# Numerische Behandlung linearer und semilinearer partieller differentiell-algebraischer Systeme mit Runge-Kutta-Methoden

**Dissertation**

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt der

Mathematisch-Naturwissenschaftlich-Technischen Fakultät
(mathematisch-naturwissenschaftlicher Bereich)
der Martin-Luther-Universität Halle-Wittenberg

von Kristian Debrabant

geb. am 12.12.1975 in Halberstadt

Gutachter:

1. Prof. Dr. K. Strehmel
2. Prof. Dr. M. Arnold
3. Prof. Dr. J. Lang
4. Prof. Dr. W. Lucht

Datum der Verteidigung: 15.10.2004

An dieser Stelle möchte ich Herrn Prof. Dr. K. Strehmel für die kontinuierliche, sehr intensive Betreuung meiner Arbeit und stets außergewöhnlich umfassende Unterstützung, die ich seit Beginn meines Studiums in Anspruch nehmen durfte, meinen herzlichen Dank aussprechen. In zahlreichen anregenden Diskussionen konnte ich von seiner Kompetenz, Geduld und Beharrlichkeit profitieren.
Mein besonderer Dank gilt auch allen Professoren und Mitarbeitern am Institut für Numerische Mathematik für das freundliche und kreative Arbeitsklima und ihre stets vorhandene Diskussions- und Hilfsbereitschaft, allen voran Herrn Prof. Dr. W. Lucht.

# Inhaltsverzeichnis

# Symbolverzeichnis

| | | |
|---|---|---|
| $\mathbb{1}_l$ | : | Einsvektor$(1,\ldots,1)^\top \in \mathbb{N}^l$ |
| $\mathcal{O}$ | : | Landausymbol: Für $f:\mathbb{R}\to\mathbb{R}^n$, $g:\mathbb{R}\to\mathbb{R}$, $g(t)\neq 0$ für $t\neq\alpha$ ist $f(t)=\mathcal{O}(g(t))$ für $t\to\alpha$, falls $\frac{\|f(t)\|}{\lvert g(t)\rvert}<\infty$ für $t\to\alpha$ |
| $\mathfrak{o}$ | : | Nullmatrix |
| $\otimes$ | : | Kronecker-Produkt |
| $\|\cdot\|$ | : | diskrete $L_2$-Norm |
| $\|\cdot\|_2$ | : | euklidische Norm |
| $\|\cdot\|_\infty$ | : | Unendlichnorm |
| $\|\cdot\|_{\mathbb{C}^n}$ | : | Vektornorm im $\mathbb{C}^n$ |
| $\|\cdot\|_{\mathbb{R}^n}$ | : | Vektornorm im $\mathbb{R}^n$ |
| $\|\cdot\|_M$ | : | Maximumnorm |
| $\alpha_{\vec{h}}$ | : | lokaler Ortsdiskretisierungsfehler zum Zeitpunkt $t_{m+1}$ |
| $\mathfrak{A}$ | : | Verfahrensmatrix des Runge-Kutta-Verfahrens |
| $b$ | : | Wichtungsvektor des Runge-Kutta-Verfahrens |
| $c$ | : | Knotenvektor des Runge-Kutta-Verfahrens |
| $\mathbb{C}$ ($\mathbb{C}^-$) | : | Menge der komplexen Zahlen (mit nichtpositivem Realteil) |
| $\delta_{m+1}$ | : | Residuenfehler des Runge-Kutta-Verfahrens zum Zeitpunkt $t_{m+1}$ |
| $\Delta_{m+1}^{(i)}$ | : | Residuenfehler der $i$-ten Stufe des Runge-Kutta-Verfahrens in $t_{m+1}$ |
| $\partial\Omega$ | : | Rand von $\Omega$ |
| $\mathrm{diag}_k\{A_k\}$ | : | Blockdiagonalmatrix mit den Blöcken $A_k$ |
| $M^{\mathfrak{D}}$ | : | Drazin-Inverse der quadratischen Matrix $M$ |
| $\vec{e}_i$ | : | $i$-ter Einheitsvektor |
| $e_{m+1}$ | : | globaler Gesamtdiskretisierungsfehler zum Zeitpunkt $t_{m+1}$ |
| $\eta_{m+1}$ | : | globaler Ortsdiskretisierungsfehler zum Zeitpunkt $t_{m+1}$ |
| $h_i$ | : | Ortsschrittweite in $x_i$-Richtung |
| $I_l$ | : | $l$-dimensionale Einheitsmatrix |
| $L_2(\Omega,\mathbb{R}^n)$ | : | Raum der über $\Omega$ quadratisch integrierbaren Funktionen mit Werten in $\mathbb{R}^n$ |
| $\lambda_{\max}(A)$ | : | Maximaler Eigenwert der Matrix $A$ |
| $\mu_2[A]$ | : | der euklidischen Vektornorm zugeordnete logarithmische Norm der Matrix $A$ |
| $\mathbb{N}$ | : | Menge der natürlichen Zahlen $\{0,1,\ldots\}$ |
| $\mathbb{N}^+$ | : | Menge der positiven natürlichen Zahlen $\{1,2,\ldots\}$ |
| $\nu_{dt}$ | : | differentieller Zeitindex des PDA-Systems |
| $\nu_{dx}$ | : | differentieller Ortsindex des PDA-Systems |
| $\Omega$ | : | offener d-dimensionaler Quader |
| $\Omega_{\vec{h}}$ | : | äquidistantes Ortsgitter mit dem Parameter $\vec{h}$, Diskretisierung von $\Omega$ |
| $p$ | : | Konsistenzordnung des Runge-Kutta-Verfahrens |
| $q$ | : | Stufenordnung des Runge-Kutta-Verfahrens |
| $\mathbb{R}$ | : | Menge der reellen Zahlen |
| $\mathfrak{Re}z$, $\mathfrak{Im}z$ | : | Real- bzw. Imaginärteil einer komplexen Zahl $z$ |
| $s$ | : | Stufenzahl des Runge-Kutta-Verfahrens |
| $\tau$ | : | Zeitschrittweite |
| $\mathbb{Z}$ | : | Menge der ganzen Zahlen |

# Kapitel 1

# Einleitung

Die mathematische Modellierung zahlreicher Probleme aus Naturwissenschaft und Technik führt auf Systeme partieller Differentialgleichungen, die aus einer Kopplung von Gleichungen unterschiedlichen Typs bestehen, zum Beispiel aus parabolischen, elliptischen und algebraischen Gleichungen. Die analytische und numerische Behandlung dieser sogenannten partiellen differentiell-algebraischen Gleichungen (PDA-Systeme, engl.: partial differential algebraic equations, PDAEs) haben in den letzten Jahren wachsendes Interesse gefunden.
Die vorliegende Arbeit liefert einen Beitrag zur numerischen Behandlung von PDA-Systemen. Im Vordergrund stehen dabei Untersuchungen zur Konvergenz von Zeitintegrationsverfahren vom Runge-Kutta-Typ.
Betrachtet wird eine spezielle Klasse von räumlich $d$-dimensionalen Anfangsrandwertaufgaben partieller differentiell-algebraischer Systeme, nämlich die semilinearen Systeme mit konstanten Koeffizienten der Form

$$A\, u_t(t,\vec{x}) + \sum_{i=1}^{d} B_i\ (u_{x_i x_i}(t,\vec{x}) + r_i u_{x_i}(t,\vec{x})) + C\, u(t,\vec{x}) = f(t,\vec{x},u),$$

wobei die Matrizen $A$ und $B_i$ singulär sein dürfen. Im Gegensatz zu Anfangsrandwertaufgaben parabolischer Differentialgleichungssysteme können bei PDA-Systemen nicht für alle Komponenten der Lösung $u(t,\vec{x})$ Anfangs- und Randbedingungen vorgeschrieben werden. Dies ist ein wesentliches Merkmal von PDA-Systemen.
Die betrachteten PDA-Systeme werden hier mit der Linienmethode numerisch gelöst: Zunächst erfolgt eine Semidiskretisierung bezüglich der räumlichen Variablen $\vec{x} = (x_1, \ldots, x_d)^\top$, das resultierende differentiell-algebraische System wird dann durch Runge-Kutta-Verfahren gelöst.
Im Fall skalarer parabolischer Anfangsrandwertprobleme ist bekannt, daß die Konvergenzordnung bezüglich der Zeit bei dieser Vorgehensweise im allgemeinen geringer ist als die klassische Konvergenzordnung des Runge-Kutta-Verfahrens, in Abhängigkeit vom Typ der Randwerte und ihrer Homogenität oder Inhomogenität kann eine Ordnungsreduktion auftreten, die Konvergenzordnung kann nichtganzzahlig sein (vgl. Sanz-Serna/Verwer/Hundsdorfer [46], Verwer [56], Ostermann/Roche [42]).
Bei differentiell-algebraischen Gleichungen wiederum hat sich das Indexkonzept auch zur Charakterisierung des Konvergenzverhaltens bewährt, die Konvergenzordnung eines Runge-Kutta-Verfahrens hängt im allgemeinen vom Index ab (vgl. Hairer/Wanner [26], Brenan/Campbell/Petzold [3]).
Durch Erweiterung des Indexkonzepts auf lineare PDA-Systeme wurde in Lucht/Strehmel/Eichler-Liebenow [37] für das BTCS- und das Crank-Nicolson-Verfahren gezeigt, daß dort ein ähnliches Verhalten auftritt.
Ziel der vorliegenden Arbeit ist die Untersuchung des Konvergenzverhaltens bei Anwendung allgemeiner Runge-Kutta-Verfahren. Dabei wird sich herausstellen, daß im allgemeinen bei-

de Ordnungsreduktionsphänomene auftreten können, das heißt, die Konvergenzordnung ist einerseits vom Typ der Randbedingungen und andererseits vom differentiellen Zeitindex des PDA-Systems abhängig.
In Kapitel 2 werden die verwendeten Hilfsmittel aus den Gebieten der Numerik gewöhnlicher Differentialgleichungen und der differentiell-algebraischen Systeme dargestellt. Der Zugriff auf dieses Material in den nachfolgenden Kapiteln wird dadurch erleichtert.
In Kapitel 3 werden die hier betrachteten Klassen semilinearer PDA-Systeme vorgestellt und einige praxisrelevante Beispiele angegeben. Ferner wird an diesen Beispielen gezeigt, wie die nicht vorschreibbaren Anfangs- und Randbedingungen bestimmt werden können.
Kapitel 4 beschäftigt sich mit der Semidiskretisierung der PDA-Systeme mittels finiter Differenzen, dem ersten Schritt der Linienmethode. Für lineare PDA-Systeme werden die Konsistenz und Konvergenz des entstehenden linearen differentiell-algebraischen Systems untersucht und zwei Konvergenzsätze angegeben. Der eine beruht auf der analytischen Darstellung der Lösung des differentiell-algebraischen Systems mittels der Drazin-Inversen, der andere benutzt eine Weierstraß-Kronecker-Zerlegung.
In Kapitel 5 werden die semidiskretisierten PDA-Systeme mittels Runge-Kutta-Methoden mit regulärer Verfahrensmatrix auch in der Zeit diskretisiert. Es werden hinreichende Bedingungen für die Konvergenz der Gesamtdiskretisierung im linearen Fall angegeben. Die Konvergenzordnung hängt dabei vom Typ der Randbedingungen, ihrer Homogenität oder Inhomogenität und dem differentiellen Zeitindex des PDA-Systems ab. Auch auf gebrochene (nichtganzzahlige) Konvergenzordnungen wird eingegangen.
Aufbauend auf den für lineare Systeme erzielten Ergebnissen werden Konvergenzaussagen für Lipschitz-stetige Funktionen $f$ hergeleitet. Abschließend werden die erhaltenen Resultate auf spezielle steifgenaue Runge-Kutta-Verfahren mit singulärer Verfahrensmatrix (z. B. die Lobatto-IIIC-Verfahren) ausgedehnt.
In Kapitel 6 wird die entwickelte Theorie am Beispiel zweier Runge-Kutta-Verfahren, des impliziten Euler-Verfahrens und des dreistufigen Radau-IIA-Verfahrens, angewendet. Beide Verfahren spielen in der numerischen Lösung von Differentialgleichungssystemen und differentiell-algebraischen Systemen eine wichtige Rolle. Die durchgeführten numerischen Experimente bestätigen die erzielten theoretischen Resultate.
Schließlich werden in Anhang B die Konvergenzresultate auf semilineare PDA-Systeme der Gestalt

$$A\,u_t(t,\vec{x}) + \sum_{i=1}^{d} B_i\,a_i(x_i)\left(\left(p_i(x_i)u_{x_i}(t,\vec{x})\right)_{x_i} + q_i(x_i)u(t,\vec{x})\right) + C\,u(t,\vec{x}) = f(t,\vec{x},u)$$

übertragen.

# Kapitel 2

# Mathematische Grundlagen

In diesem Kapitel werden für die nachfolgenden Untersuchungen wesentliche Grundlagen bereitgestellt.

## 2.1 Differentiell-algebraische Gleichungssysteme

### 2.1.1 Einführung

Klassisch verwendet man zur mathematischen Modellierung eines mechanischen Mehrkörpersystems mit $s$ Freiheitsgraden sogenannte verallgemeinerte Koordinaten oder Zustandskoordinaten, das sind Größen $q_1(t), \ldots, q_s(t)$, die die Lage des Systems zum Zeitpunkt $t$ völlig charakterisieren. Zur Vorhersage der Lage des Systems in zukünftigen Zeitpunkten benötigt man zusätzlich die verallgemeinerten Geschwindigkeiten $\dot{q}_1, \ldots, \dot{q}_s$. Insgesamt erhält man zur Beschreibung des Systemverhaltens zum Beispiel durch den Euler-Lagrange-Formalismus die Lagrange-Gleichungen zweiter Art und nach Transformation auf ein System erster Ordnung ein explizites System gewöhnlicher Differentialgleichungen

$$\dot{y}(t) = f(t, y(t)) \tag{2.1}$$

mit einem Anfangszustand

$$y(t_0) = y_0,$$

wobei $y(t)$ ein $2s$-dimensionaler Vektor ist. Neben Anfangswertproblemen kommen in der Praxis auch Zweipunktrandwertprobleme vor, bei denen die gesuchte Lösung $y(t)$ eine Randbedingung der Form

$$r(y(t_0), y(t_e)) = 0$$

erfüllt mit $t_0 \neq t_e$, wobei $r$ ein Vektor von Funktionen ist und die gleiche Dimension wie $y$ hat. Das System (2.1), das durch eine minimale Anzahl von Koordinaten charakterisiert ist, wird Zustandsform genannt.
Die Ermittlung der Zustandskoordinaten (zumeist krummlinige Koordinaten) ist oft schwierig. Einfacher und häufig auch automatisch durch Computerprogramme möglich ist die Modellierung in redundanten Koordinaten, das Ergebnis ist bei Anwendung des Euler-Lagrange-Formalismus nach Transformation auf ein System erster Ordnung ein implizites Differentialgleichungssystem der Form

$$F\left(t, y(t), \dot{y}(t)\right) = 0, \quad t \in [t_0, t_e], \quad y : [t_0, t_e] \to \mathbb{R}^n \tag{2.2}$$

mit Anfangs- bzw. Randbedingungen. Dabei ist

$$F : [t_0, t_e] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$$

stetig und stetig nach $\dot{y}(t) = \frac{dy(t)}{dt}$ differenzierbar. Die Jacobi-Matrix $F_{\dot{y}}(t, y, \dot{y})$ ist aber (in einer Umgebung der Lösung) singulär, das System ist also nicht lokal eindeutig nach $\dot{y}(t)$

auflösbar, sondern enthält auch algebraische Gleichungen. Deshalb wird (2.2) als differentiell-algebraisches System (DA-System) bezeichnet.
Die Formulierung (2.2) mit redundanten Koordinaten bezeichnet man auch als Deskriptorform. Diese wurde seit Beginn der achtziger Jahre verstärkt untersucht (vgl. Brenan/Campbell/Petzold [3], Griepentrog/März [21], Hairer/Wanner [26], Simeon [47], Arnold [2]). Sie läßt sich durch Einführung einer zusätzlichen Variablen $z$ in ein gekoppeltes System doppelter Dimension von gewöhnlichen Differentialgleichungen und nichtlinearen Gleichungen transformieren:

$$\dot{y} = z, \tag{2.3a}$$
$$0 = F(t, y, z). \tag{2.3b}$$

Damit können ohne Beschränkung der Allgemeinheit anstelle von Systemen der Gestalt (2.2) Systeme in semi-expliziter Form (2.3) betrachtet werden.
Als Lösung des DA-Systems (2.2) auf dem Intervall $[t_0, t_e]$ wird jede stetig differenzierbare Funktion

$$y(t) : [t_0, t_e] \to \mathbb{R}^n \quad \text{mit} \quad F(t, y(t), \dot{y}(t)) = 0 \quad \text{in} \quad [t_0, t_e]$$

bezeichnet. Entsprechend wird $y(t)$ als Lösung des Anfangswertproblems

$$F\left(t, y(t), \dot{y}(t)\right) = 0, \quad y(t_0) = y_0$$

bezeichnet, wenn $y$ Lösung des DA-Systems (2.2) ist und die Anfangsbedingung erfüllt.
Die Anfangsbedingung heißt konsistent, wenn das zugehörige Anfangswertproblem lösbar ist, d. h. mindestens eine Lösung besitzt.

**Bemerkung 2.1** Oftmals wird auch eine weniger strenge Definition für eine Lösung verwendet. Man fordert lediglich stetige Differenzierbarkeit in den Komponenten der Lösung, deren Ableitungen explizit in die Funktion $F$ eingehen. □

### 2.1.2 Der Differentiationsindex

Zur Klassifikation von DA-Systemen wurde von Gear [18] der Differentiationsindex eingeführt. Das DA-System (2.2) hat den Differentiationsindex $k$, wenn $k$ die kleinste Anzahl von Differentiationen ist, so daß

$$\begin{aligned} F(t, y, \dot{y}) &= 0, \\ \frac{d}{dt} F(t, y, \dot{y}) &= 0, \\ &\vdots \\ \frac{d^k}{dt^k} F(t, y, \dot{y}) &= 0 \end{aligned} \tag{2.4}$$

durch algebraische Umformungen in ein explizites gewöhnliches Differentialgleichungssystem

$$\dot{y}(t) = \phi(t, y) \tag{2.5}$$

überführt werden kann. Dieses heißt das dem DA-System zugrundeliegende Differentialgleichungssystem.
Ein gewöhnliches Differentialgleichungssystem hat also den Differentiationsindex 0.
Numerische Differentiation ist bekanntlich ein instabiler Prozeß, so daß der Differentiationsindex ein Maß für den Schwierigkeitsgrad der numerischen Behandlung des DA-Systems ist, er charakterisiert den algebraischen Anteil.
Das System (2.5) ist im allgemeinen nicht äquivalent zum System (2.4) und damit zu (2.2), weil (2.4) Zwangsbedingungen der Gestalt $r(y, t) = 0$ enthalten kann, siehe Beispiel 2.2. Die

Anfangs- oder Randwerte müssen alle Zwangsbedingungen erfüllen. In diesem Fall sind sie konsistent. Zur numerischen Lösung von Anfangswertproblemen werden stets konsistente Anfangswerte benötigt. Ihre Ermittlung stellt im allgemeinen ein schwieriges Problem dar und ist meist nicht analytisch möglich. Eine Möglichkeit zur Bestimmung konsistenter Anfangswerte von Mehrkörperproblemen ist in Simeon/Führer/Rentrop [49] angegeben und erfolgt über die Berechnung von Lösungen im statischen Gleichgewicht. Weitere Ansätze findet man zum Beispiel in Pantelides [43], Leimkuhler/Petzold/Gear [31] und Brown/Hindmarsh/Petzold [6].
Die Gewinnung des zugrundeliegenden Differentialgleichungssystems und der Zwangsbedingungen zeigt das folgende Beispiel:

**Beispiel 2.2** Gegeben sei für $y > 0$ das DA-System (vgl. März [38], Hoschek [27])

$$\dot{x} = \alpha x, \tag{2.6a}$$

$$\dot{y} = \frac{z}{y}, \tag{2.6b}$$

$$x^2 + y^2 = 1. \tag{2.6c}$$

Wegen (2.6c) liegt die Lösung auf einem Kreiszylinder mit dem Radius 1 um die $z$-Achse. Durch Differentiation dieser Zwangsbedingung und Einsetzen von (2.6a) und (2.6b) erhält man

$$\alpha x^2 + z = 0, \tag{2.7}$$

die Lösung muß also zusätzlich auf einem parabolischen Zylinder senkrecht zur $(x, z)$-Ebene liegen, dies ist eine „versteckte“ Zwangsbedingung.

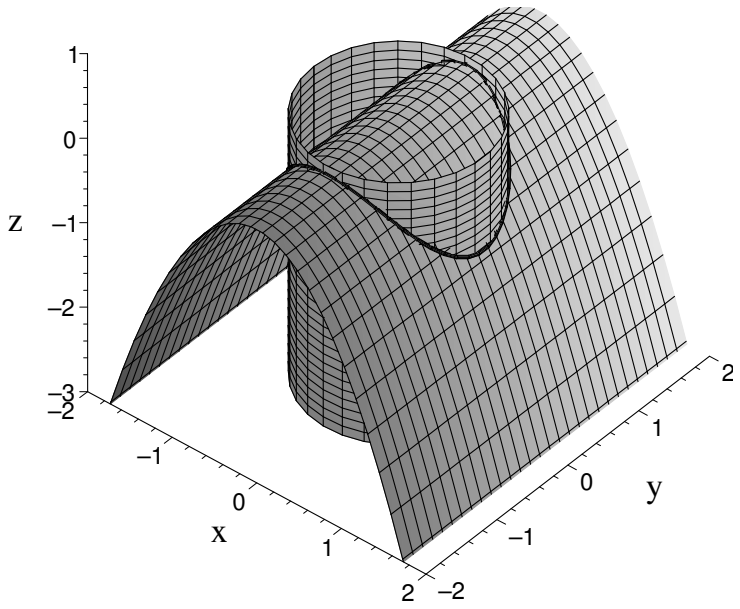


Abbildung 2.1: Schnitt von parabolischem Zylinder und Kreiszylinder

Differentiation von (2.7) und Einsetzen von (2.6a) liefert

$$\dot{z} = -2\alpha^2 x^2. \tag{2.8}$$

Das System (2.6a), (2.6b) und (2.8) ist ein explizites gewöhnliches Differentialgleichungssystem erster Ordnung, das dem DA-System (2.6) zugrundeliegt. (2.6) hat also den Differentiationsindex 2. Die Gleichungen (2.6c) und (2.7) bilden das System der Zwangsbedingungen. □

Für die numerische Lösung von DA-Systemen mit höherem Index ($k \geq 2$) ist die Rückführung des DA-Systems auf das zugrundeliegende explizite gewöhnliche Differentialgleichungssystem häufig nicht vorteilhaft, weil dessen Näherungslösungen die Zwangsbedingungen im Ausgangssystem nicht mehr exakt erfüllen, die numerische Lösung „driftet“ von der Zwangsmannigfaltigkeit ab.

**Bemerkung 2.3** Weitere Indexkonzepte sind der von Griepentrog und März [21] eingeführte Traktabilitätsindex, der anstelle von Differentiationen höherer Ordnung auf geeigneten Projektionen beruht, und der von Hairer, Lubich und Roche [24] eingeführte Störungsindex (vgl. auch Deuflhard/Bornemann [15]). Während der Differentiationsindex den Unterschied zwischen DA-Systemen und gewöhnlichen Differentialgleichungen durch Überführen der einen Klasse in die andere Klasse charakterisiert, ist der Störungsindex ein Maß für die Kondition eines gestellten Anfangswertproblems hinsichtlich der Störungen der rechten Seite des DA-Systems. Für die numerische Behandlung eines DA-Systems ist er daher wesentlicher als der Differentiationsindex. Seine Bestimmung ist aber im allgemeinen schwierig, so daß man sich häufig auf den Differentiationsindex beschränkt, welcher nach Gear [19] stets kleiner oder gleich dem Störungsindex ist.
Der von Kunkel und Mehrmann [29] für lineare DA-Systeme mit variablen Koeffizienten eingeführte Strangeness-Index verallgemeinert den Begriff des Differentiationsindex in dem Sinn, daß das DA-System auch Lösungen mit unbestimmten Komponenten haben kann. Sind der Differentiationsindex $k$ und der Strangeness-Index $s$ beide definiert und ist $k > 0$, dann gilt $k = s + 1$. □

Da für die Behandlung semilinearer partieller differentiell-algebraischer Systeme Lösungseigenschaften linearer DA-Systeme benötigt werden, sollen diese jetzt vorgestellt werden.

### 2.1.3 Lineare DA-Systeme

Betrachtet wird ein differentiell-algebraisches Anfangswertproblem der Form

$$B\dot{y} = Ay + f(t), \quad t \in [t_0, t_e], \quad y(t_0) = y_0 \tag{2.9}$$

mit konstanten Matrizen $B$, $A \in \mathbb{R}^{n,n}$, $B$ singulär, und einer gegebenen vektorwertigen Funktion

$$f : [t_0, t_e] \to \mathbb{R}^n.$$

Die Lösbarkeit von (2.9) ist abhängig von der Familie $\{A + \lambda B\}_{\lambda \in \mathbb{C}}$, die Matrizenbüschel genannt wird.

