MAX-PLANCK-INSTITUT FÜR PLASMAPHYSIK

GARCHING BEI MÜNCHEN

DIFFERENZENSCHEMATA FÜR HYPERBOLISCHE SYSTEME
(DIFFERENCE SCHEMES FOR HYPERBOLIC SYSTEMS)

W. Höhn

IPP 6/109 July 1972

Die nachstehende Arbeit wurde im Rahmen des Vertrages zwischen dem Max-Planck-Institut für Plasmaphysik und der Europäischen Atomgemeinschaft über die Zusammenarbeit auf dem Gebiete der Plasmaphysik durchgeführt.

IPP 6/109

W. Höhn

Difference Schemes for Hyperbolic Systems

July 1972

(in German)

Abstract

In the case of non-strict hyperbolic differential systems the corresponding difference schemes usually involve amplification matrices increasing like polynominals. For this weak stability an eqivalence theorem, similar to that of LAX, is established for initial value problems with time-dependent linear solution operators.

Difference schemes for hyperbolic systems using leap-frog differences - one of which is stable in the widest possible range - are studied and their stability or weak stability is investigated by direct evaluation of powers of certain normal forms of their amplification matrices.

Some numerical results demonstrate the practical usefullness of these schemes.

Inhaltsverzeichnis

Einführung

Zur numerischen Behandlung von Anfangswertproblemen werden in der Praxis gern Differenzenschemata herangezogen. Diese Methode sichert meist einen verhältnismäßig geringen Aufwand, etwa durch ihre Schnelligkeit oder bequeme Handhabung beim Rechnen.

Für den Beweis der Konvergenz von Lösungen des Differenzenschemas zu denen des entsprechenden Anfangswertproblems hat sich im linearen Falle die Theorie von LAX und RICHTMYER durchgesetzt (s.z.B. [5], [8], [1]), deren Kernstück, der Äquivalenzsatz ([8]S45,127; [1] S62), einen Zusammenhang zwischen Stabilität und Konvergenz herstellt. Im Konvergenzbegriff sind verallgemeinerte Lösungen des Anfangswertproblems mit einbezogen.

Um in der Praxis sinnvoll mit Differenzenschemata umgehen zu können, genügt aber nicht nur ihre Konvergenz allein, vielmehr benötigt man eine ausreichende Geschwindigkeit der Konvergenz, da merkliche Verkleinerung der Schrittweite h vor allem bei Problemen mit mehreren Raumdimensionen einen enormen Mehrbedarf an Speicherplatz bedeutet, oder aber das Verfahren durch Rundungsfehler ohne Erfolg bleibt.

Die Geschwindigkeit der Konvergenz läßt sich den praktischen Verhältnissen angemessen am einfachsten durch den Exponenten τ im Ausdruck $K(u) \cdot h^{\tau}$ angegeben, den man als Schranke für den Fehler beweisen muß. Die Konstante $K(u) \in \mathbb{R}^+$ ist abhängig von der zu approximierenden exakten Lösung. Damit ist die unangenehmste Tatsache für die Behandlung verallgemeinerter Lösungen bereits angesprochen, denn diese lassen sich im allgemeinen nur durch Grenzwerte von Funktionenfolgen $(u_\ell)_{\ell \in \mathbb{N}}$ gewinnen. Daß für jedes Folgenelement der Fehler

durch $K(u_\ell)\cdot h^2$ abgeschätzt werden kann, ist für den Grenzwert nutzlos da die Folge $(K(u_\ell))_{\ell\in\mathbb{N}}$ möglicherweise stark divergiert; man kann lediglich mit den bekannten Methoden überhaupt Konvergenz beweisen.

Unter diesem Gesichtspunkt scheint es gerechtfertigt, derartige Geschwindigkeitsuntersuchungen auf "klassische" Lösungen zu beschränken, solche Lösungen, die einer Konsistenzbedingung genügen. Gegenüber der LAX-RICHTMYER Theorie hat man dann die Begriffe Konvergenz und Konsistenz nicht mehr für eine Gesamtheit verallgemeinerter Lösungen zu betrachten sondern für eine einzige, feste Lösung, den Begriff Stabilität dagegen sehr wohl. Mit den so veränderten Definitionen wird im Anschnitt I ein Äquivalenzsatz auf analoge Weise bewiesen. Als Verallgemeinerungen berücksichtigt er zeitabhängige Probleme, schwache Stabilität und Geschwindigkeit der Konvergenz.

Im II. Kapitel werden diese Ergebnisse für ein von 2. Ordnung konsistentes Differenzenschema für zunächst strikt hyperbolische Systeme 1. Ordnung mit konstanten Koeffizienten verwendet, in welchen alle Ableitungen durch "Laubfrosch"-Differenzen ersetzt sind. Der Stabilitätsbereich des Schemas wird berechnet, und es wird gezeigt, daß er der größte überhaupt mögliche ist.
Schwach stabil wird dieses Schema für nicht mehr strikte sondern gewöhnlich hyperbolische Systeme. Dieser Fall wird im III. Abschnitt anhand einer für die Praxis brauchbaren Normalform für nicht mehr diagonalisierbare Amplifikationsmatrizen untersucht und die Stabilitätskonstante berechnet.

Im IV. Kapitel werden diese Ergebnisse auf andere 2-Schritt Verfahren verallgemeinert; als Beispiel wird der Stabilitätsbereich eines von 3. Ordnung konsistenten Schemas abgeschätzt.

Den Abschluß bilden numerische Resultate.

I. Ein Äquivalensatz

In einem BANACH-Raum $(B, \|\cdot\|)$, sei ein lineares Anfangswertproblem gegeben:

$$\frac{d}{dt} u(t) = A(t)u(t) \quad \text{für } t \in [0,T] \; ; \; u(0) = u_0 \tag{1}$$

mit einer Schar linearer Operatoren $A(t)$: $B' \to B$ wobei $B' \subset B$, $u(t) \in B'$ für $t \in [0,T]$, und zwar mit einem $T \in \mathbb{R}^+$.

Für alle u_0 aus einem D, $D \subset B$, existiere eine eindeutige Lösung $u(t)$ von (1). Für $t \in [0,T]$ sei $\{ E(s)_t \; ; \; s \in [t,T] \}$ eine Schar von Operatoren $E(s)_t : D_t \to B$ definiert auf:

$$D_t := \{ u_t \in B \; ; \text{ es gibt ein } u_0 \in D \text{ mit } u(t) = u_t \}$$

durch:

$$E(s)_t \, u_t := u(s+t) \qquad \text{für alle } s,t,s+t \in [0,T]$$

u_0 ist darin dem jeweiligen u_t zugehörig, $u(t)$ die Lösung von (1) zu u_0. Entsprechend zur Linearität von (1) ist $E(s)_t$ linear.

Falls $A(t)$ von der Zeit unabhängig ist, erfüllen die Lösungsoperatoren die Halbgruppeneigenschaft (S.[1], S.2,7), im zeitabhängigen Fall gilt das i.a. nicht mehr, sondern dagegen nur (auf D_t):

$$E(r)_s \, E(s)_t = E(r+s)_t \qquad \text{für alle } r,s,t, r+s+t \in [0,T]$$

Definition:

Seien $k \in \mathbb{N}$; $B^k := B \times B \times \cdots \times B = \bigotimes_{\nu=1}^{k} B$ und $\{ G(h)_t \; ; h \in [0,T] \}$ eine Schar linearer und stetiger Operatoren $G(h)_t : B^k \to B$ für $t \in [0,T]$.

$$U_h^{n+1} = G(h)_{n\cdot h} \left[U_h^n, U_h^{n-1}, \cdots, U_h^{n-k+1} \right] \tag{2}$$

mit vorgegebenen $U_h^m, U_h^{m-1}, U_h^{m-2}, \cdots, U_h^{m-k+1} \in B$, $m \in \mathbb{N}_0$, $m \geq k-1$

ist ein lineares k-Schritt Differenzenverfahren (-Schema) mit h als Schrittweite und $n \in \mathbb{N}_0$, $n \geq m$, $n \cdot h \in [0,T]$ als Variable.

($\mathbb{N}_0 := \mathbb{N} \cup \{0\}$)

Im folgenden wird für jedes $u_0 \in D = D_0$ die Lösung des Anfangswertproblems (1) mit $u(t) = E(t)_0 u_0$ bezeichnet, mit $U_h^{m\,n}$ die des Differenzenschemas (2) zu den Anfangswerten $U_h^{m\,m}, U_h^{m\,m-1}, \cdots, U_h^{m\,m-k+1} \in B$.

Der Parameter m ermöglicht den Beginn der Approximation mit dem Schema nicht nur zur Zeit t = 0 sondern für beliebige $t \in [0,T)$.

Das Schema (2) läßt sich übersichtlicher als Operation in B^k schreiben, dessen Elemente eine "$\sim$" bekommen.

$$\tilde{u} := (u_k, u_{k-1}, \cdots, u_1), \quad u_\nu \in B \text{ für } \nu = 1,2,\cdots,k$$

Die u_ν sind die Komponenten von $\tilde{u}$.

Für $t \in [0,T]$ seien $\tilde{C}(h)_t : B^k \to B^k$ definiert durch:

$$\tilde{C}(h)_t \tilde{u} := (C(h)_t [u_k, u_{k-1}, \cdots, u_1], u_k, u_{k-1}, \cdots, u_2)$$

für alle $\tilde{u} \in B^k$, solange $h, t \in [0,T]$,

und es seien $\tilde{U}_h^{m\,m} := (U_h^{m\,m}, U_h^{m\,m-1}, \cdots, U_h^{m,\,m-k+1}) \in B^k$.

Dann hat man (2) in der äquivalenten Form:

(3) $\tilde{U}_h^{m\,n+1} = \tilde{C}(h)_{n \cdot h} \tilde{U}_h^{m\,n}$ mit vorgegebenen $\tilde{U}_h^{m\,m} \in B^k$ und

$m \geq k-1$, $n \geq m$, $n \cdot h \leq T$

Für Vergleiche wird $E(\cdot)_t$ auch umgeschrieben:

Für $t \in [0,T]$ sei $D_t^k := \bigotimes_{\nu=1}^{k} D_t$ und für jedes $u_0 \in D$ wird vereinbart:

$$\tilde{u}(t) := (u(t), u(t), \cdots, u(t)) \in B^k; \quad \tilde{u}_0 := (u_0, u_0, \cdots, u_0) \in B^k$$

Falls dann $h, t+(k-1)\cdot h \in [0,T]$ so sei $\tilde{S}_t^h : D_t^k \to B^k$ definiert durch :

$$\tilde{S}_t^h \tilde{u}(t) := \left(E((k-1)h)_t\, u(t),\ E((k-2)h)_t\, u(t),\ \cdots,\ u(t) \right)$$

für alle $u_0 \in D$

Falls weiter $s, s+t, s+t-(k-1)h \in [0,T]$ so wird $\tilde{E}(s)_t : \tilde{D}_{t,h} \to B^k$ definiert durch:

$$\tilde{E}_h(s)_t\, \tilde{u} := \left(E(s)_t\, u_k,\ E(s)_{t-h},\ \cdots,\ E(s)_{t-(k-1)h}\, u_1 \right)$$

für alle $\tilde{u} \in \tilde{D}_{t,h} := \left\{ \tilde{u} \in B^k;\ u_\nu \in D_{t-(k-\nu)h} \text{ für } \nu = 1, 2, \cdots, k \right\}$

Zur Abkürzung wird noch $\tilde{u}_{t,h}(s) := \tilde{E}_h(s-(k-1)h)_{t+(k-1)h}\, \tilde{S}_t^h \tilde{u}(t)$ für $u_0 \in D$ gesetzt, falls auch $s \geq (k-1)h$. Oder anders geschrieben und für alle $t, s \in [0,T]$ mit $t+s,\ t+s-(k-1)h \in [0,T]$ gültig:

$$\tilde{u}_{t,h}(s) = \left(u(t+s),\ u(t+s-h),\ \cdots,\ u(t+s-(k-1)h) \right) \text{ für } u_0 \in D .$$

$\tilde{C}'(h)_t$, $\tilde{S}_t^h$ und $\tilde{E}_h(s)_t$ sind offenbar wieder linear.

Sei $\| \cdot \|_{\mathbb{R}^k}$ irgendeine Norm des $\mathbb{R}^k$. Dann ist die Abbildung $\| \cdot \| : B^k \to \mathbb{R} \cup \{0\}$ definiert durch

$$\| \tilde{u} \| := \left\| \left(\| u_k \|,\ \| u_{k-1} \|,\ \cdots,\ \| u_1 \| \right) \right\|_{\mathbb{R}^k} \text{ für jedes } \tilde{u} \in B^k,$$

$\tilde{u} := (u_k, u_{k-1}, \cdots, u_1)$, eine Norm in B^k. Da der $\mathbb{R}^k$ endliche Dimension besitzt, sind alle derartigen Normen äquivalent, insbesondere also äquivalent zu der durch $\| \tilde{u} \|_1 := \sum_{\nu=1}^{k} \| u_\nu \|$ definierten Norm in B^k. Sei $\| \cdot \|$ im folgenden eine solche Norm, die sich von der Norm

in B schreibtechnisch also nur durch das Argument unterscheidet, dann ist $(B^R, \|\cdot\|)$ ein BANACH-Raum.

Nach den Schreibkonventionen kommen jetzt die die Approximation betreffenden Begriffe.

Def.: (Konsistenz)

Das Schema (3) (bzw.(2)) heißt für ein $u_0 \in D$ Konsistent von der Ordnung $\rho, \rho \in \mathbb{R}^+$, wenn für ein $H_1 \in (0,T]$ und jedes $t \in [0,T]$ gilt:

$$\|[\tilde{C}(h)_t - \tilde{E}_h(h)_t]\,\tilde{E}_h(t-(k-1))_{(k-1)h}\,\tilde{S}_0^h\,\tilde{u}_0\| \le M(u_0)\cdot h^{\rho+1} \tag{4}$$

$$\text{für alle } h \in (0, H_1)\,;\ t+h,\ t-(k-1)h \in [0,T]$$

mit einer (von u_0 abhängigen) Konstanten $M(u_0) \in \mathbb{R}^+$.

Die Relation (4) läßt sich kürzer schreiben:

$$\|[\tilde{C}(h)_t - \tilde{E}_h(h)_t]\,\tilde{u}_{0,h}(t)\| \le M(u_0)\cdot h^{\rho+1} \tag{4}^1$$

Def.: (schwache Stabilität)

Das Schema (3) heißt schwach stabil vom Grade $\beta,\ \beta \in \mathbb{R}^+ \cup \{0\}$, wenn es ein $N \in \mathbb{R}^+$ und ein $H_2 \in (0,T]$ so gibt, daß für alle $n \in \mathbb{N}_0, (n \ge k-1)$, $h \in (0, H_2)$ mit $n\cdot h \in [0,T]$ gilt:

$$\Big\|\prod_{\nu=n}^{\mu} \tilde{C}(h)_{(n-\nu)h}\Big\| \le N\cdot h^{-\beta} \quad \text{für alle } \mu \in \mathbb{N},\ k-1 \le \mu \le n \tag{5}$$

Im Falle $\beta = 0$ nennt man es auch stabil.

Anmerkung:

Im zeitunabhängigen Fall hat man statt (5) einfach

$\|\tilde{C}(h)^n\| \le N\cdot h^{-\beta}$, der Index ist weggelassen,

also eine von einem Parameter, nämlich μ, weniger abhängende Relation.

Mittels (5) wird verlangt, daß sich alle bei Anwendung des Differenzenschemas zu einer Zeit $\nu \cdot h \le n \cdot h$ entstandenen Fehler bis zur Zeit $n \cdot h$ nicht zu stark, d.h. nur polynomartig, nicht exponentiell, vergrößern.

Def.:

Seien $t \in [0,T]$, $H \in (0,T]$. F_{st}^{H} ist eine Menge von Paaren dreier Folgen $(h_j)_{j\in\mathbb{N}}$, $(m_j)_{j\in\mathbb{N}}$ und $(n_j)_{j\in\mathbb{N}}$:

$$F_{st}^{H} := \{ [(h_j),(m_j),(n_j)] \,;\, h_j \in (0,H),\ m_j, n_j \in \mathbb{N}_0,\ k-1 \le m_j \le n_j,$$

$h_j \cdot (n_j+1) \in [0,T]$ für alle $j \in \mathbb{N}$, und $\lim_{j\to\infty} h_j = 0$, $\lim_{j\to\infty} m_j \cdot h_j = s$,

$\lim_{j\to\infty} n_j \cdot h_j = t \}$

Def.: (Konvergenz)

Die Schar von Lösungen $\{ U_h^{mn} \,;\, k-1 \le m \le n,\ h, n\cdot h \in [0,T] \}$ des Differenzenschemas (2) heißt von der Ordnung τ Konvergent (gegen die Lösung $u(t)$ von (1)), wenn für jedes Paar (s,t), $s,t \in [0,T]$, $s \le t$, und alle dazugehörigen $[(h_j),(m_j),(n_j)] \in F_{st}^{T}$ ein $K(u_0) \in \mathbb{R}^+$ und ein $j_0(u_0,s,t) \in \mathbb{N}$ existieren, so daß

$$\text{(6)} \qquad \left\| \left(\prod_{\nu=1}^{n_j-m_j} \tilde{C}(h_j)_{(n_j-\nu)h_j} \right) \tilde{U}_{h_j}^{m_j n_j} - \tilde{E}_{h_j,((n_j-k+1)h_j)} \tilde{\int}_0^{h_j} \tilde{u}_0 \right\| \le K(u_0) \cdot h_j^{\tau}$$

für alle $j \ge j_0(u_0,s,t)$

Falls der Laufende Index des Produkts einmal kleiner als der oben stehende Grenzindex sein sollte, so wird für das Produkt der Einheitsoperator auf B^k vereinbart.

Links in (6) steht $\tilde{U}_{h_j}^{m_j n_j}$, rechts $\tilde{u}_{m_j h_j, h_j}((n_j-m_j)h_j) = \tilde{u}_{0,h_j}(n_j h_j)$,

übersichtlicher geschrieben hat man daher statt (6):

$$\text{(6)}^1 \qquad \left\| \tilde{U}_{h_j}^{m_j n_j} - \tilde{u}_{m_j h_j}((n_j-m_j)h_j) \right\| \le K(u_0) \cdot h_j^{\tau}$$

In Übereinstimmung zur LAX-RICHTMYER-Formulierung ist hier die schwache Stabilität als eine Eigenschaft des Lösungsoperators des Differenzenschemas definiert, während dagegen Konsistenz und Konvergenz, abgesehen von der Ordnung, hier schwächere Eigenschaften sind, da sie sich nur auf ein $u_0 \in D$ beziehen.

Damit Lösungen von (2) gegen die von (1) konvergieren, wird man selbstverständlich voraussetzen, daß die Anfangswerte von (2) die exakte Lösung approximieren, und daß die Geschwindigkeit der Konvergenz am Anfang nicht kleiner ist als die gewünschte.

Def.:

Die Schar $\{ U_h^{m,n} ; k-1 \le m \le n, h, m \cdot h \in [0,T] \}$ von Lösungen von (2) approximiert am Anfang die Lösung $u(t)$ von (1) von der Ordnung μ, $\mu \in \mathbb{R}^+$, wenn für ein $H_3 \in (0,T]$ gilt:

$$\| U_h^{m,m-\nu} - u((m-\nu)h) \| \le a'(u_0) \cdot h^\mu$$

für $\nu = 0, 1, \dots, k-1$, alle $m \in \mathbb{N}_0$, $m \ge k-1$, $m \cdot h \in [0,T]$, und alle $h \in (0, H_3)$ mit einem $a'(u_0) \in \mathbb{R}^+$

oder äquivalent dazu:

(7) $$\| \tilde{U}_h^{m,m} - \tilde{u}_{mh,h}(0) \| \le a(u_0) \cdot h^\mu$$ für alle $m \in \mathbb{N}_0$ mit $m \ge k-1$, $m \cdot h \in [0,T]$ und alle $h \in (0, H_3)$ mit einem $a(u_0) \in \mathbb{R}^+$

Damit läßt sich für Lösungen von (1), die der Konsistenzbedingung (4) genügen, analog zur LAX'schen Vorgehenswese ein Äquivalenzsatz beweisen:

Satz 1: (Äquivalenzsatz)

Seien $\rho, \tau \in \mathbb{R}^+$, $\rho \ge \tau$

Für ein $u_0 \in D$ sei das Schema (2) konsistent von der

Ordnung ρ. Dann gilt:

Schwache Stabilität vom Grade $\rho - \tau$ $\Longleftrightarrow$

Jede Schar von Lösungen von (2) $\{U_h^{m,n};\ k-n \le m \le n;\ h, n\cdot h \in [0,T]\}$, die am Anfang die Lösung $u(t)$ von (1) von der Ordnung ρ approximiert, konvergiert von der Ordnung τ gegen $u(t)$.

Bemerkung:

Die im Satz garantierte Konvergenz ist eine stärkere Eigenschaft als die, die man in der Praxis normalerweise benötigt. Man braucht da nämlich Konvergenz nur für Scharen mit $m = k-n = const$, jedenfalls im Rahmen einer Theorie die Rundungsfehler ausschließt. Im zeitunabhängigen Falle verschwinden diese Unterschiede; ein ähnliches Verhalten wie bei der schwachen Stabilität.

Beweis des Satzes:

a) Das Verfahren sei konvergent.

