

Höhere Numerische Mathematik II
Evolutionsgleichungen:
Hyperbolische und parabolische Probleme

Prof. Dr. P. Spellucci

Sommersemester 2006

Inhaltsverzeichnis

1	Numerik hyperbolischer Differentialgleichungen	5
1.1	Charakteristiken hyperbolischer Differentialgleichungen	5
1.2	Numerische Charakteristikenverfahren ERG	15
1.3	Differenzenapproximationen	23
1.3.1	Hyperbolische DGLen zweiter Ordnung	23
1.3.2	Numerische Verfahren für hyperbolische Systeme erster Ordnung . .	42
1.4	Nichtlineare hyperbolische Erhaltungsgleichungen	50
1.4.1	Beispiele	50
1.4.2	Theorie	51
1.4.3	Numerische Verfahren	56
2	Parabolische Rand-Anfangswertprobleme	69
2.1	Theoretische Grundlagen	69
2.2	Differenzenapproximationen für parabolische Gleichungen	75
2.2.1	Der räumlich eindimensionale Fall	75
2.2.2	Räumlich mehrdimensionale Probleme und damit verbundene Schwierigkeiten	85
2.2.3	Ein nichtlineares parabolisches Problem ERG	88
2.3	Galerkin-Verfahren für Randanfangswertaufgaben	91
2.4	Die Lösung der eindimensionalen Wärmeleitungsgleichung mit dem Galerkinansatz.	93
2.5	Die Rothe-Methode	104

3 Die Stabilitätstheorie von Lax und Richtmyer ERG	105
3.1 Einige funktionalanalytische Grundlagen	106
3.2 Abstrakte Differenzenverfahren, Lax-Richtmyer-Theorie	116
3.3 Mehrschrittverfahren	131
3.4 Ergänzungen zur Lax-Richtmyer-Theorie	135
3.5 Kriterien für die Stabilität von Differenzenverfahren	140

Kapitel 1

Numerik hyperbolischer Differentialgleichungen

1.1 Charakteristiken hyperbolischer Differentialgleichungen

Wir beschäftigen wir uns hier mit folgenden speziellen Differentialgleichungstypen:

- **Quasilinearen (hauptsächlich semilinearen) Differentialgleichungen zweiter Ordnung für eine skalare Funktion**

$$\begin{aligned} u : G \subset \mathbb{R}^n \quad (n = 2, 3) &\rightarrow \mathbb{R} \\ \sum_{i,k=1}^n \tilde{a}_{i,k}(x, p(x)) \partial_{i,k}^2 u(x) + f(x, p(x)) &= 0 \quad x \in G \\ (\tilde{a}_{i,k} = \tilde{a}_{k,i}) & \\ \text{mit } x \in G, \quad p(x) \stackrel{\text{def}}{=} (u, \partial_1 u, \dots, \partial_n u)(x) &\in \mathbb{R}^{n+1} \end{aligned} \tag{1.1}$$

- **Quasilinearen Systemen erster Ordnung für eine vektorwertige Funktion**

$$\begin{aligned} u : G \subset \mathbb{R}^2 &\rightarrow \mathbb{R}^n \\ \partial_2 u(x, y) - A(x, y, u(x, y)) \partial_1 u(x, y) + h(x, y, u(x, y)) &= 0 \quad (x, y) \in G \subset \mathbb{R}^2 \end{aligned} \tag{1.2}$$

In beiden Fällen kommen zu der Differentialgleichung noch die angemessenen Rand- bzw. Anfangswerte hinzu, die die Aufgabe zu einer wohlgestellten machen.

Definition 1.1 Die Differentialgleichung zweiter Ordnung (1.1) heißt bezüglich einer festen Funktion $u \in C^2(G)$ in $x \in G$

hyperbolisch, wenn alle Eigenwerte von \tilde{A} sind von null verschieden sind und $n - 1$ von ihnen das gleiche Vorzeichen haben

□

Wir werden nur solche Fälle diskutieren, **in denen der Typ** sich für $x \in G$, $u \in C^2(G)$ **nicht ändert.**

Definition 1.2 Das quasilineare System (1.2) heißt bezüglich einer festen Funktion $u \in C^1(G) \rightarrow \mathbb{R}^n$ in $(x, y) \in G$

hyperbolisch, falls A n **reelle linear unabhängige** Eigenvektoren besitzt.

□

Bemerkung 1.1 Man kann natürlich eine partielle DGL zweiter Ordnung (1.1) durch Substitution in ein System erster Ordnung überführen: Setzt man

$$\begin{aligned} v_1 &\stackrel{\text{def}}{=} u \\ v_{i+1} &\stackrel{\text{def}}{=} \partial_i u \quad i = 1, \dots, n \end{aligned}$$

so ergibt sich aus (1.1) mit $v \stackrel{\text{def}}{=} (v_1, \dots, v_{n+1}) \in \mathbb{R}^{n+1}$

$$\sum_{i=1}^n \sum_{k=1}^n a_{ik} \partial_k v_{i+1} + f(\cdot, v) = 0$$

$$v_{i+1} - \partial_i v_1 = 0 \quad i = 1, \dots, n$$

$$\text{und} \quad \partial_i v_{k+1} = \partial_k v_{i+1} \quad i, k = 1, \dots, n$$

Für $n = 2$ heißt dies ausgeschrieben

$$\begin{aligned} a_{11} \partial_1 v_2 + a_{12} \partial_2 v_2 + a_{21} \partial_1 v_3 + a_{22} \partial_2 v_3 &= -f \\ \partial_1 v_1 &= v_2 \\ \partial_2 v_1 &= v_3 \end{aligned}$$

Ferner muß gelten

$$\partial_1 v_3 - \partial_2 v_2 = 0$$

Addieren wir dies etwa zur dritten Gleichung, so ergibt sich

$$\begin{bmatrix} 0 & a_{11} & a_{21} \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \partial_1 v + \begin{bmatrix} 0 & a_{12} & a_{22} \\ 0 & 0 & 0 \\ 1 & -1 & 0 \end{bmatrix} \partial_2 v = \begin{bmatrix} -f \\ v_2 \\ v_3 \end{bmatrix}$$

oder (für $a_{11} \neq 0$)

$$\partial_1 v = - \begin{bmatrix} 0 & 1 & 0 \\ a_{11}^{-1} & 0 & -a_{21} \cdot a_{11}^{-1} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & a_{12} & a_{22} \\ 0 & 0 & 0 \\ 1 & -1 & 0 \end{bmatrix} \partial_2 v + \begin{bmatrix} v_2 \\ -a_{11}^{-1} f - a_{21} a_{11}^{-1} v_3 \\ v_3 \end{bmatrix} \quad (1.3)$$

$$\partial_1 v = - \frac{1}{a_{11}} \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ -a_{21} & a_{12} + a_{21} & a_{22} \\ a_{11} & -a_{11} & 0 \end{bmatrix}}_A \partial_2 v + \begin{bmatrix} v_2 \\ -a_{11}^{-1} f - a_{21} a_{11}^{-1} v_3 \\ v_3 \end{bmatrix} \quad (1.4)$$

Ist also die Gleichung zweiter Ordnung hyperbolisch im Sinne der Definition 1.1, dann ist das System erster Ordnung (1.3) auch hyperbolisch im Sinne der Definition 1.2. Die Eigenwerte der Matrix A sind bis auf den Faktor $1/a_{11}$ $\lambda = 0$ und die Wurzeln von $\lambda^2 - \underbrace{(a_{12} + a_{21})}_{2a_{12}} \lambda + a_{11} a_{22} = 0$ also

$$\lambda_{2,3} = a_{12} \pm \sqrt{a_{12}^2 - a_{11} a_{22}} = a_{12} \pm \sqrt{-\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}} \quad .$$

□

Im Folgenden werden wir uns ausführlich beschäftigen mit hyperbolischen Systemen erster Ordnung und hyperbolischen Einzeldifferentialgleichungen zweiter Ordnung.

Grundlegend auch für das Verständnis der zugehörigen numerischen Verfahren ist in diesem Zusammenhang der Begriff der Charakteristik.

Wir betrachten zunächst als ein einfaches, jedoch nicht völlig triviales Beispiel

Beispiel 1.1 $u_t = \frac{\partial}{\partial x}(\varphi(u)), \quad \varphi \in C^\infty(\mathbb{R}),$

$$0 < \varphi_0 \leq -\varphi'(z) \leq \varphi_1 \quad \forall z \in \mathbb{R}$$

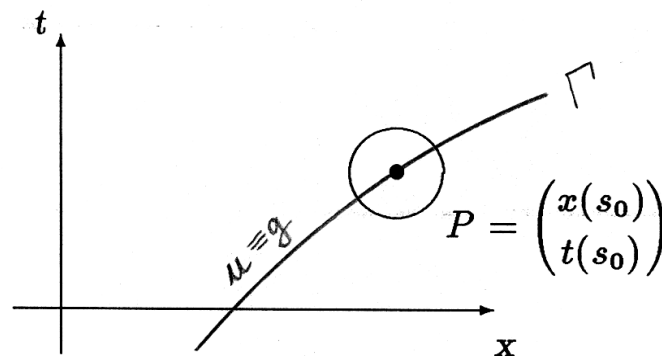


Abbildung 1.1

Γ sei eine beliebige glatte C^∞ -Kurve der (x, t) -Ebene und g eine auf einer Umgebung von Γ definierte, beliebig oft differenzierbare Funktion. Wir fragen nun nach der Möglichkeit, durch die DGL und die Anfangsvorgabe

$$u(x(s), t(s)) \equiv g(x(s), t(s))$$

mit

$$\Gamma : \begin{pmatrix} x \\ t \end{pmatrix} = \begin{pmatrix} x(s) \\ t(s) \end{pmatrix}$$

die Funktion u eindeutig (zumindest lokal) zu bestimmen.

Differentiation nach s liefert längs Γ

$$\frac{d}{ds} u(x(s), t(s)) = u_x \dot{x} + u_t \dot{t} = g_x \dot{x} + g_t \dot{t}.$$

Nach Voraussetzung über Γ ist $\begin{pmatrix} \dot{x} \\ \dot{t} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ ("glatte Kurve").

Aus der DGL haben wir

$$u_t - \varphi'(u)u_x = 0$$

d.h.

$$\begin{bmatrix} 1 & -\varphi'(u) \\ \dot{t}(s) & \dot{x}(s) \end{bmatrix} \begin{bmatrix} u_t \\ u_x \end{bmatrix} = \begin{bmatrix} 0 \\ g_x \dot{x} + g_t \dot{t} \end{bmatrix}.$$

Falls

$$\dot{x}(s) + \dot{t}(s) \overbrace{\varphi'(u(x(s), t(s)))}^{g(x(s), t(s))} \neq 0$$

zumindest für alle s in einer Umgebung von s_0 , kann man auflösen:

$$\begin{bmatrix} u_t \\ u_x \end{bmatrix} = \frac{1}{\dot{x} + \dot{t}\varphi'(g)} \begin{bmatrix} \dot{x} & \varphi'(g) \\ -\dot{t} & 1 \end{bmatrix} \begin{bmatrix} 0 \\ g_x \dot{x} + g_t \dot{t} \end{bmatrix} \quad \begin{pmatrix} x \\ t \end{pmatrix} = \begin{pmatrix} x(s) \\ t(s) \end{pmatrix}$$

und die höheren Ableitungen u_{tt}, u_{tx}, u_{xx} usw. lassen sich (an der Stelle $(x(s), t(s))$) daraus durch weitere Differentiationen erhalten. Mit allen erhaltenen Ableitungen kann man dann die formale Taylorreihe von u in einer Umgebung von $\begin{pmatrix} x(s_0) \\ t(s_0) \end{pmatrix}$ angeben (deren Konvergenz bliebe natürlich noch zu untersuchen).

Ist jedoch

$$\dot{x}(s) + \dot{t}(s)\varphi'(g(x(s), y(s))) = 0,$$

dann ist die eindeutige Auflösbarkeit nicht gegeben.

Da wegen $\varphi' \neq 0$ und $\begin{pmatrix} \dot{x} \\ \dot{t} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ \dot{t} nie null sein kann, kann man \dot{t} auf 1 normieren (d.h. t wird Parameter der Kurve). Die durch

$$\begin{aligned} \dot{t}(s) &= 1 \\ \dot{x}(s) &= -\varphi'(u(x(s), t(s))) \\ t(s_0) &= t_0 \\ x(s_0) &= x_0 \end{aligned} \tag{1.5}$$

eindeutig bestimmte Kurve nennt man eine **Charakteristik der DGL** (zur Funktion u). Offensichtlich können sich bei gegebenem u zwei Charakteristiken der DGL nicht schneiden. Bei gegebenem u gibt es eine Charakteristiken-Schar (Lösungsschar der gew. DGL (1.5)). Die durch $\begin{pmatrix} \dot{x} \\ \dot{t} \end{pmatrix}(s)$ im Punkte $\begin{pmatrix} x \\ t \end{pmatrix}(s)$ gegebene Richtung der Charakteristik nennt man **charakteristische Richtung** der DGL in diesem Punkt (bei gegebenem u).

Entsprechend lassen sich die Überlegungen anstellen bei Einzeldifferentialgleichungen zweiter Ordnung und bei quasilinearen Systemen erster Ordnung.

Es ergibt sich, daß für Einzeldifferentialgleichungen zweiter Ordnung (bei hinreichenden Regularitätsvoraussetzungen an die Koeffizienten) im Falle **zweier** unabhängiger Veränderlicher ($n = 2$) im

hyperbolischen Fall zwei Scharen von Charakteristiken

existieren. Bei drei unabhängigen Veränderlichen treten dann an die Stelle der charakteristischen Kurven charakteristische Flächen und zwar 3 Scharen im hyperbolischen Fall. Bei einem hyperbolischen System der Ordnung 1 für n gesuchte Funktionen gibt es unter einschränkenden Voraussetzungen ebenfalls stets n Charakteristiken. Genauer gilt

Satz 1.1 Gegeben sei das quasilineare System mit $A \in \mathbb{R}^{n \times n}$

$$\partial_2 u(x, y) = A(x, y, u(x, y)) \partial_1 u(x, y) + g(x, y, u(x, y)) \quad (x, y) \in G \subset \mathbb{R}^2$$

G sei einfach zusammenhängend, offen. Ferner gelte

$$A \in C^1(G \times \mathbb{R}^n), \quad g \in C^1(G \times \mathbb{R}^n),$$

Für alle $(x, y, z) \in G \times \mathbb{R}^n$ habe A n verschiedene reelle Eigenwerte. Dann laufen durch jeden Punkt von G genau n Charakteristiken. Es gibt n Scharen von Charakteristiken, die den n reellen Eigenwerten von A eineindeutig zugeordnet sind. Zwei Charakteristiken der gleichen Schar schneiden sich nie. Keine Charakteristik berührt die x -Achse. Jede Charakteristik schneidet die x -Achse höchstens einmal. \square

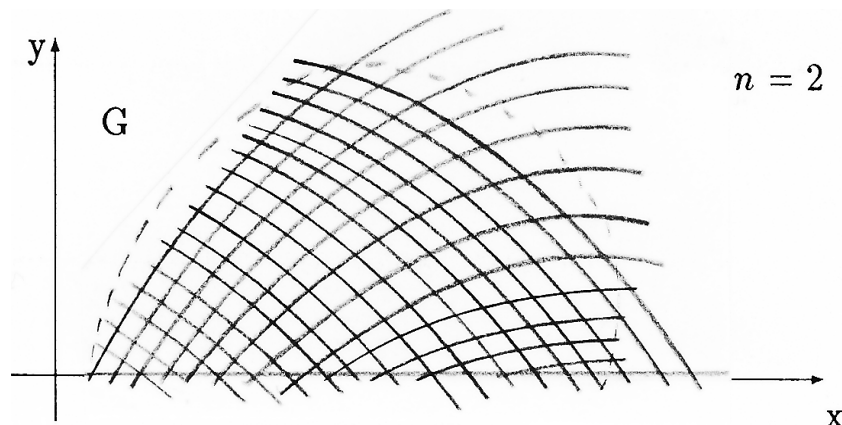


Abbildung 1.2

Zur Bedeutung der Charakteristiken betrachten wir folgendes Beispiel:

Beispiel 1.2

$$\partial_2 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \partial_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$$

wobei wir annehmen wollen, daß $g_i = g_i(x, y)$ und die feste Matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ reell diagonalisierbar ist mit Eigenwerten $\lambda_1 < \lambda_2$, $\lambda_i \neq 0$.

Mittels einer regulären Transformationsmatrix $T \in \mathbb{R}^{2 \times 2}$ wird also

$$TAT^{-1} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

und mit

$$Tu =: v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad Tg =: r$$

$$\partial_2 v_1 = \lambda_1 \partial_1 v_1 + r_1(x, y)$$

$$\partial_2 v_2 = \lambda_2 \partial_1 v_2 + r_2(x, y)$$

Wir haben somit ein zerfallendes System von 2 Gleichungen der Ordnung 1, auf das die Überlegungen aus Beispiel 1.1 entsprechend anzuwenden sind. Die beiden Charakteristiken-scharen sind also

$$x + \lambda_1 y = \text{const}$$

$$x + \lambda_2 y = \text{const}$$

oder in Parameterdarstellung

$$\left. \begin{matrix} (c_1 - \lambda_1 t, t) \\ (c_2 - \lambda_2 t, t) \end{matrix} \right\} t \in \mathbb{R}, \quad \left. \begin{matrix} c_1 \\ c_2 \end{matrix} \right\} = \text{Scharparameter} \left\{ \begin{matrix} \text{erste} \\ \text{zweite} \end{matrix} \right\} \text{Charakteristikenschar}$$

Nun betrachten wir zur DGL die Anfangsvorgaben

$$v_i(x, y) = \varphi_i(x, y) \quad i = 1, 2, \quad (x, y) \in \Gamma, \quad \Gamma = \{(x, y) : \alpha x + \beta y = 0\}$$

(entsprechend **einer** Anfangsvorgabe für die Vektorfunktion u auf Γ).

Der Einfachheit halber sei $\alpha \neq 0$, $\beta \neq 0$. Aufgrund der DGL gilt längs der Charakteristiken

$$\frac{d}{dt} v_i(c_i - \lambda_i t, t) = \partial_2 v_i - \lambda_i \partial_1 v_i = r_i(c_i - \lambda_i t, t)$$

d.h.

$$v_i(c_i - \lambda_i t, t) = \eta_i + \int_0^t r_i(c_i - \lambda_i \tau, \tau) d\tau \quad i = 1, 2$$

η_i beliebige Integrationskonstante.

Wir berücksichtigen nun die Anfangsbedingungen und schreiben Γ parametrisiert in der Form

$$\Gamma : \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} (-\beta/\alpha)t \\ t \end{pmatrix} \quad t \in \mathbb{R}$$

Längs Γ gilt:

$$\frac{d}{dt} v_i((-\beta/\alpha)t, t) = \partial_2 v_i - \frac{\beta}{\alpha} \partial_1 v_i = \frac{d}{dt} \varphi_i(-(\beta/\alpha)t, t)$$

Wir haben nun folgende Fälle zu unterscheiden:

1. Die Geraden $(c_i - \lambda_i t, t)$ und $((-\beta/\alpha)t, t)$ schneiden sich in genau einem Punkt, d.h. $\lambda_i \neq \frac{\beta}{\alpha}$

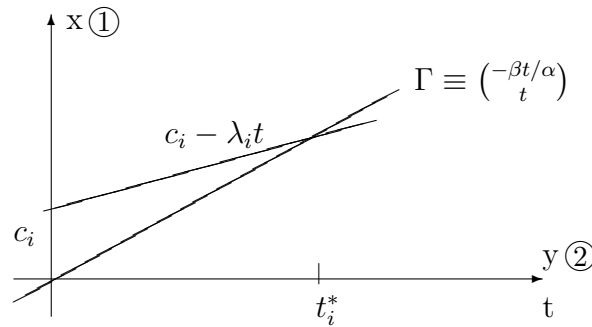


Abbildung 1.3

Dies ergibt

$$\begin{aligned} c_i - \lambda_i t_i^* &= -(\beta/\alpha)t_i^* \Rightarrow t_i^* = \frac{\alpha c_i}{\alpha \lambda_i - \beta} \\ \Rightarrow \\ v_i(c_i - \lambda_i t_i^*, t_i^*) &= \eta_i + \int_0^{t_i^*} r_i(c_i - \lambda_i \tau, \tau) d\tau = \varphi_i(-\frac{\beta}{\alpha} t_i^*, t_i^*) \\ \Rightarrow \\ \eta_i &= \varphi_i(-\frac{\beta}{\alpha} t_i^*, t_i^*) - \int_0^{t_i^*} r_i(c_i - \lambda_i \tau, \tau) d\tau = \eta_i(c_i). \end{aligned}$$

Die Integrationskonstante η_i ist somit eine eindeutige Funktion von c_i geworden und v_i **eindeutig gegeben durch**

$$\begin{aligned} v_i(x, y) &= \eta_i(x + \lambda_i y) + \int_0^y r_i(x + \lambda_i y - \lambda_i \tau, \tau) d\tau \\ (y \stackrel{def}{=} t \quad x \stackrel{def}{=} c - \lambda_i y \Rightarrow c &= x + \lambda_i y) \end{aligned}$$

2. Die Geraden $(c_i - \lambda_i t, t)$, $((-\beta/\alpha)t, t)$ haben keinen Schnittpunkt. D.h.

$$c_i \neq 0, \quad \lambda_i = \beta/\alpha$$

Dann hat die i^{te} Gleichung des entkoppelten Systems unendlich viele Lösungen, wenn man die DGL nur für das Gebiet oberhalb oder unterhalb Γ fordert.

Soll die DGL auch auf Γ gelten, muß noch

$$\frac{d}{dt}\varphi_i(-(\beta/\alpha)t, t) = r_i(-(\beta/\alpha)t, t) \quad \text{Verträglichkeitsbedingung}$$

gelten. Im Falle der Gültigkeit dieser Verträglichkeitsbedingung gibt es unendlich viele Lösungen

$$v_i(x, y; \mu) = \underbrace{\mu(\alpha x + \beta y)}_{\equiv 0 \text{ auf } \Gamma} + \varphi_i(x, y)$$

mit $\mu \in \mathbb{R}$ beliebig (Flächenschar).

3. $c_i = 0$, $\lambda_i = \beta/\alpha$, d.h. die Geraden fallen zusammen.

Die i^{te} Gleichung hat keine oder unendlich viele Lösungen, je nachdem, ob die Verträglichkeitsbedingung gilt oder nicht.

Unser hyperbolisches System hat also genau dann eine eindeutige Lösung, **wenn Γ mit keiner der Charakteristiken zusammenfällt.**

Entsprechendes gilt auch im allgemeinen Fall, wie er durch Satz 1.1 beschrieben ist.

Bei dem in diesem Beispiel vorliegenden einfachen Fall gelangen wir zu einer geschlossenen Lösungsformel für u :

$$\begin{aligned} u &= T^{-1}v, \quad v = (v_1, v_2) \\ v_i(x, y) &= \eta_i(x + \lambda_i y) + \int_0^y r_i(x + \lambda_i y - \lambda_i \tau, \tau) d\tau \\ \eta_i(x + \lambda_i y) &= \varphi_i(-\frac{\beta}{\alpha}t_i^*, t_i^*) - \int_0^{t_i^*} r_i(x + \lambda_i y - \lambda_i \tau, \tau) d\tau \end{aligned}$$

wobei t_i^* der Schnittpunktparameter der Geraden

$$(x + \lambda_i y - \lambda_i t, t) \quad \text{und} \quad \Gamma$$

ist: d.h.

$$v_i(x, y) = \varphi_i(-\frac{\beta}{\alpha}t_i^*, t_i^*) + \int_{t_i^*}^y r_i(x + \lambda_i y - \lambda_i \tau, \tau) d\tau$$

d.h. $v_i(x, y)$ ist die Summe aus dem Anfangswert auf dem Schnittpunkt der i^{ten} Charakteristik durch (x, y) mit Γ und dem Wegintegral über die Inhomogenität r_i von diesem

Schnittpunkt längs der i^{ten} Charakteristik bis zum Punkt (x, y) . Dies bedeutet z.B., daß eine Unstetigkeit von φ_i (längs Γ) längs der i^{ten} Charakteristik weitertransportiert wird.

Im vorliegenden Beispielfall ist jedes u_j eine Linearkombination der v_i und die v_i bestimmen sich aus Information längs der i^{ten} Charakteristik durch (x, y) . Dies bedeutet, daß die Lösung der DGL im Punkt (x, y) konstruiert wird aus der Information längs der Charakteristiken durch (x, y) bis zur Kurve Γ , die die Anfangswerte φ trägt. Dies erklärt die folgenden Begriffsbildungen:

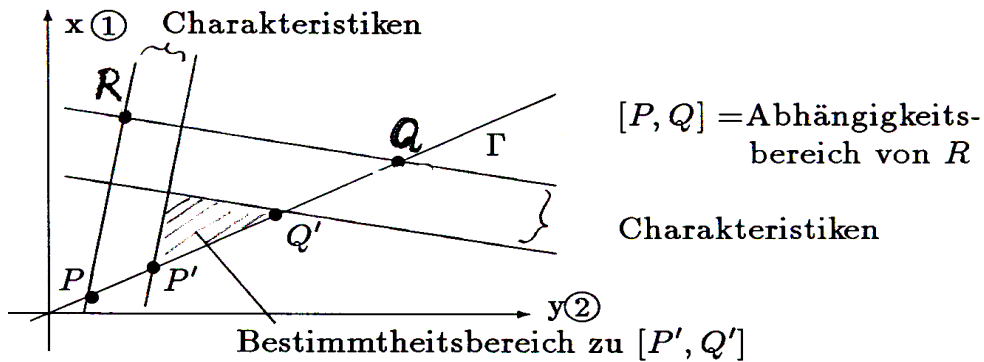


Abbildung 1.4

Definition 1.3 : ($n = 2$): Seien $R = (x, y)$ ein Punkt in G und P, Q die beiden Punkte auf Γ (Kurve mit Vorgabe der Anfangswerte, die mit keiner Charakteristik zusammenfällt), die durch den Schnitt der beiden Charakteristiken durch (x, y) mit Γ gebildet werden. Dann heißt die Strecke (P, Q) auf Γ der Abhängigkeitsbereich des Punktes R und der Bereich, der aus R, P, Q mit den Charakteristikenstücken $(R, P), (R, Q)$ und dem Stück (P, Q) von Γ als Rand gebildet wird, der Bestimmtheitsbereich der Strecke (P, Q) von Γ . Als Einflußbereich eines Punktes P auf Γ bezeichnet man die Menge aller Punkte, deren Abhängigkeitsbereich den Punkt P enthält. □

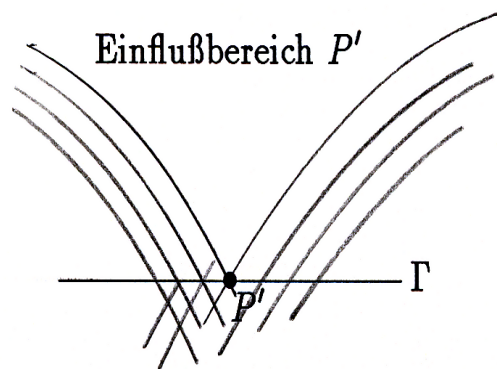


Abbildung 1.7 Charakteristiken bei variablen Koeffizienten

Bei der numerischen Lösung von hyperbolischen Gleichungen muß man die Bestimmtheitsbereiche der Strecken, die die jeweiligen Anfangswerte tragen, ganz wesentlich berücksichtigen, wie wir noch sehen werden.

Bei drei unabhängigen Veränderlichen treten an die Stelle der charakteristischen Kurven und der Kurve der Anfangswerte entsprechend charakteristische Flächen und eine Fläche mit Anfangsvorgaben. Hat man zwei unabhängige Veränderliche und ein hyperbolisches System für n Funktionen, muß man entsprechend alle n Charakteristiken berücksichtigen:

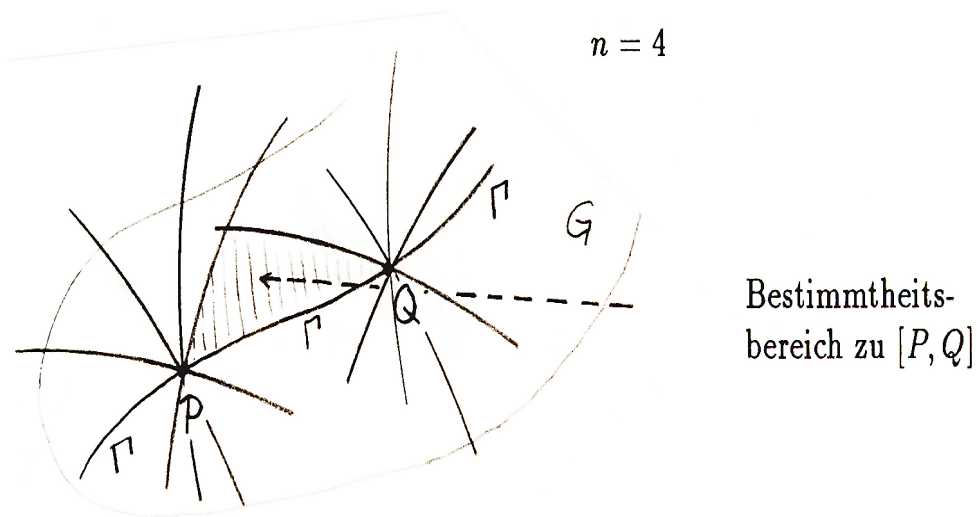


Abbildung 1.6 $n=4$ Dimension des hyperbolischen Systems

1.2 Numerische Charakteristikenverfahren ERG

In den folgenden Überlegungen wird von einem hyperbolischen System erster Ordnung für zwei gesuchte Funktionen u_1, u_2 ausgegangen:

$$\partial_2 u = A(x, y, u) \partial_1 u + g(x, y, u) \quad (x, y) \in G$$

$A \in \mathbb{R}^{2 \times 2}$ habe für alle $(x, y, z) \in G \times \mathbb{R}^2$ stets zwei reelle verschiedene Eigenwerte. Weil die Eigenwerte einer Matrix stetige Funktionen der Matrixelemente sind und im Fall eines einfachen Eigenwerts sogar differenzierbar von ihnen abhängen, wie auch ein geeignet normierter Eigenvektor, gibt es dann eine differenzierbare invertierbare Matrix T sodaß (mit $T = T(x, y, z)$)

$$T^{-1} A(x, y, z) T = \text{diag} (\lambda_1(x, y, z), \lambda_2(x, y, z)) \stackrel{\text{def}}{=} \Lambda \quad z = u$$

und

$$T^{-1} \partial_2 u = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} T^{-1} \partial_1 u + \tilde{g}, \quad T^{-1} g \stackrel{\text{def}}{=} \tilde{g}.$$

Die Charakteristiken des Systems ¹ lassen sich in Parameterdarstellung schreiben als

$$\begin{aligned} x &= \varphi_\mu(t), & y &= t \\ \varphi'_\mu(t) + \lambda_\mu(\varphi_\mu(t), t, u(\varphi_\mu(t), t)) &= 0 \\ \mu &= 1, 2. \end{aligned}$$

Der Einfachheit halber setzen wir voraus:

Γ (Kurve mit Vorgabe der Anfangswerte) ist ein nichtleeres offenes Intervall auf der x -Achse)

$$\Gamma = \{(x, y) : a < x < b, \quad y = 0\}$$

\bar{G} = Bestimmtheitsbereich von $\bar{\Gamma}$.

Dies bedeutet, daß **jede** Charakteristik durch einen Punkt von $G \cap \Gamma$ schneidet. Man kann somit die Abszisse des Schnittpunktes $(s, 0)$ als Scharparameter der Charakteristiken einführen:

$$\begin{aligned} \varphi'_\mu(t; s) + \lambda_\mu(\varphi_\mu(t; s), t, u(\varphi_\mu(t; s), t)) &= 0 \\ \varphi_\mu(0; s) &= s \quad a < s < b, \quad \mu = 1, 2 \end{aligned}$$

Die Lösungen dieser Differentialgleichungen hängen differenzierbar vom Scharparameter s ab. Durch jeden Punkt (x, y) von G laufen zwei Charakteristiken, mit den zugehörigen Scharparametern ²

$$s_1 \stackrel{\text{def}}{=} p_1(x, y), \quad s_2 \stackrel{\text{def}}{=} p_2(x, y)$$

¹Falls A unabh. von u ist, kann man mit $v \stackrel{\text{def}}{=} T^{-1}u$ noch weiter vereinfachen zu $\partial_2 v = \Lambda \partial_1 v - (\partial_2 T^{-1} - \Lambda \partial_1 T^{-1}) T v + \tilde{g}$.

²Die angegebene feste Numerierung bezieht sich auf die feste Numerierung $\lambda_1 < \lambda_2$

Es ist nach Definition

$$s = p_\mu(\varphi_\mu(t; s), t) \quad \mu = 1, 2 \quad a < s < b$$

p_1 und p_2 sind Lösungen der AWA

$$\begin{aligned} \partial_2 p_\mu(x, y) &= \lambda_\mu(x, y, u(x, y)) \partial_1 p_\mu(x, y) \\ p_\mu(x, 0) &= x \quad a < x < b \end{aligned}$$

wegen

$$0 = \frac{d}{dt}(s) = \partial_1 p_\mu(\dots) \varphi'_\mu(t; s) + \partial_2 p_\mu, \quad \varphi'_\mu = -\lambda_\mu.$$

Mit Hilfe dieser Projektionen gelangt man zu einem neuen Koordinatensystem in G , dem sogenannten **charakteristischen Koordinatensystem** mit den Koordinaten

$$\begin{aligned} \sigma &= \frac{1}{2}(p_2(x, y) + p_1(x, y)) \\ \tau &= \frac{1}{2}(p_2(x, y) - p_1(x, y)) \end{aligned}$$

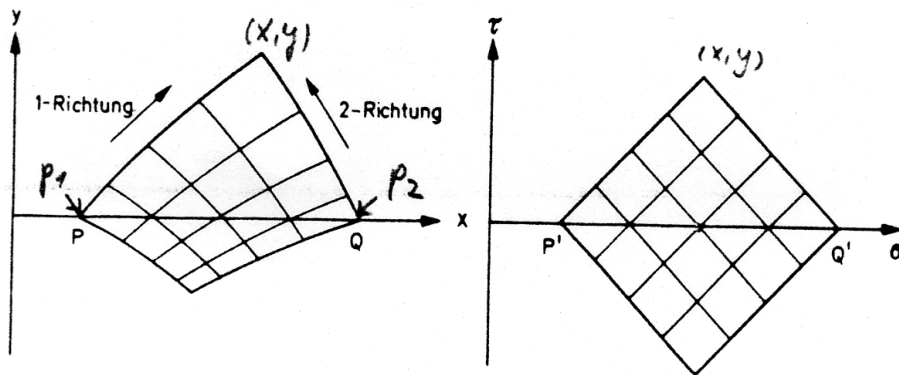


Abbildung 1.7

Die numerischen Charakteristikenverfahren bestehen nun darin, Näherungen für x, y und u auf den Gitterpunkten zu einem charakteristischen Netz

$$\{(\sigma, \tau) \mid \sigma = k \cdot h, \quad \tau = l \cdot h \quad \text{mit} \quad h > 0 \text{ fest und } l, k \in \mathbb{Z}\}$$

die in G liegen, zu bestimmen.

Das einfachste Verfahren ist ein explizites Einschrittverfahren der Ordnung 1 (entsprechend dem Euler-Verfahren zur Lösung gewöhnlicher DGLen).

Hier wird aus den Werten von (x, y, u) in den Punkten mit den charakteristischen Koordinaten $h \cdot (k - 1, l - 1)$, $h \cdot (k + 1, l - 1)$ die Information im Punkt $h \cdot (k, l)$ berechnet.

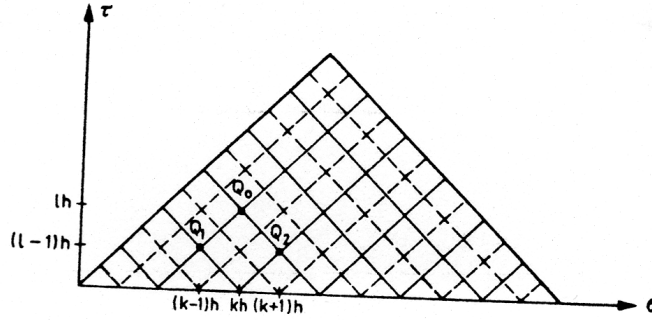


Abbildung 1.8

Seien Q_1, Q_2 die Punkte auf der Schicht $\tau = (l-1)h$ und Q_0 der Punkt auf der Schicht lh . Dann wird mit $Q_j \hat{=} (x^{(j)}, y^{(j)})$ im euklidischen Koordinatensystem

$$\begin{aligned} \frac{x^{(0)} - x^{(j)}}{y^{(0)} - y^{(j)}} &= \frac{\varphi_j(y^{(0)}) - \varphi_j(y^{(j)})}{y^{(0)} - y^{(j)}} \approx \varphi'_j(y^{(j)}) \\ &= -\lambda_j \underbrace{(\varphi_j(y^{(j)}), y^{(j)}, u(x^{(j)}, y^{(j)}))}_{x^{(j)}} \quad j = 1, 2 \end{aligned}$$

Ferner gilt

$$\begin{aligned} \frac{u_\nu(Q_0) - u_\nu(Q_j)}{y^{(0)} - y^{(j)}} &= \frac{u_\nu(\varphi_j(t_0), t_0) - u_\nu(\varphi_j(t_j), t_j)}{t_0 - t_j} \\ &\approx \partial_1 u_\nu(Q_j) \varphi'_j(t_j) + \partial_2 u_\nu(Q_j) \\ &= -\partial_1 u_\nu(Q_j) \lambda_j(Q_j, u^{(j)}) + \partial_2 u_\nu(Q_j) \end{aligned}$$

und mit

$$T^{-1}(Q_j, u^{(j)}) = (\tilde{t}_{\mu\nu}(Q_j, u^{(j)}))$$

aufgrund der DGL: $\sum_{\nu=1}^2 \tilde{t}_{\mu\nu} \partial_2 u_\nu - \lambda_\mu \sum_{\nu=1}^2 \tilde{t}_{\mu\nu} \partial_1 u_\nu = \sum_{\nu=1}^2 \tilde{t}_{\mu\nu} g_\nu$
(Für $\mu = 1$ setze $j = 1$ und substituiere $(x, y, u) = (Q_1, u^{(1)})$ entsprechend für $\mu = 2$)

$$\sum_{\nu=1}^2 \tilde{t}_{j\nu}(Q_j, u^{(j)}) \frac{u_\nu(Q_0) - u_\nu(Q_j)}{y^{(0)} - y^{(j)}} \approx \sum_{\nu=1}^2 \tilde{t}_{j\nu}(Q_j, u^{(j)}) g_\nu(Q_j, u^{(j)}) \quad j = 1, 2.$$

Dies sind nun vier lineare Gleichungen für die vier Unbekannten $x^{(0)}, y^{(0)}, u_1(Q_0), u_2(Q_0)$, die man zweckmäßig auflöst wie folgt (man ersetze \approx durch $=$)

$$\begin{bmatrix} 1 & \lambda_1^{(1)} \\ 1 & \lambda_2^{(2)} \end{bmatrix} \begin{bmatrix} x^{(0)} - x^{(1)} \\ y^{(0)} - y^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ (x^{(2)} - x^{(1)}) + \lambda_2^{(2)}(y^{(2)} - y^{(1)}) \end{bmatrix} \curvearrowright x^{(0)}, y^{(0)}$$

$$\begin{bmatrix} \tilde{t}_{11}^{(1)} & \tilde{t}_{12}^{(1)} \\ \tilde{t}_{21}^{(2)} & \tilde{t}_{22}^{(2)} \end{bmatrix} \begin{bmatrix} u_1^{(0)} - u_1^{(1)} \\ u_2^{(0)} - u_2^{(1)} \end{bmatrix} = \begin{bmatrix} (y^{(0)} - y^{(1)})(\tilde{t}_{11}^{(1)} g_1^{(1)} + \tilde{t}_{12}^{(1)} g_2^{(1)}) \\ \tilde{t}_{21}^{(2)}((y^{(0)} - y^{(2)})g_1^{(2)} + u_1^{(2)} - u_1^{(1)}) + \tilde{t}_{22}^{(2)}((y^{(0)} - y^{(2)})g_2^{(2)} + u_2^{(2)} - u_2^{(1)}) \end{bmatrix}$$

(Bezeichnungen:

$$\begin{aligned} \tilde{t}_{\nu\mu}^{(j)} &= \tilde{t}_{\nu\mu}^{(j)}(x^{(j)}, y^{(j)}, u(x^{(j)}, y^{(j)})), \\ u^{(j)} &= u(x^{(j)}, y^{(j)}) \\ g_\nu^{(j)} &= g_\nu(x^{(j)}, y^{(j)}, u(x^{(j)}, y^{(j)})) \end{aligned} \quad)$$

Für hinreichend kleines h sind die 2×2 -Matrizen dieser Systeme invertierbar, da dann

$$\begin{aligned} \lambda_2^{(2)} &\approx \lambda_2^{(1)}, & \lambda_1^{(1)} &\neq \lambda_2^{(1)} & \text{nach Vor.} \\ \tilde{t}_{21}^{(2)} &\approx \tilde{t}_{21}^{(1)}, & \tilde{t}_{22}^{(2)} &\approx \tilde{t}_{22}^{(1)}, & \tilde{T} \text{ invertierbar nach Vor.} \end{aligned}$$

Der Rechenablauf für jeden Punkt einer neuen Schicht des charakteristischen Gitters sieht also folgendermaßen aus:

1. Auswertung von $g^{(j)}, A^{(j)}, (T^{-1})^{(j)}$ und $\lambda_j^{(j)}$ für $j = 1, 2$
2. Lösung der beiden oben stehenden Gleichungssysteme.

So kann man sich von einer charakteristischen Schicht zur nächsten fortbewegen.

Selbstverständlich kann man auch zu genaueren Ansätzen gelangen, indem man die Differentialgleichungen genauer approximiert. Ersetzt man z.B. in den obigen Formeln die Argumente $x^{(j)}, y^{(j)}, u^{(j)}$ durch $(x^{(j)} + x^{(0)})/2, (y^{(j)} + y^{(0)})/2, (u^{(j)} + u^{(0)})/2$, so gelangt man zu einem Verfahren 2. Ordnung (d.h. es wird $\|u^{(j)} - u(x^{(j)}, y^{(j)})\| \leq Ch^2$).

Man hat dann allerdings ein System nichtlinearer Gleichungen in 4 Unbekannten für jeden neuen Gitterpunkt zu lösen. Dies geschieht in der Regel durch "direkte Iteration", d.h. man hat dann z.B. die oben stehenden Gleichungssysteme mehrfach zu lösen, wobei Matrizen und rechte Seite jeweils für die "alten" Werte ausgewertet werden. Diese Iteration konvergiert für genügend kleines h .

Wir diskutieren diese Vorgehensweise ausführlich am Beispiel einer hyperbolischen Einzeldifferenzialgleichung zweiter Ordnung

$$\begin{aligned} a(x, y, w)u_{xx} + b(x, y, w)u_{xy} + c(x, y, w)u_{yy} &= g(x, y, w) \\ w &\stackrel{\text{def}}{=} (u, u_x, u_y) \\ \left. \begin{aligned} b^2 - 4ac &\geq \gamma > 0 \\ |ac| &\geq \alpha > 0 \end{aligned} \right\} \forall (x, y, w) \in G \times \mathbb{R}^3 \end{aligned}$$

Längs einer Charakteristik mit dem Tangentenvektor (k'_1, k'_2) muß dann gelten:

$$\det \begin{pmatrix} a & b & c \\ k'_1 & k'_2 & 0 \\ 0 & k'_1 & k'_2 \end{pmatrix} = 0$$

$$\begin{pmatrix} k'_1 \\ k'_2 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

d.h.

$$a(k'_2)^2 - bk'_1k'_2 + c(k'_1)^2 = 0,$$

$$\begin{pmatrix} k'_1 \\ k'_2 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

und aufgrund der obigen Voraussetzung kann man z.B.

$$k'_1 \equiv 1 \quad \text{o.B.d.A.}$$

verlangen. Man erhält damit als charakteristische Richtungen im Punkt $(x, y) \in G$ die beiden Lösungen

$$\begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ -\lambda_2 \end{pmatrix}$$

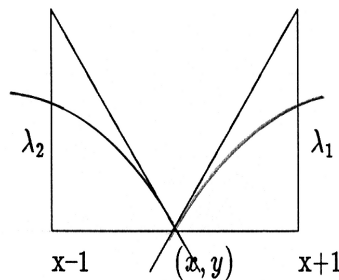


Abbildung 1.9

wobei λ_1, λ_2 die Wurzeln von

$$a\lambda^2 - b\lambda + c = 0$$

sind, d.h.

$$\lambda_{1,2} = \frac{1}{2a}(b \pm \sqrt{b^2 - 4ac})$$

G enthalte die nichtcharakteristische Anfangskurve

$$\Gamma \stackrel{\text{def}}{=} \{(x, y) : \alpha < x < \beta, \quad y = 0\}.^3$$

Zu gegebener Funktion u besitzen die Charakteristiken die Darstellung

$$\begin{pmatrix} t \\ y_j(t) \end{pmatrix} \quad \alpha < t < \beta$$

$$\dot{y}_j = (b(t, y_j(t), w_j(t)) + \theta_j \sqrt{(b^2 - 4ac)(t, y_j(t), w_j(t))}) / (2a(t, y_j(t), w_j(t)))$$

$$y_j(t_0) = 0 \quad (t_0 \text{ ist Scharparameter})$$

mit $w_j(t) \stackrel{\text{def}}{=} (u(t, y_j(t)), u_x(t, y_j(t)), u_y(t, y_j(t)))$, $\theta_1 = 1$, $\theta_2 = -1$
Längs einer solchen Charakteristik gilt, da das System

$$\begin{bmatrix} a & b & c \\ 1 & \dot{y}_j & 0 \\ 0 & 1 & \dot{y}_j \end{bmatrix} \begin{bmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{bmatrix} = \begin{bmatrix} g(t, y_j(t), w_j(t)) \\ \frac{d}{dt}u_x(t, y_j(t)) \\ \frac{d}{dt}u_y(t, y_j(t)) \end{bmatrix}$$

lösbar ist (Existenz einer eindeutigen Lösung u des AWP folgt aus den obigen Voraussetzungen), auch

$$\det \begin{bmatrix} a & g & c \\ 1 & \frac{d}{dt}u_x & 0 \\ 0 & \frac{d}{dt}u_y & \dot{y}_j \end{bmatrix} = 0$$

d.h.

$$a(t, y_j(t), w_j(t))\dot{y}_j(t) \frac{d}{dt}u_x(t, y_j(t)) - g(t, y_j(t), w_j(t))\dot{y}_j(t) + c(t, y_j(t), w_j(t)) \frac{d}{dt}u_y(t, y_j(t)) = 0, \quad j = 1, 2$$

Ferner gilt natürlich auch

$$\frac{d}{dt}u(t, y_j(t)) = u_x(t, y_j(t)) + u_y(t, y_j(t))\dot{y}_j(t) \quad j = 1, 2$$

Auf der Anfangskurve $\Gamma : \alpha < x < \beta, \quad y = 0$

seien nun die Anfangswerte

$$\begin{aligned} u(x, 0) &= \varphi_1(x) \\ u_y(x, 0) &= \varphi_2(x) \end{aligned}$$

³d.h. Γ ist der Schnitt der x -Achse mit G

vorgegeben. (Es wäre natürlich auch eine Vorgabe einer echten Linearkombination von u_x und u_y möglich. Die Vorgabe von u_x alleine ist aber längs Γ schon durch u vorhanden)

Das numerische Charakteristikenverfahren arbeitet nun wieder längs den Schichten eines charakteristischen Gitters, ausgehend von einer äquidistanten Einteilung auf Γ mit den bekannten Werten φ_1 und φ_2 , indem die Differentialgleichungen

$$\begin{aligned} \dot{y}_j &= (b + \theta_j \sqrt{(b^2 - ac)}) / (2a) & j = 1, 2 \\ a\dot{y}_j \frac{d}{dt} u_x(t, y_j(t)) + c \frac{d}{dt} u_y(t, y_j(t)) &= g(t, y_j(t), w_j(t)) \dot{y}_j(t) & j = 1, 2 \\ \frac{d}{dt} u(t, y_j(t)) &= u_x(t, y_j(t)) + u_y(t, y_j(t)) \dot{y}_j(t) & j = 1 \text{ (oder 2)} \end{aligned}$$

Gegeben:

x, y, u, u_x, u_y für Punkte P und Q

Gesucht:

x, y, u, u_x, u_y für Punkt R

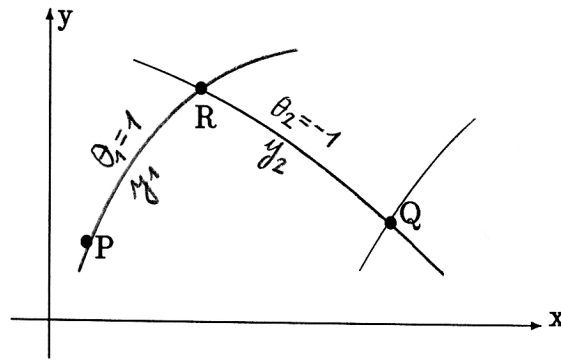


Abbildung 1.10

wie folgt approximiert werden

$$\begin{aligned} y(R) - y(P) &= \frac{1}{2}(\dot{y}_1(R) + \dot{y}_1(P))(x(R) - x(P)) \\ y(R) - y(Q) &= \frac{1}{2}(\dot{y}_2(R) + \dot{y}_2(Q))(x(R) - x(Q)) \end{aligned}$$

$$\frac{a(R)\dot{y}_j(R) + a(P_j)\dot{y}_j(P_j)}{2}(u_x(R) - u_x(P_j)) + \frac{c(R) + c(P_j)}{2}(u_y(R) - u_y(P_j)) = \frac{1}{2}(g(R) + g(P_j))(y(R) - y(P_j)) \quad j = 1, 2 \quad P_1 = P, \quad P_2 = Q$$

$$u(R) - u(P) = \frac{1}{2}(u_x(R) + u_x(P))(x(R) - x(P)) + \frac{1}{2}(u_y(R) + u_y(P))(y(R) - y(P))$$

Da im Fall einer nichtlinearen DGL g, a, b, c und damit \dot{y}_1, \dot{y}_2 noch von u, u_x und u_y abhängen, ist dies ein System von 5 im allgemeinen nichtlinearen Gleichungen in den 5 Unbekannten x, y, u, u_x und u_y für jeden Punkt R der neuen charakteristischen Schicht.

(Man beachte, daß $u_x = \varphi'_1$ auf der Anfangskurve, sodaß dort für jeden Punkt die 5 Größen x, y, u, u_x, u_y bekannt sind.)

Das nichtlineare Gleichungssystem wird durch direkte Iteration gelöst, indem man für die Unbekannten in den Argumenten der Koeffizienten $a, c, g, \dot{y}_1, \dot{y}_2$ jeweils die davorliegende Näherung einsetzt.

Als Startwerte kann man die Werte an den Punkten P oder Q einsetzen. Ein Vorteil der Charakteristikenverfahren liegt darin, daß das charakteristische Netz dem Verhalten der DGL angepaßt ist und Verfahren höherer Ordnung mit Hilfe von Extrapolationsverfahren leicht konstruierbar sind.

Bei drei oder mehr freien Variablen sind sie jedoch zu kompliziert, um effizient eingesetzt werden zu können.

1.3 Differenzenapproximationen

In diesem Abschnitt beschäftigen wir uns mit der Anwendung der bekannten Differenzenapproximationen für partielle Ableitungen auf hyperbolische Differentialgleichungen von einem naiven konstruktiven Standpunkt aus.

Mit einer genauen Untersuchung der Konvergenzbedingungen werden wir uns erst in einem späteren Kapitel beschäftigen.

Der Vorteil der Differenzenverfahren liegt in ihrer einfachen Handhabung. Ein Nachteil der verwendeten regelmäßigen Rechteckgitter ist darin zu sehen, daß sie dem Lösungsverlauf nicht so gut angepaßt sind wie z.B. die Charakteristiken-Verfahren bei hyperbolischen Problemen.

1.3.1 Hyperbolische DGLen zweiter Ordnung

Wir beginnen die Diskussion eines reinen Anfangswertproblems für die **Wellengleichung** (schwingende Saite)

$$\begin{aligned} u_{tt}(x, t) &= c^2 u_{xx}(x, t) & x \in \mathbb{R}, \quad t > 0 & \quad c > 0 \quad \text{konstant} \\ u(x, 0) &= f(x), & u_t(x, 0) &= g(x) \quad x \in \mathbb{R}. \end{aligned}$$

Wir lösen zunächst die Gleichung analytisch. Dazu beachten wir, daß die homogene Gleichung mit konstanten Koeffizienten sich auf eine noch einfachere Form transformieren läßt: Mit

$$\xi \stackrel{\text{def}}{=} x + ct, \quad \eta \stackrel{\text{def}}{=} x - ct$$

wird

$$\begin{pmatrix} x \\ t \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\xi + \eta) \\ \frac{1}{2c}(\xi - \eta) \end{pmatrix}.$$

Setzt man

$$\varphi(\xi, \eta) = u\left(\frac{1}{2}(\xi + \eta), \frac{1}{2c}(\xi - \eta)\right) = u(x, t),$$

dann wird

$$\begin{aligned} \frac{\partial}{\partial \xi} \varphi(\xi, \eta) &= (u_x \cdot \frac{1}{2} + \frac{1}{2c} \cdot u_t) \left(\frac{1}{2}(\xi + \eta), \frac{1}{2c}(\xi - \eta) \right) \\ \frac{\partial^2}{\partial \xi \partial \eta} \varphi(\xi, \eta) &= \left(\frac{1}{4} u_{xx} - \underbrace{\frac{1}{4c} u_{xt} + \frac{1}{4c} u_{tx}}_0 - \frac{1}{4c^2} u_{tt} \right) \left(\frac{1}{2}(\xi + \eta), \frac{1}{2c}(\xi - \eta) \right) \\ &\equiv 0 \end{aligned}$$

d.h.

$$\varphi(\xi, \eta) = P(\xi) + Q(\eta)$$

mit zunächst beliebigen stetig differenzierbaren Funktionen P und Q . Aufgrund der Anfangsbedingungen wird

$$\begin{aligned} P(x) + Q(x) &= f(x) \\ cP'(x) - cQ'(x) &= u_t(x, 0) = g(x) \end{aligned}$$

d.h.

$$P(x) - Q(x) = \frac{1}{c} \int_{x_0}^x g(\xi) d\xi + 2K$$

also

$$\begin{aligned} P(x) &= \frac{1}{2}f(x) + \frac{1}{2c} \int_{x_0}^x g(\xi) d\xi + K \\ Q(x) &= \frac{1}{2}f(x) - \frac{1}{2c} \int_{x_0}^x g(\xi) d\xi - K \end{aligned}$$

und

$$\begin{aligned} u(x, t) &= P(x + ct) + Q(x - ct) \\ &= \frac{1}{2}(f(x + ct) + f(x - ct)) + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\xi) d\xi. \end{aligned}$$

Dies ist die d'Alembertsche Lösungsdarstellung der Wellengleichung.

