1, \ldots, M).\end{aligned}$$

Eventuell in A vorhandene Strukturen werden nun erhalten. Ferner gilt

$$D\widetilde{T}(a) = \widetilde{B} + O(\Delta t),$$

so dass für hinreichend kleine $\Delta t > 0$ die Matrix $D\widetilde{T}(a)$ immer invertierbar ist.

© Johannes Schropp 22. Mai 2024