Wegen der Linearität sowohl des Anfangswertproblems als auch des Differenzenschemas kann man die Konvergenz von Lösungsscharen $\{U_h^{m,n};\ k-n \le m \le n;\ h, n\cdot h \in [0,T]\}$ gegen $u(t)$ auf den linearen Teilraum D^0 von B, $D^0 := \{c\cdot u_0, c \in \mathbb{R}\}$ erweitern. Benötigt wird das in diesem Beweisteil nur für das Nullelement $0 \in B$, zu dem als Anfangswert die Konstante 0, $t \in [0,T]$, als Lösung von (1) gehört. Seien $\tilde{0} := (0, 0, 0, \cdots, 0) \in B^k$ und $\{O_h^{m,n};\ k-n \le m \le n;\ h, n\cdot h \in [0,T]\}$ eine Schar von Lösungen von (2) die am Anfang 0 von der Ordnung ρ approximiert. Sei $\tilde{P}_h^{m,n}$ die Lösung von (3) mit dem Anfangswert $\tilde{P}_h^{m,n} = \tilde{J}_{(m-k+n)h}^{h}\, u_0$. $P_h^{m,n}$ stimmt dann am Anfang genau mit $u(t)$ überein, also sicher von der Ordnung ρ. $P_h^{m,n} + O_h^{m,n}$ approximiert am Anfang $u(t)$ von der Ordnung ρ. Die Elemente ohne "$\sim$" sind darin

wieder die zugehörigen Lösungen von (2). Nach Voraussetzung folgt Konvergenz für beide Scharen von der Ordnung τ, d.h.:

für jedes Paar (s,t), $s,t \in [0,T]$, $s \le t$, und alle dazugehörigen $[(h_j),(m_j),(n_j)] \in F_{st}^{T}$ existieren ein $K(u_0) \in \mathbb{R}^+$ und ein $j_0(u_0,s,t) \in \mathbb{N}$, so daß

$$\left\| \tilde{P}_{h_j}^{m_j n_j} - \tilde{u}_{0,h_j}(m_j h_j) \right\| \le K(u_0) \cdot h_j^{\tau} \qquad \text{und}$$

$$\left\| \tilde{P}_{h_j}^{m_j n_j} + \tilde{O}_{h_j}^{m_j n_j} - \tilde{u}_{0,h_j}(m_j h_j) \right\| \le K(u_0) \cdot h_j^{\tau}$$

$$\Rightarrow$$

$$\left\| \tilde{O}_{h_j}^{m_j n_j} - \tilde{o} \right\| = \left\| \tilde{O}_{h_j}^{m_j n_j} \right\| \le 2K(u_0) \cdot h_j^{\tau} \tag{8}$$

$$\Rightarrow \qquad \text{für } j \ge j_0(u_0,s,t)$$

$\{O^{mn};\ k-1 \le m \le n;\ h, n\cdot h \in [0,T]\}$ konvergiert gegen 0 von der Ordnung τ.

Sei $\tilde{V} \in B^k$ beliebig gewählt. Angenommen die Menge

$$\left\{ h^{\rho-\tau} \cdot \left\| \left(\prod_{\nu=1}^{n-m} \tilde{C}(h)_{(n-m-\nu)h} \right) \tilde{V} \right\| ;\ h \in (0,H_2),\ n,m \in \mathbb{N}_0,\ \begin{matrix} k-1 \le m \le n, \\ n\cdot h \in [0,T] \end{matrix} \right\}$$

ist für kein $H_2 \in (0,T]$ beschränkt. Dann gibt es Folgen

$(h_j),(m_j),(n_j)_{j\in\mathbb{N}}$ mit $h_j \in (0,T]$, $m_j, n_j \in \mathbb{N}_0$, $k-1 \le m_j \le n_j$; $n_j \cdot h_j \in [0,T]$

für $j \in \mathbb{N}$; mit $\lim_{j\to\infty} h_j = 0$

und $h_j^{\rho-\tau} \left\| \left(\prod_{\nu=1}^{n_j-m_j} \tilde{C}(h_j)_{(n_j-m_j-\nu)h_j} \right) \tilde{V} \right\| \xrightarrow{\text{für } j\to\infty} \infty$

und darin gibt es Teilfolgen $(h_{j_\ell}) \subset (h_j)$, $(m_{j_\ell}) \subset (m_j)$, $(n_{j_\ell}) \subset (n_j)$, $j_\ell \in \mathbb{N}$

$j_{\ell+1} > j_\ell$ für $\ell \in \mathbb{N}$ (ℓ ist Index von j) mit

$\lim_{\ell\to\infty} m_{j_\ell} h_{j_\ell} = s$, $\lim_{\ell\to\infty} n_{j_\ell} h_{j_\ell} = t$ für ein Paar (s,t) mit $s,t \in [0,T]$, $s \le t$

und

$$\tilde{W}_{j_\ell} := \left(\prod_{\nu=1}^{n_{j_\ell}-m_{j_\ell}} \tilde{C}(h_{j_\ell})_{(n_{j_\ell}-\nu)h_{j_\ell}} \right) \tilde{V} \neq \tilde{o} \qquad \text{für alle } \ell \in \mathbb{N}$$

Ohne Einschränkung der Allgemeinheit erfüllen bereits die Folgen $(h_j),(m_j),(n_j)$ diese Eigenschaften.

Sei $\tilde{V}_j := \dfrac{h_j^{\rho}\,\tilde{V}}{(h_j^{\rho-\tau}\|\tilde{W}_j\|)^{\frac{1}{2}}}$ für $j \in \mathbb{N}$.

Da der Nenner nach unten beschränkt ist, approximiert $\tilde{V}_j$ das Element $\tilde{0}$ mit h_j^{ρ} (d.h. die Relation (7) ist für die Folgen $(h_j), (m_j), (n_j)$ erfüllt). Wegen (8) gilt für alle $j \in \mathbb{N}$, $j \geq j_0$ mit einem $j_0 \in \mathbb{N}$:

$$2K(t_0)\cdot h_j^{\tau} \geq \left\| \left(\prod_{\nu=1}^{n_j - m_j} \tilde{C}(h_j)_{(n_j-\nu)h_j} \right) \tilde{V}_j \right\| =$$

$$= \frac{h_j^{\rho}\left\| \left(\prod_{\nu=1}^{n_j-m_j} \tilde{C}(h_j)_{(n_j-\nu)h_j} \right)\tilde{V}\right\|}{\left\{ h_j^{\rho-\tau}\left\| \left(\prod_{\nu=1}^{n_j-m_j} \tilde{C}(h_j)_{(n_j-\nu)h_j} \right)\tilde{V}\right\|\right\}^{\frac{1}{2}}} =$$

$$= h_j^{\rho-\frac{\rho-\tau}{2}} \|\tilde{W}_j\|^{\frac{1}{2}} = h_j^{\tau}\cdot (h_j^{\rho-\tau}\|\tilde{W}_j\|)^{\frac{1}{2}} =: h_j^{\tau}\cdot B_j$$

$\Rightarrow B_j \leq 2K(t_0)$ für genügend großes j, mit einer Folge $(B_j)_{j\in\mathbb{N}}$, die $\lim_{j\to\infty} B_j = \infty$ erfüllt und das ist ein Widerspruch.

Die Annahme war also falsch; es gibt daher ein $H_2 \in (0,T]$, so daß

$$h^{\rho-\tau}\left\| \left(\prod_{\nu=1}^{n-m} \tilde{C}(h)_{(n-\nu)h} \right)\tilde{V}\right\| \leq K_n(\tilde{V})$$

für $h \in (0, H_2)$; $n, m \in \mathbb{N}_0$, $k-1 \leq m \leq n$, $n\cdot h \in [0,T]$ mit einem $K_n(\tilde{V}) \in \mathbb{R}^+$.

Nach dem Prinzip von der gleichmäßigen Beschränktheit, angewandt auf die Menge von Operatoren in $\mathcal{B}^k$

$$\mathcal{M} := \left\{ h^{\rho-\tau}\prod_{\nu=1}^{n-m}\tilde{C}(h)_{(n-\nu)h} \,;\, h\in(0,H_2),\ n,m\in\mathbb{N}_0,\ k-1\leq m\leq n,\ n\cdot h\in[0,T] \right\}$$

gibt sein $N \in \mathbb{R}^+$, so daß

$$h^{\rho-\tau}\left\| \prod_{\nu=1}^{n-m}\tilde{C}(h)_{(n-\nu)h}\right\| \leq N \quad \text{für alle } h^{\rho}\prod_{\nu=1}^{n-m}\tilde{C}(h)_{(n-\nu)h} \in \mathcal{M},$$

d.h. schwache Stabilität vom Grade $\rho-\tau$.

Anmerkungen:

Die Konsistenz ging nicht ein in diesem Beweisteil.

Hier ist die Erfahrung ausgenutzt, daß im Beweise für den Äquivalenzsatz von LAX ([8], S.127) im entsprechenden Teil Konvergenz nur in einer Umgebung von $\tilde{0} \in \beta^k$ benötigt wird.

Hätte man genaueres Übereinstimmen von $\{ U_h^{m,n} ;\ k-r \le m \le n ,\ h, n\cdot h \in [0,T] \}$ am Anfang mit $u(t)$ gefordert, etwa von der Ordnung $p+\varepsilon$, $\varepsilon > 0$, so könnte man auf diese Weise nicht mehr die Notwendigkeit der schwachen Stabilität vom Grade $p-\tau$ für Konvergenz der Ordnung τ beweisen.

b) Das Verfahren sei schwach stabil vom Grade $p-\tau$.

Seien $s, t \in [0,T]$, $s \le t$ und $[(h_j),(m_j),(n_j)] \in \bar{F}_{s\,t}^{T}$ ein Folgentripel. u_0 erfülle die Konsistenzbedingung (4).

$\{ U_h^{m,n} ;\ k-r \le m \le n ;\ h, n\cdot h \in [0,T] \}$ sei eine Schar von Lösungen von (2), die am Anfang $u(t)$ von der Ordnung p approximiert. Dann gilt für jedes $l \in \mathbb{N}_0$ mit $l < n_j - m_j$

$$\tilde{u}_{m_j h_j, h_j}((l+1)\cdot h_j) = \tilde{E}_{h_j}(h_j)_{(m_j+l)h_j}\, \tilde{u}_{m_j h_j, h_j}(l h_j)$$

$$\tilde{U}_{h_j}^{m_j\, m_j+l+1} = \tilde{C}(h_j)_{(m_j+l)h_j}\, U_{h_j}^{m_j\, m_j+l}$$

Daraus erhält man für den Fehler

$$\tilde{Z}_{h_j}^{m_j\, m_j+l} := \tilde{U}_{h_j}^{m_j\, m_j+l} - \tilde{u}_{m_j h_j, h_j}(l h_j)$$

definiert für $l \in \mathbb{N}_0$, $l < n_j - m_j$

die rekursive Relation:

$$\tilde{Z}_{h_j}^{m_j\, m_j+l+1} = \tilde{U}_{h_j}^{m_j\, m_j+l+1} - \tilde{u}_{m_j h_j, h_j}((l+1)h_j) =$$

$$= \left[\tilde{C}(h_j)_{(m_j+l)h_j} - \tilde{E}_{h_j}(h_j)_{(m_j+l)h_j} \right] \tilde{u}_{m_j h_j, h_j}(l\cdot h_j) +$$

$$+ \tilde{C}(h_j)_{(m_j+l)h_j} \left[\tilde{U}_{h_j}^{m_j\, m_j+l} - \tilde{u}_{m_j h_j, h_j}(l h_j) \right]$$

$$= \left[\tilde{C}(h_j)_{(m_j+l)h_j} - \tilde{E}_{h_j}(h_j)_{(m_j+l)h_j} \right] \tilde{u}_{m_j h_j, h_j}(l h_j) + \tilde{C}(h_j)_{(m_j+l)h_j}\, \tilde{Z}_{h_j}^{m_j\, m_j+l}$$

Die eindeutige Lösung dieser (Operatoren-) Differenzengleichung zum Anfangswert $\tilde{\tilde{Z}}^{m_j m_j}_{\ell_j}$ ist gegeben durch:

$$\tilde{\tilde{Z}}^{m_j m_j+\ell}_{\ell_j} =$$

(9)
$$= \sum_{\mu=0}^{\ell-1} \left(\prod_{\nu=1}^{\mu} \tilde{C}^{(\ell_j)}_{(m_j+\ell-\nu)\ell_j} \right) \left[\tilde{C}^{(\ell_j)}_{(m_j+\ell-1-\mu)\ell_j} - \tilde{E}^{(\ell_j)}_{\ell_j (m_j+\ell-1-\mu)\ell_j} \right] \cdot \tilde{u}_{0,\ell_j}((m_j+\ell-1-\mu)\ell_j)$$
$$+ \prod_{\nu=1}^{\ell} \tilde{C}^{(\ell_j)}_{(m_j+\ell-\nu)\ell_j} \tilde{\tilde{Z}}^{m_j m_j}_{\ell_j} \qquad \text{für } \ell \le m_j - m_j$$

Beweis durch vollständige Induktion:

Für $\ell = 0$ ist die Aussage trivial.

Der Induktionsschritt läßt sich direkt verifizieren:

$$\left[\tilde{C}^{(\ell_j)}_{(m_j+\ell)\ell_j} - \tilde{E}^{(\ell_j)}_{\ell_j (m_j+\ell)\ell_j} \right] \tilde{u}_{m_j \ell_j, \ell_j}(\ell\ell_j) + \tilde{C}^{(\ell_j)}_{(m_j+\ell)\ell_j} \tilde{\tilde{Z}}^{m_j m_j+\ell}_{\ell_j} =$$

$$= [\cdots]\, \tilde{u}_{m_j \ell_j, \ell_j}(\ell\ell_j) + \sum_{\mu=0}^{\ell-1} \left(\tilde{C}^{(\ell_j)}_{(m_j+\ell)\ell_j} \prod_{\nu=1}^{\mu} \tilde{C}^{(\ell_j)}_{(m_j+\ell-\nu)\ell_j} \right) \cdot$$
$$\cdot \left[\tilde{C}^{(\ell_j)}_{(m+\ell-1-\mu)\ell_j} - \tilde{E}^{(\ell_j)}_{\ell_j (m_j+\ell-1-\mu)\ell_j} \right] \cdot \tilde{u}_{0,\ell_j}((m_j+\ell-1-\mu)\ell_j)$$
$$+ \tilde{C}^{(\ell_j)}_{(m_j+\ell)\ell_j} \prod_{\nu=1}^{\ell} \tilde{C}^{(\ell_j)}_{(m_j+\ell-\nu)\ell_j} \tilde{\tilde{Z}}^{m_j m_j}_{\ell_j} \rightleftharpoons$$

Es ist:

$$\tilde{C}^{(\ell_j)}_{(m_j+\ell)\ell_j} \prod_{\nu=1}^{\mu} \tilde{C}^{(\ell_j)}_{(m_j+\ell-\nu)\ell_j} = \prod_{\nu=0}^{\mu} \tilde{C}^{(\ell_j)}_{(m_j+\ell-\nu)\ell_j}$$

Summiert man von 1 bis ℓ statt von 0 bis $\ell-1$ so steht da:

$$\rightleftharpoons \left[\tilde{C}^{(\ell_j)}_{(m_j+\ell)\ell_j} - \tilde{E}^{(\ell_j)}_{\ell_j (m_j+\ell)\ell_j} \right] \tilde{u}_{m_j \ell_j, \ell_j}(\ell\ell_j) +$$

$$+ \sum_{\mu=1}^{\ell} \left(\prod_{\nu=1}^{\mu} \tilde{C}^{(\ell_j)}_{(m_j+\ell+1-\nu)\ell_j} \right) \left[\tilde{C}^{(\ell_j)}_{(m+\ell-\mu)\ell_j} - \tilde{E}^{(\ell_j)}_{\ell_j (m+\ell-\mu)\ell_j} \right] \cdot$$
$$\cdot \tilde{u}_{0,\ell_j}((m_j+\ell-\mu)\ell_j)$$
$$+ \prod_{\nu=1}^{\ell+1} \tilde{C}^{(\ell_j)}_{(m_j+\ell+1-\nu)\ell_j} \tilde{\tilde{Z}}^{m_j m_j}_{\ell_j} \rightleftharpoons$$

also nach Indizesverschiebung von μ (bzw. ν im hinteren Produkt) und damit zwangsläufig für ν (in den Produkten der Summanden).

$$\Rightarrow \tilde{Z}_{h_j}^{m_j\, m_j+\ell+1} \quad ; \text{ damit ist (9) bewiesen.}$$

Anmerkung:

Diese Herleitung begründet das Auftreten der LAX'schen Summe im Äquivalenzbeweis ([8], S.46,(3.10)), der Fehler der Anfangswerte wird in ihr berücksichtigt und die LAX'sche Methode ist durch sie auf zeitabhängige Probleme ausdehnbar.

Fortsetzung des Beweisteils b):

Aus (9) folgt für $\ell \in \mathbb{N}_0$, $\ell \leq m_j - m_j$:

$$\| \tilde{Z}_{h_j}^{m_j\, m_j+\ell} \| \leq \sum_{\mu=0}^{\ell-1} \| \prod_{\nu=1}^{n} \tilde{C}(h_j)_{(m_j+\ell-\nu)h_j} \| \cdot M(u_0) \cdot h_j^{\rho+1}$$

$$+ \| \prod_{\nu=1}^{\ell} \tilde{C}(h_j)_{(m_j+\ell-\nu)h_j} \| \cdot \| \tilde{Z}_{h_j}^{m_j\, m_j} \| \quad \Rightarrow$$

wegen der Beschränktheit der Operatoren und der Konsistenzbedingung (4)', und zwar für $h_j < H_1$.

$$\leq M(u_0) \cdot N \cdot h_j^{\rho+1} \cdot h_j^{\tau-\rho} \sum_{\mu=0}^{\ell-1} 1 + a(u_0) \cdot N \cdot h_j^{\tau-\rho} \cdot h_j^{\rho} \Rightarrow$$

wegen schwacher Stabilität (5), und der Approximation am Anfang (s.(7) , die Voraussetzung des Satzes und $\tilde{Z}_{h_j}^{m_j\, m_j} = \tilde{U}_{h_j}^{m_j\, m_j} - \tilde{u}_{m_j h_j, h_j}^{(0)}$.

$$= h_j^{\tau} \cdot N \cdot [M(u_0) \cdot h_j \cdot \ell + a(u_0)]$$

$$\leq h_j^{\tau} \cdot N \cdot [M(u_0) \cdot T + a(u_0)]$$

für $h_j < H_1$

und das ist sogar mehr als in der Konvergenzbedingung (6) gefordert wird. Q.e.d.

Anmerkungen:

Daß der Fehler (s.letzte Beweiszeile) bei festgehaltener Schrittweite linear in T ansteigt, paßt zu der praktischen Erfahrung wachsender Fehler, wenn man ein Anfangswertproblem über längere Zeiten näherungsweise berechnet.

Weniger genau stimmende Anfangswerte, etwa von der Ordnung $p-\varepsilon$, $0<\varepsilon<\tau$, hätten langsamere Konvergenz, und zwar von der Ordnung $\tau-\varepsilon$ bedingt.

In der Praxis ist ohnehin meist die Ordnung $p+1$ erforderlich.

Da dieser Beweis für eine bestimmte Lösung von (1) geführt wurde, werden von (1) nicht globale Eigenschaften der Operatorenschar $\{E(t)_0 ; t\in[0,T]\}$ wie etwa dichter Definitionsbereich in B oder gleichmäßige Beschränktheit gefordert. Dagegen gehen die "Glattheitseigenschaften" der gerade gesuchten Lösung des Anfangswertproblems im vollen Umfang ein in der Konsistenzbedingung und in den Koeffizienten vor h_j^{τ} .

Nur Konvergenz verallgemeinerter Lösungen ohne Angabe der Geschwindigkeit hätte man im zeitabhängigen Fall analog zum LAX'schen Beweise liefern können.

II. Untersuchung des 2-Schritt Schemas

$$U^{n+1} - U^{n-1} = \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \left(T_\nu - T_\nu^{-1}\right) U^n \tag{10}$$

Das Schema soll zur näherungsweisen Berechnung der Lösungen des Anfangswertproblems

$$\frac{\partial u(t,x)}{\partial t} = \sum_{\nu=1}^{N} A_\nu \frac{\partial u(t,x)}{\partial x_\nu} \quad \text{für } t \in [0,T] \text{ ; } x \in \mathbb{R}^N \tag{11}$$
$$u(0,\cdot) = u_0(\cdot)$$

für reellwertige Funktionen $u : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^m$ dienen. Die Matrizen A_ν, $\nu \in \mathcal{N}$, sind konstant, N ist die Raumdimension. Als Norm wählt man, um später mit FOURIER-Transformationen arbeiten zu können, die L_2-Norm. Sei

$$I := \bigotimes_{\nu=1}^{N} [a_\nu, b_\nu] \quad ; \; a_\ell, b_\ell \in \mathbb{R}, \; a_\ell < b_\ell \text{ für } \ell = 1,2,\dots,N$$

und im folgenden $B = L_2(I)$ d.h. $v(\cdot)$ bzgl. I periodisch, und

$$\| v(\cdot) \| := \left(\int_I (v(x), v(x))\, dx \right)^{\frac{1}{2}} \qquad \text{für alle } v(\cdot) \in B$$

Mit $(\cdot,\cdot)$ ist das Skalarprodukt zweier Vektoren $p, q \in \mathbb{R}^\ell$, $\ell \in \mathcal{N}$, gemeint: $(p,q) := \sum_{i=1}^{\ell} p_i q_i$; p_i, q_i sind die Komponenten von p bzw. q.

Zu einem $u_0(\cdot) \in B$ existiere eine eindeutige Lösung $u(t) \in B$ für $t \in [0,T]$ von (11).