**Definition 2.4** Das Matrizenbüschel $\{A + \lambda B\}$ heißt regulär, wenn ein $\lambda_r \in \mathbb{C}$ existiert, so daß $A + \lambda_r B$ regulär ist,

$$\det(A + \lambda B) \not\equiv 0,$$

andernfalls heißt es singulär. □

Ist das Matrizenbüschel $\{A + \lambda B\}$ singulär, so existieren entweder keine oder unendlich viele Lösungen von (2.9) zu einem gegebenen Anfangswert, vgl. Hairer/Wanner [26].
Für reguläre Matrizenbüschel gilt nach Weierstraß und Kronecker folgender Satz:

**Satz 2.5** Sei $\{A + \lambda B\}$ ein reguläres Matrizenbüschel. Dann existieren reguläre Matrizen $P$, $Q \in \mathbb{C}^{n,n}$, so daß

$$PAQ = \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix}, \quad PBQ = \begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix},$$

wobei die blockdiagonale Matrix $N$ durch

$$N = \operatorname{diag}(N_{m_1}, \ldots, N_{m_k}) \quad \text{mit} \quad N_{m_i} = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{pmatrix} \in \mathbb{C}^{m_i, m_i}$$

gegeben und $C$ in Jordanscher Normalform ist, also

$$C = \operatorname{diag}(R_1, \ldots, R_l) \quad \text{mit} \quad R_i = \begin{pmatrix} \kappa_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \kappa_i & 1 \\ 0 & & & \kappa_i \end{pmatrix} \in \mathbb{C}^{n_i, n_i}.$$

□

Mit diesem Satz kann (2.9) nun transformiert werden:
Multipliziert man (2.9) von links mit $P$ und setzt

$$y = Q\begin{pmatrix} u \\ v \end{pmatrix}, \quad Pf(t) = \begin{pmatrix} s(t) \\ q(t) \end{pmatrix},$$

so erhält man das entkoppelte System

$$\dot{u} = Cu + s(t), \quad N\dot{v} = v + q(t).$$

Die Gleichung für $u$ ist eine lineare Differentialgleichung erster Ordnung und besitzt für beliebige Anfangswerte $u(t_0) = u_0$ eine eindeutige Lösung.
Da $N$ eine blockdiagonale Matrix ist, wird das Gleichungssystem für $v$ erneut in $k$ Systeme der Gestalt

$$\begin{aligned} \dot{v}_{i,2} &= v_{i,1} + q_{i,1}(t), \\ &\vdots \\ \dot{v}_{i,m_i} &= v_{i,m_i-1} + q_{i,m_i-1}(t), \\ 0 &= v_{i,m_i} + q_{i,m_i}(t) \end{aligned}$$

entkoppelt, wobei zur Abkürzung

$$v_{i,j} = v_{j+\sum_{l=1}^{i-1} m_l} \quad \text{und} \quad q_{i,j} = q_{j+\sum_{l=1}^{i-1} m_l}$$

für $j = 1, \ldots, m_i$ und $i = 1, \ldots, k$ gesetzt wurden.
$v_{i,1}$ hängt folglich von der $(l-1)$-ten Ableitung von $q_{i,l}$, $l = 1, \ldots, m_i$, ab. Das Maximum $m$ aller $m_i$ ist gleich der Nilpotenz der Matrix $N$ (d. h. $N^m = 0$, $N^{m-1} \neq 0$) und wird als Nilpotenzindex $\operatorname{ind}(A, B)$ oder auch als algebraischer Index des Matrizenbüschels $\{A + \lambda B\}$ bezeichnet.
Für reguläre Matrizenbüschel gilt also: Ist $f(t)$ $(m-1)$-mal stetig differenzierbar, so existiert zu jedem Anfangswert

$$y_0 = Q\begin{pmatrix} u_0 \\ v_0 \end{pmatrix},$$

wobei $v_0$ zu $q(t)$ konsistent vorgegeben sein muß, d. h. durch $q(t)$ und dessen Ableitungen bis zur $(m-1)$-ten Ordnung an der Stelle $t_0$ bestimmt ist, eine eindeutig bestimmte Lösung $y(t, u_0)$, die aber nicht notwendig in jeder Komponente differenzierbar sein muß. Das ist ein wesentlicher Unterschied zu den Lösungseigenschaften expliziter gewöhnlicher Differentialgleichungssysteme mit konstanten Koeffizienten.

Der Nilpotenzindex ist unabhängig von der Wahl der Matrizen $P$ und $Q$ (vgl. Hairer/Wanner [26]) und stimmt mit dem Differentiationsindex und dem Störungsindex überein (vgl. Brenan/Campbell/Petzold [3]).
Für reguläre Matrizen $S$ und $T$ folgen mit $\tilde{P} = PS^{-1}$ und $\tilde{Q} = T^{-1}Q$

$$\tilde{P}(SAT)\tilde{Q} = PAQ \quad \text{und} \quad \tilde{P}(SBT)\tilde{Q} = PBQ,$$

also

$$\operatorname{ind}(A, B) = \operatorname{ind}(SAT, SBT).$$

Ist $\det(A + \lambda_r B) \neq 0$ mit $\lambda_r \in \mathbb{C}$, so gilt mit den Bezeichnungen aus Satz 2.5

$$\begin{aligned}(A + \lambda_r B)^{-1} B &= Q \left(P(A + \lambda_r B)Q\right)^{-1} PBQQ^{-1} \\ &= Q \begin{pmatrix} (C + \lambda_r I)^{-1} & 0 \\ 0 & (I + \lambda_r N)^{-1} N \end{pmatrix} Q^{-1} \\ &= Q \begin{pmatrix} (C + \lambda_r I)^{-1} & 0 \\ 0 & N \sum_{i=0}^{m-2} (-\lambda_r N)^i \end{pmatrix} Q^{-1}.\end{aligned}$$

Daraus folgt: Definiert man den Index $\operatorname{ind}(M)$ einer quadratischen Matrix $M$ als kleinste nichtnegative ganze Zahl, für die

$$\operatorname{rang}(M^k) = \operatorname{rang}(M^{k+1})$$

gilt, so ist

$$\operatorname{ind}(A, B) = \operatorname{ind}\left((A + \lambda_r B)^{-1} B\right). \tag{2.10}$$

Mit dieser Beziehung kann der Index des Matrizenbüschels ohne Transformation auf Weierstraß-Kronecker-Normalform bestimmt werden.
Eine explizite Lösungsdarstellung des DA-Systems (2.9) ohne Transformation auf Weierstraß-Kronecker-Normalform ist mittels der Drazin-Inversen möglich, vgl. Campbell/Meyer/Rose [10], Campbell [9] und Simeon/Führer/Rentrop [50]:

**Definition 2.6** Für jede quadratische Matrix $M$ ist die Drazin-Inverse $M^{\mathfrak{D}}$ von $M$ definiert als Lösung der drei Gleichungen

$$\begin{aligned} MM^{\mathfrak{D}} &= M^{\mathfrak{D}}M, \\ M^{\mathfrak{D}}MM^{\mathfrak{D}} &= M^{\mathfrak{D}}, \\ M^{\mathfrak{D}}M^{k+1} &= M^k, \quad \text{wobei} \quad k = \operatorname{ind}(M). \end{aligned}$$

□

Die Drazin-Inverse einer quadratischen Matrix $M$ existiert stets eindeutig und stimmt für reguläre $M$ mit $M^{-1}$ überein. Sie ist verträglich mit Ähnlichkeitstransformationen, d. h., mit regulären Matrizen $S$ gilt

$$(S^{-1}MS)^{\mathfrak{D}} = S^{-1}M^{\mathfrak{D}}S. \tag{2.11}$$

Wird $M$ mit einer regulären Matrix $T$ auf Jordansche Normalform transformiert, d. h.

$$M = T^{-1} \begin{pmatrix} R & 0 \\ 0 & N \end{pmatrix} T,$$

wobei die Matrix $R$ die Jordan-Blöcke zu den von Null verschiedenen Eigenwerten von $M$ und die (nilpotente) Matrix $N$ die Jordan-Blöcke der Nulleigenwerte von $M$ enthält, so folgt deshalb

$$M^{\mathfrak{D}} = T^{-1} \begin{pmatrix} R^{-1} & 0 \\ 0 & 0 \end{pmatrix} T.$$

Gilt für zwei quadratische Matrizen $M_1$, $M_2$

$$M_1 M_2 = M_2 M_1,$$

so ist auch

$$M_1 M_2^{\mathfrak{D}} = M_2^{\mathfrak{D}} M_1$$

erfüllt.
Mit der so definierten Drazin-Inversen kann man die Lösung linearer DA-Systeme mit konstanten Koeffizienten nun wie folgt darstellen:

**Satz 2.7** Sei $\{A + \lambda B\}$ ein reguläres Matrizenbüschel, und $c \in \mathbb{C}$ werde so gewählt, daß $(A + cB)$ invertierbar ist. Mit den Abkürzungen

$$\hat{A} = (A + cB)^{-1} A, \quad \hat{B} = (A + cB)^{-1} B, \quad \hat{f} = (A + cB)^{-1} f$$

und $k = \operatorname{ind}(\hat{B}) = \operatorname{ind}(A, B)$ gilt dann:
Das differentiell-algebraische Anfangswertproblem (2.9) hat genau dann eine eindeutige Lösung, wenn $f(t)$ genügend oft stetig differenzierbar ist und der Anfangswert $y_0$ sich in der Form

$$y_0 = \hat{B}\hat{B}^{\mathfrak{D}} q - (I - \hat{B}\hat{B}^{\mathfrak{D}}) \sum_{i=0}^{k-1} \left(\hat{B}\hat{A}^{\mathfrak{D}}\right)^i \hat{A}^{\mathfrak{D}} \hat{f}^{(i)}(t_0) \tag{2.12}$$

mit einem Vektor $q$ darstellen läßt, d. h. konsistent ist. Für die Lösung gilt dann

$$y(t) = e^{\hat{B}^{\mathfrak{D}}\hat{A}(t-t_0)} \hat{B}\hat{B}^{\mathfrak{D}} y_0 + \hat{B}^{\mathfrak{D}} \int_{t_0}^{t} e^{\hat{B}^{\mathfrak{D}}\hat{A}(t-s)} \hat{f}(s)\, ds - (I - \hat{B}\hat{B}^{\mathfrak{D}}) \sum_{i=0}^{k-1} \left(\hat{B}\hat{A}^{\mathfrak{D}}\right)^i \hat{A}^{\mathfrak{D}} \hat{f}^{(i)}(t). \tag{2.13}$$

Dabei sind die Gleichungen (2.12) und (2.13) unabhängig von dem gewählten $c$. □

**Bemerkung 2.8** Während $A$ und $B$ im allgemeinen nicht kommutieren, gilt wegen

$$A(A + cB)^{-1} B + cB(A + cB)^{-1} B = B(A + cB)^{-1} A + B(A + cB)^{-1} cB$$

stets

$$A(A + cB)^{-1} B = B(A + cB)^{-1} A$$

und damit

$$\hat{A}\hat{B} = \hat{B}\hat{A}.$$

□

## 2.2 Runge-Kutta-Verfahren

Eine umfangreiche Klasse von Standardverfahren zur numerischen Lösung von Anfangswertproblemen expliziter gewöhnlicher Differentialgleichungssysteme

$$\begin{aligned} \dot{y}(t) &= f(t, y), \\ y(t_0) &= y_0 \end{aligned} \tag{2.14}$$

mit $t \in [t_0, t_e]$, $y : [t_0, t_e] \to \mathbb{R}^n$ und $f : [t_0, t_e] \times \mathbb{R}^n \to \mathbb{R}^n$ sind die Runge-Kutta-Verfahren. Ein $s$-stufiges Runge-Kutta-Verfahren auf dem Punktgitter

$$I_\tau = \{t_0, t_1, \ldots, t_{M_e}\}, \quad t_0 < t_1 < \ldots < t_{M_e} \le t_e,$$

mit den Schrittweiten $\tau = t_{m+1} - t_m$ (auf den Index bei der Schrittweite $\tau$ wird wie üblich verzichtet), $m = 0, \ldots, M_e - 1$, ist gegeben durch

$$u_{m+1} = u_m + \tau \sum_{i=1}^{s} b_i k_i(t_m, u_m; \tau), \tag{2.15a}$$

$$u_{m+1}^{(i)} = u_m + \tau \sum_{j=1}^{s} a_{ij} k_j(t_m, u_m; \tau), \quad i = 1, \ldots, s, \tag{2.15b}$$

$$k_i(t_m, u_m; \tau) = f(t_m + c_i\tau, u_{m+1}^{(i)}), \quad i = 1, \ldots, s, \tag{2.15c}$$

mit der Verfahrensmatrix $\mathfrak{A} = (a_{ij})_{i,j=1,\ldots,s}$, dem Wichtungsvektor $b = (b_1, \ldots, b_s)^\top$ und dem Knotenvektor $c = (c_1, \ldots, c_s)^\top$, die das Verfahren vollständig festlegen. Die Größen $u_{m+1}^{(i)}$ werden als Stufenwerte, $k_{m+1}^{(i)}$ als Steigungswerte bezeichnet. Das Verfahren liefert mit $u_{m+1}$ eine Näherung für $y(t_{m+1})$.
Ist $\mathfrak{A}$ eine strikt untere Dreiecksmatrix, so lassen sich die Gleichungen explizit nach den zu berechnenden Werten $u_{m+1}, u_{m+1}^{(i)}$ auflösen und man nennt das Verfahren explizit, andernfalls implizit.

**Definition 2.9** Ein Runge-Kutta-Verfahren besitzt die Konsistenzordnung $p$, wenn $p$ die größte positive ganze Zahl ist, so daß für jede genügend oft differenzierbare Lösung $y(t)$ von (2.14)

$$\max_{t \in [t_0, t_e - \tau]} \|y(t + \tau) - \hat{u}(t + \tau)\|_{\mathbb{R}^n} \leq C\tau^{p+1} \quad \text{für alle} \quad \tau \in (0, \tau_{max}]$$

gilt mit einer von $\tau$ unabhängigen Konstanten $C$ und einer positiven Konstanten $\tau_{max}$, wobei $\hat{u}(t + \tau)$ das Resultat eines Schrittes des Runge-Kutta-Verfahrens mit dem Startvektor auf der exakten Lösungskurve ist, d. h. $\hat{u}(t) = y(t)$, und $\|\cdot\|_{\mathbb{R}^n}$ eine Vektornorm im $\mathbb{R}^n$ ist. □

Zur Bestimmung geeigneter Parameter für konsistente Runge-Kutta-Verfahren ($p \geq 1$) wurden von Butcher vereinfachende Bedingungen eingeführt (vgl. Butcher [7]), von denen im folgenden zwei verwendet werden. Die vereinfachende Bedingung

$$B(p): \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, p \tag{2.16}$$

bedeutet, daß die dem Runge-Kutta-Verfahren zugrundeliegende Quadraturmethode Polynome bis zum Grad $p - 1$ exakt integriert. Für ein Runge-Kutta-Verfahren der Ordnung $p$ muß sie deshalb stets erfüllt sein.
Die vereinfachende Bedingung

$$C(l): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, \ldots, s, \quad k = 1, \ldots, l \tag{2.17}$$

besagt entsprechend, daß die den Stufenwerten $u_{m+1}^{(i)}$ zugrundeliegenden Quadraturmethoden mindestens den Genauigkeitsgrad $l-1$ besitzen. Das maximale $l$, für das $B(l)$ und $C(l)$ erfüllt sind, wird als Stufenordnung $q$ des Runge-Kutta-Verfahrens bezeichnet. Ist $q \geq 1$, so sind die Stufenwerte $u_{m+1}^{(i)}$ Approximationen an die Lösung $y$ von (2.14) an der Stelle $t_m + c_i\tau$ und die Steigungswerte $k_i(t_m, u_m; \tau)$ Approximationen an $\dot{y}(t_m + c_i\tau)$, d. h., es gelten für alle $\tau \in (0, \tau_{max}]$

$$\max_{t \in [t_0, t_e - \tau]} \|y(t + c_i\tau) - \hat{u}^{(i)}(t + c_i\tau)\|_{\mathbb{R}^n} \leq C\tau^{q+1}$$

und

$$\max_{t \in [t_0, t_e - \tau]} \|\dot{y}(t + c_i\tau) - k_i(t, y(t); \tau)\|_{\mathbb{R}^n} \leq C\tau^{q+1},$$

wobei $\hat{u}^{(i)}(t + c_i\tau)$ Stufenwert eines Runge-Kutta-Schrittes mit dem Startvektor auf der exakten Lösungskurve ist.
Durch Ausnutzung entsprechender Quadraturformeln kann man $s$-stufige Runge-Kutta-Verfahren hoher Ordnung gewinnen. So erhält man durch Festlegung des Wichtungsvektors $b$ und des Knotenvektors $c$ entsprechend der Gauß-Legendre-Quadraturmethoden die Gauß-Verfahren mit der maximal möglichen Ordnung $2s$, die Verfahrensmatrix $\mathfrak{A}$ wird dabei aus der Forderung $C(s)$ bestimmt. Analog erhält man aus den Radau- bzw. Lobatto-Formeln die Radau- bzw. Lobatto-III-Verfahren mit der Ordnung $2s-1$ bzw. $2s-2$. Für die Wahl von $\mathfrak{A}$ gibt es hier verschiedene Möglichkeiten, entsprechend unterscheidet man Radau-I-, -IA-, -II-, -IIA- und Lobatto-IIIA-, -IIIB-, -IIIC-Verfahren.

**Bemerkung 2.10** Nach Strehmel/Weiner [54] sind die Verfahrensmatrizen der Gauß-, Radau-IA-, -IIA- und Lobatto-IIIC-Verfahren regulär. □

Die Funktion

$$R(z) = 1 + zb^\top (I_s - z\mathfrak{A})^{-1}\mathbb{1}_s$$

wird als Stabilitätsfunktion des Runge-Kutta-Verfahrens bezeichnet. Sie ist dadurch charakterisiert, daß das Runge-Kutta-Verfahren bei Anwendung auf die von Dahlquist [13] eingeführte Testgleichung

$$y' = \lambda y, \quad \mathfrak{Re}\lambda \le 0 \tag{2.18}$$

die Verfahrensvorschrift

$$u_{m+1} = R(z)u_m, \quad z = \tau\lambda$$

liefert.
Die exakte Lösung

$$y(t_m + \tau) = e^{\tau\lambda} y(t_m)$$

von (2.18) hat einen monoton fallenden Betrag, und es gilt

$$\lim_{\tau\mathfrak{Re}\lambda\to-\infty} y(t_m + \tau) = 0. \tag{2.19}$$

Gilt

$$|R(z)| \le 1 \quad \text{für alle } z \text{ mit } \mathfrak{Re}(z) \le 0,$$

so hat auch die numerische Lösung einen monoton fallenden Betrag, und das Runge-Kutta-Verfahren heißt A-stabil, die Stabilitätsfunktion $A$-verträglich. Beispiele dafür sind die Gauß-, Radau-IA-, -IIA- und Lobatto-IIIA-, -IIIB-, -IIIC-Verfahren.
Gilt zusätzlich

$$\lim_{\mathfrak{Re}z\to-\infty} R(z) = 0,$$

so überträgt sich außerdem die Eigenschaft (2.19), und das Verfahren wird L-stabil genannt. Erfüllt ein Runge-Kutta-Verfahren die Bedingung

$$a_{si} = b_i, \quad i = 1, \ldots, s, \tag{2.20}$$

so nennt man es steifgenau. Für steifgenaue Verfahren mit regulärer Verfahrensmatrix $\mathfrak{A}$ gilt $\lim_{z\to-\infty} R(z) = 0$. Steifgenaue A-stabile Runge-Kutta-Verfahren mit regulärer Verfahrensmatrix $\mathfrak{A}$ wie zum Beispiel die Radau-IIA- und Lobatto-IIIC-Verfahren sind also stets L-stabil.

**Bemerkung 2.11** Eine andere Darstellung der Stabilitätsfunktion, die auf Stetter [51] zurückgeht, ist

$$R(z) = \frac{\det\left(I_s - z\mathfrak{A} + z\mathbb{1}_s b^\top\right)}{\det\left(I_s - z\mathfrak{A}\right)}. \tag{2.21}$$

□

Um Diskretisierungsverfahren zur numerischen Lösung von DA-Systemen zu erhalten, kann man das System (2.3) in ein singulär gestörtes Anfangswertproblem mit dem Parameter $\varepsilon > 0$ einbetten,

$$\dot{y} = z, \tag{2.22a}$$
$$\varepsilon \dot{z} = F(t, y, z), \tag{2.22b}$$

und darauf ein Diskretisierungsverfahren für gewöhnliche Differentialgleichungssysteme anwenden. Der Grenzübergang $\varepsilon \to 0$ liefert dann ein Diskretisierungsverfahren für das DA-System. Diese Vorgehensweise bezeichnet man als „direkten Weg“, vgl. Strehmel/Weiner [54]. Dabei ist zu beachten, daß das System (2.22) für kleine $\varepsilon$ „steif“ wird, das heißt, explizite Verfahren sind wegen ihrer schlechten Stabilitätseigenschaften nicht anwendbar, sie besitzen nur ein beschränktes Stabilitätsgebiet. Auch die Lobatto-IIIB-Verfahren sind für die numerische Behandlung steifer Systeme nicht geeignet, obwohl sie A-stabil sind, vgl. Hairer/Wanner [26], S. 227.
Wendet man das Verfahren (2.15) auf (2.22) an, so erhält man mit (2.15c)

$$\begin{pmatrix} k_i \\ \varepsilon l_i \end{pmatrix} = \begin{pmatrix} v_{m+1}^{(i)} \\ F(t_m + c_i\tau, u_{m+1}^{(i)}, v_{m+1}^{(i)}) \end{pmatrix}$$

und daraus durch Grenzübergang $\varepsilon \to 0$

$$0 = F(t_m + c_i\tau, u_{m+1}^{(i)}, k_i).$$

Mit (2.15a) und (2.15b) ergibt sich insgesamt die Vorschrift

$$u_{m+1} = u_m + \tau \sum_{i=1}^{s} b_i k_i(t_m, u_m; \tau), \tag{2.23a}$$
$$u_{m+1}^{(i)} = u_m + \tau \sum_{j=1}^{s} a_{ij} k_j(t_m, u_m; \tau), \quad i = 1, \ldots, s, \tag{2.23b}$$
$$0 = F(t_m + c_i\tau, u_{m+1}^{(i)}, k_i(t_m, u_m; \tau)), \quad i = 1, \ldots, s. \tag{2.23c}$$

## 2.3 Weitere Hilfsmittel

### 2.3.1 Kronecker-Produkt

Für eine kompakte Schreibweise von Gleichungssystemen ist oft das Kronecker-Produkt vorteilhaft.
Seien $A \in \mathbb{R}^{k,l}$ und $B \in \mathbb{R}^{m,n}$ zwei Matrizen. Ihr Kronecker-Produkt $A \otimes B$ ist definiert durch

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1l}B \\ \vdots & & \vdots \\ a_{k1}B & \cdots & a_{kl}B \end{pmatrix} \in \mathbb{R}^{km,ln}.$$

Sind $A, C \in \mathbb{R}^{k,k}$ und $B, D \in \mathbb{R}^{m,m}$, so gilt

$$(A \otimes B)(C \otimes D) = AC \otimes BD.$$

Daraus folgt: Sind $A \in \mathbb{R}^{k,k}$ und $B \in \mathbb{R}^{m,m}$ regulär, so gilt

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

### 2.3.2 Matrixfunktionen

Seien $A \in \mathbb{C}^{n,n}$ und $f(z)$ eine skalare Funktion. Die Matrixfunktion $f(A)$ kann dann nach Golub/van Loan [20] wie folgt erklärt werden:

**Definition 2.12** $G$ sei ein beschränktes, einfach zusammenhängendes Gebiet in der komplexen Zahlenebene, und $\Gamma$ sei eine (mathematisch positiv orientierte) geschlossene $C^1$- Jordan-Kurve in $G$. Das von $\Gamma$ umrandete Gebiet enthalte die Eigenwerte der Matrix $A$, und $f(z)$ sei in diesem Gebiet analytisch. Dann ist $f(A)$ erklärt durch

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - A)^{-1} dz.$$

□

Ist $f(A)$ definiert und $A$ ähnlich zu einer blockdiagonalen Matrix $B$,

$$A = XBX^{-1} = X\mathrm{diag}\{B_1, \ldots, B_p\}X^{-1},\ B_i \in \mathbb{C}^{n_i,n_i}, \tag{2.24a}$$

dann folgt

$$\begin{aligned} f(A) &= \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - XBX^{-1})^{-1} dz \\ &= X \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - \mathrm{diag}\{B_1, \ldots, B_p\})^{-1} dz X^{-1}, \end{aligned}$$

also ist auch $f(A)$ blockdiagonal,

$$f(A) = Xf(B)X^{-1} = X\mathrm{diag}\{f(B_1), \ldots, f(B_p)\}X^{-1}. \tag{2.24b}$$

Zur Berechnung von $f(A)$ reicht es deshalb aus, die Funktion $f$ für Jordan-Blöcke

$$B_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ 0 & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_i,n_i}$$

berechnen zu können. Mit Definition 2.12 kann man zeigen, daß

$$f(B_i) = \begin{pmatrix} f(\lambda_i) & \frac{f^{(1)}(\lambda_i)}{1!} & \cdots & \frac{f^{(n_i-1)}(\lambda_i)}{(n_i-1)!} \\ & \ddots & \ddots & \vdots \\ 0 & & \ddots & \frac{f^{(1)}(\lambda_i)}{1!} \\ & & & f(\lambda_i) \end{pmatrix} \tag{2.25}$$

gilt, vgl. [20].