Die Charakteristiken der Gleichung in der Parametrisierung

$$\begin{pmatrix} \tau \\ y_j(\tau) \end{pmatrix}$$

bestimmen sich aus

$$\dot{y}_j(\tau) = (0 \pm \sqrt{0 - (-4c^2)}) / (-2c^2) = \pm \frac{1}{c}, \quad y_j(t_0) = 0.$$

(Man beachte: $-c^2 u_{xx} + u_{tt} = 0$ entspricht $a = -c^2$, $b = 0$, $c = 1$ in unserem Modell der Gleichung zweiter Ordnung.) Sie haben also die Form

$$\begin{pmatrix} \tau \\ \frac{1}{c}\tau + d \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} \tau \\ -\frac{1}{c}\tau + d \end{pmatrix},$$

d.h. durch den Punkt (x^*, t^*) gehen die beiden Charakteristiken

$$\begin{pmatrix} \tau \\ \frac{1}{c}(\tau + ct^* - x^*) \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} \tau \\ -\frac{1}{c}(\tau - ct^* - x^*) \end{pmatrix} \quad \tau \in \mathbb{R}$$

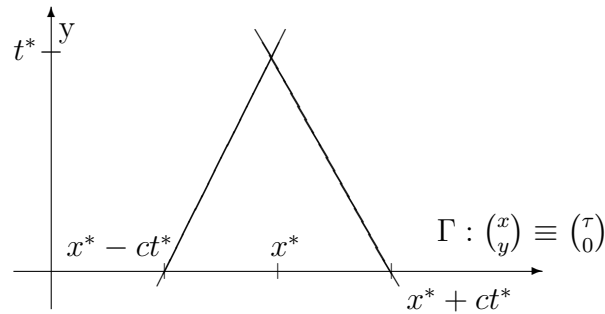


Abbildung 1.11

Aus der Darstellungsformel für u erkennt man, daß $u(x^*, t^*)$ genau durch die Anfangswerte auf dem zu (x^*, t^*) gehörenden Abhängigkeitsintervall auf Γ (x -Achse) bestimmt ist.

Zur Konstruktion einer Differenzenapproximation überziehen wir die obere Halbebene mit einem Recheckgitter

$$(x_j, t_n) = (j\Delta x, n\Delta t) \quad j \in \mathbb{Z}, \quad n \in \mathbb{N}_0 \quad \Delta x, \Delta t > 0 \text{ fest}$$

Für alle auftretenden Funktionen f etc. bedeute stets

$$f_{jn} \stackrel{\text{def}}{=} f(x_j, t_n).$$

Wir beginnen nun mit einer Taylorentwicklung der exakten Lösung: (wobei eine genügend hohe Differenzierbarkeitsordnung der Lösung u unterstellt wird)

$$u_{j,n\pm 1} = u_{j,n} \pm \Delta t (u_t)_{j,n} + \frac{1}{2} (\Delta t)^2 (u_{tt})_{j,n} \pm \frac{1}{6} (\Delta t)^3 (u_{ttt})_{j,n} + \mathcal{O}((\Delta t)^4)$$

also

$$u_{j,n+1} = 2u_{j,n} + (\Delta t)^2 (u_{tt})_{j,n} - u_{j,n-1} + \mathcal{O}((\Delta t)^4)$$

Andererseits ist aufgrund der DGL

$$(u_{tt})_{j,n} = c^2 (u_{xx})_{j,n} = c^2 \cdot \frac{1}{(\Delta x)^2} (u_{j+1,n} - 2u_{j,n} + u_{j-1,n}) + \mathcal{O}((\Delta x)^2)$$

also mit

$$\lambda \stackrel{\text{def}}{=} \frac{\Delta t}{\Delta x} \quad \text{fest}$$

$$u_{j,n+1} = 2(1 - (c\lambda)^2)u_{j,n} + (c\lambda)^2(u_{j+1,n} + u_{j-1,n}) - u_{j,n-1} + \mathcal{O}((\Delta t)^2 + (\Delta x)^2)(\Delta t)^2$$

Die Vernachlässigung des \mathcal{O} -Terms ergibt die Rechenvorschrift für Näherungen $u_{j,n}^h$ für $u_{j,n}$:

$$u_{j,n+1}^h = 2(1 - (c\lambda)^2)u_{j,n}^h + (c\lambda)^2(u_{j+1,n}^h + u_{j-1,n}^h) - u_{j,n-1}^h \quad j \in \mathbb{Z}, n \geq 1.$$

Diese Rechenvorschrift verbindet rekursiv 5 Punkte des Gitters in expliziter Weise, sodaß die Zeitschicht $n + 1$ berechenbar ist, wenn die Werte auf den Zeitschichten n und $n - 1$ bekannt sind.

Zum Start des Verfahrens benötigt man noch die Werte auf der Zeitschicht 1. Benutzen wir Taylorentwicklung bzgl. t und die Differentialgleichung sowie die Anfangswerte, dann erhalten wir

$$\begin{aligned}
u_{j,1} &= u_{j,0} + \Delta t(u_t)_{j,0} + \frac{1}{2}(\Delta t)^2(u_{tt})_{j,0} + \frac{1}{6}(\Delta t)^3(u_{ttt})_{j,0} + \mathcal{O}((\Delta t)^4) \\
&= u_{j,0} + \Delta t g_j + \frac{1}{2}(\Delta t)^2 c^2(u_{xx})_{j,0} + \frac{1}{6}(\Delta t)^3 c^2(u_{txx})_{j,0} + \mathcal{O}((\Delta t)^4) \\
&= f_j + \Delta t g_j + \frac{1}{2}(c\lambda)^2(u_{j+1,0} - 2u_{j,0} + u_{j-1,0}) \\
&\quad + \frac{1}{6}\Delta t(c\lambda)^2(g_{j+1} - 2g_j + g_{j-1}) + \mathcal{O}((\Delta t)^2 + (\Delta x)^2)(\Delta t)^2 \\
&= (1 - (c\lambda)^2)f_j + \frac{1}{2}(c\lambda)^2(f_{j+1} + f_{j-1}) + \\
&\quad \Delta t\left((1 - \frac{1}{3}(c\lambda)^2)g_j + \frac{1}{6}(c\lambda)^2(g_{j+1} + g_{j-1})\right) + (\Delta t)^2\mathcal{O}((\Delta t)^2 + (\Delta x)^2).
\end{aligned}$$

Also wählen wir

$$u_{j,1}^h = (1 - (c\lambda)^2)f_j + \frac{1}{2}(c\lambda)^2(f_{j+1} + f_{j-1}) + \Delta t\left((1 - \frac{1}{3}(c\lambda)^2)g_j + \frac{1}{6}(c\lambda)^2(g_{j+1} + g_{j-1})\right)$$

Wir betrachten zunächst diese Diskretisierung im Zusammenhang mit einem Randanfangswertproblem

$$\begin{aligned}
u_{tt} &= c^2 u_{xx}, \quad t > 0, \quad 0 \leq x \leq L, \\
u(0, t) &= 0, \quad t \geq 0 \\
u(L, t) &= 0, \quad t \geq 0 \\
u(x, 0) &= f(x) \quad 0 \leq x \leq L, \quad f(0) = f(L) = 0 \\
u_t(x, 0) &= g(x) \quad 0 \leq x \leq L, \quad g(0) = g(L) = 0
\end{aligned}$$

und zeigen

Satz 1.2 Sei $u \in C^4([0, L] \times \mathbb{R}_+)$, $(M + 1)\Delta x = L$ und

$$\lambda \leq \frac{1}{c}.$$

Dann gilt

$$\left(\frac{1}{M} \sum_{j=1}^M |u_{j,n} - u_{j,n}^h|^2\right)^{1/2} \leq K(T)(\Delta t)^2, \quad 0 \leq j \leq M+1, \quad 0 \leq n \leq n_0, \quad n_0 \Delta t \leq T.$$

□

c ist die Ausbreitungsgeschwindigkeit der Wellen, λ das Verhältnis von Zeit- zu Raumschrittweite. Die Fehlernorm auf der linken Seite dieser Ungleichung ist eine Approximation für den L_2 -Fehler einer Zeitschicht.

Beweis: Wir setzen

$$\varepsilon_{j,n} = u_{j,n} - u_{j,n}^h, \quad 0 \leq j \leq M+1.$$

Dann gilt natürlich

$$\varepsilon_{0,n} = \varepsilon_{M+1,n} = 0 \quad \forall n$$

und mit

$$\vec{\varepsilon}_n = (\varepsilon_{1,n}, \dots, \varepsilon_{M,n}, \varepsilon_{1,n-1}, \dots, \varepsilon_{M,n-1})^T$$

und den obigen Taylorentwicklungen ergibt sich die Rekursion

$$\vec{\varepsilon}_{n+1} = \begin{pmatrix} A & -I \\ I & O \end{pmatrix} \vec{\varepsilon}_n + \mathcal{O}((\Delta t)^4)$$

wobei bereits die feste Relation zwischen Δt und Δx benutzt wurde. A ist die $M \times M$ Tridiagonalmatrix

$$A = 2I + (c\lambda)^2 \text{tridiag}(1, -2, 1)$$

Der Term $\mathcal{O}((\Delta t)^4)$ bezeichnet einen Vektor, dessen Komponenten durch eine Konstante multipliziert mit $(\Delta t)^4$ abgeschätzt werden können. In diese Konstante geht das Supremum der vierten partiellen Ableitungen von u bis zur Zeitschicht $n+1$ ein. Die letzten M Komponenten dieses Vektors sind exakt null, aber dies spielt in den folgenden Abschätzungen keine Rolle. Ebenso ist wegen des Diskretisierungsfehlers der ersten Zeitschicht

$$\vec{\varepsilon}_1 = \mathcal{O}((\Delta t)^4).$$

A kann uniär diagonalisiert werden mit Hilfe einer Matrix V und mit

$$\tilde{\varepsilon}_n = \text{diag}(V, V) \vec{\varepsilon}_n$$

erhalten wir die Rekursion

$$\tilde{\varepsilon}_{n+1} = \begin{pmatrix} D & -I \\ I & O \end{pmatrix} \tilde{\varepsilon}_n + \mathcal{O}((\Delta t)^4)$$

mit der Diagonalmatrix

$$D = \text{diag}(2 - 2(c\lambda)^2(1 - \cos(j\pi/(M+1))), 1 \leq j \leq M).$$

Die Eigenwerte der $2M \times 2M$ Blockmatrix sind daher

$$\mu_{i,2,1} = \frac{1}{2}(D_{i,i} \pm \sqrt{D_{i,i}^2 - 4})$$

und weil die $D_{i,i}$ alle im offenen Intervall $] -2, 2[$ liegen, sind sie alle konjugiert komplex und vom Betrag 1. Mit einer weiteren Ähnlichkeitstransformation mit einer Permutationsmatrix P , die definiert ist durch

$$P(1, \dots, 2M)^T = (1, M+1, 2, M+2, \dots, M, 2M)$$

wird

$$P\tilde{\epsilon}_{n+1} = \text{blockdiag}\left(\left(\begin{array}{cc} D_{i,i} & -1 \\ 1 & 0 \end{array}\right), 1 \leq i \leq M\right)P\tilde{\epsilon}_n + \mathcal{O}((\Delta t)^4).$$

Wir wollen dieses System nun weiter diagonalisieren. Dazu müssen wir auch die Eigenvektormatrix der 2×2 Blockmatrizen und deren Inverse berechnen. Diese Inversen multiplizieren nämlich dann auch den Vektor $\mathcal{O}((\Delta t)^4)$. Wir erhalten als Eigenvektormatrix

$$T_i = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ \mu_i & \bar{\mu}_i \end{pmatrix} \quad (T_i)^{-1} = -\frac{1}{\sqrt{4-D_{i,i}^2}} \begin{pmatrix} \bar{\mu} & -1 \\ -\mu & 1 \end{pmatrix}$$

und mit

$$\hat{\epsilon}_n = \text{blockdiag}((T_i)^{-1})P\tilde{\epsilon}_n$$

schliesslich

$$\begin{aligned} \hat{\epsilon}_{n+1} &= \text{diag}(\mu_{i,1,2})\hat{\epsilon}_n + \text{blockdiag}((T_i)^{-1})\mathcal{O}((\Delta t)^4), \\ \hat{\epsilon}_1 &= \text{blockdiag}((T_i)^{-1})\mathcal{O}((\Delta t)^4). \end{aligned}$$

Wir benutzen nun als Norm

$$\|\cdot\| = \frac{1}{\sqrt{M}}\|\cdot\|_2.$$

In dieser Norm gilt

$$\|P\| = 1, \|V\| = 1,$$

da beide Matrizen unitär sind und ein Faktor die zugeordnete Matrixnorm nicht ändert. Für die Matrix

$$\text{blockdiag}((T_i))$$

ergibt sich in dieser Norm unter Benutzung von $\|\cdot\|_2^2 \leq \|\cdot\|_1\|\cdot\|_\infty$ die Abschätzung

$$\text{cond}(\text{blockdiag}((T_i))) \leq \frac{2\sqrt{2}}{\min_i \sqrt{4 - D_{i,i}^2}}$$

also unter Auflösung der Rekursion bezüglich n

$$\|\vec{\epsilon}_n\| \leq nC2\sqrt{2} \frac{1}{\min_i \sqrt{4 - D_{i,i}^2}} (\Delta t)^4$$

wobei C für die Konstante im Term $\mathcal{O}(\cdot)$ steht. Weil nun

$$n \leq T/(\Delta t)$$

und

$$\begin{aligned}
 4 - D_{i,i}^2 &= 4 - (2 - 2(c\lambda)^2(1 - \cos(i\pi/(M+1))))^2 \\
 &\geq 4 - (2 - 2(c\lambda)^2(1 - \cos(\pi/(M+1))))^2 \\
 &= 4 - (2 - (c\lambda)^2(\pi/(M+1))^2(1 + \mathcal{O}(1/(M+1)^2)))^2 \\
 &= 4(c\lambda)^2(\pi/(M+1))^2 + \mathcal{O}(1/(M+1)^4)
 \end{aligned}$$

gilt, ergibt sich endgültig mit einer weiteren Konstanten C_1

$$\|\vec{\epsilon}_n\| \leq TCC_1(\Delta t)^2 \quad \text{für } n\Delta t \leq T$$

wie behauptet. □

Wir betrachten weiter die Anwendung der Differenzenformel auf das reine Anfangswertproblem. Aus der Rekursionsformel folgt, daß

$$u_{jn}^h \approx u_{jn} = u(j\Delta x, n\Delta t) \stackrel{\text{def}}{=} u(x, t) \quad (u^h(x, t; \Delta x, \Delta t) \stackrel{\text{def}}{=} u_{jn}^h)$$

abhängt von den Werten $u_{j-n,0}^h, \dots, u_{j+n,0}^h$ entsprechend den Anfangswerten

$$u((j-n)\Delta x, 0), \dots, u((j+n)\Delta x, 0), \quad \text{d.h.} \quad u(x - t \cdot \frac{1}{\lambda}, 0), \dots, u(x + t \cdot \frac{1}{\lambda}, 0)$$

weil

$$n\Delta x = n\Delta t \cdot \frac{\Delta x}{\Delta t} = t/\lambda.$$

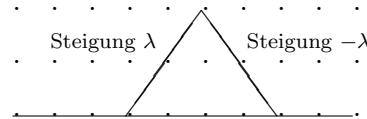


Abbildung 1.12

Man nennt $[x_j - t_n/\lambda, x_j + t_n/\lambda]$ das numerische Abhängigkeitsintervall von u_{jn}^h . Andererseits hängt $u(x, t)$ von allen Werten $u(x-ct, 0), \dots, u(x+ct, 0)$ ab, d.h. $[x-ct, x+ct]$ ist das Abhängigkeitsintervall von $u(x, t)$.

Falls $1/\lambda < c$, dann ist das numerische Abhängigkeitsintervall im analytischen Abhängigkeitsintervall echt enthalten. Mit $\Delta x \rightarrow 0$, $\Delta t \rightarrow 0$, $\Delta t/\Delta x = \lambda$ fest **kann dann das Verfahren nicht konvergieren**, weil in $u^h(x, t; \Delta x, \Delta t)$ für $\Delta t, \Delta x \rightarrow 0$ nicht alle Werte $u(x-ct, 0), \dots, u(x+ct, 0)$ eingehen und Abänderungen der Anfangswerte auch $u(x, t)$ in der Regel ändern.

Satz 1.3 *Notwendig für die Konvergenz*

$$\begin{aligned} u^h(x, t; \Delta x, \Delta t) &\longrightarrow u(x, t) \\ \Delta x &\rightarrow 0 \\ \Delta t &\rightarrow 0 \\ \lambda &= \Delta t / \Delta x = \text{const.} \end{aligned}$$

ist die Erfüllung der Courant–Friedrichs–Lewy–Bedingung

$$\lambda = \Delta t / \Delta x \leq \frac{1}{c}$$

(d.h. die t -Schritte ist im Verhältnis zur x -Schrittweite geeignet begrenzt.) □

Bemerkung 1.2 *Man kann zeigen, daß die Erfüllung der CFL-Bedingung*

Numerisches Abhängigkeitsintervall \supseteq Abhängigkeitsintervall

ganz allgemein notwendig für die Konvergenz bei Differenzenverfahren für hyperbolische Anfangswertaufgaben ist. □

Als reines Anfangswertproblem ist die Wellengleichung wenig praxisrelevant. Häufig tritt sie in Verbindung mit einem Anfangs-Randwertproblem auf, z.B.

$$\begin{aligned} u_{tt}(x, t) &= c^2(x)u_{xx}(x, t) + r(x, t) \\ u(x, 0) &= f(x), \quad u_t(x, 0) = g(x) \quad 0 \leq x \leq 1 \\ u(0, t) &= \varphi_0(t), \quad u(1, t) = \varphi_1(t) \quad t \geq 0 \\ f(0) &= \varphi_0(0), \quad f(1) = \varphi_1(0) \quad \text{Verträglichkeitsbedingungen} \end{aligned} \tag{1.6}$$

(u bedeutet hier die Auslenkung einer gespannten Saite mit orts- und zeitabhängiger Anregung r , vorgegebener Anfangsauslenkung und Anfangsgeschwindigkeit und vorgegebener Randbefestigung, die mit der Zeit veränderlich ist.)

Es treten auch Periodizitätsbedingungen auf, z.B.

$$u(x + L, t) = u(x, t) \quad \forall (x, t) \in \mathbb{R} \times \mathbb{R}_+$$

wobei dann natürlich auch die Anfangsvorgaben f und g L -periodisch sein müssen. Auch in diesem Fall kann man sich auf einen Streifen der Breite L beschränken, weil man beim numerischen Rechnen die “fehlenden” u_{jn}^h -Werte durch die Ausnutzung der Periodizität gewinnen kann: z.B.

$$u_{-1,j}^h = u_{N-1,j}^h \quad \text{und} \quad u_{N+1,j}^h = u_{1,j}^h$$

Im Zusammenhang mit einem solchen Anfangsrandwertproblem wollen wir nun sogleich einen neuen Zugang zur Gewinnung von Diskretisierungen auf Rechteckgittern kennenlernen, die sogenannte **vertikale Linienmethode** oder Semidiskretisierungsmethode. (Das Problem wird bezüglich der räumlichen Variablen diskretisiert, bleibt aber kontinuierlich in der Zeit.)

Wir betrachten weiter die obenstehende RAWA der Wellengleichung und setzen zunächst für ein Gitter in der Raumvariablen x

$$x = i\Delta x \quad 0 \leq i \leq N, \quad \Delta x = \frac{1}{N} \quad (N \geq 3)$$

$$\tilde{v}_i(t) \stackrel{\text{def}}{=} u(x_i, t)$$

Nun gilt für $u \in C^4 \quad ([0, 1] \times \mathbb{R}_+)$

$$\begin{aligned} u_{xx}(x_i, t) &= \frac{u(x_{i+1}, t) - 2u(x_i, t) + u(x_{i-1}, t))}{(\Delta x)^2} + \mathcal{O}(\Delta x^2) \\ &= \frac{\tilde{v}_{i+1}(t) - 2\tilde{v}_i(t) + \tilde{v}_{i-1}(t)}{(\Delta x)^2} + \mathcal{O}(\Delta x^2) \\ u_{tt}(x_i, t) &= \ddot{\tilde{v}}_i(t) \end{aligned}$$

Vernachlässigung der $\mathcal{O}((\Delta x)^2)$ -Terme führt dann auf das folgende Anfangswertproblem zweiter Ordnung für die gesuchten **Funktionen**

$$\begin{aligned} v_1(t), \dots, v_{N-1}(t) &: \quad (\text{als Näherung für } \tilde{v}_i(t)) \\ \ddot{v}_i(t) &= \frac{c_i^2}{(\Delta x)^2} (v_{i+1}(t) - 2v_i(t) + v_{i-1}(t)) + r(x_i, t) \quad 1 \leq i \leq N-1 \\ v_0(t) &= \varphi_0(t) \\ v_N(t) &= \varphi_1(t) \\ \left. \begin{aligned} v_i(0) &= f(i\Delta x) \\ \dot{v}_i(0) &= g(i\Delta x) \end{aligned} \right\} 1 \leq i \leq N-1 \end{aligned}$$

Dies ist also ein Anfangswertproblem für ein System gewöhnlicher Differentialgleichungen 2. Ordnung, in das die Randbedingungen schon eingearbeitet sind. Im Falle

$$c(x) = c \quad \text{konstant}, \quad r \equiv 0$$

wird das DGL-System besonders einfach:

$$\begin{aligned} \ddot{v} &= Av + b(t) \\ v(0) &= \begin{bmatrix} f_1 \\ \vdots \\ f_{N-1} \end{bmatrix} \quad \dot{v}(0) = \begin{bmatrix} g_1 \\ \vdots \\ g_{N-1} \end{bmatrix} \quad b(t) = \frac{c^2}{(\Delta x)^2} \begin{bmatrix} \varphi_0(t) \\ 0 \\ \vdots \\ 0 \\ \varphi_1(t) \end{bmatrix} \left. \vphantom{b(t)} \right\} N-3 \\ A &= -\frac{c^2}{(\Delta x)^2} (0, \dots, 0, -1, 2, -1, 0, \dots, 0) \quad (\text{tridiagonal} \in \mathbb{R}^{N-1, N-1}) \end{aligned}$$

Nach der Substitution

$$w \stackrel{\text{def}}{=} \dot{v}$$

wird daraus das System erster Ordnung für $2N - 2$ gesuchte Funktionen

$$\begin{bmatrix} \dot{v} \\ \dot{w} \end{bmatrix} = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} + \begin{bmatrix} 0 \\ b(t) \end{bmatrix}, \quad t > 0, \quad \begin{bmatrix} v(0) \\ w(0) \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_{N-1} \\ g_1 \\ \vdots \\ g_{N-1} \end{bmatrix}$$

Dieses System gewöhnlicher Differentialgleichungen kann nun im Prinzip mit den numerischen Methoden für gewöhnliche Differentialgleichungen behandelt werden. Auch im oben angedeuteten allgemeineren Fall mit variablen Koeffizienten und variabler Inhomogenität der Form $r(x, t, u, u_x, u_t)$ hat man dabei keinerlei formale Probleme. Lediglich wird dann die rechte Seite der DGL nichtlinear vom Typ $F(t, v, w)$. Die Diskretisierungsmethoden für gewöhnliche DGLen erzeugen nun ihrerseits ein t -Gitter, sodaß letztendlich wieder eine Näherungslösung u^h für u auf einem Rechteckgitter der (x, t) -Ebene erzeugt wird.

Wir erinnern nun an das schon bei gewöhnlichen DGLen diskutierte **Stabilitätsproblem**:

Die Eigenwerte der Matrix A sind die Werte $-\omega_j^2$ mit

$$\omega_j^2 = 2 \left(\frac{c}{\Delta x} \right)^2 \left(1 - \cos \left(\frac{j\pi}{N} \right) \right) \quad 1 \leq j \leq N - 1$$

und somit die Eigenwerte λ_j von $\begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}$

$$\lambda_j = \pm i\omega_j \quad i^2 = -1$$

Im Falle $\varphi_0(t) \equiv \varphi_1(t) \equiv 0$, d.h. $b(t) \equiv 0$ lautet die Lösung dieser gewöhnlichen DGL

$$\sum_{j=1}^{N-1} \left(\alpha_j z_j^{(+)} e^{i\omega_j t} + \beta_j z_j^{(-)} e^{-i\omega_j t} \right)$$

Dabei sind $z_j^{(+)}, z_j^{(-)}$ Eigenvektoren von $\begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix}$ zu $\pm i\omega_j$ und α_j, β_j bestimmen sich aus den Anfangswerten. Die Lösung enthält also für $N \gg 1$ sehr hoch oszillierende Anteile. Die Lösung der DGL beschreibt eine **ungedämpfte Schwingung**.

Man vergleiche dies mit der Lösung der partiellen Randanfangswertaufgabe durch Fourierreihenansatz: Mit

$$u_{tt} = c^2 u_{xx}, \quad 0 \leq x \leq 1, \quad u(x, 0) = f(x), \quad u_t(x, 0) = g(x)$$

ist

$$u(x, t) = \sum_{k=1}^{\infty} a_k(t) \sin(k\pi x)$$

mit

$$a_k(t) = c_k \cos(k\pi ct) + d_k \sin(k\pi ct)$$

wo

$$c_k = 2 \int_0^1 f(x) \sin(k\pi x) dx, \quad d_k = \frac{2}{k\pi c} \int_0^1 g(x) \sin(k\pi x) dx.$$

Als Integrationsmethode für das gewöhnliche DGL-System kommen nur solche Methoden in Frage, deren Bereich der absoluten Stabilität ein Intervall um 0 auf der imaginären Achse als Randstück aufweist, d.h. die Schwingungsamplituden werden weder verstärkt noch gedämpft. Im vorliegenden Fall heisst dies, daß die Schrittweite Δt so gewählt werden muß, daß

$$i\Delta t \frac{c}{\Delta x} \sqrt{2(1 + \cos(\frac{\pi}{N}))} \approx i2c \frac{\Delta t}{\Delta x}$$

auf dem Rand des Gebietes der absoluten Stabilität des Verfahrens liegen muß. Geeignet sind die **explizite Mittelpunkregel** (die **kein reelles** Intervall der absoluten Stabilität besitzt (!)) und die Trapezregel. Im Folgenden bezeichnet $v^{(i)}$, $w^{(i)}$ den Vektor der Näherungswerte für die Lösung $v(t)$, $w(t)$ der semidiskretisierten Gleichung. In der früheren Notation ist also

$$v_j^{(i)} = u_{j,i}^h$$

Für die explizite Mittelpunkregel wird

$$\begin{bmatrix} v^{(n+1)} \\ w^{(n+1)} \end{bmatrix} = \begin{bmatrix} v^{(n)} \\ w^{(n)} \end{bmatrix} + 2\Delta t \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} \begin{bmatrix} v^{(n)} \\ w^{(n)} \end{bmatrix} + 2\Delta t \begin{bmatrix} 0 \\ b(t_n) \end{bmatrix}$$

Das Stabilitätspolynom der Mittelpunkregel lautet

$$\pi(\zeta; q) = (\zeta - 1)(\zeta + 1) - 2q\zeta \quad (q \hat{=} \Delta t \lambda_j)$$

mit den Nullstellen

$$\zeta_{1,2}(q) = q \pm \sqrt{q^2 + 1}$$

und für q rein imaginär und $\pi(\zeta; q) = 0$ ist also

$$|\zeta| = 1 \text{ falls } |\Delta t \lambda_j| \leq 1$$

d.h. es liegt eine Einschränkung an das Schrittweitenverhältnis $\Delta t/\Delta x$ vor, nämlich

$$2c\Delta t/\Delta x \leq 1$$

also eine schärfere Einschränkung als die CFL-Bedingung. (Andererseits muß man aus Gründen des Phasenfehlers letztlich doch $c\Delta t/\Delta x \ll 1$ sein.)

Noch günstiger liegen die Verhältnisse bei der Trapezregel: Diese führt, angewandt auf das obige System erster Ordnung, auf das lineare Gleichungssystem

$$\left(I - \frac{\Delta t}{2} \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} \right) \begin{bmatrix} v^{(n+1)} \\ w^{(n+1)} \end{bmatrix} = \left(I + \frac{\Delta t}{2} \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} \right) \begin{bmatrix} v^{(n)} \\ w^{(n)} \end{bmatrix} + \frac{\Delta t}{2} \begin{bmatrix} 0 \\ b(t_n) + b(t_{n+1}) \end{bmatrix}$$

Ausgeschrieben lautet dies

$$\begin{aligned} v^{(n+1)} - \frac{\Delta t}{2} w^{(n+1)} &= v^{(n)} + \frac{\Delta t}{2} w^{(n)} \\ w^{(n+1)} - \frac{\Delta t}{2} A v^{(n+1)} &= w^{(n)} + \frac{\Delta t}{2} A v^{(n)} + \frac{\Delta t}{2} (b(t_n) + b(t_{n+1})) \end{aligned}$$

Multiplikation der ersten Gleichung mit $\frac{\Delta t}{2} A$ und Addition zur zweiten ergibt

$$\left(I - \frac{\Delta t^2}{4} A \right) w^{(n+1)} = \left(I + \frac{\Delta t^2}{4} A \right) w^{(n)} + \Delta t A v^{(n)} + \frac{\Delta t}{2} (b(t_n) + b(t_{n+1}))$$

oder

$$\left(I - \frac{\Delta t^2}{4} A \right) \Delta w^{(n+1)} = \frac{\Delta t^2}{2} A w^{(n)} + \Delta t A v^{(n)} + \frac{\Delta t}{2} (b(t_n) + b(t_{n+1}))$$

mit

$$\Delta w^{(n+1)} = w^{(n+1)} - w^{(n)}.$$

Dies ergibt die Rechenvorschrift für einen Zeitschritt

$$\begin{array}{ll} z^{(n)} \stackrel{def}{=} v^{(n)} + \frac{\Delta t}{2} w^{(n)} & (\hat{=} u(\cdot, t + \frac{\Delta t}{2})) \quad t = n\Delta t \\ y^{(n)} \stackrel{def}{=} \Delta t (A z^{(n)} + \frac{1}{2} (b(t_n) + b(t_{n+1}))) & (\hat{=} c^2 u_{xx}(\cdot, t + \frac{\Delta t}{2})) \\ \left(I - \frac{\Delta t^2}{4} A \right) \Delta w^{(n+1)} = y^{(n)} & \text{lösen} \\ & \text{tridiag. Gleichungssystem} \\ & \text{mit symm. pos. def. Matrix} \\ & \text{Eigenwerte} \in [1, 1 + (\frac{\Delta t}{\Delta x})^2 c^2] \\ w^{(n+1)} \stackrel{def}{=} w^{(n)} + \Delta w^{(n+1)} & (\hat{=} u_t(\cdot, t + \Delta t)) \\ v^{(n+1)} \stackrel{def}{=} v^{(n)} + \frac{\Delta t}{2} (w^{(n+1)} + w^{(n)}) & (\hat{=} u(\cdot, t + \Delta t)) \end{array}$$

Für die Wahl des Zeitschrittes Δt ist maßgeblich, wie gut

$$\frac{1 + i\omega\Delta t/2}{1 - i\omega\Delta t/2} = \exp(2i \arctan(\omega\Delta t/2)) \text{ den Wert } \exp(i\omega\Delta t)$$

approximiert. Offenbar ist das Verfahren amplitudentreu für jedes ω und Δt . Aber die numerische Phase nimmt langsamer zu als die exakte, weil

$$2 \arctan((\omega\Delta t)/2) < \omega\Delta t.$$

Der Fehler ist um so größer, je größer ω wird. Wegen

$$2 \arctan\left(\frac{\omega \Delta t}{2}\right) = \omega \Delta t - \frac{\omega^3 \Delta t^3}{12} + \frac{\omega^5 \Delta t^5}{80} \dots$$

und

$$|\omega| \leq \frac{2c}{\Delta x}$$

genügt es

$$\Delta t \leq \frac{\Delta x}{2c} \sqrt[3]{12\varepsilon}$$

zu wählen, damit ein einzelner Integrationsschritt in der $\|\cdot\|_2$ -Norm einen Fehler von höchstens ε (zwischen $v^{(n)}$ und $v(t_n)$!) erzeugt. Δt sollte also nicht zu groß gewählt werden, um die numerische Phasenverschiebung (die sogenannte numerische Dispersion) kleinzuhalten. Dies gilt erst recht bei nichtlinearen Problemen, wo die Stabilität des Verfahrens nicht so leicht zu entscheiden ist. Der Gesamtfehler des Verfahrens ist dann von der Form

$$\frac{1}{\sqrt{N}} \left\| v^{(n)} - \begin{pmatrix} u(x_1, t_n) \\ \vdots \\ u(x_{N-1}, t_n) \end{pmatrix} \right\|_2 \leq K(t_n) ((\Delta t)^2 + (\Delta x)^2)$$

(für klassische Lösungen). Man beachte, daß die beiden bisher geschilderten Verfahren nicht als Verfahren zu einer direkten Erzeugung einer Gitterfunktion $u^h(x, t; \Delta x, \Delta t)$ für $u(x, t)$ gedeutet werden können, da hier u und u_t simultan approximiert werden. Es gibt jedoch auch **direkte Integrationsverfahren** für gewöhnliche Differentialgleichungssysteme zweiter Ordnung der speziellen Gestalt

$$\ddot{y} = f(t, y) \quad f \text{ nichtlinear}$$

und lineare Differentialgleichungssysteme der Form

$$M\ddot{y} + C\dot{y} + Ky = F(t)$$

(M, C, K symm. pos. def. Matrizen), bei denen **nur** y approximiert wird. In unserem obigen speziellen Fall hat das DGL-System bereits die spezielle Form

$$\ddot{v} = F(t, v), \tag{1.7}$$

sodaß wir uns zunächst mit diesem Fall beschäftigen wollen. Ein allgemeines k -Schritt-Verfahren für (1.7) hat die Form

$$(MSV_2) \quad \sum_{i=0}^k \alpha_i u_{m+i}^h = h^2 \sum_{i=0}^k \beta_i F(t_{m+i}, u_{m+i}^h), \quad k \geq 2$$

mit $\alpha_k \neq 0$, $|\alpha_0| + |\beta_0| \neq 0$. Die Konvergenztheorie für diesen Verfahrenstyp läßt sich analog der für MSV zu $\dot{v} = F(t, v)$ entwickeln. Man kennt folgende Aussagen (zum Beweis vgl. bei Grigorieff Bd. 2)

Satz 1.4

(i) Das lineare Mehrschrittverfahren MSV_2 ist asymptotisch stabil, falls für die Nullstellen des Polynoms

$$\rho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i$$

gilt:

$$\rho(\zeta) = 0 \quad \Rightarrow \quad |\zeta| \leq 1, \quad \rho''(\zeta) \neq 0 \quad \text{falls} \quad |\zeta| = 1.$$

(ii) Das Verfahren (MSV_2) ist konsistent, falls $\rho(1) = 0$, $\rho'(1) = 0$, $\rho''(1) = 2\sigma(1)$ mit

$$\sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i$$

(iii) Die maximal erreichbare Ordnung p eines stabilen Verfahrens (MSV_2) ist

$$p = \begin{cases} k+2 & k \text{ gerade} \\ k+1 & k \text{ ungerade} \end{cases}$$

Verfahren der Ordnung $k+2$, k gerade erhält man, wenn man alle Nullstellen von ρ auf dem Einheitskreis wählt und

$$\sigma(\zeta) = \sum_{i=0}^k (f^{(i)}(1)/i!) (\zeta - 1)^i$$

mit $f(\zeta) = \rho(\zeta)/(\ln \zeta)^2$

Die Ordnung p des (MSV_2) ist dabei definiert durch

$$\sum_{i=0}^k \alpha_i y(x+ih) - h^2 \sum_{i=0}^k \beta_i f(x+ih, y(x+ih)) = C_{p+2} h^{p+2} y^{(p+2)}(x) + \mathcal{O}(h^{p+3})$$

$y'' = f(x, y)$

(iv) (MSV_2) ist konvergent genau dann, wenn es konsistent ($p \geq 1$) und stabil ist

(v) Es gibt kein auf der ganzen imaginären Achse stabiles (neutral stabiles) Verfahren vom Typ MSV_2 der Ordnung $p > 2$. \square

Typische Beispiele sind:

Das Verfahren von Störmer:

$$u_{n+1}^h - 2u_n^h + u_{n-1}^h = h^2 f_n, \quad \text{Ordnung 2}$$

Dieses Verfahren führt, angewandt auf unsere aus der Semidiskretisierung erhaltenen DGL

$$\ddot{v} = Av + b(t)$$

zurück zu unserem expliziten Differenzenverfahren.

Das Verfahren von Cowell:

$$u_{n+1}^h - 2u_n^h + u_{n-1}^h = \frac{1}{12}h^2(f_{n+1} + 10f_n + f_{n-1}), \quad \text{Ordnung 4}$$

Angewandt auf die Semidiskretisierung ergibt dies folgende Verknüpfung der Gitterfunktionswerte:

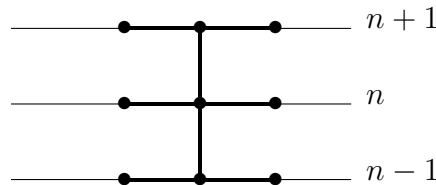


Abbildung 1.13

Für jede Zeitschicht hat man somit ein tridiagonales Gleichungssystem zu lösen. Andere mögliche implizite Verfahren sind

$$u_{n+1}^h - 2u_n^h + u_{n-1}^h = \frac{1}{2}h^2(f_{n+1} + f_{n-1}), \quad \text{Ordnung 2}$$

mit der Gitterverknüpfung

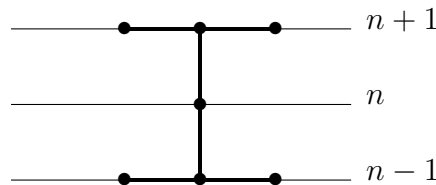


Abbildung 1.14

und

$$u_{n+1}^h - 2u_n^h + u_{n-1}^h = \frac{1}{4}h^2(f_{n+1} + 2f_n + f_{n-1}), \quad \text{Ordnung 2}$$

ebenfalls mit einer Gitterverknüpfung wie beim Cowell-Verfahren.

(Es gibt selbstverständlich Verfahren beliebig hoher Ordnung von diesem Typ, aber ihre Anwendung in diesem Zusammenhang (Semidiskretisierung einer hyperbolischen RAWA) wäre nur interessant, wenn auch die Raumableitung u_{xx} durch genauere Differenzenformeln approximiert würde. Man braucht dann zusätzliche Startschichten und spezielle Formeln

an den Rändern um z.B. $u_{xx}(\Delta x, t)$ zu approximieren. Hierdurch können die Eigenwerte des semidiskretisierten Systems ungünstig beeinflusst werden. Die Bandbreite der zu lösenden Gleichungssysteme vergrößert sich ebenfalls, was alles diesen Zugang nicht sehr attraktiv erscheinen lässt. Wir werden aber noch eine andere Form der Semidiskretisierung, nämlich mit finiten Elementen, behandeln, bei denen eine Erhöhung der Ordnung der Raumdiskretisierung vergleichsweise unproblematisch ist. Für die praktische Brauchbarkeit der Verfahren ist wiederum ihre absolute Stabilität maßgeblich, die nun durch das Stabilitätspolynom

$$\pi(\zeta; q) = \rho(\zeta) - q^2 \sigma(\zeta), \quad q \hat{=} \Delta t \lambda, \quad \lambda = \lambda_i^{\frac{1}{2}}(A),$$

wobei $\lambda_i(A)$ ein Eigenwert von A ist, beschrieben wird. Für die Formel von Störmer erhalten wir

$$\begin{aligned} \pi(\zeta; q) &= (\zeta - 1)^2 - q^2 \zeta \\ \pi(\zeta; q) = 0 &\Rightarrow \zeta = 1 + \frac{q^2}{2} \pm q \sqrt{1 + \left(\frac{q}{2}\right)^2}. \end{aligned}$$

Beim betrachteten Anwendungsfall sind die $\lambda_i(A)$ reell negativ, also interessiert der Bereich auf der imaginären Achse mit $|\zeta(q)| \leq 1$. Dies ist $[-i, i]$ entsprechend der Bedingung

$$\frac{\Delta t}{\Delta x} c \leq 1$$

die wir schon früher erhalten haben und die genau der CFL-Bedingung entspricht.

Für drei impliziten Verfahren haben wir

$$\begin{aligned} (1) \quad \pi(\zeta; q) &= (\zeta - 1)^2 - \frac{1}{12} q^2 (\zeta^2 + 10\zeta + 1) \\ (2) \quad \pi(\zeta; q) &= (\zeta - 1)^2 - \frac{1}{2} q^2 (\zeta^2 + 1) \\ (3) \quad \pi(\zeta; q) &= (\zeta - 1)^2 - \frac{1}{4} q^2 (\zeta^2 + 2\zeta + 1) \end{aligned}$$

Für rein imaginäres $q = i\omega\Delta t$ haben wir somit die Wurzeln

(1)

$$\zeta(i\omega\Delta t) = \frac{1 - \frac{5}{12}\omega^2(\Delta t)^2 \pm \sqrt{-\omega^2(\Delta t)^2 + \frac{1}{6}\omega^4(\Delta t)^4}}{1 + \frac{1}{12}\omega^2(\Delta t)^2} \quad (1.8)$$

d.h. für $\omega^2(\Delta t)^2 \leq 6$ ist $|\zeta(i\omega)| = 1$ und für $\omega^2(\Delta t)^2 > 6$

$$\max |\zeta(i\omega)| = \frac{\frac{5}{12}\omega^2 - 1 + \sqrt{\frac{1}{6}\omega^2(\omega^2 - 6)}}{\frac{1}{12}\omega^2 + 1} > 1$$

das Verfahren ist somit nur bedingt absolut stabil mit der Bedingung

$$\frac{\Delta t}{\Delta x} c \leq \sqrt{6}$$

In unserem Anwendungsfall ist dies wegen der CFL Bedingung ausreichend. Die implizite Struktur des Verfahrens zahlt sich hier offensichtlich nicht aus, weshalb es in der Praxis nicht sehr häufig angewendet wird. Daß die Zeitintegration von vierter Ordnung ist, ist aber ein Vorteil.

(2)

$$\zeta(i\omega\Delta t) = \frac{2 \pm \sqrt{4 - 4(1 + \omega^2(\Delta t)^2/2)^2}}{2(1 + \frac{1}{2}\omega^2(\Delta t)^2)} = \frac{1 \pm \sqrt{-\omega^2(\Delta t)^2 - (\frac{\omega^2(\Delta t)^2}{2})^2}}{1 + \frac{1}{2}\omega^2(\Delta t)^2}; \quad |\zeta| = 1$$

für alle $\omega\Delta t \in \mathbb{R}$, d.h. das Verfahren **ist auf der gesamten imaginären Achsen absolut (und neutral) stabil.**⁴

Die numerische Dispersion hängt davon ab, wie gut

$$\zeta(i\omega\Delta t) \quad e^{\pm i\omega\Delta t} \quad \text{approximiert.}$$

Es ist

$$\zeta(i\omega\Delta t) = \frac{1 \pm i\omega\Delta t \sqrt{1 + \frac{\omega^2(\Delta t)^2}{4}}}{1 + \frac{1}{2}\omega^2(\Delta t)^2} = e^{\pm i \arctan(\omega\Delta t \sqrt{1 + \omega^2(\Delta t)^2/4})}$$

Weil

$$\arctan(\omega \sqrt{1 + \frac{\omega^2}{4}}) < \omega$$

bewirkt das Verfahren ein Nachlaufen der hochfrequenten Lösungsanteilen.

Die genaue Reihenentwicklung zeigt

$$\arctan(\omega \sqrt{1 + \frac{\omega^2}{4}}) = \omega - \frac{5}{24}\omega^3 + \frac{209}{15 \cdot 128}\omega^5 - \dots$$

d.h., daß (wegen $\omega\Delta t \leq 2c\Delta t/\Delta x$)

$$c \frac{\Delta t}{\Delta x} \leq \sqrt[3]{\frac{3}{5}} \varepsilon$$

gewählt werden sollte, um die numerische Dispersion (pro Schritt) immer unter ε zu halten.

⁴Neutral stabil bedeutet, daß die numerische Amplitude weder gedämpft (noch verstärkt) werden, letzteres wäre Instabilität.

(3) Nun ist

$$\zeta(i\omega\Delta t) = \frac{1 - \frac{\omega^2(\Delta t)^2}{4} \pm i|\omega\Delta t|}{1 + \frac{\omega^2(\Delta t)^2}{4}}$$

d.h. $|\zeta(i\omega\Delta t)| = 1 \quad \forall \quad \omega\Delta t \in \mathbb{R}$ und

$$\zeta(i\omega\Delta t) = e^{\pm i \arctan\left(\frac{|\omega\Delta t|}{1 - \omega^2(\Delta t)^2/4}\right)}$$

Dies ergibt sich eine ähnliche Bedingung wie unter (2), um die numerische Dispersion klein zu halten.

Unter den vorgeschlagenen Varianten ist also das implizite Verfahren (2) den übrigen deutlich vorzuziehen. Unter praktischen Gesichtspunkten muß auch hierbei die Schrittweitenwahl einer Kopplungsrestriktion unterworfen werden.

Bemerkung 1.3 *Es gibt selbstverständlich uneingeschränkt stabile Verfahren der Ordnung $p > 2$ für $y'' = f(y)$. Diese sind jedoch implizit und nichtlinear in f . Ein Beispiel ist*

$$u_{n+2}^h - 2u_{n+1}^h + u_n^h = \frac{(\Delta t)^2}{12}(f_{n+2} + 10f(u_{n+1}^h - \alpha(\Delta t)^2(f_{n+2} - 2f_{n+1} + f_n)) + f_n)$$

mit $\alpha > 1/120$ und der Ordnung $p = 4$ (Chawla 1983). Man muss jedoch bedenken, daß im Zusammenhang mit der Semidiskretisierung der Wellengleichung oder verwandter Gleichungen ohnehin die CFL-Bedingung der Verwendung grosser Zeitschrittweiten entgegensteht. \square

Bemerkung 1.4 *In neuerer Zeit werden auch Ansätze des sogenannten "exponential fittings" benutzt, um die numerische Dispersion besser zu kontrollieren. Hierbei werden freie Parameter des Verfahrens so angepaßt, daß für bestimmte (für die Problemstellung relevante) Frequenzen die numerische Dispersion zu null (oder doch sehr klein) gemacht wird. (vgl. z.B. van der Houwen & Sommeier, J. Comp. Appl. Math. 1985, S. 145-161) \square*

Bemerkung 1.5 *Für die in den Ingenieur Anwendungen wichtige DGL*

$$M\ddot{y} + C\dot{y} + Ky = f(t) \quad \begin{array}{l} M, C, K \text{ symm. } n \times n \text{ Matrizen,} \\ M, K \text{ pos.def.} \end{array}$$

(die man u.a. bei der Semidiskretisierung einer linearen hyperbolischen DGL 2. Ordnung mit einem Term u_t erhält), benutzt man gerne die Methode von Newmark von der Ordnung $p = 2$.

$$\begin{aligned} & (M + \gamma\Delta t C + \beta\Delta t^2 K)u_{n+1}^h + (-2M + (1 - 2\gamma)\Delta t C + (\frac{1}{2} + \gamma - 2\beta)(\Delta t)^2 K)u_n^h \\ & \quad + (M + (\gamma - 1)\Delta t C + (\frac{1}{2} - \gamma + \beta)(\Delta t)^2 K)u_{n-1}^h \\ & = (\Delta t)^2 \left((\frac{1}{2} - \gamma + \beta)f(t_{n-1}) + (\frac{1}{2} + \gamma - 2\beta)f(t_n) + \beta f(t_{n+1}) \right) \end{aligned}$$

Im Fall $C = 0$ und $2\beta \geq \gamma \geq \frac{1}{2}$ ist die Methode uneingeschränkt neutral stabil.

(Für $M = I$, $C = 0$, $K = (c^2/(\Delta x)^2)A$ $\gamma = \frac{1}{2}$, $\beta = \frac{1}{2}$ ergibt sich die obige Formel (2)) □

1.3.2 Numerische Verfahren für hyperbolische Systeme erster Ordnung

Da man hyperbolische Einzeldifferentialgleichungen stets in ein hyperbolisches System erster Ordnung überführen kann, wie wir bereits in der Einleitung gesehen haben, stellen letztere den allgemeineren Fall dar.

Systeme erster Ordnung haben den Vorteil, daß man sie mit Einschrittverfahren behandeln kann.

Speziell für das reine AWP der Wellengleichung

$$\left. \begin{aligned} u_{tt} - c^2 u_{xx} &= 0 & x \in \mathbb{R}, \quad t > 0 \\ u(x, 0) &= f(x) \\ u_t(x, 0) &= g(x) \end{aligned} \right\} x \in \mathbb{R}$$

erhält man durch die Substitution

$$\begin{aligned} v(x, t) &\stackrel{\text{def}}{=} u(x, t) \\ w(x, t) &\stackrel{\text{def}}{=} \frac{1}{c} \int_0^x \overbrace{u_t(\xi, 0)}^{g(\xi)} d\xi + c \int_0^t u_x(x, \tau) d\tau \\ v_t(x, t) &= u_t(x, t) = \int_0^t u_{tt}(x, \tau) d\tau + g(x) \\ &= c \left(c \int_0^t u_{xx}(x, \tau) d\tau + \frac{1}{c} g(x) \right) = c w_x(x, t) \\ w_t(x, t) &= c u_x(x, t) = c v_x(x, t) \\ v(x, 0) &= u(x, 0) = f(x) \\ w(x, 0) &= \frac{1}{c} \int_0^x g(\xi) d\xi =: G(x) \end{aligned}$$

d.h.

$$\begin{pmatrix} v \\ w \end{pmatrix}_t = c \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_x \quad t > 0, x \in \mathbb{R}, \quad \begin{pmatrix} v \\ w \end{pmatrix}(x, 0) = \begin{pmatrix} f(x) \\ G(x) \end{pmatrix}$$

Allgemein werden wir zunächst lineare Systeme der Dimension 2

$$v_t = Av_x \quad x \in \mathbb{R}, t > 0, \quad v(x, 0) = g(x) \quad g: \mathbb{R} \rightarrow \mathbb{R}^2 \quad (1.9)$$

betrachten, wobei wir voraussetzen, daß gilt:

A reell diagonalisierbar mit zwei verschiedenen Eigenwerten $\neq 0$

Dies ist insbesondere erfüllt für

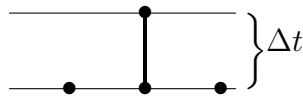
$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \quad (a - c)^2 + b^2 \geq \gamma > 0, \quad |ac - b^2| \geq \delta > 0$$

(a, b, c dürfen von x abhängen).

Die charakteristischen Richtungen dieses Systems im Punkt (x, t) sind dann $(1, 1/\lambda_1)$ und $(1, 1/\lambda_2)$, λ_1, λ_2 Eigenwerte von A .

Ein erstes naives Verfahren zur numerischen Lösung von (1.9) entsteht, indem man die t -Ableitung durch den Vorwärtsdifferenzenquotienten und die x -Ableitung durch den zentralen Differenzenquotienten ersetzt:

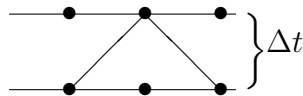
$$\begin{pmatrix} v_{j,n+1} \\ w_{j,n+1} \end{pmatrix} = \begin{pmatrix} v_{j,n} \\ w_{j,n} \end{pmatrix} + \frac{\Delta t}{2\Delta x} A \begin{pmatrix} v_{j+1,n} - v_{j-1,n} \\ w_{j+1,n} - w_{j-1,n} \end{pmatrix}, \quad \begin{pmatrix} v_{j,0} \\ w_{j,0} \end{pmatrix} = \begin{pmatrix} f_j \\ G_j \end{pmatrix}$$



Dieses “naive“ Verfahren erweist sich aber als unbrauchbar, da **instabil** für alle Verhältnisse $\Delta t : \Delta x$. Die Instabilität verschwindet, wenn man die Werte $v_{j,n}, w_{j,n}$ durch die Mittelwerte aus den Nachbarpunkten ersetzt:

Friedrichs–Verfahren

$$\begin{pmatrix} v_{j,n+1} \\ w_{j,n+1} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} v_{j+1,n} + v_{j-1,n} \\ w_{j+1,n} + w_{j-1,n} \end{pmatrix} + \frac{\Delta t}{2\Delta x} A \begin{pmatrix} v_{j+1,n} - v_{j-1,n} \\ w_{j+1,n} - w_{j-1,n} \end{pmatrix}, \quad \begin{pmatrix} v_{j,0} \\ w_{j,0} \end{pmatrix} = \begin{pmatrix} f_j \\ G_j \end{pmatrix}$$



Bei ARWA sind n Vorgaben für die Funktionen v und w an den Rändern des x -Intervalls gegeben, z.B.

$$v(0, t) = \varphi_0(t), \quad v(1, t) = \varphi_1(t),$$

sodaß man bei der Wellengleichung die obige Approximation für w an den Randnachbarpunkten nicht verwenden kann. In Abhängigkeit von den Randvorgaben muß man dann anders vorgehen.