Die Tv in (10) sind Translationsoperatoren um den Betrag $R_\nu \cdot h$, $R_\nu \in \mathbb{R}^+$, für $\nu = 1,2,\dots,N$. Seien $y, y' \in \mathbb{R}^{N+1}$, $y := (x_0, x_1, \dots, x_N)$, $y'_\nu := (x_0, x_1, \dots, x_{\nu-1}, x_\nu + R_\nu h, x_{\nu+1}, \dots, x_N)$. t wird mit x_0 identifiziert, $R_0 := 1$. Tv ist in y für alle in y und y_ν definierten Funktionen $v(\cdot)$ erklärt durch:

$$T_\nu v(y) = v(y'_\nu) \quad , \text{ und zwar für } \nu = 0,1,2,\dots,N$$

A. Konsistenz.

Von $u_0(\cdot)$ wird gefordert, daß einige partielle Ableitungen 3. Ordnung von $u(t,\cdot)$ existieren, beschränkt sind und zu B gehören, und zwar: für $i = 0,1,2,\cdots,N$ wird verlangt

$$\frac{\partial^3 u(t,\cdot)}{\partial x_i^3} \in B \quad \text{für } t \in [0,T]$$

und für $(t,x) \in [0,T] \times I_0$, I_0 eine offene Menge: $(I_0 \supset I)$

$$\left| \frac{\partial^3 u(t,x)}{\partial x_i^3} \right| \le C_i \qquad \text{mit gewissen Konstanten } C_i \in \mathbb{R}^+$$

Aus dem TAYLOR'schen Satz folgt dann für $(t,x) \in [h, T-h] \times I$ und für $\nu = 0,1,2,\cdots,N$:

$$\left| (T_\nu - T_\nu^{-1})\, u(t,x) - 2R_\nu h \frac{\partial u(t,x)}{\partial x_\nu} \right| \le \frac{R_\nu^3 h^3}{3!} \left(\left| \frac{\partial^3 u(z')}{\partial x_\nu^3} \right| + \left| \frac{\partial^3 u(z'')}{\partial x_\nu^3} \right| \right)$$

$$\le \frac{R_\nu^3}{3!} h^3 P_1 \quad \text{mit einem } P_1 \in \mathbb{R}^+$$

wobei $z' := (x_0, x_1, \cdots, x_{\nu-1}, x_\nu', x_{\nu+1}, \cdots, x_N)$ $\quad x_\nu' \in [x_\nu, x_\nu + R_\nu h]$

und

$z'' := (x_0, x_1, \cdots, x_{\nu-1}, x_\nu'', x_{\nu+1}, \cdots, x_N)$ $\quad x_\nu'' \in [x_\nu - R_\nu h, x_\nu]$

$\Rightarrow$ (12)

$$\left\| (T_\nu - T_\nu^{-1})\, u(t,\cdot) - 2R_\nu h \frac{\partial u(t,\cdot)}{\partial x_\nu} \right\| \le h^3 \frac{R_\nu^3 P_1}{6} \le h^3 P_2$$

für $t \in [h, T-h]$ und $\nu = 0,1,2,\cdots,N$ mit einem $P_2 \in \mathbb{R}^+$

Schreibt man (10) als System in $\begin{pmatrix} U^{n+1} \\ U^n \end{pmatrix}$ d.h.

$$\begin{pmatrix} U^{n+1} \\ U^n \end{pmatrix} = \begin{pmatrix} \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu}(T_\nu - T_\nu^{-1}) & I \\ I & 0 \end{pmatrix} \begin{pmatrix} U^n \\ U^{n-1} \end{pmatrix} \tag{13}$$

(genauer sind es ortsabhängige $U^n(\cdot) \in B$, I ist die Einheitsmatrix) und bezeichnet wieder die zugehörigen in Abschnitt I erklärten Lösungsoperatoren mit $\tilde{C}(h)$ bzw. $\tilde{E}(h)$ (der zeitabhängige Index ist weggelassen), dann ist für alle h, t mit $h, t, t+2h \in [0,T]$:

$$\left\| [\tilde{C}(h) - \tilde{E}(h)] \begin{pmatrix} u(t+h,\cdot) \\ u(t,\cdot) \end{pmatrix} \right\|_1 =$$

$$= \| \sum \frac{A_\nu}{R_\nu}(T_\nu - T_\nu^{-1}) u(t+h,\cdot) + u(t,\cdot) - u(t+2h,\cdot) \| + \underbrace{\| u(t+h,\cdot) - u(t+h,\cdot) \|}_{=0}$$

$$= \| \sum \frac{A_\nu}{R_\nu}(T_\nu - T_\nu^{-1}) u(t+h,\cdot) - 2h \sum A_\nu \frac{\partial}{\partial x_\nu} u(t+h,\cdot) -$$

$$u(t+2h,\cdot) - u(t,\cdot) + 2h \frac{\partial u(t+h,\cdot)}{\partial t} \| \Rightarrow$$

wegen der Differentialgleichung (11). Sei $\|A\|$ die Spektralnorm einer quadratischen Matrix A. Sie ist verträglich mit der EUKLID'ischen Vektornorm.

$$\rightarrow \leq \sum_{\nu=1}^{N} \frac{\|A_\nu\|}{R_\nu} \cdot \| (T_\nu - T_\nu^{-1}) u(t+h,\cdot) - 2h R_\nu \frac{\partial u(t+h,\cdot)}{\partial x_\nu} \|$$

$$+ \| (T_0 - T_0^{-1}) u(t+h,\cdot) - 2h \frac{\partial u(t+h,\cdot)}{\partial t} \|$$

$$\leq h^3 \cdot P_2 \cdot \left(\sum \frac{\|A_\nu\|}{R_\nu} + 1 \right) \quad \text{für } t, h \in [0,T] \text{ mit } t+2h \in [0,T]$$

wegen (12). Das ist genau die zu erfüllende Bedingung (4)' ($k=2$). $\Rightarrow$ Konsistenz von der Ordnung 2 für u_0.

B. Stabilität im strikt hyperbolischen Falle

Mit Hilfe von (10) läßt sich U^n in den Punkten eines Gitters

$$G := \{x \in \mathbb{R}^N ;\ x = h\cdot(j_1 R_1, \cdots, j_N R_N)\ ;\ j_1, \cdots, j_N \in \mathbb{Z}\} \tag{14}$$

berechnen, wenn man nur U^0 und U^1 auf diesem Gitter kennt.

Stabilität kann man mit Hilfe einer FOURIER-Transformation ermitteln (vgl.[1] S.79).

Die Funktionen $e^{i(k,x)}$, $k \in \tilde{G}$, bilden für

$$\tilde{G} := \{k \in \mathbb{R}^N ;\ k = 2\pi\left(\frac{j_1}{a_1-b_1}, \cdots, \frac{j_N}{a_N-b_N}\right) ;\ j_1, \cdots, j_N \in \mathbb{Z}\}$$

ein vollständiges Orthogonalsystem in $L_2(I)$. Nach dem Satz von RIESZ-FISCHER ist $U^n \in B$ daher eindeutig darstellbar als

$$U^n(x) = \frac{1}{\sqrt{|I|}} \sum_{k \in \tilde{G}} V^n(k)\, e^{i(k,x)} \tag{15}$$

Für die FOURIER-Koeffizienten gilt umgekehrt:

$$V^n(k) = \frac{1}{\sqrt{|I|}} \int_I U^n(x)\, e^{-i(k,x)}\, dx$$

Aus der Vollständigkeit des Orthonormalsystems folgt die PARSEVAL'sche Gleichung:

$$\|U^n\|^2 = \int_I |U^n(x)|^2 dx = \sum_{k \in \tilde{G}} |V^n(k)|^2 =: \|V^n\|^2_{\tilde{G}} \tag{16}$$

Setzt man (15) in (10) ein, so liefert ein Koeffizientenvergleich vor $e^{i(k,x)}$

$$V^{n+1}(k) - V^n(k) = \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu}\left(e^{i k_\nu R_\nu h} - e^{-i k_\nu R_\nu h}\right) V^n(k) \tag{17}$$

$$= 2i\, V^n(k) \sum \frac{A_\nu}{R_\nu} \sin k_\nu R_\nu h =: 2i\, V^n(k)\, A(k)$$

Durch die FOURIER-Transformation hat man ein Differenzenschema in $\hat{V}^n(k)$ erhalten, in das nur noch Differenzen in eine Richtung eingehen, das aber abhängig von einem Parameter ist.

Die naheliegendste Vereinfachung von (17) ist die Transformation von A(k) auf eine geeignete Normalform.

Sei (11) strikt hyperbolisch, dann ist A(k) diagonalisierbar; sei etwa

$$T^{-1}(k)\,A(k)\,T(k) = \Lambda(k) \qquad \text{für } k \in \mathbb{R}^N$$

mit nicht singulären Matrizen T(k) und mit Diagonalmatrizen $\Lambda(k)$ in denen die Eigenwerte von A(k) stehen.

Sei $\omega^n(k) := T^{-1}(k)\,\hat{V}^n(k)$

Um die Existenz von $\|\omega^0\|_{\tilde{G}}$, $\|\omega^1\|_{\tilde{G}}$ (≥ 0) zu wissen, muß man $\|T^{-1}(k)\| \leq C_{-1} = \text{const.}$ fordern, desgleichen braucht man für die andere Richtung $\|T(k)\| \leq C_{+1} = \text{const.}$, $C_{\pm 1} \in \mathbb{R}^+$ d.h., um aus der Beschränktheit von $\|\omega^n\|_{\tilde{G}}$ die von $\|\hat{V}^n\|_{\tilde{G}}$ zu folgern.

Diese Voraussetzung ist bedingt durch die Parameterabhängigkeit von (17). Im Falle HERMITE'scher Matrizen ($\Rightarrow A(k)$ HERMITEsch) gibt es sogar unitäre Matrizen $T(k)$, die $A(k)$ auf Diagonalform transformieren; die Voraussetzung ist dann also erfüllt.

$\omega^n(k)$ erfüllt das Differenzenschema (ohne Parameter k):

$$\omega^{n+1} - \omega^{n-1} = 2i\Lambda\,\omega^n \quad ; \quad \Lambda(k) = \begin{pmatrix} \lambda_1(k) & & 0 \\ & \lambda_2(k) & \\ 0 & & \lambda_m(k) \end{pmatrix},$$

welches in die Komponenten ω_ℓ von ω zerfällt:

$$\omega_\ell^{n+1} - \omega_\ell^{n-1} = 2i\lambda_\ell\,\omega_\ell^n \qquad , \ell = 1, \dots, m \tag{18}$$

Diese Differenzengleichung läßt sich bei Vorgabe von ω_ℓ^0, ω_ℓ^1 exakt lösen. Die charakeristische Gleichung von (18) ist (der Index ℓ wird unterdrückt):

$$\zeta^2 - 2i\lambda\zeta - 1 = 0$$

$$\Rightarrow \qquad \zeta_{1,2} = i\lambda \pm \sqrt{1-\lambda^2} =: \lambda_\pm$$

Für $\lambda_+ = \lambda_- =: \lambda^*$ d.h. $\lambda^* = \pm i$ hat man die Eigenlösungen $(\pm i)^n$ und $n(\pm i)^n$, $|\omega^n|$ bleibt daher in diesem Falle im allgemeinen nicht beschränkt.

Für $\lambda_+ \neq \lambda_-$ sind λ_+^n und λ_-^n die beiden Eigenlösungen von (18). Für vorgegebene Anfangswerte ω^0, ω^1 lautet daher die Lösung von (18):

$$\omega^n = a\lambda_+^n + b\lambda_-^n$$

wobei

$$a = \frac{\omega^1 - \omega^0\lambda_-}{\lambda_+ - \lambda_-} \; ; \; b = -\frac{\omega^1 - \omega^0\lambda_+}{\lambda_+ - \lambda_-} \; ; \; \lambda_+ - \lambda_- = 2\sqrt{1-\lambda^2} \neq 0$$

Offensichtlich ist für die Beschränktheit von $|\omega^n|$ im allgemeinen diejenige von λ_+^n und λ_-^n notwendig. Da λ und ω^n noch von einem Parameter K abhängen, ist zu erwarten, daß außerdem Nenner von a und b: $\lambda_+ - \lambda_- = 2\sqrt{1-\lambda^2}$ betragsmäßig nach unten durch eine von k unabhängige konstante $\varepsilon' > 0$ abzuschätzen ist: $|\lambda_+ - \lambda_-| \geq 2\varepsilon' > 0$

Zunächst hat man

$$|\lambda_+^n|, |\lambda_-^n| \leq \text{ const. zu erfüllen.}$$

$$\Longleftrightarrow \qquad |\lambda_+|, |\lambda_-| \leq 1$$

Wegen $\lambda_+ \cdot \lambda_- = -1$ ist das gleichbedeutend mit

$$|\lambda_+| = |\lambda_-| = 1$$

Es gibt daher einen Zweig des natürlichen Logarithmus, so daß $\log(-i\lambda_+)$ und $\log(-i\lambda_-)$ definiert sind.

$$-i\log(-i\lambda_\pm) = -i\log\left(\lambda \pm \sqrt{\lambda^2-1}\right) = \arccos\lambda$$

darin ist der durch die linke Seite definierte Zweig des $\arccos$ zu nehmen.

$$-i\lambda_\pm = e^{i\arccos\lambda}$$

Es soll sein $|\lambda_+| = |\lambda_-| = 1$

$\Longleftrightarrow$ $\arccos \lambda = b$

mit $b \in \mathbb{R}$, b beliebig.

$\Longleftrightarrow$ $\lambda = \cos b \quad \Rightarrow \lambda \in \mathbb{R}$

$\Longleftrightarrow$ $-1 \le \lambda \le 1$ (19)

D.h. $A(k)$ darf für alle Werte von k nur reelle Eigenwerte haben.

Das ist genau die Hyperbolizität von (11).

Um in k gleichmäßige Beschränktheit der Lösungen von (18) zu haben

wird außer (19) vorausgesetzt

$$|\lambda_+ - \lambda_-| = 2\sqrt{1-\lambda^2} \ge 2\varepsilon' > 0 \quad ; \ 0 < \varepsilon' < 1$$

$$\Longleftrightarrow \quad |\lambda| \le \sqrt{1-\varepsilon'^2} =: 1-\varepsilon \quad \Rightarrow 0 < \varepsilon < 1$$

(20) d.h. $-1+\varepsilon \le \lambda \le 1-\varepsilon$

(20) umfaßt offenbar (19).

Faßt man Beweisteile und Voraussetzungen zusammen, so erhält man die „Stabilität" von U^n:

U^0, U^1 seien vorgegeben, wegen (16) sind damit $\|V^0\|_{\tilde{G}}^2, \|V^1\|_{\tilde{G}}^2$ sowie $\|w^0\|_G^2, \|w^1\|_G^2$ bestimmt, wegen

$$|w^\alpha(k)|^2 = |T^{-1}(k) V^\alpha(k)|^2 \le \|T^{-1}(k)\|^2 |V^\alpha(k)|^2 \le C_{-1}^2 \|V^\alpha(k)\|^2$$

für $\alpha = 0, 1$

Es ist $w^n(k) = \Phi(k)\left[\left(\Lambda_+^n(k) - \Lambda_-^n(k)\right) w^1(k) + \left(\Lambda_+^{n-1}(k) - \Lambda_-^{n-1}(k)\right) w^0(k)\right]$

mit mxm Diagonalmatrizen $\Phi(k), \Lambda_+(k), \Lambda_-(k)$ definiert durch:

$$\Phi(k) := \begin{pmatrix} [\lambda_{1+}(k) - \lambda_{1-}(k)]^{-1} & & \\ & \ddots & \\ & & [\lambda_{m+}(k) - \lambda_{m-}(k)]^{-1} \end{pmatrix}; \quad \Lambda_\pm(k) := \begin{pmatrix} \lambda_{1\pm}(k) & & \\ & \ddots & \\ & & \lambda_{m\pm}(k) \end{pmatrix}$$

$$\Rightarrow |\omega^n(k)| \le \|\Phi(k)\| \cdot \left[\left(\|\Lambda_+^n(k)\| + \|\Lambda_-^n(k)\| \right) \cdot |\omega^1(k)| + \left(\|\Lambda_+^{n-1}(k)\| + \right.\right.$$

hier ist wieder die Spektralnorm gemeint. $\left.\left. + \|\Lambda_-^{n-1}(k)\| \right) |\omega^0(k)| \right] \le$

$$\le \frac{1}{\min\limits_{\ell=1,\dots,m} |\lambda_{\ell+} - \lambda_{\ell-}|} \left[\left(\max_{\ell=1,\dots,m} |\lambda_{\ell+}|^n + \max |\lambda_{\ell-}|^n \right) |\omega^1(k)| + \right.$$

$$\left. + \left(\max |\lambda_{\ell+}|^{n-1} + \max |\lambda_{\ell-}|^{n-1} \right) |\omega^0(k)| \right] =$$

$$= \frac{2}{\min |\lambda_{\ell+} - \lambda_{\ell-}|} \left[|\omega^1(k)| + |\omega^0(k)| \right] \le$$

$$\le \frac{1}{\varepsilon'} \left[|\omega^1(k)| + |\omega^0(k)| \right] =$$

$$= \frac{1}{\sqrt{1-(1-\varepsilon)^2}} \left[|\omega^1(k)| + |\omega^0(k)| \right]$$

$$\Rightarrow |\omega^n(k)|^2 \le \frac{2}{\varepsilon'^2} \left[|\omega^1(k)|^2 + |\omega^0(k)|^2 \right]$$

$$\Rightarrow \|\omega^n\|_{\tilde{G}_1}^2 \le \frac{2}{\varepsilon'^2} \left[\|\omega^1\|_{\tilde{G}_1}^2 + \|\omega^0\|_{\tilde{G}_1}^2 \right]$$

$$\Rightarrow \|U^n\|^2 = \|V^n\|_{\tilde{G}_1}^2 \le \frac{2C^2C_{-1}^2}{\varepsilon'^2} \left[\|V^1\|_{\tilde{G}_1}^2 + \|V^0\|_{\tilde{G}_1}^2 \right] = \frac{2C^2C_{-1}^2}{\varepsilon'^2} \left[\|U^1\|^2 + \|U^0\|^2 \right]$$

d.h. Stabilität

Für alle Matrizen $D \in \mathbb{R}^m \times \mathbb{R}^m$ sei $|D| := \max \{ |\lambda_1|, \lambda_\nu$ Eigenwert von $D \}$

Dann ist bewiesen der

Satz 2:

Das Differenzenschema

$$(10) \qquad U^{n+1} - U^{n-1} = \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \left(T_\nu - T_\nu^{-1} \right) U^n$$

ist stabil, wenn die folgenden drei Aussagen gelten:

a) Das Differentialgleichungssystem

$$(11) \qquad u_t(t,x) = \sum_{\nu=1}^{N} A_\nu u_{x_\nu}(t,x) \quad \text{für} \quad 0 \le t \le T \text{ und } x \in I$$

ist strikt hyperbolisch; d.h.:

Es gibt ein $\mu > 0$, so daß für alle $\xi \in W_\mu(0) := \{\xi \in \mathbb{R}^N \mid |\xi_\nu| \le \mu \text{ für } \nu = 1, \dots, N\}$ die Matrix $B(\xi) := \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \xi_\nu$ mit einer Matrix $S(\xi)$ diagonalisierbar ist, d.h. $S^{-1}(\xi)$ existiert und $S^{-1}(\xi) B(\xi) S(\xi)$ ist eine Diagonalmatrix.

b) Es gibt Konstanten $C, C_{-1} \in \mathbb{R}^+$, so daß

$$\|S^{-1}(\xi)\| \le C_{-1}; \quad \|S(\xi)\| \le C \text{ für alle } \xi \in W_\mu(0)$$

c) (eigentliche Stabilitätsbedingung)

Es gibt ein $\zeta > 0$, so daß

(21) $$|B(\xi)| \le 1 - \zeta \quad \text{für alle } \xi \in W_1(0)$$

Bemerkung 1:

Daß hier statt $W_\mu(0) = \{\xi \in \mathbb{R}^N; \xi_\nu = \sin(k_\nu R_\nu h) \text{ für } \nu = 1, 2, \dots, N; k \in \mathbb{R}^N\}$ der Würfel $W_\mu(0)$ gesetzt wurde, ist erlaubt wegen $W_1(0) = \{\xi = \frac{z}{\mu}; z \in W_\mu(0)\}$, und weil jede $m \times m$ Matrix $a \cdot M$ für $a \in \mathbb{R}$ diagonalisierbar ist durch eine Matrix S, wenn das für M gilt.

Bemerkung 2:

Falls (11) nicht hyperbolisch oder $|A(k)| \ge 1 + \zeta$ für ein $\zeta \in \mathbb{R}^+$ und für alle $k \in \overline{N} \subset \mathbb{R}^N$, wobei $\overline{N}$ nicht eine Nullmenge ist, so wächst $\|V^n\|_G$ und damit $\|U^n\|_I$ exponentiell mit $(1+\zeta)^n$, d.h. es gibt ein $b \in \mathbb{R}^+$, so daß

$$\|U^n\| \ge b\,(1+\zeta)^n$$

Bemerkung 3:

Im Falle HERMITEscher Systeme, d.h. für A_ν HERMITESCH ($\nu = 1, 2, \dots, N$) hat man nur die Bedingung c) zu erfüllen, denn $\sum_1 \frac{A_\nu}{R_\nu} \xi_\nu$ ist dann für

alle $\xi \in \mathbb{R}^N$ HERMITEsch.

Man kann noch die Aussage des Satzes abschwächen durch:

$$\xi \in W_1(0): \quad |B(\xi)| = \left|\sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \xi_\nu\right| = \left|\max_{\substack{x \in \mathbb{R}^N \\ |x|=1}} \sum_1 \frac{(A_\nu x, x)}{R_\nu} \xi_\nu\right| \leq$$

$$\leq \sum_1 \frac{|\xi_\nu|}{R_\nu} \left|\max_{\substack{x \in \mathbb{R}^N \\ |x|=1}} (A_\nu x, x)\right| = \sum_1 \frac{|\xi_\nu|}{R_\nu} |A_\nu| \leq \sum_1 \frac{|A_\nu|}{R_\nu}$$

wegen der Extremaleigenschaft HERMITEscher Formen.

$\Rightarrow$ Stabilität herrscht für

$$\sum_{\nu=1}^{N} \frac{|A_\nu|}{R_\nu} \leq 1 - \eta$$

Bemerkung 4:

Unter den Voraussetzungen des Satzes existiert insbesondere zu einem wie unter II.A. vorausgesetzten u_0 eine eindeutige Lösung von (11). Jede am Anfang von der Ordnung 2 approximierende Schar von Lösungen von (10) konvergiert daher von derselben Ordnung gegen $u(t)$. Die Wahl von $T \in \mathbb{R}^+$ ist hier willkürlich.

C. Stabilitätsbereich und Abhängigkeitsgebiet

Es läßt sich zeigen, daß das Schema (10) in der Geometrie seines Gitters das Schema mit dem größten überhaupt möglichen Stabilitätsbereich ist. Diese Eigenschaft ist von praktischer Bedeutung, da erfahrungsgemäß für die größten Zeitschrittweiten, für die das Schema noch stabil ist, der Fehler am kleinsten wird.
In ihrer klassischen Arbeit [2] haben COURANT, FRIEDRICHS und LEWY Instabilität unter dem Gesichtspunkt von Abhängigkeitsfragen behandelt. Die Lösung $u(t,x)$ des Anfangswertproblems (11) ist an der Stelle (t,x); $t \in \mathbb{R}^+$, eine eindeutige Funktion von Punkten $(0,x^*)$ aus kompakten Mengen in der Hyperebene t = 0. Der Durchschnitt aller dieser Mengen ist das Abhängigkeitsgebiet $D(t,x)$ des Punktes (t,x), s. LAX [4]. Für einen Punkt $(t_\ell, x_\ell) \in [0,T] \times G =: G_\ell^T$, $t_\ell = n\ell$ gibt es eine endliche Menge $\hat{D}_\ell(t_\ell, x_\ell)$ in der Anfangsebene t = 0, von der der Wert der Lösung von (10) in (t_ℓ, x_ℓ) (abgesehen von Punkten in der zweiten Schicht) allein abhängt. Die konvexe Hülle $D_\ell(t_\ell, x_\ell)$ von $\hat{D}_\ell(t_\ell, x_\ell)$ bezeichnet man als das Abhängigkeitsgebiet des Differenzenschemas im Gitter für den Punkt (t_ℓ, x_ℓ), kurz "Gitterabhängigkeitsgebiet". Wenn die Zeitschrittweite so groß im Verhältnis zu den Raumschrittweiten gewählt wird, daß $F_\ell(t_\ell, x_\ell) := D(t_\ell, x_\ell) \setminus D_\ell(t_\ell, x_\ell) \neq \emptyset$, so kann man den Anfangswert $u_0(\cdot)$ in $F_\ell(t_\ell, x_\ell)$ nachträglich so verändern, daß die zugehörige exakte Lösung $u(t,x)$ in (t_ℓ, x_ℓ) einen anderen Wert bekommt, während sich am Wert der Lösung des Differenzenschemas u_j^n an dieser Stelle nichts ändert ($j = (j_1, \cdots, j_N)$).