### 2.3.3 Logarithmische Matrixnorm, Theorem von J. von Neumann und Maximumnorm

Von Lozinski und Dahlquist wurde, ursprünglich zur Charakterisierung der Stabilität von Differentialgleichungssystemen, die „logarithmische Matrixnorm“ wie folgt eingeführt (vgl. Hairer/Nørsett/Wanner [25], Dekker/Verwer [14]):
Sei $\|\cdot\|_{\mathbb{C}^n}$ eine Vektornorm im $\mathbb{C}^n$, $A \in \mathbb{C}^{n,n}$. Für die zugeordnete Matrixnorm ist

$$\mu[A] = \lim_{\delta \to +0} \frac{\|I + \delta A\|_{\mathbb{C}^n} - 1}{\delta}$$

die zugeordnete „logarithmische Norm“ der Matrix A.
Dieser Grenzwert existiert für alle zugeordneten Matrixnormen und für alle Matrizen, vgl. Dekker/Verwer [14]. $\mu[A]$ ist jedoch keine Norm im klassischen Sinn (das erste Normaxiom ist verletzt), sondern ein sublineares Funktional. Für die der euklidischen Vektornorm zugeordnete Matrixnorm, die Spektralnorm, gelten

$$\|A\|_2 = \sqrt{\lambda_{\max}(\bar{A}^\top A)}, \quad \mu_2[A] = \frac{1}{2}\lambda_{\max}\left(A + \bar{A}^\top\right), \tag{2.26}$$

wobei $\lambda_{\max}(A)$ der maximale Eigenwert von $A$ ist.
Es gilt nach Hairer/Wanner [26] der folgende Satz (Version des Satzes von J. von Neumann):

**Satz 2.13** Ist für eine quadratische Matrix $A$ eine rationale Funktion $r(z)$ beschränkt für $\mathfrak{Re}\, z \le \mu_2[A]$, so gilt

$$\|r(A)\|_2 \le \sup_{\mathfrak{Re}\, z \le \mu_2[A]} |r(z)|.$$

$\varphi_r(x) = \sup_{\mathfrak{Re}\, z \le x} |r(z)|$ wird Fehlerwachstumsfunktion genannt. □

Weiterhin gilt nach [14], daß $\mu_2[A]$ die kleinste Konstante ist, so daß für alle $t \ge 0$

$$\|e^{At}\|_2 \le e^{\mu_2[A]t} \tag{2.27}$$

gilt. Für normale Matrizen $A$ (d. h. $\bar{A}^\top A = A\bar{A}^\top$) gilt in (2.27) sogar Gleichheit.
Die Berechnung der Spektralnorm ist aufwendig. Sie kann aber mit der Maximumnorm leicht abgeschätzt werden. Für eine Matrix $A = (a_{ij})_{i,j=1,\dots,n} \in \mathbb{C}^{n,n}$ ist die Maximumnorm definiert durch

$$\|A\|_M = n \max_{i,j} |a_{ij}|,$$

und es gilt

$$n^{-\frac{3}{2}}\|A\|_M \le \|A\|_2 \le n^{-\frac{1}{2}}\|A\|_M. \tag{2.28}$$

# Kapitel 3

# Partielle DA-Systeme

## 3.1 Beispiele partieller DA-Systeme

Viele Modelle in Naturwissenschaft und Technik bestehen aus einer Kopplung von Gleichungen unterschiedlichen Typs, z. B. aus elliptischen, parabolischen, hyperbolischen Differentialgleichungen und algebraischen Gleichungen. Diese gemischten Systeme werden partielle differentiell-algebraische Gleichungssysteme (PDA-Systeme) genannt.
Im folgenden werden einige praxisrelevante Beispiele von PDA-Systemen angegeben. Geeignete Anfangs- und Randbedingungen werden dabei jeweils als gegeben vorausgesetzt.

**Beispiel 3.1** Modellierung einer Populationsdynamik von $n$ Spezies in Abhängigkeit von $m$ gleichmäßig verteilten Nahrungsquellen [57]:

$$\begin{aligned} \frac{\partial u_j}{\partial t} &= D\Delta u_j + f_j(t,u,v), \quad j = 1,\dots,n, \\ \frac{\partial v_i}{\partial t} &= g_i(t,u,v), \quad i = 1,\dots,m. \end{aligned}$$

Der Dichtevektor $u = (u_1 \dots, u_n)^\top$ der Spezies genügt einer Diffusionsgleichung, der Dichtevektor $v = (v_1 \dots, v_m)^\top$ der Nahrungsquellen einer gewöhnlichen Differentialgleichung. Die positive Diffusionskonstante $D$, die Quellterme $f$ und $g$ seien gegeben. □

**Beispiel 3.2** Nichtlineares Zwei-Kompartment-Modell der Pharmakokinetik in der Leber [17]:

$$\begin{aligned} \frac{\partial u_1(t,x)}{\partial t} &= \frac{D}{V^2}\frac{\partial^2 u_1(t,x)}{\partial x^2} - \frac{Q}{V}\frac{\partial u_1(t,x)}{\partial x} - k_{12}u_1(t,x) + \varepsilon k_{21}u_2(t,x), \\ \frac{\partial u_2(t,x)}{\partial t} &= \frac{1}{\varepsilon}k_{12}u_1(t,x) - k_{21}u_2(t,x) - \frac{V_{max}}{K_m + u_2(t,x)}u_2(t,x). \end{aligned}$$

$u_1$ beschreibt die Drogenkonzentration im zentralen, $u_2$ die im peripheren Kompartment. Das Volumen $V$ des zentralen Kompartments, der korrigierte Streukoeffizient $D$, die Perfusatfließgeschwindigkeit $Q$, die maximale Abbaurate $V_{max}$, die Michaelis-Menten-Konstante $K_m$, das Verhältnis $\varepsilon$ des Volumens des peripheren Kompartments zum Volumen des zentralen Kompartments und die Transferraten $k_{12}$ vom zentralen zum peripheren Kompartment und $k_{21}$ für die Gegenrichtung seien gegeben. □

**Beispiel 3.3** Pulververbrennung [30]:

$$\begin{aligned} \frac{\partial T}{\partial t} - k\Delta T &= Qw(T,Y), \\ \frac{\partial Y}{\partial t} &= -w(T,Y). \end{aligned}$$

Dabei sind $T$ die Temperatur, dividiert durch eine Referenztemperatur, $Y$ die Konzentration des verbrennenden Reaktionsmittels, $k$ die Temperaturleitfähigkeit und $Q$ ein Wärmefreisetzungsparameter, und der Reaktionsterm hat die Gestalt $w(T,Y) = K_0 Y e^{-\frac{E}{T}}$ einer Reaktion erster Ordnung mit einer dimensionslosen Aktivierungsenergie $E$ im Arrhenius-Term. □

Während in den vorangegangenen drei Beispielen parabolische und gewöhnliche Differentialgleichungen miteinander gekoppelt sind, ist das folgende ein Beispiel für ein gekoppeltes System aus parabolischen und elliptischen Differentialgleichungen:

**Beispiel 3.4** Räuber-Beute-Modell mit unendlich schneller Reaktionsrate der Räuber [5]:

$$\begin{aligned} 0 &= D_{1i}\Delta u_i + f_i(x,y,t,u,v,), \quad i = 1,\dots,n, \\ \frac{\partial v_j}{\partial t} &= D_{2j}\Delta v_j + g_j(x,y,t,u,v), \quad j = 1,\dots,m. \end{aligned}$$

Im Gegensatz zum Modell in Beispiel 3.1 ist der Dichtevektor $u$ der Räuber hier durch eine unendlich schnelle Reaktionsrate charakterisiert, und der Dichtevektor $v$ der Beute genügt einer Diffusionsgleichung. Die Quellterme $f$ und $g$ und positive Diffusionskonstanten $D_{1i}$ und $D_{2j}$ seien gegeben.
Ist eine Beutekomponente räumlich gleichmäßig verteilt und wird die entsprechende Diffusionskonstante gleich 0 gesetzt, so kann auch hier als weiterer Typ noch eine gewöhnliche Differentialgleichung hinzukommen. □

Im folgenden Beispiel sind parabolische und gewöhnliche Differential- und algebraische Gleichungen miteinander gekoppelt:

**Beispiel 3.5** Transport-Reaktionssystem in porösem Medium [59, 44, 28]:

$$\begin{aligned} \Theta\frac{\partial u}{\partial t} - \nabla(\Theta D\nabla u) + \kappa\nabla u &= f_1(u,v,z), \\ \frac{\partial v}{\partial t} &= f_2(u,v,z), \\ 0 &= g(u,v,z). \end{aligned}$$

$u = (u_1,\dots,u_{n_1})$, $v = (v_1,\dots,v_{n_2})$ und $z = (z_1,\dots,z_{n_3})$ sind die betrachteten Spezies, $\Theta > 0$ ist der volumetrische Wasserinhalt, die symmetrische Matrix $D \in \mathbb{R}^{n_1,n_1}$ bezeichnet die Diffusion, $\kappa$ die Sickergeschwindigkeit. $\frac{\partial g}{\partial z}$ wird als regulär vorausgesetzt. □

Ein Beispiel mit hyperbolischem Charakter ist das nachstehende:

**Beispiel 3.6** Supraleitende Magnetspule [41, 12]:

$$\begin{pmatrix} 0 & 0 \\ -\frac{LC}{l^2} & \frac{L}{D} \end{pmatrix} u_{tt} - u_{xx} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} u = 0, \quad x \in (0,l), \quad t > 0.$$

$u_1(t,x)$ ist die Spannung in den Windungen der Spule, $u_2(t,x)$ die Divergenz der elektrischen Feldstärke, $l$ die Länge der Spule. $L$, $C$ und $D$ sind weitere Spulenparameter. Transformation auf ein partielles differentiell-algebraisches Gleichungssystem erster Ordnung in $t$ liefert

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{LC}{l^2} & \frac{L}{D} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} u_t - \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} u_{xx} + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} u = 0. \tag{3.1}$$

□

Weitere Beispiele dieser Art findet man unter anderem in der Verfahrenstechnik (vgl. Altenbach/Deuring/Naumenko [1]), Plasmaphysik und Knochenmodellierung (vgl. Lucht/Debrabant [34]). Auch die Navier-Stokes-Gleichungen können als PDA-System aufgefaßt werden (vgl. Lin [32], Lucht/Debrabant [34], Marszalek [40], Rabier/Rheinboldt [44]).
Weitere Typen von PDA-Systemen sind über Kontaktbedingungen gekoppelte Systeme von partiellen Differentialgleichungen und differentiell-algebraischen Gleichungen, die zum Beispiel bei der Modellierung der Wechselwirkung von Kettenwerk und Stromabnehmer auftreten, vgl. Simeon und Arnold [48], und Systeme partieller Differentialgleichungen oder partieller differentiell-algebraischer Gleichungen, deren (zeitabhängige) Randwerte durch ein DA-System gegeben sind, vgl. Günther/Wagner [23].
In den letzten Jahren hat man sich, beginnend mit den Arbeiten von Campbell und Marszalek ([11, 40]), verstärkt der numerischen Behandlung partieller differentiell-algebraischer Systeme zugewandt.

## 3.2 Aufgabenstellung

In dieser Arbeit werden räumlich $d$-dimensionale, semilineare PDA-Systeme der Form

$$A\,u_t(t,\vec{x}) + \sum_{i=1}^{d} B_i\ (u_{x_i x_i}(t,\vec{x}) + r_i u_{x_i}(t,\vec{x})) + C\,u(t,\vec{x}) = f(t,\vec{x},u) \tag{3.2a}$$

betrachtet, wobei

$$\begin{gathered} t \in (t_0,t_e), \quad \vec{x} \in \Omega{=}\Omega_1 \times \ldots \times \Omega_d, \quad \Omega_i \subset \mathbb{R}, \\ A, B_i, C \in \mathbb{R}^{n,n}, \quad r_i \in \mathbb{R}, \quad i = 1,\ldots,d, \\ u : [t_0,t_e] \times \overline{\Omega} \to \mathbb{R}^n, \quad f : [t_0,t_e] \times \overline{\Omega} \times \mathbb{R}^n \to \mathbb{R}^n \end{gathered} \tag{3.2b}$$

und $u$ hinreichend glatt sei. Die Matrizen $A$, $B_i$ und $C$ dürfen hier auch singulär sein. Die Beispiele 3.1 bis 3.6 sind sämtlich von diesem Typ, wobei in Beispiel 3.5 die Parameter $\Theta$, $D$ und $\kappa$ als konstant vorausgesetzt werden. Anfangs- bzw. Randbedingungen können im Unterschied zu parabolischen Anfangsrandwertaufgaben mit regulären Matrizen $A$ und $B_i$ nicht für alle Komponenten des Lösungsvektors vorgegeben werden, wie die folgenden drei Beispiele zeigen:

**Beispiel 3.7** (vgl. Lucht/Strehmel/Eichler-Liebenow [37])

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} u_t + \begin{pmatrix} -1 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} u_{xx} + \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} u = \begin{pmatrix} f_1(t,x) \\ f_2(t,x) \\ f_3(t,x) \end{pmatrix}.$$

Es gilt $u_2 = f_3$, $u_3 = f_2 - f_{3t} + f_{3xx}$. Die Anfangs- und Randbedingungen von $u_2$ und $u_3$ sind also durch die rechte Seite $f(t,x)$ und ihre Zeit- und Ortsableitungen eindeutig bestimmt. Man sieht, daß an die Komponenten von $f$ unterschiedliche Glattheitsforderungen zu stellen sind. □

**Beispiel 3.8** (vgl. [37])

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{A} u_t - \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} u_{xx} - \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} u = \begin{pmatrix} f_1(t,x) \\ f_2(t,x) \end{pmatrix}, \quad x \in (-l,l), \quad t \in (t_0,t_e).$$

Da $A$ regulär ist, können für alle Komponenten von $u$ Anfangsbedingungen vorgeschrieben werden. Aus der zweiten Gleichung erhält man für die Randwerte von $u_2$ die gewöhnliche Differentialgleichung

$$\frac{d}{dt} u_2(t,\pm l) = f_2(t,\pm l) + u_1(t,\pm l).$$

Nach Vorgabe der Randwerte von $u_1$ kann man diejenigen von $u_2$ aus dieser gewöhnlichen Differentialgleichung mit dem Anfangswert $u_2(t_0, \pm l)$ berechnen. Man kann aber auch die Randwerte von $u_2$ vorschreiben und diejenigen von $u_1$ dann aus dieser Gleichung durch Differentiation von $u_2(t, \pm l)$ erhalten. □

**Beispiel 3.9** Betrachtet werde das System aus Beispiel 3.6 für hinreichend glatte Lösungen $u$. Als Anfangsbedingungen werden

$$u_1(0,x) = \left(\frac{E}{l} - \frac{CDEl}{6}\right) x + \frac{CDE}{6l} x^3, \quad u_3(0,x) = 0$$

und als Randbedingungen

$$u_1(t,0) = u_2(t,0) = 0, \quad u_1(t,l) = E, \quad u_2(t,l) = CDE$$

gewählt. Dabei ist $E$ die angelegte Quellspannung.
Da die Randwerte von $u_1$ und $u_2$ jeweils konstant sind, folgt aus der dritten und vierten Gleichung von (3.1), daß $u_3$ und $u_4$ homogene Randbedingungen erfüllen. Aus der Anfangsbedingung für $u_1$ und der ersten Gleichung von (3.1) ergibt sich $u_2(0,x) = \frac{CDE}{l} x$. Aus $u_3(0,x) = 0$ folgt mit der dritten Gleichung $u_{1t}(0,x) = 0$ und daraus $u_{1xxt}(0,x) = 0$. Mit der ersten Gleichung erhält man daraus $u_{2t}(0,x) = 0$ und mit der vierten Gleichung schließlich $u_4(0,x) = 0$. □

Im folgenden werden Dirichlet-, Neumann- und periodische Randbedingungen betrachtet: Die Indexmengen $M_D$, $M_N$ und $M_P$ seien paarweise disjunkt und

$$M_D \cup M_N \cup M_P = \{1, \ldots, d\}.$$

Für $i \in M_D \cup M_N$ seien mit $l_i \in \mathbb{R},\ l_i > 0$

$$\Omega_i = (-l_i, l_i) \quad \text{und} \quad \partial_i\Omega = \{\vec{x} \in \partial\Omega :\ x_i = \pm l_i\}.$$

Für $i \in M_D$ seien auf $\partial_i\Omega$ Dirichlet- und für $i \in M_N$ Neumann-Randwerte vorgegeben.
Für $i \in M_P$ seien periodische Randbedingungen mit $\Omega_i = \mathbb{R}$ und

$$u(t,\vec{x}) = u(t,\vec{x} + 2l_i\vec{e}_i), \quad \vec{x} \in \Omega, \quad t \in [t_0, t_e] \tag{3.2c}$$

gegeben.
Im folgenden wird vorausgesetzt, daß alle Anfangswerte

$$u(t_0,\vec{x}) = \varphi(\vec{x}), \quad \vec{x} \in \overline{\Omega} \tag{3.2d}$$

und alle in die Ortsdiskretisierung eingehenden Randwerte der exakten Lösung für die numerische Berechnung bekannt sind, d. h.

$$B_i u(t,\vec{x}) = \psi_i(t,\vec{x}), \quad \vec{x} \in \partial_i\Omega, \quad t \in [t_0, t_e], \quad i \in M_D, \tag{3.2e}$$

$$\frac{\partial}{\partial x_i} B_i u(t,\vec{x}) = \chi_i(t,\vec{x}), \quad \vec{x} \in \partial_i\Omega, \quad t \in [t_0, t_e], \quad i \in M_N. \tag{3.2f}$$

**Definition 3.10** Eine Funktion

$$u(t,x) :\ [t_0, t_e] \times \overline{\Omega} \to \mathbb{R}^n$$

ist Lösung der Anfangsrandwertaufgabe (3.2) (im klassischen Sinn), wenn $u$ genügend glatt ist, die Gleichung (3.2a) punktweise erfüllt und eindeutig durch die Anfangs- und Randbedingungen bestimmt ist, vgl. Lucht/Strehmel [36]. □

**Bemerkung 3.11** Im Fall periodischer Randbedingungen folgt aus (3.2a) und (3.2d) auch die entsprechende Periodizität der rechten Seite $f$ und des Anfangswertes $\varphi$:

$$f(t,\vec{x}) = f(t,\vec{x} + 2l_i\vec{e}_i), \quad u(t_0,\vec{x}) = u(t_0,\vec{x} + 2l_i\vec{e}_i), \quad \vec{x} \in \Omega, \quad t \in [t_0, t_e].$$ □

**Bemerkung 3.12** Eine zweite Aufgabenklasse von PDA-Systemen mit variablen Koeffizienten wird in Anhang B betrachtet. □

# Kapitel 4

# Semidiskretisierung der PDA-Systeme

Eine Möglichkeit zur numerischen Behandlung von Anfangsrandwertproblemen partieller Differentialgleichungen ist die Linienmethode (method of lines, MOL), vgl. Strehmel/Weiner [53], Großmann/Roos [22]. Die Lösung erfolgt in zwei Teilschritten. Zuerst wird die Differentialgleichung semidiskretisiert. Bei der horizontalen Linienmethode (Rothe-Methode) erfolgt die Semidiskretisierung bezüglich der Zeitvariablen, das Anfangsrandwertproblem wird durch eine Folge von Randwertaufgaben approximiert, die dann mit geeigneten Verfahren gelöst werden müssen. Bei der vertikalen Linienmethode dagegen erfolgt die Semidiskretisierung bezüglich der räumlichen Variablen, zum Beispiel durch finite Differenzen oder die Methode der finiten Elemente. Parabolische Anfangsrandwertprobleme werden dadurch in ein Anfangswertproblem für ein System gewöhnlicher Differentialgleichungen erster Ordnung überführt, welches im allgemeinen steif ist, die Steifheit ist von der Feinheit der Ortsdiskretisierung abhängig. Dieses semidiskrete Problem wird anschließend durch geeignete (implizite) Diskretisierungsmethoden für Anfangswertprobleme gewöhnlicher Differentialgleichungen gelöst.
Die vertikale Linienmethode wird in Abschnitt 4.1 auf die hier betrachteten semilinearen PDA-Systeme angewendet. Die Diskretisierung bezüglich des Ortes wird dabei mittels finiter Differenzen durchgeführt. Daran anschließend wird in Abschnitt 4.2 die Konvergenz der Ortsdiskretisierung linearer PDA-Systeme sowohl auf der Grundlage der Lösungsdarstellung der linearen Fehlergleichung mittels Drazin-Inverser als auch mit einer Weierstraß-Kronecker-Transformation gezeigt. Eine lokale Konvergenzaussage für semilineare PDA-Systeme vom Zeitindex 1 findet man in Lucht/Strehmel [36].
Die Diskretisierung des entstandenen Anfangswertproblems erfolgt dann in Kapitel 5 durch implizite Runge-Kutta-Verfahren.

## 4.1 Finitisierung des Ortsraumes und Diagonalisierung des diskretisierten Ortsdifferentialoperators

Im folgenden werden für das PDA-System (3.2) Ortsdiskretisierungen für Dirichlet-, für periodische und für Neumann-Randbedingungen angegeben. Dabei werden Abschätzungen für die Eigenwerte und die Matrix der Eigenvektoren der diskretisierten Ortsdifferentialoperatoren aufgeführt, die für die nachfolgenden Konvergenzuntersuchungen benötigt werden.

### 4.1.1 Räumlich eindimensionales PDA-System

#### 4.1.1.1 Dirichlet-Randbedingungen

Um eine übersichtlichere Darstellung zu erhalten, wird zunächst der räumlich eindimensionale Fall ($d = 1$) betrachtet, also anstelle von (3.2a)

$$Au_t(t,x) + B\left(u_{xx}(t,x) + ru_x(t,x)\right) + Cu(t,x) = f(t,x,u), \quad x \in \Omega = (-l,l). \tag{4.1}$$

Es seien $N \in \mathbb{N}^+$ und

$$h = \frac{2l}{N+1}. \tag{4.2}$$

Dann kann $\Omega$ durch das äquidistante Ortsgitter

$$\Omega_h = \left\{x_k : \; x_k = -l + kh, \; k = 1, \ldots, N\right\} \tag{4.3}$$

diskretisiert werden. Mit den 3-Punkt-Differenzenapproximationen

$$u_{xx}(t,x_k) \approx \frac{1}{h^2}\left(u(t,x_{k+1}) - 2u(t,x_k) + u(t,x_{k-1})\right), \quad k = 1, \ldots, N,$$

für die zweite Ortsableitung und den Approximationen

$$\begin{aligned} u_x(t,x_k) \approx \frac{1}{h}\left(\delta u(t,x_{k+1}) + (1-2\delta)u(t,x_k) + (\delta-1)u(t,x_{k-1})\right), \\ k = 1, \ldots, N, \quad \delta \in [0,1], \end{aligned} \tag{4.4}$$

für die erste Ortsableitung ($x_0 = -l$, $x_{N+1} = l$) ergibt sich aus Gleichung (4.1) für jeden Gitterpunkt $x_k$ die semidiskrete Gleichung

$$\begin{aligned} A\,\dot{u}_k(t) + \frac{1}{h^2}B\Big(&\left(1 + hr\delta\right)u_{k+1}(t) - \left(2 - hr(1-2\delta)\right)u_k(t) \\ &+ \left(1 + hr(\delta-1)\right)u_{k-1}(t)\Big) + C\,u_k(t) = f_k(t,u_k), \end{aligned} \tag{4.5}$$

wobei $u_k(t)$ Näherung der Funktion $u(t,x)$ in den einzelnen Gitterpunkten $x_k$, $k = 1, \ldots, N$, und $f_k(t,u_k) = f(t,x_k,u_k)$ sind und für $Bu_0$ und $Bu_{N+1}$ die vorgegebenen Randwerte eingesetzt werden.