Bei einer Randanfangswertaufgabe für die Wellengleichung

$$\begin{pmatrix} v \\ w \end{pmatrix}_t = c \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_x \quad t > 0, \quad 0 \leq x \leq 1, \quad \begin{pmatrix} v \\ w \end{pmatrix}(x, 0) = \begin{pmatrix} f(x) \\ G(x) \end{pmatrix}$$

$$v(0, t) = \varphi_0(t), \quad v(1, t) = \varphi_1(t)$$

kann man z.B. diskretisieren wie folgt (Courant, Friedrichs, Lewy 1928)

$$v_{j,n+1} = v_{j,n} + c \frac{\Delta t}{\Delta x} (w_{j+\frac{1}{2},n} - w_{j-\frac{1}{2},n}) \quad j = 1, \dots, m \quad \Delta x = \frac{1}{m+1}$$

$$v_{0,n+1} = \varphi_0(t_{n+1}), \quad v_{m+1,n+1} = \varphi_1(t_{n+1})$$

$$w_{j-\frac{1}{2},n+1} = w_{j-\frac{1}{2},n} + c \frac{\Delta t}{\Delta x} (v_{j,n+1} - v_{j-1,n+1}) \quad j = 1, \dots, m+1$$

Das Friedrichs–Verfahren konvergiert nur von erster Ordnung in Δt . Das folgende Verfahren für ein allgemeines hyperbolisches System mit konstanten Koeffizienten konvergiert von zweiter Ordnung in Δt und Δx :

$$u_{i,n+1}^h = u_{i,n}^h + \frac{\Delta t}{2\Delta x} A(u_{i+1,n}^h - u_{i-1,n}^h) + \frac{1}{2} \left(\frac{\Delta t}{\Delta x} \right)^2 A^2(u_{i+1,n}^h - 2u_{i,n}^h + u_{i-1,n}^h)$$

$u_{i,n}^h$ steht hier als Näherung für den Vektor $u(x_i, t_n)$, DGL $u_t = Au_x$. Dies ist das

Lax–Wendroff–Verfahren

Dieses Verfahren ist stabil für

$$\frac{\Delta t}{\Delta x} \rho(A) \leq 1$$

Man kann es auf folgende Art herleiten: Die rechte Seite Au_x wird mit dem zentralen Differenzenquotienten diskretisiert, d.h. von zweiter Ordnung in Δx . Für die linke Seite u_t benutzt man die Diskretisierung

$$u(x, t_{n+1}) = u(x, t_n) + u_t(x, t_n)\Delta t + \frac{1}{2}u_{tt}(x, t_n)(\Delta t)^2 + \mathcal{O}((\Delta t)^3)$$

und sodann

$$u_{tt} = A u_{xt} = A u_{tx} = A \frac{\partial}{\partial x} (A u_x) = A^2 u_{xx}$$

und für u_{xx} nun den symmetrischen Differenzenquotienten zweiter Ordnung.

Im Spezialfall $n = 1$ heißt die DGL

$$u_t = a(x)u_x \quad |a'(x)| \leq K$$

die Konvektionsgleichung. Häufig verwendete Diskretisierungen dieser Gleichung sind:

Friedrichs–Verfahren:

$$u_{i,n+1}^h = \frac{1}{2}(u_{i-1,n}^h + u_{i+1,n}^h) + \frac{\Delta t}{2\Delta x} a(x_i)(u_{i+1,n}^h - u_{i-1,n}^h)$$

$$\frac{\Delta t}{\Delta x} \leq 1 / \sup |a(x)|$$

Verfahren von Courant–Isaacson und Rees:

$$u_{i,n+1}^h = u_{i,n}^h + \frac{\Delta t}{\Delta x} \left(a^+(x_i)(u_{i+1,n}^h - u_{i,n}^h) + a^-(x_i)(u_{i,n}^h - u_{i-1,n}^h) \right)$$

$$\frac{\Delta t}{\Delta x} \leq 1 / \sup |a(x)|$$

$$a^+(x) = \max(a(x), 0)$$

$$a^-(x) = \min(a(x), 0)$$

(Man beachte, daß hier die x -Diskretisierung von Verlauf der Charakteristiken abhängt!)

Lax–Wendroff–Verfahren:

$$u_{i,n+1}^h = u_{i,n}^h + \frac{\Delta t}{2\Delta x} a(x_i)(u_{i+1,n}^h - u_{i-1,n}^h) + \frac{1}{2} \left(\frac{\Delta t}{\Delta x} \right)^2 a(x_i) \cdot \\ \cdot \left(a(x_{i+\frac{1}{2}})(u_{i+1,n}^h - u_{i,n}^h) - a(x_{i-\frac{1}{2}})(u_{i,n}^h - u_{i-1,n}^h) \right)$$

$$\frac{\Delta t}{\Delta x} \leq 1/\sup |a(x)|$$

Die Konvergenzuntersuchungen für alle diese Verfahren werden wir im Rahmen der allgemeinen Theorie in Kapitel 3 behandeln. Diese allgemeine Theorie behandelt in erster Linie Probleme mit konstanten Koeffizienten und variable Koeffizienten erfordern zusätzliche Betrachtungen. Oft ist es aber möglich, mit Methoden, die auf den speziellen Fall zugeschnitten sind, einfacher zum Ziel zu gelangen. Dies sei hier an einem Beispiel vorgeführt.

Beispiel 1.3 $u_t = -a(x, t)u_x \quad 0 \leq x \leq L, \quad 0 \leq t \leq T$

$$u(x, 0) = f(x), \quad 0 \leq x \leq L$$

$$u(0, t) = g(t), \quad 0 \leq t \leq T.$$

Es gelte:

$$f(0) = g(0)$$

$$f, g \quad \text{stetig}$$

$$a \in C^2(\mathbb{R}_+ \times \mathbb{R}_+)$$

$$0 < a_0 \leq a(x, t) \leq a_1 \quad (x, t) \in [0, L] \times [0, T]$$

$\lambda := \Delta t/\Delta x$ fest, $\Delta x = \frac{1}{M}$, $\Delta t = \frac{1}{N}$, $T = N\Delta t$, $L = M\Delta x$,
Gitterwerte $x_j = j\Delta x$, $t_n = n\Delta t$.

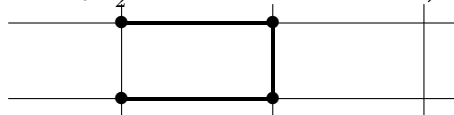
Wir benutzen folgende Differenzenapproximation:

$$u_{j,n+1}^h = u_{j,n}^h - \frac{1}{2}\lambda \left(a_{j-\frac{1}{2},n} (u_{j,n}^h - u_{j-1,n}^h) + a_{j-\frac{1}{2},n+1} (u_{j,n+1}^h - u_{j-1,n+1}^h) \right)$$

$$u_{0,n}^h = g(t_n)$$

$$u_{0,n+1}^h = g(t_{n+1})$$

$$u_{j,0}^h = f(x_j)$$



Bestimmung der Konsistenzordnung:

$$\begin{aligned}
u_{j,n+1} - u_{j,n} &= (u_{tt})_{j,n+\frac{1}{2}} \frac{(\Delta t)^2}{8} + \mathcal{O}((\Delta t)^3) \\
a_{j-\frac{1}{2},n}(u_{j,n} - u_{j-1,n}) &= a_{j-\frac{1}{2},n} \left((u_{xx})_{j-\frac{1}{2},n} \frac{(\Delta x)^2}{8} + \mathcal{O}((\Delta x)^4) \right) \\
a_{j-\frac{1}{2},n+1}(u_{j,n+1} - u_{j-1,n+1}) &= a_{j-\frac{1}{2},n+1} \left((u_{xx})_{j-\frac{1}{2},n+1} \frac{(\Delta x)^2}{8} + \mathcal{O}((\Delta x)^4) \right) \\
\frac{1}{2} \left(a_{j-\frac{1}{2},n}(u_{j,n} - u_{j-1,n}) + a_{j-\frac{1}{2},n+1}(u_{j,n+1} - u_{j-1,n+1}) \right) &= \\
&= a_{j-\frac{1}{2},n+\frac{1}{2}} (u_{xx})_{j-\frac{1}{2},n+\frac{1}{2}} \frac{(\Delta x)^2}{8} + \mathcal{O}((\Delta t)^2(\Delta x)^2) + \mathcal{O}((\Delta x)^4) \\
(u_{tt})_{j,n+\frac{1}{2}} \frac{(\Delta t)^2}{8} &= (u_{tt})_{j-\frac{1}{2},n+\frac{1}{2}} \frac{(\Delta t)^2}{8} + \mathcal{O}(\Delta x(\Delta t)^2).
\end{aligned}$$

Es ist aber aufgrund der DGL

$$\begin{aligned}
u_{tt} = - \left(a(x,t) u_x \right)_t &= -a_t(x,t) u_x - a(x,t) u_{xt} \\
&= -a_t(x,t) u_x + a(x,t) \left(a(x,t) u_x \right)_x \\
&= -a_t(x,t) u_x + a(x,t) a_x(x,t) u_x + a^2(x,t) u_{xx}.
\end{aligned}$$

Somit ist im Falle

$$a(x,t) \equiv 1 = \lambda$$

das Verfahren von zweiter Ordnung und in allen anderen Fällen von erster Ordnung konsistent ($h \hat{=} \Delta t = \mathcal{O}(\Delta x)$).

Sei

$$\varepsilon_{j,n} = u_{j,n} - u_{j,n}^h.$$

Dann gilt die Rekursion $j = 1, \dots, N$

$$\begin{aligned}
\varepsilon_{j,n+1} &= \varepsilon_{j,n} - \frac{1}{2} \left(\gamma_{j,n}(\varepsilon_{j,n} - \varepsilon_{j-1,n}) + \gamma_{j,n+1}(\varepsilon_{j,n+1} - \varepsilon_{j-1,n+1}) \right) + \Delta t \tau_{j,n+1} \\
\tau_{j,n+1} &= \mathcal{O}(\Delta t) \quad \text{und} \\
\gamma_{j,k} &:= \lambda a_{j-\frac{1}{2},k} > \lambda a_0 > 0 \\
\varepsilon_{0,k} &= 0 \quad \forall k, \\
\varepsilon_{j,0} &= 0 \quad \forall j.
\end{aligned}$$

Setzt man also $\varepsilon_n := (\varepsilon_{1,n}, \dots, \varepsilon_{M,n})^T$, dann wird

$$(I + \Gamma_{n+1})\varepsilon_{n+1} = (I - \Gamma_n)\varepsilon_n + \Delta t \tau_{n+1}, \quad n = 0, 1, \dots, N-1$$

wo

$$\Gamma_k = \begin{pmatrix} \gamma_{1,k} & 0 & \cdots & \cdots & 0 \\ -\gamma_{2,k} & \gamma_{2,k} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\gamma_{M,k} & \gamma_{M,k} \end{pmatrix}$$

bzw. mit

$$\begin{aligned}\tilde{\varepsilon}_n &:= (I + \Gamma_n)\varepsilon_n \\ \tilde{\varepsilon}_{n+1} &= (I - \Gamma_n)(I + \Gamma_n)^{-1}\tilde{\varepsilon}_n + \Delta t \tau_{n+1}, \quad n = 0, 1, \dots, N-1.\end{aligned}$$

Wir wollen wieder eine Abschätzung in der Norm

$$\|\cdot\| = \frac{1}{\sqrt{M}} \|\cdot\|_2$$

anstreben. Jedenfalls ist

$$\sup_{\substack{n \in \mathbb{N} \\ n\Delta t = T \\ \Delta t \rightarrow 0}} \|\tau_{n+1}\| \leq \tau = \mathcal{O}(\Delta t).$$

Wenn

$$\|(I - \Gamma_n)(I + \Gamma_n)^{-1}\| \leq 1 + K\Delta t, \quad (1.10)$$

K unabhängig von n und Δt , dann wird schließlich

$$\begin{aligned}\|\varepsilon_N\| &\leq \Delta t \tau \sum_{\nu=0}^{N-1} (1 + K\Delta t)^\nu \\ &\leq \Delta t \tau N (1 + K\frac{T}{N})^N \leq \tau T e^{KT},\end{aligned}$$

womit die Konvergenz in $\|\cdot\|$ bewiesen wäre.

Dabei tritt keine Einschränkung an λ auf. Da das Gleichungssystem für ε_{n+1} eine untere Dreiecksmatrix als Koeffizientenmatrix hat, ist das Verfahren quasi explizit.

Es bleibt die Abschätzung (1.10) zu zeigen. Dazu beachtet man zuerst

$$\|(I - \Gamma_n)(I + \Gamma_n)^{-1}\| = \|(I - \Gamma_n)(I + \Gamma_n)^{-1}\|_2.$$

Ferner nach Definition $\|A\|_2 = \varrho^{\frac{1}{2}}(AA^T)$, also

$$\begin{aligned}\|(I - \Gamma_n)(I + \Gamma_n)^{-1}\|_2^2 &= \varrho\left((I - \Gamma_n)(I + \Gamma_n)^{-1}(I + \Gamma_n^T)^{-1}(I - \Gamma_n^T)\right) \\ &= \varrho\left((I + \Gamma_n)^{-1}(I + \Gamma_n^T)^{-1}(I - \Gamma_n^T)(I - \Gamma_n)\right) \\ &= \varrho\left(((I + \Gamma_n^T)(I + \Gamma_n))^{-1}(I - \Gamma_n^T)(I - \Gamma_n)\right)\end{aligned}$$

Sei μ ein Eigenwert von $((I + \Gamma_n^T)(I + \Gamma_n))^{-1}(I - \Gamma_n^T)(I - \Gamma_n)$.

Dann gilt mit geeignetem x und $x^H x = 1$

$$((I - \Gamma_n^T)(I - \Gamma_n)x = \mu(I + \Gamma_n^T)(I + \Gamma_n)x$$

d.h.

$$\begin{aligned}
\mu &= \frac{x^T(I - \Gamma_n^T)(I - \Gamma_n)x}{x^T(I + \Gamma_n^T)(I + \Gamma_n)x} \\
&= \frac{1 - x^T(\Gamma_n^T + \Gamma_n)x + x^T\Gamma_n^T\Gamma_n x}{1 + x^T(\Gamma_n^T + \Gamma_n)x + x^T\Gamma_n^T\Gamma_n x} \\
&= 1 - 2\frac{x^T(\Gamma_n^T + \Gamma_n)x}{1 + x^T\Gamma_n^T\Gamma_n x + x^T(\Gamma_n^T + \Gamma_n)x}.
\end{aligned}$$

Wenn also gezeigt ist, daß

$$x^T(\Gamma_n^T + \Gamma_n)x \geq -\tilde{K}\Delta t, \quad (1.11)$$

dann

$$\mu \leq 1 + \frac{2\tilde{K}\Delta t}{1 - \tilde{K}\Delta t} \leq 1 + 4\tilde{K}\Delta t$$

falls $\tilde{K}\Delta t \leq \frac{1}{2}$, also Δt hinreichend klein.

Damit wird dann

$$\|(I - \Gamma_n)(I + \Gamma_n)^{-1}\| \leq \sqrt{1 + 4\tilde{K}\Delta t} \leq 1 + 2\tilde{K}\Delta t =: 1 + K\Delta t.$$

Somit bleibt (1.11) zu zeigen.

Aber

$$\Gamma_k + \Gamma_k^T = \begin{pmatrix} 2\gamma_{1,k} & -\gamma_{2,k} & 0 & \cdots & 0 \\ -\gamma_{2,k} & 2\gamma_{2,k} & -\gamma_{3,k} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\gamma_{M,k} \\ 0 & \cdots & 0 & -\gamma_{M,k} & 2\gamma_{M,k} \end{pmatrix}$$

und nach dem Kreisesatz von Gerschgorin und der Minimax-Eigenschaft des Rayleighquotienten

$$\begin{aligned}
x^H(\Gamma_k + \Gamma_k^T)x &\geq -\sup_{k,i} |\gamma_{i+1,k} - \gamma_{i,k}| \\
&= -\lambda \sup_{k,i} |a_{i+\frac{1}{2},k} - a_{i-\frac{1}{2},k}| \\
&\geq -\lambda\beta\Delta x = -\beta\Delta t,
\end{aligned}$$

wo $\beta = \max_{(x,t) \in [0,L] \times [0,T]} \left| \frac{\partial}{\partial x} a(x,t) \right|$

Schließlich bleibt zu zeigen, daß

$$\|\varepsilon_n\| \leq K^* \|\tilde{\varepsilon}_n\|. \quad (K^* \text{ unabh. von } n, \Delta t)$$

Aber

$$\varepsilon_n = (I + \Gamma_n)^{-1} \tilde{\varepsilon}_n$$

d.h.

$$\begin{aligned}\|\varepsilon_n\| &\leq \varrho^{1/2} \left((I + \Gamma_n)(I + \Gamma_n^T) \right)^{-1} \|\tilde{\varepsilon}_n\| \\ &= \frac{1}{\lambda_{\min}^{1/2}((I + \Gamma_n)(I + \Gamma_n^T))} \|\tilde{\varepsilon}_n\| \leq \frac{1}{1 - \beta\Delta t} \|\tilde{\varepsilon}_n\|\end{aligned}$$

d.h. für $\beta\Delta t \leq \frac{1}{2}$ (z.B.) kann man $K^* = 2$ wählen, womit alles bewiesen ist. □

1.4 Nichtlineare hyperbolische Erhaltungsgleichungen

Quelle für diesen Abschnitt: [7], [10]. Weiterführende Literatur ist [17], [18]

1.4.1 Beispiele

Eine eindimensionale Strömung wird im einfachsten Fall beschrieben durch die raum- und zeitabhängige Dichte und Geschwindigkeit. Bei fester Zeit t trägt ein Intervall $[x_1, x_2]$ die Masse

$$\int_{x_1}^{x_2} \varrho(x, t) dx .$$

Den Punkt x durchfließt im Zeitintervall $[t_1, t_2]$ die Masse

$$\int_{t_1}^{t_2} \varrho(x, t) v(x, t) dt .$$

Nach dem Gesetz von der Erhaltung der Masse muss gelten:

Die Änderung der Masse eines Raumintervalls zwischen zwei Zeitpunkten ist gleich der Differenz der Massenflüsse an den Intervallenden genommen über dieses Zeitintervall

D.h.

$$\int_{x_1}^{x_2} \varrho(x, t_2) dx - \int_{x_1}^{x_2} \varrho(x, t_1) dx = \int_{t_1}^{t_2} \varrho(x_1, t) v(x_1, t) dt - \int_{t_1}^{t_2} \varrho(x_2, t) v(x_2, t) dt$$

also

$$\int_{x_1}^{x_2} \int_{t_1}^{t_2} \frac{\partial}{\partial t} \varrho(x, t) dt dx = - \int_{t_1}^{t_2} \int_{x_1}^{x_2} \frac{\partial}{\partial x} (\varrho(x, t) v(x, t)) dx dt$$

In differentieller Form lautet dies

$$\frac{\partial}{\partial t} \varrho = - \frac{\partial}{\partial x} (\varrho v) .$$

Man benötigt eine Beziehung zwischen v und ϱ , um daraus eine abgeschlossene Gleichung für ϱ zu erhalten. Ein vernünftiges Modell im Bereich der Verkehrsflüsse ist z.B.

$$v = v(\varrho) = v_{\max} \left(1 - \frac{\varrho}{\varrho_{\max}} \right)$$

Damit wird die Differentialgleichung zu

$$\frac{\partial}{\partial t} \varrho + \frac{\partial}{\partial x} (f(\varrho)) = 0$$

mit der Flussfunktion

$$f(\varrho) = \varrho v_{\max} \left(1 - \frac{\varrho}{\varrho_{\max}}\right).$$

Dies ist die Standardform einer skalaren Erhaltungsgleichung. In der Praxis spielen vor allem Systeme solcher Gleichungen eine wichtige Rolle, z.B. die Eulergleichungen der Gasdynamik

$$\begin{aligned} \frac{\partial}{\partial t} \begin{pmatrix} \varrho \\ \varrho v \\ E \end{pmatrix} + \frac{\partial}{\partial x} \begin{pmatrix} \varrho v \\ p + \varrho v^2 \\ v(E + p) \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ p - (\gamma - 1)(E - \frac{1}{2}\varrho v^2) &= 0 \end{aligned}$$

Hierbei bedeuten ϱ , v , p , E die Dichte, Geschwindigkeit, Druck und die totale Energie und γ ist eine Materialkonstante.

1.4.2 Theorie

Im allgemeinen Fall hat ein System von Erhaltungsgleichungen folgende Form:

$$\begin{aligned} u &= u(x, t) : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^n \\ F_i &= F_i(u) : \mathbb{R}^n \rightarrow \mathbb{R}^n \\ u_t + \sum_{i=1}^d (F_i(u))_{x_i} &= 0. \end{aligned} \tag{1.12}$$

Definition 1.4 Das System (1.12) heißt hyperbolisch, falls folgendes erfüllt ist:

Sei

$$A_i(u) \stackrel{\text{def}}{=} \mathcal{J}_{F_i}(u) = \left(\frac{\partial (F_i)_j}{\partial u_k} \right)_{1 \leq j, k \leq n}$$

Dann hat die Matrix

$$B(u, w) \stackrel{\text{def}}{=} \sum_{i=1}^d w_i A_i(u)$$

für alle $w \neq 0$ und alle $u \in \mathbb{R}^n$ ein vollständiges reelles Eigensystem. Sind alle Eigenwerte stets paarweise verschieden, heißt das System strikt hyperbolisch.

Wir betrachten nun den skalaren Fall $n = d = 1$. Das reine Anfangswertproblem hat die Form

$$\begin{aligned} \frac{\partial}{\partial t} u + \frac{\partial}{\partial x} (f(u)) &= 0 \\ u(x, 0) &= u_0(x), \quad x \in \mathbb{R} \end{aligned}$$

Den linearen Fall $f(u) = au$ mit festem a haben wir bereits in Abschnitt 1.1 behandelt. Hier interessiert uns der nichtlineare Fall. Es können ganz neue Phänomene auftreten, insbesondere muss der Lösungsbegriff neu im Sinne einer schwachen Lösung verallgemeinert werden. Während im linearen Fall bei glatten Daten und Kompatibilität in den Anfangs-Randbedingungen eine klassische differenzierbare Lösung vorliegt, wovon wir in den vorausgegangenen Abschnitten auch immer ausgegangen sind, ist dies nun nicht mehr notwendig der Fall. Wir betrachten als Beispiel die Burgers Gleichung

$$u_t + \left(\frac{u^2}{2}\right)_x = 0.$$

Eine charakteristische Kurve in der Parametrisierung $(x(t), t)$ ist hier gegeben durch das Anfangswertproblem

$$\dot{x}(t) = u(x(t), t) \quad x(0) = x_0$$

und $u(x(t), t)$ ist längs der Charakteristik konstant, also

$$u(x(t), t) = u(x(0), 0) = u_0(x_0)$$

Also wird

$$x(t) = x_0 + u_0(x_0)t$$

d.h. die Charakteristiken sind Geraden. Ist nun

$$u'_0(x) > 0 \quad \forall x$$

dann können sich zwei Charakteristiken niemals schneiden und für glatte Anfangsfunktion u_0 erhält man eine eindeutige klassische Lösung der Aufgabe. Ist aber $u'_0 < 0$, dann gibt es immer eine Lösung von

$$x_0 + u_0(x_0)t^* = x_1 + u_0(x_1)t^*, \quad t^* > 0, \quad x_1 > x_0$$

und dies bedeutet, daß in $(x(t^*), t^*)$ die Lösung unstetig wird. Physikalisch bedeutet dies einen Schock. Die gleichen Überlegungen gelten ganz allgemein für eine skalare Erhaltungsgleichung

$$u_t + (f(u))_x = 0, \quad u(x, 0) = u_0(x)$$

und wir haben

Satz 1.5 Sei $f'' > 0$ und $u'_0 > 0$ auf \mathbb{R} . Dann hat die Anfangswertaufgabe der Erhaltungsgleichung für alle Zeiten eine eindeutige klassische Lösung.

Wir untersuchen weiter den Fall nichtklassischer Lösungen. Es ist eine naheliegende Idee, schwache Lösungen durch Multiplikation der DGL mit einer Testfunktion und Integration zu charakterisieren: Ist u eine klassische Lösung, dann gilt

$$\int_{\mathbb{R}} \int_0^{\infty} (\phi(x, t)u_t(x, t) + \phi(x, t)(f(u))_x) dt dx = 0 \quad \forall \phi \in C^1(\mathbb{R} \times \mathbb{R}_+)$$

wobei ϕ noch kompakten Träger haben muss. Der erste Summand wird nun bezüglich t , der zweite bezüglich x partiell integriert

$$\begin{aligned} & - \int_{\mathbb{R}} \int_0^{\infty} (\phi_t(x, t) u(x, t)) dt dx + \int_{\mathbb{R}} (\phi \cdot u)(x, t) \Big|_{t=0}^{t=\infty} dx \\ & - \int_{\mathbb{R}} \int_0^{\infty} (\phi_x(x, t) f(u(x, t))) dt dx + \int_0^{\infty} (\phi \cdot f(u))(x, t) \Big|_{x=-\infty}^{x=\infty} dt = 0 \end{aligned}$$

Berücksichtigt man nun, daß ϕ kompakten Träger hat, gelangt man zu

Definition 1.5 $u = u(x, t)$ heißt schwache Lösung von

$$u_t + (f(u))_x = 0, \quad u(x, 0) = u_0(x), \quad x \in \mathbb{R}, \quad t \geq 0 \quad (1.13)$$

falls

$$\int_{\mathbb{R}} \int_0^{\infty} (u \phi_t + f(u) \phi_x) dt dx + \int_{\mathbb{R}} u_0(x) \phi(x, 0) dx = 0 \quad \forall \phi \in C_0^1(\mathbb{R} \times \mathbb{R}_+).$$

Ein Beispiel für das Auftreten einer solchen schwachen Lösung ist ein ‘‘Riemannproblem‘‘ mit den Anfangswerten

$$\begin{aligned} u_t + (f(u))_x &= 0 \\ u(x, 0) = u_0(x) &= \begin{cases} u_l & \text{falls } x < 0, \\ u_r & \text{falls } x > 0 \end{cases}. \end{aligned}$$

Man rechnet mit Hilfe der Definition nach, daß für $u_l > u_r$ die Schockwelle

$$u(x, t) = \begin{cases} u_l & \text{falls } x < st \\ u_r & \text{falls } x > st \end{cases}$$

eine schwache Lösung dieses Problems ist, wobei die Schockgeschwindigkeit s die sogenannte **Rankine-Hugoniot Bedingung**

$$s = \frac{f(u_l) - f(u_r)}{u_l - u_r}$$

erfüllt. Man beachte, daß s als ein Zwischenwert von f' geschrieben werden kann, f' ist monoton wachsend, da wir f als konvex vorausgesetzt haben. Es gilt ganz allgemein

Satz 1.6 Sei u eine schwache Lösung des Anfangswertproblems (1.13). u sei unstetig längs der Kurve $(x(t), t)$ und $s(t) = \dot{x}(t)$. Ist

$$u_l(t) \stackrel{\text{def}}{=} \lim_{\varepsilon \searrow 0} u(x(t) - \varepsilon, t)$$

$$u_r(t) \stackrel{\text{def}}{=} \lim_{\varepsilon \searrow 0} u(x(t) + \varepsilon, t)$$

dann gilt

$$s(t) = \frac{f(u_r(t)) - f(u_l(t))}{u_r(t) - u_l(t)} .$$

Die Charakterisierung einer schwachen Lösung legt diese jedoch nicht eindeutig fest und es gibt in der Tat Probleme des Typs (1.13) mit nichteindeutigen Lösungen:

Beispiel 1.1

$$\begin{aligned} u_t + uu_x &= 0 \\ u(x, 0) &= \begin{cases} 0, & x < 0, \\ 1, & x > 0 \end{cases} \end{aligned}$$

Die Charakteristiken durch den Punkt $(x_0, 0)$ sind hier

$$(x_0, t) \text{ für } x_0 < 0 \text{ und } (x(t) = x_0 + t, t) \text{ für } x_0 > 0 .$$

Für jedes $\alpha \in]0, 1[$ ist

$$u(x, t) = \begin{cases} 0, & \text{für } x < \alpha t/2, \\ \alpha, & \text{für } \alpha t/2 < x < (1 + \alpha)t/2 \\ 1, & \text{für } x > (1 + \alpha)t/2 \end{cases}$$

eine schwache Lösung, wie man durch Einsetzen und explizites Ausintegrieren in der Definition bestätigt. Jede dieser unendlich vielen unstetigen Lösungen erfüllt auch die Rankine-Hugoniot-Bedingung. \square

Man muß also nach zusätzlichen Charakterisierungen suchen, die eine physikalisch sinnvolle Lösung eindeutig festlegen (die Dichte eines Gases kann z.B. nicht mehrdeutig sein). Um eine solche Bedingung zu finden, gehen wir zunächst von einer klassischen Lösung aus. Ist H eine konvexe zweimal stetig differenzierbare Funktion, dann ist

$$(H(u(x, t)))_t = H'(u(x, t))u_t = -(H'f')(u(x, t))u_x .$$

F heisst Entropie-Fluß und H Entropie, falls gilt

$$F' = H'f'$$

(z.B. mit $H(w) = \frac{1}{2}w^2 : F(w) = \int_0^w v f'(v)dv$). Für eine solches Entropie-Entropieflußpaar gilt also

$$\begin{aligned} (H(u))_t + (F(u))_x &= H'(u)u_t + F'(u)u_x = \\ H'(u)u_t + H'(u)f'(u)u_x &= H'(u)(u_t + (f(u))_x) = 0 . \end{aligned}$$

Wir definieren u^ε als die eindeutige Lösung von

$$(H(u^\varepsilon))_t + (F(u^\varepsilon))_x = \varepsilon u_{xx}^\varepsilon H'(u^\varepsilon) = \varepsilon (H(u^\varepsilon))_{xx} - \varepsilon H''(u^\varepsilon)(u_x^\varepsilon)^2, \quad u^\varepsilon(x, 0) = u_0(x) .$$

Man beachte, daß für streng konvexes H (also $H' \neq 0$) u^ε sich berechnet aus

$$(u^\varepsilon)_t = \varepsilon (u^\varepsilon)_{xx} - f'(u^\varepsilon)(u^\varepsilon)_x$$

also aus einer parabolischen Randanfangswertaufgabe, die unabhängig ist vom gewählten Entropie-Entropiefluß-Paar. Multiplikation mit einer nichtnegativen Testfunktion ϕ aus $C^1(\mathbb{R} \times \mathbb{R}_+)$ mit kompaktem Träger und partielle Integration ergibt

$$\begin{aligned} &\int_{\mathbb{R}} \int_0^\infty (H(u^\varepsilon(x, t))\phi_t(x, t) + F(u^\varepsilon(x, t))\phi_x(x, t)) dt dx + \int_{\mathbb{R}} H(u_0(x, 0))\phi(x, 0) dx \\ &= \varepsilon \int_{\mathbb{R}} \int_0^\infty (H(u^\varepsilon(x, t))_x \phi_x(x, t) + H''(u^\varepsilon)(u_x^\varepsilon)^2 \phi(x, t)) dt dx \end{aligned}$$

Wegen $H'' \geq 0$ und $\phi \geq 0$ ist der zweite Summand auf der rechten Seite nichtnegativ. Man kann nun zeigen, daß in $L_\infty(\mathbb{R} \times \mathbb{R})$ die Funktion u^ε für $\varepsilon \rightarrow 0$ einen Grenzwert u hat und für diesen gilt dann

$$\int_{\mathbb{R}} \int_0^\infty (H(u(x, t))\phi_t(x, t) + F(u(x, t))\phi_x(x, t)) dt dx + \int_{\mathbb{R}} H(u_0(x, 0))\phi(x, 0) dx \geq 0 \quad \forall \phi \geq 0 . \tag{1.14}$$

Man bezeichnet eine schwache Lösung der Erhaltungsgleichung als Entropielösung, wenn sie die Bedingung (1.14) für jedes Entropie-Entropieflußpaar erfüllt. Für die Entropielösung kann man Eindeutigkeit zeigen. Dazu gilt

Satz 1.7 (Kruskov) *Das skalare Anfangswertproblem*

$$u_t + (f(u))_x = 0, \quad u(x, 0) = u_0(x)$$

mit $f \in C^1(\mathbb{R})$ und $u_0 \in L_\infty(\mathbb{R})$ hat eine eindeutige Entropielösung $u \in L_\infty(\mathbb{R} \times \mathbb{R}_+)$ mit folgenden Eigenschaften:

1. $\|u(\cdot, t)\|_{L_\infty} \leq \|u_0\|_{L_\infty}$ (L_∞ -Stabilität)
2. $u_0 \geq v_0 \Rightarrow u(\cdot, t) \geq v(\cdot, t)$ (Monotonie bezüglich des Anfangswertes)
3. $u_0 \in BV(\mathbb{R}) \Rightarrow u(\cdot, t) \in BV(\mathbb{R})$ und $TV(u(\cdot, t)) \leq TV(u_0)$ (TV-Stabilität)
4. $u_0 \in L_1(\mathbb{R}) \Rightarrow \int_{\mathbb{R}} u(x, t) dx = \int_{\mathbb{R}} u_0(x) dx \quad \forall t \geq 0$. (Konservativität)

In diesem Satz werden zwei hier noch nicht definierte Begriffe benutzt: BV der Raum der Funktionen von beschränkter Variation und TV die Totalvariation

Definition 1.6 Sei $u \in L_\infty(\Omega)$, $\Omega \subset \mathbb{R}$ offen. Dann heißt

$$TV(u) \stackrel{def}{=} \limsup_{\varepsilon \rightarrow 0} \int_{\Omega} \frac{|u(x+\varepsilon) - u(x)|}{\varepsilon} dx$$

die Totalvariation (oder totale Variation) von u . Der Raum der Funktionen von beschränkter Variation ist

$$BV(\Omega) = \{u : u \in L_\infty(\Omega), TV(u) < \infty\}$$

Bemerkung 1.6 Der Satz gilt auch für skalare Gleichungen mit mehreren Dimensionen im Raum. Er gilt aber nicht für Systeme, für die es keine vergleichbaren Aussagen gibt.

Bemerkung 1.7 Für stückweise stetige Lösungen skalarer Erhaltungsgleichungen gilt auch noch eine L_1 -Stabilität

$$\|u(\cdot, t + \tau)\|_{L_1} \leq \|u(\cdot, t)\|_{L_1}$$

wie Lax 1972 gezeigt hat.

1.4.3 Numerische Verfahren

Bereits in Abschnitt 1.3 haben wir Differenzenverfahren für hyperbolische Gleichungen 1. Ordnung beschrieben. Diese sollen hier weiter untersucht werden. Eine allgemeine Theorie der Konvergenz solcher Verfahren wird in einem späteren Kapitel dargestellt. Hier ist es jedoch zweckmässig, eine spezielle Darstellung zu wählen. Im Folgenden ist

$$u_{i,j}^h \text{ Näherung für } u(x_i, t_j)$$

und h bedeutet den Diskretierungsparameter, unter dem man sich z.B. Δt vorstellen kann, da wegen der CFL-Bedingung ohnehin eine Kopplung zwischen Δt und Δx besteht. Wir betrachten hier explizite Einschrittverfahren (d.h. Verfahren, die nur zwei aufeinanderfolgende Zeitschichten verkoppeln) der allgemeinen Gestalt

$$u_{i,j+1}^h = \Psi(u_{i-m,j}^h, \dots, u_{i+m,j}^h)$$

mit einer stetigen Funktion

$$\Psi : \mathbb{R}^{2m+1} \rightarrow \mathbb{R}.$$

In den Beispielen von Abschnitt 1.3 war stets $m = 1$. Wir nehmen weiter an, $\Delta t = h$ sei fest und ebenfalls

$$\lambda = \frac{\Delta t}{\Delta x} \text{ fest.}$$

Oft ist es zweckmässig, die Struktur der Funktion Ψ als Integrator in einem Einschrittverfahren (wie im Fall gewöhnlicher Differentialgleichungen) genauer zu beschreiben:

Definition 1.7 Ein Differenzenverfahren besitzt die Erhaltungseigenschaft, wenn es eine Funktion

$$F : \mathbb{R}^{2m} \rightarrow \mathbb{R} ,$$

die sogenannte "numerische Flußfunktion" gibt, mit

$$\Psi(u_{-m}, \dots, u_m) = u_0 - \lambda(F(u_{-m+1}, \dots, u_m) - F(u_{-m}, \dots, u_{m-1})) .$$

Das Verfahren heißt dann auch **konservativ**.

Der Grund für diese Bezeichnung ist offensichtlich: es gilt dann

$$\sum_{i \in \mathbb{Z}} u_{i,j}^h = \sum_{i \in \mathbb{Z}} u_{i,0}^h \quad \forall j \in \mathbb{N} .$$

Wegen der entsprechenden Eigenschaft der wahren Lösung (der Erhaltungseigenschaft) wird man nur solche Verfahren in die Betrachtung einbeziehen. Subtrahieren wir die Verfahrensvorschrift eines konservativen Verfahrens von der Differentialgleichung, dann erhalten wir

$$u_t(x_i, t_j) - \frac{u_{i,j+1}^h - u_{i,j}^h}{\Delta t} = (f'(u)u_x)(x_i, t_j) - \sum_{k=1}^{2m} \partial_k F(u_{i-m,j}^h, \dots, u_{i+m-1,j}^h) \frac{u_{i-m+k,j}^h - u_{i-m+k-1,j}^h}{\Delta x} + \mathcal{O}((\Delta x)^2) .$$

und wenn wir vernünftigerweise annehmen, daß die Differenzenquotienten gegen die wahren Ableitungen konvergieren, wenn Δt und Δx gegen null gehen, dann sehen wir, daß

$$\sum_{k=1}^{2m} \partial_k F(u, \dots, u) = f'(u)$$

gelten sollte. Hinreichend dafür ist die Konsistenzbedingung:

Definition 1.8 Die numerische Flußfunktion F heißt konsistent mit der kontinuierlichen Flußfunktion f wenn gilt

$$F(u, \dots, u) = f(u) .$$

Bemerkung 1.8 Konsistenz impliziert also für Ψ

$$\Psi(u, \dots, u) = u .$$

Beispiel 1.2 Für das Friedrichs-Verfahren ist

$$u_{i,j+1}^h = u_{i,j}^h - \frac{\lambda}{2}(f(u_{i+1,j}^h) - f(u_{i-1,j}^h)) + \frac{u_{i+1,j}^h - u_{i,j}^h - (u_{i,j}^h - u_{i-1,j}^h)}{2}$$

und daher

$$F(u, v) = \frac{1}{2\lambda}(u - v) + \frac{1}{2}(f(u) + f(v))$$

und somit natürlich

$$F(u, u) = f(u) .$$

Die Konsistenzordnung eines Verfahrens wird wie üblich definiert mithilfe des ‘‘Einsetzfehlers‘‘ einer glatten Lösung in die Verfahrensfunktion:

Definition 1.9 Sei u eine hinreichend glatte Lösung von $u_t + (f(u))_x = 0$. $p \in \mathbb{N}$ heißt die Konsistenzordnung des Verfahrens Ψ , falls

$$u(x, t + \Delta t) - \Psi(u(x - m\Delta x, t), \dots, u(x + m\Delta x, t)) = \mathcal{O}((\Delta t)^{p+1})$$

für $\Delta t \rightarrow 0$ und $\Delta t/\Delta x$ fest. Für $p \geq 1$ heißt das Verfahren konsistent. Die Grösse

$$\begin{aligned} \tau(u, \Delta t, \Delta x) &\stackrel{\text{def}}{=} \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} \\ &\quad + \frac{1}{\Delta x}(F(u(x - (m-1)\Delta x, t), \dots, u(x + m\Delta x, t)) - \\ &\quad \quad F(u(x - m\Delta x, t), \dots, u(x + (m-1)\Delta x, t))) \end{aligned}$$

heißt der lokale Abschneidefehler eines Verfahrens in Erhaltungsform.

Bemerkung 1.9 Ein Verfahren hat also die Konsistenzordnung p genau dann, wenn

$$|\tau(u, \Delta t, \Delta x)| = \mathcal{O}((\Delta t)^p) .$$

Es soll nun die Struktur von τ näher untersucht werden. Dazu benutzen wir Taylorentwicklung höherer Ordnung. Wir setzen dabei voraus, daß $F \in C^2(\mathbb{R}^{2m})$. Zur Abkürzung benutzen wir die Notation

$$u_i = u(x + ih, t), \quad u = u_0 = u(x, t), \quad h = \Delta x, \quad k = \Delta t .$$

Wegen der Festlegung von λ ist u.a.

$$h = \mathcal{O}(k) \quad \text{und} \quad k = \mathcal{O}(h)$$

sodaß wir nach Zweckmässigkeit das eine durch das andere ersetzen können. Wir benötigen u.a. eine Darstellung von u_{tt} mittels der Differentialgleichung:

$$\begin{aligned}
u_{tt} &= -\frac{\partial}{\partial t}(f'(u)u_x) \\
&= -f''(u)u_t u_x - f'(u)u_{xt} \\
&= f''(u)f'(u)(u_x)^2 - f'(u)u_{tx} \\
&= f''(u)f'(u)(u_x)^2 - f'(u)(-f''(u)(u_x)^2 - f'(u)u_{xx}) \\
&= 2f''(u)f'(u)(u_x)^2 + (f'(u))^2 u_{xx} \\
&= \frac{\partial}{\partial x}((f'(u))^2 u_x) .
\end{aligned}$$

Dann ist (man beachte $u = u_0$)

$$\begin{aligned}
\tau(u, k, h) &= u_t + \frac{1}{2}k u_{tt} + \mathcal{O}(k^2) \\
&\quad + \frac{1}{h}(F(u_{-m+1}, \dots, u_m) - F(u_0, \dots, u_0) - (F(u_{-m}, \dots, u_{m-1}) - F(u_0, \dots, u_0))) .
\end{aligned}$$

Weiter benutzen wir

$$\begin{aligned}
u_t &= -f'(u)u_x = -\left(\sum_{i=1}^{2m} \partial_i F(u, \dots, u)\right)u_x \\
u_{-m+k} - u_0 &= h(-m+k)u_x + \frac{h^2}{2}(-m+k)^2 u_{xx} + \mathcal{O}(h^3) \\
F(u_{-m+1}, \dots, u_m) - F(u_0, \dots, u_0) &= \sum_{i=1}^{2m} \partial_i F(u, \dots, u)(u_{-m+i} - u_0) \\
&\quad + \frac{1}{2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} \partial_i \partial_j F(u, \dots, u)(u_{-m+i} - u_0)(u_{-m+j} - u_0) + \mathcal{O}(h^3) \\
F(u_{-m}, \dots, u_{m-1}) - F(u_0, \dots, u_0) &= \sum_{i=1}^{2m} \partial_i F(u, \dots, u)(u_{-m+i-1} - u_0) \\
&\quad + \frac{1}{2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} \partial_i \partial_j F(u, \dots, u)(u_{-m+i+1} - u_0)(u_{-m+j+1} - u_0) + \mathcal{O}(h^3)
\end{aligned}$$

Dies ergibt

$$\begin{aligned}
\frac{1}{h}(F(u_{-m+1}, \dots, u_m) - F(u_{-m}, \dots, u_{m-1})) &= \\
\sum_{i=1}^{2m} \partial_i F(u, \dots, u)((-m+i) - (-m+i-1))u_x + \\
\frac{h}{2} \sum_{i=1}^{2m} \partial_i F(u, \dots, u)((-m+i)^2 - (-m+i-1)^2)u_{xx} + \\
\frac{h}{2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} \partial_i \partial_j F(u, \dots, u)(u_x)^2((-m+i)(-m+j) - (-m+i-1)(-m+j-1)) + \mathcal{O}(h^2) &=
\end{aligned}$$

$$\begin{aligned}
& \left(\sum_{i=1}^{2m} \partial_i F(u, \dots, u) \right) u_x + \\
& \frac{h}{2} \sum_{i=1}^{2m} \partial_i F(u, \dots, u) (2i - 2m - 1) u_{xx} + \\
& \frac{h}{2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} \partial_i \partial_j F(u, \dots, u) (u_x)^2 (i + j - 2m - 1) + \mathcal{O}(h^2)
\end{aligned}$$

Andererseits ist

$$\begin{aligned}
\frac{k}{2} \frac{1}{\lambda} \frac{\partial}{\partial x} \left(\sum_{i=1}^{2m} \partial_i F(u, \dots, u) (2i - 2m - 1) u_x \right) &= \frac{h}{2} \left(\sum_{i=1}^{2m} \partial_i F(u, \dots, u) (2i - 2m - 1) u_{xx} + \right. \\
& \left. \sum_{i=1}^{2m} \sum_{j=1}^{2m} \partial_i \partial_j F(u, \dots, u) (2i - 2m - 1) (u_x)^2 \right)
\end{aligned}$$

Schliesslich gilt noch

$$\sum_{i=1}^{2m} (2i - 2m - 1) \partial_i \left(\sum_{j=1}^{2m} \partial_j F(u, \dots, u) \right) = \sum_{i=1}^{2m} \sum_{j=1}^{2m} \partial_i \partial_j F(u, \dots, u) (i + j - 2m - 1),$$

denn vertauscht man auf der linken Seiten die Summationsindizes, muss man auch $2i$ gegen $2j$ austauschen und es gilt ja

$$\partial_i \partial_j F = \partial_j \partial_i F.$$

Dies alles in die Entwicklung von τ eingesetzt ergibt

$$\tau(u, k, h) = \frac{k}{2} \frac{\partial}{\partial x} \left(\left((f'(u))^2 + \frac{1}{\lambda} \left(\sum_{i=1}^{2m} \partial_i F(u, \dots, u) (2i - 2m - 1) \right) \right) u_x \right) + \mathcal{O}(k^2).$$

Mit der Definition

$$B(u, \lambda) \stackrel{def}{=} -\frac{1}{2} ((f'(u))^2) + \frac{1}{\lambda} \sum_{i=1}^{2m} \partial_i F(u, \dots, u) (2i - 2m - 1)$$

können wir dann schreiben

$$\tau(u, k, h) = -k \frac{\partial}{\partial x} (B(u, \lambda) u_x) + \mathcal{O}(k^2).$$

Wenn wir annehmen, daß auf der Zeitschicht j

$$u_{i,j}^h = u_{i,j} = u(x_i, t_j)$$

gilt, dann ist nach Definition von τ

$$u_{i,j+1} - u_{i,j+1}^h = k \tau(u_{i,j}, k, h)$$

und damit ist $u_{i,j+1}^h$ eine $\mathcal{O}(k^2)$ -Approximation der Lösung der Differentialgleichung

$$v_t + (f(v))_x - k \frac{\partial}{\partial x} (B(v, \lambda) v_x) = 0 \quad (1.15)$$

mit den Anfangswerten $v(x_i, t_j) = u(x_i, t_j)$. Die Gleichung (1.15) heißt die dem Verfahren zugeordnete modifizierte Gleichung. Dies ist eine parabolische Gleichung, der Term $k \frac{\partial}{\partial x} (B(v, \lambda) v_x)$ ist der Diffusionsterm. Für $B > 0$ hat diese Gleichung ein Glättungs- und Dämpfungsverhalten, während für $B < 0$ (dies entspricht dem Lösen einer Gleichung mit $B > 0$ rückwärts in der Zeit) Instabilität vorliegt. Dementsprechend wird man nach Verfahren suchen, für die $B \geq 0$ gilt.

Definition 1.10 *Der Term*

$$k \frac{\partial}{\partial x} (B(u, \lambda) u_x)$$

heißt die numerische Viskosität des Verfahrens.

Ist diese numerische Viskosität negativ, dann ist zugeordnete modifizierte Gleichung instabil, das Verfahren also sicher unbrauchbar.

Beispiel 1.3 *Für das naive Verfahren*

$$u_{i,j+1}^h = u_{i,j}^h - \frac{\lambda}{2} (u_{i+1,j}^h - u_{i-1,j}^h)$$

ist

$$F(u, v) = \frac{1}{2} (f(u) + f(v))$$

und für die Konvektionsgleichung

$$u_t + au_x = 0$$

ergibt sich somit

$$B(u, \lambda) = -\frac{1}{2} (a^2 + \frac{1}{\lambda} (-a + a)) = -a^2/2$$

und dies zeigt bereits die Instabilität, die wir später mit funktionalanalytischen Methoden noch einmal beweisen werden. Für das Friedrichsverfahren ist

$$B(u, \lambda) = -\frac{1}{2} (a^2 + \frac{1}{\lambda} (-\frac{1}{2\lambda} - \frac{1}{2\lambda})) = \frac{1}{2} (\frac{1}{\lambda^2} - a^2)$$

*und dies ist genau unter der CFL-Bedingung nichtnegativ. Für das Lax-Wendroff-Verfahren ist $B(u, \lambda) = 0$, es ist das einzige Verfahren mit $m = 1$ von der Ordnung 2. Es hat aber den Nachteil, daß bei Schocklösungen starke Oszillationen in der numerischen Näherung auftreten. Das **upwind**-Verfahren*

$$u_{i,j+1}^h = \begin{cases} u_{i,j}^h - \lambda a (u_{i,j}^h - u_{i,j-1}^h) & \text{wenn } a > 0 \\ u_{i,j}^h - \lambda a (u_{i+1,j}^h - u_{i,j}^h) & \text{wenn } a < 0 \end{cases}$$

das man in kompakter Form als

$$u_{i,j+1}^h = u_{i,j}^h - \lambda a \frac{u_{i+1,j}^h - u_{i-1,j}^h}{2} + \frac{|\lambda a|}{2} (u_{i+1,j}^h - 2u_{i,j}^h + u_{i-1,j}^h)$$

schreiben kann, ergibt sich

$$B(u, \lambda) = \frac{|a|}{2\lambda} (1 - |\lambda a|)$$

und dies ist unter der CFL-Bedingung wieder positiv.

Es gibt weitere Verfahren, die in diesem Zusammenhang häufig benutzt werden. Das Godunov-Verfahren überträgt die upwind-Idee auf den nichtlinearen Fall. Die Lösung wird auf einer Gitterlinie $t = \text{const}$ stückweise konstant approximiert durch die Integralmittel über eine Gitterweite, sogenannte Zellmittel. Die Anfangswerte sind dabei

$$u_{i,0}^h = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x) dx .$$

Dann wird für jeden Gitterpunkt ein Riemannproblem (mit einer Sprungstelle in den Gittermittelpunkten) exakt gelöst. Dies ist einfach im linearen Fall. Das Riemannproblem lautet

$$u_t + (f(u))_x = 0, \quad u_0(x) = \begin{cases} u_l = u_{i,0} & \text{für } x \leq x_{i+\frac{1}{2}} \\ u_r = u_{i+1,0} & \text{für } x > x_{i+\frac{1}{2}} \end{cases}$$

Für konvexes f sind die Lösungen entweder Schocks mit der Geschwindigkeit

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l}$$

oder Verdünnungswellen mit Ausbreitungsgeschwindigkeiten $f'(u_l)$ und $f'(u_r)$. Die Schocks können nicht interagieren, wenn

$$k \sup_w |f'(w)| \leq h/2 ,$$

also symbolisch unter der Bedingung "CFL/2". Im allgemeinen Schritt tritt die Zeitschicht $t = t_j$ an die Stelle der Zeitschicht $t = 0$. Nun integriert man die Differentialgleichung über dem Rechteck $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t_j, t_{j+1}]$:

$$\frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t_{j+1}) dx - \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t_j) dx = -\frac{1}{h} \left(\int_{t_j}^{t_{j+1}} f(u(x_{i+\frac{1}{2}}, t)) dt - \int_{t_j}^{t_{j+1}} f(u(x_{i-\frac{1}{2}}, t)) dt \right)$$

und nimmt neue Zellmittel auf der nächsten Zeitschicht, also

$$u_{i,j+1}^h = u_{i,j}^h - \frac{1}{h} \left(\int_{t_j}^{t_{j+1}} f(u(x_{i+\frac{1}{2}}, t)) dt - \int_{t_j}^{t_{j+1}} f(u(x_{i-\frac{1}{2}}, t)) dt \right) .$$

Man muß also die Lösung der Riemannprobleme kennen, um die Integrale ausführen zu können. Wir setzen weiter strikt konvexes f voraus, und untersuchen Fallunterscheidungen. Man beachte im Folgenden, daß der Fall $u_l = u_r$ trivial ist (die Lösung ist die Konstante u_l).

1. $f'(u_l) \geq 0$ und $f'(u_r) \geq 0$.

(a) $f'(u_l) < f'(u_r)$. Wegen der Konvexität von f und $u_l < u_r$ ist dann die Lösung eine nach rechts laufende Verdünnungswelle. Dies ist die Entropielösung.