Bei Verfeinerung des Gitters kann daher U_j^n nicht gegen $u(t,x)$ konvergieren, da es unendlich viele exakte Lösungen von (11) gibt, die mit U_j^0, U_j^1 für alle möglichen j beliebig genau übereinstimmen.

Die Forderung $D(t_h, x_h) \subset D_h(t_h, x_h)$ für alle $(t_h, x_h) \in G_h^T$ bedeutet daher eine unmittelbar einleutende, "natürliche" Bedingung für Stabilität eines Schemas. Was über sie hinausgeht wird man als ein Maß für Ungereimtheiten betrachten dürfen, die durch die Konstruktion des Differenzenschemas, nicht durch das zu approximierende Anfangswertproblem, herbeigeführt sind.

Da das System (11) linear und homogen mit konstanten Koeffizienten ist, genügt es das Abhängigkeitsgebiet D des Punktes $(1, 0) \in \mathbb{R}^{N+1}$ zu betrachten. Dazu werden solche G_h^T gewählt, die diesen Punkt enthalten; das zugehörige Gitterabhängigkeitsgebiet sei D_h.

Wegen der Konvexität von D_h ist

$$D \subset D_h \iff H \subset D_h \tag{22}$$

worin H die konvexe Hülle von D ist.

D_h läßt sich leicht bestimmen: zur Berechnung von U_j^{n+1} werden im Schema (10) die Werte $T_\nu U_j^n$, $T_\nu^{-1} U_j^n$, U_j^{n-1} für $\nu = 1, 2, \dots, N$ benötigt. Stellt man sich G_h^T noch um eine Schicht $\{-h\} \times G$ erweitert vor und läßt das Schema (10) in ihr beginnen, so ist D_h der N-dimensionale Rhombus mit den Eckpunkten

$(R_1, 0, \dots, 0)$, $(-R_1, 0, \dots, 0)$
$(0, R_2, 0, \dots, 0)$, $(0, -R_2, 0, \dots, 0)$
$\vdots$ $\vdots$
$(0, 0, \dots 0, R_N)$, $(0, 0, \dots, 0, -R_N)$

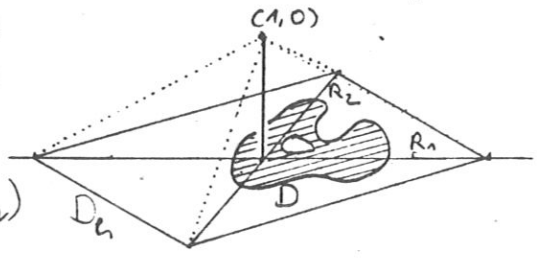


Fig. 1

d.h. die konvexe Hülle der aus diesen Punkten bestehenden Menge (s.Figur 1).

H ist von LAX für das Problem (11) in [4] durch die Trägerfunktion "f_H" bestimmt worden, die für eine beliebige, beschränkte Menge $M \subset \mathbb{R}^N$ so definiert ist:

$$f_M(k) := \sup_{x \in M} k \cdot x \qquad \text{für alle } k \in \mathbb{R}^N$$

Mit $k \cdot x$ ist das Skalarprodukt $k \cdot x = \sum_{i=1}^{N} k_i \cdot x_i$ gemeint.
Da M beschränkt, existiert $f_M(k)$ immer.
f_M ist durch die konvexe Hülle von M eindeutig bestimmt und umgekehrt. Für zwei beschränkte konvexe Mengen M, M' gilt genau dann

$$M' \subset M''$$

wenn $f_{M'}(k) \leq f_{M''}(k)$ für alle $k \in \mathbb{R}^N$

Die Notwendigkeit dieser Bedingung ist anhand der Definition von f_M unmittelbar einzusehen. Daß sie auch hinreichend ist, folgt genauso wie die eindeutige Bestimmtheit aus der Theorie der konvexen Mengen (s. LAX [4]).
Da D_Δ und H konvex sind, lautet die COURANT-FRIEDRICHS-LEWY-(CFL) Bedingung (22):

(23) $$f_H(k) \leq f_{D_\Delta}(k) \qquad \text{für alle } k \in \mathbb{R}^N$$

Wegen der Homogenität von f_M für nicht-negative Parameter:

$$f_M(a \cdot k) = a \cdot f_M(k) \qquad \text{für } a \in \mathbb{R}^+ \cup \{0\},\ k \in \mathbb{R}^N$$

ist (23) äquivalent zu

(23') $$f_H(k) \leq f_{D_\Delta}(k) \qquad \text{für alle } k \in P$$

mit einer Menge P, die die Bedingung

$$\mathbb{R}^N = \{ k' \in \mathbb{R}^N ; \text{ es gibt } k \in P, b \in \mathbb{R}^+ \cup \{0\} \text{ mit } k' = b \cdot k \} \text{ erfüllt.}$$

Eine solche Menge P, die von jeder Halbgeraden mit 0 als Anfangspunkt geschnitten wird, ist zum Beispiel der Rand ∂Q_R des N-dimensionalen Quaders Q_R :

$$Q_R := \{ k \in \mathbb{R}^N ; |k_\nu| \le \frac{1}{R_\nu} \text{ für } \nu = 1, 2, \cdots, N \} ; \; R := (R_1, R_2, \cdots, R_N) \quad (R_\nu > 0 \text{ für } \nu = 1, \cdots, N)$$

Darauf läßt sich f_{D_k} bestimmen:

Behauptung:

$$f_{D_k}(k) = 1 \qquad \text{für alle} \quad k \in \partial Q_R$$

Beweis:

Wegen der Symmetrie des Rhombus D_k und des Quaders Q_R gibt es zu jedem $x \in D_k$ ein $x^+ \in D_k$ mit $x_\nu^+ = |x_\nu|$, zu jedem $k \in \partial Q_R$ ein $k^+ \in \partial Q_R$ mit $k_\nu^+ = |k_\nu|$ und für $y \in D_k$, $k \in \partial Q_R$ ein $\tilde{y} \in D_k$ mit $|\tilde{y}_\nu| = |y_\nu|$ und $\tilde{y}_\nu \cdot k_\nu \ge 0$ für $\nu = 1, 2, \cdots, N$.

$$\Rightarrow \quad y \cdot k \le \tilde{y} \cdot k = y^+ \cdot k^+$$

$$\Rightarrow \quad f_{D_k}(k) = f_{D_k^+}(k^+) \qquad \text{für } k \in \partial Q_R$$

mit $D_k^+ := \{ x \in D_k, x_\nu \ge 0 \text{ für } \nu = 1, 2, \cdots, N \}$

oder anders geschrieben:

$$f_{D_k}(k) = f_{D_k^+}(k^+) \qquad \text{für } k^+ \in \partial Q_R^+$$

mit analogen ∂Q_R^+ :

$$\partial Q_R^+ := \{ k \in \partial Q_R, k_\nu \ge 0 \text{ für } \nu = 1, 2, \cdots, N \}$$

D_k^+ ist das von der Hyperebene $\sum_{\nu=1}^{N} \frac{x_\nu}{R_\nu} = 1$ und den Koordinantenebenen $x_\varkappa = 0$, $\varkappa = 1, 2, \cdots, N$ erzeugte Simplex:

$$D_k^+ = \{ x \in \mathbb{R}^N, \sum_{\nu=1}^{N} \frac{x_\nu}{R_\nu} \le 1 ; x_\varkappa \ge 0 \text{ für } \varkappa = 1, 2, \cdots, N \}$$

Sei $k \in \partial Q_R^+$ fest. Dann gibt es ein $\ell \in \mathbb{N}$, $\ell \leq N$ mit

$k_\ell = \frac{1}{R_\ell}$.

Für $x \in D_k^+$ ist:

$$x_\ell \leq R_\ell \cdot \left(1 - \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{N} \frac{x_\nu}{R_\nu}\right)$$

$$\Rightarrow \quad k \cdot x = \sum_{1}^{N} k_\nu x_\nu \leq \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{N} k_\nu x_\nu + \frac{1}{R_\ell} \cdot R_\ell \left(1 - \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{N} \frac{x_\nu}{k_\nu}\right)$$

$$= \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{N} \left(k_\nu - \frac{1}{R_\nu}\right) x_\nu + 1$$

$$\leq 1 \qquad \text{da } k_\nu \leq \frac{1}{R_\nu},\ x_\nu \geq 0 \text{ für } \nu = 1, 2, \cdots, N$$

Für $x_{R_\ell} := (0, \cdots, 0, R_\ell, 0, \cdots, 0) \in D_k^+$, d.h. die ℓ-te Komponente ist größer als Null, ist aber $k \cdot x_{R_\ell} = 1$

$\Rightarrow$ Für alle $k \in \partial Q_R^+$ existiert sogar $\max_{x \in D_k^+} k \cdot x$ und ist gleich **1**.

Es ist also

$$f_{D_k^+}(k^+) = f_{D_k}(k) = 1 \quad \text{für alle } k \in \partial Q_R \qquad \text{q.e.d.}$$

LAX hat nun in [4] die Trägerfunktion f_H berechnet zu:

$$f_H(k) = \lambda_{\max}\left(\sum_{\nu=1}^{N} k_\nu \cdot A_\nu\right)$$

worin $\lambda_{\max}(M)$ den größten Eigenwert der $m \times m$ Matrix M bezeichnet.

Wegen (23)' ist daher die CFL-Bedingung äquivalent zu:

$$\lambda_{\max}\left(\sum k_\nu \cdot A_\nu\right) \leq 1 \quad \text{für alle } k \in \partial Q_R$$

$$\Leftrightarrow \quad \lambda_{\max}\left(\sum k_\nu \cdot \frac{A_\nu}{R_\nu}\right) \leq 1 \quad \text{für alle } k \in \partial W_1 \ (:= \partial W_1(0))$$

mit einem Würfel $W_1 := \{k \in \mathbb{R}^N;\ |k_\nu| \leq 1 \text{ für } \nu = 1, 2, \cdots, N\}$

$$\Leftrightarrow \quad \lambda_{\max}\left(\sum k_\nu \cdot \frac{A_\nu}{R_\nu}\right) \leq 1 \quad \text{für alle } k \in W_1$$

$\longrightarrow$

Denn für $k' \in W_1$ gibt es ein $b \in [0,1]$ und ein $k \in \partial W_1$ mit $k' = b \cdot k$.

Wegen $\lambda_{max}(-B) = -\lambda_{min}(B)$ und $k \in W_1 \Leftrightarrow -k \in W_1$ folgt:

$$\Rightarrow\Leftarrow \quad \lambda_{min}\left(\sum_{\nu=1}^{N} k_\nu \frac{A_\nu}{R_\nu}\right) \geq -1 \quad \text{für } k \in W_1$$

Man hat damit bewiesen:

Satz 3: Das Schema (10) erfüllt genau dann die CFL-Bedingung für (11), wenn

$$\left|\sum_{\nu=1}^{N} k_\nu \frac{A_\nu}{R_\nu}\right| \leq 1 \quad \text{für alle} \quad k \in W_1$$

Dies ist bis auf ein beliebig klein wählbares $\zeta > 0$ genau (21), die Stabilitätsbedingung c) des Satzes 1. Aus den Ergebnissen des folgenden Abschnitts wird man ersehen, daß das ζ die Folge schwacher d.h. polynomartiger Instabilitäten, im Gegensatz zu den exponentiellen, ist.

III. Schwache Stabilität

In diesem Abschnitt wird der Fall der nicht nur strikten Hyperbolizität von (11) untersucht. (10) ist dann nicht mehr wie in II. stabil sondern i.a. schwach stabil.

Man wählt wieder die Form (13) des Schemas (10):

(13) $$\tilde{U}^{n+1} = \begin{pmatrix} \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu}(T_\nu - T_\nu^{-1}) & I \\ I & 0 \end{pmatrix} \tilde{U}^n \quad \text{mit} \quad \tilde{U}^n := \begin{pmatrix} U^{n+1} \\ U^n \end{pmatrix}$$

$$= \tilde{C}(h)\,\tilde{U}^n$$

$\tilde{U}^n = \tilde{U}^n(\cdot)$ ist darin eine Funktion abhängig von $x \in \mathbb{R}^N$.

Sei jetzt einfach $h \in \mathbb{R}^+$ ohne die Beschränkung $h \in [0,T]$.

Mit der Argumentation in (15)-(17) erhält man für die Transformierte:

$$\tilde{V}^{n+1} = \underbrace{\begin{pmatrix} 2i \sum \frac{A_\nu}{R_\nu} \sin k_\nu R_\nu h & I \\ I & 0 \end{pmatrix}}_{=: \tilde{A}_h(k)} \tilde{V}^n(k) \quad \text{für } k \in \mathbb{R}^N$$

$$\text{mit } \tilde{V}^n(k) = \begin{pmatrix} V^{n+1}(k) \\ V^n(k) \end{pmatrix}$$

Um schwache Stabilität zu beweisen hat man daher die Potenzen der $2m \times 2m$ Matrix $\tilde{A}_h(k)$ zu betrachten. Dazu bietet sich die Wahl einer Normalform an. Hier wird vorläufig die von JORDAN genommen, und zwar seien:

$$A_h(k) := \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \sin k_\nu R_\nu h \,,$$

$J_h(k) = T_h^{-1}(k) A_h(k) T(k)$ mit regulären Matrizen $T_h(k)$ für $k \in \mathbb{R}^N$

$J_h(k)$ ist eine $m \times m$ Diagonalübermatrix:

$$J_h(k) := \mathrm{Diag}\left(J_{h,1}(k), J_{h,2}(k), \cdots, J_{h,r}(k)\right)$$

mit $m_\nu \times m_\nu$ Matrizen $J_{h,\nu}(k)$ für $\nu = 1, 2, \cdots, r$,

und $\sum_{\nu=1}^{r} m_\nu = m$.

Die $J_{h,\nu}(k)$ sind wieder Diagonalübermatrizen.

$$J_{h,\nu}(k) := \mathrm{Diag}\left(J_{h,\nu,1}(k), J_{h,\nu,2}(k), \cdots, J_{h,\nu,s}(k)\right)$$

mit $m_{\nu\mu} \times m_{\nu\mu}$ Matrizen $J_{h,\nu,\mu}(k)$ für $\mu = 1, 2, \cdots, s$,

und $\sum_{\mu=1}^{s} m_{\nu\mu} = m_\nu$. $J_{h,\nu,\mu}(k)$ hat schließlich die Gestalt:

$$J_{h,\nu,\mu}(k) = \lambda_{h,\nu}(k)\, I_{m_{\nu\mu}} + N_{m_{\nu\mu}}$$

Die $\lambda_{h,1}(k), \cdots, \lambda_{h,r}(k)$ sind die r verschiedenen Eigenwerte von $A_h(k)$; $I_{m_{\nu\mu}}$ ist die $m_{\nu\mu} \times m_{\nu\mu}$ Einheitsmatrix und $N_{m_{\nu\mu}}$ mit derselben Zeilenlänge quadratisch und aus lauter Nullen außer $m_{\nu\mu} - 1$ Einsen in der ersten oberen Nebendiagonale bestehend:

$$I_{m_{\nu\mu}} := \underbrace{\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}}_{m_{\nu\mu}} ; \quad N_{m_{\nu\mu}} := \underbrace{\begin{pmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}}_{m_{\nu\mu}}$$

$N_{m_{\nu\mu}}$ ist also nilpotent.

Dann gilt:
$$\begin{pmatrix} T_h^{-1}(k) & 0 \\ 0 & T_h^{-1}(k) \end{pmatrix} \tilde{A}_h(k) \begin{pmatrix} T_h(k) & 0 \\ 0 & T_h(k) \end{pmatrix} =$$

$$= \begin{pmatrix} T_h^{-1}(k) & 0 \\ 0 & T_h^{-1}(k) \end{pmatrix} \begin{pmatrix} 2i\, A_h(k) & I \\ I & 0 \end{pmatrix} \begin{pmatrix} T_h(k) & 0 \\ 0 & T_h(k) \end{pmatrix} =$$

$$= \begin{pmatrix} 2i\, J_h(k) & I \\ I & 0 \end{pmatrix} =: \tilde{J}_h(k)$$

Das Verhalten von $\tilde{A}_{\ell}(k)^n$ ist daher bis auf Transformationsmatrizen durch das von $\tilde{J}_h(k)^n$ gegeben. Die Multiplikation der Untermatrizen gebildet aus $\tilde{J}_{\ell}(k)$ durch Wegnahme

der Zeilen von der 1.

bis $\sum_{x=1}^{\nu} m_{\nu x} + \sum_{x=1}^{\mu-1} m_{\nu x}$ ten,

der $\sum_{x=1}^{\nu} m_{\nu x} + \sum_{x=1}^{\mu} m_{\nu x} + 1$ ten

bis zur

$\sum_{x=1}^{\nu} m_{\nu x} + \sum_{x=1}^{\mu-1} m_{\nu x} + m$ ten

und von der

$\sum_{x=1}^{\nu} m_{\nu x} + \sum_{x=1}^{\mu} m_{\nu x} + 1 + m$ ten

bis zur $2m$ ten Zeile

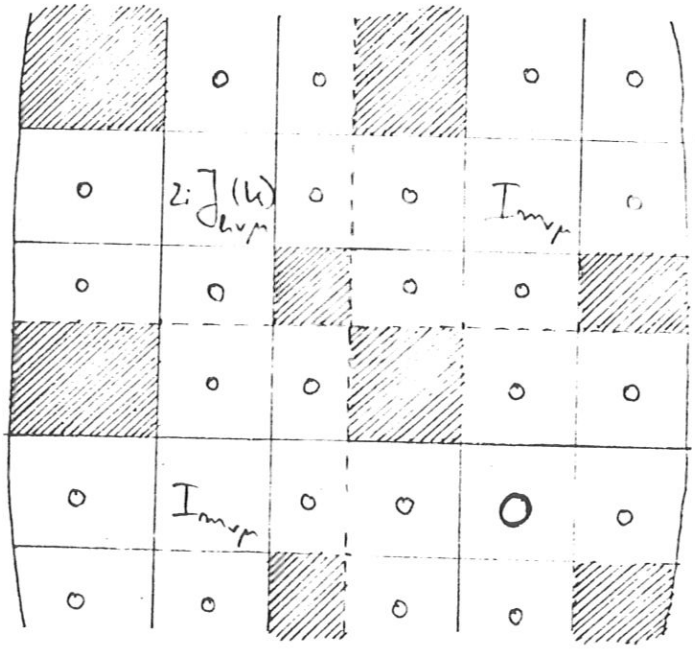


Fig. 2

und derselben Spalten, d.h. die Multiplikation der Matrizen

$$\begin{pmatrix} 2i\,J_{\lambda,\nu,\mu}(k) & I_{m_{\nu\mu}} \\ I_{m_{\nu\mu}} & 0 \end{pmatrix}$$

ist voneinander unabhängig, wegen des Schachbrettaufbaus von $\tilde{J}_h(k)$ (s.Fig. 2). Man kann sich daher auf die Potenzen dieser Matrizen beschränken.

Hilfssatz:

Das charakteristische und das Minimalpolynom der 2px2p Matrix ($p \in \mathbb{N}$)

$$\tilde{J} := \begin{pmatrix} 2i J_\lambda & I_p \\ I_p & 0 \end{pmatrix} \quad \text{mit} \quad J_\lambda := \lambda I_p + N_p = \begin{pmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \text{ für } \lambda \in \mathbb{C}$$

sind gegeben durch $(\zeta^2 - 2i\lambda\zeta - 1)^p$.

Korollar:

Für $\lambda \neq \pm 1$ ist die JORDAN'sche Normalform $\tilde{J}_0$ von $\tilde{J}$ gegeben durch $\begin{pmatrix} J_{\lambda_+} & 0 \\ 0 & J_{\lambda_-} \end{pmatrix}$ mit $\lambda_\pm := i\lambda \pm \sqrt{1-\lambda^2}$ (bis auf Vertauschung von $J_{\lambda_+}, J_{\lambda_-}$),

und für $\lambda = \pm 1$ durch $\tilde{J}_0 = \pm i I_{2p} + N_{2p}$

Die Lösungen der charakeristischen Gleichung $(\zeta^2 - 2i\lambda\zeta - 1)^p = 0$ sind nämlich λ_+ und λ_-, und zwar jeweils von der Vielfachheit p.

$$\lambda_+ \lambda_- = -1 \quad , \quad \lambda_+ + \lambda_- = 2i\lambda$$

Da das Minimalpolynom mit den charakteristischen Polynom zusammenfällt ist die Länge der JORDAN-Matrizen dann p bzw. 2p. Für den Beweis des Hilfssatzes ist nützlich das

Lemma:

$$D_{2p} := \det \left(\begin{array}{c|c} a I_p + b N_p & I_p \\ \hline I_p & c I_p \end{array} \right) =$$

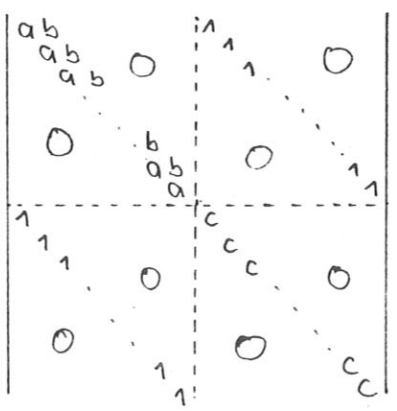
für $a, b, c \in \mathbb{C}, p \in \mathbb{N}$

Beweis des Lemmas durch vollständige Induktion über p:

Für p = 1 ist $D_2 = \det \begin{pmatrix} a & 1 \\ 1 & c \end{pmatrix} = (ac-1)^1$

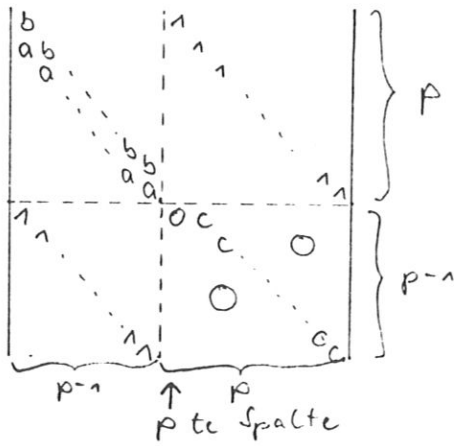
Die Behauptung sei für p-1 bewiesen:

Sei $|D| := \det(D)$ für jede quadratische Matrix D.