**Bemerkung 4.1** Für $\delta = 0$ liefert die Approximation (4.4) den rückwärtsgenommenen, für $\delta = \frac{1}{2}$ den zentralen und für $\delta = 1$ den vorwärtsgenommenen Differenzenquotienten. □

Durch Taylor-Entwicklung bis zur $m$-ten Ordnung im Gitterpunkt $x_k$ erhält man

$$\begin{aligned} &\frac{1}{h^2}\Big(\left(1 + hr\delta\right)u(t,x_{k+1}) - \left(2 - hr(1-2\delta)\right)u(t,x_k) + \left(1 + hr(\delta-1)\right)u(t,x_{k-1})\Big) \\ &= \frac{1}{h^2}\sum_{i=1}^{m}\frac{h^i}{i!}\frac{\partial^i}{\partial x^i}u(t,x_k)\left(1 + hr\delta + (-1)^i(1 + hr(\delta-1))\right) \\ &\quad + \frac{h^{m-1}}{(m+1)!}\left((1 + hr\delta)\frac{\partial^{m+1}}{\partial x^{m+1}}u(t,\zeta_{1k}(t)) + (-1)^{(m+1)}(1 + hr(\delta-1))\frac{\partial^{m+1}}{\partial x^{m+1}}u(t,\zeta_{2k}(t))\right) \\ &= ru_x(t,x_k) + u_{xx}(t,x_k) + h^{p_1}\gamma_k(t), \end{aligned} \tag{4.6}$$

dabei ist

$$p_1 = \begin{cases} 2\text{: } r = 0 \text{ (kein konvektiver Term) oder } \delta = \frac{1}{2} \text{ (zentraler Differenzenquotient)} \\ 1\text{: sonst} \end{cases} \tag{4.7}$$

die Approximationsordnung der verwendeten Diskretisierung, und $\gamma_k$ ist für $p_1 = 1$ ($m = 2$) durch

$$\begin{aligned} \gamma_k(t) = \frac{r(2\delta-1)}{2}\frac{\partial^2}{\partial x^2}u(t,x_k) + \frac{1}{6}\Big\{&(1 + hr\delta)\frac{\partial^3}{\partial x^3}u(t,\zeta_{1k}(t)) \\ &- \left(1 + hr(\delta-1)\right)\frac{\partial^3}{\partial x^3}u(t,\zeta_{2k}(t))\Big\} \end{aligned} \tag{4.8a}$$

und für $p_1 = 2$ ($m = 3$) durch

$$\gamma_k(t) = \frac{r}{6}\frac{\partial^3}{\partial x^3}u(t,x_k) + \frac{1}{24}\left\{(1+hr\delta)\frac{\partial^4}{\partial x^4}u(t,\zeta_{1k}(t)) + (1+hr(\delta-1))\,\frac{\partial^4}{\partial x^4}u(t,\zeta_{2k}(t))\right\} \tag{4.8b}$$

definiert mit den Zwischenwerten

$$\zeta_{1k}(t) \in (x_k, x_{k+1}), \quad \zeta_{2k}(t) \in (x_{k-1}, x_k), \quad k = 1,..,N. \tag{4.9}$$

Existiert

$$K = \max_{(t,x)\in[0,t_e]\times\overline{\Omega}} \left\{ \begin{array}{lcl} \frac{|r|}{2}\|\frac{\partial^2}{\partial x^2}u(t,x)\| + \frac{1+l|r|}{3}\|\frac{\partial^3}{\partial x^3}u(t,x)\| & : & p_1 = 1 \\ \frac{|r|}{6}\|\frac{\partial^3}{\partial x^3}u(t,x)\| + \frac{1+l|r|}{12}\|\frac{\partial^4}{\partial x^4}u(t,x)\| & : & p_1 = 2 \end{array} \right\} < \infty, \tag{4.10}$$

so gilt

$$\|\gamma_k(t)\|_\infty \le K, \quad k = 1,\ldots,N. \tag{4.11}$$

Aus dem System semidiskreter Gleichungen (4.5) erhält man mit dem in Abschnitt 2.3.1 eingeführten Kronecker-Produkt die kompakte Schreibweise

$$(I_N \otimes A)\,\dot{U}(t) + \left(\frac{1}{h^2}P \otimes B + I_N \otimes C\right)\,U(t) = F(t,U) - \omega(t) =: \tilde{F}(t,U). \tag{4.12}$$

Dabei ist $I_N$ die $N$-dimensionale Einheitsmatrix,

$$U(t) = \left(u_1^\top(t),\ldots,u_N^\top(t)\right)^\top \in \mathbb{R}^{Nn}$$

und

$$F(t,U) = \left(f_1^\top(t,u_1),\ldots,f_N^\top(t,u_N)\right)^\top$$

enthalten die Vektoren $u_k$ und $f_k$ in den Gitterpunkten,

$$\omega(t) = \left(\frac{1+hr(\delta-1)}{h^2}\psi^\top(t,-l),0,\ldots,0,\frac{1+hr\delta}{h^2}\psi^\top(t,l)\right)^\top \tag{4.13}$$

enthält die vorgegebenen Randwerte. Der zum Differentialoperator $\left(\frac{\partial^2}{\partial x^2} + r\frac{\partial}{\partial x}\right)$ zugehörige diskrete Operator $\frac{1}{h^2}P$ ist durch

$$P = \begin{pmatrix} -(2-hr(1-2\delta)) & 1+hr\delta & & \\ 1+hr(\delta-1) & -(2-hr(1-2\delta)) & 1+hr\delta & \\ & & \cdots & \\ & & 1+hr(\delta-1) & -(2-hr(1-2\delta)) \end{pmatrix} \tag{4.14}$$

mit $P \in \mathbb{R}^{N,N}$ gegeben.
Ferner sei

$$U(t_0) \in \mathbb{R}^{nN} \tag{4.15}$$

ein (konsistenter) Anfangsvektor (vgl. Abschnitt 2.1).
Für singuläres $A$ ist (4.12) mit (4.15) ein Anfangswertproblem eines DA-Systems.
Da im Abschnitt 5.2 für die Untersuchung der Konvergenz der Gesamtdiskretisierung die Eigenwerte und die Norm einer diagonalisierenden Matrix von $\frac{1}{h^2}P$ benötigt werden, sollen

diese nun angegeben bzw. abgeschätzt werden.
Nach (A.15) ergeben sich die Eigenwerte von $\frac{1}{h^2}P$ zu

$$\begin{aligned}\lambda_j &= -\frac{2-hr(1-2\delta)}{h^2} + 2\frac{1+hr\delta}{h^2}\sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}\cos\frac{j\pi}{N+1} \\ &= \frac{r}{h} - 2\frac{1+hr\delta}{h^2}\left(1-\sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}\right) - 4\frac{1+hr\delta}{h^2}\sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}\sin^2\frac{j\pi}{2(N+1)}, \\ &\quad j=1,\dots,N. \end{aligned} \tag{4.16}$$

Sie sind sämtlich voneinander verschieden.
Nach (A.13) ist $S_P = (v_{jk})^\top_{j,k=1,\dots,N}$ mit

$$v_{jk} = c_j d_h^k \sin\frac{jk\pi}{N+1}, \quad d_h = \sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}, \tag{4.17}$$

die zugehörige Matrix der Eigenvektoren. Dann gilt

$$\frac{1}{h^2}PS_P = S_P\mathrm{diag}\{\lambda_1,\dots,\lambda_N\}.$$

$S_P$ läßt sich schreiben als

$$S_P = D_S Q_S$$

mit

$$D_S = \mathrm{diag}\{d_h, d_h^2, \dots, d_h^N\}$$

und

$$Q_S = (q_{jk})_{j,k=1,\dots,N}, \quad q_{jk} = c_k \sin\frac{jk\pi}{N+1}. \tag{4.18}$$

$Q_S$ entspricht der Matrix der Eigenvektoren von $\frac{1}{h^2}P$ für den Fall, daß $d_h = 1$ gilt, d. h. $r = 0$ (kein konvektiver Term). In diesem Fall ist $P$ symmetrisch, und da die Eigenwerte sämtlich voneinander verschieden sind, sind die Eigenvektoren untereinander orthogonal. Durch Wahl von

$$\begin{aligned}\frac{1}{c_k^2} &= \sum_{j=1}^N \sin^2\frac{jk\pi}{N+1} = \sum_{j=0}^N\left(\frac{1}{2}-\frac{1}{2}\cos\frac{2jk\pi}{N+1}\right) = \frac{N+1}{2} - \frac{1}{2}\mathfrak{Re}\sum_{j=0}^N e^{\frac{2jk\pi}{N+1}i} \\ &= \frac{N+1}{2} - \frac{1}{2}\mathfrak{Re}\frac{e^{2k\pi i}-1}{e^{\frac{2k\pi}{N+1}i}-1} = \frac{N+1}{2}, \quad k=1,\dots,N,\end{aligned}$$

in (4.18) kann man sie bezüglich der euklidischen Norm normieren und erhält in diesem Fall, daß $Q_S$ eine (symmetrische) Orthogonalmatrix ist,

$$Q_S = Q_S^\top = Q_S^{-1}.$$

Für genügend kleine $h$ ($|r|h < 1$) ist $d_h > 0$ und also $S_P$ regulär, es gilt

$$(S_P^{-1})_{jk} = \sqrt{\frac{2}{N+1}}\frac{1}{d_h^k}\sin\frac{jk\pi}{N+1}. \tag{4.19}$$

Die euklidische Matrixnorm $\|\cdot\|$ ist invariant unter orthogonalen Transformationen, so daß

$$\|S_P\| = \|D_S Q_S\| = \|D_S\| \quad \text{und} \quad \|S_P^{-1}\| = \|Q_S^{-1}D_S^{-1}\| = \|D_S^{-1}\| \tag{4.20}$$

folgen. Da $D_S$ eine Diagonalmatrix ist, gelten

$$\|D_S\| = \max\left\{d_h, d_h^N\right\} \quad \text{und} \quad \|D_S^{-1}\| = \max\left\{\frac{1}{d_h}, \frac{1}{d_h^N}\right\}. \tag{4.21}$$

Aus

$$\lim_{h\to 0} \frac{hr}{1+hr\delta} = 0$$

folgt (vgl. Mangoldt/Knopp [39])

$$\lim_{h\to 0} \left(1 - \frac{hr}{1+hr\delta}\right)^{-\frac{1+hr\delta}{hr}} = e$$

und daraus

$$\lim_{h\to 0} \left(1 - \frac{hr}{1+hr\delta}\right)^{\frac{2l(1+hr\delta)}{h}} = e^{-2lr}.$$

Da

$$\lim_{h\to 0} \left(1 - \frac{hr}{1+hr\delta}\right)^{2lr\delta+1} = 1$$

gilt, erhält man

$$\lim_{h\to 0} \left(1 - \frac{hr}{1+hr\delta}\right)^{\frac{2l}{h}-1} = e^{-2lr}$$

und schließlich

$$\lim_{N\to\infty} d_h^N = e^{-lr}.$$

Für genügend kleine $h$ folgt damit für $r \geq 0$

$$\frac{1}{2}e^{-lr} \leq d_h^N \leq d_h \leq 1 \tag{4.22a}$$

und für $r < 0$

$$1 < d_h < d_h^N \leq \frac{3}{2}e^{-lr} \tag{4.22b}$$

und damit

$$\frac{1}{2}e^{-l|r|} \leq \max\{d_h, d_h^N\} \leq \frac{3}{2}e^{l|r|}.$$

Aus (4.20) und (4.21) folgen für genügend kleine $h$

$$\frac{1}{2}e^{-l|r|} \leq \|S_P\| \leq \frac{3}{2}e^{l|r|}, \quad \frac{2}{3}e^{-l|r|} \leq \|S_P^{-1}\| \leq 2e^{l|r|}. \tag{4.23}$$

Im folgenden wird eine Ortsdiskretisierung für periodische Randbedingungen betrachtet.

#### 4.1.1.2 Periodische Randbedingungen

Wird anstelle einer Dirichlet-Randbedingung eine periodische Randbedingung

$$u(t,x) = u(t,x+2l), \quad x \in \mathbb{R}, \quad t \in [t_0, t_e] \tag{4.24}$$

vorgegeben, so ist eine Diskretisierung zum Beispiel mit $N \in \mathbb{N}^+$ und

$$h = \frac{2l}{N} \tag{4.25}$$

auf dem Ortsgitter $\left\{x_k : \ x_k = -l + kh, \ k \in \mathbb{Z}\right\}$ möglich. Da aus (4.24) und (4.25)

$$u(t,x_k) = u(t,x_{k+N}) \quad \text{für alle} \quad k \in \mathbb{Z}$$

folgt, reicht es aus, die Diskretisierung auf dem Ortsgitter (4.3) mit $h = \frac{2l}{N}$ durchzuführen und die erhaltene Näherungslösung periodisch fortzusetzen. In diesem Fall bleiben die Gleichungen (4.5) - (4.12) mit $\omega(t) \equiv 0$ (d. h. $F(t,U) \equiv \tilde{F}(t,U)$) gültig, und (4.14) wird durch

$$P = \begin{pmatrix} -(2-hr(1-2\delta)) & 1+hr\delta & & 1+hr(\delta-1) \\ 1+hr(\delta-1) & -(2-hr(1-2\delta)) & 1+hr\delta & \\ & & \cdots & \\ 1+hr\delta & & 1+hr(\delta-1) & -(2-hr(1-2\delta)) \end{pmatrix} \tag{4.26}$$

ersetzt.
Aus Lemma A.1 folgt für die Eigenwerte von $\frac{1}{h^2}P$

$$\lambda_j = -\frac{2+hr(2\delta-1)}{h^2}\left(1-\cos\frac{2j\pi}{N}\right) + i\frac{r}{h}\sin\frac{2j\pi}{N}, \quad j = 1,\ldots,N, \tag{4.27}$$

also für genügend kleine $h$ ($h < -\frac{2}{r}$ für $r < 0$)

$$\mathfrak{Re}\lambda_j \leq 0.$$

Die Matrix der Eigenvektoren $S_P = (v_{jk})_{j,k=1,\ldots,N}$ mit

$$v_{jk} = \frac{1}{\sqrt{N}} e^{\frac{2jk\pi}{N}i} \tag{4.28}$$

ist nach Lemma A.1 eine Orthogonalmatrix, und damit gilt

$$\|S_P\| = \|S_P^{-1}\| = 1. \tag{4.29}$$

#### 4.1.1.3 Neumann-Randbedingungen

Die Neumann-Randbedingungen (3.2f) sollen mittels

$$\begin{aligned} \chi(t,-l) &= Bu_x(t,-l) \approx B\tfrac{u(t,x_1)-u(t,x_0)}{h}, \\ \chi(t,l) &= Bu_x(t,l) \approx B\tfrac{u(t,x_{N+1})-u(t,x_N)}{h} \end{aligned} \tag{4.30}$$

approximiert werden.
Entsprechend Thomas [55] wäre die sich daraus ergebende Ortsdiskretisierung auf dem Gitter (4.3) mit $x_0 = -l$, $x_{N+1} = l$ nicht konsistent (siehe Definition 4.7). Wie dort soll für Neumann-Randbedingungen deshalb ein sogenanntes Offsetgitter verwendet werden,

$$\Omega_h = \left\{x_k : \; x_k = -l - \frac{h}{2} + kh, \; k = 1,\ldots,N\right\}, \tag{4.31}$$

mit

$$h = \frac{2l}{N} \tag{4.32}$$

und $x_0 = -l - \frac{h}{2}$, $x_{N+1} = l + \frac{h}{2}$. Dabei wird vorausgesetzt, daß die exakte Lösung $u$ auf dem erweiterten Ortsintervall

$$\Omega' = (-l-h_0, l+h_0) \tag{4.33}$$

mit einem $h_0 > 0$ existiert und hinreichend glatt ist.
Es gelten weiterhin die semidiskreten Gleichungen (4.5), wobei entsprechend (4.30)

$$Bu_0(t) = Bu_1(t) - h\chi(t,-l), \tag{4.34a}$$

$$Bu_{N+1}(t) = Bu_N(t) + h\chi(t,l) \tag{4.34b}$$

gesetzt werden.
Für $k = 2, \ldots, N-1$ gilt wieder die Taylor-Entwicklung (4.6). Für $k = 1$ erhält man dagegen für die Approximation (4.34a) mit den Zwischenwerten $\zeta_{11} \in (x_1, x_2)$, $\zeta_{21} \in (-l, x_1)$

$$\begin{aligned}
&\frac{1}{h^2}\Big( (1+hr\delta)\, Bu(t,x_2) - (2-hr(1-2\delta))\, Bu(t,x_1) \\
&+ (1+hr(\delta-1))\,(Bu(t,x_1) - h\chi(t,-l))\Big) \\
&= \frac{1+hr\delta}{h^2} B\left( hu_x(t,x_1) + \frac{h^2}{2}u_{xx}(t,x_1) + \frac{h^3}{6}u_{xxx}(t,\zeta_{11}(t))\right) \\
&\quad - \frac{1+hr\delta-hr}{h} B\left( u_x(t,x_1) - \frac{h}{2}u_{xx}(t,x_1) + \frac{h^2}{8}u_{xxx}(t,\zeta_{21}(t))\right) \\
&= B\left( ru_x(t,x_1) + u_{xx}(t,x_1) + h^{p_1}\gamma_1(t)\right) \qquad (4.35)
\end{aligned}$$

mit $p_1 = 1$ und

$$\gamma_1(t) = r\left(\delta - \frac{1}{2}\right) u_{xx}(t,x_1) + \frac{1+hr\delta}{6}u_{xxx}(t,\zeta_{11}(t)) - \frac{1+hr(\delta-1)}{8}u_{xxx}(t,\zeta_{21}(t)), \quad (4.36)$$

für $k = N$ gilt analog

$$\begin{aligned}
&\frac{1}{h^2}\Big( (1+hr\delta)\,(Bu(t,x_N) + h\chi(t,l)) - (2-hr(1-2\delta))\, Bu(t,x_N) \\
&+ (1+hr(\delta-1))\, Bu(t,x_{N-1})\Big) \\
&= B\left( ru_x(t,x_N) + u_{xx}(t,x_N) + h^{p_1}\gamma_N(t)\right) \qquad (4.37)
\end{aligned}$$

mit $p_1 = 1$ und

$$\gamma_N(t) = r\left(\delta - \frac{1}{2}\right) u_{xx}(t,x_N) + \frac{1+hr\delta}{8}u_{xxx}(t,\zeta_{1N}(t)) - \frac{1+hr(\delta-1)}{6}u_{xxx}(t,\zeta_{2N}(t)) \quad (4.38)$$

mit $\zeta_{1N} \in (x_N, l)$, $\zeta_{2N} \in (x_{N-1}, x_N)$.
Es folgt deshalb aus (4.10) wieder (4.11), und mit

$$\omega(t) = \left( -\frac{1+hr(\delta-1)}{h}\chi^\top(t,-l), 0, \ldots, 0, \frac{1+hr\delta}{h}\chi^\top(t,l)\right)^\top$$

anstelle (4.13) und

$$P = \begin{pmatrix} -(1+hr\delta) & 1+hr\delta & & \\ 1+hr(\delta-1) & -(2-hr(1-2\delta)) & 1+hr\delta & \\ & & \cdots & \\ & & 1+hr(\delta-1) & -(1+hr(\delta-1)) \end{pmatrix} \quad (4.39)$$

anstelle (4.14) erhält man wieder das DA-System (4.12).
Für die Eigenwerte von $\frac{1}{h^2}P$ folgt aus (A.19)

$$\begin{aligned}
\lambda_j &= -\frac{2-hr(1-2\delta)}{h^2} + 2\frac{1+hr\delta}{h^2}\sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}\cos\frac{j\pi}{N} \\
&= \frac{r}{h} - 2\frac{1+hr\delta}{h^2}\left(1 - \sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}\right) - 4\frac{1+hr\delta}{h^2}\sqrt{\frac{1+hr(\delta-1)}{1+hr\delta}}\sin^2\frac{j\pi}{2N}, \qquad (4.40a) \\
&\quad j = 1, \ldots, N-1, \\
\lambda_N &= 0. \qquad (4.40b)
\end{aligned}$$

Die zugehörige Matrix der Eigenvektoren ist nach (A.18) durch $S_P = (v_{jk})_{j,k=1,\ldots,N}^{\top}$ mit

$$v_{jk} = c'_j d_h^k \left( \sin \frac{jk\pi}{N} - d_h \sin \frac{j(k-1)\pi}{N} \right), \quad j = 1, \ldots, N-1, \tag{4.41a}$$

$$v_{Nk} = c_N \tag{4.41b}$$

für $k = 1, \ldots, N$ und $d_h$ aus (4.17) gegeben.
Auch für die Matrizen $S_P$ und $S_P^{-1}$ kann man wieder zeigen, daß ihre euklidischen Matrixnormen für genügend kleine $h$ nach oben und gegen Null beschränkt sind. Zur einfacheren Darstellung soll dies hier nur für den Fall $r = 0$, das heißt $d_h = 1$, gezeigt werden:
In diesem Fall gilt für die Eigenvektoren von $\frac{1}{h^2} P$

$$\begin{aligned} v_{jk} &= c'_j \left( \sin \frac{jk\pi}{N} - \sin \frac{j(k-1)\pi}{N} \right) = 2c'_j \sin \frac{j\pi}{2N} \cos \frac{j(2k-1)\pi}{2N} \\ &= c_j \cos \frac{j(2k-1)\pi}{2N}, \quad j = 1, \ldots, N-1, \quad k = 1, \ldots, N, \end{aligned} \tag{4.42a}$$

$$v_{Nk} = c_N, \quad k = 1, \ldots, N, \tag{4.42b}$$

und $P$ ist symmetrisch. Da die Eigenwerte sämtlich voneinander verschieden sind, sind die Eigenvektoren orthogonal. Durch Wahl von

$$\begin{aligned} \frac{1}{c_j^2} &= \sum_{k=1}^{N} \cos^2 \frac{j(2k-1)\pi}{2N} = \frac{N}{2} + \frac{1}{2} \sum_{k=1}^{N} \cos \frac{j(2k-1)\pi}{N} \\ &= \frac{N}{2} + \frac{1}{2} \mathfrak{Re}\, e^{\frac{j\pi i}{N}} \sum_{k=0}^{N-1} e^{\frac{2jk\pi i}{N}} = \frac{N}{2}, \quad j = 1, \ldots, N-1, \\ \frac{1}{c_N^2} &= N \end{aligned}$$

kann man sie orthonormieren, und es gilt mit dieser Wahl der Konstanten wieder

$$\|S_P\| = \|S_P^{-1}\| = 1. \tag{4.43}$$

### 4.1.2 Verallgemeinerung auf räumlich mehrdimensionales PDA-System

Sei nun $d$ beliebig. Dann kann das in (3.2b) gegebene Gebiet $\Omega$ diskretisiert werden durch

$$\Omega_{\vec{h}} = \Omega_{h_1} \times \ldots \times \Omega_{h_d}, \tag{4.44}$$

wobei $\Omega_{h_i}$ für Dirichlet- und periodische Randbedingungen (d. h. $i \in M_D \cup M_P$) gemäß (4.3) und für Neumann-Randbedingungen ($i \in M_N$) gemäß (4.31) gewählt wird. Dabei ist der Vektor der Ortsschrittweiten $\vec{h} = (h_1, \ldots, h_d)$ gegeben durch $h_i = \frac{2l_i}{N_i+1}$ für Dirichlet-Randbedingungen und $h_i = \frac{2l_i}{N_i}$ für periodische und Neumann-Randbedingungen, $N_i \in \mathbb{N}^+$, $i = 1, \ldots, d$. Sei $v(t, \vec{x}) \in L_2(\Omega, \mathbb{R}^n)$ für $t \in [t_0, t_e]$. Auf dem Ortsgitter $\Omega_{\vec{h}}$ aus (4.44) mit dem Gitterparameter $\vec{h}$ erhält man die zugeordnete Gitterfunktion $v_{\vec{h}}(t)$ aus dem Raum $L_{2,\vec{h}}(\Omega, \mathbb{R}^n)$ der Gitterfunktionen. Man fordert nun, daß die Norm $\|\cdot\|_{L_{2,\vec{h}}}$ in diesem Raum so beschaffen ist, daß für alle $v(t, \vec{x}) \in L_2(\Omega, \mathbb{R}^n)$ und jedes $t \in [t_0, t_e]$ für die zugeordnete Gitterfunktion $v_{\vec{h}}(t) \in L_{2,\vec{h}}(\Omega, \mathbb{R}^n)$

$$\|v_{\vec{h}}(t)\|_{L_{2,\vec{h}}} \to \|v(t, \vec{x})\|_{L_2} \quad \text{für} \quad \|\vec{h}\| \to 0$$

gilt, d. h. die Normen aufeinander abgestimmt sind, vgl. Samarskij [45]. Daraus ergibt sich die Definition der „diskreten $L_2$-Norm“:

**Definition 4.2** Ist

$$v_{\vec{h}} = \left(v_{1,1,\ldots,1}^\top, \ldots, v_{N_1,1,\ldots,1}^\top, v_{1,2,\ldots,1}^\top, \ldots, v_{N_1,\ldots,N_d}^\top\right)^\top \in \mathbb{R}^{nN_1\cdot\ldots\cdot N_d},$$

dann ist die diskrete $L_2$-Norm von $v_{\vec{h}}$ gegeben durch

$$\|v_{\vec{h}}\| = \sqrt{\prod_{i=1}^{d} h_i v_{\vec{h}}^\top v_{\vec{h}}}.$$

□

Die der diskreten $L_2$-Norm zugeordnete Matrixnorm ist die Spektralnorm. Beide sind invariant unter orthogonalen Transformationen.
Für jeden Gitterpunkt $\vec{x}_{\vec{k}}$=$(x_{1,k_1}, \ldots, x_{d,k_d}) \in \Omega_{\vec{h}}$, $\vec{k}$=$(k_1, \ldots, k_d)$, seien

$$u_{\vec{k}}(t) = u_{k_1,\ldots,k_d}(t) \approx u(t, \vec{x}_{\vec{k}})$$

eine Näherung für die exakte Lösung im Gitterpunkt $\vec{x}_{\vec{k}}$ und

$$f_{\vec{k}}(t, u_{\vec{k}}) = f_{k_1,\ldots,k_d}(t, u_{\vec{k}}) = f(t, \vec{x}_{\vec{k}}, u_{\vec{k}}).$$