(b) $f'(u_l) > f'(u_r)$. Die Lösung ist ein nach rechts laufender Schock mit der Geschwindigkeit

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l}.$$

Also ist in beiden Fällen

$$u(x_{i+\frac{1}{2}}, t) = u_l.$$

2. $f'(u_l) \geq 0$, $f'(u_r) < 0$. Wegen der Konvexität ist nun $u_l > u_r$. Also liegt eine nach rechts oder links laufende Schockwelle vor, d.h.

$$u(x_{i+\frac{1}{2}}, t) = \begin{cases} u_l & \text{für } s \geq 0 \\ u_r & \text{für } s \leq 0 \end{cases}.$$

3. $f'(u_l) < 0$, $f'(u_r) \leq 0$.

(a) $f'(u_l) < f'(u_r)$, also wegen der Konvexität $u_l < u_r$, d.h. Verdünnungswelle nach links.

(b) $f'(u_l) > f'(u_r)$, also $u_l > u_r$, also eine Schockwelle nach links,

also ist

$$u(x_{i+\frac{1}{2}}, t) = u_r$$

4. $f'(u_l) < 0$, $f'(u_r) > 0$, also $u_r > u_l$, also eine Verdünnungswelle, wobei linker Punkt nach links und rechter Punkt nach rechts läuft, also kann gewählt werden

$$u(x_{i+\frac{1}{2}}, t) = u_s, \text{ mit } f'(u_s) = 0.$$

Dies ergibt folgende Verfahrensvorschrift

$$u_{i,j+1}^h = u_{i,j}^h - \lambda(F(u_{i,j}^h, u_{i+1,j}^h) - F(u_{i-1,j}^h, u_{i,j}^h))$$

mit dem numerischen Fluß

$$F(v, w) = \begin{cases} f(v) & , \text{ falls } v \geq w \text{ und } f(v) \geq f(w) \\ f(w) & , \text{ falls } v \geq w \text{ und } f(v) \leq f(w) \\ f(v) & , \text{ falls } v \leq w \text{ und } f'(v) \geq 0 \\ f(w) & , \text{ falls } v \leq w \text{ und } f'(w) \leq 0 \\ f((f')^{-1}(0)) & , \text{ sonst} \end{cases}.$$

Das Enquist-Osher-Schema schliesslich lautet

$$u_{i,j+1}^h = u_{i,j}^h - \lambda(F(u_{i,j}^h, u_{i+1,j}^h) - F(u_{i-1,j}^h, u_{i,j}^h))$$

mit

$$F(v, w) = \frac{1}{2}(f(v) + f(w)) - \int_v^w |f'(s)| ds .$$

Anders formuliert

$$u_{i,j+1}^h = u_{i,j}^h - \lambda \left(\int_{u_{i-1,j}^h}^{u_{i,j}^h} (f')_+(s) ds + \int_{u_{i,j}^h}^{u_{i+1,j}^h} (f')_-(s) ds \right)$$

wo wie üblich $a_+ = \max\{a, 0\}$, $a_- = \min\{a, 0\}$.

Neben der Erhaltungseigenschaft interessieren in der Praxis auch die Übertragung der Eigenschaften "TV-Stabilität" und " L_1 -Stabilität". Die zugehörigen diskreten Grössen sind

$$\|u_{\cdot,j}^h\|_{L_1} \stackrel{def}{=} \sum_{i \in \mathbb{Z}} |u_{i,j}^h| \quad \text{diskrete } L_1\text{-Norm}$$

und

$$\|u_{\cdot,j}^h\|_{TV} \stackrel{def}{=} \sum_{i \in \mathbb{Z}} |u_{i+1,j}^h - u_{i,j}^h| \quad \text{diskrete Totalvariation}$$

TVD steht dabei für "total variation diminishing". Eine gewisse Klasse von Verfahren hat in der Tat alle diesen wünschenswerten Eigenschaften:

Definition 1.11 *Ein Einzschrittverfahren*

$$u_{i,j+1}^h = \Psi(u_{i-m,j}^h, \dots, u_{i+m,j}^h)$$

heißt *monoton*, wenn Ψ in allen Argumenten eine nichtfallende Funktion ist.

Satz 1.8 Für ein monotones Schema in Erhaltungsform mit $m = 1$ gilt

1.

$$\min\{u_{i-1,k}^h, u_{i,k}^h, u_{i+1,k}^h\} \leq u_{i,k+1}^h \leq \max\{u_{i-1,k}^h, u_{i,k}^h, u_{i+1,k}^h\}$$

also L_∞ -Stabilität,

2.

$$\|u_{\cdot,k+1}^h - v_{\cdot,k+1}^h\|_{L_1} \leq \|u_{\cdot,k}^h - v_{\cdot,k}^h\|_{L_1}$$

falls $\|u_{\cdot,k}^h\|_{L_1}, \|v_{\cdot,k}^h\|_{L_1} < \infty$, also diskrete L_1 -Stabilität und

3.

$$\|u_{\cdot,k+1}^h\|_{TV} \leq \|u_{\cdot,k}^h\|_{TV} \quad \text{falls } \|u_{\cdot,k}^h\|_{L_1} < \infty$$

also die TVD-Eigenschaft.

Zum Beweis siehe bei Crandall und Madja: Monotone difference approximations for scalar conservation laws. Math. Comp. 34, (1980), 1-21.

Satz 1.9 Gegeben sei ein monotones Schema in Erhaltungsform mit $m = 1$ und einem stetigen, konsistenten numerischen Fluß F . Dann konvergiert die numerische Lösung gegen die Entropielösung der Erhaltungsgleichung und die Konvergenzordnung ist maximal eins

Beweis bei Hyman, Harten und Lax: On finite difference approximations and entropy conditions for shocks. Comm Pure Appl. Math. 19, (1976), 297–332.

Bemerkung 1.10 Das Friedrichsverfahren ist monoton für

$$\lambda \max_u \{|f'(u)|\} \leq 1,$$

das Lax-Wendroff-Verfahren mit

$$F(u, v) = \frac{1}{2}(f(u) + f(v)) - \frac{\lambda}{2} \left(\frac{f(u) - f(v)}{u - v} \right)^2 (v - u)$$

dagegen nicht.

Das Lax-Wendroff-Verfahren hat den grossen Vorteil, für glatte Funktionen von zweiter Ordnung konsistent zu sein und unter der CFL-Bedingung auch zu konvergieren. In Schocks

entstehen aber starke Oszillationen der numerischen Lösung. Man will nun Verfahren konstruieren, die ebenfalls glatte Lösungen von zweiter Ordnung approximieren, Schocks aber gut reproduzieren, d.h. weder ausglätten, wie es ein Verfahren mit hoher Viskosität (wie Lax-Friedrich) tut, noch Oszillationen erzeugen. Dies gelingt durch Hinzunahme weiterer Gitterpunkte (also $m > 1$.) Ein möglicher Ansatz sei hier kurz beschrieben. Dies sind die sogenannten "flux-limiter"-Verfahren. Der Ansatz besteht darin, den numerischen Fluß des Verfahrens zu modifizieren durch den Ansatz

$$F(u, v) = F^L(u, v) + \phi(\theta)(F^H(u, v) - F^L(u, v)) .$$

Dabei ist F^L z.B. die Flußfunktion eines genügend viskosen Verfahrens der Ordnung eins und F^H z.B. die Flußfunktion des Lax-Wendroff-Verfahrens. $\phi(\theta)$ ist eine lösungsabhängige Funktion, die in einem Schock null, aber für glatte Lösungen eins wird. Zum Messen von Unstetigkeiten benutzt man den Vergleich numerischer Steigungen:

$$\theta_{i+\frac{1}{2},j,+} = \frac{u_{i,j}^h - u_{i-1,j}^h}{u_{i-1,j}^h - u_{i-2,j}^h} \quad \text{falls } f'((u_{i+1,j}^h + u_{i,j}^h)/2) > 0$$

bzw.

$$\theta_{i+\frac{1}{2},j,-} = \frac{u_{i+2,j}^h - u_{i+1,j}^h}{u_{i+1,j}^h - u_{i,j}^h} \quad \text{falls } f'((u_{i+1,j}^h + u_{i,j}^h)/2) < 0 .$$

Das Vorzeichen richtet sich also nach dem Vorzeichen von a bzw. $f'(\frac{u_{i+1,j}^h + u_{i,j}^h}{2})$. In Bereichen, in denen die numerische Lösung die wahre Lösung von zweiter Ordnung approximiert und glatt ist, ist θ dann von der Größenordnung $1 + \Delta x$, weicht aber nahe einem Schock von 1 stark ab. Für ϕ macht man nun einen geeigneten Funktionsansatz, und es gilt dazu folgender Satz:

Satz 1.10 *Das durch F^L gegebene Verfahren sei monoton, konservativ und konsistent von erster Ordnung, das durch F^H gegebene Verfahren sei von zweiter Ordnung konsistent für glatte Lösungen. Die CFL-Bedingung sei erfüllt. Die Funktion ϕ erfülle folgende Eigenschaften:*

1. $\phi \in C^1(\mathcal{U}(1))$ mit einer geeigneten Umgebung \mathcal{U} von 1 .
2. $\phi(1) = 1$.
3. $0 \leq \phi(\theta) \leq 2$, $0 \leq \frac{\phi(\theta)}{\theta} \leq 2$, $\forall \theta \in \mathbb{R}$.

Dann ist das Verfahren konsistent, konsistent von zweiter Ordnung für glatte Lösungen und hat die TVD-Eigenschaft.

Folgende Ansätze für den flux-limiter ϕ erfüllen diese Bedingung:

1. $\phi_\alpha(\theta) = \max\{0, \min\{1, \alpha\theta\}, \min\{\alpha, \theta\}\}$.

Dieser Ansatz ist als "superbee" bekannt. Hier ist $\alpha \in]1, 2]$ fest gewählt.

2. $\phi(\theta) = \frac{\theta+|\theta|}{1+|\theta|}$.

Dieser Ansatz stammt von van Leer.

Es gibt weitere Verfahren dieser Art, auch für Systeme von Erhaltungsgleichungen. Siehe dazu [17], [18].

Kapitel 2

Parabolische Rand-Anfangswertprobleme

2.1 Theoretische Grundlagen

Quelle für diesen Abschnitt ist [7] und [3]. Bei parabolischen DGLen in zwei freien Veränderlichen ist in jedem Punkt (x, t) des betrachteten Gebietes nur eine charakteristische Richtung vorhanden. Es gibt also für solche Gleichungen kein charakteristisches Netz und dementsprechend auch kein Charakteristiken-Verfahren.

Wir beginnen die Diskussion mit dem einfachsten Beispiel, der Wärmeleitungsgleichung

$$\begin{aligned}u_t &= u_{xx} & x \in \mathbb{R}, \quad t > 0 \\u(x, 0) &= g(x), & |g(x)e^{-|x|}| \leq K\end{aligned}$$

Eine Charakteristik mit der Parameterdarstellung $\begin{pmatrix} k_1(t) \\ k_2(t) \end{pmatrix}$ erfüllt hier die DGL

$$(k_2')^2 \equiv 0,$$

d.h. die Charakteristikenschar ist $\begin{pmatrix} t \\ c \end{pmatrix}$ c = Scharparameter. Die Anfangswertaufgabe gibt hier die Anfangswerte auf einer charakteristischen Kurve vor!

Wenn man sich auf die Lösung mit

$$|u(x, t)e^{-|x|}| \leq C$$

beschränkt, ist die Anfangswertaufgabe wohlgestellt und ihre Lösung wird gegeben durch

$$u(x, t) = \begin{cases} \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\tau)^2}{4t}\right) g(\tau) d\tau & t > 0 \\ g(x) & t = 0 \end{cases}$$

(vgl. z.B. bei Hellwig, Partielle DGLen, Teubner 1960, Kap. I.4)

Aus der Lösungsformel geht hervor, daß für jeden noch so kleinen Wert $t > 0$ das Abhängigkeitsintervall von (x, t) die gesamte x -Achse ist. Man erkennt ferner, daß mit wachsendem t die Lösung $u(x, t)$ zunehmend glatter wird.

Das reine Anfangswertproblem der Wärmeleitungsgleichung ist praktisch uninteressant. Diese Gleichung tritt immer im Zusammenhang mit Randanfangswertaufgaben auf, z.B.

$$\begin{aligned} u_t &= u_{xx}, & 0 \leq x \leq L, & & 0 < t < T \\ u(x, 0) &= f(x) & 0 \leq x \leq L & & f \in C[0, 1] \\ u(0, t) &= \varphi_0(t), & u(L, t) &= \varphi_1(t) & \varphi \in C[0, T] \end{aligned}$$

(Wärmeleitung in einem isolierten homogenen Stab mit vorgegebener Anfangstemperturverteilung und Vorgabe der Temperatur an den Stabenden.)

Unter der Verträglichkeitsbedingung

$$f(0) = \varphi_0(0), \quad f(1) = \varphi_1(0)$$

ist das Problem sachgemäß gestellt (vgl. Petrovsky 1954, 38, [13]).

Dies gilt jedoch nicht, wenn man die Zeitrichtung umkehrt, also das gleiche Problem mit $-T < t < 0$ betrachtet!

Im vorliegenden Fall erfüllt die Lösung u sogar ein Maximum / Minimum-Prinzip:

Satz 2.1 *Es sei u eine Lösung von $u_t = u_{xx}$ auf $G =]0, L[\times]0, T[$ und $u \in C^2(G) \cap C(\bar{G})$. Dann gilt*

$$\begin{aligned} \max_{(x,t) \in \bar{G}} u(x, t) &= \max_{(x,t) \in \partial G_0} u(x, t) \\ \min_{(x,t) \in \bar{G}} u(x, t) &= \min_{(x,t) \in \partial G_0} u(x, t) \end{aligned}$$

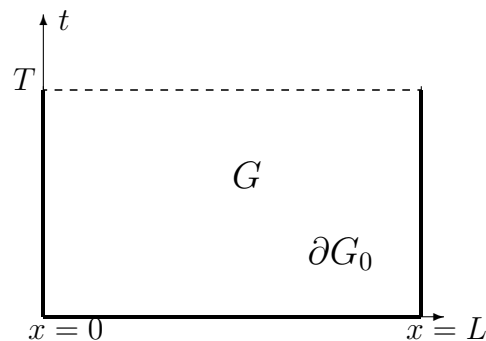


Abbildung 2.1

mit $\partial G_0 = \partial G \setminus \{T\} \times]0, L[$

Beweis: Setze $v(x, t) = u(x, t) - \varepsilon \cdot t$, dann ist $v_{xx} = v_t + \varepsilon$

Ann. 1: $\exists P \in G : v(P) = \max_{(x,t) \in \bar{G}} v(x,t)$.
 $\Rightarrow v_x(P) = v_t(P) = 0$ und $v_{xx}(P) \leq 0$ im Widerspruch!
 zu $v_{xx}(P) = v_t(P) + \varepsilon = \varepsilon > 0$.

Ann. 2: $\exists P \in]0, L[\times \{T\}$ mit $v(P) = \max_{(x,t) \in \bar{G}} v(x,t)$.
 Dann ist $v_t(P) \geq 0$, $v_x(P) = 0$ und $v_{xx}(P) \leq 0$,
 aber $v_{xx}(P) = v_t(P) + \varepsilon \geq \varepsilon$ Widerspruch!

Mit $\varepsilon \rightarrow 0$ folgt die entsprechende Aussage für u . Die Minimaussage beweist man entsprechend. \square

Im Allgemeinen lautet ein parabolisches Randanfangswertproblem

$$\frac{\partial}{\partial t} u - L(u) = f \text{ auf } \Omega \times]0, T[\stackrel{\text{def}}{=} G$$

mit den Anfangsbedingungen (Anfangszustand)

$$u(x, 0) = u_0(x) \text{ für } x \in \Omega$$

und zusätzlichen Randbedingungen für $\partial\Omega \times [0, T[$. Diese Randbedingungen können vom Dirichlet-, von Neumann- oder auch Robin-Typ sein. L ist dabei ein (gleichmäßig) elliptischer Differentialoperator auf Ω . Ω sei dabei ein Lipschitzgebiet (vergl. die Definition in Höhere Numerische Mathematik I, WS2005/2006) Klassische Lösungen wie im Fall der oben diskutierten Wärmeleitungsgleichung bedingen starke Voraussetzungen sowohl an f , die Koeffizienten von L , das Gebiet Ω und noch zusätzlich die Rand- und Anfangsbedingungen und ihre Kopplung. Bei homogenen Dirichletrandbedingungen auf $\partial\Omega \times [0, T[$ muss man z.B. auch fordern, daß $u_0(\partial\Omega) \equiv 0$. Ohne solche Bedingungen kann man keine Beschränktheit der Ableitungen auf \bar{G} erwarten. Wir beschreiben hier noch kurz den Begriff der schwachen Lösung solcher Probleme. Ist u eine klassische Lösung und $a(., .)$ die zu L gehörige Bilinearform, dann ist

$$\frac{d}{dt}(u(., t), v) + a(u(., t), v) = (f(., t), v) \quad \forall v \in V$$

wobei $(., .)$ das L_2 -Skalarprodukt auf Ω und V ein zu L passender Raum von Testfunktionen auf Ω ist. Die Ableitung $\frac{d}{dt}u(., t)$ wirkt hier also als lineares Funktional auf V . Die Abbildung $x \rightarrow u(x, t)$ wird bei festem t als eine Abbildung in einen Funktionenraum über Ω interpretiert und die partielle Differentialgleichung so zu einer gewöhnlichen Differentialgleichung in einem Hilbertraum.

Definition 2.1 Sei X ein Banachraum mit der Norm $\|\cdot\|_X$. Der Raum $L_2(0, T; X)$ ist der Raum aller Funktionen $[0, T] \rightarrow X$ mit

$$\|u\| \stackrel{\text{def}}{=} \left(\int_0^T \|u(., t)\|_X^2 dt \right)^{1/2} < \infty$$

Mit dieser Norm ist $L_2(0, T; X)$ selbst ein Banachraum. Ist $f \in L_2(G)$ dann ist auch $f \in L_2(0, T; L_2(\Omega))$ weil nach dem Satz von Fubini

$$\|f\|_{L_2(G)}^2 = \int_0^T \int_{\Omega} f^2(x, t) dx dt$$

Ist X reflexiv und separabel und X' der Dualraum zu X , dann ist $L_2(0, T; X')$ der Dualraum zu $L_2(0, T; X)$. Ein Evolutionstripel ist gegeben durch V, H, V' mit

$$V \subset H \subset V'$$

und

- V ist ein separabler, reflexiver Banachraum ,
- H ist ein separabler Hilbertraum ,
- V ist dicht und stetig eingebettet in H , d.h. $\|v\|_H \leq C \cdot \|v\|_V \forall v \in V$.

Beispiel 2.1 $V = H_0^1(\Omega)$, $H = L_2(\Omega)$, $V' = H^{-1}(\Omega)$.

Definition 2.2 Die Funktion $u \in L_2(0, T; V)$ besitzt die verallgemeinerte Ableitung $w \in L_2(0, T; V')$, symbolisch $w = u'$, wenn gilt

$$\int_0^T \phi'(t) u(\cdot, t) dt = - \int_0^T \phi(t) w(\cdot, t) dt \quad \forall \phi \in C_0^\infty(0, T)$$

Dabei ist die Gleichheit im Sinne des Raumes V zu verstehen.

Nun definiert man

Definition 2.3

$$W_2^1(0, T; V, H) \stackrel{def}{=} \{u \in L_2(0, T; V) : \exists u' \in L_2(0, T; V')\} \cap C(G)$$

mit der Norm

$$\|u\|_{W_2^1} \stackrel{def}{=} \|u\|_{L_2(0, T; V)} + \|u'\|_{L_2(0, T; V')} .$$

Für $u \in W_2^1(0, T; V, H)$ ist die Abbildung $t \mapsto u(\cdot, t)$ stetig, sodaß die Forderung

$$u(\cdot, 0) = u_0(\cdot)$$

sinnvoll ist. Es gilt auf $W_2^1(0, T; V, H)$ die Formel der partiellen Integration

$$(u(\cdot, t), v(\cdot, t)) - (u(\cdot, s), v(\cdot, s)) = \int_s^t \left((u'(\cdot, \tau), v(\cdot, \tau)) + (u(\cdot, \tau), v'(\cdot, \tau)) \right) d\tau .$$

Mit diesen Vorbereitungen lautet die schwache Form einer parabolischen RAWA: Sei V ein Sobolev-Raum zwischen $H_0^1(\Omega)$ und $H^1(\Omega)$ (je nach Randbedingungen) und $H = L_2(\Omega)$, ferner $a(\cdot, \cdot)$ eine koerzive beschränkte Bilinearform auf $V \times V$, $u_0 \in H$ und $f \in L_2(0, T; V')$. Gesucht ist $u \in W_2^1(0, T; V, H)$ mit $u(\cdot, 0) = u_0$ sodaß gilt

$$\frac{d}{dt}(u(\cdot, t), v) + a(u(\cdot, t), v) = \langle f(\cdot, t), v \rangle \quad \forall v \in V \quad (2.1)$$

Die Randbedingungen sind also hier in den Sobolevraum V eingearbeitet, während die Anfangsbedingung explizit formuliert ist: wir haben eine gewöhnliche Differentialgleichung in einem Hilbertraum. Es gilt

Satz 2.2 *Das Problem (2.1) besitzt genau eine Lösung.*

(Zum Beweis siehe z.B. bei Zeidler: Nonlinear functional analysis.) Die Lösung des Randanfangswertproblems mit $f \in L_2(G)$ erlaubt eine a priori Abschätzung, die zeigt, daß das Problem wohlgestellt ist. Wegen der stetigen Einbettung von $L_2(G)$ in V ist nun

$$\langle f(\cdot, t), v \rangle = (f(\cdot, t), v)$$

Setzt man $v = u(\cdot, t)$ dann ergibt sich

$$\frac{1}{2} \frac{d}{dt} \|u(\cdot, t)\|_{L_2}^2 + a(u(\cdot, t), u(\cdot, t)) = (f(\cdot, t), u(\cdot, t))$$

und wegen $a(\cdot, \cdot) \geq 0$ und der Schwarzschen Ungleichung

$$\frac{d}{dt} \|u(\cdot, t)\|_{L_2} \leq \|f(\cdot, t)\|_{L_2}$$

Dies, integriert bezüglich t ergibt

$$\|u(\cdot, t)\|_{L_2} \leq \|u_0\|_{L_2} + \int_0^t \|f(\cdot, s)\|_{L_2} ds$$

Parabolische Randanfangswertaufgaben haben eine Glättungseigenschaft: auch wenn die Anfangs- und Randwerte inkompatibel sind, also keine beschränkten Ableitungen auf \bar{G} existieren, so sind doch die Ableitungen für $T > 0$ in der Regel beschränkt. Z.B. gilt

Satz 2.3 Die eindeutige schwache Lösung u der parabolischen Randanfangswertaufgabe

$$\begin{aligned}\frac{\partial}{\partial t}u - \Delta u &= 0 && \text{auf } \Omega \times]0, T[\\ u &= 0 && \text{auf } \partial\Omega \times]0, T[\\ u(\cdot, 0) &= u_0 && \text{mit } u_0 \in L_2(\Omega)\end{aligned}$$

erfüllt die Wachstumsbedingung

$$\|u(\cdot, t)\|_{H^k(\Omega)} \leq Ct^{-\frac{1}{2}k} \|u_0\|_{L_2} \text{ für } t \geq \delta > 0 \text{ mit } C = C(\delta) .$$

2.2 Differenzenapproximationen für parabolische Gleichungen

2.2.1 Der räumlich eindimensionale Fall

Differenzenverfahren für die RAWA der Wärmeleitungsgleichung oder auch für die allgemeinere Gleichung

$$u_t = \frac{\partial}{\partial x}(a(x)u_x) + f(x, t, u)$$

erhält man durch die Methode der Semidiskretisierung: Wir besprechen hier zunächst die Semidiskretisierung im Raum, die dem Ansatz auch die Bezeichnung “vertikale Linienmethode“ gegeben hat.

Man setze

$$\tilde{v}_i(t) \stackrel{\text{def}}{=} u(x_i, t)$$

und ersetze die räumliche Ableitung durch einen Differenzenquotienten, z.B.

$$\begin{aligned} u_{xx}(x_i, t) &= \frac{u(x_{i-1}, t) - 2u(x_i, t) + u(x_{i+1}, t))}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2) \\ \frac{\partial}{\partial x}(a(x)u_x)(x_i, t) &= \frac{1}{(\Delta x)^2} \left(a_{i+\frac{1}{2}}(u(x_{i+1}, t) - u(x_i, t)) - a_{i-\frac{1}{2}}(u(x_i, t) - u(x_{i-1}, t)) \right) \\ &\quad + \mathcal{O}((\Delta x)^2) \end{aligned}$$

(oder auch bessere Approximationen) und erhält dann ein System gewöhnlicher Differentialgleichungen **erster Ordnung** der Form

$$\dot{v} = \frac{1}{(\Delta x)^2} Av + F(t, v) \tag{2.2}$$

(v_i steht für die Approximation an \tilde{v}_i , die durch die Vernachlässigung der \mathcal{O} -Terme entsteht)

Im Falle $u_t = u_{xx}$ mit den Dirichletdaten φ_0 bzw. φ_1 bei $x = 0$ bzw. $x = 1$ und der Verwendung des symmetrischen Differenzenquotienten wird

$$A = (0, \dots, 0, 1, -2, 1, 0, \dots, 0), \quad F(t, v) = \frac{1}{(\Delta x)^2} \begin{pmatrix} \varphi_0(t) \\ 0 \\ \vdots \\ 0 \\ \varphi_1(t) \end{pmatrix} \begin{matrix} \leftarrow \\ \\ \\ \\ \leftarrow \end{matrix} \begin{matrix} \text{Randvorgabe} \\ \text{für } u \text{ bei } x = 0 \\ \text{und } x = 1 \end{matrix}$$

Auch in den allgemeineren Fällen entsteht F aus dem Vektor $(f(x_1, t, v_1), \dots, f(x_m, t, v_m))^T$ und den Termen, die aus $v_0 = \tilde{v}_0 = \varphi_0$ und $v_{m+1} = \tilde{v}_{m+1} = \varphi_1$ stammen. m ist dabei die

Anzahl der inneren x -Knoten, auf $[0, 1]$ und $\Delta x = 1/(m+1)$. Die Vektorfunktion F ist (entsprechend f) eine in v "glatte" Funktion, d.h. $\|\frac{\partial}{\partial v} F(t, v)\|$ ist nicht sehr groß. Dagegen ist der lineare homogene Teil der DGL

$$\dot{v} = \frac{1}{(\Delta x)^2} A v$$

sehr steif, wenn Δx klein ist. Die Eigenwerte von $\frac{1}{(\Delta x)^2} A$ sind

$$2 \cdot \frac{1}{(\Delta x)^2} \left(\cos\left(\frac{i\pi}{m+1}\right) - 1 \right) \quad i = 1, \dots, m$$

liegen also in $[-\frac{4}{(\Delta x)^2}, -\pi^2 + \frac{\pi^4}{12(m+1)^2} \dots]$ (für $x \in [0, 1]$, d.h. $\Delta x = \frac{1}{m+1}$). Im Prinzip kann man jedes für gewöhnliche DGLen erster Ordnung konvergente Diskretisierungsverfahren auf die gewöhnliche DGL (2.2) anwenden, hat dann aber wegen deren Steifheit mit den üblichen Schwierigkeiten zu rechnen. Wir diskutieren nun einige in diesem Zusammenhang gebräuchliche Verfahren. Wir betrachten folgendes Einschrittverfahren für eine DGL $y' = f(t, y)$:

$$u_{n+1}^h - u_n^h = h(\alpha f_{n+1} + (1-\alpha)f_n)$$

also in der Notation der MSV $\rho(\zeta) = \zeta - 1$ und $\sigma(\zeta) = \alpha\zeta + 1 - \alpha$. Dieses Verfahren ist für alle α konsistent von der Ordnung 1 und asymptotisch stabil sowie von der Ordnung 2 für $\alpha = \frac{1}{2}$. Wir erhalten für

$$\begin{array}{lll} \alpha = 0 : & \text{Euler vorwärts} & \\ \alpha = \frac{1}{2} : & \text{Trapezregel} & (A\text{-stabil}) \\ \alpha = 1 : & \text{Euler rückwärts} & (A\text{-stabil}) \end{array}$$

In diesem Zusammenhang diese Verfahren die Bezeichnung

- explizites Differenzenverfahren
- Crank-Nicholson-Verfahren
- voll implizites Differenzenverfahren

Es ergibt sich bei diesem Verfahren folgende Verknüpfung der Gitterfunktionswerte:

$$\begin{array}{cccc} \alpha & 1 & -2 & 1 & \text{Zeitschicht } n+1 \\ & \bullet & \text{---} & \bullet & \\ 1-\alpha & 1 & -2 & 1 & \text{Zeitschicht } n \\ & \bullet & \text{---} & \bullet & \end{array}$$

Setzt man

$$u_n^h = \begin{pmatrix} v_1(t_n) \\ \vdots \\ v_m(t_n) \end{pmatrix} = \begin{pmatrix} u_{1,n}^h \\ \vdots \\ u_{m,n}^h \end{pmatrix} \quad u_{i,j}^h = \text{Näherung für } u(x_i, t_j)$$

dann erhält man für $u_t = u_{xx}$, $u(0, t) = u(1, t) = 0$ folgendes Gleichungssystem für u_{n+1}^h :

$$(I - \alpha \frac{\Delta t}{(\Delta x)^2} A) u_{n+1}^h = (I + (1 - \alpha) \frac{\Delta t}{(\Delta x)^2} A) u_n^h$$

mit

$$A = \text{tridiag}(0, \dots, 0, 1, -2, 1, 0, \dots, 0).$$

Die Matrix auf der linken Seite ist für $\alpha > 0$ irreduzibel diagonaldominant und im Falle $\alpha = 0$ die Identität, d.h. u_{n+1}^h ist für $\alpha \geq 0$ wohldefiniert. Aufgrund der Herleitung ist klar, daß das Verfahren konsistent ist von der Ordnung $\Delta t + (\Delta x)^2$ für $\alpha \geq 0$, $\alpha \neq \frac{1}{2}$ und $(\Delta t)^2 + (\Delta x)^2$ für $\alpha = \frac{1}{2}$.

Für die praktische Brauchbarkeit ist aber nicht nur die Ordnungsaussage wichtig, sondern auch die Güte der Approximation

$$\frac{1 + (1 - \alpha) \frac{\Delta t}{(\Delta x)^2} \lambda_j}{1 - \alpha \frac{\Delta t}{(\Delta x)^2} \lambda_j} \quad \text{für} \quad e^{\frac{\Delta t}{(\Delta x)^2} \lambda_j}$$

wobei λ_j für einen beliebigen Eigenwert von A steht, d.h.

$$\frac{\Delta t}{(\Delta x)^2} \lambda_j \in \approx \left[-\frac{4\Delta t}{(\Delta x)^2}, -\pi\Delta t \right]$$

Damit ist bereits klar, daß das explizite Verfahren nur für $\Delta t/(\Delta x)^2 \ll 1$ und auch das Crank–Nicholson–Verfahren nur für $\Delta t/(\Delta x)^2 \lesssim 1$ brauchbar sein kann, während das vollimplizite Verfahren keine solche Schrittweitereinschränkung aufweist. Speziell für den einfachen Fall $u_t = u_{xx}$ kann man für die hergeleiteten Verfahren direkt einen Konvergenzbeweis erhalten. Zu diesem Zweck untersuchen wir zunächst den lokalen Diskretisierungsfehler: Es ist

$$u_{j,n+1} - u_{j,n} = \Delta t \cdot (u_t)_{j,n} + \frac{(\Delta t)^2}{2} (u_{tt})_{j,n} + \mathcal{O}((\Delta t)^3)$$

Andererseits gilt aufgrund der DGL

$$u_t = u_{xx} \Rightarrow u_{tt} = u_{xxt} = u_{txx} = u_{xxxx}$$

d.h.

$$u_{j,n+1} - u_{j,n} = \Delta t (u_{xx})_{j,n} + \frac{\Delta t^2}{2} (u_{xxxx})_{j,n} + \mathcal{O}((\Delta t)^3).$$

Ebenso erhält man, wenn man die Entwicklung für den Punkt (x_j, t_n) an der Stelle (x_j, t_{n+1}) ausführt

$$u_{j,n+1} - u_{j,n} = \Delta t (u_{xx})_{j,n+1} - \frac{\Delta t^2}{2} (u_{xxxx})_{j,n+1} + \mathcal{O}((\Delta t)^3).$$

Für die Diskretisierung der räumlichen Ableitung gilt

$$(u_{xx})_{j,m} = \frac{u_{j+1,m} - 2u_{j,m} + u_{j-1,m}}{(\Delta x)^2} - \frac{(\Delta x)^2}{12} (u_{xxxx})_{j,m} + \mathcal{O}((\Delta x)^4)$$

für $m \in \mathbb{N}$. (Hierbei ist allerdings vorausgesetzt, daß $u \in C^6$)
Einsetzen in die obige Gleichung ergibt mit

$$\begin{aligned} \rho &\stackrel{def}{=} \frac{\Delta t}{(\Delta x)^2} \\ u_{j,n+1} - u_{j,n} &= \rho(u_{j+1,n} - 2u_{j,n} + u_{j-1,n}) + \Delta t \left(\frac{\Delta t}{2} - \frac{(\Delta x)^2}{12} \right) (u_{xxxx})_{j,n} \\ &\quad + \Delta t (\mathcal{O}(\Delta x^4) + \mathcal{O}(\Delta t^2)) \\ &= \rho(u_{j+1,n+1} - 2u_{j,n+1} + u_{j-1,n+1}) - \Delta t \left(\frac{\Delta t}{2} + \frac{(\Delta x)^2}{12} \right) (u_{xxxx})_{j,n+1} \\ &\quad + \Delta t (\mathcal{O}(\Delta x^4) + \mathcal{O}(\Delta t^2)) \end{aligned}$$

Multiplikation der ersten Gleichung mit $1 - \alpha$, der zweite mit α und Addition ergibt

$$\begin{aligned} u_{j,n+1} - u_{j,n} &= \rho \left((1 - \alpha)(u_{j+1,n} - 2u_{j,n} + u_{j-1,n}) + \alpha(u_{j+1,n+1} - 2u_{j,n+1} + u_{j-1,n+1}) \right) \\ &\quad + \frac{(\Delta t)^2}{2} \left((1 - \alpha)(u_{xxxx})_{j,n} - \alpha(u_{xxxx})_{j,n+1} \right) \\ &\quad - \Delta t \frac{(\Delta x)^2}{12} \left((1 - \alpha)(u_{xxxx})_{j,n} + \alpha(u_{xxxx})_{j,n+1} \right) \\ &\quad + \Delta t \left(\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4) \right) \end{aligned} \left. \vphantom{u_{j,n+1} - u_{j,n}} \right\} \stackrel{def}{=} \Delta t \tau_{j,n+1}$$

Man erkennt, daß für $\alpha = \frac{1}{2}$ und für $\alpha = 0$, $\rho = \frac{1}{6}$ der lokale Diskretisierungsfehler eine spezielle Form annimmt, nämlich

$$\tau_{j,n+1} = \begin{cases} \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) & \alpha = \frac{1}{2}, \quad \rho \text{ beliebig} \\ \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4) & \alpha = 0, \quad \rho = \frac{1}{6} \\ \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) & \text{sonst} \end{cases}$$

Setzen wir

$$\varepsilon_{j,n} \stackrel{def}{=} u_{j,n}^h - u_{j,n}, \quad \varepsilon_n \stackrel{def}{=} (\varepsilon_{1,n}, \dots, \varepsilon_{m,n})^T, \quad \tau_n \stackrel{def}{=} (\tau_{1,n}, \dots, \tau_{m,n})$$

so ergibt sich schließlich die Rekursionsformel für den globalen Diskretisierungsfehler

$$\begin{aligned} (I - \alpha \rho A) \varepsilon_{n+1} &= (I + (1 - \alpha) \rho A) \varepsilon_n + \Delta t \cdot \tau_{n+1} \\ \varepsilon_0 &= 0 \end{aligned}$$

Die Matrix $I - \alpha \rho A$ ist für $\alpha > 0$ irreduzibel diagonaldominant und für $\alpha = 0$ die Identität. Wir erhalten

$$\varepsilon_{n+1} = (I - \alpha \rho A)^{-1} (I + (1 - \alpha) \rho A) \varepsilon_n + \Delta t \cdot (I - \alpha \rho A)^{-1} \tau_{n+1}$$

Sei

$$\begin{aligned} A &= V \Lambda V^T \quad \text{mit reell orthogonalem } V \text{ und den} \\ &\quad \text{Eigenwerten } \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m) \text{ von } A \\ &\quad \lambda_i = 2 \left(\cos\left(\frac{i\pi}{m+1}\right) - 1 \right) \in] -4, 0[\end{aligned}$$

Dann wird

$$V^T \varepsilon_{n+1} = (I - \alpha \rho \Lambda)^{-1} (I + (1 - \alpha) \rho \Lambda) (V^T \varepsilon_n) + \Delta t (I - \alpha \rho \Lambda)^{-1} V^T \tau_{n+1}$$

Für die Normabschätzung erweist sich eine skalierte $\|\cdot\|_2$ -Norm als zweckmäßig, weil für $m \rightarrow \infty$ (d.h. $\Delta x \rightarrow 0$) in der Regel auch $\|\tau_n\|_2 \rightarrow \infty$.

Sei also im Folgenden

$$\|x\| \stackrel{\text{def}}{=} \frac{1}{\sqrt{m}} \|x\|_2 \quad \text{für } x \in \mathbb{R}^m \quad \text{“diskrete } L_2\text{-Norm”}$$

(Wenn $x_i = x(t_i)$ und $t_i - t_{i+1} = \frac{1}{m+1}$, $x_0 = x_{m+1} = 0$, $t_0 = 0$, dann ist

$$\|x\| = \left(\sum_{i=1}^m x_i^2 \frac{1}{m} \right)^{\frac{1}{2}} = \left(\sum_{i=0}^m x_i^2 \left(\frac{1}{m+1} + \frac{1}{m(m+1)} \right) \right)^{\frac{1}{2}} = \left(\int_0^1 x^2(t) dt + \mathcal{O}\left(\frac{1}{m}\right) \right)^{\frac{1}{2}}$$

d.h. dies ist in der vorliegenden Anwendung ein vernünftiges Größenmaß für die Länge von x).

Damit ergibt sich

$$\|\varepsilon_{n+1}\| \leq \max_i \left| \frac{1 + (1 - \alpha) \rho \lambda_i}{1 - \alpha \rho \lambda_i} \right| \|\varepsilon_n\| + \frac{\Delta t}{1 - \alpha \rho \lambda_1} \|\tau_{n+1}\|$$

Man beachte daß

$$\begin{aligned} \left| \frac{1 + (1 - \alpha) \rho \lambda_i}{1 - \alpha \rho \lambda_i} \right| &= \left| 1 + \frac{\rho \lambda_i}{1 - \alpha \rho \lambda_i} \right| \\ &= \left| 1 - \frac{\rho |\lambda_i|}{1 + \alpha \rho |\lambda_i|} \right| \in \left[1 - \frac{4\rho}{1 + 4\alpha\rho}, 1 \right]. \end{aligned}$$

Wegen der Wahl von $\|\cdot\|$ ist für $\Delta x \rightarrow 0$ für $u \in C^6(G)$

$$\tau_{n+1} = \begin{cases} \mathcal{O}((\Delta t)^2) + \mathcal{O}((\Delta x)^2) & \alpha = \frac{1}{2}, \quad \rho \text{ beliebig} \\ \mathcal{O}((\Delta t)^2) + \mathcal{O}((\Delta x)^4) & \alpha = 0, \quad \rho = \frac{1}{6} \\ \mathcal{O}(\Delta t) + \mathcal{O}((\Delta x)^2) & \text{sonst} \end{cases}$$

Zunächst ist immer

$$1 - \alpha \rho \lambda_1 \geq 1.$$

Wir wissen bereits aus der Diskussion der Konvergenz von Einschrittverfahren für gewöhnliche DGLen, daß der Vorfaktor von $\|\varepsilon_n\|$ die Form $1 + \mathcal{O}(\Delta t)$ haben muß, weil anders für $\Delta t \rightarrow 0$, $n\Delta t \rightarrow T$ keine Konvergenz eintreten kann. Wir diskutieren deshalb nun diesen Vorfaktor. Es gilt für $-4 < x < 0$ (x entspricht λ_i)

$$\begin{aligned} \alpha = 0 & \quad 1 + \rho x \in [1 - 4\rho, 1] \subset [-1, 1] \text{ falls } \rho \leq 1/2 \\ 0 < \alpha < \frac{1}{2} & \quad \frac{1 + (1 - \alpha)\rho x}{1 - \alpha \rho x} \in \left[\frac{1 - 4(1 - \alpha)\rho}{1 + 4\alpha\rho}, 1 \right] \end{aligned}$$

Für $\alpha > 0$ ist dies eine in $x \leq 0$ monoton fallende Funktion mit der waagerechten Asymptote $1 - 1/\alpha$. Nun ist

$$\frac{1 - 4(1 - \alpha)\rho}{1 + 4\alpha\rho} \geq -1 \quad \Leftrightarrow \quad \rho \leq \frac{1}{2(1 - 2\alpha)} \quad (0 \leq \alpha < \frac{1}{2})$$

während für $\alpha \geq \frac{1}{2}$ keine Einschränkung an das Schrittweitenverhältnis $\rho = \Delta t / (\Delta x)^2$ vorliegt, um diese Bedingung zu erreichen. Insbesondere für $\alpha = 0$ (entspricht "Euler vorwärts") haben wir also eine sehr starke Einschränkung an die t -Schrittweite. Dieses Resultat war uns natürlich schon aus der Untersuchung der absoluten Stabilität des Euler-Verfahrens bekannt. Die gefundene Stabilitätsbedingung nennt man die L_2 -Stabilitätsbedingung entsprechend der oben skizzierten Bedeutung der Norm $\|\cdot\|$. Wir erhalten somit

Satz 2.4 Sei $u(x, t)$ die Lösung von

$$\begin{aligned} u_t &= u_{xx} & 0 \leq x \leq L, & \quad 0 < t < T \\ u(x, 0) &= f(x) & 0 \leq x \leq L, \\ u(0, t) &= \varphi_0(t), \quad u(L, t) = \varphi_1(t) \end{aligned}$$

und $u \in C^6(]0, L[\times]0, T[)$. Ferner bezeichne

$$u_n \stackrel{\text{def}}{=} \left(u(x_1, t_n), \dots, u(x_m, t_n) \right) \quad \text{mit} \quad x_i = i\Delta x, \quad \Delta x = \frac{L}{m+1}$$

und u_n^h sei definiert durch

$$u_0^h \stackrel{\text{def}}{=} (f(x_1), \dots, f(x_m))^T$$

$$(I - \alpha\rho A)u_{n+1}^h = (I + (1 + \alpha)\rho A)u_n^h + \rho \left((1 - \alpha) \begin{pmatrix} \varphi_0(t_n) \\ 0 \\ \vdots \\ 0 \\ \varphi_1(t_n) \end{pmatrix} + \alpha \begin{pmatrix} \varphi_0(t_{n+1}) \\ 0 \\ \vdots \\ 0 \\ \varphi_1(t_{n+1}) \end{pmatrix} \right)$$

Dann gilt mit $n\Delta t \nearrow T^* \leq T$, $\Delta t \rightarrow 0$, $\Delta x \rightarrow 0$

$$\|u_n^h - u_n\| \leq T^* \sup_{t \leq T^*} \|\tau(t)\| = \begin{cases} \mathcal{O}((\Delta t)^2) + \mathcal{O}((\Delta x)^2) & \alpha = \frac{1}{2} \\ \mathcal{O}((\Delta t)^2) + \mathcal{O}((\Delta x)^4) & \alpha = 0, \quad \rho = \frac{1}{6} \\ \mathcal{O}(\Delta t) + \mathcal{O}((\Delta x)^2) & \text{sonst} \end{cases}$$

falls zusätzlich gilt:

$$\rho = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2(1 - 2\alpha)} \quad \text{für} \quad 0 \leq \alpha < \frac{1}{2}.$$

□

Eine Fehlerabschätzung in der diskreten L_2 -Norm entspricht in der Regel nicht den Wunschvorstellungen eines Anwenders, vielmehr möchte man eine Aussage über den maximal auftretenden Fehler an einer einzelnen Stelle (x, t) haben. Hierzu muß man ε_n in der $\|\cdot\|_\infty$ -Norm abschätzen.

Wir kehren also zurück zur Darstellung

$$\varepsilon_{n+1} = (I - \alpha\rho A)^{-1}(I + (1 - \alpha)\rho A)\varepsilon_n + \Delta t(I - \alpha\rho A)^{-1}\tau_{n+1}$$

Hierin ist $(I - \alpha\rho A)^{-1} \geq 0$ (komponentenweise) da $I - \alpha\rho A$ eine irreduzibel diagonaldominante L -Matrix (also eine M -Matrix) bzw. die Identität ist, für $\alpha \geq 0$ und $\rho > 0$ beliebig. Wir zeigen nun, daß

$$\|(I - \alpha\rho A)^{-1}\|_\infty \leq 1.$$

Sei $\zeta \stackrel{\text{def}}{=} \|(I - \alpha\rho A)^{-1}\|_\infty > 1$ angenommen. Dann gilt

$$\zeta = |z_{i_0}| = \max_i |z_i| > 1 \quad \text{mit} \quad z = (I - \alpha\rho A)^{-1}e \geq 0, \quad e = (1, \dots, 1)^T$$

d.h. $z_i = |z_i|$ und

$$\begin{aligned} e &= (I - \alpha\rho A)z \\ 1 = (e)_{i_0} &= (-\alpha\rho z_{i_0-1} + (1 + 2\alpha\rho)z_{i_0} - \alpha\rho z_{i_0+1}) \\ &= z_{i_0} + \alpha\rho \underbrace{(z_{i_0} - z_{i_0-1})}_{\geq 0} + \alpha\rho \underbrace{(z_{i_0} - z_{i_0+1})}_{\geq 0} \\ &\geq z_{i_0} > 1 \quad \text{Widerspruch!} \end{aligned}$$

Ferner ist

$$\|(I + (1 - \alpha)\rho A)\|_\infty = 2(1 - \alpha)\rho + |1 - 2(1 - \alpha)\rho| = 1$$

falls

$$\rho \leq \frac{1}{2(1 - \alpha)} \quad 0 \leq \alpha < 1, \quad \text{sonst } \rho \text{ beliebig.}$$

Dies ergibt:

Satz 2.5 Die Aussage von Satz 2.4 gilt auch in der Norm $\|\cdot\|_\infty$, falls die Bedingung an die Schrittweitenkopplung ersetzt wird durch

$$\rho = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2(1 - \alpha)} \quad \text{für } 0 \leq \alpha < 1$$

□

Bemerkung 2.1 *Unter den Bedingungen des Satzes 2.5 erfüllt die Gitterfunktion u_{ij}^h das diskrete Maximum–Minimum–Prinzip, das dem kontinuierlichen Maximum–Minimum–Prinzip von Satz 2.1 entspricht*

$$\min_{\substack{0 \leq i \leq m \\ 0 \leq j \leq n}} (u_{i,j}^h) = \min_{\substack{0 \leq i \leq m \\ 0 \leq j \leq n}} \max \{u_{0,j}^h, u_{m+1,j}^h, u_{i,0}^h\}$$

(Den Beweis kann man induktiv über die Zeitschichten führen.) □

Man kann auch für variable Koeffizienten einen direkten Konvergenzbeweis für die Standarddiskretisierung durchführen, ein Beispiel einer solchen Vorgehensweise sei hier angefügt:

Beispiel 2.2

$$\begin{aligned} u_t &= a(x, t)u_{xx} & 0 \leq x \leq L, \quad 0 \leq t \leq T \\ u(x, 0) &= f(x) \\ u(0, t) &= \varphi_0(t) \text{ mit } f(0) = \varphi_0(0), \\ u(L, t) &= \varphi_1(t) \text{ mit } f(L) = \varphi_1(0) \end{aligned}$$

Voraussetzungen: $0 < a^* \leq a(x, t) \leq a^{**}$ und $\|\nabla a(x, t)\| \leq \beta$ für alle $(x, t) \in \bar{G} = [0, L] \times [0, T]$.

Wir betrachten wieder das übliche Einschrittverfahren (mit der Abkürzung $\delta^2 v_i = (v_{i+1} - 2v_i + v_{i-1})/(\Delta x)^2$)

$$u_{j,n+1}^h - u_{j,n}^h = \alpha \rho a_{j,n+1} \delta^2 u_{j,n+1}^h + (1-\alpha) \rho a_{j,n} \delta^2 u_{j,n}^h + \rho \left(\alpha \begin{pmatrix} \varphi_0(t_{n+1}) \\ 0 \\ \vdots \\ 0 \\ \varphi_1(t_{n+1}) \end{pmatrix} + (1-\alpha) \begin{pmatrix} \varphi_0(t_n) \\ 0 \\ \vdots \\ 0 \\ \varphi_1(t_n) \end{pmatrix} \right)$$

mit

$$\begin{aligned} 1 &\leq j \leq N \\ 0 &\leq n \leq M-1 \end{aligned}$$

wo $\rho \stackrel{\text{def}}{=} \Delta t / (\Delta x)^2$, $0 \leq \alpha \leq 1$, $\Delta x = \frac{L}{N+1}$, $\Delta t = \frac{T}{M}$.

Mit

$$\begin{aligned} \gamma_{j,n} &\stackrel{\text{def}}{=} \rho a_{j,n} \\ A_n &\stackrel{\text{def}}{=} \underset{j}{\text{tridiag}}(-\gamma_{j,n}, 2\gamma_{j,n}, -\gamma_{j,n}) \in \mathbb{R}^{N \times N} \\ \varepsilon_{j,n} &\stackrel{\text{def}}{=} u_{j,n} - u_{j,n}^h \\ \tau_n &\stackrel{\text{def}}{=} (\tau_{1,n}, \dots, \tau_{N,n})^T \end{aligned}$$

ergibt sich nun

$$(I + \alpha A_{n+1})\varepsilon_{n+1} = (I - (1 - \alpha)A_n)\varepsilon_n + \Delta t \tau_{.,n+1} .$$

Für τ_n gilt dabei die gleiche Abschätzung wie im Fall $u_t = u_{xx}$ (s.o.).

Verlangt man nun

$$\rho \leq \frac{1}{2a^{**}(1 - \alpha)},$$

so beweist man wörtlich wie zuvor die Konvergenz in $\|\cdot\|_\infty$. Zum Beweis der Konvergenz in $\frac{1}{\sqrt{N}}\|\cdot\|_2$ benutzt man, daß A_n ähnlich ist zu der symmetrischen Matrix

$$\text{tridiag}_j(-\sqrt{\gamma_{j-1,n}\gamma_{j,n}}, 2\gamma_{j,n}, -\sqrt{\gamma_{j,n}\gamma_{j+1,n}}) \stackrel{def}{=} B_n$$

mittels der diagonalen Ähnlichkeitstransformation $B_n = D_n A_n D_n^{-1}$

$$D_n = \text{diag}(1, \sqrt{\gamma_{2,n}/\gamma_{1,n}}, \dots, \sqrt{\gamma_{N,n}/\gamma_{1,n}}).$$

Mit

$$\begin{aligned} C_n &\stackrel{def}{=} (I - (1 - \alpha)B_n)(I + \alpha B_n)^{-1} \\ \Gamma_n &= (D_n D_{n+1}^{-1} - I)/\Delta t \\ \tilde{\varepsilon}_n &\stackrel{def}{=} D_n(I + \alpha A_n)\varepsilon_n \end{aligned}$$

wird dann

$$\tilde{\varepsilon}_{n+1} = (I + \Delta t \Gamma_n)^{-1} C_n \tilde{\varepsilon}_n + \Delta t (I + \Delta t \Gamma_n)^{-1} D_n \tau_{.,n+1}$$

$\|C_n\| \leq 1$, $\|\Gamma_n\| \leq K$ für $\rho \leq \frac{1}{2a_1(1-\alpha)}$ wenn $0 \leq \alpha < 1/2$, ρ beliebig für $\alpha \geq 1/2$, womit man wiederum die Konvergenz direkt gezeigt hat. □

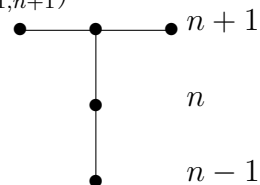
Bisher haben wir nur die allereinfachsten Diskretisierungen für die semidiskretisierte DGL

$$\dot{v} = \frac{1}{(\Delta x)^2} A v + F(t, v)$$

besprochen. Selbstverständlich ist jedes (Zeit-)Diskretisierungsverfahren, das auf der ganzen negativen reellen Achse absolut stabil (also A_0 -stabil) ist, im Prinzip geeignet. Besonders das Verfahren der rückwärtsgenommenen Differenzenquotienten (BDF, Gear-Formeln) bietet sich an. Da die betrachtete Raumdiskretisierung nur von 2. Ordnung konsistent ist, ist in diesem Zusammenhang nur das 2-Schritt-Verfahren 2. Ordnung interessant:

$$\frac{3}{2}(u_{j,n+1}^h - u_{j,n}^h) - \frac{1}{2}(u_{j,n}^h - u_{j,n-1}^h) = \rho(u_{j+1,n+1}^h - 2u_{j,n+1}^h + u_{j-1,n+1}^h)$$

$$\begin{aligned} u_{0,n+1}^h &= \varphi_0(t_{n+1}), & u_{m+1,n+1}^h &= \varphi_1(t_{n+1}) \\ u_{i,0}^h &= f(x_i) \\ \rho &= \frac{\Delta t}{(\Delta x)^2} \end{aligned}$$



Dieses Verfahren ist konsistent und konvergent in der diskreten L_2 -Norm mit einem Fehler $\mathcal{O}((\Delta x)^2) + \mathcal{O}((\Delta t)^2)$ ohne Einschränkung an ρ . Sein Nachteil besteht darin, daß man die erste Zeitschicht $n = 1$ separat berechnen muß (wie bei jedem 2-Schritt-Verfahren). Man könnte natürlich daran denken, die durch Semidiskretisierung entstandene DGL durch ein explizites Mehrschrittverfahren der Ordnung 2 (in Δt) zu integrieren, weil man sich dadurch die Auflösung linearer Gleichungssysteme spart. Eine naheliegende Methode wäre die Anwendung der Mittelpunkregel:

$$u_{n+1}^h = u_{n-1}^h + 2\Delta t f(t_n, u_n^h),$$

in der hier vorliegenden Notation

$$u_{j,n+1}^h = u_{j,n-1}^h + 2\rho(u_{j+1,n}^h - 2u_{j,n}^h + u_{j-1,n}^h) \quad \begin{array}{l} 1 \leq j \leq m \\ n = 1, 2, \dots \end{array}$$

Wir wissen aber bereits, daß die explizite Mittelpunkregel kein Intervall absoluter Stabilität der Form $[-\alpha, 0[$ besitzt. Dies wirkt sich so aus, daß für diese Anwendung mit $\Delta x \rightarrow 0$, $\Delta t \rightarrow 0$ für keinen Wert von $\rho = \Delta t/(\Delta x)^2$ Konvergenz eintritt. (Dies ist kein Widerspruch zur asymptotischen Stabilität (D -Stabilität) der Mittelpunkregel. Bei festem $\Delta x > 0$ konvergiert das Verfahren mit $\Delta t \rightarrow 0$ gegen die Lösung der **semidiskretisierten DGL**).