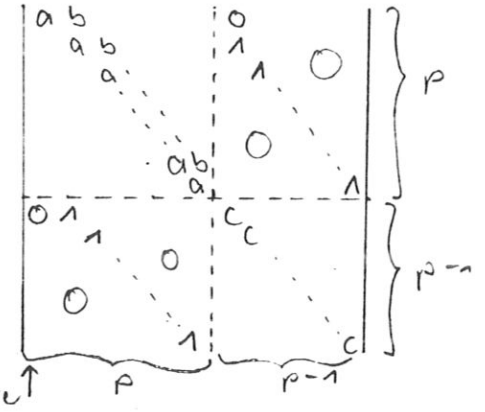
$D_{2p} =$ 

Entwicklung nach der p + 1 ten Zeile ←

$$D_{2p} = (-1)^{p+1+1} \cdot$$

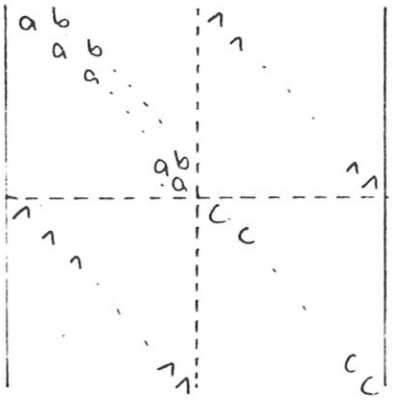


$$+ c \cdot (-1)^{p+1+p+1} \cdot$$



Die entstandenen Determinanten lassen sich nach der $p$ten bzw. 1. Spalte entwickeln, wobei dieselbe Restdeterminante übrig bleibt.

$$D_{2p} = [(-1)^p \cdot (-1)^{p+1} + c \cdot a] \cdot$$



$$= (ac-1) \cdot D_{2p-2}$$

$$= (ac-1)^p$$ nach Induktionsvoraussetzung.

Anmerkung: b geht nicht ein in die Determinante. q.e.d. (Lemma)

Beweis des Hilfssatzes:

Setzt man $a = 2i\lambda - b$, $b = 2i$, $c = -b$, so hat man die Determinante von

$$\left(\begin{array}{c|c} 2i\,\mathcal{J}_\lambda - b\,I_\rho & I_\rho \\ \hline I_\rho & -b\,I_\rho \end{array}\right) \quad \text{zu } (b^2 - 2i\lambda b - 1)^\rho \text{ bestimmt.}$$

Damit ist die Gestalt des charakteristischen Polynoms geklärt.

Das Minimalpolynom muß im Falle $\lambda_+ \neq \lambda_-$, d.h. $\lambda \neq \pm 1$,

den Faktor $b^2 - 2i\lambda b - 1$ enthalten, da das charakeristische Polynom dann diese beiden verschiedenen Nullstellen besitzt.

Es ist:

$$\tilde{\mathcal{J}}^2 = \left(\begin{array}{c|c} (2i\,\mathcal{J}_\lambda)^2 + I_\rho & 2i\,\mathcal{J}_\lambda \\ \hline 2i\,\mathcal{J}_\lambda & I_\rho \end{array}\right)$$

und weiter

$$\tilde{\mathcal{J}}^2 - 2i\lambda\tilde{\mathcal{J}} - I_{2\rho} = \left(\begin{array}{c|c} (2i\,\mathcal{J}_\lambda)^2 - (2i)^2\lambda\,\mathcal{J}_\lambda & 2i\,\mathcal{J}_\lambda - 2i\lambda\,I_\rho \\ \hline 2i\,\mathcal{J}_\lambda - 2i\lambda I_\rho & 0 \end{array}\right)$$

Für $\mathcal{J}_\lambda$ hat man: $\mathcal{J}_\lambda^2 = \lambda^2 I_\rho + 2\lambda N_\rho + N_\rho^2$

$$\Rightarrow \mathcal{J}_\lambda^2 - 2\mathcal{J}_\lambda = \lambda^2 I_\rho + 2\lambda N_\rho + N_\rho^2 - \lambda^2 I_\rho - 2\lambda N_\rho =$$
$$= N_\rho(2\lambda + N_\rho) = N_\rho\,\mathcal{J}_\lambda$$

$$\Rightarrow \tilde{\mathcal{J}}^2 - 2i\lambda\tilde{\mathcal{J}} - I_{2\rho} = \left(\begin{array}{c|c} -4N_\rho\,\mathcal{J}_\lambda & 2i\,N_\rho \\ \hline 2i\,N_\rho & 0 \end{array}\right) = 2i\begin{pmatrix} N_\rho & 0 \\ 0 & N_\rho \end{pmatrix}\begin{pmatrix} 2i\,\mathcal{J}_\lambda & I_\rho \\ I_\rho & 0 \end{pmatrix}$$

$$= 2i\begin{pmatrix} N_\rho & 0 \\ 0 & N_\rho \end{pmatrix}\tilde{\mathcal{J}}$$

Die beiden letzten Matrizen darf man vertauschen, denn N_ρ kommutiert mit $\mathcal{J}_\lambda = \lambda I_\rho + N_\rho$. $\tilde{\mathcal{J}}$ ist eine nicht-singuläre Matrix, denn sie besitzt nur die Eigenwerte λ_+, λ_- mit $\lambda_+ \cdot \lambda_- = -1 \Rightarrow$

$$\lambda_+, \lambda_- \neq 0$$

N_ρ dagegen ist nilpotent, und zwar ist $\varkappa = \rho$ in $N_\rho^{\,\varkappa} = 0$ der kleinste Exponent, der diese Relation erfüllt. Daher verschwindet

$(\tilde{J}^2 - 2i\lambda\tilde{J} - I_{2\rho})^{\varkappa}$ zuerst für $\varkappa = \rho$, d.h.:

$(\zeta^2 - 2i\lambda\zeta - 1)^{\rho}$ ist auch das Minimalpolynom für $\lambda \neq \pm 1$

Im Falle $\lambda = +1$ oder -1 hat das charakteristische Polynom die Gestalt:

$$(\zeta \mp i)^{2\rho}$$

Man weiß daher nur, daß das Minimalpolynom den linearen Faktor $\zeta \mp i$ enthält und irgendeine Potenz mit einem Exponenten $\leq 2\rho$ davon ist.

Der Fall gerader Exponenten läßt sich wie oben behandeln:

$$(\tilde{J} \mp iI_{2\rho})^{2\varkappa} = (\tilde{J}^2 \mp 2i\tilde{J} - I_{2\rho})^{\varkappa} = (2i)^{\varkappa} \begin{pmatrix} N_\rho & 0 \\ 0 & N_\rho \end{pmatrix}^{\varkappa} \tilde{J}^{\varkappa}$$

und das verschwindet zuerst für $\varkappa = \rho$. Falls

$$(\tilde{J} \mp iI_{2\rho})^{\tau} = 0$$

für ein ungerades $\tau \in \mathbb{N}$, $\tau < 2\rho$,

so müßte insbesondere $(\tilde{J} \mp iI_{2\rho})^{2\rho-1}$ gleich der Nullmatrix sein; für jene gilt:

$$(\tilde{J} \mp iI_{2\rho})^{2\rho-1} = (\tilde{J} \mp iI_{2\rho})^{2(\rho-1)} (\tilde{J} \mp iI_{2\rho}) =$$

$$= (2i)^{\rho-1} \tilde{J}^{\rho-1} \left(\begin{array}{c|c} N_\rho^{\rho-1} & 0 \\ \hline 0 & N_\rho^{\rho-1} \end{array}\right) \left(\begin{array}{c|c} 2i J_{\pm 1} \mp iI_\rho & I_\rho \\ \hline I_\rho & \mp iI_\rho \end{array}\right) =$$

$$= (2i)^{\rho-1} \tilde{J}^{\rho-1} \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & 0 \end{array}\right) \left(\begin{array}{c|c} \pm iI_\rho + 2iN_\rho & I_\rho \\ \hline I_\rho & \mp iI_\rho \end{array}\right) =$$

ρte 2ρte Spalte

$$= (2i)^{\rho-1} \tilde{J}^{\rho-1} \left(\begin{array}{c|c} 0 & \pm i \\ \hline 0 & \mp i \end{array}\right) \begin{array}{l} \leftarrow 1. \\ \\ \leftarrow \rho+1 \text{ te} \end{array} \quad \text{Zeile}$$

Der 2ρ-Vektor $\begin{pmatrix} \pm i \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \pm i \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \mp i \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ ist ein Eigenvektor von $\tilde{J}$ zum Eigen-

wert $\pm i$:

$$\left(\begin{array}{c|c} \pm 2i\, I_\rho + 2i\, N_\rho & I_\rho \\ \hline I_\rho & 0 \end{array}\right) \begin{pmatrix} \pm i \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \pm i \\ 0 \end{pmatrix} = \pm i \begin{pmatrix} \pm i \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix}$$

$$\Rightarrow \quad \left(\tilde{J} \mp i I_{2\rho}\right)^{2\rho-1} = (2i)^{\rho-1}(\pm i)^{\rho-1} \left(\begin{array}{c|c} 0 & \pm i \\ \hline 0 & 1 \end{array} \; \begin{array}{c|c} 1 & \\ \hline & \mp i \end{array}\right) \neq 0$$

$\Rightarrow$

$\left(\tilde{J} \mp i I_{2\rho}\right)^{\tau}$ verschwindet also nicht für ungerade $\tau < 2\rho$.

Damit ist der Hilfssatz bewiesen.

Die Potenzen von $\tilde{J}$ sind bis auf eine nicht-singuläre Transformationsmatrix $\tilde{T}$ gegeben durch $\tilde{J}_0^n$; aus

$$\tilde{J} = \tilde{T}^{-1} \tilde{J}_0 \tilde{T} \quad \text{folgt} \quad \tilde{J}^n = \tilde{T}^{-1} \tilde{J}_0^n \tilde{T} \quad \text{für } n \in \mathbb{N}$$

Polynomartiges Anwachsen der Komponenten der Matrix $\tilde{A}_4(k)$

läßt sich daher für feste Werte der Parameter recht leicht kontrollieren.

Zunächst hat man von den Eigenwerten λ von $A_4(k)$ vorauszusetzen

$$|\lambda| \leq 1 \quad , \lambda \in \mathbb{R} ,$$

da sonst (s.II.B.) für die Eigenwerte λ_+, λ_- von $\tilde{A}_4(k)$ gilt:

$$|\lambda_+|, |\lambda_-| > 1 \quad ,$$

so daß zum Beispiel die Elemente in der Hauptdiagonale von $\tilde{J}_0^n$,

die gleich λ_+^n bzw. λ_-^n sind, exponentiell divergieren. (Hier geht

auch die Hyperbolizität wieder ein: $\lambda \in \mathbb{R}$).

Für JORDAN Matrizen $\mu I_\rho + N_\rho$, $\mu \in \mathbb{C}$, gilt nämlich:

(24) $$(\mu I_\rho + N_\rho)^n = \sum_{\nu=0}^{n} \binom{n}{\nu} \mu^{n-\nu} N_\rho^{\nu} = \sum_{\nu=0}^{\min(n,\rho-1)} \binom{n}{\nu} \mu^{n-\nu} N_\rho^{\nu}$$

für $n \in \mathbb{N}, \mu \in \mathbb{C}$

da N_ρ nilpotent ist mit $N_\rho^{\rho} = 0$ $(N_\rho^0 := I_\rho)$.

Dann hat man $A_L(k)$ nach Eigenwerten + 1 oder - 1 zu untersuchen;

sei q deren größter Elementarteilerexponent im Minimalpolynom.

Aus dem Korollar des Hilfssatzes weiß man, daß die JORDANsche Normalform der zugehörigen Matrix $\tilde{J}$ die Matrix $\pm I_{2q} + N_{2q}$ ist, wegen (24) enthält $\tilde{J}_0^n$ daher in der rechten oberen Ecke ein Element $(\pm 1)^{n-2q+1} \binom{n}{2q-1}$,

für $n \geq 2q-1$

Das ist ein Polynom in n vom Grade $2q-1$, also dasjenige mit dem höchsten Grad.

Es bleiben noch die Eigenwerte λ von $A_L(k)$ mit $-1 < \lambda < 1$.

Sei ρ ihr größter Elementarteilerexponent, das zugehörige $\tilde{J}$ hat jetzt die JORDAN'sche Normalform

$$\begin{pmatrix} J_{\lambda_+} & 0 \\ 0 & J_{\lambda_-} \end{pmatrix} \qquad \text{mit } J_{\lambda_\pm} = \lambda_\pm I_\rho + N_\rho$$

$J_{\lambda_\pm}^n$ hat in der rechten oberen Ecke ein Element $(\lambda_\pm)^{n-\rho+1} \cdot \binom{n}{\rho-1}$

für $n \geq \rho - 1$, das ist das Polynom in n mit dem höchsten Grad unter allen Elementen von $\tilde{J}^n$. Da $|\lambda_\pm| = 1$, ändert die $n-\rho+1$-te Potenz von $\lambda_\pm$ daran nichts, d.h. $\lambda_\pm^n$ konvergiert nicht gegen 0 und divergiert nicht gegen ∞, sondern bleibt betragsmäßig nach oben und nach unten beschränkt.

Da man durch die Wahl größerer Gitterverhältnisse R_ν die Eigenwerte von $A_L(k)$ verkleinern kann, sind die ersten beiden der drei Fälle leicht auszuschließen, etwa durch Auferlegung einer Bedingung (21). Der dritte Fall ist durch das Problem (11) unweigerlich bedingt. Es wird daher i.a.

Elemente in $\tilde{A}_L(K)$ geben, die Polynome vom Grade $p-1$ in n sind, mit dem obigen p. Von diesen Elementen hat man schwache Stabilität desselben Grades $p-1$ zu erwarten, wegen $h^{1-p} = \frac{n^{p-1}}{t^{p-1}}$ für alle in (5) zugelassenen $h, n, t = nh$.

Diese Argumentation vermag zwar eine plausible Erklärung für das zu erwartende Verhalten der Potenzen der Amplifikationsmatrix $\tilde{A}_L(K)$ geben, mathematisch ist sie schlicht unbefriedigend, da die Parameter unberücksichtigt geblieben sind. Die Schwäche dieser Vorgehensweise liegt in der Unkenntnis der Transformationsmatrizen

$$\tilde{T}_L(K) := \begin{pmatrix} T_L(K) & 0 \\ 0 & T_L(K) \end{pmatrix} .$$

Die in solchen Situationen häufig gestellte Forderung der gleichmäßigen Beschränktheit dieser Matrizen in einer geeigneten Norm, und zwar gleichmäßig beschränkt in K, ist eine reichlich realitätsfremde Auflage, wenn sich der maximale Elementarteiler~~-~~exponent für einen Eigenwert auch wirklich ändert. Die Eigenwerte von $\tilde{A}_L(K)$ selbst kann man noch als stetig von K abhängig auffassen, da die Koeffizienten des sie zu Nullstellen habenden charakteristischen Polynoms stetige Funktionen in K sind. An einer solchen Sprungstelle K^* der Elementarteilerexponenten wird man aber wohl kaum eine Schar von derartigen Matrizen $T_L(K)$

$$\{T_L(K) T_L^{-1}(K) \text{ existiert, } J_L(K) = T_L^{-1}(K) A_L(K) T_L(K) \text{ für alle } K \text{ aus einer Umgebung von } K^*\}$$

finden, die zusammen mit der analogen Schar für $T_L^{-1}(K)$ dort stetig oder auch nur gleichmäßig beschränkt ist in der Norm, wenn so etwas nicht überhaupt grundsätzlich unmöglich ist.

Wählt man dagegen eine gegenüber der von JORDAN etwas veränderte Normalform, nämlich eine solche, in der statt der $J_{\lambda,\nu,\mu}^{(k)}$ Matrizen der Form

$$L_{\lambda,\nu,\mu}(k) := \lambda_{\lambda,\nu}(k)\, I_{m_{\nu\mu}} + c_{\lambda,\nu,\mu}(k) N_{m_{\nu\mu}} \text{ stehen,}$$

so lassen sich damit in der Praxis brauchbare Kriterien für die schwache Stabilität von $\tilde{A}_h(k)$ entwickeln. Ein Beispiel wird noch später gegeben, welches auch der Leistungsfähigkeit des Schemas (10) angemessen ist, das ja in der Regel von 2. Ordnung konsistent ist, also nur schwache Stabilität für Grade < 2 kompensieren kann.

Zunächst ist das auch tatsächlich eine Normalform, denn

$L_\lambda := \lambda I_\rho + c N_\rho$ mit $\lambda, c \in \mathbb{C}$, $c \neq 0$, $\rho \in \mathbb{N}$ hat das

charakteristische und das Minimalpolynom $(z-\lambda)^\rho$, und ist daher für $c \neq 0$ der JORDAN Matrix $\lambda I_\rho + N_\rho$ ähnlich. Denn für das charakteristische Polynom sieht man das sofort, und der Ausdruck

$(\lambda I_\rho + c N_\rho - \lambda I_\rho)^\varkappa = c^\varkappa N_\rho^\varkappa$ verschwindet zuerst für $\varkappa = \rho$ falls $c \neq 0$

Der nächste Schritt besteht in der direkten Ausrechnung der Potenzen von

$$\tilde{L} := \begin{pmatrix} 2i\, L_\lambda & I_\rho \\ I_\rho & 0 \end{pmatrix}$$

Dazu ist einiger rechnerischer Aufwand erforderlich.

Lemma: Sei $\rho \in \mathbb{N}$.

Die Potenzen der 2px2p Matrix $\tilde{\mathcal{K}} := \begin{pmatrix} \chi_\rho & I_\rho \\ I_\rho & 0 \end{pmatrix}$ mit einer pxp Matrix $\chi_\rho \in \mathbb{C}^\rho \times \mathbb{C}^\rho$ sind gegeben durch:

(25) $$\tilde{\mathcal{K}}^n = \begin{pmatrix} \mathcal{K}_n & \mathcal{K}_{n-1} \\ \mathcal{K}_{n-1} & \mathcal{K}_{n-2} \end{pmatrix} \quad \text{für} \quad n \in \mathbb{N}$$

Darin sind die $\mathcal{K}_\nu$ $p \times p$ Matrizen, und zwar:

$$\mathcal{K}_\ell := \begin{cases} \sum_{\substack{\nu=0 \\ 2\nu \le \ell}}' \binom{\ell-\nu}{\nu} X_p^{\ell-2\nu} & \text{für } \ell \in \mathbb{N} \cup \{0\} \\ 0 \quad \text{für} \quad \ell = -1 \end{cases}$$

Die Summation ist dabei über alle ganzen Zahlen von 0 bis zu der zu führen, die gerade noch kleiner oder gleich $\frac{\ell}{2}$ ist.

Beweis durch vollständige Induktion über n:

Die Behauptung (25) ist offenbar richtig für $n=1$ ($X_p^0 := I_p$).

Der Index p wird jetzt weggelassen. Für $n = 2$ hat man:

$$\mathcal{K}_2 = I + X^2, \; \mathcal{K}_1 = X, \; \mathcal{K}_0 = I \quad \text{und es ist auch}$$

$$\tilde{\mathcal{K}}^2 = \begin{pmatrix} X^2 + I & X \\ X & I \end{pmatrix}$$

Sei (25) für $\nu \in \mathbb{N}$, $\nu \le n$ bewiesen. Dann ist:

$$\tilde{\mathcal{K}}^{n+1} = \tilde{\mathcal{K}} \cdot \tilde{\mathcal{K}}^n = \begin{pmatrix} X & I \\ I & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{K}_n & \mathcal{K}_{n-1} \\ \mathcal{K}_{n-1} & \mathcal{K}_{n-2} \end{pmatrix} \text{ und das soll sein}$$

gleich $$\overset{!}{=} \begin{pmatrix} \mathcal{K}_{n+1} & \mathcal{K}_n \\ \mathcal{K}_n & \mathcal{K}_{n-1} \end{pmatrix}$$

In der 2. Zeile stehen rechts und links dieselben Matrizen. Man hat nur noch zu beweisen:

$$\mathcal{K}_{n+1} = X\mathcal{K}_n + \mathcal{K}_{n-1} \qquad \text{und noch einmal}$$

$$\mathcal{K}_n = X\mathcal{K}_{n-1} + \mathcal{K}_{n-2}$$

$$X\mathcal{K}_n + \mathcal{K}_{n-1} = \sum_{\substack{\nu=0 \\ 2\nu\le n}} \binom{n-\nu}{\nu} X^{n-2\nu+1} + \sum_{\substack{\nu=0 \\ 2\nu\le n-1}} \binom{n-1-\nu}{\nu} X^{n+1-2\nu} \geqq$$

Indexverschiebung in der 2. Summe:

$$\geqq \sum_{\substack{\nu=0 \\ 2\nu\le n}} \binom{n-\nu}{\nu} X^{n-2\nu+1} + \sum_{\substack{\nu=1 \\ 2\nu\le n+1}} \binom{n-\nu}{\nu-1} X^{n+1-2\nu} =$$

$$= X^{n+1} + \sum_{\substack{\nu=1 \\ 2\nu\le n}} \binom{n+1-\nu}{\nu} X^{n+1-2\nu} + \sum_{n<2\nu\le n+1} \binom{n-\nu}{\nu-1} X^{n+1-2\nu} =$$

wegen $\binom{n-\nu}{\nu} + \binom{n-\nu}{\nu-1} = \binom{n+1-\nu}{\nu}$

$$= \sum_{\substack{\nu=0 \\ 2\nu\le n+1}} \binom{n+1-\nu}{\nu} X^{n+1-2\nu} = \mathcal{K}_{n+1}$$

denn

$$\binom{n-\frac{n+1}{2}}{\frac{n+1}{2}-1} = \binom{\frac{n}{2}-\frac{1}{2}}{\frac{n}{2}-\frac{1}{2}} = 1 = \binom{n+1-\frac{n+1}{2}}{\frac{n+1}{2}} \quad \text{falls } n \text{ ungerade}$$

q.e.d. (Lemma)

Das Lemma gestattet einen Einblick in die Struktur von $\tilde{L}^n$:

$$\tilde{L}^n = \begin{pmatrix} \mathcal{K}_n & \mathcal{K}_{n-1} \\ \mathcal{K}_{n-1} & \mathcal{K}_{n-2} \end{pmatrix} \qquad \text{für } n \in \mathbb{N}$$

mit $$\mathcal{K}_\ell := \sum_{\substack{\nu=0 \\ 2\nu\le\ell}} \binom{\ell-\nu}{\nu} (2i)^{\ell-2\nu} L_\lambda^{\ell-2\nu} \quad \text{für} \quad \ell \in \mathbb{N}_0$$

und $\mathcal{K}_\ell = 0$ für $\ell = -1$

Sei $\ell \in \mathbb{N}_0$. Es ist $L_\lambda = \lambda I_p + c N_p$.

$$\Rightarrow \mathcal{K}_\ell = \sum_{\substack{\nu=0 \\ 2\nu\le\ell}} \binom{\ell-\nu}{\nu} (2i)^{\ell-2\nu} (\lambda I_p + c N_p)^{\ell-2\nu} =$$

$$= \sum_{\substack{\nu=0 \\ 2\nu\le\ell}} \binom{\ell-\nu}{\nu} (2i)^{\ell-2\nu} \sum_{\varkappa=0}^{\min(\ell-2\nu,\,p-1)} \binom{\ell-2\nu-\varkappa}{\varkappa} \lambda^{\ell-2\nu-\varkappa} c^\varkappa N^\varkappa =$$

$$= \sum_{\varkappa=0}^{p-1} c^\varkappa N^\varkappa \sum_{\substack{\nu=0 \\ 2\nu\le\ell-\varkappa}} \binom{\ell-\nu}{\nu} \binom{\ell-2\nu-\varkappa}{\varkappa} (2i)^{\ell-2\nu} \lambda^{\ell-2\nu-\varkappa}$$

Für die 2.Summe wird darin 0 vereinbart, falls $\ell - \varkappa < 0$

$\mathcal{K}_\ell$ ist also eine Dreiecksmatrix :  . In jeder oberen Nebendiagonale sind alle Elemente gleich. Sei $k_{\ell\varkappa}$ ein Element in der $\varkappa$ten ($\varkappa = 0,1,2,\cdots,p-1$; für $\varkappa = 0$ hat man die Hauptdiagonale), oberen Nebendiagonalen, dann ist:

$$k_{\ell\varkappa} = (2ic)^\varkappa \sum_{\substack{\nu=0 \\ 2\nu\le\ell-\varkappa}} \binom{\ell-\nu}{\nu} \binom{\ell-2\nu-\varkappa}{\varkappa} (2i\lambda)^{\ell-2\nu-\varkappa}$$

An dieser Darstellung kann man allerdings das Verhalten der Elemente von $\tilde{L}^n$ noch nicht ablesen; es bedarf noch einer Umschreibung, die die Eigenwerte λ_+, λ_- einführt.