Mit

$$N = N_d \cdot \ldots \cdot N_1$$

erhält man in Analogie zum räumlich eindimensionalen Fall ein DA-System der Gestalt

$$M\dot{U}(t) = DU(t) + \tilde{F}(t, U), \tag{4.45}$$

wobei

$$M = I_N \otimes A,$$

$$D = -\left(\sum_{i=1}^{d} \frac{1}{h_i^2} I_{N_d\cdot\ldots\cdot N_{i+1}} \otimes P_i \otimes I_{N_{i-1}\cdot\ldots\cdot N_1} \otimes B_i + I_N \otimes C\right)$$

und

$$\tilde{F}(t, U) = F(t, U) - \omega(t)$$

sind. Die $I_{N_i\cdot\ldots\cdot N_j}$ sind dabei Einheitsmatrizen der Dimension $N_i \cdot N_{i-1} \cdot \ldots \cdot N_j$,

$$U(t) = \left(u_{1,1,\ldots,1}^\top(t), \ldots, u_{N_1,1,\ldots,1}^\top(t), u_{1,2,\ldots,1}^\top(t), \ldots, u_{N_1,\ldots,N_d}^\top(t)\right)^\top \in \mathbb{R}^{Nn}$$

und

$$F(t, U) = \left(f_{1,1,\ldots,1}^\top(t, u_{1,1,\ldots,1}), \ldots, f_{N_1,\ldots,N_d}^\top(t, u_{N_1,\ldots,N_d})\right)^\top$$

enthalten die Vektoren $u_{\vec{k}}$ und $f_{\vec{k}}$ in den Gitterpunkten,

$$\omega(t) = \left(r_{1,1,\ldots,1}^\top(t), \ldots, r_{N_1,1,\ldots,1}^\top(t), r_{1,2,\ldots,1}^\top(t), \ldots, r_{N_1,\ldots,N_d}^\top(t)\right)^\top \tag{4.46}$$

mit

$$r_{\vec{k}}(t) = \sum_{i\in M_D} \left\{ \begin{array}{lll} \frac{1+h_i r_i(\delta_i - 1)}{h_i^2} \psi_i(t, \vec{x}_{k_1,\ldots,k_{i-1},0,k_{i+1},\ldots,k_d}) : & k_i = 1 \\ \frac{1+h_i r_i \delta_i}{h_i^2} \psi_i(t, \vec{x}_{k_1,\ldots,k_{i-1},N_i+1,k_{i+1},\ldots,k_d}) & : \quad k_i = N_i \\ 0 & : \quad \text{sonst} \end{array} \right\}$$

$$+ \sum_{i\in M_N} \left\{ \begin{array}{lll} -\frac{1+h_i r_i(\delta_i - 1)}{h_i} \chi_i(t, x_{1,k_1}, \ldots, x_{i-1,k_{i-1}}, -l_i, x_{i+1,k_{i+1}}, \ldots) : & k_i = 1 \\ \frac{1+h_i r_i \delta_i}{h_i} \chi_i(t, x_{1,k_1}, \ldots, x_{i-1,k_{i-1}}, l_i, x_{i+1,k_{i+1}}, \ldots) & : \quad k_i = N_i \\ 0 & : \quad \text{sonst} \end{array} \right\}$$

enthält die vorgegebenen Randwerte, $U(t_0) \in \mathbb{R}^{nN}$ sei ein (konsistenter) Anfangsvektor. Die Matrizen $P_i \in \mathbb{R}^{N_i,N_i}$ sind durch (4.14), (4.26) bzw. (4.39) gegeben, $i = 1, \ldots, d$. Es gilt mit entsprechend (4.17), (4.28) bzw. (4.41) definierten Matrizen $S_{P_i} = (v_{jk})_{j,k=1,\ldots,N_i}^{\top}$ stets

$$S_{P_i}^{-1} \frac{1}{h_i^2} P_i S_{P_i} = \operatorname{diag}\{\lambda_{i1}, \ldots, \lambda_{iN_i}\} \tag{4.47}$$

mit $\lambda_{i,j}$ aus (4.16), (4.27) bzw. (4.40), und es existieren nach (4.23), (4.29) bzw. (4.43) positive Konstanten $C_{1i}$, $C_{2i}$ mit

$$C_{1i} \leq \max\{\|S_{P_i}\|, \|S_{P_i}^{-1}\|\} \leq C_{2i}. \tag{4.48}$$

Ausgehend von dem DA-System (4.45) kann der differentielle Zeitindex des PDA-Systems (3.2a) in Anlehnung an Lucht/Strehmel [36] nun wie folgt definiert werden:

**Definition 4.3** Kann $h_0 > 0$ so gewählt werden, daß der Differentiationsindex des DA-Systems (4.45) für alle $\vec{h}$ mit $0 < h_i \leq h_0$, $i = 1, \ldots, d$, existiert und unabhängig von $\vec{h}$ ist, so wird dieser Index als differentieller Zeitindex $\nu_{dt}$ des PDA-Systems bezeichnet. □

**Bemerkung 4.4** Im linearen Fall ($f$ unabhängig von $u$) ist dies gleichbedeutend damit, daß ein $h_0 > 0$ existiert, so daß für alle $\vec{h}$ mit $0 < h_i \leq h_0$, $i = 1, \ldots, d$, das Matrizenbüschel $\{D + \lambda M\}$ regulär und sein Nilpotenzindex unabhängig von $\vec{h}$ ist. Dann ist der differentielle Zeitindex $\nu_{dt}$ gleich diesem Nilpotenzindex. □

**Bemerkung 4.5** Die PDA-Systeme in den Beispielen 3.1, 3.2, 3.3 und 3.8 haben den differentiellen Zeitindex 0, das System in Beispiel 3.5 hat den differentiellen Zeitindex 1, und die PDA-Systeme in den Beispielen 3.6 und 3.7 haben den differentiellen Zeitindex 2. □

## 4.2 Konsistenz und Konvergenz der Semidiskretisierung linearer PDA-Systeme

Zur Definition der Konsistenz der Ortsdiskretisierung wird zunächst der lokale Ortsdiskretisierungsfehler eingeführt:

**Definition 4.6** Sei $U_{\vec{h}}(t)$ der Vektor der auf $\Omega_{\vec{h}}$ eingeschränkten exakten Lösung des PDA-Systems (3.2). Dann heißt

$$\alpha_{\vec{h}}(t) = M\, \dot{U}_{\vec{h}}(t) - D\, U_{\vec{h}}(t) - \tilde{F}(t, U_{\vec{h}}) \tag{4.49}$$

lokaler Ortsdiskretisierungsfehler. □

Der lokale Ortsdiskretisierungsfehler $\alpha_{\vec{h}}$ stellt nach (4.45) den Defekt der Gitterfunktion $U_{\vec{h}}$ bezüglich des semidiskreten Systems dar.

**Definition 4.7** Die Semidiskretisierung heißt konsistent mit der Ordnung $(p_1, \ldots, p_d)$, falls

$$\max_{t \in [t_0, t_e]} \|\alpha_{\vec{h}}(t)\| = \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i}) \quad \text{für} \quad \vec{h} \to 0$$

gilt. □

Mit den Taylor-Entwicklungen (4.6) bzw. (4.35) und (4.37) folgt

$$\alpha_{\vec{h}}(t) = \sum_{i=1}^{d} h_i^{p_i} (I_N \otimes B_i) \gamma^{(i)}(t), \tag{4.50}$$

wobei $p_i \in \{1,2\}$ ist in Abhängigkeit von der jeweiligen Ortsdiskretisierung und $\gamma_{\vec{k}}^{(i)}(t)$ gleich den entsprechend (4.8) bzw. (4.36) und (4.38) definierten $\gamma_{k_i}$ sind. Analog zu (4.11) existieren Konstanten $K_i$ mit

$$\|\gamma^{(i)}(t)\|_\infty \le K_i, \tag{4.51}$$

das heißt

$$\alpha_{\vec{h}}(t) = \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i}).$$

Um die Güte der Semidiskretisierung abschätzen zu können, wird neben dem lokalen Ortsdiskretisierungsfehler der globale Ortsdiskretisierungsfehler untersucht:

**Definition 4.8** Seien $U(t)$ die exakte Lösung des DA-Systems (4.45) zu gegebenem konsistentem Anfangswert und $U_{\vec{h}}(t)$ der Vektor der auf $\Omega_{\vec{h}}$ eingeschränkten exakten Lösung des PDA-Systems (3.2). Dann heißt der Vektor

$$\eta_{\vec{h}}(t) = U(t) - U_{\vec{h}}(t)$$

globaler Ortsdiskretisierungsfehler. □

**Definition 4.9** Die Semidiskretisierung heißt konvergent mit der Ordnung $(p_1, \ldots, p_d)$, falls

$$\max_{t\in[t_0,t_e]} \|\eta_{\vec{h}}(t)\| = \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i}) \quad \text{für} \quad \vec{h} \to 0$$

gilt. □

Für semidiskretisierte lineare PDA-Systeme, das heißt

$$f(t,\vec{x},u) \equiv f(t,\vec{x}) \quad \text{bzw.} \quad \tilde{F}(t,U) \equiv \tilde{F}(t), \tag{4.52}$$

kann die Konvergenz wie folgt gezeigt werden:
Aus der Definitionsgleichung (4.49) für den lokalen Ortsdiskretisierungsfehler, dem DA-System (4.45) und der Definition des globalen Ortsdiskretisierungsfehlers erhält man das lineare DA-System

$$M\dot{\eta}_{\vec{h}} = D\eta_{\vec{h}} - \alpha_{\vec{h}}(t) \tag{4.53a}$$

mit der Anfangsbedingung

$$\eta_{\vec{h}}(t_0) = U(t_0) - U_{\vec{h}}(t_0). \tag{4.53b}$$

Im folgenden wird vorausgesetzt, daß der differentielle Zeitindex $\nu_{dt}$ des PDA-Systems entsprechend Bemerkung 4.4 existiert. Dann ist das Matrizenbüschel $\{D+\lambda M\}$ regulär, und nach Satz 2.7 kann die eindeutige Lösung von (4.53) bei genügender Glattheit der exakten Lösung $u(\vec{x},t)$ und damit des lokalen Ortsdiskretisierungsfehlers $\alpha_{\vec{h}}(t)$ durch

$$\begin{aligned}\eta_{\vec{h}}(t) = {} & e^{\hat{M}^{\mathfrak{D}}\hat{D}(t-t_0)}\hat{M}\hat{M}^{\mathfrak{D}}\left(U(t_0) - U_{\vec{h}}(t_0)\right) - \hat{M}^{\mathfrak{D}}\int_{t_0}^{t} e^{\hat{M}^{\mathfrak{D}}\hat{D}(t-s)}\hat{\alpha}_{\vec{h}}(s)\,ds \\ & + \left(I_{Nn} - \hat{M}\hat{M}^{\mathfrak{D}}\right)\sum_{i=0}^{\nu_{dt}-1}\left(\hat{M}\hat{D}^{\mathfrak{D}}\right)^i \hat{D}^{\mathfrak{D}}\hat{\alpha}_{\vec{h}}^{(i)}(t)\end{aligned} \tag{4.54}$$

dargestellt werden, wobei mit einem $c \in \mathbb{C}$, für das $(D+cM)$ regulär ist, $\hat{M}$ und $\hat{D}$ durch

$$\hat{D} = (D+cM)^{-1}D, \quad \hat{M} = (D+cM)^{-1}M \tag{4.55}$$

und $\hat{\alpha}_{\vec{h}}(t)$ durch

$$\hat{\alpha}_{\vec{h}}(t) = (D + cM)^{-1}\alpha_{\vec{h}}(t) \tag{4.56}$$

definiert sind (es gilt nach Bemerkung 4.4 und der Beziehung (2.10) für den Index eines Matrizenbüschels $\mathrm{ind}(\hat{M}) = \mathrm{ind}(D, M) = \nu_{dt}$).
Durch direkte Abschätzung der rechten Seite der Fehlergleichung (4.54) wie in Eichler-Liebenow [16] würde man nur schwer nachprüfbare Bedingungen für die Konvergenz der Ortsdiskretisierung erhalten, weil die Dimensionen der zu berechnenden Matrizen für gegen Null gehende Ortsschrittweiten gegen Unendlich gehen. Mittels einer Diagonalisierung dieser Matrizen läßt sich erreichen, daß nur noch Matrizen fester Dimension, d. h., die Dimension ist unabhängig von der Ortsschrittweite $\vec{h}$, für die Konvergenzuntersuchungen betrachtet werden müssen.
Sei

$$Q = S_{P_d} \otimes \ldots \otimes S_{P_1} \otimes I_n \tag{4.57}$$

mit $S_{P_i}$ aus (4.47). Dann gelten

$$Q^{-1}MQ = M \quad \text{und} \quad Q^{-1}DQ = \mathrm{diag}_{\vec{k}}\{D_{\vec{k}}\}, \tag{4.58}$$

wobei

$$D_{\vec{k}} = -\left(\sum_{v=1}^{d} \lambda_{v,k_v} B_v + C\right) \tag{4.59}$$

mit $\lambda_{i,j}$ aus (4.47) und

$$\mathrm{diag}_{\vec{k}}\{D_{\vec{k}}\} = \mathrm{diag}\{D_{1,\ldots,1}, \ldots, D_{N_1,1,\ldots,1}, D_{1,2,1,\ldots,1}, \ldots, D_{N_1,\ldots,N_d}\}.$$

Aus den Gleichungen (4.55) für $\hat{D}$ und $\hat{M}$ folgen deshalb

$$Q^{-1}\hat{D}Q = \mathrm{diag}_{\vec{k}}\{(D_{\vec{k}} + cA)^{-1}D_{\vec{k}}\} = \mathrm{diag}_{\vec{k}}\{\hat{D}_{\vec{k}}\}, \tag{4.60a}$$

$$Q^{-1}\hat{M}Q = \mathrm{diag}_{\vec{k}}\{(D_{\vec{k}} + cA)^{-1}A\} = \mathrm{diag}_{\vec{k}}\{\hat{M}_{\vec{k}}\}, \tag{4.60b}$$

wobei

$$\hat{D}_{\vec{k}} = (D_{\vec{k}} + cA)^{-1}D_{\vec{k}}, \quad \hat{M}_{\vec{k}} = (D_{\vec{k}} + cA)^{-1}A. \tag{4.61}$$

Da die Drazin-Inverse nach (2.11) verträglich mit Ähnlichkeitstransformationen ist, ergibt sich daraus

$$Q^{-1}\hat{D}^{\mathfrak{D}}Q = \mathrm{diag}_{\vec{k}}\{\hat{D}^{\mathfrak{D}}_{\vec{k}}\}, \quad Q^{-1}\hat{M}^{\mathfrak{D}}Q = \mathrm{diag}_{\vec{k}}\{\hat{M}^{\mathfrak{D}}_{\vec{k}}\}. \tag{4.62}$$

Aus der Fehlergleichung (4.54) erhält man durch Einsetzen der Beziehungen (4.56), (4.58), (4.60) und (4.62) und Berücksichtigung der Kommutativität von $\hat{M}^{\mathfrak{D}}_{\vec{k}}$ und $\hat{D}_{\vec{k}}$

$$\begin{aligned}
\eta_{\vec{h}}(t) = {} & Q\left(\mathrm{diag}_{\vec{k}}\left\{e^{\hat{M}^{\mathfrak{D}}_{\vec{k}}\hat{D}_{\vec{k}}(t-t_0)}\hat{M}_{\vec{k}}\hat{M}^{\mathfrak{D}}_{\vec{k}}\right\}\right)Q^{-1}\left(U(t_0) - U_{\vec{h}}(t_0)\right) \\
& -Q\int_{t_0}^{t}\left(\mathrm{diag}_{\vec{k}}\left\{e^{\hat{M}^{\mathfrak{D}}_{\vec{k}}\hat{D}_{\vec{k}}(t-s)}\hat{M}^{\mathfrak{D}}_{\vec{k}}(D_{\vec{k}} + cA)^{-1}\right\}\right)Q^{-1}\alpha_{\vec{h}}(s)\,ds \\
& +Q\sum_{i=0}^{\nu_{dt}-1}\left(\mathrm{diag}_{\vec{k}}\left\{(I_n - \hat{M}_{\vec{k}}\hat{M}^{\mathfrak{D}}_{\vec{k}})\left(\hat{M}_{\vec{k}}\hat{D}^{\mathfrak{D}}_{\vec{k}}\right)^i\hat{D}^{\mathfrak{D}}_{\vec{k}}(D_{\vec{k}} + cA)^{-1}\right\}\right)Q^{-1}\alpha^{(i)}_{\vec{h}}(t)
\end{aligned} \tag{4.63}$$

mit

$$\nu_{dt} = \mathrm{ind}(\hat{M}) = \mathrm{ind}(\mathrm{diag}_{\vec{k}}\{\hat{M}_{\vec{k}}\}) = \max_{\vec{k}}\{\mathrm{ind}(\hat{M}_{\vec{k}})\} = \max_{\vec{k}}\{\mathrm{ind}(D_{\vec{k}}, A)\}. \tag{4.64}$$

Da für alle Matrizen $S$, $T$

$$\|S \otimes T\| = \sqrt{\lambda_{\max}\left(\left(\bar{S}^\top S\right) \otimes \left(\bar{T}^\top T\right)\right)} = \sqrt{\lambda_{\max}\left(\bar{S}^\top S\right) \lambda_{\max}\left(\bar{T}^\top T\right)} = \|S\| \, \|T\|$$

gilt, folgt aus der Definitionsgleichung (4.57) der Matrix $Q$ und der für $\|S_{P_i}\|$ und $\|S_{P_i}^{-1}\|$ geltenden Abschätzung (4.48), daß positive Konstanten $C_1$, $C_2$ mit

$$C_1 \leq \max\{\|Q\|, \|Q^{-1}\|\} \leq C_2 \tag{4.65}$$

existieren.
Wird $A_k \in \mathbb{C}^{n_{1k}s_1, n_{2k}s_2}$ durch $A_k = \left((A_k)_{j_1 j_2}\right)_{j_i=1,\ldots,s_i}$ in $s_1 \times s_2$ Blockmatrizen der Dimension $n_{1k} \times n_{2k}$ unterteilt, so gilt für die Spektralnorm

$$\begin{aligned}
\| \left(\mathrm{diag}_k\{(A_k)_{j_1 j_2}\}\right)_{j_i=1,\ldots,s_i} \| &= \sqrt{\lambda_{\max}\left(\left(\mathrm{diag}_k\{\sum_{i=1}^{s_1} \left(\bar{A}_k^\top\right)_{ji} (A_k)_{il}\}\right)_{j,l=1,\ldots,s_2}\right)} \\
&= \max_k \sqrt{\lambda_{\max}\left(\sum_{i=1}^{s_1} \left(\left(\bar{A}_k^\top\right)_{ji} (A_k)_{il}\right)_{j,l=1,\ldots,s_2}\right)} \\
&= \max_k \sqrt{\lambda_{\max}\left(\bar{A}_k^\top A_k\right)} = \max_k \|A_k\|. \qquad (4.66)
\end{aligned}$$

Mit der Gleichung (4.50) für den lokalen Ortsdiskretisierungsfehler und den Beziehungen (4.65), (4.66) und (2.27) für die Spektralnorm folgt aus der Fehlergleichung (4.63) der folgende Konvergenzsatz:

**Satz 4.10** Die Ortsdiskretisierung sei mit der Ordnung $(p_1, \ldots, p_d)$ konsistent. Gilt

$$\|U(t_0) - U_{\vec{h}}(t_0)\| = \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i}) \quad \text{für} \quad \vec{h} \to 0$$

und sind die logarithmische Matrixnorm $\mu_2[\hat{M}_{\vec{k}}^{\mathfrak{D}} \hat{D}_{\vec{k}}]$ sowie die Matrizen

$$\hat{M}_{\vec{k}} \hat{M}_{\vec{k}}^{\mathfrak{D}}, \quad \hat{M}_{\vec{k}}^{\mathfrak{D}} (D_{\vec{k}} + cA)^{-1} B_j, \quad (I_n - \hat{M}_{\vec{k}} \hat{M}_{\vec{k}}^{\mathfrak{D}}) \left(\hat{M}_{\vec{k}} \hat{D}_{\vec{k}}^{\mathfrak{D}}\right)^i \hat{D}_{\vec{k}}^{\mathfrak{D}} (D_{\vec{k}} + cA)^{-1} B_j$$

normmäßig nach oben beschränkt für alle $\vec{k}$, $i = 0, \ldots, \nu_{dt} - 1$, $j = 1, \ldots, d$ und $\vec{h} \to 0$, so ist die Ortsdiskretisierung auch konvergent mit der Ordnung $(p_1, \ldots, p_d)$. □

**Bemerkung 4.11** Die Spektralnormen der Matrizen in Satz 4.10 brauchen nicht berechnet zu werden, da nach Ungleichung (2.28) $\|K\|$ für eine Matrix $K$ genau dann beschränkt ist, wenn alle Elemente von $K$ beschränkt sind. □

**Bemerkung 4.12** Verwendet man als konsistenten Anfangswert des DA-Systems (4.45)

$$U(t_0) = \hat{M} \hat{M}^{\mathfrak{D}} U_{\vec{h}}(t_0) - (I_{Nn} - \hat{M} \hat{M}^{\mathfrak{D}}) \sum_{i=0}^{\nu_{dt}-1} \left(\hat{M} \hat{D}^{\mathfrak{D}}\right)^i \hat{D}^{\mathfrak{D}} (D + cM)^{-1} \tilde{F}^{(i)}(t) \tag{4.67}$$

gemäß (2.12), so ist unter den übrigen Voraussetzungen des Satzes 4.10 die Voraussetzung $\|U(t_0) - U_{\vec{h}}(t_0)\| = \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i})$ für $\vec{h} \to 0$ erfüllt, da mit

$$U_{\vec{h}}(t_0) = \hat{M} \hat{M}^{\mathfrak{D}} U_{\vec{h}}(t_0) - (I_{Nn} - \hat{M} \hat{M}^{\mathfrak{D}}) \sum_{i=0}^{\nu_{dt}-1} \left(\hat{M} \hat{D}^{\mathfrak{D}}\right)^i \hat{D}^{\mathfrak{D}} (D + cM)^{-1} \left(\tilde{F}^{(i)}(t) + \alpha_{\vec{h}}^{(i)}(t)\right)$$

aus (4.67) und der für den lokalen Ortsdiskretisierungsfehler geltenden Gleichung (4.50) bei genügend glatter exakter Lösung

$$\|U(t_0) - U_{\vec{h}}(t_0)\| \leq \sum_{i=1}^{d} \sum_{j=0}^{\nu_{dt}-1} \max_{\vec{k}} \left\| (I_n - \hat{M}_{\vec{k}} \hat{M}_{\vec{k}}^{\mathfrak{D}}) \left( \hat{M}_{\vec{k}} \hat{D}_{\vec{k}}^{\mathfrak{D}} \right)^j \hat{D}_{\vec{k}}^{\mathfrak{D}} (D_{\vec{k}} + cA)^{-1} B_i \right\| \mathcal{O}(h_i^{p_i})$$

folgt. □

**Beispiel 4.13** Betrachtet wird das lineare PDA-System

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{=A} u_t(t,x) + \underbrace{\begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}}_{=B} u_{xx}(t,x) + \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}}_{=C} u(t,x) = f(t,x).$$

Es gilt nach (4.59) mit $d = 1$ und $\lambda_k = \lambda_{1,k}$

$$D_k = \begin{pmatrix} -1 & \lambda_k - 1 & 0 \\ 0 & 0 & \lambda_k - 1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Für alle $c$, für die $(D_k + cA)$ regulär ist (d. h. $c \neq 1$), folgen aus (4.61)

$$\begin{aligned} \hat{D}_k &= \begin{pmatrix} c-1 & \lambda_k - 1 & 0 \\ 0 & c & \lambda_k - 1 \\ 0 & -1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} -1 & \lambda_k - 1 & 0 \\ 0 & 0 & \lambda_k - 1 \\ 0 & -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{c-1} & 0 & \frac{\lambda_k - 1}{c-1} \\ 0 & 0 & -1 \\ 0 & \frac{1}{\lambda_k - 1} & \frac{c}{\lambda_k - 1} \end{pmatrix} \begin{pmatrix} -1 & \lambda_k - 1 & 0 \\ 0 & 0 & \lambda_k - 1 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{1-c} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{c}{1-\lambda_k} & 1 \end{pmatrix}, \\ \hat{M}_k &= \begin{pmatrix} \frac{1}{c-1} & 0 & \frac{\lambda_k - 1}{c-1} \\ 0 & 0 & -1 \\ 0 & \frac{1}{\lambda_k - 1} & \frac{c}{\lambda_k - 1} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{c-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{\lambda_k - 1} & 0 \end{pmatrix}. \end{aligned}$$

Daraus folgen

$$\hat{D}_k^{\mathfrak{D}} = \hat{D}_k^{-1} = \begin{pmatrix} 1-c & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{c}{\lambda_k - 1} & 1 \end{pmatrix}, \quad \hat{M}_k^{\mathfrak{D}} = \begin{pmatrix} c-1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat{M}_k \hat{M}_k^{\mathfrak{D}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

und

$$\mu_2[\hat{M}_k^{\mathfrak{D}} \hat{D}_k] = \mu_2 \left[ \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right] = -1$$

sowie

$$(D_k + cA)^{-1} B = \begin{pmatrix} \frac{1}{c-1} & 0 & \frac{\lambda_k - 1}{c-1} \\ 0 & 0 & -1 \\ 0 & \frac{1}{\lambda_k - 1} & \frac{c}{\lambda_k - 1} \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{1-c} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\lambda_k} \end{pmatrix}.$$

Ferner erhält man aus (4.64)

$$\operatorname{ind}(D_k, A) = \operatorname{ind}(\hat{M}_k) = 2 = \nu_{dt},$$

und es gilt

$$(I_n - \hat{M}_{\vec{k}}\hat{M}^{\mathfrak{D}}_{\vec{k}})\left(\hat{M}_{\vec{k}}\hat{D}^{\mathfrak{D}}_{\vec{k}}\right)^i \hat{D}^{\mathfrak{D}}_{\vec{k}}(D_{\vec{k}}+cA)^{-1}B = \begin{pmatrix} 0&0&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix} -1 & 0 & 0\\ 0 & 0 & 0\\ 0 & \frac{1}{\lambda_k-1} & 0\end{pmatrix}^i \begin{pmatrix} 0&1&0\\0&0&0\\0&0&\frac{1}{1-\lambda_k}\end{pmatrix}$$
$$= \begin{cases} \begin{pmatrix} 0&0&0\\0&0&0\\0&0&\frac{1}{1-\lambda_k}\end{pmatrix}: \ i=0 \\ \begin{pmatrix} 0&0&0\\0&0&0\\0&0&0\end{pmatrix} \quad : \ i\geq 1 \end{cases}.$$

Da wegen $\lambda_k \leq 0$ die Norm aller dieser Matrizen für $h \to 0$ beschränkt bleibt, ist nach Satz 4.10 die Ortsdiskretisierung mit Anfangswert (4.67) für dieses Beispiel konvergent. □