Durch eine scheinbar geringfügige Modifikation des Verfahrens gelangt man zu einem durchaus brauchbaren 2-Schrittverfahren der Ordnung $\mathcal{O}\left(\left(\frac{\Delta t}{\Delta x}\right)^2\right) + \mathcal{O}\left((\Delta t)^2 + (\Delta x)^2\right)$, dem Verfahren von **Du Fort und Frankel**.

Man ersetze $u_{j,n}^h$ durch den Mittelwert $\frac{1}{2}(u_{j,n+1}^h + u_{j,n-1}^h)$:

$$u_{j,n+1}^h = u_{j,n-1}^h + 2\rho(u_{j+1,n}^h - u_{j,n+1}^h - u_{j,n-1}^h + u_{j-1,n}^h) \\ j = 1, \dots, m, \quad n = 1, 2, \dots$$

Dieses Verfahren ist allerdings nur für $\lim_{\substack{\Delta t \rightarrow 0 \\ \Delta x \rightarrow 0}} \frac{\Delta t}{\Delta x} = 0$ konsistent zu $u_t \stackrel{\text{def}}{=} u_{xx}$:

Es ist nämlich

$$\begin{aligned} \frac{1}{2\Delta t}(u_{j,n+1} - u_{j,n-1}) &= (u_t)_{j,n} + \frac{(\Delta t)^2}{6}(u_{ttt})_{j,n} + \mathcal{O}((\Delta t)^4) \\ \frac{1}{(\Delta x)^2}(u_{j+1,n} - 2u_{j,n} + u_{j-1,n}) &= (u_{xx})_{j,n} + \frac{(\Delta x)^2}{12}(u_{xxx})_{j,n} + \mathcal{O}((\Delta x)^4) \\ 2u_{j,n} - (u_{j,n+1} + u_{j,n-1}) &= -(\Delta t)^2(u_{tt})_{j,n} + \mathcal{O}((\Delta t)^4) \end{aligned}$$

d.h.

$$\begin{aligned} \frac{1}{2\Delta t}(u_{j,n+1} - u_{j,n-1}) - \frac{1}{(\Delta x)^2}(u_{j+1,n} - u_{j,n+1} - u_{j,n-1} + u_{j-1,n}) &= \\ = (u_t)_{j,n} - (u_{xx})_{j,n} + \left(\frac{\Delta t}{\Delta x}\right)^2 (u_{tt})_{j,n} + \mathcal{O}((\Delta t)^2 + (\Delta x)^2 + \frac{(\Delta t)^4}{(\Delta x)^2}) \end{aligned}$$

d.h. für $\lim_{\substack{\Delta t \rightarrow 0 \\ \Delta x \rightarrow 0}} \frac{\Delta t}{\Delta x} = \beta \neq 0$ ist das Verfahren konsistent zu

$$u_t - u_{xx} + \beta^2 u_{tt} = 0 \quad (\text{hyperbolische Gleichung})$$

und zwar in $(\Delta t)^2 + (\Delta x)^2$.

Zum Stabilitätsnachweis vgl. Kapitel 3.

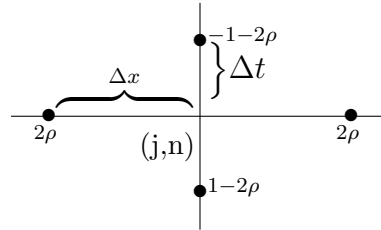


Abbildung 2.2

Bemerkung 2.2 Bei der Diskretisierung von von Neumann oder Robin-Randbedingungen benutzt man gerne die Methode der versetzten Gitter: Es ist dann

$$x_j = (j + \frac{1}{2})h, \quad -1 \leq j \leq N$$

und die Richtungsableitungen werden durch den symmetrischen Differenzenquotienten der Werte auf den beiden äusseren Gitterlinien und die Funktionswerte durch die Mittelwerte der Werte auf den beiden äusseren Gitterlinien angenähert. Man hat dann auch in diesen Werten Konsistenz der Ordnung 2. Oder man benutzt zwei zusätzliche äussere Gitterlinien und dazu die Differentialgleichung auch auf den Rändern $x = x_0$ und $x = x_N$.

2.2.2 Räumlich mehrdimensionale Probleme und damit verbundene Schwierigkeiten

Die gleiche Vorgehensweise ist wörtlich übertragbar auf den räumlich mehrdimensionalen Fall. Da man in der Regel mit einem impliziten Verfahren arbeiten wird, ist dann pro Zeitschritt ein sehr grosses lineares (oder im Fall einer nichtlinearen Aufgabe sogar nichtlineares) Gleichungssystem zu lösen, das nun nicht mehr die angenehme Struktur einer tridiagonalen oder doch zumindest sehr schmalen Bandmatrix hat. Wenn keine gemischten partiellen Ableitungen vorliegen, kann man dann vorteilhaft die Technik der sogenannten "Operator-Faktorisierung" anwenden. Im Zusammenhang mit parabolischen Problemen ist dies als sogenannte ADI-Methode bekannt. Wir beschreiben kurz den zweidimensionalen Fall mit dem Crank-Nicholson-Verfahren: Wir gehen aus von

$$\begin{aligned} u_t &= u_{xx} + u_{yy} & (x, y) &\in]0, 1[\times]0, 1[, \\ u(t, x, y) &= 0 & \text{falls } x &\in \{0, 1\} \text{ oder } y \in \{0, 1\} \\ u(0, x, y) &= u_0 & x, y &\in [0, 1] \times [0, 1] \end{aligned} \quad (2.3)$$

und benutzen die gleiche Vorgehensweise wie im eindimensionalen Fall, und ersetzen Δu auf der rechten Seite von (2.3) durch den Fünfpunkteoperator. Wir gelangen zu folgendem linearen Gleichungssystem (mit $\Delta x = \Delta y$ und $\rho = \Delta t/(\Delta x)^2$)

$$(I - \frac{\rho}{2}A)u^{(n+1)} = (I + \frac{\rho}{2}A)u^{(n)}$$

Dabei ist $u^{(n)}$ der Vektor der inneren räumlichen Gitterwerte, also von der Länge m^2 mit $\Delta x = \Delta y = \frac{1}{m+1}$. A ist die bekannte Block-Tridiagonalmatrix

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 & \cdots & 0 \\ A_{21} & A_{22} & A_{23} & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & A_{m-1,m} \\ 0 & \cdots & 0 & A_{m,m-1} & A_{m,m} \end{bmatrix}$$

$$A_{ij} = +I_m \quad i \neq j \quad A_{ii} = \text{tridiag}(+1, -4, +1)$$

Pro Zeitschritt hätte man also einen ganz beträchtlichen Rechenaufwand zu leisten. Die Idee des Verfahrens der alternierenden Richtungen ist die folgende: Obiges Gleichungssystem ist die Zusammenfassung der m^2 Gleichungen $1 \leq i, j \leq m$

$$\frac{(u_{ij}^h)^{(n+1)} - (u_{ij}^h)^{(n)}}{\Delta t} = \frac{1}{2(\Delta x)^2} \left\{ (u_{i-1,j}^h)^{(n+1)} - 2(u_{i,j}^h)^{(n+1)} + (u_{i+1,j}^h)^{(n+1)} + (u_{i-1,j}^h)^{(n)} - 2(u_{i,j}^h)^{(n)} + (u_{i+1,j}^h)^{(n)} \right\} + \frac{1}{2(\Delta y)^2} \left\{ (u_{i,j-1}^h)^{(n+1)} - 2(u_{i,j}^h)^{(n+1)} + (u_{i,j+1}^h)^{(n+1)} + (u_{i,j-1}^h)^{(n)} - 2(u_{i,j}^h)^{(n)} + (u_{i,j+1}^h)^{(n)} \right\}$$

Wenn wir die Differenzenoperatoren zweiter Ordnung in der Richtung x mit δ_{xx} und in y -Richtung entsprechend δ_{yy} schreiben, dann heißt dies

$$(u_{ij}^h)^{(n+1)} = (u_{ij}^h)^{(n)} + \frac{\rho}{2}(\delta_{xx}(u_{ij}^h)^{(n+1)} + \delta_{xx}(u_{ij}^h)^{(n)}) + \frac{\rho}{2}(\delta_{yy}(u_{ij}^h)^{(n+1)} + \delta_{yy}(u_{ij}^h)^{(n)}) .$$

Da diese Approximation von zweiter Ordnung in Δt und $\Delta x, \Delta y$ konsistent ist, stört es die Konsistenzordnung nicht, wenn wir diese Formel abändern zu

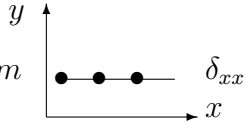
$$(u_{ij}^h)^{(n+1)} = (u_{ij}^h)^{(n)} + \frac{\rho}{2}(\delta_{xx}(u_{ij}^h)^{(n+1)} + \delta_{xx}(u_{ij}^h)^{(n)}) + \frac{\rho}{2}(\delta_{yy}(u_{ij}^h)^{(n+1)} + \delta_{yy}(u_{ij}^h)^{(n)}) + \frac{\rho^2}{4}\delta_{xx}\delta_{yy}((u_{ij}^h)^{(n)} - (u_{ij}^h)^{(n+1)})$$

Die Operatoren δ_{xx} und δ_{yy} sind linear und vertauschbar. Deshalb erhält man

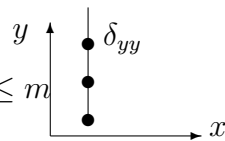
$$(1 - \frac{\rho}{2}\delta_{xx})\left((1 - \frac{\rho}{2}\delta_{yy})(u_{ij}^h)^{(n+1)}\right) = (1 + \frac{\rho}{2}\delta_{yy})\left((1 + \frac{\rho}{2}\delta_{xx})(u_{ij}^h)^{(n)}\right) \quad (2.4)$$

Die Idee, dieses Gleichungssystem (das explizit ausgeschrieben eher komplizierter ist als das ursprüngliche) ökonomisch zu lösen, besteht nun in der Einführung von Zwischengrößen.

Wenn wir die inneren Gitterpunkte zeilenweise numerieren, ist das lineare Gleichungssystem

$$(1 - \frac{\tau}{2}\delta_{xx})(u_{ij}^h)^{(n+\frac{1}{2})} = (1 + \frac{\tau}{2}\delta_{xx})(u_{ij}^h)^{(n)} \quad 1 \leq i, j \leq m$$


tridiagonal. Danach numerieren wir die Variablen nun spaltenweise und das Gleichungssystem

$$(1 - \frac{\tau}{2}\delta_{yy})(u_{ij}^h)^{(n+1)} = (1 + \frac{\tau}{2}\delta_{yy})(u_{ij}^h)^{(n+\frac{1}{2})} \quad 1 \leq i, j \leq m$$


ist erneut tridiagonal. Damit ist aber die Lösung von (2.4) bestimmt. Dieses Verfahren ist konsistent von zweiter Ordnung in Δt , Δx , Δy . Eine genauere mathematische Analyse der Konvergenzbedingungen erfolgt in Kapitel 3. Solche (approximativen) Faktorisierungsmethoden sind auch für andere Problemstellungen bekannt und ein wesentliches Hilfsmittel, um räumlich mehrdimensionale Probleme praktisch behandelbar zu machen. Bei drei Raumvariablen benutzt man gern die Approximation von Douglas und Gunn

$$1 - \frac{\tau}{2}(\delta_{xx} + \delta_{yy} + \delta_{zz}) = (1 - \frac{\tau}{2}\delta_{xx})(1 - \frac{\tau}{2}\delta_{yy})(1 - \frac{\tau}{2}\delta_{zz}) + \mathcal{O}(\tau^2)$$

und rechnet gemäss

$$\begin{aligned} (1 - \frac{\tau}{2}\delta_{xx})u^{(n+1/3)} &= (1 + \frac{\tau}{2}(\delta_{xx} + 2\delta_{yy} + 2\delta_{zz}))u^{(n)} \\ (1 - \frac{\tau}{2}\delta_{yy})u^{(n+2/3)} &= u^{(n+1/3)} - \frac{\tau}{2}\delta_{yy}u^{(n)} \\ (1 - \frac{\tau}{2}\delta_{zz})u^{(n+1)} &= u^{(n+2/3)} - \frac{\tau}{2}\delta_{zz}u^{(n)} \end{aligned}$$

Nun hat man also drei tridiagonale Gleichungssysteme zu lösen.

Die Faktorisierung ist jedoch unmöglich, wenn gemischte partielle Ableitungen vorliegen. Einen elliptischen Differentialoperator zweiter Ordnung mit konstanten Koeffizienten kann man zwar stets auf die Normalform $-\Delta u$ transformieren, bei variablen Koeffizienten ist dies jedoch nicht möglich. Um den hohen Aufwand durch die Lösung der extrem grossen Gleichungssysteme im allgemeinen mehrdimensionalen Fall zu vermeiden, muß man sich auf explizite Verfahren beschränken. Es ist deshalb von Interesse, daß es explizite Integrationsverfahren für gewöhnliche Differentialgleichungen gibt, die ein sehr grosses Gebiet der absoluten Stabilität besitzen, sodaß man auf eine Zeitschrittweitenrestriktion

$$\frac{\Delta t}{(\Delta x)^2} \leq C$$

mit einer grossen Konstanten C geführt wird. Solche Verfahren wurden von Sommeijer und van der Houwen angegeben. (P.J. van der Houwen and B.P. Sommeijer: On the internal stability of explicit, m-stage Runge-Kutta methods for large m-values. Z. Angew. Math. Mech. 60 (1980), pp.479-485. J.G. Verwer, W.H. Hundsdorfer und B.P. Sommeijer: Convergence properties of the Runge-Kutta-Chebyshev method. Numer. Math. 57 (1990), pp.157-178.) Es sind explizite Runge-Kutta-Verfahren zweiter Ordnung mit einer hohen Stufenzahl m , die auf der Quadratur mit den Tschebyscheff-Knoten aufbauen. Hier wächst C quadratisch mit der Stufenzahl.

Bemerkung 2.3 *Einerseits stellt die vertikale Linienmethode eine einfache und wirksame Lösungsmöglichkeit für parabolische Randanfangswertprobleme dar, da sie die hoch entwickelten Techniken zur Lösung gewöhnlicher Differentialgleichungen für den Bereich der partiellen Differentialgleichungen nutzbar macht. Andererseits besitzt sie auch einen wesentlichen Nachteil: In der Praxis wird man kaum mit äquidistanten Gittern arbeiten können und die Bereiche, in denen zur genügend genauen Beschreibung der Lösung ein verfeinertes Gitter notwendig ist, werden in der Regel zeitabhängig sein. Dem steht aber die zeitunabhängige Vorwahl der "senkrechten" Linien entgegen, und man muss bei einer Verschiebung der Gitterverfeinerung zu Interpolationsmethoden greifen, um das neue Gitter zu generieren. Die Interpolationsfehler können aber ungünstige Auswirkungen auf die numerische Stabilität und Fehlerdämpfung der Methode haben*

2.2.3 Ein nichtlineares parabolisches Problem ERG

In den vorausgegangenen Abschnitten und auch in der allgemeinen Theorie von Kapitel 3 werden nur lineare Probleme behandelt. Hier soll an einem speziellen Fall gezeigt werden, wie auch nichtlineare Probleme analysiert und numerisch gelöst werden können. Wir gehen aus von folgender Aufgabenstellung

$$\begin{aligned} u_t &= \psi(x, t, u, u_x, u_{xx}), & 0 \leq x \leq L, & \quad 0 \leq t \leq T \\ u(x, 0) &= f(x) & 0 \leq x \leq L \\ u(0, t) &= \varphi_0(t) \\ u(L, t) &= \varphi_1(t) & t \geq 0. \end{aligned}$$

Es gelte für $\psi : [0, L] \times [0, T] \times \mathbb{R}^3 \rightarrow \mathbb{R}$

$$0 < \gamma_0 \leq \partial_5 \psi(\dots) \leq \gamma_1, \quad |\partial_3 \psi(\dots)|, |\partial_4 \psi(\dots)| \leq K.$$

Es sei

$$\begin{aligned} x_j &= j\Delta x & j = 0, \dots, M+1, & \quad M+1 = \frac{L}{\Delta x} \\ t_n &= n\Delta t, & n = 0, \dots, N, & \quad N = \frac{T}{\Delta t} \\ \varrho &\stackrel{\text{def}}{=} \frac{\Delta t}{(\Delta x)^2} > 0 & \text{fest.} \end{aligned}$$

Semidiskretisierung und Anwendung eines linearen Einschrittverfahren auf die entstehende steife nichtlineare gewöhnliche DGL (für M Funktionen) führt auf

$$u_{j,n+1}^h = u_{j,n}^h + \Delta t(\alpha\psi_{j,N+1} + (1-\alpha)\psi_{j,n}), \quad 1 \leq j \leq M, \quad n = 0, \dots, N-1$$

mit

$$\psi_{j,k} = \psi(x_j, t_k, u_{j,k}^h, \frac{u_{j+1,k}^h - u_{j-1,k}^h}{2\Delta x}, \frac{u_{j+1,k}^h - 2u_{j,k}^h + u_{j-1,k}^h}{(\Delta x)^2}).$$

Der Abschneidefehler ergibt sich wie im linearen Fall zu

$$\tau_{j,n+1} = \begin{cases} \mathcal{O}(\Delta t + (\Delta x)^2) & \text{für } \alpha \neq \frac{1}{2} \\ \mathcal{O}((\Delta t)^2 + (\Delta x)^2) & \text{für } \alpha = \frac{1}{2} \end{cases}$$

(Taylorentwicklung von ψ).

Somit für

$$\begin{aligned} \varepsilon_{j,k} &\stackrel{def}{=} u_{j,k} - u_{j,k}^h \\ \varepsilon_{j,n+1} &= \varepsilon_{j,n} + \Delta t \left(\alpha \left(\partial_3 \psi(\dots) \varepsilon_{j,n+1} + \partial_4 \psi(\dots) \frac{\varepsilon_{j+1,n+1} - \varepsilon_{j-1,n+1}}{2\Delta x} \right. \right. \\ &\quad \left. \left. + \partial_5 \psi(\dots) \frac{\varepsilon_{j+1,n+1} - 2\varepsilon_{j,n+1} + \varepsilon_{j-1,n+1}}{(\Delta x)^2} \right) \right. \\ &\quad \left. + (1-\alpha) \left(\partial_3 \psi(\dots) \varepsilon_{j,n} + \partial_4 \psi(\dots) \frac{\varepsilon_{j+1,n} - \varepsilon_{j-1,n}}{2\Delta x} \right. \right. \\ &\quad \left. \left. + \partial_5 \psi(\dots) \frac{\varepsilon_{j+1,n} - 2\varepsilon_{j,n} + \varepsilon_{j-1,n}}{(\Delta x)^2} \right) \right) + \Delta t \tau_{j,n+1} \\ \varepsilon_{j,0} &= 0, \\ \varepsilon_{0,n} &= \varepsilon_{M+1,n} = 0. \end{aligned}$$

Die partiellen Ableitungen von ψ sind dabei an geeigneten Zwischenstellen zu nehmen. In Vektor-Matrix-Schreibweise ergibt dies

$$A_{n+1}\varepsilon_{n+1} = B_n\varepsilon_n + \Delta t \tau_n$$

Dabei ist

$$\begin{aligned} A_{n+1} &= \text{tridiag} \left(+\alpha \varrho(\Delta x/2) \partial_4 \psi(\dots) - \alpha \varrho \partial_5 \psi(\dots), \right. \\ &\quad \left. 1 - \alpha(\Delta x)^2 \varrho \partial_3 \psi(\dots) + 2\alpha \varrho \partial_5 \psi(\dots), \right. \\ &\quad \left. - \alpha \varrho(\Delta x/2) \partial_4 \psi(\dots) - \alpha \varrho \partial_5 \psi(\dots) \right) \\ B_n &= \text{tridiag} \left(-(1-\alpha) \varrho(\Delta x/2) \partial_4 \psi(\dots) + (1-\alpha) \varrho \partial_5 \psi(\dots), \right. \\ &\quad \left. 1 + (1-\alpha)(\Delta x)^2 \varrho \partial_3 \psi(\dots) - 2\varrho \partial_5 \psi(\dots)(1-\alpha), \right. \\ &\quad \left. (1-\alpha) \varrho(\Delta x/2) \partial_4 \psi(\dots) + (1-\alpha) \varrho \partial_5 \psi(\dots) \right). \end{aligned}$$

Die Argumente in den partiellen Ableitungen von ψ stimmen dabei jeweils in einer Zeile überein. Es gelte

$$\Delta x < \min\{1, 2\gamma_0/K\}, \quad \Delta t < 1/K.$$

Dann ist A_{n+1} eine irreduzibel diagonaldominante L -Matrix, also eine M -Matrix.

$$\begin{aligned} A_{n+1}e &\geq (1 - \alpha\varrho(\Delta x)^2K)e, & e = (1, \dots, 1)^T \\ \Rightarrow e &\geq (1 - \alpha\varrho(\Delta x)^2K)A_{n+1}^{-1}e \end{aligned}$$

d.h.

$$\|A_{n+1}^{-1}\|_\infty \leq \frac{1}{1 - \alpha\Delta tK}$$

und

$$\|B_n\|_\infty \leq 1 + (1 - \alpha)\Delta tK.$$

Die letzte Abschätzung setzt voraus, daß

$$1 - \varrho(1 - \alpha)(2\gamma_1 + (\Delta x)^2K) \geq 0$$

d.h.

$$\varrho \leq \frac{1}{2(1 - \alpha)(\gamma_1 + (\Delta x)^2K/2)}$$

d.h.

$$\Delta t \leq \frac{(\Delta x)^2}{2(1 - \alpha)(\gamma_1 + (\Delta x)^2K/2)}.$$

Diese Bedingung ist nicht wesentlich stärker als im linearen Fall. Im ganzen hat man damit

$$\|\varepsilon_{n+1}\|_\infty \leq \frac{1 + (1 - \alpha)K\Delta t}{1 - \alpha\Delta tK} \|\varepsilon_n\|_\infty + \frac{1}{1 - \alpha\Delta tK} \Delta t \|\tau_n\|_\infty$$

d.h. Konvergenz in $\|\cdot\|_\infty$ von der Ordnung von τ_n .

Man beachte, daß für

$$K \gg \gamma_0$$

schon Δx sehr klein gewählt werden muß. Die Forderung an Δt ist dann außer im “vollimpliziten” Fall (Euler rückwärts) extrem restriktiv. Man bezeichnet die Gleichung in diesem Fall als Konvektions- Diffusionsgleichung (vom praktischen Standpunkt aus verhält sich die Gleichung wie eine hyperbolische). Für diesen Gleichungstyp gibt es speziell angepaßte Verfahren, die die ungünstige Schrittweitenforderung vermeiden. Man benutzt dabei für die Diskretisierung von u_x nicht den symmetrischen Differenzenquotienten zweiter Ordnung, sondern eine Linearkombination

$$(1 - \vartheta_j) \frac{u_j - u_{j-1}}{\Delta x} + \vartheta_j \frac{u_{j+1} - u_j}{\Delta x}$$

Ein günstiger Wert für ϑ_j ist

$$\vartheta_j = \begin{cases} \max\{\frac{1}{2}, 1 - \partial_5\psi_{j,n}/(\Delta x\partial_4\psi_{j,n})\} & \text{für } \partial_4\psi_{j,n} \geq 0 \\ \min\{\frac{1}{2}, \partial_5\psi_{j,n}/(\Delta x|\partial_4\psi_{j,n}|)\} & \text{für } \partial_4\psi_{j,n} < 0. \end{cases}$$

2.3 Galerkin–Verfahren für Randanfangswertaufgaben

Das Konzept der Semidiskretisierung gilt heute allgemein als bestes Mittel zur Lösung von Randanfangswertaufgaben partieller Differentialgleichungen, bei denen eine Veränderliche die Rolle der Zeit spielt (d.h. die Lösungen verhalten sich als Funktionen dieser Veränderlichen “stabil”). Die in Abschnitt 2.2 betrachtete Form der Methode (Diskretisierung und Differenzenquotienten) erfordert aber letztlich, daß das Gebiet in den Raumvariablen von einfacher Gestalt ist (Vereinigung von Rechtecksgebieten), weil man sonst die gleichen Probleme wie bei der Lösung elliptischer RWA mit Differenzenformeln auf Nichtrechtecksgebieten bekommt. Diese Schwierigkeit wird umgangen bei der Galerkinmethode, bei der die partielle DGL bzgl. der Raumvariablen mittels der Methode der finiten Elemente diskretisiert wird.

Ausgangspunkt ist eine parabolische oder hyperbolische RAWA

$$\left. \begin{aligned} -u_t(x, t) \\ (-u_{tt}(x, t)) \end{aligned} \right\} &= Lu(x, t) + f(x, t) & (x, t) \in B \times]0, T] \\ u(x, 0) &= u_0(x) & x \in B \\ (u_t(x, 0) &= u_1(x)) & x \in B \\ u(x, t) &= \psi(x, t) & x \in \partial B, \quad t \in]0, T]. \end{aligned}$$

Dabei ist B ein wegzusammenhängendes Lipschitzgebiet in \mathbb{R}^d ($d = 1, 2, 3$), nicht notwendig einfach zusammenhängend, mit stückweise glattem Rand ∂B und L ein linearer, gleichmäßig elliptischer Differentialoperator (bzgl. x), d.h.

$$(Lu, v) = [u, v], \quad [., .] \text{ Skalarprodukt}$$

für $u, v \in D_L$, wobei D_L eine dichte Teilmenge eines geeigneten Hilbertraumes $\mathcal{H} \subset L_2(B)$ und $(., .)$ das L_2 -Skalarprodukt auf B ist.

(Auch semilineare Probleme lassen sich leicht behandeln, wie in HNMI beschrieben.)

Für unsere Anwendungen ist \mathcal{H} stets ein Sobolev-Raum, also etwa $H_0^1(B)$. Inhomogene Randwerte werden dann in die Ansatzfunktionen eingearbeitet. Wir betrachten im Folgenden den Fall $\psi \equiv 0$. Man wählt nun wieder eine Schar endlich-dimensionaler Teilräume $S_h \subset \mathcal{H}$ mit Scharparameter $0 < h \leq h_0$ und betrachtet die schwache Form der DGL und ihre Semidiskretisierung:

$$-(u_t, v) = (Lu, v) + (f, v) \quad 0 \leq t \leq T, \quad \forall v \in \mathcal{H}$$

bzw. mit

$$\begin{aligned} u^h &\stackrel{\text{def}}{=} \sum_{i=1}^n \alpha_i(t) \varphi_i(x), \quad n = \dim S_h, \quad \{\varphi_1, \dots, \varphi_n\} \text{ Basis von } S_h \\ -(\sum_{i=1}^n \dot{\alpha}_i \varphi_i, \varphi_j) &= [\sum_{i=1}^n \alpha_i \varphi_i, \varphi_j] + (f, \varphi_j), \quad j = 1, \dots, n. \end{aligned}$$

(Im hyperbolischen Fall erhält man entsprechend eine DGL 2. Ordnung

$$-\left(\sum_{i=1}^n \ddot{\alpha}_i \varphi_i, \varphi_j\right) = \left[\sum_{i=1}^n \alpha_i \varphi_i, \varphi_j\right] + (f, \varphi_j).$$

Aus den Anfangsvorgaben für $t = 0$ erhält man die Anfangswerte für $\alpha_i(0)$ (und $\dot{\alpha}_i(0)$ im hyperbolischen Fall). Man erhält damit eine Anfangswertaufgabe für ein System gewöhnlicher Differentialgleichungen erster (oder zweiter) Ordnung

$$\begin{aligned} -M\dot{a}(t) &= Ka(t) + F(t), & a(0) &= a_0 \\ \text{(bzw. } -M\ddot{a}(t) &= Ka(t) + F(t), & a(0) &= a_0, \quad \dot{a}(0) = a_1). \end{aligned}$$

Wenn die Koeffizienten des Differentialoperators L nicht von t abhängen, sind M und K konstante Matrizen und F der Vektor

$$F(t) = ((f(\cdot, t), \varphi_j))_{j=1}^n.$$

M ist die sogenannte Massenmatrix und K die Gesamtsteifigkeitsmatrix des Finite-Element-Ansatzes:

$$M_{ij} = (\varphi_i, \varphi_j), \quad K_{ij} = [\varphi_i, \varphi_j].$$

M und K sind beide symmetrisch und positiv definit und bei Verwendung eines Finite-Element-Ansatzes dünn besetzt (bandförmig). Die Lösungen der gewöhnlichen DGL

$$-M\dot{a}(t) = Ka(t), \quad a(0) = a_0,$$

können dargestellt werden in der Form

$$a(t) = \sum_{j=1}^n z_j e^{\lambda_j t}$$

wobei z_1, \dots, z_n geeignete konstante Vektoren und λ_j die Eigenwerte des verallgemeinerten Eigenwertproblems

$$-\lambda Mx = Kx$$

sind. Diese Eigenwerte sind sämtlich reell und negativ und der betragsgrößte hat die Größenordnung n^p wobei p die Ordnung des Differentialoperators ist, d.h. die Differentialgleichung ist für größeres n sehr steif. Entsprechendes gilt für die Lösung von

$$-M\ddot{a} = Ka \quad (\text{zugehörige Eigenwerte} \quad -\lambda^2 Mx = Kx)$$

vgl. dazu die Ausführungen in Kapitel 1.

Die Ordnung der Zeitdiskretisierung der gewöhnlichen DGL paßt man der Ordnung der Raumdiskretisierung an.

Z.B. bei $d = 1$ ($u_t = u_{xx}$ z.B) und kubischen Splines als Ansatzfunktionen (mit einem Approximationsfehler $\mathcal{O}(\Delta x^4)$ in der L_2 -Norm) wird man die DGL

$$-M\dot{a}(t) = Ka(t) + F(t)$$

etwa mit dem A -stabilen Rosenbrock-Wanner-Verfahren der Ordnung 4 integrieren und erhält dann einen Gesamtfehler $\mathcal{O}(\Delta t^4 + \Delta x^4)$ in der L_2 -Norm. Bei $d = 2$ und quadratischen finiten Elementen auf einer regulären Triangulierung wird man entsprechend mit einem A -stabilen Verfahren dritter Ordnung rechnen.

Im hyperbolischen Fall muß man mit einem der speziellen Verfahren für Differentialgleichungen zweiter Ordnung arbeiten, die in Kapitel 1 angegeben sind.

2.4 Die Lösung der eindimensionalen Wärmeleitungsgleichung mit dem Galerkinansatz.

Das oben beschriebene Konzept soll nun für die spezielle RAWA

$$\begin{aligned} -u_t &= -\frac{\partial}{\partial x}(a(x)\frac{\partial}{\partial x}u) + c(x)u - f(x,t) \\ u(a,t) &= u(b,t) = 0 & t \geq 0 \\ u(x,0) &= g(x) & a \leq x \leq b, \quad g(a) = g(b) = 0 \end{aligned}$$

näher untersucht werden. Als Hilbertraum benutzen wir

$$\mathcal{H} \stackrel{def}{=} \mathcal{K}_0^1[a, b]$$

d.h. die homogenen Randbedingungen werden in die Ansatzfunktionen eingearbeitet. Unter den Voraussetzungen

$$\begin{aligned} a &\in C^1[a, b], & a(x) &\geq \alpha > 0 \\ c &\in C[a, b], & c(x) &\geq 0 \end{aligned}$$

erfüllt dann

$$Lu \stackrel{def}{=} -\frac{\partial}{\partial x}(a(x)\frac{\partial}{\partial x}u) + c(x)u$$

die Voraussetzungen

$$(Lu, v) = [u, v] = \int_a^b \{a(x)u'(x)v'(x) + c(x)u(x)v(x)\}dx$$

für alle $u, v \in C_0^2[a, b]$ und

$$[u, v] \geq \frac{\alpha}{1 + (b - a)^2} \|u\|_{2,1}^2 \quad \forall u \in \mathcal{K}_0^1[a, b].$$

Sei nun S_h ein endlichdimensionaler Teilraum von \mathcal{H} und $\{\varphi_1, \dots, \varphi_n\}$ Basis von S_h .

Wir setzen

$$\begin{aligned} M &= ((\varphi_i, \varphi_j))_{1 \leq i, j \leq n}, & (\cdot, \cdot) & \text{ } L_2\text{-Skalarprodukt} \\ K &= ([\varphi_i, \varphi_j])_{1 \leq i, j \leq n}, & [\cdot, \cdot] & \text{ obiges "Energie"-Skalarprodukt} \\ -M\dot{a}(t) &= Ka(t) + F(t) & t & \geq 0 \end{aligned}$$

mit

$$\begin{aligned} F(t) &= ((\varphi_i, f(\cdot, t)))_{i=1}^n \\ Ma(0) &= c = ((\varphi_i, g))_{i=1}^n \end{aligned}$$

und mit der Lösung $a(t) = (\alpha_1(t), \dots, \alpha_n(t))^T$ des obigen Anfangswertproblems

$$u^h(x, t) \stackrel{\text{def}}{=} \sum_{i=1}^n \alpha_i(t) \varphi_i(x).$$

Wir bezeichnen u^h als Galerkin-Approximation für u . Man beachte, daß u^h automatisch als kontinuierliche Funktion von x und t definiert ist und daß u^h von der gleichen Ordnung bzgl. x differenzierbar ist wie die Ansatzfunktionen, nach t jedoch einmal mehr als die Inhomogenität f des Ausgangsproblems.

Wir zeigen nun

Satz 2.6 Für $t \in [0, T]$ sei die Projektion Pu von u auf S_h dargestellt als

$$(Pu)(x, t) = \sum_{i=1}^n \beta_i(t) \varphi_i(x)$$

mit

$$\begin{aligned} [u - Pu, \varphi_j] &= 0 \quad j = 1, \dots, n \quad \text{d.h.} \\ ([u, \varphi_j])_{j=1}^n &= K(\beta_1, \dots, \beta_n)^T \end{aligned}$$

(Damit sind also die Koeffizienten β_j differenzierbare Funktionen von t . Sie sind aber von den Koeffizienten $\alpha_j(t)$, die Lösung der obigen AWA sind, verschieden). Ferner sei

$$\varepsilon(x, t) \stackrel{\text{def}}{=} u(x, t) - u^h(x, t; h), \quad s(x, t) = u(x, t) - (Pu)(x, t).$$

Dann gilt mit

$$\lambda \stackrel{\text{def}}{=} \frac{\alpha}{1 + (b - a)^2}$$

$$\begin{aligned} \|\varepsilon(\cdot, t)\|_{2,0} &\leq \|s(\cdot, t)\|_{2,0} + e^{-\lambda t} \|Pu(\cdot, 0) - u^h(\cdot, 0; h)\|_{2,0} \\ &\quad + \int_0^t e^{\lambda(\tau-t)} \left\| \frac{d}{dt} s(\cdot, t) \Big|_{t=\tau} \right\|_{2,0} d\tau \end{aligned}$$

Beweis: Es gilt

$$\begin{aligned} (u_t, v) + [u, v] &= (f, v) & \forall v \in \mathcal{H} \\ (u_t^h, \sum \gamma_j \varphi_j) + [u^h, \sum \gamma_j \varphi_j] &= (f, \sum \gamma_j \varphi_j) & \text{für beliebige } \gamma_1, \dots, \gamma_n. \end{aligned}$$

Somit für $v = \sum \gamma_j \varphi_j = Pu - u^h$ d.h. $\gamma_j = \beta_j - \alpha_j$:

$$(u_t - u_t^h, \sum \gamma_j \varphi_j) + [u - u^h, \sum \gamma_j \varphi_j] = 0 \quad \forall t \in [0, T].$$

Aber

$$\begin{aligned} [u - u^h, \sum \gamma_j \varphi_j] &= [u - Pu + Pu - u^h, \sum \gamma_j \varphi_j] \\ &= [u - Pu, \sum \gamma_j \varphi_j] + [Pu - u^h, Pu - u^h] = [Pu - u^h, Pu - u^h] \end{aligned}$$

d.h.

$$\begin{aligned} \left(\frac{d}{dt} \varepsilon, Pu - u^h\right) + [Pu - u^h, Pu - u^h] &= 0 \\ \left(\frac{d}{dt} s, Pu - u^h\right) + [Pu - u^h, Pu - u^h] &= \left(\frac{d}{dt} (s - \varepsilon), Pu - u^h\right) \\ &= \left(\frac{d}{dt} (u^h - Pu), Pu - u^h\right). \end{aligned}$$

Nun gilt

$$\begin{aligned} \left(\frac{d}{dt}(u^h - Pu), Pu - u^h \right) &= \int_a^b -\frac{d}{dt} \left(\sum \gamma_i(t) \varphi_i(x) \right) \cdot \left(\sum \gamma_j(t) \varphi_j(x) \right) dx \\ &= -\frac{1}{2} \frac{d}{dt} \|Pu - u^h\|_{2,0}^2 \\ &= -\|Pu - u^h\|_{2,0} \frac{d}{dt} \|Pu - u^h\|_{2,0}. \end{aligned}$$

Somit wegen

$$\begin{aligned} [Pu - u^h, Pu - u^h] &\geq \lambda \|Pu - u^h\|_{2,0}^2 \\ \|Pu - u^h\|_{2,0} \frac{d}{dt} \|Pu - u^h\|_{2,0} + \lambda \|Pu - u^h\|_{2,0}^2 &\leq \left\| \frac{d}{dt} s \right\|_{2,0} \|Pu - u^h\|_{2,0} \end{aligned}$$

d.h.

$$\frac{d}{dt} \|Pu - u^h\|_{2,0} + \lambda \|Pu - u^h\|_{2,0} \leq \left\| \frac{d}{dt} s \right\|_{2,0}.$$

Multiplikation mit $e^{\lambda\tau}$ und Integration von 0 bis t ergibt

$$\begin{aligned} \int_0^t e^{\lambda\tau} \frac{d}{d\tau} \|(Pu - u^h)(\cdot, \tau)\|_{2,0} d\tau + \lambda \int_0^t e^{\lambda\tau} \|(Pu - u^h)(\cdot, \tau)\|_{2,0} d\tau &\leq \\ &\int_0^t e^{\lambda\tau} \left\| \frac{d}{d\tau} s(\cdot, \tau) \right\|_{2,0} d\tau \\ \int_0^t \frac{d}{d\tau} \left(e^{\lambda\tau} \|(Pu - u^h)(\cdot, \tau)\|_{2,0} \right) d\tau &\leq \int_0^t e^{\lambda\tau} \left\| \frac{d}{d\tau} s(\cdot, \tau) \right\|_{2,0} d\tau \\ e^{\lambda t} \|(Pu - u^h)(\cdot, t)\|_{2,0} - \|(Pu - u^h)(\cdot, 0)\|_{2,0} &\leq \int_0^t e^{+\lambda\tau} \left\| \frac{d}{d\tau} s(\cdot, \tau) \right\|_{2,0} d\tau \\ \|(Pu - u^h)(\cdot, t)\|_{2,0} &\leq e^{-\lambda t} \|(Pu - u^h)(\cdot, 0)\|_{2,0} + \int_0^t e^{\lambda(\tau-t)} \left\| \frac{d}{d\tau} s(\cdot, \tau) \right\|_{2,0} d\tau. \end{aligned}$$

Aber

$$\varepsilon(x, t) = u(x, t) - u^h(x, t; h) = (u - Pu)(x, t) + Pu(x, t) - u^h(x, t; h)$$

d.h.

$$\begin{aligned} \|\varepsilon(\cdot, t)\|_{2,0} &\leq \|(u - Pu)(\cdot, t)\|_{2,0} + e^{-\lambda t} \|(Pu - u^h)(\cdot, 0)\|_{2,0} + \\ &\int_0^t e^{\lambda(\tau-t)} \left\| \frac{d}{d\tau} (u - Pu)(\cdot, \tau) \right\|_{2,0} d\tau. \end{aligned}$$

□

Bemerkung 2.4 Dieser Beweis funktioniert auch im räumlich höherdimensionalen Fall wörtlich, unter Beachtung von

$$[\cdot, \cdot] \geq \lambda \|\cdot\|_{2,0}^2.$$

Aus Aussagen über die Approximationsgüte des Unterraumes S_h für \mathcal{H} (d.h. $(u - Pu)(x, t)$) und dem Anfangsfehler $(Pu - u^h)(x, 0)$ gelangt man so unmittelbar zu Fehlerabschätzungen für die Galerkin-Approximation.

Man kann nun wiederum die Theorie der finite Element-Approximation benutzen, um zu konkreten Fehlerabschätzungen bei gegebener Form von S_h zu gelangen:

Im Folgenden sei S_h der Raum der stetigen, stückweise linearen Funktionen v auf einer Zerlegung

$$a = x_0 < x_0 + h < x_0 + 2h \cdots < b = a + (n + 1)h$$

mit $v(a) = v(b) = 0$.

Satz 2.7 Für u , die Lösung der RAWA, gelte $u \in H_0^4([a, b[\times]0, T])$. Dann ergibt sich

$$\|\varepsilon(\cdot, t)\|_{2,0} \leq K(T)h^2 \quad \text{für } 0 \leq t \leq T,$$

wobei ε der globale Diskretisierungsfehler der Galerkinapproximation mit stückweise linearen Ansatzfunktionen auf einer äquidistanten Zerlegung von $[a, b]$ und $K(T)$ eine geeignete Funktion unabhängig von h ist.

Beweis: Im Folgenden beachte man, daß wegen des Sobolev'schen Einbettungssatzes

$$u \in C^2([a, b[\times]0, T])$$

gilt. Insbesondere sind die gemischten zweiten partiellen Ableitungen gleich. Wir zeigen zunächst

$$\|u(\cdot, t) - Pu(\cdot, t)\|_{2,0} \leq h^2 K_1 \left\| \frac{\partial^2}{(\partial x)^2} u(\cdot, t) \right\|_{2,0}.$$

Sei dazu $z = z(\cdot, t)$ die Lösung der Zweipunkttrandwertaufgabe (t fest)

$$Lz = u - Pu, \quad z(a, t) = z(b, t) = 0.$$

Dann gilt: z ist schwache Lösung, somit

$$[z, v] = (u - Pu, v) \quad \forall v \in \mathcal{H}.$$

Mit $v \stackrel{def}{=} u - Pu$ und $v^h \in S_h$ bel. gilt (weil $[v^h, u - Pu] = 0$)

$$\begin{aligned} [z, u - Pu] &= \|u - Pu\|_{2,0}^2 \\ &= [z - v^h, u - Pu]. \end{aligned}$$

Sei nun u_I die stückweise linear Interpolierende zu u . Dann wird

$$[u - u_I, u - u_I]^{1/2} \leq hK_2 \|u_{xx}(\cdot, t)\|_{2,0}.$$

weil

$$\frac{d}{dx}(u - u_I)(x, t) = h \int_0^1 \int_0^1 u_{xx}(x + \xi(\tau - \nu)h, t)(\tau - \nu) d\xi d\tau$$

mit $x = x_i + \nu h$, $0 \leq \nu \leq 1$ als Definition für ν . Dies ergibt sich wegen der Stetigkeit von $u_{xx}(\cdot, \cdot)$ aufgrund der Voraussetzungen aus der Abschätzung des Interpolationsfehlers unmittelbar. Andererseits ist

$$\begin{aligned} [u - u_I, u - u_I] &= [u - Pu + \overbrace{Pu - u_I}^{\in S_h}, u - Pu + Pu - u_I] \\ &= [u - Pu, u - Pu] + [Pu - u_I, Pu - u_I], \end{aligned}$$

weil $[u - Pu, v^h] = 0$ für $v^h \in S_h$ und $v^h \stackrel{def}{=} Pu - u_I \in S_h$.

Also ist

$$[u - Pu, u - Pu] \leq [u - u_I, u - u_I]$$

d.h.

$$[u - Pu, u - Pu]^{1/2} \leq hK_2 \|u_{xx}(\cdot, t)\|_{2,0}.$$

Ebenso gilt natürlich für die oben definierte Hilfsfunktion z

$$[z - Pz, z - Pz]^{1/2} \leq hK_2 \|z_{xx}(\cdot, t)\|_{2,0}.$$

Aber (man setze oben $v^h \stackrel{def}{=} Pz$) wegen der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} \|u - Pu\|_{2,0}^2 &= [z - Pz, u - Pu] \leq [z - Pz, z - Pz]^{1/2} [u - Pu, u - Pu]^{1/2} \\ &\leq h^2 K_2^2 \|u_{xx}(\cdot, t)\|_{2,0} \|z_{xx}(\cdot, t)\|_{2,0} \\ &\leq h^2 K_2^2 \|u_{xx}(\cdot, t)\|_{2,0} \cdot C \|(u - Pu)(\cdot, t)\|_{2,0}. \end{aligned}$$

Die letzte Abschätzung ergibt sich folgendermassen: Zunächst ist aufgrund des Lax-Milgram Lemmas $z(\cdot, t) \in H_0^1([a, b])$ und

$$\|z\|_{2,1} \leq \|u - Pu\|_{2,0}/\lambda$$

Weil aber $(u - Pu)(\cdot, t) \in H_0^1([a, b])$ gilt auch noch z'' in $L_2[a, b]$ und

$$\|z''\|_{L_2} \leq \frac{1}{\alpha} \left(\frac{c^* + a^{**}}{\lambda} + 1 \right) \|u - Pu\|_{2,0}$$

wo

$$c^* = \max\{c(x) : x \in [a, b]\} \text{ und } a^{**} = \max\{a'(x) : x \in [a, b]\}.$$

Damit ist die erste Teilbehauptung gezeigt.

Die Projektion $Pu = \sum_{j=1}^n \beta_j(t) \varphi_j(x)$ ist definiert durch das lineare Gleichungssystem

$$\left([u, \varphi_j] \right)_{j=1}^n = \left([\varphi_i, \varphi_j] \right) \begin{pmatrix} \beta_1(t) \\ \vdots \\ \beta_n(t) \end{pmatrix}.$$

Nach Voraussetzung existiert $u_{xt} = u_{tx}$, d.h. man kann die β_j differenzieren und auf der linken Seite die t -Ableitung in das Skalarprodukt hineinziehen:

$$\frac{\partial}{\partial t}(Pu) = P\left(\frac{\partial}{\partial t}u\right)$$

d.h. mit einer Wiederholung der obigen Abschätzung für $u_t - Pu_t$ haben wir

$$\left\| \frac{\partial}{\partial t}(u - Pu)(\cdot, t) \right\| \leq h^2 K_2^2 C \|u_{txx}(\cdot, t)\|_{2,0}.$$

Nach Konstruktion von $a(0)$ ist $u^h(x, 0)$ die Projektion von $u(x, 0)$ auf S_h im L_2 -Skalarprodukt. Somit gilt mit geeignetem K_3

$$\|u(\cdot, 0) - u^h(\cdot, 0)\|_{2,0} \leq \|u(\cdot, 0) - u_I(\cdot, 0)\|_{2,0} \leq h^2 K_3 \|u_{xx}(\cdot, 0)\|_{2,0}.$$

Für $(Pu)(x, 0) - u^h(x, 0)$ ergibt sich somit

$$\begin{aligned} \|Pu(\cdot, 0) - u^h(\cdot, 0)\|_{2,0} &\leq \|Pu(\cdot, 0) - u(\cdot, 0)\|_{2,0} + \|u(\cdot, 0) - u^h(\cdot, 0)\|_{2,0} \\ &\leq \|Pu(\cdot, 0) - u(\cdot, 0)\|_{2,0} + \|u(\cdot, 0) - u_I(\cdot, 0)\|_{2,0} \\ &\leq h^2 (K_3 + CK_2^2) \|u_{xx}(\cdot, 0)\|_{2,0}. \end{aligned}$$

Somit folgt aus der Abschätzung in Satz 2.6

$$\begin{aligned} \|\varepsilon(\cdot, t)\|_{2,0} &\leq \|(u - Pu)(\cdot, t)\|_{2,0} + e^{-\lambda t} \|(Pu - u^h)(\cdot, 0)\|_{2,0} \\ &\quad + \int_0^t e^{\lambda(\tau-t)} \left\| \frac{d}{d\tau}(u - Pu)(\cdot, \tau) \right\|_{2,0} d\tau \\ &\leq h^2 K_2^2 C \|u_{xx}(\cdot, t)\|_{2,0} + e^{-\lambda t} h^2 (K_3 + CK_2^2) \|u_{xx}(\cdot, 0)\|_{2,0} \\ &\quad + h^2 K_2^2 C \int_0^t e^{\lambda(\tau-t)} \|u_{txx}(\cdot, \tau)\|_{2,0} d\tau \\ &= \mathcal{O}(h^2). \end{aligned}$$

Aus der Darstellung des Fehlers erkennt man überdies, daß die Anfangsfehler weggedämpft werden. Normalerweise nimmt auch $\|u_{xx}(\cdot, t)\|_{2,0}$ und $\|u_{txx}(\cdot, t)\|_{2,0}$ mit wachsendem t ab. \square

Bemerkung 2.5 Für einen kubischen Spline-Ansatz mit Gitterweite h erhält man entsprechend eine Fehlerabschätzung $\mathcal{O}(h^4)$. \square

Bemerkung 2.6 Auch obiger Beweis lässt sich auf den räumlich mehrdimensionalen Fall übertragen, allerdings benötigt man wegen der Sobolevschen Einbettungssätze im dreidimensionalen Fall die höhere Regularitätsannahme $u \in H_0^5(\Omega \times]0, T[)$.

Es entsteht so ein gewöhnliches DGL-System

$$\begin{aligned} -M\dot{a}(t) &= Ka(t) + F(t) \\ Ma(0) &= c \end{aligned}$$

mit großen Bandmatrizen M und K . Es ist selbstverständlich **nicht** sinnvoll, dieses DGL-System in die explizite Gestalt $\dot{a}(t) = M^{-1}K(a) + M^{-1}F(t)$ zu überführen, da $M^{-1}K$ voll besetzt wäre. Weil das DGL-System steif ist, kommen nur Einschritt- oder Mehrschrittverfahren in Frage, die auf der ganzen negativen reellen Achse absolut stabil sind. Im

Zusammenhang mit linearen finiten Elementen bietet sich das 2-Schritt BDF-Verfahren (Gear) an. Das Crank-Nicholson-Verfahren, das hier

$$-(M + \frac{\Delta t}{2}K)a_{n+1} = (-M + \frac{\Delta t}{2}K)a_n + \frac{\Delta t}{2}(F(t_{n+1}) + F(t_n))$$

lautet, ist nicht so beliebt wegen des schwach oszillierenden Verhaltens der diskretisierten Lösung.

Es gibt jedoch auch Einschrittverfahren, die von höherer als erster Ordnung konsistent und auf der negativen reellen Achse absolut stabil sind, und auch die zu den höheren Eigenwerten von $M^{-1}K$ gehörenden Lösungsanteile schnell dämpfen. Diese Verfahren leiten sich aus rationalen Approximationen der Exponentialfunktion her mit Nennergrad $>$ Zählergrad.

Es sind dies einmal die sogenannten Padé-Approximationen, die definiert sind durch

$$e^{-x} = \frac{P_n(x)}{Q_m(x)} + \mathcal{O}(x^{m+n+1}), \quad P_n \in \Pi_n, \quad Q_m \in \Pi_m$$

mit $m > n$, z.B.

$$\frac{1}{1+x+x^2/2} \quad \text{und} \quad \frac{1-x/3}{1+\frac{2}{3}x+\frac{1}{6}x^2}$$

(das Crank-Nicholson-Verfahren gehört zur Padé-Approximation mit $n = m = 1$ $(1 - x/2)/(1 + x/2)$) und die zuerst von Norsett benutzte Approximation

$$e^{-x} = \frac{P_{n-1}(x)}{(1+\beta x)^n} + \mathcal{O}(x^{n+1}), \quad \beta > 0$$

z.B. mit $n = 2$: $\beta = 1 + \frac{1}{2}\sqrt{2}$, $P_1(x) = 1 + (1 + \sqrt{2})x$.

Auch die Rosenbrock-Wanner-Formeln (siehe Höhere Numerische Mathematik I) beruhen auf einer Norsett-Approximation.

Der Ansatz von Norsett hat den Vorteil, daß man pro Zeitschritt mehrere lineare Gleichungssysteme mit der gleichen Matrix lösen muß, statt mehrerer verschiedener Matrizen bei den Padé-Approximationen (wenn man den Nenner in Linearfaktoren zerlegt hat).