Transformiert man $\tilde{L}$ auf seine JORDAN'sche Normalform und berechnet mit dem Ergebnis die Potenzen von $\tilde{L}$, so gewinnt man für $p = 2, 3$ überschaubare Ausdrücke für $k_{\ell\varkappa}$. Von $p = 4$ ab wird das Verfahren recht umfangreich; insbesondere ist für allgemeine $p \in \mathbb{N}$ zwar noch die Matrix $\tilde{T}$ zu berechnen, die $\tilde{L}$ auf JORDAN'sche

Normalform transformiert, die Inversion von $\tilde{T}$ aber ist recht kompliziert. Die für $p=2,3$ erhaltenen Formeln für $K_{\ell x}$ lassen glücklicherweise eine rein formal begründete Extension für größere p zu. Dies zur Motivation des folgenden

Hilfssatzes:

(26) Für $n, x, n-x \in \mathbb{N} \cup \{0\}$ und a,b aus einem kommutativen Ring mit Einselement gilt:

$$\text{(i)} \quad (a-b)^{2x+1} \sum_{\substack{v=0 \\ 2v \le n-x}} \binom{n-v}{v}\binom{n-2v}{x}(a+b)^{n-x-2v}(-ab)^v =$$

$$\text{(ii)} \quad = (a-b)^{2x+1} \sum_{\mu=0}^{n-x} \binom{n-\mu}{x}\binom{x+\mu}{x} a^{n-x-\mu} b^\mu =$$

$$\text{(iii)} \quad = \sum_{\tau=0}^{x} \binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau}\left[a^{n+1+x-2\tau} - b^{n+1+x-2\tau}\right](-ab)^\tau$$

Die Summen in (i) und (ii) sind ebenfalls gleich.

Beweis durch Doppelinduktion über n und x.

Wenn man sich anhand der Matrizenmultiplikation klargemacht hat, wie der Induktionsschritt auszusehen hat, besteht der Beweis nur noch aus Rechnerei.

Für den Induktionsanfang hat man zunächst den Fall $x = 0$ zu beweisen:

$$(a-b) \sum_{\substack{v=0 \\ 2v \le n}} \binom{n-v}{v}(a+b)^{n-2v}(-ab)^v = (a-b) \sum_{\mu=0}^{n} a^{n-\mu} b^\mu =$$

$$= a^{n+1} - b^{n+1}$$

Die 2. Gleichheit ist die bekannte Darstellung der geometrischen Reihe, die der beiden Summen (bezeichnet mit M_n) folgt wie im Lemma durch Induktion über n :

für $n = 0$ ist : $1 = 1$

$n = 1$: $a+b = a+b$

Induktionsschritt: $M_{n+1} = (a+b) M_n - ab M_{n-1}$

(i) Beh.: M_n erfüllt den Induktionsschritt für $n \in \mathbb{N}$

Beweis:

$$(a+b) M_n - ab M_{n-1} =$$

$$= (a+b) \sum_{\substack{v=0 \\ 2v \le n}} \binom{n-v}{v} (a+b)^{n-2v} (-ab)^v - ab \sum_{\substack{v=0 \\ 2v \le n-1}} \binom{n-1-v}{v} (a+b)^{n-1-2v} (-ab)^v =$$

Indexverschiebung wie oben

$$= (a+b)^{n+1} + \sum_{\substack{v=1 \\ 2v \le n}} \left[\binom{n-v}{v} + \binom{n-v}{v-1} \right] (a+b)^{n+1-2v} (-ab)^v$$

$$+ \sum_{n < 2v < n+1} \binom{n-v}{v-1} (a+b)^{n+1-2v} (-ab)^v =$$

$$= M_{n+1}$$

(ii) Beh.: M_n erfüllt den Induktionsschritt für $n \in \mathbb{N}$:

Beweis:

$$(a+b) M_n - ab M_{n-1} =$$

$$= (a+b) \sum_{\mu=0}^{n} a^{n-\mu} b^\mu - ab \sum_{\mu=0}^{n-1} a^{n-1-\mu} b^\mu =$$

$$= a^{n+1} + b^{n+1} + 2 \sum_{\mu=0}^{n-1} ab \cdot a^{n-1-\mu} b^\mu - ab \sum_{\mu=0}^{n-1} a^{n-1-\mu} b^\mu =$$

$$= \sum_{\mu=0}^{n+1} a^{n+1-\mu} b^\mu = M_{n+1}$$

q. e.d. (Induktionsanfang 1. Teil)

Man braucht weiter für den Induktionsanfang (26) für die Parameter $\varkappa = n$ und $\varkappa = n-1$. Der Fall ($\varkappa = n$) ist leicht einzusehen, in (i) und (ii) ergeben die Summen beide den Wert 1 und in (iii) steht die Binominialentwicklung von $(a-b)^{2n+1}$.

Sei $\varkappa = n-1$, $n \in \mathbb{N}$. In(i),(ii) stehen die Ausdrücke

(i) $\quad (a-b)^{2n-1} \cdot n \cdot (a+b) \quad =$

(ii) $\quad (a-b)^{2n-1} \cdot [na + nb]$

die beiden Summen sind also wieder gleich.

Die Übereinstimmung mit (iii) sieht man so ein:

$$n(a-b)^{2n-1}(a+b) = n \cdot \sum_{\nu=0}^{2n-1} \binom{2n-1}{\nu} a^{2n-1-\nu}(-b)^{\nu}(a+b) =$$

$$= n \sum_{\nu=0}^{2n-1} \binom{2n-1}{\nu} a^{2n-\nu}(-b)^{\nu} - n \sum_{\nu=0}^{2n-1} \binom{2n-1}{\nu} a^{2n-1-\nu}(-b)^{\nu+1} =$$

$$= n(a^{2n}+b^{2n}) + n \sum_{\nu=1}^{2n-1} \left[\binom{2n-1}{\nu} - \binom{2n-1}{\nu-1}\right] a^{2n-\nu}(-b)^{\nu} =$$

nach Indexverschiebung in der 2. Summe.

$$= \sum_{\nu=0}^{2n} \binom{2n}{\nu}(n-\nu)\, a^{2n-\nu}(-b)^{\nu} \quad , \text{denn } n \cdot [\cdots] = \binom{2n}{\nu}(n-\nu)$$

$$= \sum_{\tau=0}^{n-1} \binom{2n}{\tau}(n-\tau)\left[a^{2n-2\tau} - b^{2n-2\tau}\right](-ab)^{\tau}$$

Für $\varkappa = 0,\ n-1,\ n$ ist der Hilfssatz also richtig, das ist der Induktionsanfang.

Induktionsschritt:

Der Ausdruck unter (i), (ii), (iii) im Hilfssatz wird jetzt jeweils mit $M_{n\,x}$ bezeichnet für alle $n, x, n-x \in \mathbb{N}_0$. Für $n, x, n-x \in \mathbb{N}$ wird dann in allen drei Fällen der Induktionsschritt

(27) $$M_{n+1\,x} = (a+b)M_{n\,x} - ab\,M_{n-1\,x} + (a-b)^2 M_{n\,x-1}$$ bewiesen,

aus welchem dann die Gleichheit von (i),(ii),(iii) in (26) folgt. In (i),(ii) sind die Zusammenhänge einfacher, dort bezeichne $N_{n\,x}$ jeweils die Summe. Sei also $n, x, n-x \in \mathbb{N}$.

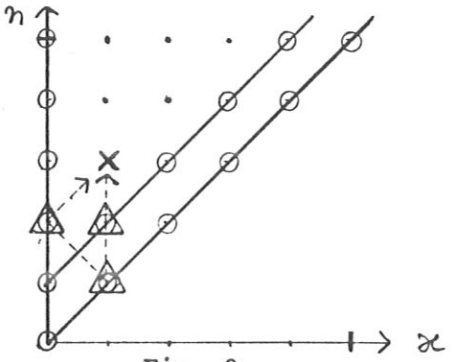


Fig. 3

(i) Es muß (27) gezeigt werden. Wegen

$$M_{n\,x} = (a-b)^{2x+1} N_{n\,x}$$ folgt das aus

(28) $$N_{n+1\,x} = (a+b)\,N_{n\,x} - ab\,N_{n-1\,x} + N_{n\,x-1}$$

Es ist

$$(a+b)N_{n\,x} - ab\,N_{n-1\,x} + N_{n\,x-1} =$$

$$= \sum_{\substack{v=0 \\ 2v \le n-x}} \binom{n-v}{v}\binom{n-2v}{x}(a+b)^{n-x-2v+1}(-ab)^v +$$

$$+ \sum_{\substack{v=0 \\ 2v \le n-1-x}} \binom{n-1-v}{v}\binom{n-1-2v}{x}(a+b)^{n-1-x-2v}(-ab)^{v+1} +$$

$$+ \sum_{\substack{v=0 \\ 2v \le n-x+1}} \binom{n-v}{v}\binom{n-2v}{x-1}(a+b)^{n-x-2v+1}(-ab)^v \Rightarrow$$

Indexverschiebung in der 2. Summe:

$$\Rightarrow \quad \cdots \quad +$$

$$+ \sum_{\substack{v=1 \\ 2v \le n+1-x}} \binom{n-v}{v-1}\binom{n-2v+1}{x}(a+b)^{n-x-2v+1}(-ab)^v +$$

$$+ \cdots =$$

$$= \left[\binom{n}{x}+\binom{n}{x-1}\right](a+b)^{n-x+1} + \sum_{n-x<2v\le n-x+1} \left[\binom{n-v}{v-1}\binom{n-2v+1}{x}+\binom{n-v}{v}\binom{n-2v}{x-1}\right] \cdot$$

$$\cdot (a+b)^{n-x-2v+1}(-ab)^v +$$

$$+ \sum_{\substack{v=1 \\ 2v \le n-x}} \left[\binom{n-v}{v}\binom{n-2v}{x}+\binom{n-v}{v-1}\binom{n-2v+1}{x}+\binom{n-v}{v}\binom{n-2v}{x-1}\right](a+b)^{n-x-2v+1}(-ab)^v =$$

$$= \binom{n+1}{x}(a+b)^{n-x+1} + \sum_{n-x<2v\le n-x+1} \underbrace{\left[\binom{n-v}{v-1}\binom{n-n+x-1+1}{x}+\binom{n-v}{v}\binom{n-n+x-1}{x-1}\right]}_{=\binom{n+1-v}{v}} \cdot$$

$$\cdot (a+b)^{n-x-2v+1}(-ab)^v +$$

$$+ \sum_{\substack{v=1 \\ 2v \le n-x}} \binom{n+1-v}{v}\binom{n+1-2v}{x}(a+b)^{n+1-x-2v}(-ab)^v \Rightarrow$$

Weil in der Randsumme allenfalls ein Term mit $v = \frac{n-x+1}{2}$

beiträgt, und wegen:

$$\binom{n-v}{v}\binom{n-2v}{x} + \binom{n-v}{v-1}\binom{n-2v+1}{x} + \binom{n-v}{v}\binom{n-2v}{x-1} =$$

$$= \left[\binom{n-v}{v} + \binom{n-v}{v-1}\right]\binom{n-2v+1}{x} = \binom{n+1-v}{v}\binom{n+1-2v}{x}$$

$$\Rightarrow \sum_{\substack{v=0 \\ 2v \le n+1-x}} \binom{n+1-v}{v}\binom{n+1-2v}{x}(a+b)^{n+1-x-2v}(-ab)^v = N_{n+1,x}$$

Die Relation (27) ist also erfüllt.

(ii)

Wieder genügt es statt (27) die einfachere Relation (28) zu zeigen.

Es ist:

$$(a+b)\,N_{n\,x} - a\,b\,N_{n-1\,x} + N_{n\,x-1} =$$

$$= \sum_{\mu=0}^{n-x} \binom{n-\mu}{x}\binom{x+\mu}{x} a^{n-x-\mu} b^{\mu}(a+b) \; +$$

$$- \sum_{\mu=0}^{n-x-1} \binom{n-1-\mu}{x}\binom{x+\mu}{x} a^{n-\mu-x} b^{\mu+1} \; +$$

$$+ \sum_{\mu=0}^{n+1-x} \binom{n-\mu}{x-1}\binom{x-1+\mu}{x-1} a^{n+1-x-\mu} b^{\mu} =$$

$$= \sum_{\mu=0}^{n-x} \binom{n-\mu}{x}\binom{x+\mu}{x} a^{n+1-x-\mu} b^{\mu} + \sum_{\mu=0}^{n-x} \binom{n-\mu}{x}\binom{x+\mu}{x} a^{n-x-\mu} b^{\mu+1} +$$

$$- \ldots +$$

$$+ \ldots =$$

$$= \sum_{\mu=0}^{n-x} \left[\binom{n-\mu}{x}\binom{x+\mu}{x} + \binom{n-\mu}{x-1}\binom{x-1+\mu}{x-1}\right] a^{n+1-x-\mu} b^{\mu} + \binom{n}{x-1} b^{n+1-x} +$$

$$+ \sum_{\mu=0}^{n-1-x} \underbrace{\left[\binom{n-\mu}{x} - \binom{n-1-\mu}{x}\right]}_{=\binom{n-1-\mu}{x-1}} \binom{x+\mu}{x} a^{n-x-\mu} b^{\mu+1} + \binom{n}{x} b^{n+1-x} =$$

$$= \ldots + \binom{n+1}{x} b^{n+1-x} +$$

$$+ \sum_{\mu=1}^{n-x} \binom{n-\mu}{x-1}\binom{x+\mu-1}{x} a^{n+1-x-\mu} b^{\mu} \;\geqq$$

nach Indexverschiebung

$$\geq \left[\binom{n}{x}+\binom{n}{x-1}\right] a^{n+1-x} + \binom{n+1}{x} b^{n+1-x} +$$

$$+\sum_{\mu=1}^{n-x}\left[\binom{n-\mu}{x}\binom{x+\mu}{x}+\binom{n-\mu}{x-1}\binom{x-1+\mu}{x-1}+\binom{n-\mu}{x-1}\binom{x+\mu-1}{x}\right] a^{n+1-x-\mu} b^{\mu} =$$

$$=\sum_{\mu=0}^{n+1-x}\binom{n+1-\mu}{x}\binom{x+\mu}{x} a^{n+1-x-\mu} b^{\mu} = M_{n+1\,x}$$

wegen

$$\binom{n-\mu}{x}\binom{x+\mu}{x} + \binom{n-\mu}{x-1}\left[\binom{x-1+\mu}{x-1} + \binom{x-1+\mu}{x}\right] =$$

$$= \left[\binom{n-\mu}{x} + \binom{n-\mu}{x-1}\right]\binom{x+\mu}{x} = \binom{n+1-\mu}{x}\binom{x+\mu}{x}$$

Hier ist (27) also auch erfüllt.

(iii)

In diesem Falle zeigt man am besten (27) direkt:

$$(a+b) M_{n\,x} - ab\, M_{n-1\,x} + (a-b)^2 M_{n\,x-1} =$$

$$=\sum_{\tau=0}^{x}\binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau}\left[a^{n+1+x-2\tau} - b^{n+1+x-2\tau}\right](-ab)^{\tau}(a+b) +$$

$$+\sum_{\tau=0}^{x}\binom{n+x}{\tau}\binom{n-1-\tau}{x-\tau}\left[a^{n+x-2\tau} - b^{n+x-2\tau}\right](-ab)^{\tau+1} +$$

$$+\sum_{\tau=0}^{x-1}\binom{n+x}{\tau}\binom{n-\tau}{x-1-\tau}\left[a^{n+x-2\tau} - b^{n+x-2\tau}\right](-ab)^{\tau}(a-b)^2 \geq$$

Es ist

$$\left[a^{n+1+x-2\tau}-b^{n+1+x-2\tau}\right](-ab)^{\tau}(a+b) =$$

$$=(-1)^{\tau}\left[a^{n+2+x-\tau}b^{\tau}+a^{n+1+x-\tau}b^{\tau+1}-b^{n+1+x-\tau}a^{\tau+1}+\right.$$
$$\left.-b^{n+2+x-\tau}a^{\tau}\right]$$

$$\left[a^{n+x-2\tau}-b^{n+x-2\tau}\right](-ab)^{\tau}\cdot(a^2-2ab+b^2)=$$

$$=(-1)^{\tau}\left[a^{n+2+x-\tau}b^{\tau}-2a^{n+1+x-\tau}b^{\tau+1}+a^{n+x-\tau}b^{\tau+2}+\right.$$
$$\left.-b^{n+x-\tau}a^{\tau+2}+2b^{n+1+x-\tau}a^{\tau+1}-b^{n+2+x-\tau}a^{\tau}\right]$$

Setzt man in die Summen ein und ordnet nach Potenzen von a und b,

so steht da:

$$\Rightarrow\sum_{\tau=0}^{x}\left\{(-1)^{\tau}\binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau}\left[a^{n+2+x-\tau}b^{\tau}-b^{n+2+x-\tau}a^{\tau}\right]+\right.$$
$$\left.+\left[(-1)^{\tau}\binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau}+(-1)^{\tau+1}\binom{n+x}{\tau}\binom{n-1-\tau}{x-\tau}\right]\left[a^{n+1+x-\tau}b^{\tau+1}-b^{n+1+x-\tau}a^{\tau+1}\right]\right\}+$$
$$+\sum_{\tau=0}^{x-1}(-1)^{\tau}\binom{n+x}{\tau}\binom{n-\tau}{x-1-\tau}\cdot\left[a^{n+2+x-\tau}b^{\tau}-2a^{n+1+x-\tau}b^{\tau+1}+\right.$$
$$\left.+a^{n+x-\tau}b^{\tau+2}-b^{n+x-\tau}a^{\tau+2}+2b^{n+1+x-\tau}a^{\tau+1}-b^{n+2+x-\tau}a^{\tau}\right]\Rightarrow$$

Jetzt werden für die einzelnen Summationsterme die Indizes verschoben, so daß a und b nur noch in Monomen der Form

$a^{n+2+x-\tau}b^{\tau}$ und $b^{n+2+x-\tau}a^{\tau}$ auftreten.

$$\Rightarrow\sum_{\tau=0}^{x}(-1)^{\tau}\binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau}\left[a^{n+2+x-\tau}b^{\tau}-b^{n+2+x-\tau}a^{\tau}\right]+$$
$$+\sum_{\tau=1}^{x+1}\left[(-1)^{\tau-1}\binom{n+x+1}{\tau-1}\binom{n+1-\tau}{x+1-\tau}+(-1)^{\tau}\binom{n+x}{\tau-1}\binom{n-\tau}{x+1-\tau}\right]\cdot\left[a^{n+2+x-\tau}b^{\tau}\right.$$
$$\left.-b^{n+2+x-\tau}a^{\tau}\right]+$$
$$+\sum_{\tau=0}^{x-1}(-1)^{\tau}\binom{n+x}{\tau}\binom{n-\tau}{x-1-\tau}\left[a^{n+2+x-\tau}b^{\tau}-b^{n+2+x-\tau}a^{\tau}\right]+$$
$$+\sum_{\tau=1}^{x}(-1)^{\tau}\binom{n+x}{\tau-1}\binom{n+1-\tau}{x-\tau}\cdot 2\left[a^{n+2+x-\tau}b^{\tau}-b^{n+2+x-\tau}a^{\tau}\right]+$$
$$+\sum_{\tau=2}^{x+1}(-1)^{\tau}\binom{n+x}{\tau-2}\binom{n+2-\tau}{x+1-\tau}\left[a^{n+2+x-\tau}b^{\tau}-b^{n+2+x-\tau}a^{\tau}\right]\Rightarrow$$

$$= \left[\binom{n}{x} + \binom{n}{x-1}\right] \cdot \left[a^{n+2+x} - b^{n+2+x}\right] +$$

$$+ \left[-(n+x+1)\binom{n-1}{x-1} + \binom{n+1}{x} - \binom{n-1}{x} - (n+x)\binom{n-1}{x-2} + 2\binom{n}{x-1}\right] \cdot \left[a^{n+1+x}b - b^{n+1+x}a\right] +$$

$$+ (-1)^x \left[\binom{n+x+1}{x} - (n+1+x)\binom{n+x+1}{x-1} + (n-x)\binom{n+x}{x-1} + 2\binom{n+x}{x-1} + \right.$$

$$\left. + (n+2-x)\binom{n+x}{x-2}\right] \cdot \left[a^{n+2}b^x - b^{n+2}a^x\right] +$$

$$+ (-1)^x \underbrace{\left[\binom{n+x+1}{x} - \binom{n+x}{x} - \binom{n+x}{x-1}\right]}_{=0} \cdot \left[a^{n+1}b^{x+1} - b^{n+1}a^{x+1}\right] +$$

$$+ \sum_{\tau=2}^{x-1} (-1)^\tau \left[\binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau} - \binom{n+x+1}{\tau-1}\binom{n+1}{x-\tau+1} + \binom{n+x}{\tau-1}\binom{n-\tau}{x-\tau+1} + \right.$$

$$\left. + \binom{n+x}{\tau}\binom{n-\tau}{x-1-\tau} + 2\binom{n+x}{\tau-1}\binom{n+1}{x-\tau} + \binom{n+x}{\tau-2}\binom{n+2-\tau}{x+1-\tau}\right] \cdot$$

$$\cdot \left[a^{n+2+x-\tau}b^\tau - b^{n+2+x-\tau}a^\tau\right]$$

Darin wird vereinbart: $\binom{\ell}{x-2} = 0$ für $\ell \in \mathbb{N}_0$, falls $x = 1$.