Auf der Grundlage einer Weierstraß-Kronecker-Transformation soll nun ein zweiter Konvergenzsatz hergeleitet werden, der bei bekannten Transformationsmatrizen einfacher als Satz 4.10 ist.
Wegen (4.58) und $M = I_N \otimes A$ folgt

$$\det(D+\lambda M) = \det\left(Q^{-1}(D+\lambda M)Q\right) = \det\left(\operatorname{diag}_{\vec{k}}\{D_{\vec{k}}+\lambda A\}\right) = \prod_{\vec{k}} \det(D_{\vec{k}}+\lambda A),$$

das Matrizenbüschel $\{D+\lambda M\}$ ist nach Definition 2.4 also genau dann regulär, wenn die Matrizenbüschel $\{D_{\vec{k}}+\lambda A\}$ alle regulär sind.
Dann existieren nach Satz 2.5 reguläre Matrizen $P_{\vec{k}}$ und $Q_{\vec{k}}$ mit

$$P_{\vec{k}}AQ_{\vec{k}} = \operatorname{diag}\{I_{n_{\vec{k}1}},\ldots,I_{n_{\vec{k}s_{\vec{k}}}},N_{m_{\vec{k}1}},\ldots,N_{m_{\vec{k}l_{\vec{k}}}}\}, \tag{4.68a}$$
$$P_{\vec{k}}D_{\vec{k}}Q_{\vec{k}} = \operatorname{diag}\{R_{\vec{k}1},\ldots,R_{\vec{k}s_{\vec{k}}},I_{m_{\vec{k}1}},\ldots,I_{m_{\vec{k}l_{\vec{k}}}}\}, \tag{4.68b}$$

wobei

$$R_{\vec{k}i} = \begin{pmatrix} \kappa_{\vec{k}i} & 1 & & 0\\ & \ddots & \ddots & \\ & & \kappa_{\vec{k}i} & 1\\ 0 & & & \kappa_{\vec{k}i}\end{pmatrix} \in \mathbb{C}^{n_{\vec{k}i},n_{\vec{k}i}}, \quad N_{m_{\vec{k}i}} = \begin{pmatrix} 0 & 1 & & 0\\ & \ddots & \ddots & \\ & & 0 & 1\\ 0 & & & 0\end{pmatrix} \in \mathbb{C}^{m_{\vec{k}i},m_{\vec{k}i}}. \tag{4.68c}$$

Daraus folgt für den differentiellen Zeitindex

$$\nu_{dt} = \max_{\vec{k}}\{m_{\vec{k}i} : i = 1,\ldots,l_{\vec{k}}\}.$$

Aus den Gleichungen (4.61) für $\hat{D}_{\vec{k}}$ und $\hat{M}_{\vec{k}}$ erhält man

$$\hat{D}_{\vec{k}} = Q_{\vec{k}}\operatorname{diag}\{\ldots,(R_{\vec{k}j_1}+cI_{n_{\vec{k}j_1}})^{-1}R_{\vec{k}j_1},\ldots,(I_{m_{\vec{k}j_2}}+cN_{m_{\vec{k}j_2}})^{-1},\ldots\}Q_{\vec{k}}^{-1}$$

und

$$\hat{M}_{\vec{k}} = Q_{\vec{k}}\operatorname{diag}\{\ldots,(R_{\vec{k}j_1}+cI_{n_{\vec{k}j_1}})^{-1},\ldots,(I_{m_{\vec{k}j_2}}+cN_{m_{\vec{k}j_2}})^{-1}N_{m_{\vec{k}j_2}},\ldots\}Q_{\vec{k}}^{-1}$$

und daraus

$$\hat{D}^{\mathfrak{D}}_{\vec{k}} = Q_{\vec{k}}\operatorname{diag}\{\ldots,(R_{\vec{k}j_1}+cI_{n_{\vec{k}j_1}})R^{\mathfrak{D}}_{\vec{k}j_1},\ldots,I_{m_{\vec{k}j_2}}+cN_{m_{\vec{k}j_2}},\ldots\}Q_{\vec{k}}^{-1}$$

und

$$\hat{M}^{\mathfrak{D}}_{\vec{k}} = Q_{\vec{k}}\operatorname{diag}\{\ldots,R_{\vec{k}j_1}+cI_{n_{\vec{k}j_1}},\ldots,\mathfrak{o},\ldots\}Q_{\vec{k}}^{-1}$$

sowie

$$\begin{aligned}
\hat{M}_{\vec{k}}^{\mathfrak{D}}\hat{D}_{\vec{k}} &= Q_{\vec{k}}\mathrm{diag}\{\ldots,R_{\vec{k}j_1},\ldots,\mathfrak{o},\ldots\}Q_{\vec{k}}^{-1},\\
\hat{M}_{\vec{k}}\hat{M}_{\vec{k}}^{\mathfrak{D}} &= Q_{\vec{k}}\mathrm{diag}\{\ldots,I_{n_{\vec{k}j_1}},\ldots,\mathfrak{o},\ldots\}Q_{\vec{k}}^{-1},\\
\hat{M}_{\vec{k}}^{\mathfrak{D}}(D_{\vec{k}}+cA)^{-1} &= Q_{\vec{k}}\mathrm{diag}\{\ldots,I_{n_{\vec{k}j_1}},\ldots,\mathfrak{o},\ldots\}P_{\vec{k}},\\
\hat{M}_{\vec{k}}\hat{D}_{\vec{k}}^{\mathfrak{D}} &= Q_{\vec{k}}\mathrm{diag}\{\ldots,R_{\vec{k}j_1}^{\mathfrak{D}},\ldots,N_{m_{\vec{k}j_2}},\ldots\}Q_{\vec{k}}^{-1},\\
\hat{D}_{\vec{k}}^{\mathfrak{D}}(D_{\vec{k}}+cA)^{-1} &= Q_{\vec{k}}\mathrm{diag}\{\ldots,R_{\vec{k}j_1}^{\mathfrak{D}},\ldots,I_{m_{\vec{k}j_2}},\ldots\}P_{\vec{k}}.
\end{aligned}$$

Einsetzen in die Fehlergleichung (4.63) liefert mit der für den lokalen Ortsdiskretisierungsfehler geltenden Gleichung (4.50)

$$\begin{aligned}
\eta_{\vec{h}}(t) = {} & Q\ \mathrm{diag}_{\vec{k}}\left\{Q_{\vec{k}}\mathrm{diag}\left\{\ldots,e^{R_{\vec{k}j_1}(t-t_0)},\ldots,\mathfrak{o},\ldots\right\}Q_{\vec{k}}^{-1}\right\}Q^{-1}\left(U(t_0)-U_{\vec{h}}(t_0)\right)\\
& -Q\sum_{v=1}^{d}\int_{t_0}^{t}\mathrm{diag}_{\vec{k}}\left\{Q_{\vec{k}}\mathrm{diag}\left\{\ldots,e^{R_{\vec{k}j_1}(t-s)},\ldots,\mathfrak{o},\ldots\right\}P_{\vec{k}}B_v\right\}Q^{-1}h_v^{p_v}\gamma^{(v)}(s)\ ds\\
& +Q\sum_{v=1}^{d}\sum_{i=0}^{\nu_{dt}-1}\mathrm{diag}_{\vec{k}}\left\{Q_{\vec{k}}\mathrm{diag}\left\{\ldots,\mathfrak{o},\ldots,N_{m_{\vec{k}j_2}}^{i},\ldots\right\}P_{\vec{k}}B_v\right\}Q^{-1}h_v^{p_v}\frac{d^i}{dt^i}\gamma^{(v)}(t).
\end{aligned}$$

Mit den Beziehungen (4.65) und (4.66) für die diskrete $L_2$-Norm und der Darstellung (2.25) für die Matrixfunktion eines Jordan-Blocks folgt daraus der folgende Konvergenzsatz:

**Satz 4.14** Die Ortsdiskretisierung sei mit der Ordnung $(p_1,\ldots,p_d)$ konsistent. Für $h_v\to 0$ $(N_v\to\infty)$, $v=1,\ldots,d$, gelte $\|U(t_0)-U_{\vec{h}}(t_0)\| = \sum_{i=1}^{d}\mathcal{O}(h_i^{p_i})$, und es seien für alle $\vec{k}$ mit $k_v=1,\ldots,N_v$ die folgenden Voraussetzungen erfüllt:

(a) $\|Q_{\vec{k}}\mathrm{diag}\{N_{n_{\vec{k}1}}^{i},\mathfrak{o},\ldots,\mathfrak{o}\}Q_{\vec{k}}^{-1}\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N_{n_{\vec{k}s_{\vec{k}}}}^{i},\mathfrak{o},\ldots,\mathfrak{o}\}Q_{\vec{k}}^{-1}\|,$
$\|Q_{\vec{k}}\mathrm{diag}\{N_{n_{\vec{k}1}}^{i},\mathfrak{o},\ldots,\mathfrak{o}\}P_{\vec{k}}B_v\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N_{m_{\vec{k}l_{\vec{k}}}}^{i}\}P_{\vec{k}}B_v\|,$
sind für $i=0,\ldots,\nu_{dt}-1$ beschränkt,

(b) in (4.68c) sind $\kappa_{\vec{k}i}$ für $i=1,\ldots,s_{\vec{k}}$ nach oben beschränkt.

Dann ist die Ortsdiskretisierung auch konvergent mit der Ordnung $(p_1,\ldots,p_d)$. □

**Beispiel 4.15** Betrachtet werde das PDA-System aus Beispiel 4.13. Mit

$$P_k=\begin{pmatrix}1&0&\lambda_k-1\\0&1&0\\0&0&-1\end{pmatrix},\quad Q_k=\begin{pmatrix}1&0&0\\0&0&1\\0&\frac{1}{\lambda_k-1}&0\end{pmatrix}$$

gelten

$$P_kAQ_k=\begin{pmatrix}1&0&0\\0&0&1\\0&0&0\end{pmatrix},\quad P_kD_kQ_k=\begin{pmatrix}-1&0&0\\0&1&0\\0&0&1\end{pmatrix}$$

und

$$P_kB=\begin{pmatrix}0&-1&0\\0&0&-1\\0&0&0\end{pmatrix},\quad Q_k^{-1}=\begin{pmatrix}1&0&0\\0&0&\lambda_k-1\\0&1&0\end{pmatrix}.$$

Da $\|Q_k\|$ und $\|P_kB\|$ damit beschränkt sind und

$$Q_k\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}Q_k^{-1}=\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}$$

gilt, ist die Ortsdiskretisierung mit Anfangswert (4.67) nach Satz 4.14 konvergent. □

# Kapitel 5

# Diskretisierung des MOL-DA-Systems

Nachdem in Kapitel 4 die betrachteten PDA-Systeme mittels finiter Differenzen semidiskretisiert wurden, wird im folgenden die Zeitdiskretisierung des MOL-DA-Systems mittels Runge-Kutta-Methoden betrachtet. Ziel dieses Kapitels ist die Untersuchung der Konvergenzordnung der entstehenden Gesamtdiskretisierung. Zunächst für lineare Probleme und darauf aufbauend auch für semilineare Probleme werden hinreichende Konvergenzbedingungen angegeben. Konvergenzuntersuchungen für eine spezielle Klasse quasilinearer PDA-Systeme findet man in Lucht/Debrabant [35].
Bei parabolischen und hyperbolischen Problemen ist für Runge-Kutta-Verfahren hoher Ordnung das Phänomen der Ordnungsreduktion bekannt, in Abhängigkeit von den (räumlichen) Randbedingungen konvergiert das Verfahren mit einer geringeren Konvergenzordnung als erwartet, vgl. Strehmel/Weiner [53] und Calvo/Palencia [8]. Auch gebrochene (nicht ganzzahlige) Konvergenzordnungen können auftreten, vgl. Sanz-Serna/Verwer/Hundsdorfer [46] und Verwer [56]. Untersuchungen dazu für parabolische Probleme in abstrakter Formulierung mit impliziten Runge-Kutta-Verfahren findet man zum Beispiel in Brenner/Crouzeix/Thomée [4], Ostermann/Roche [42] und Lubich/Ostermann [33]). Diese können nicht unmittelbar auf die hier betrachteten PDA-Systeme angewendet werden, weil die dortigen Voraussetzungen, insbesondere die Regularität der Matrix $A$ und daß der Differentialoperator

$$\mathcal{L}=\sum_{i=1}^{d} B_i(\partial_{x_i x_i} + r_i \partial_{x_i}) + C$$

positiv definit oder Generator einer beschränkten analytischen Halbgruppe ist, dessen Resolventenmenge 0 enthält, im allgemeinen nicht erfüllt sind.

## 5.1 Zeitdiskretisierung durch Runge-Kutta-Verfahren

Die Anwendung eines $s$-stufigen impliziten Runge-Kutta-Verfahrens (2.23) auf das DA-System (4.45) liefert

$$U_{m+1} = U_m + \tau \sum_{i=1}^{s} b_i K_{m+1}^{(i)}, \tag{5.1a}$$

$$U_{m+1}^{(i)} = U_m + \tau \sum_{j=1}^{s} a_{ij} K_{m+1}^{(j)}, \quad i = 1, \ldots, s, \tag{5.1b}$$

$$M K_{m+1}^{(i)} = D U_{m+1}^{(i)} + \tilde{F}(t_m + c_i \tau, U_{m+1}^{(i)}), \quad i = 1, \ldots, s. \tag{5.1c}$$

Als Startvektor wird der auf das Ortsgitter eingeschränkte exakte Anfangswert $U_{\vec{h}}(t_0)$ des PDA-Systems verwendet,

$$U_0 = U_{\vec{h}}(t_0) = \left(\varphi^\top(x_{11},\ldots,x_{d1}),...,\varphi^\top(x_{1N_1},\ldots,x_{dN_d})\right)^\top \in \mathbb{R}^{nN}. \tag{5.2}$$

Unter Verwendung des Kronecker-Produkts folgen aus (5.1)

$$\begin{aligned} U_{m+1} &= U_m + \tau\left(b^\top \otimes I_{Nn}\right) K_{m+1}, && \text{(5.3a)}\\ S_{m+1} &= \mathbb{1}_s \otimes U_m + \tau\left(\mathfrak{A} \otimes I_{Nn}\right) K_{m+1}, && \text{(5.3b)}\\ \left(I_s \otimes M\right) K_{m+1} &= \left(I_s \otimes D\right) S_{m+1} + \bar{F}(t_{m+1}, S_{m+1}), && \text{(5.3c)} \end{aligned}$$

wobei $\mathbb{1}_s$ der $s$-dimensionale Vektor $(1,\ldots,1)^\top$ und

$$\begin{aligned} K_{m+1} &= \left({K_{m+1}^{(1)}}^\top,\ldots,{K_{m+1}^{(s)}}^\top\right)^\top,\\ S_{m+1} &= \left({U_{m+1}^{(1)}}^\top,\ldots,{U_{m+1}^{(s)}}^\top\right)^\top,\\ \bar{F}(t_{m+1},S_{m+1}) &= \left(\tilde{F}\left(t_m + c_1\tau, U_{m+1}^{(1)}\right)^\top,\ldots,\tilde{F}\left(t_m + c_s\tau, U_{m+1}^{(s)}\right)^\top\right)^\top \end{aligned}$$

sind.
Im folgenden sollen zunächst der Einfluß von Störungen auf die Lösung des Systems (5.3) und darauf aufbauend die Konvergenzordnung dieses Verfahrens bei Anwendung auf semidiskretisierte lineare PDA-Systeme (d. h. $\tilde{F}(t,U) \equiv \tilde{F}(t)$) untersucht werden. Unter Verwendung der für lineare Systeme erzielten Ergebnisse werden dann auch für semilineare Systeme hinreichende Konvergenzbedingungen angegeben.

## 5.2 Konvergenz der Gesamtdiskretisierung

### 5.2.1 Einfluß von Störungen in den Runge-Kutta-Gleichungen bei linearen PDA-Systemen

In diesem Abschnitt soll der Einfluß von Störungen auf die numerische Lösung (5.3a) und die Stufenwerte (5.3b) untersucht werden. Dies stellt die Grundlage für eine Abschätzung des globalen Diskretisierungsfehlers dar und liefert damit das erforderliche Hilfsmittel für die Untersuchung der Konvergenz des Verfahrens.
Betrachtet wird ein gestörtes Runge-Kutta-Verfahren

$$\begin{aligned} \hat{U}_{m+1} &= \hat{U}_m + \tau\left(b^\top \otimes I_{Nn}\right)\hat{K}_{m+1} + \theta_{m+1}, && \text{(5.4a)}\\ \hat{S}_{m+1} &= \mathbb{1}_s \otimes \hat{U}_m + \tau\left(\mathfrak{A} \otimes I_{Nn}\right)\hat{K}_{m+1} + \Theta_{m+1}, && \text{(5.4b)}\\ \left(I_s \otimes M\right)\hat{K}_{m+1} &= \left(I_s \otimes D\right)\hat{S}_{m+1} + \bar{F}(t_{m+1}) + \rho_{m+1}. && \text{(5.4c)} \end{aligned}$$

Die Störungen $\theta_{m+1} \in \mathbb{R}^{Nn}$, $\Theta_{m+1},\ \rho_{m+1} \in \mathbb{R}^{sNn}$ können dabei zum Beispiel als Rundungsfehler oder als Defekt der exakten Lösung bei Einsetzen in das System (5.3) interpretiert werden.
Subtrahiert man (5.3c) von (5.4c) sowie (5.3b) von (5.4b), so folgen

$$\begin{aligned} \left(I_s \otimes M\right)\left(\hat{K}_{m+1} - K_{m+1}\right) &= \left(I_s \otimes D\right)\left(\hat{S}_{m+1} - S_{m+1}\right) + \rho_{m+1},\\ \hat{S}_{m+1} - S_{m+1} &= \mathbb{1}_s \otimes \left(\hat{U}_m - U_m\right) + \tau\left(\mathfrak{A} \otimes I_{Nn}\right)\left(\hat{K}_{m+1} - K_{m+1}\right) + \Theta_{m+1} \end{aligned} \tag{5.5}$$

und daraus durch Einsetzen der zweiten in die erste Gleichung

$$(I_s \otimes M)\left(\hat{K}_{m+1} - K_{m+1}\right) = (\mathbb{1}_s \otimes D)\left(\hat{U}_m - U_m\right) + \tau\,(\mathfrak{A} \otimes D)\left(\hat{K}_{m+1} - K_{m+1}\right) + (I_s \otimes D)\,\Theta_{m+1} + \rho_{m+1}. \tag{5.6}$$

Für $\bar{N} \in \mathbb{N}$ und beliebige Matrizen $O,\ K \in \mathbb{R}^{\bar{N},\bar{N}}$ werden zur Abkürzung die Matrix $G(O,K)$ und für reguläre $G(O,K)$ auch die Matrizen $J(O,K)$, $R(O,K)$, $L(O,K)$ und $H(O,K)$ wie folgt definiert:

$$G(O,K) = I_s \otimes O - \mathfrak{A} \otimes K, \tag{5.7}$$
$$J(O,K) = (b^\top \otimes I_{\bar{N}})G(O,K)^{-1}, \tag{5.8}$$
$$R(O,K) = I_{\bar{N}} + J(O,K)(\mathbb{1}_s \otimes K), \tag{5.9}$$
$$L(O,K) = J(O,K)(I_s \otimes K), \tag{5.10}$$
$$H(O,K) = G(O,K)^{-1}(\mathbb{1}_s \otimes O). \tag{5.11}$$

**Bemerkung 5.1** Im Spezialfall $\bar{N} = 1$, $O = 1$ und $K = \tau\lambda$ liefert $R(O,K)$ die klassische Stabilitätsfunktion $R(1,\tau\lambda) = R(\tau\lambda)$ des Runge-Kutta-Verfahrens, so daß $R(O,K)$ als verallgemeinerte Stabilitätsfunktion angesehen werden kann. □

Mit (5.7) läßt sich (5.6) schreiben als

$$G(M,\tau D)\left(\hat{K}_{m+1} - K_{m+1}\right) = (\mathbb{1}_s \otimes D)\left(\hat{U}_m - U_m\right) + (I_s \otimes D)\,\Theta_{m+1} + \rho_{m+1}. \tag{5.12}$$

Einsetzen dieser Gleichung in die mit der Matrix $G(M,\tau D)$ von links multiplizierte Gleichung (5.5) liefert unter Berücksichtigung von (5.7)

$$\begin{aligned} G(M,\tau D)\left(\hat{S}_{m+1} - S_{m+1}\right) &= (G(M,\tau D) + \tau\,(\mathfrak{A} \otimes D))\left(\mathbb{1}_s \otimes \left(\hat{U}_m - U_m\right) + \Theta_{m+1}\right) \\ &\quad + \tau\,(\mathfrak{A} \otimes I_{Nn})\,\rho_{m+1} \\ &= (I_s \otimes M)\left(\mathbb{1}_s \otimes \left(\hat{U}_m - U_m\right) + \Theta_{m+1}\right) + \tau\,(\mathfrak{A} \otimes I_{Nn})\,\rho_{m+1}. \end{aligned} \tag{5.13}$$

Im folgenden wird vorausgesetzt, daß die Matrix $G(M,\tau D)$ regulär ist (vgl. Bemerkung 5.8). Dann sind das gestörte und das ungestörte System (5.3) und (5.4) im linearen Fall eindeutig lösbar. Subtrahiert man (5.4a) von (5.3a), so erhält man

$$\hat{U}_{m+1} - U_{m+1} = \hat{U}_m - U_m + \tau\left(b^\top \otimes I_{Nn}\right)\left(\hat{K}_{m+1} - K_{m+1}\right) + \theta_{m+1}$$

und daraus durch Einsetzen der mit $G(M,\tau D)^{-1}$ von links multiplizierten Gleichung (5.12) und der durch (5.8) - (5.10) definierten Größen

$$\hat{U}_{m+1} - U_{m+1} = R(M,\tau D)\left(\hat{U}_m - U_m\right) + L(M,\tau D)\Theta_{m+1} + \tau J(M,\tau D)\rho_{m+1} + \theta_{m+1}, \tag{5.14}$$

wobei der Einfachheit halber die Zeitschrittweite $\tau$ von nun an stets als konstant vorausgesetzt wird, d. h. $\tau = \frac{t_e - t_0}{M_e}$.

**Bemerkung 5.2** Für singuläre $A$ folgt aus der Regularität von $G(M,\tau D)$, daß die Verfahrensmatrix $\mathfrak{A}$ regulär sein muß: Andernfalls würden von Null verschiedene Vektoren $v_1 \in \mathbb{R}^{nN}$, $v_2 \in \mathbb{R}^s$ existieren mit $Mv_1 = 0$, $\mathfrak{A}v_2 = 0$, daraus würde dann $G(M,\tau D)(v_2 \otimes v_1) = 0$ folgen. Für Beispiele von Verfahren mit regulären Verfahrensmatrizen siehe Bemerkung 2.10.
Das Konvergenzverhalten einer speziellen Klasse von Runge-Kutta-Verfahren mit singulärer Verfahrensmatrix, auch für singuläre Matrizen $A$, wird in Abschnitt 5.2.5 untersucht. □

Die Ausführung eines Rekursionsschrittes in (5.14) liefert

$$\begin{aligned}\hat{U}_{m+1} - U_{m+1} = {} & R(M,\tau D)^2 \left(\hat{U}_{m-1} - U_{m-1}\right) + L(M,\tau D)\Theta_{m+1} + \tau J(M,\tau D)\rho_{m+1} \\ & + \theta_{m+1} + R(M,\tau D)\Big(L(M,\tau D)\Theta_m + \tau J(M,\tau D)\rho_m + \theta_m\Big),\end{aligned}$$

und durch weitere Rekursion sowie aus (5.13) und (5.11) ergibt sich das folgende Lemma:

**Lemma 5.3** Es sei $G(M,\tau D)$ regulär. Für die Differenz der Lösung des gestörten Systems (5.4) und des Systems (5.3) gilt dann

$$\begin{aligned}\hat{U}_{m+1} - U_{m+1} = {} & R(M,\tau D)^{m+1}\left(\hat{U}_0 - U_0\right) + \sum_{i=0}^{m} R(M,\tau D)^i L(M,\tau D)\Theta_{m+1-i} \\ & + \tau \sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D)\rho_{m+1-i} + \sum_{i=0}^{m} R(M,\tau D)^i \theta_{m+1-i}\end{aligned}$$

und für die Differenz der Stufenwerte

$$\begin{aligned}\hat{S}_{m+1} - S_{m+1} = {} & H(M,\tau D)\left(\hat{U}_m - U_m\right) + G(M,\tau D)^{-1}\left(I_s \otimes M\right)\Theta_{m+1} \\ & + \tau G(M,\tau D)^{-1}\left(\mathfrak{A} \otimes I_{Nn}\right)\rho_{m+1}.\end{aligned}$$

□

Dieses Störungslemma wird bei den nachfolgenden Konvergenzuntersuchungen für lineare PDA-Systeme Anwendung finden.

### 5.2.2 Konvergenzuntersuchungen für lineare PDA-Systeme

In diesem Abschnitt soll, aufbauend auf den Ergebnissen des vorherigen, die Konvergenz des Verfahrens (5.1) für lineare PDA-Systeme untersucht werden.