Die Approximation

$$e^{-x} = \frac{1 + (1 + \sqrt{2})x}{(1 + (1 + \frac{1}{2}\sqrt{2})x)^2} + \mathcal{O}(x^3)$$

wird benutzt zur Approximation von $e^{-tM^{-1}K}$ in der Lösungsdarstellung

$$a(t) = e^{-tM^{-1}K} \left(M^{-1}c - \int_0^t e^{\tau M^{-1}K} M^{-1}F(\tau) d\tau \right)$$

mit den Entsprechungen

$$0 \hat{=} t_i, \quad t \hat{=} h, \quad a(h) \hat{=} a_{i+1}, \quad M^{-1}c \hat{=} a_i.$$

Anwendung der Trapezregel auf das Integral und Einsetzen der Approximationen liefert die Näherungsformel

$$\tilde{a}_{i+1} = (I + h(1 + 1/\sqrt{2})M^{-1}K)^{-2}(I + h(1 + \sqrt{2})M^{-1}K)(\tilde{a}_i - \frac{h}{2}M^{-1}F(t_i)) - \frac{h}{2}M^{-1}F(t_{i+1})$$

mit der Konsistenzordnung h^2 . \tilde{a}_{i+1} ist Näherung für $a(t_{i+1})$.

Gerechnet wird dies zweckmäßig in den Schritten: $i = 0, 1, 2, \dots$

$$\begin{aligned} Mw_i &= F_{i+1} \\ \left(M + (1 + 1/\sqrt{2})h K\right)u_i &= M\tilde{a}_i - \frac{h}{2}F_i \\ \left(M + (1 + 1/\sqrt{2})h K\right)\tilde{a}_{i+1} &= \left(M + (1 + \sqrt{2})h K\right)u_i - \frac{h}{2}(M + (1 + 1/\sqrt{2})h K)w_i \end{aligned}$$

mit $M\tilde{a}_0 = c$, d.h. man hat ausschließlich Gleichungssysteme mit der Koeffizientenmatrix $M + (1 + 1/\sqrt{2})h K$ (symm. pos. def.) bzw. M zu lösen.

Bemerkung 2.7 Die gewöhnliche Differentialgleichung, die wir hier erhalten haben, ist nicht in der expliziten Standardform

$$\dot{a} = G(t, a)$$

wegen des Auftretens der Massenmatrix M auf der linken Seite. Dies ist kein Problem, wenn man implizite Integratoren benutzt, weil man in diesem Fall ohnehin pro Zeitschritt lineare Gleichungssysteme mit dünn besetzten (in der Regel Band-)Matrizen zu lösen hat. Wenn man sich mit linearen stetigen finiten Elementen begnügt, kann man das Auftreten von M durch die sogenannte ‘‘lumping’’-Technik umgehen. Hierbei wird die Massenmatrix M durch eine Diagonalmatrix

$$\tilde{M} \stackrel{\text{def}}{=} \text{diag}\left(\sum_i M_{j,i}\right)$$

ersetzt. Dies entspricht der Auswertung der Integrale (φ_i, φ_j) durch eine Quadraturformel der Ordnung zwei, die die asymptotische Konvergenzordnung nicht verschlechtert. Diese Quadraturformel ist im räumlich eindimensionalen Fall

$$\int_a^b f(x)dx = h \sum_{j=1}^n f(x_j)$$

also wegen $f(a) = f(b) = 0$ für Funktionen aus $\mathcal{K}_0^1([a, b])$ gerade die Trapezregel. Es gilt nämlich offensichtlich

$$\varphi_i(x_j)\varphi_k(x_j) = 0 \quad \forall j \text{ falls } i \neq k$$

Bemerkung 2.8 Alle Überlegungen und Aussagen dieses Abschnittes lassen sich problemlos auf den räumlich mehrdimensionalen Fall übertragen. Man vergleiche die Entsprechungen im statischen elliptischen Fall, wie sie in HNMI dargestellt wurden. Solange man nur lineare Elemente benutzt, kann man auch hier (wörtlich) die Lumping-Technik anwenden.

Bemerkung 2.9 *Mit wachsender Feinheit der räumlichen Diskretisierung wächst auch die Steifheit der gewöhnlichen DGL über alle Grenzen. Es ist deshalb nicht klar, wie die Ordnungsaussage für den Zeitintegrator zu bewerten ist und tatsächlich kennt man das Phänomen der sogenannten Ordnungsreduktion bei den Zeitintegratoren. Massgeblich ist hier nicht die asymptotische Konsistenzordnung, sondern die sogenannte B-Konvergenzordnung. Diese beträgt z.B. bei den auf der Gaussquadratur beruhenden impliziten Runge-Kutta-Verfahren mit m Stufen nicht $2m$, sondern nur m . Untersuchungen dieser Art sind Gegenstand aktueller Forschung*

Bemerkung 2.10 *Es ist nicht zwingend notwendig, zur Lösung des Zeitintegrationsproblems einen stetigen Zeitintegrator anzuwenden. Man kann auch daran denken, die schwache Gleichung selbst auch in der Zeit im schwachen Sinne zu lösen. Hierbei ist es nicht einmal notwendig, stetige Ansatzfunktionen zu benutzen (oben haben wir ja sogar glatte Koeffizientenfunktionen $a(t)$ eingeführt.) Man gelangt dann zur sogenannten diskontinuierlichen Galerkinmethode. Hierbei wird die Lösung u auf jedem Zeitintervall durch ein Polynom eines festen Grades q mit Werten im Sobolevraum \mathcal{H} dargestellt, es wird aber keine Stetigkeit in t an den Gitterpunkten des Zeitgitters verlangt. Das Zeitgitter sei*

$$0 = t_0 \leq t_1 \dots \leq t_M = T .$$

Sei

$$W_{h,t} = \{w : w|_{[t_i, t_{i+1}]}(x, t) = \sum_{j=0}^q t^j v_{i,j}(x) \text{ mit } v_{i,j} \in \mathcal{S}_h \forall i, j\}, \quad i = 0, \dots, M-1 .$$

Dann lautet das Problem nun : Gesucht $u^h \in W_{h,t}$ mit

$$\int_0^T \left(\left(\frac{d}{dt} u^h, w \right) + a(u^h, w) \right) dt + \sum_{i=1}^{M-1} (u^h(\cdot, t_i + 0) - u^h(\cdot, t_i - 0), w(\cdot, t_i + 0)) + (u^h(\cdot, 0+) - u_0(\cdot), w(\cdot, 0+)) = \int_0^T \langle f(\cdot, t), w \rangle dt \quad \forall w \in W_{h,t} .$$

Die Ableitung nach t ist hier natürlich eine schwache Ableitung und das Integral entsprechend auszuwerten. Da die Darstellungen von w auf den einzelnen Zeitintervallen voneinander unabhängig sind, kann man wieder zu einer schrittweisen Integration in der Zeit gelangen. Setzt man

$$U^0(t_0) \stackrel{\text{def}}{=} u_0(\cdot)$$

und

$$U^m \stackrel{\text{def}}{=} u|_{[t_{m-1}, t_m]}^h, \quad m = 1, \dots,$$

dann erhält man für $m = 1, \dots, M$

$$\int_{t_{m-1}}^{t_m} \left\{ \left(\frac{d}{dt} U^m, v \right) + a(U^m, v) \right\} dt + (U^m(t_{m-1} + 0) - U^{m-1}(t_{m-1}), v) = \int_{t_{m-1}}^{t_m} \langle f, v \rangle dt$$

Mit einem Basisansatz für $W_{h,t}$ gelangt man dann schliesslich zum einem Blockgleichungssystem für die Koeffizienten $v_{i,0}, \dots, v_{i,q}$ des Ansatzes. Wegen der schwierigeren Zeita-daption bietet dieser Ansatz jedoch keine entscheidenden Vorteile. Details siehe in der Spezialliteratur.

2.5 Die Rothe-Methode

Wir gehen wieder aus von der schwachen Form der Gleichung

$$\frac{d}{dt}(u(\cdot, t), v) + a(u(\cdot, t), v) = \langle f(\cdot, t), v \rangle \quad \forall v \in V.$$

Als Funktionenraum V werden wir z.B. wieder $V = H_0^1(\Omega)$ haben. Bei der Rothe-Methode wird diese parabolische Randanfangswertaufgabe mittels Zeitdiskretisierung durch eine Schar elliptischer Randwertaufgaben ersetzt und die Lösung dadurch approximiert. Es wird ein Zeitgitter

$$t_0 = 0 < t_1 < \dots < T$$

erzeugt. Für $t = t_0$ ist die Lösung aus dem Anfangswert bekannt. Sei nun die Lösung $u(x, t)$ zum Zeitpunkt $t = t_j$ bereits approximiert durch eine Funktion $u_j^h(x)$ mit $x \in \Omega$. Nun wird die Näherung $u_{j+1}^h(x)$ für $u(x, t_{j+1})$ definiert als Lösung der elliptischen Randwertaufgabe

$$\left(\frac{u_{j+1}^h - u_j^h}{t_{j+1} - t_j}, v \right) + a(u_{j+1}^h, v) = \langle f(t_{j+1}), v \rangle \quad \forall v \in V.$$

Diese Diskretisierung entspricht dem impliziten Euler-Verfahren. Natürlich kann man auch bessere Formeln für die Approximation der Zeitableitung benutzen. Man beachte, daß die Randbedingungen in die Definition von V eingearbeitet sind. Dies ist nun eine elliptische Randwertaufgabe zur Berechnung von u_{j+1}^h , die mit den bereits besprochenen Methoden behandelbar ist. Die Vorgehensweise heißt in Analogie zur "vertikalen Linienmethode" auch "horizontale Linienmethode". Es ist zu beachten, daß hier u_j^h die Rolle einer Inhomogenität spielt und bei der Umschreibung in ein elliptisches Problem in Standardform der (kleine) Faktor $t_{j+1} - t_j$ als Multiplikator bei der Bilinearform $a(\cdot, \cdot)$ auftritt, was zu weiteren numerischen Problemen führen kann. Wegen der vorausgesetzten Koerzivität von a ist klar, daß das Verfahren wohldefiniert ist. Mit Hilfe der Funktionen $u_j^h(x)$ kann man dann leicht eine auf ganz G definierte Approximation an u erhalten durch lineare Interpolation in der Zeit:

$$u^h(x, t) \stackrel{\text{def}}{=} \sum_{i=0}^M \varphi_i(t) u_i^h(x)$$

wobei φ_i die Basisfunktion der stetigen stückweise linearen Interpolation zum Knoten t_i auf dem Zeitgitter ist. Bei der Lösung des elliptischen Randwertproblems für eine Zeitschicht hat man nun freie Hand in der Wahl eines geeigneten Raunggitters bzw. räumlicher Adaption, was ein technischer Vorteil ist. Auf der Basis der bereits bekannten Abschätzungen kann man dann leicht eine Gesamtfehlerabschätzung für die Lösung der diskretisierten Randwertaufgabe erhalten, wenn man eine Aussage für $\|u - u^h\|$ mit der oben definierten kontinuierlichen Näherungslösung hat. Dazu erhält man für die obige Diskretisierung der Zeit nach "Euler implizit" zunächst

$$\|u - u^h\|_{W_2^1(0, T; V, H)} = \mathcal{O}(\sup |t_{j+1} - t_j|)$$

Ein Beweis kann auf der apriori Abschätzung aus Abschnitt 2.1 aufgebaut werden.

Kapitel 3

Die Stabilitätstheorie von Lax und Richtmyer ERG

Quelle für das Folgende ist [12], [4] und [15]. Im folgenden Kapitel beschäftigen wir uns mit einer allgemeinen Konvergenztheorie für Diskretisierungen von Anfangs- und Randanfangswertaufgaben, die in ihrer wesentlichen Struktur ganz der Dahlquist'schen Theorie für gewöhnliche DGLen entspricht. Während wir aber dort allgemeine nichtlineare DGLen behandeln konnten, beschränken wir uns hier auf lineare DGLen mit konstanten Koeffizienten. Auch für partielle DGLen mit speziellen Nichtlinearitäten und mit variablen Koeffizienten sind viele Resultate bekannt (vgl. etwa bei Ansorge 1978 [1] und Ansorge & Hass 1970 [2]). Wir werden darauf aber nur an Hand spezieller Beispiele eingehen.

Im Gegensatz zur Theorie von Dahlquist, bei der **punktweise** Fehlerabschätzungen für die diskrete Lösung (η_i) das Ziel sind, werden in dieser Theorie die diskreten Näherungen als Funktionen im gleichen Funktionenraum (einem Banachraum) eingebettet, in dem auch die wahre Lösung der DGL gesucht wird. Deshalb müssen wir hier zunächst einige der Hilfsmittel aus der Funktionalanalysis bereitstellen, die wir für die Theorie benötigen.

3.1 Einige funktionalanalytische Grundlagen

Definition 3.1 Es sei B ein Vektorraum über dem Körper K (\mathbb{R} bzw. \mathbb{C}). Auf B sei eine Norm definiert, d.h.

$$\|\cdot\| : B \rightarrow \mathbb{R}_+ \text{ mit } \begin{cases} (i) & \|a\| = 0 \Leftrightarrow a = 0, \\ (ii) & \|\lambda a\| = |\lambda| \|a\| \quad (\lambda \in K, a \in B), \\ (iii) & \|a + b\| \leq \|a\| + \|b\| \end{cases}$$

$(B, \|\cdot\|)$ heißt **Banachraum**, falls der Häufungswert jeder Cauchyfolge aus B selbst in B liegt, d.h. der lineare Raum B in der durch $\|\cdot\|$ induzierten Topologie vollständig ist. \square

Ob ein linearer Raum ein Banachraum ist, hängt wesentlich von der gewählten Norm ab. So ist z.B. mit $I = [a, b]$

$$(C^0(I, \mathbb{R}), \|\cdot\|_\infty) \quad \text{ein Banachraum}$$

aber

$$(C^0(I, \mathbb{R}), \|\cdot\|_2) \quad \text{kein Banachraum.}$$

(Betrachte dazu etwa das Beispiel $I = [0, 1]$,

$$f_\nu(x) = \begin{cases} (2x)^\nu & 0 \leq x \leq \frac{1}{2} \\ 1 & \frac{1}{2} \leq x \leq 1 \end{cases}$$

$\{f_\nu\}$ ist eine Cauchy-Folge in $\|\cdot\|_2$, aber die Grenzfunktion ist nicht stetig bei $x = \frac{1}{2}$)

Definition 3.2 Es sei $(B, \|\cdot\|)$ ein Banachraum. D_1, D_2 seien **Teilmengen** von B . $D_1 \subset D_2$. D_1 heißt **dicht** in D_2 , wenn

$$\forall \varepsilon > 0 \quad \forall d_2 \in D_2 \quad \exists d_1 \in D_1 : \quad \|d_1 - d_2\| < \varepsilon$$

(m.a.W. $\bar{D}_1 = D_2$) \square

In den weiteren Betrachtungen spielen dichte Untervektorräume eines Banachraumes eine wichtige Rolle. Es sollen deshalb hier einige Beispiele solcher Räume folgen. Als Banachraum B betrachten wir

$$B = (C^0(K, \mathbb{C}), \|\cdot\|_\infty), \quad K \subset \mathbb{R}^m \quad \text{kompakt}$$

(Für $b \in B$ ist $\|b\|_\infty = \max_{x \in K} |b(x)|$)

1. $\bigcup_{\nu \in \mathbb{N}_0} \Pi_\nu(K)$ liegt dicht in B (**Approximationssatz von Weierstrass**)
2. $C^k(K, \mathbb{C})$ $k \in \mathbb{N}$ liegt dicht in B
3. $C^\infty(K, \mathbb{C})$ liegt dicht in B
4. $V = \{f \in C^\infty([a, b], \mathbb{C}) : f^{(\nu)}(a) = f^{(\nu)}(b) = 0, \nu \in \mathbb{N}\}$ liegt dicht in B für $K = [a, b]$
5. $C^\infty(\mathbb{R}, \mathbb{C})$ liegt dicht in $C^0(\mathbb{R}, \mathbb{C})$ mit $\|\cdot\|_\infty$ als Norm.

Jetzt sei

$$B = L_2(G, \mathbb{C}^n), \quad G \text{ ein Gebiet des } \mathbb{R}^m$$

Dabei bezeichnet L_2 den Banachraum der über G quadratintegrablen Funktionen:

$$\|b\|_2 = \left(\int_G b(x)^H b(x) dx \right)^{\frac{1}{2}}$$

wobei **zwei solche Funktionen identifiziert werden**, falls

$$\int_G (b_1(x) - b_2(x))^H (b_1(x) - b_2(x)) dx = 0$$

(d.h. $L_2(G, \mathbb{C}^n)$ ist ein Quotientenraum). Dafür gilt:

1. $C_0^\infty(G, \mathbb{C})$ dicht in $L_2(G, \mathbb{C})$
2. $\bigcup_{\nu \in \mathbb{N}_0} \Pi_\nu(\mathbb{C}^m)$ dicht in $L_2(G, \mathbb{C})$ falls G beschränkt.

In den für uns wichtigen Banachräumen kann man also Elemente (Funktionen) durch unendlich oft differenzierbare Funktionen beliebig genau approximieren.

Dies wird für uns wesentlich sein, wenn wir die Lösung einer AWA oder RAWA durch die Lösungen entsprechender Aufgaben mit beliebig glatten Anfangs- und Randwerten annähern wollen. Für die Lösung der Probleme mit "glatten Daten" können wir dann die entsprechenden Existenz- Eindeutigkeits- und Regularitätsaussagen benutzen.

Definition 3.3 Es seien B_1, B_2 Banachräume und D ein Untervektorraum von B_1 . Eine Abbildung

$$A : D \rightarrow B_2$$

heißt **linearer Operator**, falls für alle $a, b \in D$ und alle $\lambda, \mu \in K$ gilt

$$A(\lambda a + \mu b) = \lambda Aa + \mu Ab$$

Ein linearer Operator heißt **beschränkt**, falls

$$\exists \alpha > 0 : \frac{\|A(a)\|^{(2)}}{\|a\|^{(1)}} \leq \alpha \quad \forall a \in D.$$

$$\|A\| \stackrel{\text{def}}{=} \sup_{a \in D} \frac{\|A(a)\|^{(2)}}{\|a\|^{(1)}}$$

heißt die **Norm von** A (durch die Normen $\|\cdot\|^{(i)}$ auf B_i induzierte Operatornorm). Im folgenden sei

$$L(D, B_2) \stackrel{\text{def}}{=} \{A : D \rightarrow B_2 \text{ linear und beschränkt}\}$$

□

Satz 3.1 Es seien B_1, B_2 Banachräume, D dicht in B_1 und $A \in L(D, B_2)$. Dann gibt es genau ein $\hat{A} \in L(B_1, B_2)$ mit $\hat{A}|_D \equiv A$. Es gilt $\|\hat{A}\| = \|A\|$. (Fortsetzungssatz) □

Diesen Satz werden wir so benutzen, daß wir die Zuordnung

$$\left. \begin{array}{l} \text{DGL +} \\ \text{Anfangs- und Randwerte} \end{array} \right\} \longrightarrow \text{Lösung}$$

als linearen Operator auffassen. D spielt dabei die Rolle des Unterraums “DGL + “glatte” Anfangs-/Randwerte”, für die Existenz– Eindeutigkeits– und Regularitätsaussagen vorliegen, B_1 den entsprechenden Raum mit **allen** in Frage kommenden Anfangs– und Randwerten (z.B. $u_t = u_{xx}$ mit L_2 -Anfangswerten)

Definition 3.4 B_1, B_2 seien Banachräume und $M \subset L(B_1, B_2)$. M heißt **gleichmäßig beschränkt**, wenn die Menge $\{\|A\| : A \in M\}$ beschränkt ist. □

Satz 3.2 (Prinzip der gleichmäßigen Beschränktheit):

B_1, B_2 seien Banachräume, $M \subset L(B_1, B_2)$. Es gelte

$$\sup_{A \in M} \|A(b)\| < \infty \quad \forall b \in B_1.$$

Dann ist M gleichmäßig beschränkt. □

Wir werden die Differenzenapproximationen ebenfalls als Funktionen im gleichen Banachraum wie die Lösung u betrachten. Den Differenzenapproximationen sind dann also auch “Lösungsoperatoren” zugeordnet, die aber noch von einem Parameter abhängen, nämlich der Diskretisierungsschrittweite Δt ($\Delta t/\Delta x$ bzw. $\Delta t/(\Delta x)^2$ konstant).

Wir haben also damit eine Menge von Lösungsoperatoren vor uns, auf die wir das Prinzip der gleichmäßigen Beschränktheit anwenden werden.

Die Bedeutung dieses Prinzips für unsere Anwendung macht folgender Satz deutlich

Satz 3.3 B_1, B_2 seien Banachräume, A, A_i ($i \in \mathbb{N}$) $\in L(B_1, B_2)$.

Ferner $\exists A^{-1}, \forall i : \exists A_i^{-1} \in L(B_2, B_1)$. Die Folge der A_i approximiere A punktweise, d.h.

$$\|Ax - A_i x\| \xrightarrow{i \rightarrow \infty} 0 \quad (\forall x \in B_1)$$

Sei $Ax = f$ und x_i definiert durch

$$A_i x_i = f$$

Dann gilt $\forall f \in B_2 : x_i \xrightarrow{i \rightarrow \infty} x \Leftrightarrow \{A_i : i \in \mathbb{N}\}$ gleichmäßig beschränkt. □

Die Approximationseigenschaft in diesem Satz entspricht der Konsistenz der Differenzenapproximationen.

Definition 3.5 B sei ein Banachraum und $[T_1, T_2]$ ein reelles Intervall.

$$u : [T_1, T_2] \rightarrow B$$

heißt Frechét-differenzierbar in $t_0 \in [T_1, T_2]$ falls gilt:

$$\exists a \in B : \lim_{\substack{h \rightarrow 0 \\ t_0+h \in [T_1, T_2]}} \|u(t_0+h) - u(t_0) - ha\|/|h| = 0$$

a heißt dann die Ableitung von u in t_0 : $a \stackrel{\text{def}}{=} u'(t_0)$.

u heißt differenzierbar auf $[T_1, T_2]$, falls es für jedes $t \in [T_1, T_2]$ differenzierbar ist.

u heißt gleichmäßig differenzierbar, falls

$$\forall \varepsilon > 0 \quad \exists \delta(\varepsilon) > 0 : \quad |h| \leq \delta(\varepsilon) \text{ und } t \in [T_1, T_2] \Rightarrow \\ \|u(t+h) - u(t) - hu'(t)\| < \varepsilon|h|$$

u heißt stetig differenzierbar, falls $u'(t)$ stetig ist. □

Satz 3.4 Eine stetig differenzierbare Funktion (auf einem kompakten Intervall) ist gleichmäßig differenzierbar. □

Ferner benötigen wir zur Behandlung inhomogener Anfangswertaufgaben den Begriff des Integrals stetiger Abbildungen eines Intervalls in einen Banachraum:

Definition 3.6 $I = [T_1, T_2]$ sei ein reelles Intervall und B ein Banachraum.

$v \in C^0(I, B)$, $u : I \rightarrow B$ heißt Stammfunktion zu v auf I , falls u differenzierbar auf I ist und $u'(t) = v(t) \quad \forall t \in I$. □

Für die Konstruktion und das Rechnen mit diesen Stammfunktionen gelten wörtlich die bekannten Regeln aus der reellen Analysis, z.B.

$$\frac{d}{dx} w(x) = \int_{T_1}^{T_2} \partial_2 f(t, x) dt \quad \text{falls} \quad w(x) = \int_{T_1}^{T_2} f(t, x) dt \quad x \in [T_3, T_4]$$

und $f, \partial_2 f$ stetig auf $[T_1, T_2] \times [T_3, T_4]$

$$\int_{T_1}^{T_2} C(v(t)) dt = C \int_{T_1}^{T_2} v(t) dt \quad \text{falls} \quad v \in B, C \in L(B, B)$$

usw.

Bei der Behandlung von Differentialgleichungen mit impliziten Differenzenverfahren treten in diesem Zusammenhang lineare Gleichungen in einem Banachraum auf:

$$R(y) = S(x) \quad \text{mit} \quad R, S \in L(B, B).$$

Die Lösbarkeit eines solchen Systems wird gewöhnlich mittels folgenden Satzes bewiesen:

Satz 3.5

a) Es sei $D \in L(B, B)$, $\|D\| < 1$ und $R = I + D$.
Dann existiert $R^{-1} \in L(B, B)$ mit $R \circ R^{-1} = R^{-1} \circ R = I$.

b) Ist $R \in C^0([0, 1], L(B, B))$
(d.h. R hängt stetig von einem Parameter $t \in [0, 1]$ ab) und $R(0) = I$ und ist

$$\{\|R^{-1}(t)\| : t \in [0, 1] \text{ und } \exists R^{-1}(t)\} \text{ beschränkt,}$$

dann existiert $R^{-1}(t)$ für jedes $t \in [0, 1]$. □

Wir kommen nunmehr zur Definition “sachgemäß gestellter Anfangswertaufgaben” und behandeln dann die Begriffe Konsistenz, Stabilität und Konvergenz. Dabei gehen wir von einer abstrakten Formulierung unserer partiellen hyperbolischen oder parabolischen Gleichung als gewöhnliche Differentialgleichung in einem Banachraum aus, wobei der lineare Operator A die Ableitung bezüglich der räumlichen Veränderlichen darstellt. Bei einer Randanfangswertaufgabe werden die Randwerte in der Definition des Raumes (B) bzw. in die Inhomogenität eingearbeitet. Selbstverständlich können wir mit dieser Theorie nur lineare DGLen mit zeitunabhängigen Koeffizienten behandeln.

Definition 3.7 Es seien B ein Banachraum, D_A ein Untervektorraum von B , $T > 0$ und

$$A : D_A \rightarrow B$$

ein linearer Operator. Mit $P(B, T, A)$ bezeichnen wir folgendes Problem:

Gegeben ist $c \in D_A$. Gesucht ist eine differenzierbare Abbildung

$u : [0, T] \rightarrow D_A$ (Lösung) mit

$$u'(t) = A(u(t)) \quad t \in [0, T] \quad (\text{“klassische Lösung”, “eigentliche Lösung”})$$

$$u(0) = c \quad \square$$

Die Menge aller Anfangswerte c , für die die Aufgabe lösbar ist, bildet einen Untervektorraum von D_A . Wir gehen von einer eindeutigen Lösbarkeit aus und haben dann eine Zuordnung

$$c \rightarrow u_c(t_0) \quad \text{für jedes } t_0 \in [0, T].$$

Dies ist nun ebenfalls ein (t -abhängiger) linearer Operator. Dieser "Lösungsoperator" wird zur Definition einer sachgemäß gestellten Aufgabe herangezogen:

Definition 3.8 $P(B, T, A)$ heißt sachgemäß gestellt (engl.: well posed), wenn es $D_{E_0} \subset D_A$ gibt und eine Schar von linearen Operatoren

$$M_0 = \{E_0(t) : t \in [0, T]\} \quad E_0(t) : D_{E_0} \rightarrow D_A$$

mit

- (i) $\bar{D}_{E_0} = B$
- (ii) $\forall c \in D_{E_0} \quad \exists_1 u_c(t)$ Lösung von $P(B, T, A)$, $u_c(t) \stackrel{\text{def}}{=} (E_0(t))(c)$
- (iii) M_0 ist gleichmäßig beschränkt \square

Folgerung: Ist $P(B, T, A)$ sachgemäß gestellt, dann hängt die Lösung u_c von $P(B, T, A)$ lipschitzstetig von c ab. $E_0(t)$ kann auf B in eindeutiger Weise fortgesetzt werden. Die Fortsetzung heie $E(t)$. Die zu beliebigen Anfangswerten $c \in B$ gehrenden "Lsungen" $E(t)c$ bezeichnet man als "verallgemeinerte Lsungen" der zugrundeliegenden DGL.

Definition 3.9 Sei $E(t)$ die Fortsetzung von $E_0(t)$ auf B und

$$M = \{E(t) : t \in [0, T]\}$$

Die Abbildung

$$c \rightarrow E(t)c \quad \text{mit} \quad E(\cdot)(c) : [0, T] \rightarrow B$$

heißt verallgemeinerte Lsung von $P(B, T, A)$ zum Anfangswert c . \square

Satz 3.6 Sei $P(B, T, A)$ sachgemäß gestellt. Dann gilt

(i) $\forall c \in B, \quad \forall \varepsilon > 0 \quad \exists \tilde{c} \in D_{E_0}$:

$$\|E(t)(c) - E_0(t)(\tilde{c})\| < \varepsilon \quad (\forall t \in [0, T])$$

(ii) $E(\cdot)(c) \in C^0([0, T], B) \quad \forall c \in B$

(iii) $E(r+s) = E(r)E(s) \quad \forall r, s: \quad r, s, r+s \in [0, T]$

(iv) $E(t) \cdot (A(c)) = A(E(t)(c)) \quad \forall c \in D_{E_0}$
(Differentiation nach der Raumvariablen und Integration bzgl. t vertauschbar)

(v) Zu $c \in D_{E_0}$ ist $E_0(t)(c)$ sogar **stetig** differenzierbar.

Beweis:

(i) Sei $\varepsilon > 0$ und $\|E(t)\| < L \quad \forall E(t) \in M, \quad \forall t \in [0, T]$

Wähle $\tilde{c} \in D_{E_0}$ mit $\|\tilde{c} - c\| < \varepsilon/L$.

Dann

$$\|E(t)(c) - E_0(t)(\tilde{c})\| = \|E(t)(c) - E(t)(\tilde{c})\| \leq L\|c - \tilde{c}\| < \varepsilon$$

(ii) Sei $t_0 \in [0, T]$ bel. und $\varepsilon > 0$ vorgegeben. Ferner sei $\tilde{c} \in D_{E_0}$ gewählt mit

$$\|E(t)(c) - E_0(t)(\tilde{c})\| < \varepsilon/3 \quad \forall t \in [0, T]$$

$E_0(t)\tilde{c}$ ist differenzierbar und deshalb stetig, d.h.

$$\exists \delta(t_0, \varepsilon) > 0: \quad \|(E_0(s) - E_0(t_0))(\tilde{c})\| < \varepsilon/3$$

$$\text{falls } |s - t_0| < \delta, \quad s \in [0, T]$$

$$\begin{aligned} \|(E(s) - E(t_0))(c)\| &\leq \|E(s)(c) - E_0(s)(\tilde{c})\| + \|E_0(s)(\tilde{c}) - E_0(t_0)(\tilde{c})\| \\ &\quad + \|E_0(t_0)(\tilde{c}) - E(t_0)(c)\| < \varepsilon \end{aligned}$$

(iii) o.B.d.A. $0 < r \leq s$.

Neben $P(B, T, A)$ betrachten wir noch $P(B, r, A), \quad P(B, s, A)$. Offenbar gilt (der Index r und s bezieht sich auf $P(B, r, A)$ bzw. $P(B, s, A)$)

$$D_{E_0} \subset D_{E_0, s} \subset D_{E_0, r} \subset D_A \subset B$$

und $P(B, r, A), \quad P(B, s, A)$ sind sachgemäß gestellt.

$$\begin{aligned} t \in [0, r], \quad c \in D_{E_0} &: & E_{0,r}(t)(c) &= E_0(t)(c) \\ t \in [0, r], \quad c \in B &: & E_r(t)(c) &= E(t)(c) \\ E_0(s)(c) \in D_{E_0, r} & & & \text{(weil } s + r \leq T) \end{aligned}$$

Sei $c \in D_{E_0}$: Dann ist

$$\begin{aligned} E(r)(E(s)(c)) &= E(r)(E_0(s)(c)) = E_{0,r}(r)(E_0(s)(c)) \\ &= E_0(r+s)(c) = E(r+s)(c) \end{aligned}$$

Für $c \in B$ beliebig, benutze $\tilde{c} \in D_{E_0}$ mit $\|c - \tilde{c}\| < \varepsilon$, ε bel. klein.

(iv)

$$\begin{aligned} &\|E(t)(A(c)) - A(E(t)(c))\| \leq \\ &\|E(t)(A(c)) - \frac{1}{h}E(t)(E(h)(c) - c)\| + \|\frac{1}{h}E(t)(E(h)(c) - c) - A(E(t)(c))\| \\ &\leq \frac{1}{|h|}\|E(t)\| \|E(h)(c) - c - hA(c)\| + \\ &+ \frac{1}{|h|}\|E(h)(E(t)(c)) - E(t)(c) - hA(E(t)(c))\| \\ &= \frac{1}{|h|}\|E(t)\| \|u(h) - u(0) - hu'(0)\| + \frac{1}{|h|}\|u(t+h) - u(t) - hu'(t)\| \end{aligned}$$

(für $c \in D_{E_0}$) und $h \rightarrow 0$ liefert die Behauptung.

(v) für $c \in D_{E_0}$ ist

$$u'(t) = A(u(t)) = A(E(t)(c)) = E(t)(A(c))$$

Aber $E(t)(A(c))$ ist verallgemeinerte Lösung und somit $u' \in C^0([0, T], B)$. □

Bemerkung 3.1 Aus (iii) und (iv) folgt

$$\frac{u(t+h) - u(t)}{h} - Au(t) = E(t) \left(\frac{E(h) - I}{h} - A \right) u_0$$

für $u_0 \in D_{E_0}$. □

Beispiel 3.1 Hyperbolische Gleichung

$$\begin{aligned} u_t &= u_x & 0 \leq t \leq T & & x \in \mathbb{R} \\ u(x, 0) &= c(x), & & & c \in C_{2\pi}(\mathbb{R}) \end{aligned}$$

mit der Lösung $u(x, t) = c(x+t)$.

Differenzierbarkeit ist natürlich nur für differenzierbares c gegeben, d.h. wir haben

$$\begin{aligned} A &= \frac{\partial}{\partial x} \\ B &= (C_{2\pi}(\mathbb{R}), \|\cdot\|_\infty) \\ D_A &= B \cap C^1(\mathbb{R}) \end{aligned}$$

offensichtlich ist $\|E_0(t)\| = \|E(t)\| = 1 \quad \forall t \in [0, T] \quad (\forall T)$ somit $P(B, T, A)$ sachgemäß gestellt. Für jedes $c \in B$ ist $u(x, t) = c(x+t)$ somit verallgemeinerte Lösung der AWA. □

Beispiel 3.2 Parabolische Gleichung, RAWA

$$\begin{aligned}u_t &= u_{xx} \\u(0, t) &= u(1, t) = 0 \\u(x, 0) &= c(x)\end{aligned}$$

Nun ist $A = \frac{\partial^2}{\partial x^2}$. Als Banachraum B wollen wir jedoch

$$B = \{u : u \in L_2[0, 1] \text{ mit } u(0) = u(1) = 0\}$$

wählen, d.h. die Randbedingungen sind in B eingearbeitet.

Natürlich muß dann $c(0) = c(1) = 0$ sein. Dagegen ist

$$D_A = C_0^2[0, 1].$$

Die Anfangsvorgabe c kann dargestellt werden als

$$c(x) = \sum_{k=1}^{\infty} \gamma_k \sin(k\pi x)$$

Falls $c \in C_0^2[0, 1]$ dann ist $\gamma_k = \mathcal{O}(\frac{1}{k^2})$ und wenn $\sum_{k=1}^{\infty} k^2 \gamma_k^2 < \infty$ ist umgekehrt $c \in C_0^2[0, 1]$. Die Lösungen von $u_t = u_{xx}$ sind dann klassische Lösungen:

$$E_0(t)c(x) = u(x, t) = \sum_{k=1}^{\infty} e^{-k^2\pi^2 t} \gamma_k \sin(k\pi x)$$

Somit gilt

$$\|E_0(t)\| = \sup_{(\gamma_1, \dots)} \left(\sum_{k=1}^{\infty} \left(e^{-k^2\pi^2 t} \gamma_k \right)^2 \right)^{\frac{1}{2}} / \left(\sum_{k=1}^{\infty} \gamma_k^2 \right)^{\frac{1}{2}} \leq 1 \quad \text{für } t \in [0, T]$$

d.h. die Aufgabe ist sachgemäß gestellt. Verallgemeinerte Lösungen existieren somit für jede auf $[0, 1]$ quadratintegrale Anfangsvorgabe c mit $c(0) = c(1) = 0$. Die unendliche Reihe $\sum_{k=1}^{\infty} e^{-k^2\pi^2 t} \gamma_k \sin(k\pi x)$ braucht aber für solche verallgemeinerten Lösungen gar nicht punktweise zu konvergieren! \square

Beispiel 3.3 Wie Beispiel 3.2, jedoch mit $u_t = -u_{xx}$, d.h. Umkehr der Zeitrichtung: “parabolisch rückwärts in der Zeit”.

An der Darstellung von $E_0(t)$ sieht man, daß das Problem nicht korrekt gestellt ist (man betrachte $t < 0$) $e^{-k^2\pi^2 t} \rightarrow \infty!$ \square

3.2 Abstrakte Differenzenverfahren, Lax-Richtmyer-Theorie

Wir kommen nunmehr zur Definition eines Differenzenverfahrens in einem Banachraum und den Begriffen **Konsistenz**, **Stabilität**, **Konvergenz**:

Definition 3.10 Es seien $P(B, T, A)$ eine sachgemäß gestellte Anfangswertaufgabe und $M = \{E(t) : t \in [0, T]\}$ die Schar der zugehörigen verallgemeinerten Lösungsoperatoren und $0 < h_0 \leq T$.

(i) Eine Schar von linearen, beschränkten Operatoren $C(h) : B \rightarrow B$

$$M_D = \{C(h) : h \in]0, h_0]\}$$

heißt **Differenzenverfahren** zu $P(B, T, A)$, falls $\|C(\cdot)\|$ beschränkt ist auf $[h_1, h_0]$ für alle $h_1 \in]0, h_0]$.

(ii) M_D heißt **konsistent**, wenn $\exists D_C \subset B, \bar{D}_C = B$

$$\frac{1}{|h|} \|(C(h) - E(h))(E(t)(c))\| \xrightarrow{h \rightarrow 0} 0 \quad \text{gleichmäßig auf } [0, T]$$

für alle $c \in D_C$.

(iii) M_D heißt **stabil**, falls

$$\{C(h)^n : h \in]0, h_0], n \in \mathbb{N}, nh \leq T\}$$

gleichmäßig beschränkt ist.

(iv) M_D heißt **konvergent**, wenn

$$\lim_{\substack{j \rightarrow \infty \\ n_j h_j \rightarrow t \\ h_j \rightarrow 0}} \|C(h_j)^{n_j}(c) - E(t)(c)\| = 0 \quad (\forall c \in B, \forall t \in [0, T])$$

□

Der obige Begriff **Differenzenverfahren** ist sehr abstrakt und erfaßt praktisch jede (noch so unsinnige) Konstruktion. Konkreter hat man sich $C(h)$ vorzustellen wie folgt:

Der Operator A in

$$u'(t) = A(u(t)) \quad u = (u_1, \dots, u_n)^T$$

ist ein Differentialoperator bezüglich einer oder mehreren Raumvariablen:

$$A = \sum_{|\alpha| \leq m} P_\alpha(x) D^\alpha \quad m = \text{Ordnung von } A$$

mit

$$\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d \quad |\alpha| = \sum_{i=1}^d \alpha_i$$

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d}$$

$$P_\alpha(x) \in \mathbb{C}^{n \times n}$$

(A hängt nicht von t ab, sonst gilt die sogenannte Halbgruppeneigenschaft $E(r)E(s) = E(r+s)$ in der Regel nicht.)

A wird nun durch Differenzenquotienten bezüglich der Raumvariablen approximiert. Eine solche Differenzenapproximation baut auf auf Linearkombinationen von Translationsoperatoren bezüglich der Raumvariablen:

$$D(x, h) = \sum_{j \in J} d_j T^j \quad j = (j_1, \dots, j_d) \in \mathbb{Z}^d$$

mit

$$d_j = d_j(x, h) \in \mathbb{C}^{n \times n}$$

$$T^j = T_1^{j_1} \cdots T_d^{j_d}$$

$$T_k^{j_k} u(x_1, \dots, x_d) = u(x_1, \dots, x_{k-1}, x_k + j_k \Delta x_k, \dots, x_{k+1}, \dots, x_d)$$

$$\Delta x_k = g_k(h)$$

(d.h. man hat eine **feste Kopplung** zwischen der Schrittweite h in t -Richtung und den Diskretisierungsschrittweiten Δx_k in den Raumrichtungen.)

Mit dieser Definition von D hat man nun Differenzenoperationen nicht nur auf einem Gitter, wie wir sie früher betrachtet haben, sondern für beliebige Elemente von B .

Für $\nu = 0, 1, \dots, \lfloor \frac{T}{h} \rfloor$ betrachtet man nun die Approximation

$$D_1(x, h)u^{\nu+1} = D_0(x, h)u^\nu$$

(für $D_1 \neq I$ also ein implizites Differenzenverfahren) mit invertierbarem D_1 , also letztlich mit u^ν als Näherung für $u(\nu h)$

$$u^{\nu+1} = (D_1(x, h))^{-1}(D_0(x, h)u^\nu),$$

während

$$u((\nu+1)h) = E(h)(u(\nu h))$$

Der Zusammenhang mit A wird hergestellt durch

$$\frac{u(\nu h + h) - u(\nu h)}{h} - Au(\nu h) = \frac{1}{h}(E(h) - I - hA)(u(\nu h))$$

$$\frac{u^{\nu+1} - u^\nu}{h} = \frac{D_1^{-1} \circ D_0 - I}{h} u^\nu$$

Wir werden also erwarten, daß

$$\frac{1}{|h|}(D_1^{-1} \circ D_0 - (E(h)))(c) \rightarrow 0 \quad \text{für } h \rightarrow 0 \quad (c \in B)$$

Der lineare Operator $C(h)$ wird somit

$$C(h) = D_1^{-1} \circ D_0$$

Die Abhängigkeit von x , d.h. von B , tritt dabei formal nicht in Erscheinung, weil bezüglich der Integration in t -Richtung x wie ein "Parameter" wirkt.

Der folgende Satz lautet nun wörtlich wie in der Dahlquist-Theorie der gewöhnlichen Differentialgleichungen:

Satz 3.7 (Äquivalenzsatz von Lax)

$P(B, T, A)$ sei korrekt gestellt und M_D sei konsistent. Dann ist M_D genau dann konvergent, wenn es stabil ist.

Beweis: Aus Konvergenz folgt Stabilität:

Annahme: M_D konvergent, aber nicht stabil. Mit Hilfe des Prinzips der gleichmäßigen Beschränktheit soll ein Widerspruch konstruiert werden. Nach Annahme gibt es eine Folge h_j und eine zugehörige Folge natürlicher Zahlen n_j mit

$$\begin{aligned} n_j h_j &\rightarrow t \in [0, T], \\ \{ \|C(h_j)^{n_j}\| \} &\text{ nicht beschränkt.} \end{aligned}$$

o.B.d.A. sei auch $h_j \rightarrow h$ konvergent.

1. Fall: $h > 0 \Rightarrow n_j = n \quad (j \geq j_0)$

$\Rightarrow h_j \in [h/2, h_0] \quad j \geq j_0$. Aber $\|C(\tau)\|$ beschränkt für $\tau \in [h/2, h_0]$.

Also für $j \geq j_0$

$$\|C(h_j)^{n_j}\| = \|C(h_j)^n\| \leq \|C(h_j)\|^n \leq K^n \quad \text{Widerspruch!}$$

2. Fall: $h = 0$

Nun gilt wegen der Konvergenz

$$\|C(h_j)^{n_j}(u_0) - E(t)(u_0)\| \xrightarrow{j \rightarrow \infty} 0 \quad \forall u_0 \in B$$

Somit $\exists j_0(u_0)$:

$$\|C(h_j)^{n_j}(u_0) - E(t)(u_0)\| < 1 \quad \forall j \geq j_0(u_0)$$

d.h.

$$\|C(h_j)^{n_j}(u_0)\| < 1 + \|E(t)(u_0)\|$$

Setze

$$\beta(u_0) \stackrel{\text{def}}{=} \max_{j \leq j_0(u_0)} \{1 + \|E(t)(u_0)\|, \|C(h_j)^{n_j}(u_0)\|\}$$

Somit gilt

$$\forall u_0 \in B, \quad \forall j \in \mathbb{N} \quad \|C(h_j)^{n_j}(u_0)\| \leq \beta(u_0)$$

Nach dem Prinzip der gleichmäßigen Beschränktheit gilt somit

$$\|C(h_j)^{n_j}\| \leq K \quad \text{mit } K \text{ geeignet. Widerspruch!}$$

Aus Stabilität und Konsistenz folgt Konvergenz:

Sei $c \in D_C$, $\{h_j\}$ eine Nullfolge und n_j eine Folge natürlicher Zahlen mit

$$h_j n_j \rightarrow t \in [0, T], \quad 0 \leq h_j n_j \leq T.$$

Setze

$$\psi_j(c) \stackrel{\text{def}}{=} (C(h_j)^{n_j} - E(t))(c) \quad j \in \mathbb{N}.$$

Zu zeigen ist $\psi_j(c) \rightarrow 0 \quad j \rightarrow \infty$, zunächst für $c \in D_c$, später für $c \in B$. Für $\psi_j(c)$ gilt wegen der Halbgruppeneigenschaft folgende Darstellung

$$\psi_j(c) = \sum_{k=0}^{n_j-1} C(h_j)^k (C(h_j) - E(h_j)) E((n_j - 1 - k)h_j)(c) + E(\delta_j) (E(n_j h_j - \delta_j) - E(t - \delta_j))(c)$$

mit

$$\delta_j = \min\{t, n_j h_j\}$$

Sei $\varepsilon > 0$ beliebig vorgegeben und $\forall j : \|C(h_j)^{n_j}\| \leq K_C$ mit $0 \leq n_j h_j \leq T$. Wegen der vorausgesetzten Konsistenz gilt

$$\exists j_1(\varepsilon, c) : \quad \|(C(h_j) - E(h_j)) \underbrace{E((n_j - 1 - k)h_j)(c)}_{\doteq E(t_{j,k})(c) \text{ mit } c \in D_C}\| \leq \varepsilon h_j \quad \text{für } j \geq j_1(\varepsilon, c)$$

Da $P(B, T, A)$ sachgemäß gestellt ist, gilt

$$\|E(\tau)\| \leq K_E \quad \forall \tau \in [0, T]$$

Wegen Satz 3.6 (ii) gilt

$$\exists j_2(\varepsilon, c) : \quad \|(E(n_j h_j - \delta_j) - E(t - \delta_j))(c)\| < \varepsilon \quad \text{für } j \geq j_2(\varepsilon, c)$$

Damit gilt für $j \geq \max\{j_1(\varepsilon, c), j_2(\varepsilon, c)\}$ und $c \in D_C$

$$\|\psi_j(c)\| \leq n_j K_C \varepsilon h_j + K_E \varepsilon \leq (K_C T + K_E) \varepsilon$$

Sei nun $\tilde{c} \in B$ beliebig. Dann wird mit $c \in D_C : \|c - \tilde{c}\| < \varepsilon$

$$\begin{aligned} \psi_j(\tilde{c}) &= C(h_j)^{n_j}(\tilde{c}) - E(t)(\tilde{c}) \\ &= C(h_j)^{n_j}(c) - E(t)(c) + C(h_j)^{n_j}(\tilde{c} - c) + E(t)(c - \tilde{c}) \end{aligned}$$

d.h.

$$\|\psi_j(\tilde{c})\| \leq (K_C T + K_E) \varepsilon + K_C \varepsilon + K_E \varepsilon \quad \text{für } j \geq \max\{j_1(\varepsilon, c), j_2(\varepsilon, c)\}$$

Da $\varepsilon > 0$ beliebig war, folgt die Behauptung. \square

Bemerkung 3.2 *In obigen Betrachtungen ist nirgends von einer Konvergenzordnung die Rede. Geht man obigen Beweis einmal durch, so sieht man, daß dies auch nicht möglich ist für $n_j h_j \neq t$ wegen des zweiten Termes in der Darstellung von $\psi_j(c)$.*

Falls jedoch

$$n_j h_j = t \quad j \geq j_0$$

und

$$\frac{1}{|h|} \|(C(h) - E(h))E(t)u_0\| = \mathcal{O}(h^p) \quad \forall u_0 \in D_C,$$

dann konvergiert das Verfahren für $u_0 \in D_C$ von der Ordnung h^p .

Man beachte, daß eine Genauigkeitsforderung an die Diskretisierung in den Raumvariablen hierin implizit enthalten ist. Die Kopplungsbedingungen an das Verhältnis t -Schrittweite / x_i -Schrittweite sind implizit enthalten in der Stabilitätsbedingung. \square

Bemerkung 3.3 *Im ersten Teil des Beweises von Satz 3.7 wird die Konsistenz nicht benutzt. Tatsächlich gibt es inkonsistente und dennoch konvergente Differenzenverfahren, die aber natürlich keinerlei praktische Bedeutung haben. Ein Beispiel stammt von Spijker (1967) und ist in Ansorge-Hass Seite 75ff. beschrieben.* \square

Bemerkung 3.4 *Der Äquivalenzsatz läßt sich auch ohne die Halbgruppeneigenschaft beweisen (wichtig für Fälle, in denen A (d.h. die Koeffizienten der partiellen DGL) von t abhängt). Der Beweis wird dann komplizierter, vgl. bei Ansorge-Hass Seite 63ff.* \square

Bemerkung 3.5 *Eine der wesentlichen Voraussetzungen im Beweis von Satz 3.7 war die gleichmäßige Beschränktheit von $E(t)$. In der Literatur über partielle DGLen findet man Existenz- Eindeutigkeits- und Regularitätsaussagen für die Lösung u , in der Regel aber keine Abschätzungen für $\|E(t)\|$. Dann kann der folgende Satz weiterhelfen:* \square

Satz 3.8 Die Aufgabe $P(B, T, A)$ erfülle die Voraussetzungen (i) und (ii) von Definition 3.8 (eindeutige klassische Lösbarkeit auf einem dichten Teilraum D_{E_0} von B). Ferner gebe es eine Schar von Operatoren $C(h) \in L(B, B)$

$$M_D = \{C(h) : h \in]0, h_0]\}$$

mit

$$(i) \lim_{h \rightarrow 0} \frac{1}{|h|} \|C(h)(E_0(t)(u_0)) - E_0(t+h)(u_0)\| \rightarrow 0$$

gleichmäßig auf $[0, T]$ für $u_0 \in D_{E_0}$

(ii) $\{\|C(h)^n\| : h \in]0, h_0], nh \leq T\}$ ist beschränkt.
Dann ist $P(B, T, A)$ sachgemäß gestellt, d.h.

$$\|E_0(\tau)\| \leq K_E \quad \forall \tau \in [0, T]$$

(und damit $\|E(\tau)\| \leq K_E \quad \forall \tau \in [0, T]$).

Beweis: Sei $\|C(h)^n\| \leq K_C \quad \forall h \in]0, h_0], nh \leq T$.

Setze $h = t/m$ mit $m \in \mathbb{N}$ für $t \in [0, T]$, d.h. $\delta = 0$ im Beweis des vorausgegangenen Satzes.

Für $u_0 \in D_{E_0}$, $u_0 \neq 0$ gilt dann (vgl. Darstellung von $\psi_j(c)$ oben)

$$\begin{aligned} \|E_0(\underbrace{mh}_t)(u_0)\| &\leq \|C(h)^m(u_0)\| + \sum_{k=0}^{m-1} \|C(h)^k\| \|C(h) - E_0(h)\| \\ &\quad \cdot \|E_0((m-1-k)h)(u_0)\| \\ &\leq K_C \|u_0\| + mK_C \|C(h)(E_0(\underbrace{(m-1-k)h}_{\stackrel{\text{def}}{=} \tau})(u_0)) \\ &\quad - E_0(\tau+h)(u_0)\| \\ &\leq K_C \|u_0\| + mK_C \|u_0\| h = (1+T)K_C \|u_0\| \end{aligned}$$

falls h so klein ist, daß wegen (i)

$$\|(C(h)E(\tau) - E(\tau+h))(u_0)\| \leq h \|u_0\| \quad \forall \tau, \tau+h \in [0, T]$$

Damit ist

$$\|E_0(\tau)\| \leq (1+T)K_C \quad \forall \tau \in [0, T]$$

□

Beispiel 3.4

$$\begin{aligned} B &= \{u \in C_{2\pi}(\mathbb{R}), \|\cdot\|_\infty\} \\ A &= \frac{\partial^2}{\partial x^2} \end{aligned}$$

$$\begin{aligned} u_t &= u_{xx} & t \in [0, T] \\ u(x, 0) &= u_0(x) & x \in \mathbb{R} \end{aligned}$$

$P(B, T, A)$ ist korrekt gestellt, denn $D_A = C_{2\pi}^2(\mathbb{R})$ ist dicht in $C_{2\pi}(\mathbb{R})$ und mit $D_{E_0} = D_A$ kann man schreiben

$$\begin{aligned} u_0(x) &= \frac{\alpha_0}{2} + \sum_{k=1}^{\infty} \left(\alpha_k \cos(kx) + \beta_k \sin(kx) \right) \\ u(x, t) &= \frac{\alpha_0}{2} + \sum_{k=1}^{\infty} e^{-k^2 t} \left(\alpha_k \cos(kx) + \beta_k \sin(kx) \right) \end{aligned}$$

und ersichtlich ist

$$\|u(\cdot, t)\|_{\infty} \leq \frac{|\alpha_0|}{2} + \sum_{k=1}^{\infty} (|\alpha_k| + |\beta_k|) \leq \frac{|\alpha_0|}{2} + 2C \frac{\pi^2}{6} \stackrel{\text{def}}{=} \beta(u_0)$$

wobei C eine von $\|u_0''(x)\|_{\infty}$ abhängende Konstante ist mit

$$|\alpha_k|, |\beta_k| \leq \frac{C}{k^2} \quad k = 1, 2, \dots$$

Also ist $\{E_0(t)\}$ gleichmäßig beschränkt.