Die Randterme lassen sich ohne Umwege in die gewünschte Form bringen:

$$\binom{n}{x} + \binom{n}{x-1} = \binom{n+1}{x}$$

$$-(n+x+1)\binom{n-1}{x-1} + \binom{n+1}{x} - (n+x)\binom{n-1}{x-2} - 2\binom{n}{x-1} =$$

$$= -(n+x)\left[\binom{n-1}{x-1} + \binom{n-1}{x-2}\right] + \binom{n+1}{x} - \binom{n-1}{x-1} - \binom{n-1}{x} - 2\binom{n}{x-1} =$$

$$= -(n+x)\binom{n}{x-1} + \underbrace{\binom{n+1}{x} - \binom{n}{x-1}}_{-\binom{n+1}{x}} = -(n+1+x)\binom{n}{x-1}$$

$$\binom{n+x+1}{x} - (n+1-x)\binom{n+x+1}{x-1} + (n-x)\binom{n+x}{x-1} + 2\binom{n+x}{x-1} + (n+2-x)\binom{n+x}{x-2} =$$

$$= \binom{n+x+1}{x} - \binom{n+x+1}{x-1} + 2\binom{n+x}{x-1} + 2\binom{n+x}{x-2}$$

$$+ (n-x)\underbrace{\left[-\binom{n+x+1}{x-1} + \binom{n+x}{x-1} + \binom{n+x}{x-2}\right]}_{=0} =$$

$$= \binom{n+x+1}{x} - \binom{n+x+1}{x-1} + 2\binom{n+x+1}{x-1} =$$

$$= \binom{n+2+x}{x}$$

Die in der Summe auftretenden Binominalkoeffizienten zerlegt man dagegen am besten:

$$\binom{n+x+1}{\tau}\binom{n+x+1}{x-\tau} - \binom{n+x+1}{\tau-1}\binom{n+1-\tau}{x+1-\tau} + \binom{n+x}{\tau-1}\binom{n-\tau}{x+1-\tau} +$$

$$+ \binom{n+x}{\tau}\binom{n-\tau}{x-1-\tau} + 2\binom{n+x}{\tau-1}\binom{n+1-\tau}{x-\tau} + \binom{n+x}{\tau-2}\binom{n+2-\tau}{x+1-\tau} =$$

$$= \left[\binom{n+x}{\tau} + \binom{n+x}{\tau-1}\right]\binom{n-\tau}{x-\tau} - \left[\binom{n+x}{\tau-1} + \binom{n+x}{\tau-2}\right]\cdot\left[\binom{n-\tau}{x+1-\tau} + \binom{n-\tau}{x-\tau}\right] +$$

$$+ \binom{n+x}{\tau-1}\binom{n-\tau}{x+1-\tau} + \binom{n+x}{\tau}\binom{n-\tau}{x-1-\tau} + 2\binom{n+x}{\tau-1}\binom{n+1-\tau}{x-\tau} +$$

$$+ \binom{n+x}{\tau-2}\left[\binom{n-\tau}{x-1-\tau} + 2\binom{n-\tau}{x-\tau} + \binom{n-\tau}{x+1-\tau}\right] =$$

Die "Hälfte" vom ersten Term, der ganze zweite Term, der dritte und die "Hälfte" vom sechsten Term heben sich heraus. Es bleibt übrig:

$$= \binom{n+x}{\tau}\left[\binom{n-\tau}{x-\tau} + \binom{n-\tau}{x-1-\tau}\right] + 2\binom{n+x}{\tau-1}\binom{n+1-\tau}{x-\tau} +$$

$$+ \binom{n+x}{\tau-2}\left[\binom{n-\tau}{x-1-\tau} + \binom{n-\tau}{x-\tau}\right] =$$

$$= \left[\binom{n+x}{\tau} + 2\binom{n+x}{\tau-1} + \binom{n+x}{\tau-2}\right]\cdot\binom{n+1-\tau}{x-\tau} =$$

$$= \binom{n+2+x}{\tau}\binom{n+1-\tau}{x-\tau}$$

Es folgt:

$$\geq \sum_{\tau=0}^{x_1} (-1)^\tau \binom{n+2+x}{\tau}\binom{n+1-\tau}{x-\tau}\left[a^{n+2+x-\tau}b^\tau - b^{n+2+x-\tau}a^\tau\right] =$$

$$= M_{n+1\,x}$$

Damit ist der Hilfssatz bewiesen.

Mit seiner Hilfe läßt sich die Potenzbeschränktheit von

$$\tilde{L} = \left(\begin{array}{c|c} 2i\lambda I_\rho + 2iN_\rho & I_\rho \\ \hline I_\rho & 0 \end{array}\right)$$ vollständig überblicken.

Mit $a=\lambda_+$, $b=\lambda_-$ $\Rightarrow a+b=2i\lambda$, $-ab=1$ kann man direkt in ihn eingehen und erhält für $k_{n\varkappa}$ die Darstellung

(29) (i) $$k_{n\varkappa} = (2ic)^{\varkappa} \sum_{\substack{\nu=0 \\ 2\nu\le n-\varkappa}}^{1} \binom{n-\nu}{\nu}\binom{n-2\nu}{\varkappa}(2i\lambda)^{n-\varkappa-2\nu} =$$

(ii) $$= (2ic)^{\varkappa} \sum_{\mu=0}^{n-\varkappa} \binom{n-\mu}{\varkappa}\binom{\varkappa+\mu}{\varkappa}\lambda_+^{n-\varkappa-\mu}\lambda_-^{\mu} =$$

(iii) $$= \frac{(2ic)^{\varkappa}}{(\lambda_+-\lambda_-)^{2\varkappa+1}} \sum_{\tau=0}^{\varkappa} \binom{n+\varkappa+1}{\tau}\binom{n-\tau}{\varkappa-\tau}\left[\lambda_+^{n+1+\varkappa-2\tau}-\lambda_-^{n+1+\varkappa-2\tau}\right]$$

wovon die letzte Beziehung (iii) nur für $\lambda_+ \neq \lambda_-$ gilt. Aus ihr sieht man im Falle $|\lambda|>1$, d.h. $|\lambda_+|$ oder $|\lambda_-|>1$, wegen $|\lambda_+|\cdot|\lambda_-|=1$ sofort das exponentielle Anwachsen aller nicht verschwindenden Elemente von $\tilde{L}^n$ bedingt durch den führenden Term $n^{\varkappa}\lambda_+^{n+\varkappa+1}$ bzw. $n^{\varkappa}\lambda_-^{n+\varkappa+1}$, der für genügend großes n alle anderen dem Betrage nach majorisiert.

Falls $\lambda_- = \lambda_+$ d.h. $\lambda = \pm 1$ oder $\lambda = -1$, erhält man aus der Darstellung (ii) für $k_{n\varkappa}$:

$$k_{n\varkappa} = (\pm 1)^{n-\varkappa}(2c)^{\varkappa} i^n \sum_{\mu=0}^{n-\varkappa}\binom{n-\mu}{\varkappa}\binom{\varkappa+\mu}{\varkappa}$$

$$= (\pm 1)^{n-\varkappa}(2c)^{\varkappa} i^n \binom{n+\varkappa+1}{2\varkappa+1}$$

Die darin verwandte elementare Reaktion zwischen Binominalkoeffizienten beweist man am einfachsten durch Koeffizientenvergleich vor dem Monom $a^{\varkappa}b^{\varkappa}$ in der Summe $\sum_{\mu=0}^{n-\varkappa}(a+1)^{n-\mu}(b+1)^{\varkappa+\mu}$ und in einem anderen Ausdruck für sie, den man über Umformungen mittels der Darstellung von geometrischen Reihen erhält. Man hat dann noch das allgemeine Additionstheorem für Binominalkoeffizienten anzuwenden.

In der allgemeineren Form:

$$\binom{n+p}{m} = \sum_{k=0}^{m} \binom{n-k}{n-m}\binom{p+k-1}{k}$$

findet man die obige Relation

z.B. in [9] auf S. 8 (3b).

Für genügend große n wachsen daher die Elemente k_{nx} dem Betrage nach wie n^{2x+1}, abgesehen vom konstanten Faktor $2^x|c|^x$. Das ist das Ergebnis, welches aus den Überlegungen anhand der JORDAN'schen Normalform zu erwarten war.

Es bleibt der hauptsächlich interessierende Fall $-1<\lambda<1$ zu untersuchen, in welchem $\lambda_+ \neq \lambda_-$, $|\lambda_+|=|\lambda_-|=1$. Man schätzt $|k_{nx}|$ mittels (29) (iii) ab, indem $|\lambda_+^\ell - \lambda_-^\ell| \leq 2$ gesetzt wird. Die noch verbleibende, unhandliche Summe gestattet auch eine bequeme Umformung.

Lemma: Für $n, x, n-x \in \mathbb{N}_0$ gilt:

$$\sum_{\tau=0}^{x} \binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau} = \frac{2^x}{x!}\prod_{\tau=0}^{x-1}(n+x-2\tau) \tag{30}$$

Für das Produkt wird der Wert 1 vereinbart, falls $x=0$, dann stimmt (30) offenbar.

Beweis:

Setzt man in (26) (i) und (iii) für a und b die Zahlen a = 1, b = -1 ein, so steht da:

$$2^{2x+1}\sum_{\substack{\nu=0 \\ 2\nu\leq n-x}} \binom{n-\nu}{\nu}\binom{n-2\nu}{x}\cdot 0^{n-x-2\nu} = \sum_{\tau=0}^{x}\binom{n+x+1}{\tau}\binom{n-\tau}{x-\tau}\left[1-(-1)^{n+1+x-2\tau}\right]$$

Darin verschwinden beide Seiten, falls $n-\varkappa$ oder dazu äquivalent $n+\varkappa$ ungerade ist. Für $n-\varkappa$ gerade, und das ist dasselbe wie $n+\varkappa$ gerade, verschwindet in der linken Summe nur der Term mit $\nu = \frac{n-\varkappa}{2}$ nicht, während rechts $1-(-1)^{n+1+\varkappa-2\tau} = 2$ ist, d.h.:

$$2^{2\varkappa+1}\binom{n-\frac{n-\varkappa}{2}}{\frac{n-\varkappa}{2}}\binom{n-n+\varkappa}{\varkappa} = 2\sum_{\tau=0}^{\varkappa}{}'\binom{n+\varkappa+1}{\tau}\binom{n-\tau}{\varkappa-\tau} =$$

$$= 2^{2\varkappa+1}\binom{\frac{n+\varkappa}{2}}{\frac{n-\varkappa}{2}} = \frac{2^{2\varkappa+1}}{\varkappa!}\prod_{\nu=0}^{\varkappa-1}\left(\frac{n+\varkappa}{2}-\nu\right) = \frac{2^{\varkappa+1}}{\varkappa!}\cdot\prod_{\nu=0}^{\varkappa-1}(n+\varkappa-2\nu)$$

Hält man $\varkappa$ fest und betrachtet die rechte und die linke Seite von (30) als Polynome in n, so stimmen diese Polynome im Grade, bei beiden $\varkappa$, und für unendlich viele Werte von n überein, und zwar für $n \in \mathbb{N}_0$ mit $n+\varkappa$ gerade. Ihre Differenz ist daher ein Polynom höchstens vom Grade $\varkappa$, das an unendlich vielen Stellen verschwindet, also das Nullpolynom. Damit ist (30) auch für $n+\varkappa$ ungerade bewiesen.

Für $k_{n\varkappa}$ hat man daher im Falle $-1 < \lambda < 1$:

$$|k_{n\varkappa}| \le \frac{|2c|^{\varkappa}\cdot 2^{\varkappa+1}}{|\lambda_+-\lambda_-|^{2\varkappa+1}\cdot\varkappa!}\prod_{\nu=0}^{\varkappa-1}(n+\varkappa-2\nu) \le \frac{2\cdot|4c|^{\varkappa}}{|\lambda_+-\lambda_-|^{2\varkappa+1}\cdot\varkappa!}\cdot(n+1)^{\varkappa} \tag{31}$$

wegen $(n+1+a)(n+1-a) = (n+1)^2-a^2 \le (n+1)^2$ für $a \in \mathbb{R}$. Der Grad des Polynoms, welches die Elemente von k_n dem Betrage nach abschätzt, ist hier um $\varkappa+1$ kleiner als für $\lambda_+ = \lambda_-$.

Um aus (31) schwache Stabilität vom Grade $\varkappa$ abzuleiten, hat man zunächst eine Norm $\|\cdot\|$ für die Matrizen $\mathfrak{Z}^n$ zu wählen; hier wird wieder die sich anbietende Spektralnorm genommen.

Für jede $\ell \times \ell$ Matrix Z, $\ell \in \mathbb{N}$, also $Z = (z_{\mu\nu})_{\mu,\nu=1,\cdots,\ell}$ mit $z_{\mu\nu} \in \mathbb{C}$ für alle $\nu,\mu = 1,\cdots,\ell$ sei Z_+ definiert durch:

$$Z_+ := (|z_{\mu\nu}|)_{\mu,\nu=1,\cdots,\ell}$$

und wie üblich die komplex konjugierte Matrix $\overline{Z}$ sowie die Transponierte Z^T durch:

$$\overline{Z} := (\overline{z}_{\mu\nu})_{\mu,\nu=1,\cdots,\ell} \quad ; \quad Z^T := (z_{\nu\mu})_{\mu,\nu=1,\cdots,\ell}$$

Dann ist (s.z.B. [10] S.175) allgemein:

$$|Z| \leq |Z_+| \qquad \text{, und weiter:}$$

$$|Z\,\overline{Z}^T| \leq |(Z\cdot\overline{Z}^T)_+| \leq |Z_+(\overline{Z}^T)_+| = |Z_+(Z_+)^T|$$

d.h.:

$$\|Z\| \leq \|Z_+\|$$

Mit derselben Argumentation weiß man von zwei $\ell \times \ell$ Matrizen X, Y mit $X = X_+$, $Y - X = (Y-X)_+$ die Eigenschaft:

$$\|X\| \leq \|Y\|$$

Damit läßt sich die Spektralnorm von

$$\tilde{L}^n = \begin{pmatrix} \mathcal{H}_n & \mathcal{H}_{n-1} \\ \mathcal{H}_{n-1} & \mathcal{H}_{n-2} \end{pmatrix} \qquad \text{für } n \geq 2 \text{ abschätzen.}$$

Zunächst wird eine zusätzliche Bedingung an die Eigenwerte λ_+, λ_- gestellt, die später wie in II B gleichmäßige Beschränktheit der

Norm in K garantiert, und zwar

(20) $\quad -1+\varepsilon \le \lambda \le 1-\varepsilon \qquad$ für ein $\varepsilon \in \mathbb{R}^+,\ \varepsilon < 1$

In (31) kann die rechte Seite für größeres $\varkappa$ noch kleiner werden, falls η klein genug. Damit die unschöne Korrektur"für genügend großes η " nicht mitgeschleppt werden muß, wird ein $\alpha \in \mathbb{R}^+ \cup \{0\}$ eingeführt, und zwar sei $\varepsilon' \in \mathbb{R}^+,\ 0 < \varepsilon' < 1$. und

$$\sqrt{1-\varepsilon'^2} := 1-\varepsilon$$

$$\Rightarrow \quad |\lambda_+ - \lambda_-| = 2\left|\sqrt{1-\lambda^2}\right| \ge 2\sqrt{1-(1-\varepsilon)^2} = 2\sqrt{\varepsilon'^2} = 2\varepsilon'$$

sowie

$$\alpha := \max\left[|c|,\ \varepsilon'^2(p-1)\right] \quad \Rightarrow \quad \frac{\alpha}{\varepsilon'^2(p-1)}^{(n+1)} \ge 0$$

für $n \ge 0,\ p \ge 2$

(32) $\Rightarrow$

$$|k_{n\varkappa}| \le \frac{\alpha^{p-1}}{(p-1)!\,\varepsilon'^{2p-1}} \cdot (n+1)^{p-1} =: m_{n+} \qquad \text{für } \varkappa = 0,1,\dots,p-1$$

und das gilt auch für $p = 1$ wegen (29)$^{\text{iii}}$, d.h. für $p \in \mathbb{N},\ n \in \mathbb{N}_0$.

(Es wird $0^0 = 1$ vereinbart).

Seien $M_1, M_{n+} \in \mathbb{R}^{2p} \times \mathbb{R}^{2p}$ definiert durch

$$M_1 := \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}}_{2p} \Big\}\, 2p \qquad M_{n+} := m_{n+} \cdot M_1$$

Wegen (32) folgt zunächst für $n \ge 2$

$$M_{n+} - \left(\tilde{L}^n\right)_+ = \left(M_{n+} - \left(\tilde{L}^n\right)_+\right)_+$$

Das gilt auch für $n=1$ und $n=0$, denn

$$m_{n+1}, m_{0+} \geq 1 \qquad (\text{für alle } p \in \mathbb{N})$$
$$\Rightarrow \|\tilde{L}^n\| \leq \|M_{n+}\| = m_{n+}\,\|M_1\| = 2p\, m_{n+}$$

M_1 hat offenbar $2p$ und sonst die Null $2p-1$ mal zum Eigenwert; aus der besonderen Gestalt dieser Matrix folgt dann der Wert $2p$ für ihre Spektralnorm.

$$\Rightarrow$$

(33) $$\|\tilde{L}^n\| \leq 2p\,\frac{\alpha^{p-1}}{(p-1)!\;\varepsilon'^{2p-1}}$$ für $p \in \mathbb{N},\ n \in \mathbb{N}_0$, $\lambda \in \mathbb{R},\ |\lambda| \leq 1-\varepsilon$, $\varepsilon' = \sqrt{1-(1-\varepsilon)^2}$

Diese Abschätzung wird zu der der Potenzen von $\tilde{A}_h(k)$ dienen.

Der Parameterbereich der Matrizen

$$A_h(k) = \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \sin h R_\nu k_\nu$$ ist zwar der $\mathbb{R}^N$ für k, $\mathbb{R}^+$ für h, aber die beschränkte Funktion Sinus reduziert ihn mit der einfachen Transformation $\xi \in \mathbb{R}^N$, $\xi_\nu := \sin h R_\nu k_\nu$ für $k \in \mathbb{R}^N$ auf W_1, d.h. die Matrizen

$$B(\xi) := A_h(k)$$ besitzen kompakten Parameterbereich.

Läßt man in der JORDAN'schen Normalform $J_h(k)$ von $A_h(k)$ statt der $J_{h,\nu,\mu}(k)$ die Matrizen $L_{h,\nu,\mu}(k) =: L_{\nu,\mu}(\xi)\, z_h$ mit $c_{h,\nu,\mu}(k) \in \mathbb{C}$ in der ersten oberen Nebendiagonale, so hat man eine andere Normalform, sie heiße $L(\xi) := : L_h(k)$, mit den Eigenwerten $\lambda(\xi)$ in der Hauptdiagonalen.