**Definition 5.4** Sei

$$e_{m+1} = U_{\vec{h}}(t_{m+1}) - U_{m+1} \tag{5.15}$$

der globale Gesamtdiskretisierungsfehler zum Zeitpunkt $t = t_{m+1}$. Gilt dann für jeden Zeitpunkt $t$, daß

$$\|e_{m+1}\| = \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i}) + \mathcal{O}(\tau^{p^\star}) \quad \text{für} \quad (m+1)\tau = \text{konst.},\ \tau,\ h_i \to 0$$

ist, so heißt die Gesamtdiskretisierung konvergent mit der Ordnung $(p_1,\ldots,p_d,p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

Setzt man die auf das Ortsgitter eingeschränkte exakte Lösung $U_{\vec{h}}(t)$ des PDA-Systems (3.2) und ihre erste Zeitableitung $\dot{U}_{\vec{h}}(t)$ in die Verfahrensvorschrift (5.1a), (5.1b) ein, so erhält man die Residuenfehler:

**Definition 5.5** Mit

$$\delta_{m+1} = U_{\vec{h}}(t_m + \tau) - U_{\vec{h}}(t_m) - \tau \sum_{i=1}^{s} b_i \dot{U}_{\vec{h}}(t_m + c_i\tau) \tag{5.16a}$$

bzw.

$$\Delta_{m+1}^{(i)} = U_{\vec{h}}(t_m + c_i\tau) - U_{\vec{h}}(t_m) - \tau \sum_{j=1}^{s} a_{ij} \dot{U}_{\vec{h}}(t_m + c_j\tau) \tag{5.16b}$$

werden die Residuenfehler des Verfahrens bzw. der Stufenwerte des Verfahrens bezeichnet. □

Aus dieser Definition und der Gleichung (4.49) für den lokalen Ortsdiskretisierungsfehler folgt, daß die Vektoren

$$\begin{aligned}
\hat{U}_m &= U_{\vec{h}}(t_m), \quad m = 0, \ldots, M_e, \\
\hat{S}_{m+1} &= \left( U_{\vec{h}}(t_m + c_1\tau)^\top, \ldots, U_{\vec{h}}(t_m + c_s\tau)^\top \right)^\top, \quad m = 0, \ldots, M_e - 1, \\
\hat{K}_{m+1} &= \left( \dot{U}_{\vec{h}}(t_m + c_1\tau)^\top, \ldots, \dot{U}_{\vec{h}}(t_m + c_s\tau)^\top \right)^\top, \quad m = 0, \ldots, M_e - 1,
\end{aligned}$$

das System (5.4) mit den Störungen

$$\theta_{m+1} = \delta_{m+1}, \tag{5.17a}$$

$$\Theta_{m+1} = \Delta_{m+1} = \left( {\Delta_{m+1}^{(1)}}^\top, \ldots, {\Delta_{m+1}^{(s)}}^\top \right)^\top, \tag{5.17b}$$

$$\rho_{m+1} = \alpha_{\vec{h}m+1} = \left( \alpha_{\vec{h}}(t_m + c_1\tau)^\top, \ldots, \alpha_{\vec{h}}(t_m + c_1\tau)^\top \right)^\top \tag{5.17c}$$

erfüllen. Da für den gemäß (5.2) gewählten Startvektor $U_0 = U_{\vec{h}}(t_0)$ die Beziehung $e_0 = 0$ gilt, folgt für den globalen Gesamtdiskretisierungsfehler aus Lemma 5.3

$$\begin{aligned}
e_{m+1} = &\sum_{i=0}^{m} R(M, \tau D)^i L(M, \tau D) \Delta_{m+1-i} + \tau \sum_{i=0}^{m} R(M, \tau D)^i J(M, \tau D) \alpha_{\vec{h}m+1-i} \\
&+ \sum_{i=0}^{m} R(M, \tau D)^i \delta_{m+1-i}.
\end{aligned} \tag{5.18}$$

Führt man den Verbundvektor

$$\Gamma_{m+1}^{(v)} = \left( \gamma^{(v)}(t_m + c_1\tau)^\top, \ldots, \gamma^{(v)}(t_m + c_s\tau)^\top \right)^\top \tag{5.19}$$

ein, so erhält man aus (4.50) für den lokalen Ortsdiskretisierungsfehler

$$\alpha_{\vec{h}m+1} = \sum_{v=1}^{d} h_v^{p_v} (I_{sN} \otimes B_v) \Gamma_{m+1}^{(v)}. \tag{5.20}$$

Das betrachtete Runge-Kutta-Verfahren habe die Konsistenzordnung $p$. Die Zeitableitungen der exakten Lösung seien bis zur $(p+1)$-ten Ordnung beschränkt.
Eine Taylor-Entwicklung von $U_{\vec{h}}(t_m + c_i\tau)$ und $\dot{U}_{\vec{h}}(t_m + c_j\tau),\ j = 1, \ldots, s$, an der Stelle $t_m$ bis zur Ordnung $p$ liefert für den Residuenfehler $\Delta_{m+1}^{(i)}$ der $i$-ten Stufe

$$\begin{aligned}
\Delta_{m+1}^{(i)} = &U_{\vec{h}}(t_m) + \sum_{r=1}^{p} \frac{(c_i\tau)^r}{r!} U_{\vec{h}}^{(r)}(t_m) + \frac{(c_i\tau)^{p+1}}{(p+1)!} U_{\vec{h}}^{(p+1)}(\xi_0) \\
&- U_{\vec{h}}(t_m) - \tau \sum_{j=1}^{s} a_{ij} \left( \sum_{r=1}^{p} \frac{(c_j\tau)^{r-1}}{(r-1)!} U_{\vec{h}}^{(r)}(t_m) + \frac{(c_j\tau)^p}{p!} U_{\vec{h}}^{(p+1)}(\xi_j) \right) \\
= &\sum_{r=1}^{p} \left[ \frac{(c_i\tau)^r}{r!} U_{\vec{h}}^{(r)}(t_m) - \tau \sum_{j=1}^{s} a_{ij} \frac{(c_j\tau)^{r-1}}{(r-1)!} U_{\vec{h}}^{(r)}(t_m) \right] \\
&+ \tau^{p+1} \left( \frac{c_i^{p+1}}{(p+1)!} U_{\vec{h}}^{(p+1)}(\xi_0) - \sum_{j=1}^{s} a_{ij} \frac{c_j^p}{p!} U_{\vec{h}}^{(p+1)}(\xi_j) \right)
\end{aligned}$$

mit den Zwischenwerten $\xi_0, \xi_j \in [t_m, t_{m+1}],\ j = 1, \ldots, s$. Führt man den Vektor

$$\tilde{c}^i = (c_1^i, \ldots, c_s^i)^\top \tag{5.21}$$

ein, so läßt sich $\Delta_{m+1}^{(i)}$ in der Form

$$\Delta_{m+1}^{(i)} = \sum_{r=1}^{p} \frac{\tau^r}{r!} \left[\tilde{c}_i^r - r\left(\mathfrak{A}\tilde{c}^{r-1}\right)_i\right] U_{\vec{h}}^{(r)}(t_m) + \mathcal{O}(\tau^{p+1}) \tag{5.22}$$

schreiben.
Das Runge-Kutta-Verfahren besitze ferner die Stufenordnung $q$, das heißt, die vereinfachende Bedingung $C(q)$ ist erfüllt (siehe (2.17)). Dann folgt für den in (5.17b) eingeführten Verbundvektor der Residuenfehler der Stufen aus (5.22)

$$\Delta_{m+1} = \sum_{r=q+1}^{p} \frac{\tau^r}{r!} \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes U_{\vec{h}}^{(r)}(t_m) + \mathcal{O}(\tau^{p+1}). \tag{5.23}$$

**Bemerkung 5.6** Nach (5.23) besitzt der gemäß Definition 5.5 definierte Residuenfehler der Stufenwerte mindestens die Konsistenzordnung $q$ (nach Definition ist $q \leq p$). Daher rührt auch die Bezeichnung Stufenordnung. □

Durch Einsetzen der für den lokalen Ortsdiskretisierungsfehler und den Residuenfehler der Stufen gewonnenen Gleichungen (5.20) und (5.23) in die Beziehung (5.18) für den globalen Gesamtdiskretisierungsfehler ergibt sich

$$\begin{aligned} e_{m+1} = &\sum_{i=0}^{m} R(M,\tau D)^i L(M,\tau D) \left( \sum_{r=q+1}^{p} \frac{\tau^r}{r!} \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes U_{\vec{h}}^{(r)}(t_{m-i}) + \mathcal{O}(\tau^{p+1}) \right) \\ &+\tau \sum_{v=1}^{d} h_v^{p_v} \sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D)(I_{sN} \otimes B_v)\Gamma_{m+1-i}^{(v)} + \sum_{i=0}^{m} R(M,\tau D)^i \delta_{m+1-i}. \end{aligned}$$

Unter der Voraussetzung, daß für alle hinreichend kleinen $\tau$ und $\vec{h}$ und $r = q+1, \ldots, p$ Matrizen $W_{r0}(M,\tau D) \in \mathbb{R}^{Nn,Nn}$ existieren mit

$$\left(I_{Nn} - R(M,\tau D)\right) W_{r0}(M,\tau D) D = J(M,\tau D) \left(\left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes I_{Nn}\right) D \tag{5.24}$$

(vergleiche Bemerkung 5.9), erhält man unter Verwendung der in (5.10) definierten Matrix $L(M,\tau D)$

$$\begin{aligned} e_{m+1} = &\tau \sum_{v=1}^{d} h_v^{p_v} \sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D) \left(I_{sN} \otimes B_v\right) \Gamma_{m+1-i}^{(v)} \\ &+ \sum_{i=0}^{m} R(M,\tau D)^i \delta_{m+1-i} + \sum_{i=0}^{m} R(M,\tau D)^i L(M,\tau D) \mathcal{O}(\tau^{p+1}) \\ &+ \sum_{i=0}^{m} R(M,\tau D)^i \left(I_{Nn} - R(M,\tau D)\right) \sum_{r=q+1}^{p} \frac{\tau^{r+1}}{r!} W_{r0}(M,\tau D) D U_{\vec{h}}^{(r)}(t_{m-i}). \end{aligned} \tag{5.25}$$

Der letzte Summand

$$\kappa = \sum_{i=0}^{m} R(M,\tau D)^i \left(I_{Nn} - R(M,\tau D)\right) \sum_{r=q+1}^{p} \frac{\tau^{r+1}}{r!} W_{r0}(M,\tau D) D U_{\vec{h}}^{(r)}(t_{m-i})$$

wird nun in eine für die nachfolgenden Abschätzungen günstigere Form gebracht (eine ähnliche Umformung findet man in Brenner/Crouzeix/Thomée [4]):
Durch Vertauschen der Summationsreihenfolge erhält man

$$\kappa = \sum_{r=q+1}^{p} \frac{\tau^{r+1}}{r!} \sum_{i=0}^{m} R(M,\tau D)^{m-i} \left(I_{Nn} - R(M,\tau D)\right) W_{r0}(M,\tau D) D U_{\vec{h}}^{(r)}(t_i).$$

Daraus folgt

$$\kappa = \sum_{r=q+1}^{p} \frac{\tau^{r+1}}{r!} \Bigg( \sum_{i=0}^{m-1} R(M,\tau D)^{m-i} W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_i) + W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_m)$$
$$- \sum_{i=1}^{m} R(M,\tau D)^{m-i+1} W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_i) - R(M,\tau D)^{m+1} W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_0) \Bigg).$$

Durch Verschieben des Summationsindex und Zusammenfassen ergibt sich

$$\kappa = \sum_{r=q+1}^{p} \frac{\tau^{r+1}}{r!} \Bigg( \sum_{i=0}^{m-1} R(M,\tau D)^{m-i} W_{r0}(M,\tau D) D \left( U_{\vec{h}}^{(r)}(t_i) - U_{\vec{h}}^{(r)}(t_{i+1}) \right)$$
$$+ W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_m) - R(M,\tau D)^{m+1} W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_0) \Bigg)$$

und daraus

$$\kappa = \sum_{r=q+1}^{p} \frac{\tau^{r+1}}{r!} \Bigg( - \sum_{i=0}^{m-1} R(M,\tau D)^{m-i} W_{r0}(M,\tau D) \int_{t_i}^{t_{i+1}} DU_{\vec{h}}^{(r+1)}(s) ds$$
$$+ W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_m) - R(M,\tau D)^{m+1} W_{r0}(M,\tau D) DU_{\vec{h}}^{(r)}(t_0) \Bigg). \qquad (5.26)$$

Sind $\bar{N} \in \mathbb{N}$ und $O,\ K \in \mathbb{R}^{\bar{N},\bar{N}}$ beliebig, so wird für $r = q+1, \ldots, p$ und $\alpha_r \in \mathbb{R}$ unter der Voraussetzung, daß $K^{1+\alpha_r}$ existiert, mit $W_{r\alpha_r}(O,K)$ eine Matrix bezeichnet, die

$$(I_{\bar{N}} - R(O,K))\, W_{r\alpha_r}(O,K) K^{1+\alpha_r} = J(O,K) \left( \left[ \tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1} \right] \otimes I_{\bar{N}} \right) K \qquad (5.27)$$

erfüllt. Die Matrixpotenzfunktion ist dabei entsprechend Abschnitt 2.3.2 definiert.

**Bemerkung 5.7** Die Matrizen $W_{r0}$ und $W_{r\alpha_r}$ sind Verallgemeinerungen der in Ostermann/ Roche [42] für Konvergenzuntersuchungen von Runge-Kutta-Methoden für skalare parabolische Differentialgleichungen eingeführten rationalen Funktionen

$$W_r(z) = \frac{b^\top (I - z\mathfrak{A})^{-1} (\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1})}{1 - R(z)}, \quad r \geq 1.$$

□

Anstelle der Existenz von Matrizen $W_{r0}(M,\tau D)$, die (5.24) erfüllen, wird nun vorausgesetzt, daß $\alpha_r \in \mathbb{R}$, $\alpha_r \geq -1$ so gewählt werden können, daß für alle hinreichend kleinen $\tau$ und $\vec{h}$ für $r = q+1, \ldots, p$ Matrizen $W_{r\alpha_r}(M,\tau D)$ gemäß (5.27) existieren. Dann gilt entsprechend (5.26)

$$\kappa = \sum_{r=q+1}^{p} \frac{\tau^{r+1+\alpha_r}}{r!} \Bigg( - \sum_{i=0}^{m-1} R(M,\tau D)^{m-i} W_{r\alpha_r}(M,\tau D) \int_{t_i}^{t_{i+1}} D^{1+\alpha_r} U_{\vec{h}}^{(r+1)}(s) ds$$
$$+ W_{r\alpha_r}(M,\tau D) D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t_m) - R(M,\tau D)^{m+1} W_{r\alpha_r}(M,\tau D) D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t_0) \Bigg).$$

Einsetzen in die Gleichung (5.25) für den globalen Gesamtdiskretisierungsfehler liefert

$$e_{m+1} = \tau \sum_{v=1}^{d} h_v^{p_v} \sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D) \left( I_{sN} \otimes B_v \right) \Gamma_{m+1-i}^{(v)}$$

$$+\sum_{i=0}^{m} R(M,\tau D)^i \delta_{m+1-i} + \sum_{i=0}^{m} R(M,\tau D)^i L(M,\tau D)\mathcal{O}(\tau^{p+1})$$
$$+\sum_{r=q+1}^{p} \frac{\tau^{r+1+\alpha_r}}{r!}\Bigg( -\sum_{i=0}^{m-1} \int_{t_i}^{t_{i+1}} R(M,\tau D)^{m-i} W_{r\alpha_r}(M,\tau D) D^{1+\alpha_r} U_{\vec{h}}^{(r+1)}(s)ds$$
$$+W_{r\alpha_r}(M,\tau D) D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t_m) - R(M,\tau D)^{m+1} W_{r\alpha_r}(M,\tau D) D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t_0)\Bigg). \tag{5.28}$$

Die sich daraus durch direkte Abschätzung ergebenden Konvergenzbedingungen wären nicht unmittelbar zu überprüfen, weil, wie in Abschnitt 4.2, die Dimensionen der dann zu berechnenden Matrizen für gegen Null gehende Ortsschrittweiten gegen Unendlich gehen. Deshalb sollen auch hier durch Diagonalisierung die Normen dieser Matrizen auf die Normen von Matrizen der Dimension $n$ des Ausgangssystems zurückgeführt werden. Da in Abschnitt 5.2.3 durch Weierstraß-Kronecker-Transformation eine weitere Diagonalisierung erfolgt, werden die entsprechenden Gleichungen zunächst mit allgemeinen Matrizen aufgeschrieben.
Es seien $P$ und $S$ reguläre $(\bar{N}\times\bar{N})$-Matrizen mit

$$POS = \mathrm{diag}_k\{A_k\}=\mathrm{diag}\{A_1,\ldots,A_{\bar{n}}\},\quad PKS = \mathrm{diag}_k\{C_k\}=\mathrm{diag}\{C_1,\ldots,C_{\bar{n}}\}, \tag{5.29a}$$
$$A_k, C_k \in \mathbb{C}^{n_k,n_k},\quad k=1,\ldots,\bar{n},$$

und

$$\bar{N} = \sum_{k=1}^{\bar{n}} n_k. \tag{5.29b}$$

Dann folgen für die in (5.7) - (5.10) definierten Größen

$$\begin{aligned}
(I_s\otimes P)G(O,K)(I_s\otimes S) &= G(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\}), && (5.30a)\\
S^{-1}J(O,K)(I_s\otimes P^{-1}) &= (b^\top\otimes I_{\bar{N}})(I_s\otimes S^{-1})G(O,K)^{-1}(I_s\otimes P^{-1})\\
&= (b^\top\otimes I_{\bar{N}})G(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\})^{-1}\\
&= J(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\}), && (5.30b)\\
S^{-1}R(O,K)S &= I_{\bar{N}} + \tau S^{-1}J(O,K)(I_s\otimes P^{-1})(\mathbb{1}_s\otimes(PKS))\\
&= I_{\bar{N}} + \tau J(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\})(\mathbb{1}_s\otimes\mathrm{diag}_k\{C_k\})\\
&= R(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\}), && (5.30c)\\
S^{-1}L(O,K)(I_s\otimes S) &= L(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\}). && (5.30d)
\end{aligned}$$

**Bemerkung 5.8** Wegen (5.30a) und (4.58) gilt

$$\det G(M,\tau D) = \prod_{\vec{k}} \det G(A,\tau D_{\vec{k}}),$$

$G(M,\tau D)$ ist also genau dann regulär, wenn dies für alle Matrizen $G(A,\tau D_{\vec{k}})$ gilt. □

Sind

$$\mathfrak{A}_1 = (a_{1ij})_{i,j=1,\ldots,s},\quad \mathfrak{A}_2 = (a_{2ij})_{i,j=1,\ldots,s} \in \mathbb{C}^{s,s}$$

und ist für alle $k$

$$(\mathfrak{A}_1\otimes A_k + \mathfrak{A}_2\otimes C_k)$$

invertierbar, dann sei

$$(\mathfrak{A}_1\otimes A_k + \mathfrak{A}_2\otimes C_k)^{-1} =: T^{(k)} = \left(T_{ij}^{(k)}\right)_{i,j=1,\ldots,s} \tag{5.31}$$

mit $(T_{ij}^{(k)}) \in \mathbb{C}^{n_k,n_k}$, das heißt, $T^{(k)}$ wird in $s \times s$ quadratische Blockmatrizen der Dimension $n_k$ unterteilt. Damit gilt

$$\begin{aligned}
&(\mathfrak{A}_1 \otimes \operatorname{diag}_k\{A_k\} + \mathfrak{A}_2 \otimes \operatorname{diag}_k\{C_k\}) \left(\operatorname{diag}_k\{T_{ij}^{(k)}\}\right)_{i,j=1,\dots,s} \\
&= (\operatorname{diag}_k\{a_{1ij}A_k + a_{2ij}C_k\})_{i,j=1,\dots,s} \left(\operatorname{diag}_k\{T_{ij}^{(k)}\}\right)_{i,j=1,\dots,s} \\
&= \left(\operatorname{diag}_k\left\{\sum_{l=1}^{s} (a_{1il}A_k + a_{2il}C_k)\, T_{lj}^{(k)}\right\}\right)_{i,j=1,\dots,s} = I_{\bar{N}},
\end{aligned}$$

also

$$(\mathfrak{A}_1 \otimes \operatorname{diag}_k\{A_k\} + \mathfrak{A}_2 \otimes \operatorname{diag}_k\{C_k\})^{-1} = \left(\operatorname{diag}_k\{T_{ij}^{(k)}\}\right)_{i,j=1,\dots,s}. \tag{5.32}$$

Setzt man

$$G(A_k, C_k)^{-1} =: \left(\left(G(A_k, C_k)^{-1}\right)_{ij}\right)_{i,j=1,\dots,s},$$

dann folgen deshalb

$$G(\operatorname{diag}_k\{A_k\}, \operatorname{diag}_k\{C_k\})^{-1} = \left(\operatorname{diag}_k\left\{\left(G(A_k, C_k)^{-1}\right)_{ij}\right\}\right)_{i,j=1,\dots,s}, \tag{5.33a}$$

$$\begin{aligned}
J(\operatorname{diag}_k\{A_k\}, \operatorname{diag}_k\{C_k\}) &= \left(b^\top \otimes I_{\bar{N}}\right) \left(\operatorname{diag}_k\left\{\left(G(A_k, C_k)^{-1}\right)_{ij}\right\}\right)_{i,j=1,\dots,s} \\
&= \left(\operatorname{diag}_k\left\{\left(\left(b^\top \otimes I_{n_k}\right) G(A_k, C_k)^{-1}\right)_j\right\}\right)_{j=1,\dots,s} \\
&= \left(\operatorname{diag}_k\left\{\left(J(A_k, C_k)\right)_j\right\}\right)_{j=1,\dots,s},
\end{aligned} \tag{5.33b}$$

$$\begin{aligned}
R(\operatorname{diag}_k\{A_k\}, \operatorname{diag}_k\{C_k\}) &= I_{\bar{N}} + \left(\operatorname{diag}_k\left\{\left(J(A_k, C_k)\right)_j\right\}\right)_{j=1,\dots,s} (\mathbb{1}_s \otimes \operatorname{diag}_k C_k) \\
&= \operatorname{diag}_k\left\{I_{n_k} + J(A_k, C_k)\,(\mathbb{1}_s \otimes C_k)\right\} \\
&= \operatorname{diag}_k\{R(A_k, C_k)\},
\end{aligned} \tag{5.33c}$$

$$\begin{aligned}
L(\operatorname{diag}_k\{A_k\}, \operatorname{diag}_k\{C_k\}) &= \left(\operatorname{diag}_k\left\{\left(J(A_k, C_k)\,(I_s \otimes C_k)\right)_j\right\}\right)_{j=1,\dots,s} \\
&= \left(\operatorname{diag}_k\left\{\left(L(A_k, C_k)\right)_j\right\}\right)_{j=1,\dots,s}.
\end{aligned} \tag{5.33d}$$

Nach (4.58) können in (5.29a) $P = Q^{-1}$, $O = M$, $S = Q$ und $K = \tau D$ gewählt werden, und man erhält aus (5.30) und (5.33)

$$(I_s \otimes Q^{-1}) G(M, \tau D)^{-1} (I_s \otimes Q) = \left(\operatorname{diag}_{\vec{k}}\left\{\left(G(A, \tau D_{\vec{k}})^{-1}\right)_{ij}\right\}\right)_{i,j=1,\dots,s}, \tag{5.34}$$

$$Q^{-1} J(M, \tau D)(I_s \otimes Q) = \left(\operatorname{diag}_{\vec{k}}\left\{\left(J(A, \tau D_{\vec{k}})\right)_j\right\}\right)_{j=1,\dots,s}, \tag{5.35}$$

$$Q^{-1} R(M, \tau D) Q = \operatorname{diag}_{\vec{k}}\left\{R(A, \tau D_{\vec{k}})\right\}, \tag{5.36}$$

$$Q^{-1} L(M, \tau D)\,(I_s \otimes Q) = \left(\operatorname{diag}_{\vec{k}}\left\{\left(L(A, \tau D_{\vec{k}})\right)_j\right\}\right)_{j=1,\dots,s}. \tag{5.37}$$

Nach (2.24) sind Matrixfunktionen verträglich mit Ähnlichkeitstransformationen, so daß sich aus (4.58) für alle $\alpha_r$, für die $D^{\alpha_r}$ existiert,

$$Q^{-1} D^{\alpha_r} Q = \operatorname{diag}_{\vec{k}}\{D_{\vec{k}}^{\alpha_r}\} \tag{5.38}$$

ergibt. Durch Einsetzen von (5.36), (5.38) und (5.35) in (5.27) folgt für die Matrizen $W_{r\alpha_r}(M, \tau D)$

$$\begin{aligned}
&\operatorname{diag}_{\vec{k}}\{I_n - R(A, \tau D_{\vec{k}})\} \left(Q^{-1} W_{r\alpha_r}(M, \tau D) Q\right) \tau^{\alpha_r} \operatorname{diag}_{\vec{k}}\{D_{\vec{k}}^{1+\alpha_r}\} \\
&= \operatorname{diag}_{\vec{k}}\{J(A, \tau D_{\vec{k}}) \left(\left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes I_n\right) D_{\vec{k}}\}
\end{aligned}$$

und daraus, daß $W_{r\alpha_r}(M,\tau D)$ genau dann existiert, wenn für alle $\vec{k}$ analog zu (5.27) Matrizen $W_{r\alpha_r}(A,\tau D_{\vec{k}})$ existieren, in diesem Fall kann

$$W_{r\alpha_r}(M,\tau D)=Q\mathrm{diag}_{\vec{k}}\left\{W_{r\alpha_r}(A,\tau D_{\vec{k}})\right\}Q^{-1} \tag{5.39}$$

gesetzt werden.