Wir betrachten nun folgende Differenzenapproximation (das explizite Differenzenverfahren) mit $\rho = \Delta t / (\Delta x)^2$

$$u^{\nu+1}(x) = u^{\nu}(x) + \rho(u^{\nu}(x - \Delta x) - 2u^{\nu}(x) + u^{\nu}(x + \Delta x))$$

d.h. ($h \hat{=} \Delta t$)

$$C(h)u(x) = u(x) + \rho(u(x - \Delta x) - 2u(x) + u(x + \Delta x))$$

Wie bereits früher nachgerechnet wurde, ist mit $\rho = \text{const}$ für $u \in C^4(\mathbb{R}) \cap B \stackrel{\text{def}}{=} D_C$

$$\frac{1}{|h|} \underbrace{\left(E(h)u - C(h)u \right)}_{u(x, t+h)} = \mathcal{O}(h)$$

d.h. die Konsistenzbedingung gilt auf D_C .

Nun ist die Stabilität zu prüfen. Sei zunächst $\rho \leq 1/2$ Dann ist

$$|C(h)u(x)| \leq (1 - 2\rho)|u(x)| + \rho|u(x + \Delta x)| + \rho|u(x - \Delta x)| \leq \|u\|_{\infty}$$

d.h. $\|C(h)\|_{\infty} \leq 1$ und somit ist das Verfahren stabil.

Sei jetzt $\rho > 1/2$. Wir konstruieren ein spezielles Beispiel, mit dessen Hilfe wir beweisen, daß $\|C^n(h)\|$ nicht beschränkt sein kann.

Als $u_0(x)$ wählen wir die stückweise linear stetige Interpolierende (Streckenzug) zu den Daten $(i\Delta x, (-1)^{i+1})$ $i \in \mathbb{Z}$ mit $\Delta x = \frac{\pi}{m}$ und $m \in \mathbb{N}$. Es gilt $u_0 \in B$, $\|u_0\|_\infty = 1$. (u_0 besitzt sogar eine konvergente Fourierreihe)

Es gilt

$$u_0(x) = -u_0(x - \Delta x) = -u_0(x + \Delta x)$$

Somit

$$\begin{aligned} u^\nu(x) &= u^{\nu-1}(x) + \rho(u^{\nu-1}(x - \Delta x) - 2u^{\nu-1}(x) + u^{\nu-1}(x + \Delta x)) \\ &= (1 - 4\rho)u^{\nu-1}(x) = (1 - 4\rho)^\nu u_0(x) \quad (u^0 \stackrel{\text{def}}{=} u_0) \end{aligned}$$

und für $\rho > \frac{1}{2}$ somit

$$\|C^n(h)\|_\infty \geq \|C^n(h)u_0\|_\infty = \|u^n\|_\infty = |1 - 4\rho|^n$$

$\Rightarrow M_D$ ist nicht stabil \Rightarrow Verfahren konvergiert nicht. □

Beispiel 3.5 Beispiel 3.4 mit $B = \{C_{2\pi}(\mathbb{R}^d), \|\cdot\|_\infty\}$

$$\begin{aligned} u_t &= \sum_{k=1}^d u_{x_k x_k} \\ u(x, 0) &= u_0(x) \quad (x \in \mathbb{R}^d) \end{aligned}$$

Bei der entsprechend übertragenen expliziten Differenzenapproximation ergibt sich jetzt Stabilität für

$$\rho = \Delta t / (\Delta x_k)^2 \quad k = 1, \dots, d$$

bei $\rho \leq \frac{1}{2d}$, Instabilität für $\rho > \frac{1}{2d}$. □

Beispiel 3.6 ADI für $u_t = \Delta u$ ($d = 2$). Wir greifen das in Abschnitt 2.2 beschriebene Verfahren auf (2.3) und wollen uns nun im folgenden Beispiel damit beschäftigen, dieses Verfahren in die vorliegende Theorie einzupassen.

Als Grundraum wählen wir

$$\begin{aligned} B &= \{u \in C_0^0([0, 1] \times [0, 1]), \|\cdot\|_\infty\} \\ D_A &= \{C_0^2([0, 1] \times [0, 1]), \|\cdot\|_\infty\} \end{aligned}$$

D_A ist dicht in B . Als D_C ergibt sich

$$D_C = \{C_0^6([0, 1] \times [0, 1]), \|\cdot\|_\infty\}$$

(bei der Herleitung der Konsistenzordnung wurde die 6. partielle Ableitung in x (bzw. y -) Richtung benutzt.)

Die Konsistenz ergibt sich unmittelbar aus den bereits früher hergeleiteten Beziehungen. Nun zur Stabilität!

Wenn man x, y als kontinuierliche Variable interpretiert (Einschränkung auf das $(\Delta x, \Delta y)$ -Gitter in $[0, 1] \times [0, 1]$ erfolgt erst beim Rechnen) hat man

$$\begin{aligned} u^{n+\frac{1}{2}}(x, y) &= u^n(x, y) + \frac{\rho}{2} \left(u^{n+\frac{1}{2}}(x - \Delta x, y) - 2u^{n+\frac{1}{2}}(x, y) + u^{n+\frac{1}{2}}(x + \Delta x, y) \right) \\ &\quad + \frac{\rho}{2} \left(u^n(x - \Delta x, y) - 2u^n(x, y) + u^n(x + \Delta x, y) \right) \\ u^{n+1}(x, y) &= u^{n+\frac{1}{2}}(x, y) + \frac{\rho}{2} \left(u^{n+\frac{1}{2}}(x, y - \Delta y) - 2u^{n+\frac{1}{2}}(x, y) + u^{n+\frac{1}{2}}(x, y + \Delta y) \right) \\ &\quad + \frac{\rho}{2} \left(u^{n+1}(x, y - \Delta y) - 2u^{n+1}(x, y) + u^{n+1}(x, y + \Delta y) \right) \end{aligned}$$

und damit für $\rho \leq 1$

$$(1 + \rho) \|u^{n+\frac{1}{2}}\|_\infty \leq (1 - \rho) \|u^n\|_\infty + \rho \|u^{n+\frac{1}{2}}\|_\infty + \rho \|u^n\|_\infty$$

d.h.

$$\|u^{n+\frac{1}{2}}\|_\infty \leq \|u^n\|_\infty$$

und entsprechend

$$\|u^{n+1}\|_\infty \leq \|u^{n+\frac{1}{2}}\|_\infty$$

d.h. $\|C(\Delta t)\|_\infty \leq 1$ und damit Stabilität und Konvergenz in $\|\cdot\|_\infty$.

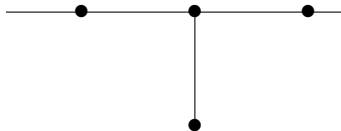
Für $d \geq 2$ erhält man bei dieser Vorgehensweise im allgemeinen Fall die Stabilitätsbedingung $\rho \leq \frac{d}{2(d-1)}$ □

Beispiel 3.7 Vollimplizites Verfahren für das reine Anfangswertproblem einer parabolischen DGL

$$\begin{aligned} \partial_2 u(x, t) &= \partial_1(a(x)\partial_1 u(x, t)) \\ u(x, 0) &= u_0(x) \quad x \in \mathbb{R} \end{aligned}$$

mit $a \in C^\infty(\mathbb{R}, \mathbb{R})$, $a' \in C_0^\infty(\mathbb{R}, \mathbb{R})$ (d.h. kompakter Träger)
 $a(x) > 0 \quad x \in \mathbb{R}$. $B \stackrel{\text{def}}{=} L_2(\mathbb{R}, \mathbb{C})$, $D_A = C_0^2(\mathbb{R}, \mathbb{C})$, $D_C = C_0^4(\mathbb{R}, \mathbb{C})$

Vollimplizites Verfahren:



Definiere mit $T_{\Delta x} =$ Shiftoperator in der Raumvariablen x um Δx :

$$\begin{aligned} H(\Delta x) &\stackrel{\text{def}}{=} a\left(x - \frac{\Delta x}{2}\right)T_{\Delta x}^{-1} - \left(a\left(x + \frac{\Delta x}{2}\right) + a\left(x - \frac{\Delta x}{2}\right)\right)I + a\left(x + \frac{\Delta x}{2}\right)T_{\Delta x} \\ \rho &\stackrel{\text{def}}{=} \frac{h}{(\Delta x)^2} \quad (h \text{ entspricht } \Delta t) \\ C(h) &\stackrel{\text{def}}{=} (I - \rho H(\Delta x))^{-1} \\ u^{\nu+1} &= C(h)u^\nu \quad u^0 = u_0 \end{aligned}$$

(Dies entspricht der Differenzgleichung

$$\frac{1}{\Delta x} \left(a\left(x_i + \frac{\Delta x}{2}\right) \frac{u_{i+1}^{\nu+1} - u_i^{\nu+1}}{\Delta x} - a\left(x_i - \frac{\Delta x}{2}\right) \frac{u_i^{\nu+1} - u_{i-1}^{\nu+1}}{\Delta x} \right) = \frac{u_i^{\nu+1} - u_i^\nu}{\Delta t}$$

$$i \in \mathbb{Z}, \quad \nu = 0, 1, 2, \dots .)$$

Behauptung: Das Verfahren ist für alle $\rho \in \mathbb{R}_+$ stabil und von erster Ordnung konsistent.

Wir zeigen zunächst die Stabilität, indem wir beweisen, daß

$$\|C(h)\| \leq 1.$$

Wir beweisen für $f \in C^0(\mathbb{R}, \mathbb{C}) \cap L_2(\mathbb{R}, \mathbb{C})$, daß

$$\|C(h)f\| \leq \|f\|.$$

Da $C^0(\mathbb{R}, \mathbb{C}) \cap L_2(\mathbb{R}, \mathbb{C})$ dicht in $L_2(\mathbb{R}, \mathbb{C})$ folgt dann die erste Teilbehauptung aus dem Fortsetzungssatz.

Mit $g \stackrel{\text{def}}{=} C(h)f$ ist

$$f(x) = g(x) - \rho \underbrace{\left(a\left(x + \frac{\Delta x}{2}\right)(g(x + \Delta x) - g(x)) - a\left(x - \frac{\Delta x}{2}\right)(g(x) - g(x - \Delta x)) \right)}_{\stackrel{\text{def}}{=} s(x)}$$

Somit

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \|f\|^2 = \|g\|^2 - \rho \int_{-\infty}^{\infty} \overline{g(x)} s(x) dx - \rho \int_{-\infty}^{\infty} g(x) \overline{s(x)} dx + \rho^2 \|s\|^2$$

Es ist aber

$$\int_{-\infty}^{\infty} \overline{g(x)} s(x) dx = \int_{-\infty}^{\infty} \underbrace{\left(a\left(x + \frac{\Delta x}{2}\right) \overline{g(x)} (g(x + \Delta x) - g(x)) \right)}_{\text{Subst. } y \stackrel{\text{def}}{=} x + \Delta x/2} dx -$$

benötigte Stetigkeit \rightarrow Subst. $y \stackrel{\text{def}}{=} x + \Delta x/2$

$$\begin{aligned}
& \underbrace{-a(x - \frac{\Delta x}{2})\overline{g(x)}(g(x) - g(x - \Delta x))}_{\text{Subst. } y \stackrel{\text{def}}{=} x - \Delta x/2} dx \\
& = \int_{-\infty}^{\infty} \left(a(x)\overline{g}(x - \frac{\Delta x}{2})(g(x + \frac{\Delta x}{2}) - g(x - \frac{\Delta x}{2})) - \right. \\
& \quad \left. - a(x)\overline{g}(x + \frac{\Delta x}{2})(g(x + \frac{\Delta x}{2}) - g(x - \frac{\Delta x}{2})) \right) dx \\
& = - \int_{-\infty}^{\infty} a(x) |g(x + \frac{\Delta x}{2}) - g(x - \frac{\Delta x}{2})|^2 dx = \int_{-\infty}^{\infty} g(x)\overline{s(x)} dx,
\end{aligned}$$

d.h.

$$\begin{aligned}
\|f\|^2 &= \|g\|^2 + 2\rho \int_{-\infty}^{\infty} \underbrace{a(x)}_{> 0, \text{ beschränkt, da } a' \text{ kompakten Träger hat}} |g(x + \frac{\Delta x}{2}) - g(x - \frac{\Delta x}{2})|^2 dx + \rho^2 \|s\|^2 \\
&\Rightarrow \\
\|f\| &\geq \|g\|, \quad \text{d.h. } \|C(h)\| \leq 1 \quad \Rightarrow \quad \text{Stabilität.}
\end{aligned}$$

Nun zur Konsistenz! Dazu wählen wir $v \in C_0^4(\mathbb{R}, \mathbb{C})$.

Wir wollen zeigen, daß

$$\frac{1}{h} \|E(t+h)(v) - C(h)E(t)(v)\| \leq Kh \quad \text{für } K = K(v) \text{ und } 0 \leq t \leq T.$$

Es ist mit

$$\begin{aligned}
v(\cdot, t+h) - C(h)v(\cdot, t) &= g(\cdot) \\
C^{-1}(h)g(\cdot) &= (I - \rho H(\Delta x))v(\cdot, t+h) - v(\cdot, t)
\end{aligned}$$

d.h.

$$\|g\| \leq \|(I - \rho H(\Delta x))v(\cdot, t+h) - v(\cdot, t)\|$$

Aber mit

$$\begin{aligned}
s_1(x + \Delta x, t+h) &= a(x + \frac{\Delta x}{2})(v(x + \Delta x, t+h) - v(x, t+h)) \\
s(x + \Delta x, t+h) &= s_1(x + \Delta x, t+h) - s_1(x - \Delta x, t+h) \\
\frac{1}{h} \left(v(x, t+h) - \rho \left(s(x + \Delta x, t+h) \right) - v(x, t) \right) &= \\
&= v_t(x, t+h) - v_{tt}(x, t + \theta_1 h)h/2 - \frac{1}{(\Delta x)^2} s(x + \Delta x, t+h).
\end{aligned}$$

Ferner ist nach dem Taylor'schen Satz (Entwicklung bzgl. der Variablen θ) mit

$$\begin{aligned}
\varphi(\theta) &\stackrel{\text{def}}{=} s(x + \theta\Delta x, t+h) : \\
\varphi(1) &= \varphi(0) + \varphi'(0) + \varphi''(0)/2 + \varphi'''(0)/6 + \varphi^{(4)}(\tilde{\theta})/24 \quad \text{mit } \tilde{\theta} \in]0, 1[
\end{aligned}$$

Ferner gilt mit den Abkürzungen (' bedeutet Differentiation bezüglich x)

$$\begin{aligned}
a_{\pm} &\stackrel{def}{=} a(x \pm \theta \frac{\Delta x}{2}, t + h), \\
v_{\pm} &\stackrel{def}{=} v(x \pm \theta \Delta x, t + h), \\
a &\stackrel{def}{=} a(x), \\
v &\stackrel{def}{=} v(x, t + h) \\
\varphi(\theta) &= a_- v_- - (a_+ + a_-)v + a_+ v_+ \\
\varphi'(\theta) &= \Delta x (-\frac{1}{2} a'_- v_- - a_- v'_- - \frac{1}{2} (a'_+ - a'_-)v + \frac{1}{2} a'_+ v_+ + a_+ v'_+) \\
\varphi''(\theta) &= (\Delta x)^2 (\frac{1}{4} a''_- v_- + a'_- v'_- + a_- v''_- - \frac{1}{4} (a''_+ + a''_-)v + \frac{1}{4} a''_+ v_+ + a'_+ v'_+ + a_+ v''_+) \\
\varphi'''(\theta) &= (\Delta x)^3 (-\frac{1}{8} a'''_- v_- - \frac{3}{4} a''_- v'_- - \frac{3}{2} a'_- v''_- - a_- v'''_- - \frac{1}{8} (a'''_+ - a'''_-)v \\
&\quad + \frac{1}{8} a'''_+ v_+ + \frac{3}{4} a''_+ v'_+ + \frac{3}{2} a'_+ v''_+ + a_+ v'''_+) \\
\varphi^{(4)}(\theta) &= (\Delta x)^4 \left(\frac{1}{16} a^{(4)}_- v_- + \frac{1}{2} a'''_- v'_- + \frac{3}{2} a''_- v''_- + 2a'_- v'''_- + a_- v^{(4)}_- \right. \\
&\quad \left. - \frac{1}{16} (a^{(4)}_+ + a^{(4)}_-)v + \frac{1}{16} a^{(4)}_+ v_+ + \frac{1}{2} a'''_+ v'_+ + \frac{3}{2} a''_+ v''_+ + 2a'_+ v'''_+ + a_+ v^{(4)}_+ \right)
\end{aligned}$$

d.h.

$$\varphi(0) = 0, \quad \varphi'(0) = 0, \quad \frac{1}{2} \varphi''(0) = (\Delta x)^2 (a'v' + av''), \quad \varphi'''(0) = 0.$$

Somit wird

$$-\frac{1}{(\Delta x)^2} s(x + \Delta x, t + h) = -(a'v' + av'')(x, t + h) - \frac{1}{24} (\Delta x)^2 \chi(x + \theta \Delta x, t + h)$$

Aufgrund der Voraussetzungen an a und v kann man zeigen, daß χ betragsmäßig abschätzbar ist durch eine bezüglich x quadratintegrierbare Funktion, d.h.

$$\exists K : \quad |\chi(x + \theta_x \Delta x, t + h)| \leq K(x) \quad \forall \quad 0 \leq t + h \leq T,$$

mit

$$\int_{-\infty}^{\infty} K^2(x) dx < \infty.$$

Ferner ist auch

$$\int_{-\infty}^{\infty} v_{tt}^2(x, \cdot) dx < K_1 < \infty$$

und somit

$$\begin{aligned}
\frac{1}{h} \|E(t + h)(v) - C(h)E(t)(v)\| &\leq hK_1 + (\Delta x)^2 \|K\|/24 \\
&= h(K_1 + \|K\|/24 \cdot \frac{1}{\rho})
\end{aligned}$$

d.h. es ist Konsistenz bewiesen und für $\eta_j h_j = t \in [0, T]$ und $u_0 \in D_C$ konvergiert das Verfahren von erster Ordnung in h , falls $\rho = \frac{\Delta t}{(\Delta x)^2} > 0$ fest, sonst aber beliebig. \square

Bemerkung 3.6 Das vollimplizite Verfahren für das reine Anfangswertproblem der parabolischen DGL ist natürlich nur von theoretisch-modellhaften Charakter, da es praktisch nicht durchführbar ist. Hier hängen ja die Werte auf der ersten Zeitschicht auch bei festem endlichen Δx von allen u_0 -Werten auf dem Anfangsgitter ab, d.h. man hätte ein unendliches Gleichungssystem pro Zeitschicht. \square

Als weiteres Beispiel soll nun noch eine einfache hyperbolische Gleichung dienen, nämlich die Konvektionsgleichung mit variablen Koeffizienten:

$$\partial_2 u(x, t) = a(x) \partial_1 u(x, t), \quad a \in C^\infty(\mathbb{R}, \mathbb{R}), \quad 0 < |a(x)| \leq K \quad \forall x \in \mathbb{R}.$$

Beispiel 3.8 Friedrichs-Verfahren:

$$\frac{1}{h} \left(u(x, t+h) - \frac{1}{2} (u(x+\Delta x, t) + u(x-\Delta x, t)) \right) \approx \frac{a(x)}{2\Delta x} \left(u(x+\Delta x, t) - u(x-\Delta x, t) \right)$$

Also

$$C(h) = \frac{1 - \rho a(x)}{2} T_{\Delta x}^{-1} + \frac{1 + \rho a(x)}{2} T_{\Delta x} \quad \text{mit } \rho = \frac{h}{\Delta x}$$

Als Banachraum legen wir wieder $B = L_2(\mathbb{R}, \mathbb{R})$ zugrunde. $D_A = C^1(\mathbb{R}, \mathbb{R}) \cap L_2(\mathbb{R}, \mathbb{R})$
Behauptung: Das Friedrichs-Verfahren ist konsistent von der Ordnung 1 und stabil für $\rho \leq \frac{1}{K}$. Wir beweisen zunächst die Konsistenz. Dazu wählen wir $D_C = C_0^2(\mathbb{R}, \mathbb{R})$. Ähnlich wie in Beispiel 3.7 setzen wir (mit $v \in D_C$)

$$\varphi(\theta) \stackrel{\text{def}}{=} \frac{1}{2} \left(v(x+\theta\Delta x, t) + v(x-\theta\Delta x, t) + \rho a(x) (v(x+\theta\Delta x, t) - v(x-\theta\Delta x, t)) \right)$$

mit

$$\begin{aligned} \varphi(1) &= C(h)v(x, t), \\ \varphi(1) &= \varphi(0) + \varphi'(0) + \varphi''(\theta)/2, \quad \theta \in]0, 1[, \\ \varphi(0) &= v(x, t), \\ \varphi'(0) &= \frac{1}{2}\Delta x \left(\partial_1 v(x, t) - \partial_1 v(x, t) + 2\rho a(x) \partial_1 v(x, t) \right) = h a(x) \partial_1 v(x, t). \end{aligned}$$

Mit $v \in C_0^2(\mathbb{R}, \mathbb{R})$ ist auch $E(t)v \in C_0^2(\mathbb{R}, \mathbb{R})$,¹ d.h. für $t \in [0, T]$ kann man abschätzen ($\Delta x \leq 1$)

$$\begin{aligned} \frac{1}{(\Delta x)^2} |\varphi''(\theta)| &\leq L(x) \quad \text{mit } L \in L_2(\mathbb{R}, \mathbb{R}) \\ L(x) &= \left(\sup_{\substack{y \in [-1, 1] \\ t \in [0, T]}} |v_{xx}(x+y, t)| \right) \cdot (1 + \rho K) \end{aligned}$$

¹ $u(x, t) = w \left(\int_0^x \frac{1}{a(\xi)} d\xi + t \right)$ mit $w \left(\underbrace{\int_0^x \frac{1}{a(\xi)} d\xi}_{\varphi(x)} \right) = u_0(x)$ d.h. $w(x) \stackrel{\text{def}}{=} u_0(\varphi^{-1}(x))$.

Somit wird

$$\begin{aligned} \frac{1}{h}|v(x, t+h) - C(h)v(x, t)| &= \frac{1}{h}|v(x, t) + h\partial_2 v(x, t) + \frac{h^2}{2}\partial_2^2 v(x, t + \theta_1 h) \\ &\quad - v(x, t) - ha(x)\partial_1 v(x, t) - \varphi''(\theta)/2| \leq \frac{h}{2}|\partial_2^2 v(x, t + \theta_1 h)| + \frac{1}{h}(\Delta x)^2 L(x)/2 \\ &\leq \frac{h}{2}|\partial_2^2 v(x, t + \theta_1 h)| + h L(x)/(2\rho^2) \end{aligned}$$

und da auch $\partial_2^2 v(x, t + \theta_1 h)$ quadratintegabel ist (bzgl. x), folgt die erste Teilbehauptung.

Zum Beweis der Stabilität sei $f \in L_2(\mathbb{R}, \mathbb{R})$. Es ist

$$\begin{aligned} C(h)(f)(x) &= \frac{1}{2}(f(x + \Delta x) + f(x - \Delta x)) + \frac{\rho}{2}a(x)(f(x + \Delta x) - f(x - \Delta x)) \\ \|C(h)f\|^2 &= \frac{1}{4} \int_{-\infty}^{\infty} |f(x + \Delta x) + f(x - \Delta x)|^2 dx \\ &\quad + \frac{\rho^2}{4} \int_{-\infty}^{\infty} a^2(x) |f(x + \Delta x) - f(x - \Delta x)|^2 dx \\ &\quad + \frac{\rho}{2} \int_{-\infty}^{\infty} a(x) (|f(x + \Delta x)|^2 - |f(x - \Delta x)|^2) dx \\ &\leq \frac{1}{2} \int_{-\infty}^{\infty} (|f(x + \Delta x)|^2 + |f(x - \Delta x)|^2) dx \\ &\quad + \frac{\rho}{2} \int_{-\infty}^{\infty} (a(x - \Delta x) |f(x)|^2 - a(x + \Delta x) |f(x)|^2) dx \end{aligned}$$

(wegen $|\rho a(x)| \leq 1$ und $|\alpha + \beta|^2 + |\alpha - \beta|^2 = 2(|\alpha|^2 + |\beta|^2)$ und der Argumentverschiebung im 2. Integral)

$$\leq \|f\|^2 + \rho \Delta x \tilde{K} \|f\|^2$$

(wegen $|a(x - \Delta x) - a(x + \Delta x)| = |2\Delta x a'(x + \theta\Delta x)| \leq 2\Delta x \tilde{K}$). Also wegen $\rho \Delta x = \Delta t = h$

$$\|C(h)f\|^2 \leq (1 + h\tilde{K})\|f\|^2$$

Dies ergibt

$$\|C(h)\| \leq \sqrt{1 + h\tilde{K}} \leq \exp(h\tilde{K}/2)$$

und somit

$$\|C(h)^n\| \leq \|C(h)\|^n \leq \exp(T\tilde{K}/2) \stackrel{\text{def}}{=} K^* \quad \text{für } nh \leq T.$$

Damit ist die Stabilität des Verfahrens bewiesen. □

Ein anderes bewährtes Verfahren für die Konvektionsgleichung ist das Verfahren von Courant-Isaacson-Rees: (CIR)

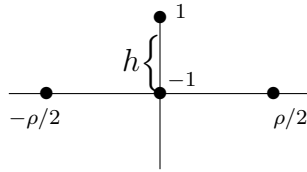
$$\begin{aligned} \frac{1}{h}(u(x, t+h) - u(x, t)) &\approx \frac{1}{\Delta x} \left((a(x))_+ (u(x + \Delta x, t) - u(x, t)) + (a(x))_- \right. \\ &\quad \left. \cdot (u(x, t) - u(x - \Delta x, t)) \right) \end{aligned}$$

Dieses Verfahren hat die gleichen Stabilitätseigenschaften wie das Friedrichs-Verfahren und ist ebenfalls konsistent von der Ordnung 1.

$$(y)_+ = \begin{cases} y & \text{für } y \geq 0 \\ 0 & \text{sonst} \end{cases} \quad \text{und entsprechend} \quad (y)_- = \begin{cases} 0 & \text{für } y \geq 0 \\ y & \text{sonst.} \end{cases}$$

Das “naive” Verfahren

$$u(x, t + h) - u(x, t) = \frac{1}{2}\rho a(x)(u(x + \Delta x, t) - u(x - \Delta x, t))$$



ist dagegen instabil. Das Friedrichs-Verfahren und das CIR-Verfahren kann man auch so schreiben:

$$\begin{aligned} u(x, t + h) - u(x, t) &= \frac{1}{2}\rho a(x) \left(u(x + \Delta x, t) - u(x - \Delta x, t) \right) \\ &\quad + \frac{1}{2} \left(u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t) \right) \end{aligned}$$

bzw.:

$$\begin{aligned} u(x, t + h) - u(x, t) &= \frac{1}{2}\rho a(x) \left(u(x + \Delta x, t) - u(x - \Delta x, t) \right) \\ &\quad + \frac{1}{2}\rho |a(x)| \left(u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t) \right) \end{aligned}$$

(Übg.) Man erkennt, daß gegenüber dem instabilen naiven Verfahren ein Term $\frac{1}{2}(\Delta x)^2 \partial_1^2 u(x + \theta \Delta x, t)$ bzw. $\frac{1}{2}\rho |a(x)| (\Delta x)^2 \partial_1^2 u(x + \theta \Delta x, t)$ hinzukommt, den man als Diskretisierung von $\varepsilon(x, h) \partial_1^2 u(x, t)$ mit $\varepsilon \rightarrow 0$ auffassen kann, also eine Diskretisierung eines **parabolischen** (und damit glättenden) Anteils. Man nennt solche Terme **numerische Viskosität**. Sie beeinträchtigen die Konsistenzordnung nicht und wirken genügend glättend, um das naive Verfahren zu stabilisieren. Bei Verfahren höherer Ordnung ist die Bestimmung geeigneter Terme für numerische Viskosität schwierig.

3.3 Mehrschrittverfahren

Die Theorie in Abschnitt 3.1 wurde nur für Einschrittverfahren formuliert. Mehrschrittverfahren, von denen ja bereits einige Beispiele gebracht wurden, paßt man in diese Theorie ein, indem man sie in der von gewöhnlichen DGLen her schon bekannten Weise als Einschrittverfahren für eine Funktion in einem Produktraum umschreibt.

Sei die Anfangswertaufgabe

$$\begin{aligned} u_t &= Au & 0 \leq t \leq T & \quad D_A \subset B \quad \text{dicht in } B \\ u(t, 0) &= u_0, & u_0 \in B \end{aligned}$$

vorgegeben und das Mehrschrittdifferenzenverfahren habe die Form

$$D_q u^{q+\nu} + D_{q-1} u^{q-1+\nu} + \dots + D_0 u^\nu = 0, \quad \nu = 0, 1, \dots \quad (3.1)$$

Dabei sind die D_i lineare und beschränkte Operatoren auf B und D_q sei umkehrbar. (Die D_i bauen sich wieder aus Linearkombinationen von Potenzen des Shiftoperators $T_{\Delta x}$ auf. Bei gegebener fester Beziehung zwischen $h = \Delta t$ und den Δx_j sind die D_i dann noch vom Parameter h abhängig.)

Mit

$$C_j \stackrel{\text{def}}{=} -D_q^{-1} \cdot D_j \quad j = 0, \dots, q-1 \quad C_j = C_j(\Delta t)$$

erhält man

$$u^{q+\nu} = \sum_{i=1}^q C_{q-i} u^{q+\nu-i}$$

Nun betrachten wir den Produktraum B^q mit der Norm $w \in B^q$:

$$\| |w| \|^2 \stackrel{\text{def}}{=} \sum_{i=1}^q \|w_i\|^2 \quad \| \cdot \| \quad \text{Norm in } B$$

Setzen wir

$$w^{\nu+1} \stackrel{\text{def}}{=} \begin{bmatrix} u^{q+\nu} \\ \vdots \\ u^{\nu+1} \end{bmatrix} \quad (\in B^q)$$

dann ist das Mehrschrittverfahren in B äquivalent zum Einschrittverfahren

$$w^{\nu+1} = C(\Delta t) w^\nu$$

mit

$$C(\Delta t) = \begin{bmatrix} C_{q-1} & C_{q-2} & \dots & C_0 \\ I & 0 & & 0 \\ 0 & I & & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & I \end{bmatrix}$$

$C(\Delta t)$ ist nach Definition und nach Konstruktion von $||| \cdot |||$ ein beschränkter linearer Operator von B^q in B^q .

Um das Verfahren zu starten, benötigt man den Anfangsvektor w^0 . Wir stellen uns zunächst vor, die exakte Lösung sei auf ersten q Zeitschichten bekannt, d.h.

$$w^0 = S \begin{pmatrix} u^0 \\ \vdots \\ u^0 \end{pmatrix} \quad \text{mit} \quad S \stackrel{\text{def}}{=} \begin{pmatrix} E((q-1)h) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & E(h)I \end{pmatrix}$$

Dann wird

$$w^\nu = C(h)^\nu S \begin{pmatrix} u^0 \\ \vdots \\ u^0 \end{pmatrix}$$

während die entsprechend gebildete Größe aus den wahren Werten geschrieben werden kann als

$$w(\nu h) = E_q(h)^\nu S \begin{pmatrix} u^0 \\ \vdots \\ u^0 \end{pmatrix} \quad E_q(h) = \begin{pmatrix} E(h) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & E(h) \end{pmatrix}$$

Es ist also offensichtlich das Verhalten von

$$C(h)^\nu - E_q(h)^\nu$$

entscheidend für das Verfahren. Berechnet man die ersten q Zeitschichten mit einem anderen Verfahren, dann hat man in der Formel für w^ν noch einen zusätzlichen Term $C(h)^\nu \Delta w^0$. Wir sehen voraus, daß

$$|||\Delta w^0||| \xrightarrow{h \rightarrow 0} 0.$$

Dann beeinflusst dieser Fehler Stabilität und Konvergenz des Verfahrens nicht, allenfalls die Konvergenzordnung.

Definition 3.11 Das Verfahren (3.1) heißt konsistent zur AWA, wenn es eine Menge $D_C \subset B$, $\bar{D}_C = B$ gibt mit $E(t)u_0$ klassische Lösung für jedes $u_0 \in D_C$ und

$$\lim_{h \rightarrow \infty} \frac{1}{|h|} |||(C(h) - E_q(h))S\tilde{u}(t)||| = 0 \quad \text{gleichmäßig in } t \in [0, T]$$

wobei $\tilde{u}(t) \stackrel{\text{def}}{=} E_q(t) \begin{pmatrix} u_0 \\ \vdots \\ u_0 \end{pmatrix} \in B^q$. □

Definition 3.12 Das Verfahren (3.1) heißt konvergent, wenn für jede Nullfolge $\{h_j\}$ und $\{n_j\} \subset \mathbb{N}$ mit $n_j h_j \rightarrow t \in [0, T]$ gilt

$$\lim_{\substack{j \rightarrow \infty \\ \Delta w^0 \rightarrow 0}} \left\| \left\| C(h_j)^{n_j} \left(S \begin{pmatrix} u_0 \\ \vdots \\ u_0 \end{pmatrix} + \Delta w^0 \right) - E_q(t) \begin{pmatrix} u_0 \\ \vdots \\ u_0 \end{pmatrix} \right\| \right\| = 0$$

□

Definition 3.13 Die Schar von Differenzenoperatoren

$$M_D = \{C(h) : 0 < h \leq h_0\}$$

heißt stabil, falls $\|C(h)^n\| \leq K$ für $0 < h \leq h_0$ und $0 \leq nh \leq T$ mit einer von h und n unabhängigen Konstanten K . □

Satz 3.9 Die AWA sei sachgemäß gestellt und das Mehrschrittverfahren sei konsistent. Dann ist das Verfahren genau dann konvergent, wenn M_D stabil ist.

Beweis: analog zum Beweis in Satz 3.1.1. Die zusätzlichen Betrachtungen sind analog zu denen bei MSV für gewöhnliche DGLen. Der Beweis kann z.B. im Buch von Richtmyer und Morton (Abschnitt 7.3) nachgelesen werden. □

Wir betrachten jetzt einige Beispiele von MSV.

Zunächst formen wir MSV für die Aufgabe

$$\begin{aligned} u_t &= u_{xx} & x \in \mathbb{R}, \quad 0 < t \leq T \\ u(x, 0) &= u_0 \end{aligned}$$

auf ESV um:

Beispiel 3.9 Das "5-Punkt-Verfahren" (Semidiskretisierung mit expliziter Mittelpunkregel) ergibt

$$\frac{u^{\nu+2} - u^\nu}{2h} = \frac{1}{(\Delta x)^2} (T_{\Delta x}^{-1} - 2I + T_{\Delta x}) u^{\nu+1}$$

oder

$$u^{\nu+2} = u^\nu + 2\rho (T_{\Delta x}^{-1} - 2I + T_{\Delta x}) u^{\nu+1} \quad \rho = \frac{h}{(\Delta x)^2}, \quad h \hat{=} \Delta t$$

d.h.

$$C(h) = \begin{bmatrix} 2\rho(T_{\Delta x}^{-1} - 2I + T_{\Delta x}) & I \\ I & 0 \end{bmatrix}$$

(Wir werden in Abschnitt 3.5 zeigen, daß $M_D = \{C(h) : 0 < h \leq h_0\}$ nie stabil ist.)

□

Beispiel 3.10 Das Verfahren von Du Fort und Frankel

$$\begin{aligned} \frac{u^{\nu+2} - u^\nu}{2h} &= \frac{1}{(\Delta x)^2} \left((T_{\Delta x}^{-1} + T_{\Delta x})u^{\nu+1} - (u^{\nu+2} + u^\nu) \right) \\ &\left(= \frac{1}{(\Delta x)^2} \left((T_{\Delta x}^{-1} + T_{\Delta x})u^{\nu+1} - 2Iu^{\nu+1} - (u^{\nu+2} - 2u^{\nu+1} + u^\nu) \right) \right) \\ &= \frac{1}{(\Delta x)^2} \left((T_{\Delta x}^{-1} - 2I + T_{\Delta x})u^{\nu+1} - (\Delta x)^2 u_{tt}(x, \theta_\nu) \right) \quad \theta_\nu \in]t_\nu, t_{\nu+2}[\\ &\text{d.h. hier wirkt ein hyperbolischer Term stabilisierend!} \end{aligned}$$

ergibt

$$u^{\nu+2} = \frac{2\rho}{1+2\rho} (T_{\Delta x} + T_{\Delta x}^{-1})u^{\nu+1} + \frac{1-2\rho}{1+2\rho} u^\nu$$

d.h.

$$C(h) = \begin{bmatrix} \frac{2\rho}{1+2\rho} (T_{\Delta x} + T_{\Delta x}^{-1}) & \frac{1-2\rho}{1+2\rho} I \\ I & 0 \end{bmatrix}$$

□

Beispiel 3.11 Semidiskretisierung und 2-Schritt BDF (Gear)

$$\frac{3}{2h} (u^{\nu+2} - u^{\nu+1}) - \frac{1}{2h} (u^{\nu+1} - u^\nu) = \frac{1}{(\Delta x)^2} (T_{\Delta x}^{-1} - 2I + T_{\Delta x})u^{\nu+2}$$

d.h.

$$C(h) = \begin{pmatrix} 3u^{\nu+2} - 4u^{\nu+1} + u^\nu = 2\rho(T_{\Delta x}^{-1} - 2I + T_{\Delta x})u^{\nu+2} \\ 4(-2\rho(T_{\Delta x}^{-1} + T_{\Delta x}) + (3+4\rho)I)^{-1} & -(\dots)^{-1} \\ I & 0 \end{pmatrix}$$

□

3.4 Ergänzungen zur Lax-Richtmyer-Theorie

Wie wir zuvor gesehen haben, spielt die Stabilität der Schar von Differenzenoperatoren $M_D = \{C(h) : 0 < h \leq h_0\}$ eine entscheidende Rolle. In diesem Zusammenhang ist es wichtig, daß man bei einer "leichten Abänderung" des Differenzschemas die Stabilität nicht verliert:

Satz 3.10 (Kreiss): *Es sei $P(B, T, A)$ eine sachgemäß gestellte Aufgabe $M_D = \{C(h) : 0 < h \leq h_0\}$ sei ein stabiles Differenzenverfahren zu $P(B, T, A)$ und $\{Q(h) : 0 < h \leq h_0\}$ irgendeine Schar von gleichmäßig beschränkten linearen Operatoren $B \rightarrow B$. Dann ist*

$$\tilde{M}_D = \{C(h) + hQ(h) : 0 < h \leq h_0\}$$

ebenfalls stabil.

Beweis: Nach Voraussetzung gilt

$$\begin{aligned} \|C(h)^n\| &\leq K_1 \quad \text{für } n \in \mathbb{N}, \quad h \in]0, h_0] \quad \text{und } nh \leq T \\ \|Q(h)\| &\leq K_2 \quad \text{für } 0 < h \leq h_0. \end{aligned}$$

Nun gilt:

$$\begin{aligned} (C(h) + hQ(h))^m &= (C(h) + hQ(h)) \cdots (C(h) + hQ(h)) \\ &= \sum_{\nu=0}^m h^\nu \sum_{k=1}^{\binom{m}{\nu}} P_{k,\nu} \\ P_{k,\nu} &= \text{Produkt aus } \nu \text{ Faktoren } Q(h) \text{ und } m - \nu \text{ Faktoren } C(h) \\ &\quad \max \nu + 1 \text{ Potenzen von } C(h) \text{ trennen die Faktoren } Q(h) \end{aligned}$$

(Man beachte, daß $C(h)$ und $Q(h)$ nicht vertauschbar zu sein brauchen, daß also die binomische Formel **nicht** notwendig gilt!)

Anwendung der Dreiecksungleichung liefert:

$$\begin{aligned} \|(C(h) + hQ(h))^m\| &\leq \sum_{\nu=0}^m h^\nu \sum_{k=1}^{\binom{m}{\nu}} K_1^{\nu+1} K_2^\nu = K_1 \sum_{\nu=0}^m \binom{m}{\nu} (hK_1K_2)^\nu \\ &= K_1(1 + hK_1K_2)^m \leq K_1 \exp(mhK_1K_2) \leq K_1 \exp(TK_1K_2) \quad \square \end{aligned}$$

Die große formale Einfachheit der Lax-Richtmyer-Theorie in der bisher formulierten Form ist durch einige einschränkende Voraussetzungen erkauft worden:

1. Es werden nur reine Anfangswertaufgaben betrachtet. Dies kann man durch Einarbeitung von Randbedingungen in den Raum B erreichen, erhält dann aber bei inhomogenen Randbedingungen inhomogene Anfangswertaufgaben, vgl. dazu unten.

2. Es werden nur **lineare** Anfangswertaufgaben mit zeitunabhängigen Koeffizienten betrachtet. Man kann jedoch wesentliche Teile der Theorie (hinreichende Konvergenzbedingungen) auf viel allgemeinere Fälle ausdehnen, vergleiche dazu bei Ansorge.
3. Alle Differenzenoperatoren sind auf dem gleichen Banachraum B wie die AWA selbst definiert und bilden ihn in sich ab, man hat es also nicht mit Gitterfunktionen zu tun, wie wir sie zuvor bei gew. DGLen betrachtet haben
4. Die Differenzenoperatoren sind für kontinuierlich variierende Schrittweite definiert

Von den Voraussetzungen 3. und 4. kann man sich durch Einführung "streng endlicher" Differenzenverfahren befreien. Ein solches Resultat ist (vgl. für den Beweis bei Meiss-Marcowitz):

Definition 3.14 Sei $h_0 > 0$ fest. $P(B, T, A)$ sei sachgemäß gestellt.

(i) Die Folge

$$M_D = \{(B_\nu, r_\nu, C_\nu) \quad \nu \in \mathbb{N}_0\}$$

heißt **streng endliches Differenzenverfahren**, wenn gilt

B_ν ist ein endlich-dimensionaler Banachraum mit der Norm $\|\cdot\|^{(\nu)}$

r_ν ist eine lineare Abbildung (Restriktion) von B in B_ν und

$$\lim_{\nu \rightarrow \infty} \|r_\nu(c)\|^{(\nu)} = \|c\| \quad (\forall c \in B)$$

$C_\nu : B_\nu \rightarrow B_\nu$ linear, $C_\nu = C_\nu(h_\nu)$ und $h_{\nu+1} = h_\nu/k_\nu$, $k_\nu \in \mathbb{N} \setminus \{1\}$

(Konkret ist C_ν definiert mit Hilfe von Zeitschrittweiten h_ν und Shiftoperatoren $T_{(\Delta x_k), \nu}$. Die Gitter $\{jh_\nu\}$ sind kohärent.)

(ii) M_D heißt **konsistent**, wenn es einen in B dichten Teilraum D_C gibt mit $D_C \subset D_{E_0}$, so daß für $c \in D_C$ gilt

$$\lim_{\nu \rightarrow \infty} \frac{1}{h_\nu} \|C_\nu \circ r_\nu E_0(t)(c) - r_\nu \circ E_0(t + h_\nu)(c)\|^{(\nu)} = 0 \quad \text{glm. in } [0, T]$$

(iii) M_D heißt **stabil**, wenn die Menge

$$\{\|C_\nu^n\|^{(\nu)} : \nu \in \mathbb{N}_0, \quad n \in \mathbb{N}, \quad nh_\nu \leq T\}$$

beschränkt ist.

(iv) M_D heißt **konvergent**, wenn für alle $t \in [0, T]$ mit $t = n_\nu h_\nu$ für ein $n_\nu \in \mathbb{N}_0$ und ein $\nu \in \mathbb{N}_0$ gilt:

$$\lim_{\mu \rightarrow \infty} \|C_\mu^{n_\mu} \circ r_\mu(c) - r_\mu \circ E_0(t)(c)\|^{(\mu)} = 0$$

$$n_\mu h_\mu = n_\nu h_\nu$$

□

Es gilt:

Satz 3.11 *Ist ein streng endliches Differenzenverfahren konsistent und stabil, dann ist es auch konvergent. (Beweis siehe bei Meis-Marcowitz) \square*

Wir betrachten nun noch inhomogene lineare Anfangswertaufgaben

$$\begin{aligned} u'(t) &= A(u(t)) + g(t) & t \in [0, T] \\ u(0) &= u_0 \end{aligned} \quad (3.2)$$

Es ergibt sich, daß die Konsistenz und Stabilität für das homogene Problem ($g \equiv 0$) ausreicht, um die Konvergenz auch für die inhomogene Aufgabe sicherzustellen. $P(B, T, A)$ sei im folgenden eine sachgemäß gestellte Anfangswertaufgabe (3.2).

Von der Inhomogenität

$$g: [0, T] \rightarrow B$$

setzen wir Stetigkeit voraus und benutzen als Norm

$$\|g\| = \max_{t \in [0, T]} \|g(t)\|,$$

wobei $\|\cdot\|$ die Norm in B ist.

Bemerkung 3.7 *In den Anwendungen tritt zunächst in der Regel ein Anfangsrandwertproblem auf, das geschrieben werden kann als*

$$\begin{aligned} \dot{u}(t) &= A(u(t)) + g(t) & t \in [0, T] \\ A_1 u(t) &= h(t) & (\text{Randwerte}) \\ u(0) &= u_0 \quad u_0 \in B & (\text{Anfangswerte}) \end{aligned}$$

Dabei ist A_1 ein linearer Operator, der die Projektion von B auf die Randwerte (in den Raumkoordinaten) darstellt.

Die Randvorgaben sind hier also inhomogen. Sei nun eine hinreichend glatte Funktion w bekannt mit

$$A_1 w(t) = h(t) \quad t \in [0, T]$$

und

$$v \stackrel{\text{def}}{=} u - w.$$

Dann bekommt man

$$\begin{aligned} \dot{v}(t) &= Av(t) + \tilde{g}(t) & t \in [0, T] \\ A_1 v(t) &= 0 \\ v(0) &= u_0 - w(0) \end{aligned}$$

mit

$$\tilde{g}(t) = g(t) + Aw(t) - \dot{w}(t)$$

(d.h. w muß in D_A liegen). Definiert man nun

$$\tilde{D}_A \stackrel{\text{def}}{=} D_A \cap \{v \in B : A_1 v = 0\}$$

dann kann man schreiben: Gesucht $\tilde{v} \in \tilde{D}_A$:

$$\begin{aligned}\tilde{v}(t) &= A\tilde{v}(t) + \tilde{g}(t) \\ \tilde{v}(0) &= u_0 - w(0)\end{aligned}$$

d.h. man hat jetzt ein reines Anfangswertproblem in einem modifizierten Funktionenraum, allerdings mit inhomogener DGL auch für $g \equiv 0$. \square

Wir betrachten also weiter unser inhomogenes Anfangswertproblem (3.2). Wir erinnern zunächst an die Lösungsformel für eine skalare lineare inhomogene Differentialgleichung

$$\begin{aligned}\dot{y}(t) &= \lambda y(t) + g(t), & y(0) &= y_0 \\ y(t) &= e^{\lambda t} y_0 + \int_0^t e^{(t-s)\lambda} g(s) ds\end{aligned}$$

Dem Term $e^{\lambda t}$ entspricht in unserem Fall hier der Operator $E(t)$ und man nennt

$$u(t) \stackrel{\text{def}}{=} E(t)(u_0) + \int_0^t E(t-s)g(s) ds \quad (3.3)$$

die **verallgemeinerte Lösung** der AWA (3.2). Diese verallgemeinerte Lösung ist (im Sinne der Differenzierbarkeit in B) bei beliebigem $g \in C([0, T], B)$ nicht differenzierbar. Nur wenn g "genügend glatt" ist, kann man tatsächlich differenzieren:

Definition 3.15 *Es sei*

$$\begin{aligned}D_{\tilde{A}} &\stackrel{\text{def}}{=} \{c \in B : u(t) = E(t)(c) \text{ ist differenzierbar für } t = 0\} \\ \tilde{A} &: A|_{D_{\tilde{A}}} \rightarrow B \quad \text{und } c \mapsto u'(0)\end{aligned}$$

\square

Bemerkung 3.8 *Für $c \in D_{\tilde{A}}$ ist u überall in $[0, T]$ differenzierbar wegen*

$$\frac{1}{h}(u(t_1 + h) - u(t_1)) = E(t_1) \frac{1}{h}(E(h)(c) - c)$$

\square

Mit beliebigem $g \in C([0, T], B)$ definieren wir

$$\tilde{g}(t) \stackrel{\text{def}}{=} \int_0^\infty \varphi(r) E(r)(g(t)) \, dr$$

Dabei sei $Tr(\varphi) \subset]0, T[$ und $\varphi \in C^\infty(\mathbb{R}, \mathbb{R})$. Man nennt \tilde{g} eine Regularisierung von g . Man kann φ so konstruieren, daß bei gegebenem $\varepsilon > 0$ $\|\tilde{g}(t) - g(t)\| < \varepsilon$ $\forall t \in [0, T]$. Benutzt man \tilde{g} als Inhomogenität anstelle von g , dann hat man Differenzierbarkeit der verallgemeinerten Lösung (3.3), d.h. auf $D_{\tilde{A}}$

$$\frac{d}{dt} \left(\int_0^t E(t-s)(\tilde{g}(s)) \, ds \right) = \tilde{g}(t) + \tilde{A} \left(\int_0^t E(t-s)(\tilde{g}(s)) \, ds \right)$$

d.h. letztendlich für $u_0 \in D_{\tilde{A}}$

$$\dot{u}(t) = \tilde{A}u(t) + \tilde{g}(t)$$

Den Beweis dieser Beziehung und des nachfolgenden Satzes kann man bei Meis-Marcowitz nachlesen.

Satz 3.12 Gegeben sei $u_0 \in B$, $g \in C([0, T], B)$ und ein konsistentes und stabiles Differenzenverfahren $M_D = \{C(h) : 0 < h \leq h_0\}$ für die sachgemäß gestellte Anfangswertaufgabe $P(B, T, A)$.

Ferner gelte $\lim_{\substack{h_j \rightarrow 0 \\ n_j \rightarrow \infty}} h_j n_j = t \in [0, T]$, und $\Theta \in [0, 1]$ fest. Dann konvergiert die Lösung $u^{(n_j)}$ der Differenzgleichung

$$\begin{aligned} u^{(0)} &= u_0 \\ u^{(\nu)} &= C(h_j)(u^{(\nu-1)}) + h_j g(\nu h_j - \Theta h_j), \quad \nu = 1, \dots, n_j \end{aligned}$$

für $j \rightarrow \infty$ gegen die verallgemeinerte Lösung (3.3) der AWA (3.2). □

3.5 Kriterien für die Stabilität von Differenzenverfahren

Aus der bisher dargestellten Theorie folgt, daß es genügt, Konsistenz und Stabilität eines Differenzenverfahrens zu beweisen, um die Konvergenz sicherzustellen. Die Konsistenz folgt in einfachster Weise aus der Verfahrenskonstruktion (mittels des Taylor'schen Satzes). Es bleibt somit "nur" noch die Aufgabe, nach einfachen hinreichenden bzw. notwendigen Kriterien für die Stabilität, d.h. für die gleichmäßige Beschränktheit der Operatorenfamilie

$$\{C(h)^n : 0 < h \leq h_0 \quad nh \leq T\}$$

zu suchen. Im Spezialfall einer DGL mit konstanten Koeffizienten gelingt dies mittels der Methode der Fouriertransformation, die wir im Folgenden besprechen werden. Auch für nichtkonstante Koeffizienten und sogar für nichtlineare Probleme gibt es einzelne Resultate, in diesen Fällen läßt sich aber oft die Stabilität im Einzelfall direkt einfacher zeigen, als das Verfahren erst in die allgemeine Theorie einzupassen.

Wir beginnen unsere Darstellung mit einigen Ergebnissen aus der Theorie der Fouriertransformation. Die gesamte Darstellung kann ganz analog auf mehrere räumliche Veränderliche übertragen werden.

Im folgenden ist B einer der Räume

$$L_2(\mathbb{R}, \mathbb{C}^n), \quad L_2((0, 2\pi), \mathbb{C}^n), \quad L_2((0, \pi), \mathbb{C}^n).$$

Damit, der allgemeinen Theorie entsprechend, der Translationsoperator $T_{\Delta x}$ auch im Falle $L_2((0, 2\pi), \mathbb{C}^n)$ und $L_2((0, \pi), \mathbb{C}^n)$ beliebig angewendet werden kann, denken wir uns die Funktionen entsprechend fortgesetzt:

$$\left. \begin{array}{l} f(x + 2\pi) = f(x) \\ f(x + 2\pi) = f(x) \\ f(x) = -f(-x) \end{array} \right\} \quad (\forall x \in \mathbb{R}) \quad \text{für } f \in L_2((0, 2\pi), \mathbb{C}^n)$$

$$\left. \begin{array}{l} f(x) = -f(-x) \end{array} \right\} \quad (\forall x \in \mathbb{R}) \quad \text{für } f \in L_2((0, \pi), \mathbb{C}^n)$$

Durch diese Fortsetzung wird der Raum $L_2((0, \pi), \mathbb{C}^n)$ zu einem abgeschlossenen Teilraum von $L_2((0, 2\pi), \mathbb{C}^n)$.