Satz 4.

a) Das Differentialgleichungssystem

(11) $$u_t(t,x) = \sum_{\nu=1}^{N} A_\nu u_{x_\nu}(t,x) \qquad \text{für } 0 \le t \le T$$

sei hyperbolisch, d.h. die Eigenwerte $\lambda(\xi)$ der Matrix

$$B(\xi) := \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} \xi_\nu \qquad \text{für } \xi \in W_1$$

sind reell, diese erfüllen weiter für ein $\varepsilon \in \mathbb{R}^+$, $0 < \varepsilon < 1$ die Bedingung

(20) $$-1+\varepsilon \le \lambda(\xi) \le 1-\varepsilon \qquad \text{für } \xi \in W_1 .$$

$B(\xi)$ besitzt für jedes ξ eine Normalform $L(\xi)$, $\xi \in W_1$.
Sei $\varepsilon' := \sqrt{1-(1-\varepsilon)^2}$.

b) Unter den für $\xi \in W_1$ auftretenden Elementarteilerexponenten von $B(\xi)$ sei p, $p \in \mathbb{N}$, der größte.

c) Die Elemente $c(\xi)$ in der ersten oberen Nebendiagonalen besitzen eine Schranke $\alpha \in \mathbb{R}^+ \cup \{0\}$

$$|c(\xi)| \le \alpha \qquad \text{für alle } \xi \in W_1 \text{ mit } \alpha \le \varepsilon'^{2(p-1)}$$

d) Es gebe eine Schar von regulären Transformationsmatrizen $T(\xi)$ mit $L(\xi) = T^{-1}(\xi) B(\xi) T(\xi)$ und konstanten $C, C_{-1} \in \mathbb{R}^+$ so daß

$$\|T(\xi)\| \le C, \quad \|T^{-1}(\xi)\| \le C_{-1} \qquad \text{für alle } \xi \in W_1$$

Dann ist das Differenzenschema

(10) $$U^{n+1} - U^{n-1} = \sum_{\nu=1}^{N} \frac{A_\nu}{R_\nu} (T_\nu - T_\nu^{-1}) U^n$$

schwach stabil vom Grade p - 1. Für seine Amplifikationsmatrix gilt:

Beweis:

Seien $\tilde{B}(\xi) := \begin{pmatrix} 2; B(\xi) & I_m \\ I_m & 0 \end{pmatrix}$, $\tilde{T}(\xi) := \begin{pmatrix} T(\xi) & 0 \\ 0 & T(\xi) \end{pmatrix}$,

$\tilde{L}(\xi) := \begin{pmatrix} 2; L(\xi) & I_m \\ I_m & 0 \end{pmatrix}$, $\tilde{L}_{\nu,\mu}(\xi) := \begin{pmatrix} 2; L_{\nu,\mu}\xi & I_{m-1} \\ I_{m-1} & 0 \end{pmatrix}$

Dann ist

$$\|\tilde{B}(\xi)^n\| = \|\tilde{T}(\xi)^{-1}\tilde{L}(\xi)^n T(\xi)\| \le \|\tilde{T}(\xi)^{-1}\| \cdot \|\tilde{L}(\xi)^n\| \cdot \|T(\xi)\| \le$$
$$\le C \cdot C_{-1} \|\tilde{L}(\xi)^n\|$$

Es ist weiter:

$$\|\tilde{L}(\xi)^n\| = |\tilde{L}(\xi)^n \cdot \overline{\tilde{L}(\xi)^n}^T|$$

$\tilde{L}(\xi)^n$, $\overline{\tilde{L}(\xi)^n}^T$ haben beide denselben Aufbau wie $\tilde{J}_4(4)$ in Fig. 2, die Multiplikation der durch Streichen von Zeilen und Spalten entstehenden Untermatrizen $\tilde{L}_{\nu,\mu}(\xi)$ bzw. $\overline{L_{\nu,\mu}(\xi)}^T$ ist daher jeweils für ein Indexpaar ν,μ von dem anderen unabhängig. Daher hat auch die Matrix $\tilde{L}(\xi)^n \cdot \overline{\tilde{L}(\xi)^n}^T - 2I_{2m}$ dieselbe Struktur. Das charakteristische Polynom von $\tilde{L}(\xi)^n \cdot \overline{\tilde{L}(\xi)^n}^T$ zerfällt deshalb in die charakteristischen Polynome der

$$L_{\nu,\mu}(\xi)^n \cdot \overline{\tilde{L}_{\nu,\mu}(\xi)^n}^T$$

Mit anderen Worten: die Spektralnorm von $\tilde{L}(\xi)^n$ ist kleiner oder gleich der größten Spektralnorm der $\tilde{L}_{\nu,\mu}(\xi)^n$. Wegen (33) und aus der Definition von d in b) sowie p in c) folgt

$$\|\tilde{L}(\xi)^n\| \le 2p \frac{d^{p-1}}{(p-1)!\,\varepsilon^{2p-1}} \cdot (n+1)^{p-1}$$

für $\xi \in \omega_1, n \in \mathbb{N}_0, -1+\varepsilon \le \lambda \le 1-\varepsilon$

Damit ist (34) bewiesen.

Bem. 1

Falls die Amplifikationsmatrix $\tilde{A}_h(\mu)$ ein Minimalpolynom mit höchstens quadratischen Faktoren besitzt, und sich für sie die Bedingungen c) und d) erfüllen lassen, so folgt Konvergenz von der Ordnung 1 aus dem Äquivalenzsatz für solche $u_0(\cdot)$, wie sie in II A vorausgesetzt sind.

Bem. 2

Mittels apriori-Abschätzungen der Lösungen von (11) lassen sich oft die dritten Ableitungen von $u(\cdot,\cdot)$ die (11) auch lösen, eingrenzen. In solchen Fällen kann man die Konstante P_2 in (12) und damit $M(u_0)$ in (4) berechnen. Verwendet man im Beweis des Äquivalenzsatzes noch (34), so erhält man direkte Fehlerschranken für die Näherungslösung.

Beispiel: $u_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} u_x + \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} u_y$

Es ist: $B(\xi) = \begin{pmatrix} s_1+s_2 & s_1-s_2 \\ 0 & s_1+s_2 \end{pmatrix}$. Sei $s_\nu = \xi_\nu / R_\nu$, $\nu = 1,2$. Auf JORDAN'sche Normalform transformiert z.B. die Diagonalmatrix $\begin{pmatrix} s_1-s_2 & 0 \\ 0 & 1 \end{pmatrix}$:

$$\frac{1}{s_1-s_2}\begin{pmatrix} 1 & 0 \\ 0 & s_1-s_2 \end{pmatrix}\begin{pmatrix} s_1+s_2 & s_1-s_2 \\ 0 & s_1+s_2 \end{pmatrix}\begin{pmatrix} s_1-s_2 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} s_1+s_2 & 1 \\ 0 & s_1+s_2 \end{pmatrix},$$

die Transformation ist so also nur für $s_1 \neq s_2$ möglich. Wählt man dagegen $\tilde{R}_1 = R_2 = 1 - \frac{\varepsilon}{2}$, $\alpha = 1$, $p = 2$, so sind die Voraussetzungen des Satzes erfüllt.

IV Erweiterung auf ähnliche Schemata

Der Satz 4 zusammen mit der Kenntnis von Instabilitäten für den Fall, in welchem seine Voraussetzung a) nicht gegeben ist, lassen sich ohne Schwierigkeiten auf alle Schemata anwenden, die den Aufbau

(35) $$U^{n+1} - U^{n-1} = \sum_{\nu=1}^{N} \sum_{\varkappa=-l}^{l} A_{\nu\varkappa}(h) T_\nu^\varkappa U^n \qquad l \in \mathbb{N}_0$$ besitzen.

Statt $A_l(h)$ hat man dann die Matrizen

$$\sum_{\nu=1}^{N} \sum_{\varkappa=-l}^{l} A_{\nu\varkappa}(h)\, e^{i\varkappa l R_\nu} L_\nu$$ zu untersuchen.

Genausogut aber auch auf Schemata der Form:

(36) $$U^{n+1} - e^{i\pi\varphi} U^{n-1} = \sum_{\nu=1}^{N} \sum_{\varkappa=-l}^{l} A_{\nu\varkappa}(h) T_\nu^\varkappa U^n \; ; \; \varphi \in \mathbb{R}$$

mit anderen Matrizen $A_{\nu\varkappa}(h)$, denn (36) geht, durch die Transformation

$$U^n =: e^{i\pi \frac{\varphi}{l} n} V^n$$ in (35) über.

Anders ausgedrückt: die beiden $2m \times 2m$ Matrizen

$$\begin{pmatrix} A & I_m \\ e^{i\pi\varphi} I_m & 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} e^{-i\pi\frac{\varphi}{l}} A & I_m \\ I_m & 0 \end{pmatrix}$$ sind bis auf

einen komplexen Faktor vom Betrage 1 unitär kongruent wegen

$$\begin{pmatrix} e^{i\pi\frac{\varphi}{l}} I_m & 0 \\ 0 & I_m \end{pmatrix} \begin{pmatrix} A & I_m \\ e^{i\pi\varphi} I_m & 0 \end{pmatrix} \begin{pmatrix} e^{-i\pi\frac{\varphi}{l}} I_m & 0 \\ 0 & I_m \end{pmatrix} = e^{i\pi\frac{\varphi}{l}} \begin{pmatrix} e^{-i\pi\frac{\varphi}{l}} A & I_m \\ I_m & 0 \end{pmatrix}$$

Damit hat man eine große Auswahl zur Konstruktion stabiler, hochkonsistenter 2-Schritt Schemata.

Das Beispiel für hyperbolische Systeme mit konstanten Koeffizienten

$$U^{n+1} + U^{n-1} = \Big\{ 2 + \sum_{\nu=1}^{N} \frac{A_\nu^2}{R_\nu^2} \left(T_\nu - 2 + T_\nu^{-1} \right) + \tag{37}$$

$$+ \sum_{\substack{\nu<\ell \\ \nu,\ell=1}}^{N} \left(A_\nu A_\ell + A_\nu A_\ell \right) \frac{T_\nu - T_\nu^{-1}}{2R_\nu} \cdot \frac{T_\ell - T_\ell^{-1}}{2R_\ell} \Big\} U^n$$

gehört zur Klasse (36). Für ausreichend oft differenzierbare Lösungen des Systems (11) gilt:

$$U_{tt} = \sum_{\nu,\ell=1}^{N} A_\nu A_\ell \, U_{x_\nu x_\ell}$$

und für sie ist in (37) der Abschneidefehler ein Term $O(h^4)$.

Die Methode, bei einem zeitabhängigen System 1. Ordnung zum iterierten überzugehen, garantiert das Auftreten der zweiten Ableitungen in der Zeit, die sich besonders gut durch zentrierte Differenzen annähern lassen. Ersetzt man die übrigen Ableitungen durch ebenso genaue Differenzen so erhält man Konsistenz 3. Ordnung für genügend glatte Lösungen.

Die Stabilitätsuntersuchung wird für (37) nur im HERMITEschen Falle durchgeführt. Sei $\bar{A}_\nu^T = A_\nu$ für $\nu = 1, \cdots, N$. Dann ist

$$2A_\ell(k) := 2I_m - 4 \sum_{\nu=1}^{N} \frac{A_\nu^2}{R_\nu^2} \sin^2\left(h R_\nu \frac{k_\nu}{2}\right) +$$

$$- \sum_{\nu<\ell}^{N} \frac{A_\nu A_\ell + A_\ell A_\nu}{R_\nu R_\ell} \sin(h R_\nu k_\nu) \cdot \sin(h R_\ell k_\ell) \rightrightarrows$$

HERMITEsch und für ihre Eigenwerte $2\lambda_\ell(k)$ hat man die Relation (20) zu etablieren.

$$\rightrightarrows 2I_m - 4 \sum_{1} \frac{A_\nu^2}{R_\nu^2} \sin^2\left(h R_\nu \frac{k_\nu}{2}\right) - \left(\sum_{1}^{N} \frac{A_\nu}{R_\nu} \sin h R_\nu k_\nu \right)^2$$

$$+ \sum_{\nu=1}^{N} \frac{A_\nu^2}{R_\nu^2} \sin^2 h R_\nu k_\nu =$$

$$= 2I_m - 4 \sum_{1} \frac{A_\nu^2}{R_\nu^2} \sin^2(h \cdot R_\nu k_\nu) - \left(\sum \frac{A_\nu}{R_\nu} \sin h R_\nu k_\nu \right)^2 \rightrightarrows$$

wegen: $4\sin^2\alpha - \sin^2 2\alpha = 4\sin^4\alpha$ für $\alpha \in \mathbb{R}$

$$\Rightarrow: 2I_m - 2F_h(4)$$

Da $F_h(4)$ HERMITEsch ist, sind seine Eigenwerte in der Menge

$$L := \{(y, F_h(4)y),\ y \in \mathbb{R}^m,\ |y| = 1\}$$ zu finden.

Nun sind sowohl $\frac{A_\nu^2}{R_\nu^2}\sin^2 h R_\nu \frac{4_\nu}{2}$ für $\nu = 1, \cdots, N$

als auch $\left(\sum_1 \frac{A_\nu}{R_\nu}\sin h R_\nu 4_\nu\right)^2$ positiv semidefinit.

$\Rightarrow$

Für $r \in L$ gilt: $r \geq 0$

$\Rightarrow (|A_h(4)| \leq 1-\varepsilon \iff |F_h(4)| \leq 2-\varepsilon)$

für $0 < \varepsilon < 1$

Es gilt:

$$|F_h(4)| \leq 4\sum \frac{|A_\nu|^2}{R_\nu^2} + \left(\sum \frac{|A_\nu|}{R_\nu}\right)^2$$

wählt man $R_\nu \geq |A_\nu| \sqrt{\frac{4N+N^2}{2-\varepsilon}}$ so folgt:

$$|F_h(4)| \leq 4N\frac{2-\varepsilon}{4N+N^2} + N^2\frac{2-\varepsilon}{4N+N^2} = 2-\varepsilon$$

Man hat also das Ergebnis:

Für HERMITEsche Systeme ist (37) stabil, falls

$$R_\nu \geq |A_\nu| \cdot \sqrt{\frac{4N+N^2}{2-\varepsilon}}$$ gewählt wird.

V. Numerische Resultate

In diesem Abschnitt werden die beiden Schemata (10) und (37), im folgenden ZENT und DZENT genannt, mit dem von 2. Ordnung konsistenten 1-Schritt Schema (3.3) aus [6], im folgenden LAXW, in ihrer Genauigkeit für drei hyperbolische Systeme verglichen. Das LAX-WENDROFF Schema LAXW erhält man aus

$$u(t+h) = u(t) + h\, u_t(t) + \frac{h^2}{2} u_{tt}(t) + O(h^3)$$

worin alle auftretenden Zeitableitungen mittels (11) durch Raumableitungen ersetzt werden und dann $\frac{\partial}{\partial x_i}$ gegen $\frac{T_i - T_i^{-1}}{2R_i h} =: \Delta_i$, $\frac{\partial^2}{\partial x_i \partial x_\ell}$ gegen $\Delta_i \Delta_\ell$ (für $i \neq \ell$) und $\frac{\partial^2}{\partial x_i^2}$ gegen $\frac{T_i - 2 + T_i^{-1}}{R_i^2 h^2}$ ausgewechselt werden. LAXW benötigt als 1-Schritt Verfahren etwa halb so viel Speicherplatz wie ZENT und DZENT, dagegen etwas mehr Rechenaufwand als DZENT und erheblich mehr als ZENT.

Bei den folgenden Gleichungen, die ersten zwei inhomogene, symmetrische Systeme mit konstanten bzw. mit orts- und zeitabhängigen Koeffizienten, die dritte quasilinear, wurden für jeden Zeitschritt die maximal auftretenden Fehler der Näherungslösungen zur jeweilig gewählten exakten Lösung niedergeschrieben.

Die Anfangswerte für LAXW in der ersten Zeitschicht $(t=0)$ und für ZENT und DZENT in den ersten beiden Schichten wurden exakt vorgegeben. Um die Wirkung von Anfangsfehlern zu testen, wurden für ZENT und DZENT in einem weiteren Experiment in der ersten Schicht die exakten und in der zweiten die von LAXW berechneten Werte genommen. Die zugehörigen Fehler sind unter ZENTL bzw. DZENTL aufgeführt. Auf den Rändern des Raumgitters sind in allen Fällen die Werte der exakten Lösung genommen.

Die Raumgitter wurden in den drei Beispielen festgehalten und die Zeitschrittweite h bis an die Stabilitätsgrenze variiert.

1. Beispiel

$$u_t(t,x,y) = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} u_x(t,x,y) + \begin{pmatrix} 3 & 1 \\ 1 & -2 \end{pmatrix} u_y(t,x,y) + \\ - \begin{pmatrix} (2\cos x + 3\cos y)\cdot \cos t \\ (\sin x + 2\sin y)\cdot \sin t \end{pmatrix}$$

Exakte Lösung

$$u(t,x,y) = \begin{pmatrix} (\sin x + \sin y)\cdot \cos t \\ (\cos x + \cos y)\cdot \sin t \end{pmatrix}$$

In diesem wie auch im folgenden Beispiel wurde als Raumgitter die diskrete Menge

$$\left\{(x,y)\in \mathbb{R}^2;\ x = m\cdot\frac{2\pi}{100},\ y = n\cdot\frac{2\pi}{100},\ m,n\in \mathbb{N}_0,\ m,n\le 10\right\}$$

gewählt.

Die Indizes in den Tabellen geben die Komponenten der Vektoren an.

In den Tabellen auf Seite 73 zeigt sich fast kein Unterschied bei den Fehlern von ZENT und ZENTL. Die Genauigkeit des LAX-WENDROFF-Schemas reicht also aus, um die Anfangswerte zu berechnen. Zwischen DZENT und DZENTL gibt es schon größere Differenzen, in der ersten Komponente störend größere. Dieselben Unterschiede traten bei allen Rechnungen auf, ZENTL und DZENTL sind daher in den folgenden Tabellen nicht aufgeführt.

ZENT ist deutlich schlechter als LAXW und LAXW erheblich schlechter als DZENT in diesem Beispiel. Bemerkenswert sind die kleinen Fehler von DZENT für größtmögliche Schrittweiten h. Während für $h = 0.015$ LAXW und ZENT schon instabil sind, liefert DZENT besonders gute Ergebnisse. Für $h = 0.02$ ist auch DZENT instabil. Nahe der Stabili-

tätsgrenze ist das Verfahren also sehr genau. Daß ZENT noch vor DZENT instabil wird, mag an der Inhomogenität des Problems oder am "Festhalteeffekt" der Randwerte liegen.

2. Beispiel

$$u_t = \begin{pmatrix} t & y \\ y & t \end{pmatrix} u_x + \begin{pmatrix} t & x \\ x & t \end{pmatrix} u_y + (y+x) \cdot \begin{pmatrix} \sin(x+y+t^2) \\ -\cos(x+y+t^2) \end{pmatrix}$$

Exakte Lösung

$$u(t,x,y) = \begin{pmatrix} \sin(x+y+t^2) \\ \cos(x+y+t^2) \end{pmatrix}$$

Nach der Tabelle auf Seite 74 fällt DZENT in diesem Beispiel wegen der Größe seiner Fehler aus dem Rahmen. ZENT ist klar besser als LAXW. Nach [3] und [7] (s.auch [8] S.119) läßt sich die Stabilität von LAXW unter anderem für dieses Beispiel beweisen. Das dabei verwandte KREISSsche Matrix Theorem läßt sich sicher nicht auf ZENT und DZENT anwenden, da die Amplifikationsmatrizen dieser beiden Schemata im Stabilitätsbereich nur Eigenwerte des Absolutbetrages 1. haben, sie beide also niemals dissipativ sind (s.[8] s. 109). Die für Stabilität nicht notwendige Dissipativitätsbedingung ist also auch kein Garant für die Qualität eines Verfahrens.

Für $h = 0.02$ wurde ZENT, für $h = 0.025$ wurden auch LAXW und DZENT instabil.

3. Beispiel

$$u_t = u\,u_x$$

Exakte Lösung: $u(t,x) = \frac{x}{1-t} \quad ; \quad t \neq 1$

Ein Vergleich der beiden Tabellen liefert für um einen Faktor 10

veränderte Schrittweite stark veränderte Fehler. DZENT ist für $h=0.005$ das schlechteste, für $h=0.05$ das beste Verfahren. Umgekehrt verhält es sich mit ZENT, das im Falle $h=0.005$ die insgesamt kleinsten Fehler aufweist. Das Ergebnis legt dar, wie sehr im quasilinearen Falle die Qualität von Differenzenverfahren nicht nur vom gerade zu behandelnden System sondern auch von der Wahl der Schrittweite abhängt.

Maximale Fehler in 10^{-3} Einheit

Tabellen zum Beispiel 1

h = 0,01	LAXW1	ZENT1	ZENTL1	DZENT1	DZENTL1	LAXW2	ZENT2	ZENTL2	DZENT2	DZENTL2
t = 0.1	0.58	0.99	1.06	0.02	0.29	0.15	0.26	0.27	0.00	0.11
0.2	1.07	2.08	2.11	0.03	0.13	0.26	0.52	0.52	0.01	0.04
0.3	1.05	2.19	2.15	0.01	0.36	0.30	0.57	0.58	0.02	0.12
0.4	1.01	2.06	2.02	0.00	0.26	0.30	0.86	0.87	0.05	0.06
0.5	0.96	1.92	1.95	0.01	0.18	0.32	0.95	0.94	0.13	0.17
0.6	0.90	2.19	2.20	0.02	0.29	0.33	1.03	1.02	0.13	0.14
0.9	0.70	3.05	2.99	0.04	0.11	0.46	1.55	1.54	0.05	0.12

h = 0.015	LAXW	ZENT	DZENT	LAXW	ZENT	DZENT
t = 0.15	0.70	1.56	0.003	0.17	0.40	0.004
0.3	0.78	18.88	0.003	0.23	4.61	0.011
0.45	langsam	45763.05	0.009	langsam	10803.35	0.031
0.6	instabil	schnell	0.017	instabil	schnell	0.043
0.9		instabil	0.006		instabil	0.031
1.2			0.008			0.016
1.5	430.44		0.004	103.73		0.012

Maximale Fehler in 10^{-3} Einheit

Tabelle zum Beispiel 2

h = 0.015	LAXW	ZENT	DZENT	LAX	ZENT	DZENT
t = 0.3	0.87	0.24	10.79	2.21	0.19	38.03
0.6	3.60	0.36	32.60	2.85	0.31	81.21
0.9	12.88	0.34	57.64	4.26	1.31	87.72
1.2	13.70	1.31	39.52	7.60	1.69	23.06
1.5	4.06	2.19	19.98	12.97	1.64	55.90

Tabellen zum Beispiel 3

h = 0.005	LAXW	ZENT	DZENT
t = 0.1	0.02	0.01	0.15
0.2	0.06	0.02	0.65
0.3	0.10	0.03	1.47
0.4	0.16	0.05	2.59
0.5	0.24	0.09	3.96
0.6	0.36	0.30	158.83

h = 0.05	LAXW	ZENT	DZENT
t = 0.1	0.15	0.28	0.02
0.2	0.53	0.64	0.10
0.3	1.14	1.35	0.29
0.4	2.18	2.54	0.68
0.5	4.23	4.87	1.45
0.6	8.37	9.79	3.08
0.7	18.95	22.90	7.59
0.8	56.86	70.38	25.33
0.9	353.62	489,92	229.50

Literaturverzeichnis

[1] ANSORGE, R. und HASS, R.: Konvergenz von Differenzenverfahren für lineare und nichtlineare Anfangswertaufgaben. Lecture notes in mathematics. Springer 1970, Heidelberg.

[2] COURANT, R., FRIEDRICHS, K.O., und LEWY, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. Math.Ann., vol.100 (1928), p.32

[3] KREISS, H.O.: On difference approximations of the dissipative type for hyperbolic differential equations. Comm.Pure Appl.Math., vol.17(1964), p.335

[4] LAX, P.D.:Differential equations, difference equations and matrix theory. Comm.Pure Appl.Math.,vol.11(1958), p.175

[5] LAX, P.D. and RICHTMYER, R.D.:Survey of the stability of linear finite difference equations. Comm.Pure Appl. Math.,vol. 9 (1956), p.267

[6] LAX, P.D. and WENDROFF, B.: Difference schemes for hyperbolic equations with high order of accuracy.Comm.Pure Appl. Math., vol.17(1964), p.381

[7] PARLETT, B.N.:Accuracy and dissipation of difference schemes. Comm. Pure Appl. Math., vol. 19 (1966), p.111

[8] RICHTMYER, R.D. and MORTON, K.W.: Difference methods for initial value problems. Interscience Publishers 1967 2nd ed., New York

[9] RIORDAN, J.: Combinitorial Identities. WILEY 1968, New York

[10] GRÖBNER, W.:Matrizenrechnung. B.I.1966, Mannheim