Funktionen f aus $L_2((0, 2\pi), \mathbb{C}^n)$ und $L_2((0, \pi), \mathbb{C}^n)$ sollen entsprechend dieser Einbettung k -mal stetig differenzierbar heißen, wenn ihre Fortsetzungen auf \mathbb{R} k -mal stetig differenzierbar sind, d.h. also insbesondere auch periodisch. Für $L_2((0, \pi), \mathbb{C}^n)$ ergeben sich dann auch automatisch Nullrandbedingungen für die geradzahigen Ableitungen:

$$f^{(2\nu)}(0) = f^{(2\nu)}(\pi) = 0 \quad \nu = 0, \dots, (k/2) \quad f \in C^k(\mathbb{R}, \mathbb{C}^n) \cap L_2((0, \pi), \mathbb{C}^n)$$

Satz 3.13 Die Abbildung (Fourierentwicklung)

$$\mathfrak{F}_{2\pi,n} : L_2((0, 2\pi), \mathbb{C}^n) \rightarrow l_2(\mathbb{C}^n)$$

$$\left(l_2(\mathbb{C}^n) = \left\{ \{a(\nu)\}_{\nu \in \mathbb{Z}}, \quad a(\nu) \in \mathbb{C}^n \quad \sum_{\nu=-\infty}^{\infty} |a(\nu)|^2 < \infty, \quad i = 1, \dots, n \right\} \right)$$

definiert durch

$$f \mapsto \{a(\nu) : \nu \in \mathbb{Z}\}$$

$$a(\nu) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(x) \exp(-i\nu x) dx \quad i = \sqrt{-1}$$

ist ein normtreuer Isomorphismus des $L_2((0, 2\pi), \mathbb{C}^n)$ auf den $l_2(\mathbb{C}^n)$ (d.h. $\mathfrak{F}_{2\pi,n}$ ist linear, surjektiv, injektiv und erfüllt $\|\mathfrak{F}_{2\pi,n}(f)\| = \|f\|$)

Ist $a \in l_2(\mathbb{C}^n)$ gegeben und setzt man

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} \sum_{\nu=-\mu}^{\mu} a(\nu) \exp(i\nu x) \quad \mu = 0, 1, \dots$$

dann ist die Folge $\{f_\mu\}$ eine Cauchyfolge in $L_2((0, 2\pi), \mathbb{C}^n)$. Sei f der Grenzwert von $\{f_\mu\}$. Die Zuordnung

$$a \mapsto f$$

wird als inverse Fourierentwicklung bezeichnet ($\mathfrak{F}_{2\pi,n}^{-1}$).

Wenn f stetig und von beschränkter Variation ist auf $[0, 2\pi]$, dann konvergiert $f_\mu \rightarrow f$ sogar gleichmäßig.

Falls $f \in L_2((0, \pi), \mathbb{C}^n)$, dann gilt $a(\nu) = -a(-\nu) \quad \forall \nu \in \mathbb{Z}$. □

Für periodische Funktionen bzw. periodische Fortsetzungen von auf einem Intervall definierten quadratintegralen Funktionen hat man somit in der Fourierentwicklung ein einfaches formales Hilfsmittel zur Beschreibung. Dieses Hilfsmittel werden wir dann auf Funktionen $(C(h))^k(u_0)$ mit $u_0 \in B$ anwenden.

Sind die Funktionen nicht periodisch, dann tritt an die Stelle der Fourierentwicklung die Fouriertransformation.

Rein formal ist die Fouriertransformation einer quadratintegralen Funktion f definiert durch

$$g(u) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-iu x) f(x) dx \quad (\stackrel{\text{def}}{=} \mathfrak{F}_n(f)(u)) \quad f : \mathbb{R} \rightarrow \mathbb{R}^n$$

Für jede Stelle x , an der f endlich ist, gilt dann

$$f(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ixu) g(u) du \quad (\stackrel{\text{def}}{=} \mathfrak{F}_n^{-1}(g)(x))$$

Genauer:

Satz 3.14 (Fouriertransformation) Es sei $f \in L_2(\mathbb{R}, \mathbb{C}^n)$ und

$$\begin{aligned} g_\mu(y) &\stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\mu}^{\mu} f(x) \exp(-iyx) \, dx \\ f_\mu(y) &\stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\mu}^{\mu} g(x) \exp(iyx) \, dx \end{aligned} \quad \mu \in \mathbb{N}, \quad y \in \mathbb{R}$$

Dann sind die Folgen g_μ und f_μ Cauchyfolgen in $L_2(\mathbb{R}, \mathbb{C}^n)$ mit Grenzwerten g und f . Die Zuordnung $f \mapsto g$ und $g \mapsto f$ sind normtreue Automorphismen von $L_2(\mathbb{R}, \mathbb{C}^n)$ auf sich. Die zweite Abbildung ist die Umkehrung der ersten.

Wenn f bzw. g kompakten Träger haben, konvergieren die Folgen auch punktweise, sonst im allgemeinen nur im Mittel. \square

Für die Fourierentwicklung und die Fouriertransformation gelten einige nützliche Rechenregeln:

Satz 3.15 Für $\Delta x \in \mathbb{R}_+$ sei $\varepsilon_{\Delta x} : \mathbb{R} \rightarrow \mathbb{C}$ erklärt durch $\varepsilon_{\Delta x}(x) \stackrel{\text{def}}{=} \exp(ix\Delta x)$.

Dann gilt:

$$\begin{aligned} \mathfrak{F}_{2\pi,n}(T_{\Delta x}(f))(\nu) &= \varepsilon_{\Delta x}(\nu) \mathfrak{F}_{2\pi,n}(f)(\nu) \quad (\nu \in \mathbb{Z}), \quad f \in L_2((0, 2\pi), \mathbb{C}^n) \\ \mathfrak{F}_n(T_{\Delta x}(f))(u) &= \varepsilon_{\Delta x}(u) \mathfrak{F}_n(f)(u) \quad (f \in L_2(\mathbb{R}, \mathbb{C}^n)) \end{aligned}$$

\square

Bemerkung 3.9 Man kann völlig analog auch Fourierentwicklungen und -Transformationen für mehrere Raumvariablen betrachten, doch werden wir dies hier nicht benutzen. \square

Wir betrachten nun zu einer linearen Anfangswertaufgabe eine (implizite) Differenzenapproximation der Form

$$C(h) = \left(\sum_{\nu=-k}^k B_\nu(x, h) T_{\Delta x}^\nu \right)^{-1} \cdot \left(\sum_{\nu=-k}^k A_\nu(x, h) T_{\Delta x}^\nu \right)$$

Dabei sind B_ν, A_ν reelle $n \times n$ -Matrizen und mit $\lambda \in \mathbb{R}_+$ fest.

$$\begin{aligned} \Delta x &= \frac{h}{\lambda} \quad (\text{hyperbolischer Fall}) \\ \Delta x &= \sqrt{\frac{h}{\lambda}} \quad (\text{parabolischer Fall}) \end{aligned}$$

Bei der Anwendung der Fouriertransformation denken wir uns die Variable x in den Koeffizienten "eingefroren". Die Rechtfertigung dafür liefert der später angeführte Satz von Lax-Nirenberg.

Definition 3.16 Die durch die Vorschrift

$$G(h, y, x) := \left(\sum_{\nu=-k}^k \exp(i\nu y \Delta x) B_\nu(x, h) \right)^{-1} \cdot \left(\sum_{\nu=-k}^k \exp(i\nu y \Delta x) A_\nu(x, h) \right)$$

definierte $n \times n$ -Matrix heißt die dem Differenzenverfahren

$$M_D = \{C(h) : 0 < h \leq h_0\}$$

zugeordnete **Amplifikationsmatrix**. Dabei ist

$$x \in \mathcal{I} = \mathbb{R} \quad \text{oder} \quad (0, 2\pi) \quad \text{oder} \quad (0, \pi)$$

$$y \in \mathbb{F} = \begin{cases} \mathbb{R} & \text{für } \mathcal{I} = \mathbb{R} \\ \mathbb{Z} & \text{für } \mathcal{I} = (0, 2\pi) \\ \mathbb{N} & \text{für } \mathcal{I} = (0, \pi) \end{cases}$$

Sind die Koeffizientenmatrizen des Differenzenverfahren alle x -unabhängig, so schreibt man auch kurz $G(h, y)$ statt $G(h, y, x)$. \square

Satz 3.16 Ein Differenzenverfahren mit ortsunabhängigen Koeffizienten ist genau dann stabil, wenn es eine Konstante $K \in \mathbb{R}_+$ gibt mit

$$\|G(h, y)^\nu\|_2 \leq K \quad 0 < \nu h \leq T, \quad \nu \in \mathbb{N}, \quad h \leq h_0, \quad y \in \mathbb{F}. \quad (3.4)$$

Beweis:

1.Fall $B = L_2(\mathbb{R}, \mathbb{C}^n)$, $\mathcal{I} = \mathbb{R}$, $\mathbb{F} = \mathbb{R}$.
(Anfangswertproblem auf der reellen Achse)

a. (3.4) ist hinreichend.

Sei $g \stackrel{\text{def}}{=} (C(h))(f)$ mit $f \in B$. Dann folgt wegen der Ortsunabhängigkeit

$$\begin{aligned} \mathfrak{F}_n \left(\left(\sum_{\nu=-k}^k B_\nu(h) T_{\Delta x}^\nu \right) g(\cdot) \right) &= \sum_{\nu=-k}^k B_\nu(h) \varepsilon_{\Delta x}^\nu(\cdot) \mathfrak{F}_n(g)(\cdot) \\ &= \mathfrak{F}_n \left(\left(\sum_{\nu=-k}^k A_\nu(h) T_{\Delta x}^\nu \right) f(\cdot) \right) \\ &= \sum_{\nu=-k}^k A_\nu(h) \underbrace{\varepsilon_{\Delta x}^\nu(\cdot)}_{\exp(i(\cdot)\nu\Delta x)} \mathfrak{F}_n(f)(\cdot) \end{aligned}$$

also

$$\mathfrak{F}_n(g) = G(h, \cdot) \mathfrak{F}_n(f)$$

und entsprechend

$$\mathfrak{F}_n\left((C(h))^\nu(f)\right) = G^\nu(h, \cdot)\mathfrak{F}_n(f)$$

Da \mathfrak{F}_n normtreu ist, folgt

$$\begin{aligned} \|(C(h))^\nu(f)\| &= \|\mathfrak{F}_n\left((C(h))^\nu(f)\right)\| \\ &= \|G^\nu(h, \cdot)\mathfrak{F}_n(f)\| \\ &\leq \sup_{y \in \mathbb{R}} \|G^\nu(h, y)\| \cdot \|\mathfrak{F}_n(f)\| \leq K\|f\| \end{aligned}$$

und da f beliebig war, folgt die Behauptung.

b. (3.4) ist notwendig.

Indirekter Beweis: Es existiere für beliebiges $K \in \mathbb{R}_+$ ein $w \in \mathbb{R}$, ein $h_1 \in]0, h_0[$, ein $l \in \mathbb{N}$ mit $h_1 l \leq T$, so daß

$$\|G(h_1, w)^l\|_2 > K.$$

Sei

$$S(y) \stackrel{\text{def}}{=} G(h_1, y)^l, \quad \text{d.h.} \quad S(w) = G(h_1, w)^l$$

und

$$\lambda \stackrel{\text{def}}{=} \|S(w)\|_2 > K$$

Nun gilt:

$$(a) \exists v \in \mathbb{C}^n \text{ mit } v^H S^H(w) S(w) v = \lambda^2 v^H v$$

$$(b) \exists \hat{f} \in L_2(\mathbb{R}, \mathbb{C}^n), \quad \hat{f} \text{ stetig mit } \hat{f}(w) = v$$

d.h.

$$\hat{f}(w)^H S^H(w) S(w) \hat{f}(w) > K^2 \hat{f}(w)^H \hat{f}(w)$$

Da \hat{f} stetig sein soll $\exists [w_0, w_1]$ mit $w_0 < w < w_1$ und

$$\hat{f}(u)^H S^H(u) S(u) \hat{f}(u) > K^2 \hat{f}(u)^H \hat{f}(u) \quad \forall u \in [w_0, w_1]$$

Sei

$$f(y) \stackrel{\text{def}}{=} \begin{cases} \hat{f}(y) & y \in [w_0, w_1] \\ 0 & \text{sonst} \end{cases}$$

$$g \stackrel{\text{def}}{=} \mathfrak{F}_n^{-1}(f)$$

Nach Konstruktion ist

$$\|S(\cdot)f\|_{L_2}^2 > K^2 \|f\|_{L_2}^2$$

Somit

$$\begin{aligned} K\|g\| &= K\|f\| < \|S(\cdot)f\| = \|G(h_1, \cdot)^l \mathfrak{F}_n(g)\| \\ &= \|\mathfrak{F}_n(C(h_1)^l g)\| = \|(C(h_1))^l g\| \end{aligned}$$

Das Differenzenverfahren kann also nicht stabil sein.

2. Fall: $B = L_2((0, 2\pi), \mathbb{C}^n)$, $\mathbb{F} = \mathbb{Z}$ bzw. $B = L_2((0, \pi), \mathbb{C}^n)$, $\mathbb{F} = \mathbb{N}$.

Dann ist

$$\begin{aligned}\mathfrak{F}_{2\pi,n}(T_{\Delta x}^j(f))(\nu) &= \exp(i\nu j \Delta x) \mathfrak{F}_{2\pi,n}(f)(\nu) \\ \sum_{j=-k}^k B_j(h) \mathfrak{F}_{2\pi,n}(T_{\Delta x}^j(f))(\nu) &= \mathfrak{F}_{2\pi,n}\left(\sum_{j=-k}^k B_j(h) T_{\Delta x}^j(f)\right)(\nu) \\ &= \sum_{j=-k}^k B_j(h) \exp(i\nu j \Delta x) \mathfrak{F}_{2\pi,n}(f)(\nu)\end{aligned}$$

Analog

$$\mathfrak{F}_{2\pi,n}\left(\sum_{j=-k}^k A_j(h) T_{\Delta x}^j(g)\right)(\nu) = \sum_{j=-k}^k A_j(h) \exp(i\nu j \Delta x) \mathfrak{F}_{2\pi,n}(g)(\nu)$$

Definiert man also f durch

$$f \stackrel{def}{=} (C(h))(g)$$

dann ist

$$\begin{aligned}\mathfrak{F}_{2\pi,n}(f)(\nu) &= \left(\sum_{j=-k}^k B_j(h) \exp(i\nu j \Delta x)\right)^{-1} \left(\sum_{j=-k}^k A_j(h) \exp(i\nu j \Delta x)\right) \mathfrak{F}_{2\pi,n}(g)(\nu) \\ &= G(h, \nu) \mathfrak{F}_{2\pi,n}(g)(\nu)\end{aligned}$$

und entsprechend für

$$\begin{aligned}f &\stackrel{def}{=} (C(h))^m(g) \\ \mathfrak{F}_{2\pi,n}(f)(\nu) &= G(h, \nu)^m \mathfrak{F}_{2\pi,n}(g)(\nu)\end{aligned}$$

Somit

$$\begin{aligned}\|f\|^2 &= \sum_{\nu \in \mathbb{F}} \|\mathfrak{F}_{2\pi,n}(f)(\nu)\|^2 = \sum_{\nu \in \mathbb{F}} \|G(h, \nu)^m \mathfrak{F}_{2\pi,n}(g)(\nu)\|^2 \\ &\leq \sup_{\nu \in \mathbb{F}} \|G(h, \nu)^m\|^2 \sum_{\nu \in \mathbb{F}} \|\mathfrak{F}_{2\pi,n}(g)(\nu)\|^2 \\ &\leq \left(\sup_{\nu \in \mathbb{F}} \|G(h, \nu)^m\|\right)^2 \|g\|^2\end{aligned}$$

d.h. aus der gleichmäßigen Beschränktheit von $\|G(h, \nu)^m\|$ folgt die Stabilität.

Umgekehrt existiert zu jedem ν und m mit $hm \leq T$ ein $g_{\nu,m} \in \mathbb{C}^n$ mit

$$\|G(h, \nu)^m g_{\nu,m}\| = \|G(h, \nu)^m\|, \quad \|g_{\nu,m}\| = 1$$

Sei

$$g(x) \stackrel{def}{=} \frac{1}{\sqrt{2\pi}} g_{\nu,m} \exp(i\nu x)$$

Dann wird

$$\|g\| = \|g_{\nu,m}\| = 1$$

und

$$\begin{aligned} \|C(h)^m\| &\geq \|C(h)^m(g)\| = \|f\| = \mathfrak{F}_{2\pi,n}(f)(\nu) \\ &= \|G(h,\nu)^m \mathfrak{F}_{2\pi,n}(g)(\nu)\| = \|G(h,\nu)^m g_{\nu,m}\| = \|G(h,\nu)^m\| \end{aligned}$$

d.h. aus der Unbeschränktheit von $\|G(h,\nu)^m\|$ folgt die Instabilität des Verfahrens M_D . \square

Bemerkung 3.10 *Aufgrund der Einbettung durch periodische Fortsetzung ist ein Verfahren, das in $L_2(\mathbb{R}, \mathbb{C}^n)$ stabil ist, auch in $L_2((0, 2\pi), \mathbb{C}^n)$ bzw. $L_2((0, \pi), \mathbb{C}^n)$ stabil. (Die Umkehrung gilt nicht, doch sind diese pathologischen Fälle ohne praktisches Interesse). Man untersucht deshalb die Stabilität der Verfahren stets anhand der Matrizen $G(h, y)$ mit $y \in \mathbb{R}$. \square*

Die Frage der Stabilität von Differenzenverfahren ist damit zurückgeführt auf die Frage der Beschränktheit von Matrizenpotenzen in der euklidischen Norm, die immer noch schwierig genug zu behandeln ist. Zunächst ein einfaches notwendiges Kriterium:

Satz 3.17 (von Neumann-Bedingung)

Es sei M_D ein Differenzenverfahren mit ortsunabhängigen Koeffizienten. Es seien

$$\begin{array}{ll} \lambda_j(h, y) & \text{die Eigenwerte von } G(h, y) \\ \varrho(h, y) = \max_{j=1(1)n} |\lambda_j(h, y)| & \text{der Spektralradius von } G(h, y). \end{array}$$

Notwendig für die Stabilität des Differenzenverfahrens ist

$$\varrho(h, y) \leq 1 + Kh, \quad (\forall y, \quad 0 < h \leq h_0)$$

Beweis: M_D sei stabil und somit nach Satz 3.16

$$\varrho^m(h, y) \leq \|G(h, y)^m\| \leq \tilde{K} \quad \forall y, \forall m, h: \quad mh \leq T$$

Somit

$$\varrho(h, y) \leq \tilde{K}^{1/m}$$

o.B.d.A. kann $\tilde{K} \geq 1$ angenommen werden. Da m maximal gewählt werden kann mit $mh \leq T$, d.h. $mh > T - h$, kann man weiter abschätzen

$$\varrho(h, y) \leq \tilde{K}^{\frac{h}{mh}} \leq \tilde{K}^{\frac{h}{T-h}} \stackrel{\text{def}}{=} f(h)$$

Es ist $f(0) = 1$, f konvex auf $]0, T[$, wie man durch direkten Nachweis von $f'' \geq 0$ sieht, somit

$$f(h) \leq 1 + \underbrace{\frac{f(h_0) - 1}{h_0}}_K h \quad \text{für } 0 < h \leq h_0 \quad \square$$

In speziellen Fällen ist die von Neumann-Bedingung auch hinreichend, nämlich immer dann, wenn der Spektralradius selbst eine Norm ist:

Satz 3.18 *Es sei M_D ein konsistentes Differenzenverfahren mit ortsunabhängigen Koeffizienten. Es seien $G(h, y)$ die Amplifikationsmatrix zu M_D und*

$$\begin{aligned} \lambda_j(h, y) & \text{ die Eigenwerte von } G(h, y) \\ \varrho(h, y) & \text{ der Spektralradius von } G(h, y). \end{aligned}$$

und

$$\varrho(h, y) \leq 1 + Kh \quad \text{für alle } y \in \mathbb{F} \text{ und } 0 < h \leq h_0$$

Ferner gelte:

1.

$$\begin{aligned} \|G(h, y)\|_\infty & \leq K^* \\ |\lambda_j(h, y)| & \leq \mu < 1 \quad \forall y \in \mathbb{F}, \quad 0 < h \leq h_0, \quad j = 2, \dots, n \end{aligned}$$

oder

2.

$$GG^H = G^H G \quad (\forall y \in \mathbb{F}, \quad 0 < h \leq h_0)$$

oder

3. Die Matrizen $A_\nu(h), B_\nu(h)$ sind simultan diagonalisierbar, die Transformationsmatrizen $T(h)$ und $T^{-1}(h)$ seien gleichmäßig beschränkt in $]0, h_0]$.

Dann ist M_D stabil in $L_2(\mathbb{R})$.

Beweis:

Zu (1):

Zunächst wird $G(h, y)$ auf Schur-Normalform transformiert. Ist G eine $m \times m$ -Matrix, dann existiert $U \in \mathbb{C}^{m \times m}$ unitär mit

$$U^H(h, y) G(h, y) U(h, y) = R(h, y)$$

$R(h, y)$ obere Dreiecksmatrix. Auf der Diagonalen von R treten also die Eigenwerte von G auf, o.B.d.A. mit der Numerierung

$$\begin{aligned} \lambda_1 & = R_{11} = 1 + \mathcal{O}(h) \\ \lambda_j & = R_{jj}, \quad j = 2, \dots, m. \end{aligned}$$

Es sei h_0 so klein, daß $\lambda_1 > (1 + \mu)/2$.

Der nächste Transformationsschritt besteht nun in der Herstellung einer Block-Dreiecksmatrix.

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \boxed{} \\ \vdots & & & \\ \vdots & & & \\ 0 & & & \end{pmatrix}.$$

Setze dazu

$$R = \begin{pmatrix} \lambda_1 & r^H \\ 0 & \hat{R} \end{pmatrix}$$

$$\lambda_1 x^H - x^H \hat{R} = -r^H,$$

d.h.

$$x^H = -r^H (\lambda_1 I - \hat{R})^{-1}.$$

Definiere

$$T = \begin{pmatrix} 1 & x^H \\ 0 & I_{m-1} \end{pmatrix}.$$

Es ist

$$T^{-1} = \begin{pmatrix} 1 & -x^H \\ 0 & I_{m-1} \end{pmatrix}$$

(alle Matrizenkoeffizienten hängen von h und y ab!)

Dann wird

$$\begin{aligned} T^{-1}RT &= \begin{pmatrix} 1 & -x^H \\ 0 & I_{m-1} \end{pmatrix} \begin{pmatrix} \lambda_1 & \lambda_1 x^H + r^H \\ 0 & \hat{R} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \lambda_1 x^H + r^H - x^H \hat{R} \\ 0 & \hat{R} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & 0 \\ 0 & \hat{R} \end{pmatrix}. \end{aligned}$$

Im ganzen haben wir dann

$$T^{-1}(h, y) U^H(h, y) G(h, y) U(h, y) T(h, y) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \hat{R} \end{pmatrix}.$$

Weil die Elemente unitärer Matrizen betragsmässig kleinergleich 1 sind, gilt für die Außerdiagonalelemente von \hat{R} die Abschätzung

$$|\hat{R}_{ij}(h, y)| \leq m^2 \max_{i,j} |(G(h, y))_{i,j}| \leq m^2 K^* \stackrel{def}{=} K^{**},$$

$\forall h \in]0, h_0]$, $i = 2, \dots, m$, $j = i+1, \dots, m$. Die gleiche Abschätzung gilt auch für die Elemente von r^H .

Für die Elemente von x gilt wegen der rekursiven Berechnung (mit $x = (\xi_2, \dots, \xi_m)^T$)

$$\begin{aligned} (\bar{\lambda}_1 - \hat{R}^H)x &= -r \\ |\xi_2| &\leq K^{**}/\left(\frac{1+\mu}{2} - \mu\right) = 2K^{**}/(1-\mu) \\ |\xi_j| &\leq K^{**}\left(1 + \sum_{i=2}^{j-1} |\xi_i|\right)/\left(\frac{1+\mu}{2} - \mu\right) \quad j = 3, \dots, m \end{aligned}$$

und wegen $\mu < 1$ die universelle Abschätzung

$$|\xi_j| \leq m! 2^{m-1}/(1-\mu)^{m-1} K^{**} \stackrel{def}{=} K^{***}, \quad j = 2, \dots, m.$$

Somit gilt

$$\|T(h, y)\|_1 \|T^{-1}(h, y)\|_1, \quad \|T(h, y)\|_\infty, \quad \|T^{-1}(h, y)\|_\infty \leq K^{(4)} \stackrel{def}{=} K^{***} + 1.$$

Sei schließlich mit $\delta \stackrel{def}{=} \frac{1-\mu}{2(m-1)(K^{**}+1)} < 1 \quad (m \geq 2)$

$$D = \text{diag}\left(1, \delta, \delta^2, \dots, \delta^{m-1}\right).$$

Dann ist

$$D^{-1}T^{-1}U^H G U T D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{R} & \\ 0 & & & \end{pmatrix}$$

und

$$\hat{R}_{ij} = \hat{R}_{ij} \delta^{j-i}$$

somit

$$\|\hat{R}\|_1, \quad \|\hat{R}\|_\infty \leq \mu + (m-2)K^{**}\delta = \mu + \frac{1}{2} \frac{m-2}{m-1} \frac{K^{**}}{K^{**}+1} (1-\mu) < \frac{1+\mu}{2} < 1,$$

d.h. mit

$$\begin{aligned} A &\stackrel{def}{=} D^{-1}T^{-1}U^H G U T D \\ \|A\|_\infty &= |\lambda_1| \leq 1 + Kh, \\ \|A\|_1 &= |\lambda_1| \end{aligned}$$

d.h.

$$\|A^n\|_\infty, \quad \|A^n\|_1 \leq \exp(KT)$$

für $nh \leq T, \quad 0 < h \leq h_0, \quad n \in \mathbb{N}$.

Dies ergibt

$$\|A^n\|_2 \leq \left(\|A^n\|_\infty \|A^n\|_1\right)^{1/2} \leq \exp(KT)$$

und wegen

$$\begin{aligned} \|G(h, y)^n\|_2 &\leq \|A^n\|_2 \text{cond}_{\|\cdot\|_2}(U) \text{cond}_{\|\cdot\|_2}(T) \text{cond}_{\|\cdot\|_2}(D) \\ \text{cond}_{\|\cdot\|_2}(U) &= 1 \\ \text{cond}_{\|\cdot\|_2}(D) &= \delta^{1-m} \\ \text{cond}_{\|\cdot\|_2}(T) &\leq \left(\text{cond}_{\|\cdot\|_\infty}(T) \text{cond}_{\|\cdot\|_1}(T)\right)^{1/2} \leq K^{(4)} \end{aligned}$$

ergibt sich

$$\|G(h, y)^n\|_2 \leq \exp(KT)K^{(4)}\delta^{1-m}$$

$\forall y \in \mathbb{F}, \quad 0 < h \leq h_0, \quad n \in \mathbb{N}: nh \leq T.$

Zu (2):

G normal $\iff G$ unitär diagonalisierbar.

$$\begin{aligned} \|G^n\|_2 &= \|U^H G^n U\|_2 = \|(U^H G U)^n\|_2 = \|\Lambda_G^n\|_2 \\ |\lambda_1^n| &\leq \exp(KT) \end{aligned}$$

für $0 \leq nh \leq T, \quad 0 < h \leq h_0, \quad n \in \mathbb{N}, \quad y \in \mathbb{F}.$

Zu (3):

Sei $T(h)$ die Matrix, die die $A_\nu(h), B_\nu(h)$ simultan diagonalisiert.

$\Lambda_{A_\nu}(h), \Lambda_{B_\nu}(h), \Lambda_G(h)$ seien die entsprechenden Diagonalmatrizen. Dann ist

$$\begin{aligned} \|G^n(h, y)\|_2 &= \|T(h)(T^{-1}(h)G(h, y)T(h))^n T^{-1}(h)\|_2 \\ &\leq \text{cond}_{\|\cdot\|_2}(T(h)) |\lambda_1^n(h, y)| \leq \tilde{K} \exp(KT), \end{aligned}$$

da $\text{cond}_{\|\cdot\|_2}(T(h)) \leq \tilde{K}$ für $0 \leq h \leq h_0$ nach Voraussetzung. □

Beispiel 3.12

$$\begin{aligned} u_t &= a^2 u_{xx} + bu_x + cu, \quad 0 \leq t \leq T, \quad 0 \leq x \leq 2\pi \\ \left(\begin{aligned} u(x, 0) &= u_0(x) \in L_2([0, 2\pi]) \end{aligned} \right) \\ \lambda &\stackrel{\text{def}}{=} \frac{\Delta t}{(\Delta x)^2} \quad \text{fest.} \quad h = \Delta t \end{aligned}$$

Explizites Verfahren:

$$\begin{aligned} u_{n+1}(x) &= u_n(x) + \lambda a^2 (T_{\Delta x}^{-1} - 2I + T_{\Delta x})u_n(x) + \frac{b}{2}\sqrt{h\lambda}(T_{\Delta x} - T_{\Delta x}^{-1})u_n(x) + chu_n(x) \\ G(h, y) &= (1 + ch - 2\lambda a^2) + 2\lambda a^2 \cos(y\Delta x) + b\sqrt{h\lambda} i \sin(\Delta x \cdot y) \\ &\in [1 + ch - 4\lambda a^2, 1 + ch] \times [-ib\sqrt{h\lambda}, ib\sqrt{h\lambda}] \end{aligned}$$

G skalar, also normal

$$\begin{aligned} \varrho^2(G) &= (1 + 2\lambda a^2(\cos(y\Delta x) - 1) + ch)^2 + b^2 h \lambda \\ &= (1 + 2\lambda a^2(\cos(y\Delta x) - 1))^2 + \mathcal{O}(h) \\ &\leq 1 + \mathcal{O}(h) \quad \text{für } \lambda \leq 1/(2a^2) \end{aligned}$$

(die Terme bu_x und cu sind also für die Stabilität (dies ist eine asymptotische Eigenschaft!) unerheblich, aber natürlich bei konkreten Werten von $h = \Delta t$ und Δx nicht für die Genauigkeit) □

Beispiel 3.13

$$u_t = u_{xx}, \quad 0 \leq t \leq T, \quad \Delta x = \sqrt{h/\lambda}$$

Naive Diskretisierung der Konsistenz $\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)$:

$$u_{n+1}(x) = u_{n-1}(x) + 2\lambda(T_{\Delta x} - 2I + T_{\Delta x}^{-1})u_n(x)$$

Als Einschrittverfahren geschrieben lautet dies

$$\begin{pmatrix} u_{n+1} \\ u_n \end{pmatrix} (x) = \begin{pmatrix} 2\lambda(T_{\Delta x} - 2I + T_{\Delta x}^{-1}) & I \\ I & 0 \end{pmatrix} \begin{pmatrix} u_n \\ u_{n-1} \end{pmatrix} (x)$$

$$G(h, y) = \begin{pmatrix} 4\lambda(\cos(y\Delta x) - 1) & 1 \\ 1 & 0 \end{pmatrix}$$

Für $\cos(y\Delta x) \neq 1$ erfüllen die beiden Eigenwerte λ_1, λ_2 von G die Bedingungen

$$\lambda_1 \lambda_2 = -1, \quad \lambda_1 + \lambda_2 = 4\lambda(\cos(y\Delta x) - 1) \in [-8\lambda, 0] < 0,$$

das Verfahren kann also für keinen Wert von λ stabil sein. □

Beispiel 3.14

$$u_t = u_{xx}, \quad 0 \leq t \leq T, \quad \Delta x = \sqrt{h/\lambda}$$

Verfahren von du Fort und Frankel:

Als Einschrittverfahren geschrieben

$$\begin{pmatrix} u_{n+1} \\ u_n \end{pmatrix} (x) = \begin{pmatrix} \frac{2\lambda}{1+2\lambda}(T_{\Delta x} + T_{\Delta x}^{-1}) & \frac{1-2\lambda}{1+2\lambda}I \\ I & 0 \end{pmatrix} \begin{pmatrix} u_n \\ u_{n-1} \end{pmatrix} (x)$$

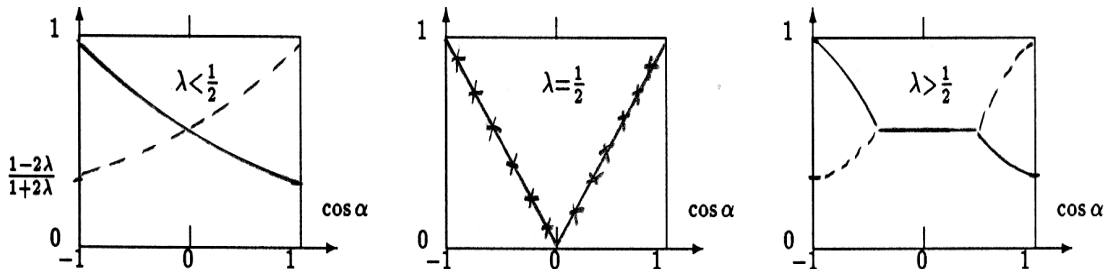
Es wird

$$G(h, y) = \begin{pmatrix} \frac{4\lambda}{1+2\lambda} \cos(y\Delta x) & \frac{1-2\lambda}{1+2\lambda} \\ 1 & 0 \end{pmatrix}$$

Die Eigenwerte λ_1 und λ_2 von $G(h, y)$ sind (mit $\alpha \stackrel{\text{def}}{=} y\Delta x = y\sqrt{h/\lambda}$)

$$\lambda_{1,2} = \frac{2\lambda \cos \alpha \pm \sqrt{1 - 4\lambda^2 \sin^2 \alpha}}{1 + 2\lambda}$$

Als Funktionen von $\cos \alpha$ verlaufen $|\lambda_1|, |\lambda_2|$ wie folgt:



Wegen Satz 3.18 (1) folgt somit die Stabilität des Verfahrens für alle Werte von $\lambda > 0$ \square

Beispiel 3.15 $u_t = Au_x$ $A \neq 0$ reell symmetrisch

$$\lambda \stackrel{\text{def}}{=} h/\Delta x$$

Naïves Differenzenverfahren:

$$\begin{aligned} u_{n+1}(x) &= \left(I + \frac{\lambda}{2} A (T_{\Delta x} - T_{\Delta x}^{-1}) \right) u_n(x) \\ G(h, y) &= I + i\lambda A \sin \alpha, \quad \alpha = y\Delta x = yh/\lambda \end{aligned}$$

G ist normal, da A reell symmetrisch ist, und

$$\varrho(G(h, y)) = (1 + \lambda^2 \varrho^2(A) \sin^2 \alpha)^{1/2}$$

Wegen $\sin^2(\alpha) = 1$ für geeignetes y und $\varrho(A) > 0$ liegt somit Stabilität genau für

$$\lambda^2 = \mathcal{O}(h)$$

vor, d.h. wir haben eine extreme, dem Problem nicht angepasste, Schrittweitenbegrenzung. \square

Beispiel 3.16 $u_t = Au_x$, $A \neq 0$ reell symmetrisch

$$\lambda = h/\Delta x \quad \text{fest.}$$

Friedrichs-Verfahren:

$$\begin{aligned} u_{n+1}(x) &= \left(\frac{1}{2}(T_{\Delta x} + T_{\Delta x}^{-1}) + \frac{\lambda}{2} A (T_{\Delta x} - T_{\Delta x}^{-1}) \right) u_n(x) \\ G(h, y) &= \cos \alpha \cdot I + i\lambda \sin \alpha \cdot A, \quad \alpha = y\Delta x \end{aligned}$$

G ist normal. Ferner gilt

$$\varrho^2(G) = \cos^2 \alpha + \lambda^2 \sin^2 \alpha \varrho^2(A).$$

Stabilität liegt also vor für

$$\lambda \leq 1/\varrho(A),$$

ein sehr viel besseres Resultat als in Beispiel 3.15. \square

Beispiel 3.17 $u_t = Au_x$ A $n \times n$ reell diagonalisierbar.

$$\lambda \stackrel{\text{def}}{=} h/\Delta x \quad \text{fest.}$$

Lax–Wendroff–Verfahren:

$$C(h) = I + \frac{1}{2}\lambda A(T_{\Delta x} - T_{\Delta x}^{-1}) + \frac{1}{2}\lambda^2 A^2(T_{\Delta x} - 2I + T_{\Delta x}^{-1})$$

Amplifikationsmatrix $G(h, y)$:

$$G(h, y) = I + i\lambda \sin \alpha \cdot A - \lambda^2 A^2(1 - \cos \alpha), \quad \alpha = y\Delta x$$

Eigenwerte von $G(h, y)$:

$$\lambda_j = 1 - \lambda^2 \mu_j^2(1 - \cos \alpha) + i\lambda(\sin \alpha)\mu_j, \quad \mu_j = \text{Eigenwerte von } A$$

Wegen Satz 3.18 (3) liefert die von Neumann-Bedingung eine hinreichende und notwendige Stabilitätsbedingung:

$$\begin{aligned} |\lambda_j|^2 &= \lambda^2 \sin^2 \alpha \mu_j^2 + (1 - \lambda^2 \mu_j^2(1 - \cos \alpha))^2 \\ &= 4\lambda^2 \sin^2\left(\frac{\alpha}{2}\right) \cos^2\left(\frac{\alpha}{2}\right) \mu_j^2 + \left(1 - \lambda^2 \mu_j^2 \left(\cos^2\left(\frac{\alpha}{2}\right) + \sin\left(\frac{\alpha}{2}\right) - (\cos^2\left(\frac{\alpha}{2}\right) - \sin^2\left(\frac{\alpha}{2}\right))\right)\right)^2 \\ &= 4\lambda^2 \sin^2\left(\frac{\alpha}{2}\right) \mu_j^2 - 4\lambda^2 \sin^4\left(\frac{\alpha}{2}\right) \mu_j^2 + 1 - 4\lambda^2 \sin^2\left(\frac{\alpha}{2}\right) \mu_j^2 + 4\lambda^4 \mu_j^4 \sin^4\left(\frac{\alpha}{2}\right) \mu_j^2 \\ &= 1 - 4\lambda^2 \mu_j^2 \sin^4\left(\frac{\alpha}{2}\right)(1 - \lambda^2 \mu_j^2) \leq 1 \end{aligned}$$

für

$$\lambda \leq \frac{1}{\varrho(A)}$$

□

In Satz 3.18 geht man davon aus, daß eine Abschätzung

$$\varrho(G(h, y)) \leq 1 + Kh$$

unmittelbar möglich ist. Selbst wenn dies erfüllt ist, kann der direkte Nachweis doch sehr schwierig sein. Es ist deshalb nützlich, auch andere hinreichende Kriterien für die Beschränktheit von $\|G^n(h, y)\|$ zu besitzen. Eines liefert der folgende Satz von Lax und Wendroff: (Comm. on pure and applied Math. 17, (1964), 381-398)

Satz 3.19 Es sei $A \in \mathbb{C}^{n \times n}$ und mit

$$\begin{aligned} R(x; A) &\stackrel{\text{def}}{=} \frac{x^H A x}{x^H x} \\ |R(x; A)| &\leq 1 \quad x \in \mathbb{C}^n, \quad x \neq 0 \end{aligned}$$

Dann existiert eine Konstante $K(n)$ mit

$$\|A^m\|_2 \leq K(n) \quad \forall m \in \mathbb{N}.$$

(Der Beweis wird über die Schursche Transformation auf Dreiecksgestalt und zunächst für $n = 2$ geführt. Dann Induktion über die Dimension n .) □

Satz 3.20 *Lax–Wendroff–Bedingung:*

Hinreichend für die Stabilität eines Differenzenverfahrens (für eine AWA mit konstanten Koeffizienten) ist

$$|R(x; G(h, y))| \leq 1 + Kh \quad \forall x \in \mathbb{C}^n, \quad x \neq 0$$

Beweis: Man setze $A \stackrel{\text{def}}{=} e^{-Kh} G(h, y)$ und wende Satz 3.19 an:

$$R(x; A) \leq e^{-Kh}(1 + Kh) \leq 1.$$

Damit

$$\|A^m\|_2 \leq K^*(n) \quad \forall m \in \mathbb{N}$$

und deshalb

$$\|G(h, y)^m\|_2 \leq e^{Khm} K^*(n) \leq e^{KT} K^*(n)$$

□

Eine Anwendung dieses Satzes liefert

Beispiel 3.18 *Hyperbolisches System in zwei Ortsvariablen:*

$$\begin{aligned} u_t &= Au_x + Bu_y \\ u &= u(x, y, t) \quad G \subset \mathbb{R}^2 \times [0, T] \rightarrow \mathbb{R}^n \\ A, B &\text{ reell symmetrisch} \end{aligned}$$

Differenzenverfahren:

$$\begin{aligned} u_{n+1}(x, y) &= \left(\frac{1}{4}(T_{\Delta x} + T_{\Delta x}^{-1} + T_{\Delta y} + T_{\Delta y}^{-1}) \right. \\ &\quad \left. + \frac{1}{2}(\lambda_1 A(T_{\Delta x} - T_{\Delta x}^{-1}) + \lambda_2 B(T_{\Delta y} - T_{\Delta y}^{-1})) \right) u_n(x, y) \end{aligned}$$

(dies ist die unmittelbare Übertragung des Friedrich-Verfahrens auf 2 räumliche Variablen)
und

$$\lambda_1 = \frac{h}{\Delta x}, \quad \lambda_2 = \frac{h}{\Delta y}, \quad h \hat{=} \Delta t$$

Die Amplifikationsmatrix G hängt nun von zwei Parametern r, s ab und

$$\alpha_1 = r\Delta x, \quad \alpha_2 = s\Delta y$$

$$G(h, r, s) = \frac{\cos \alpha_1 + \cos \alpha_2}{2} I + i(\lambda_1 \sin \alpha_1 A + \lambda_2 \sin \alpha_2 B).$$

Da A und B reell symmetrisch sind, ist G normal und $u^H A u, u^H B u \in \mathbb{R}$. Also

$$\begin{aligned} |R(u; G(h, r, s))|^2 &\leq \frac{1}{4}(\cos^2 \alpha_1 + 2|\cos \alpha_1 \cos \alpha_2| + \cos^2 \alpha_2) \\ &\quad + \lambda_1^2 \sin^2 \alpha_1 \varrho^2(A) + 2\lambda_1 \lambda_2 |\sin \alpha_1 \sin \alpha_2| \varrho(A) \varrho(B) + \lambda_2^2 \sin^2 \alpha_2 \varrho^2(B). \end{aligned}$$

Für

$$\lambda_1 \leq \frac{1}{2\varrho(A)}, \quad \lambda_2 \leq \frac{1}{2\varrho(B)}$$

folgt

$$|R(u; G(h, r, s))|^2 \leq \frac{1}{2}(\cos^2 \alpha_1 + \cos^2 \alpha_2 + \sin^2 \alpha_1 + \sin^2 \alpha_2) = 1$$

(weil $2|\cos \alpha_1 \cos \alpha_2| \leq \cos^2 \alpha_1 + \cos^2 \alpha_2$ und $2|\sin \alpha_1 \sin \alpha_2| \leq \sin^2 \alpha_1 + \sin^2 \alpha_2$).

Somit liefern die Sätze 3.18, 3.19 und 3.20 die Stabilität des Verfahrens. \square

Die bisherigen Resultate gelten nur für DGLen mit konstanten Koeffizienten, während die allgemeine Theorie auch raumabhängige Koeffizienten zuließ. Tatsächlich kann man auch für x -abhängige Koeffizienten Stabilitätskriterien mittels der (x -abhängigen) Amplifikationsmatrix $G(h, y, x)$ erhalten.

Ein entsprechender Satz stammt von Lax und Nirenberg und lautet

Satz 3.21 Es sei $M_D = \{C(h) : h > 0\}$ mit

$$C(h) = \sum_{\mu=-k}^k B_\mu(x) T_{\Delta x}^\mu$$

und $\Delta x = h/\lambda$, λ fest, ein Differenzenverfahren zu einer sachgemäß gestellten Aufgabe $P(L_2(\mathbb{R}, \mathbb{R}^n), T, A)$. Ferner gelte:

- (1) $B_\mu \in C^2(\mathbb{R}) \quad \mu = -k, \dots, k$
- (2) $\|B_\mu^{(\nu)}(x)\|_\infty \leq K \quad \nu = 0, 1, 2, \quad \mu = -k, \dots, k, \quad x \in \mathbb{R}$
- (3) $\|G(h, y, x)\|_2 \leq 1 \quad (h > 0, \quad y \in \mathbb{R}, \quad x \in \mathbb{R})$

Dann ist M_D stabil. (Beweis z.B. bei Meis-Marcowitz) \square

Bemerkung 3.11 Bei mehreren Raumveränderlichen wird die Amplifikationsmatrix völlig analog definiert. Ist etwa mit $x \in \mathbb{R}^d$, $\Delta x \in \mathbb{R}^d$

$$C(x, h) = \sum_s B_s(x, h) T_{\Delta x}^s, \quad s \in \mathbb{Z}^d$$

$$T_{\Delta x}^s = T_{\Delta x_1}^{s_1} \cdots T_{\Delta x_d}^{s_d},$$

dann wird

$$G(h, y, x) = \sum_s B_s(x, h) \exp(i \sum_{j=1}^d s_j y_j \Delta x_j),$$

wobei wiederum

$$\lambda_j = h/\Delta x_j \quad \text{bzw.} \quad \lambda_j = h/\Delta x_j^2 \quad \text{konstant, } j = 1, \dots, d$$

angenommen ist. Die Sätze übertragen sich auf diesen Fall entsprechend. \square

Beispiel 3.19 Friedrichs-Verfahren für ein regulär hyperbolisches symmetrisches System

$$u_t = A(x, y)u_x + B(x, y)u_y \quad (x, y) \in \mathbb{R}^2, \quad t \in [0, T]$$

$A(x, y)$, $B(x, y)$ seien reell symmetrisch und erfüllen die Voraussetzungen von Satz 3.21 (1) und (2) entsprechend.

Ferner seien die Eigenwerte von

$$P(x, y, k_1, k_2) = k_1 A(x, y) + k_2 B(x, y)$$

für $|k_1| + |k_2| \neq 0$ paarweise verschieden. Dann ist das gestellte Problem regulär hyperbolisch.

Die Wiederholung der Überlegungen zu Beispiel 3.18 führt nun unter der Voraussetzung

$$\lambda_1 \leq \frac{1}{2 \sup \varrho(A(x, y))}, \quad \lambda_2 \leq \frac{1}{2 \sup \varrho(B(x, y))}$$

zur Stabilitätsaussage. □

In Satz 3.21 sind h -abhängige Koeffizienten $B_\mu(x, h)$ nicht zugelassen, wie man sie bei der Anwendung des Lax–Wendroff–Verfahrens auf Systeme mit variablen Koeffizienten erhält. Hier hilft jedoch Satz 3.10 (Satz von Kreiss) weiter:

Beispiel 3.20 Lax–Wendroff–Verfahren bei variablen Koeffizienten, $d = 1$:

$$u_t = A(x)u_x, \quad x \in \mathbb{R}, \quad 0 \leq t \leq T.$$

$A(x)$ reell symmetrisch mit paarweise verschiedenen Eigenwerten $\neq 0$ und

$$\|A(x) - A(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}.$$

Differenzenverfahren:

$$\begin{aligned} u_{n+1}(x) &= (I + \frac{\lambda}{2}A(x)(T_{\Delta x} - T_{\Delta x}^{-1}) + \frac{\lambda^2}{2}A(x)(A(x + \frac{\Delta x}{2})(T_{\Delta x} - I) \\ &\quad - A(x - \frac{\Delta x}{2})(I - T_{\Delta x}^{-1})))u_n(x) \\ C(h) &= I + \frac{\lambda}{2}A(x)(T_{\Delta x} - T_{\Delta x}^{-1}) + \frac{\lambda^2}{2}A(x)(A(x + \frac{h}{2\lambda})(T_{\Delta x} - I) \\ &\quad - A(x - \frac{h}{2\lambda})(I - T_{\Delta x}^{-1})) \\ &= \underbrace{I + \frac{\lambda}{2}A(x)(T_{\Delta x} - T_{\Delta x}^{-1}) + \frac{\lambda^2}{2}A^2(x)(T_{\Delta x} - 2I + T_{\Delta x}^{-1})}_{\tilde{C}(h)} + \mathcal{O}(h) \end{aligned}$$

Stabilitätsnachweis für $\tilde{C}(h)$:

Zugeordnete Amplifikationsmatrix

$$\begin{aligned} G(h, y, x) &= I + i\lambda A(x) \sin \alpha + \lambda^2 A^2(x) \left(\overbrace{\cos \alpha}^{\cos^2(\frac{\alpha}{2}) - \sin^2(\frac{\alpha}{2})} - 1 \right) \\ &= I - 2 \sin^2(\frac{\alpha}{2}) \lambda^2 A^2(x) + 2i\lambda A(x) \sin(\frac{\alpha}{2}) \cos(\frac{\alpha}{2}). \end{aligned}$$

Seien $\mu_j(x)$ die Eigenwerte von $A(x)$. Dann sind die Eigenwerte $\lambda_j(h, y, x)$ von $G(h, y, x)$

$$\lambda_j(h, y, x) = 1 - 2 \sin^2\left(\frac{\alpha}{2}\right) \lambda^2 \mu_j^2(x) + 2i \lambda \mu_j(x) \sin\left(\frac{\alpha}{2}\right) \cos\left(\frac{\alpha}{2}\right).$$

G ist normal und deshalb $\|G(h, y, x)\|_2 \leq 1$ falls

$$|\lambda_j(h, y, x)| \leq 1 \quad \forall j$$

d.h.

$$\lambda \leq \frac{1}{\sup_x \varrho(A(x))}$$

(vgl. die entsprechende Rechnung in Beispiel 3.17)

□

Literaturverzeichnis

- [1] Ansorge, R.: Differenzenapproximationen partieller Anfangswertaufgaben, Teubner, Stuttgart (1978)
- [2] Ansorge, R.; Hass, R.: Konvergenz von Differenzenverfahren für lineare und nichtlineare Anfangswertaufgaben, Springer: lecture Notes in Mathematics 159, (1970)
- [3] v. Finckenstein, K: Einführung in die Numerische Mathematik Bd 2, Hanser-Verlag, München (1978)
- [4] v. Finckenstein, K: Numerische Behandlung von Anfangs-Randwertproblemen partieller Differentialgleichungen, Skriptum zur Vorlesung WS 1982/83, THD, FB4.
- [5] Gladwell, I; Wait, R.(eds.): A survey of Numerical methods for Partial Differential Equations, Clarendon Press, Oxford (1979)
- [6] Grigorieff, R. D. : Numerik gewöhnlicher Differentialgleichungen, Bd. 2 Teubner, Stuttgart , (1977)
- [7] Großmann, Ch.; Roos, H.G.: Numerik partieller Differentialgleichungen. Teubner (Studienbücher) (1992).
- [8] Hellwig: Partielle Differentialgleichungen. Stuttgart: Teubner (1960).
- [9] Houwen & Sommeier: J. Comp. Appl. Math. (1985), S. 145–161.
- [10] Klar, A.: Skriptum zur Vorlesung “Höhere Numerische Mathematik II“, TUD, SS 2002.
- [11] Noye, J.(ed): Computational Techniques for Differential Equations, North Holland (Math. Studies 83), (1984)
- [12] Meis, Th.; Marcowitz, U: Numerische Behandlung partieller Differentialgleichungen, Springer (1978).
- [13] Petrovsky, I.G.: Lectures on partial differential equations. New York-London: Interscience Publishers (1954).

- [14] Quarteroni, A.;Valli, A.: Numerical Approximation of Partial Differential Equations. Springer (1994).
- [15] Richtmyer, R.D.; Morton, K.W.: Difference Methods for Initial Value Problems, Wiley, New York, (1967).
- [16] Toernig, W.; Spellucci, P.: Numerische Mathematik für Ingenieure und Physiker ,Bd 2, 2. Aufl. Springer (1990)
- [17] Le Vegue, R.J.: Finite volume methods for hyperbolic problems. Cambridge Univ. Press (2002).
- [18] Le Vegue, R.J.: Numerical methods for conservation laws. 2nd ed. Birkhäuser (1992).

Index

- Abhängigkeitsbreich, 13
- Abschneidefehler, 58
- ADI, 85
- Amplifikationsmatrix, 143, 153–156
- Anfangs-Randwertproblem, 30

- BDF2, 84, 100
- Bestimmtheitsbereich, 13
- Burgers-Gleichung, 52

- CFL, 30
- Charakteristik, 7
- Charakteristikenverfahren, 16
- Courant-Friedrichs-Lewy, 30, 43
- Courant-Isaacson-Rees, 44
- Cowell, 37
- Crank-Nicholson, 76, 85, 100

- d’Alembert, 24
- Differentialgleichung, quasilineare, zweiter Ordnung, 5
- Differentialgleichung, steife, 92
- Differenzenverfahren, explizites, 76
- Differenzenverfahren, vollimplizites, 76
- Diskretisierung, naive, 43
- Dispersion, numerische, 35
- Du Fort und Frankel, 84

- Einflussbereich, 13
- Einschrittverfahren, 56
- Enquist-Osher, 64
- Entropie, 55
- Entropiefluß, 55
- Entropielösung, 55
- Erhaltungsgleichung, 51

- finite Elemente, 91

- Flussfunktion, 51
- flux limiter, 67
- Flußfunktion, numerische, 57
- Fortsetzungssatz, 108
- Fourierentwicklung, 141
- Fouriertransformation, 142
- Friedrichsverfahren, 43, 44

- Galerkin-Approximation, 94
- Galerkinmethode, diskontinuierliche, 102
- Gesamtsteifigkeitsmatrix, 92
- Godunov, 62

- hochoszillatorisch, 32
- hyperbolisch, 6

- konservativ, 57
- konsistent, 58
- Konsistenzordnung, 58
- Koordinatensystem, charakteristisches, 16

- Lax-Wendroff, 44, 45
- Linienmethode, horizontale, 104
- Linienmethode, vertikale, 31, 75
- lumping, 101

- Massenmatrix, 92
- Mittelpunktregel, 33
- monoton, 64

- neutral stabil, 36
- Norsett, 100

- Operator-Splitting, 85
- Ordnungsreduktion, 102

- Padé-Approximation, 100

- Rankine-Hugoniot Bedingung, 53

Riemannproblem, 62
Rosenbrock-Wanner, 100
Rothe-Methode, 104

Saite, schwingende, 23
Semidiskretisierung, 31, 75
Störmer, 37
Stabilitätsproblem, 32
superbee, 67
Systeme, quasilineare, 5

Totalvariation, 56
Trapezregel, 34
TVD, 64

upwind, 61

van Leer, 67
Variation, beschränkte, 56
Verfahren, konservatives, 57
Verfahren, monotones, 64
von Neumann, 146

Weierstrass, 107
Wellengleichung, 23

Zellmittel, 62