**Bemerkung 5.9** Ist $(I_{\bar{N}} - R(O,K))$ regulär, so folgt aus der $R(O,K)$ definierenden Gleichung (5.9) auch die Regularität von $K$, und es gelten für $r = q+1,\ldots,p$

$$W_{r0}(O,K) = (I_{\bar{N}} - R(O,K))^{-1} J(O,K)\left(\left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right]\otimes I_{\bar{N}}\right) \tag{5.40a}$$

und

$$W_{r\alpha_r}(O,K) = W_{r0}(O,K)K^{-\alpha_r}, \tag{5.40b}$$

die Matrix $W_{r\alpha_r}(O,K)$ existiert damit in diesem Fall für alle $\alpha_r \in \mathbb{R}$.
Aus (5.36) folgt, daß $(I_{Nn} - R(M,\tau D))$ genau dann regulär ist, wenn dies für alle Matrizen $(I_n - R(A,\tau D_{\vec{k}}))$ gilt. Im regulären Fall erfüllen die durch (5.40) definierten Matrizen $W_{r\alpha_r}(M,\tau D)$ und $W_{r\alpha_r}(A,\tau D_{\vec{k}})$ nach (5.35), (5.36) und (5.38)

$$\begin{aligned} W_{r\alpha_r}(M,\tau D) &= QQ^{-1}W_{r0}(M,\tau D)QQ^{-1}(\tau D)^{-\alpha_r}QQ^{-1} \\ &= Q\mathrm{diag}_{\vec{k}}\{(I_n - R(A,\tau D_{\vec{k}}))^{-1}J(A,\tau D_{\vec{k}})\left(\left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right]\otimes I_n\right)(\tau D_{\vec{k}})^{-\alpha_r}\}Q^{-1} \end{aligned}$$

und damit ebenfalls (5.39).
Die Existenz von $W_{r(-1)}(A,\tau D_{\vec{k}})$ kann unter der Voraussetzung, daß der Realteil der in (4.68c) eingeführten $\kappa_{\vec{k}i}$ nicht positiv ist, auch bei singulärer Matrix $(I_n - R(A,\tau D_{\vec{k}}))$ zum Beispiel für die Radau-IIA- und Lobatto-IIIC-Verfahren gesichert werden, vgl. Abschnitt 5.2.3. □

Die Bestimmung von $W_{r\alpha_r}(A,\tau D_{\vec{k}})$ bei singulärer Matrix $(I_n - R(A,\tau D_{\vec{k}}))$ wird im folgenden Beispiel gezeigt:

**Beispiel 5.10** Gegeben seien die Verfahrensmatrix $\mathfrak{A}$, der Wichtungsvektor $b$ und der Knotenvektor $c$ des dreistufigen Radau-IIA-Verfahrens (siehe Abschnitt 6.2) sowie das PDA-System

$$u_t + \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} u_{xx} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} u = f.$$

Dann gilt nach (4.59) mit $d=1$ und $\lambda_k{=}\lambda_{1,k}$

$$D_k = \begin{pmatrix} \lambda_k & 0 \\ -1 & 0 \end{pmatrix}$$

und nach (2.24) unter der Voraussetzung $\lambda_k \neq 0$ für $\alpha_r \neq -1$

$$D_k^{1+\alpha_r} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\lambda_k} & \frac{1}{\lambda_k} \end{pmatrix}\begin{pmatrix} \lambda_k & 0 \\ 0 & 0 \end{pmatrix}^{1+\alpha_r}\begin{pmatrix} 1 & 0 \\ 1 & \lambda_k \end{pmatrix} = \begin{pmatrix} \lambda_k^{1+\alpha_r} & 0 \\ -\lambda_k^{\alpha_r} & 0 \end{pmatrix}.$$

Aus (5.7) - (5.9) erhält man

$$\begin{aligned} I_2 - R(A,\tau D_k) &= -\tau J(A,\tau D_k)(\mathbb{1}_3\otimes D_k) = -\tau(b^\top\otimes I_2)(I_3\otimes A - \tau\mathfrak{A}\otimes D_k)^{-1}(\mathbb{1}_3\otimes D_k) \\ &= \frac{(6\tau\lambda_k - \tau^2\lambda_k^2 - 60)\tau}{-60 + 36\tau\lambda_k - 9\tau^2\lambda_k^2 + \tau^3\lambda_k^3}\begin{pmatrix} -\lambda_k & 0 \\ 1 & 0 \end{pmatrix} \end{aligned}$$

und

$$J(A,\tau D_k)\left(\left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right]\otimes D_k\right) = \frac{3}{5}\,\frac{j_r}{-60 + 36\tau\lambda_k - 9\tau^2\lambda_k^2 + \tau^3\lambda_k^3}\begin{pmatrix} \lambda_k & 0 \\ -1 & 0 \end{pmatrix}$$

mit

$$j_r = \begin{cases} \frac{\tau\lambda_k}{2} & : \ r = 4 \\ \frac{7\tau\lambda_k - 5}{5} & : \ r = 5 \end{cases} .$$

Mit dem Ansatz

$$W_{r\alpha_r}(A, \tau D_k) = \begin{pmatrix} a_r & b_r \\ c_r & d_r \end{pmatrix}$$

folgt aus (5.27) für $\alpha_r \neq -1$

$$(\tau\lambda_k)^{\alpha_r}(6\tau\lambda_k - \tau^2\lambda_k^2 - 60) \begin{pmatrix} -\lambda_k(a_r\lambda_k - b_r) & 0 \\ a_r\lambda_k - b_r & 0 \end{pmatrix} = \frac{3}{5} j_r \begin{pmatrix} \lambda_k & 0 \\ -1 & 0 \end{pmatrix}.$$

Diese Gleichung ist zum Beispiel für $a_r = c_r = d_r = 0$, $b_r = \frac{3}{5} \ \frac{1}{(\tau\lambda_k)^{\alpha_r}} \ \frac{j_r}{6\tau\lambda_k - \tau^2\lambda_k^2 - 60}$ erfüllt. Im Fall $\alpha_r = -1$ kann man analog vorgehen. Für $\lambda_k < 0$ ist $\|W_{r\alpha_r}(A, \tau D_k)\|$ damit für $\alpha_4 \in [-1, 1]$ und $\alpha_5 \in [-1, 0]$ für alle $\tau$ und $h$ beschränkt. □

Insgesamt erhält man für den globalen Gesamtdiskretisierungsfehler durch Diagonalisierung der in (5.28) auftretenden Matrizen gemäß (5.35) - (5.39) die Darstellung

$$\begin{aligned} e_{m+1} = \tau \sum_{v=1}^{d} h_v^{p_v} Q \sum_{i=0}^{m} \left( \mathrm{diag}_{\vec{k}} \left\{ \left( R(A, \tau D_{\vec{k}})^i J(A, \tau D_{\vec{k}}) \right)_j B_v \right\} \right)_{j=1,\ldots,s} \left( I_s \otimes Q^{-1} \right) \Gamma_{m+1-i}^{(v)} \\ + Q \sum_{i=0}^{m} \mathrm{diag}_{\vec{k}} \left\{ R(A, \tau D_{\vec{k}})^i \right\} Q^{-1} \delta_{m+1-i} \\ + Q \sum_{i=0}^{m} \left( \mathrm{diag}_{\vec{k}} \left\{ \left( R(A, \tau D_{\vec{k}})^i L(A, \tau D_{\vec{k}}) \right)_j \right\} \right)_{j=1,\ldots,s} (I_s \otimes Q^{-1}) \mathcal{O}(\tau^{p+1}) \\ + Q \sum_{r=q+1}^{p} \frac{\tau^{r+1+\alpha_r}}{r!} \Bigg( \mathrm{diag}_{\vec{k}} \left\{ W_{r\alpha_r}(A, \tau D_{\vec{k}}) \right\} Q^{-1} D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t_m) \\ - \sum_{i=0}^{m-1} \int_{t_i}^{t_{i+1}} \mathrm{diag}_{\vec{k}} \left\{ R(A, \tau D_{\vec{k}})^{m-i} W_{r\alpha_r}(A, \tau D_{\vec{k}}) \right\} Q^{-1} D^{1+\alpha_r} U_{\vec{h}}^{(r+1)}(s) ds \\ - \mathrm{diag}_{\vec{k}} \left\{ R(A, \tau D_{\vec{k}})^{m+1} W_{r\alpha_r}(A, \tau D_{\vec{k}}) \right\} Q^{-1} D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t_0) \Bigg). \end{aligned} \tag{5.41}$$

Analog zu Gleichung (5.22) für den Residuenfehler der Stufen erhält man für den Residuenfehler des Verfahrens

$$\delta_{m+1-i} = \sum_{r=1}^{p} \frac{\tau^r}{r!} \left[ 1 - r b^\top \tilde{c}^{r-1} \right] U_{\vec{h}}^{(r)}(t_{m-i}) + \mathcal{O}(\tau^{p+1}). \tag{5.42}$$

Da das Runge-Kutta-Verfahren die Konsistenzordnung $p$ hat, gelten die Konsistenzbedingungen (2.16), und es folgt

$$\delta_{m+1-i} = \mathcal{O}(\tau^{p+1}).$$

Mit Bemerkung 5.8 sowie den für die Spektralnorm geltenden Beziehungen (4.65) und (4.66) ergibt sich damit unter der Voraussetzung, daß die $(p+1)$-ten Ableitungen der exakten Lösung nach der Zeit und für $p_i = 1$ die dritten und für $p_i = 2$ die vierten Ableitungen der exakten Lösung nach $x_i$ beschränkt sind, $i = 1, \ldots, d$, der folgende Konvergenzsatz:

**Satz 5.11** Sei $p^\star \in [1, \min(p, r + 1 + \alpha_r : \ r = q + 1, \ldots, p)]$ mit $\alpha_r \in \mathbb{R}$, $\alpha_r \geq -1$, die größte reelle Zahl, so daß mit $\bar{M} \in \mathbb{N}$ für $\tau \to 0$ $(M_e \to \infty)$, $h_v \to 0$ $(N_v \to \infty)$, $v = 1, \ldots, d$, $r = q + 1, \ldots, p$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die folgenden Voraussetzungen erfüllt sind:

(a) $G(A, \tau D_{\vec{k}})$ ist regulär,

(b) $W_{r\alpha_r}(A, \tau D_{\vec{k}})$ existiert gemäß (5.27),

(c) für $i = 0, \ldots, \bar{M}$ sind $\|\tau R(A, \tau D_{\vec{k}})^i J(A, \tau D_{\vec{k}}) (I_s \otimes B_v)\|$, $\|R(A, \tau D_{\vec{k}})^i \tau^{p+1-p^\star}\|$ und $\|R(A, \tau D_{\vec{k}})^i L(A, \tau D_{\vec{k}}) \tau^{p+1-p^\star}\|$ beschränkt,

(d) für $i = \bar{M}+1, \ldots, M_e - 1$ sind $\|R(A, \tau D_{\vec{k}})^i J(A, \tau D_{\vec{k}}) (I_s \otimes B_v)\|$, $\|R(A, \tau D_{\vec{k}})^i \tau^{p-p^\star}\|$ und $\|R(A, \tau D_{\vec{k}})^i L(A, \tau D_{\vec{k}}) \tau^{p-p^\star}\|$ beschränkt,

(e) $\|\tau^{r+1+\alpha_r-p^\star} R(A, \tau D_{\vec{k}})^i W_{r\alpha_r}(A, \tau D_{\vec{k}})\|$ ist für $i = 0, \ldots, M_e$ beschränkt,

(f) $\|D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t)\|$ und $\|D^{1+\alpha_r} U_{\vec{h}}^{(r+1)}(t)\|$ sind für $t \in [t_0, t_e]$ beschränkt.

Dann ist das Diskretisierungsverfahren (5.3) für lineare Systeme konvergent mit der Ordnung $(p_1, \ldots, p_d, p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

**Bemerkung 5.12** Die Terme in den Voraussetzungen (c) und (d) in Satz 5.11 unterscheiden sich nur um den Faktor $\tau$, weil zum Beispiel $\sum_{i=0}^{m} \max_{\vec{k}} \|R(A, \tau D_{\vec{k}})^i\| \mathcal{O}(\tau^{p+1})$ abgeschätzt wird durch

$$\left( \sum_{i=0}^{\bar{M}} \max_{\vec{k}} \|R(A, \tau D_{\vec{k}})^i \tau^{p+1-p^\star}\| + (t_e - t_0) \max_{i=\bar{M}+1}^{M_e-1} \max_{\vec{k}} \|R(A, \tau D_{\vec{k}})^i \tau^{p-p^\star}\| \right) \mathcal{O}(\tau^{p^\star}).$$

□

**Bemerkung 5.13** Ist bei Gültigkeit der übrigen Voraussetzungen des Satzes 5.11 die Beschränktheit der Normen nur unter einer Bedingung an $\tau$ und $\vec{h}$, zum Beispiel von der Form

$$c_0 \leq \frac{\tau}{h_i^2} \leq c_1, \quad i = 1, \ldots, d,$$

erfüllt, so ist das Verfahren bedingt konvergent mit der Ordnung $(p_1, \ldots, p_d, p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

Die Beschränktheit der Matrixnormen in Satz 5.11 ist für ein gegebenes Runge-Kutta-Verfahren und ein gegebenes lineares PDA-System nach Bemerkung 4.11 einfach überprüfbar. Es bleibt für $l = q+1, \ldots, p+1$ zu untersuchen, für welche $\alpha \geq -1$ die Matrixnorm $\|D^{1+\alpha} U_{\vec{h}}^{(l)}(t)\|$ beschränkt bleibt. Für $\alpha = -1$ ist diese Voraussetzung stets erfüllt. Genügt $B_i \frac{\partial^l u}{\partial t^l}$, $i = 1, \ldots, d$, homogenen Dirichlet-, homogenen Neumann- oder periodischen Randbedingungen, so gilt nach den Taylor-Entwicklungen (4.6), (4.35) und (4.37)

$$DU_{\vec{h}}^{(l)}(t) = \sum_{i=1}^{d} (I_N \otimes B_i) \left( U_{\vec{h}x_i x_i}^{(l)}(t) + U_{\vec{h}x_i}^{(l)}(t) \right) + \sum_{i=1}^{d} \mathcal{O}(h_i^{p_i})$$

mit

$$U_{\vec{h}x_i}(t) = \left( \frac{\partial}{\partial x_i} u(t, \vec{x}_{1\ldots1}), \ldots, \frac{\partial}{\partial x_i} u(t, \vec{x}_{N_1 \ldots N_d}) \right)$$

und $U_{\vec{h}x_i x_i}(t)$ entsprechend, und es kann $\alpha = 0$ gewählt werden. Diese Abschätzungen können weiter verfeinert werden. Dazu wird der folgende Satz benötigt:

**Satz 5.14** Sei $\Omega_Q = (a_1, b_1) \times \ldots \times (a_d, b_d) \subset \mathbb{R}^d$ ein offener $d$-dimensionaler Quader. Existieren für eine Funktion $\mathfrak{f} : \overline{\Omega}_Q \to \mathbb{R}$ die partiellen Ableitungen $d$-ter Ordnung und sind stetig, so ist für alle Zerlegungen $\mathfrak{Z}$ von $\overline{\Omega}_Q$ mit den Teilpunkten

$$\{x_{11}, \ldots, x_{1n_1}\} \times \ldots \times \{x_{d1}, \ldots, x_{dn_d}\},$$

$$x_{ji} < x_{j(i+1)}, \quad i = 1, \ldots, n_j - 1, \quad x_{j1} = a_j, \quad x_{jn_j} = b_j, \quad j = 1, \ldots, d,$$

die Summe

$$S_{\mathfrak{Z}} = \sum_{\substack{k_i=1,\ldots,n_i-1 \\ i=1,\ldots,d}} |\mathfrak{f}_{[k_1,\ldots,k_d]}|$$

mit

$$\mathfrak{f}_{[k_1,\ldots,k_d]} = \sum_{\substack{\nu_i=0,1 \\ i=1,\ldots,d}} (-1)^{\sum_{i=1}^{d} \nu_i} \mathfrak{f}(x_{1(k_1+\nu_1)}, \ldots, x_{d(k_d+\nu_d)})$$

beschränkt. □

**Bemerkung 5.15** Ist $d = 1$, so entspricht die obige Eigenschaft der Funktion $\mathfrak{f}$ dem für Funktionen über einem eindimensionalen abgeschlossenen Intervall in $\mathbb{R}$ bekannten Begriff der beschränkten Schwankung. □

Zum Beweis von Satz 5.14 wird zunächst die folgende Verallgemeinerung des Satzes von Rolle auf $d \geq 2$ Dimensionen bewiesen:

**Lemma 5.16** Existieren die partiellen Ableitungen $d$-ter Ordnung von $\mathfrak{f}$ in $\overline{\Omega}_Q$ und sind stetig und gilt

$$\sigma_{\mathfrak{f}} = (-1)^d \sum_{\substack{\nu_i=0,1 \\ i=1,\ldots,d}} (-1)^{\sum_{i=1}^{d} \nu_i} \mathfrak{f}(a_1 + \nu_1(b_1 - a_1), \ldots, a_d + \nu_d(b_d - a_d)) = 0, \tag{5.43}$$

so existiert $\vec{\xi} \in \Omega_Q$ mit $\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{\xi}) = 0$. □

**Beweis:** Sei $\mathfrak{I} = \int_{a_d}^{b_d} \ldots \int_{a_1}^{b_1} \frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{x})\, dx_1 \ldots dx_d$. Im Fall $d = 1$ gilt

$$\mathfrak{I} = \int_{a_1}^{b_1} \frac{\partial \mathfrak{f}}{\partial x_1}\, dx_1 = \mathfrak{f}(b_1) - \mathfrak{f}(a_1) = \sigma_{\mathfrak{f}}.$$

Für $d > 1$ folgt durch vollständige Induktion ebenfalls $\mathfrak{I} = \sigma_{\mathfrak{f}}$:

$$\begin{aligned}
\mathfrak{I} &= \int_{a_d}^{b_d} \ldots \int_{a_1}^{b_1} \frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{x})\, dx_1 \ldots dx_d \\
&= \int_{a_d}^{b_d} (-1)^{d-1} \sum_{\substack{\nu_i=0,1 \\ i=1,\ldots,d-1}} (-1)^{\sum_{i=1}^{d-1} \nu_i} \mathfrak{f}_{x_d}(a_1 + \nu_1(b_1 - a_1), \ldots, a_{d-1} + \nu_{d-1}(b_{d-1} - a_{d-1}), x_d)\, dx_d \\
&= (-1)^{d-1} \sum_{\substack{\nu_i=0,1 \\ i=1,\ldots,d-1}} (-1)^{\sum_{i=1}^{d-1} \nu_i} \Big( \mathfrak{f}(a_1 + \nu_1(b_1 - a_1), \ldots, a_{d-1} + \nu_{d-1}(b_{d-1} - a_{d-1}), b_d) \\
&\qquad - \mathfrak{f}(a_1 + \nu_1(b_1 - a_1), \ldots, a_{d-1} + \nu_{d-1}(b_{d-1} - a_{d-1}), a_d) \Big) \\
&= \sigma_{\mathfrak{f}}.
\end{aligned}$$

Damit gilt $\mathfrak{I} = \sigma_{\mathfrak{f}} = 0$. Wäre $\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{x}) > 0$ bzw. $\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{x}) < 0$ für alle $\vec{x} \in \Omega_Q$, so wäre

$$\mathfrak{I} = \int\limits_{\overline{\Omega}_Q} \frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{x})\, d\vec{x} > 0 \quad \text{bzw.} \quad \mathfrak{I} < 0.$$

Wegen der Stetigkeit von $\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}$ folgt daraus die Behauptung. □

Mit diesem Lemma erhält man folgende Verallgemeinerung des Mittelwertsatzes der Differentialrechnung:

**Lemma 5.17** Existieren die partiellen Ableitungen $d$-ter Ordnung von $\mathfrak{f}$ in $\overline{\Omega}_Q$ und sind stetig, dann existiert $\vec{\xi} \in \Omega_Q$ mit

$$\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{\xi}) = \frac{\sigma_{\mathfrak{f}}}{\prod\limits_{i=1}^{d} (b_i - a_i)}$$

mit $\sigma_{\mathfrak{f}}$ aus (5.43). □

**Beweis:** Es sei $g(\vec{x}) = \sigma_{\mathfrak{f}} \prod\limits_{i=1}^{d} \frac{x_i - a_i}{b_i - a_i}$. Dann erfüllt $(\mathfrak{f}(\vec{x}) - g(\vec{x}))$ wegen $\sigma_g = \sigma_{\mathfrak{f}}$ die Voraussetzungen des Lemmas 5.16, weshalb $\vec{\xi} \in \Omega_Q$ mit

$$\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{\xi}) = \frac{\partial^d g}{\partial x_1 \ldots \partial x_d}(\vec{\xi}) = \frac{\sigma_{\mathfrak{f}}}{\prod\limits_{i=1}^{d} (b_i - a_i)}$$

existiert. □

**Bemerkung 5.18** Eine Abschwächung der Voraussetzungen der Lemmata 5.16 und 5.17 ist möglich, wird hier aber nicht benötigt. □

Damit läßt sich nun Satz 5.14 leicht beweisen:

**Beweis:** Ist $M' = \max\limits_{\vec{x} \in \overline{\Omega}_Q} |\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{x})|$, so gilt nach Lemma 5.17

$$\begin{aligned} S_{\mathfrak{Z}} &= \sum_{\substack{k_i = 1, \ldots, n_i - 1 \\ i = 1, \ldots, d}} |\mathfrak{f}_{[k_1, \ldots, k_d]}| = \sum_{\substack{k_i = 1, \ldots, n_i - 1 \\ i = 1, \ldots, d}} \prod_{i=1}^{d} (x_{i(k_i+1)} - x_{ik_i}) \left| \frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}(\vec{\xi}_{k_1 \ldots k_d}) \right| \\ &\leq M' \sum_{\substack{k_i = 1, \ldots, n_i - 1 \\ i = 1, \ldots, d}} \prod_{i=1}^{d} (x_{i(k_i+1)} - x_{ik_i}) = M' \prod_{i=1}^{d} (b_i - a_i). \end{aligned}$$

□

Nun kann folgender Satz gezeigt werden:

**Satz 5.19** Gilt mit einer von $\vec{h}$ unabhängigen Konstanten $C_1$ für alle hinreichend kleinen $\vec{h}$

$$|(D_{\vec{k}}^{\beta})_{ij}| \leq C_1 \left( 1 + \sum_{v=1}^{d} |\lambda_{v,k_v}|^{\beta} \right), \quad i, j = 1, \ldots, n, \tag{5.44}$$

für alle $\vec{k}$ und $\lambda_{v,k_v}$ aus (4.47) und ist im PDA-System (3.2) $r_i = 0$ für $i \in M_N$, so ist für alle Funktionen $\mathfrak{f} : [t_0, t_e] \times \overline{\Omega} \to \mathbb{R}^n$, deren Ableitung $\frac{\partial^d \mathfrak{f}}{\partial x_1 \ldots \partial x_d}$ in $\overline{\Omega}$ existiert und stetig ist, für die zugehörige Gitterfunktion

$$\mathfrak{F}(t) = \left( \mathfrak{f}_{1,1,\ldots,1}(t)^{\top}, \ldots, \mathfrak{f}_{N_1,1,\ldots,1}(t)^{\top}, \mathfrak{f}_{1,2,\ldots,1}(t)^{\top}, \ldots, \mathfrak{f}_{N_1,\ldots,N_d}(t)^{\top} \right)^{\top}$$

mit $\mathfrak{f}_{\vec{k}}(t) = \mathfrak{f}(t, \vec{x}_{\vec{k}})$

$$\|D^\beta \mathfrak{F}(t)\| \le C_\beta \quad \text{für alle} \quad \beta < \frac{1}{4} \tag{5.45}$$

mit von $\vec{h}$ unabhängigen Konstanten $C_\beta$. □

**Beweis:** Nach (5.38) gilt

$$\|D^\beta \mathfrak{F}(t)\| = \|Q \mathrm{diag}_{\vec{k}}\{D_{\vec{k}}^\beta\} Q^{-1} \mathfrak{F}(t)\|,$$

wobei entsprechend (4.57) $Q = S_{P_d} \otimes \ldots \otimes S_{P_1} \otimes I_n$ ist. Mit den für $S_{P_i}$ geltenden Abschätzungen (4.23), (4.29) und (4.43) erhält man daraus

$$\|D^\beta \mathfrak{F}(t)\| \le C_2 \|\mathrm{diag}_{\vec{k}}\{D_{\vec{k}}^\beta\} \left( S_{P_d}^{-1} \otimes \ldots \otimes S_{P_1}^{-1} \otimes I_n \right) \mathfrak{F}(t)\|$$

mit $C_2 = \frac{3}{2} e^{\sum\limits_{i \in M_d} l_i |r_i|}$ bzw.

$$\begin{aligned} \|D^\beta \mathfrak{F}(t)\|^2 &\le C_2^2 \prod_{i=1}^d h_i \sum_{\vec{k}_1} \|D_{\vec{k}_1}^\beta \sum_{\vec{k}_2} \prod_{i=1}^d (S_{P_i}^{-1})_{k_{1i} k_{2i}} \mathfrak{f}_{\vec{k}_2}\|^2 \\ &\le C_2^2 \prod_{i=1}^d h_i \sum_{\vec{k}_1} \|D_{\vec{k}_1}^\beta\|^2 \cdot \| \sum_{\vec{k}_2} \prod_{i=1}^d (S_{P_i}^{-1})_{k_{1i} k_{2i}} \mathfrak{f}_{\vec{k}_2}\|^2. \end{aligned}$$

Durch Einsetzen der Gleichungen (4.19), (4.28) und (4.42) für die Komponenten der Matrizen $S_{P_i}^{-1}$ sowie (4.2), (4.25) und (4.32) für $h_i$ folgt

$$\begin{aligned} \|D^\beta \mathfrak{F}(t)\|^2 \le C_3^2 \sum_{\vec{k}_1} \|D_{\vec{k}_1}^\beta\|^2 \cdot \Big\| \sum_{\vec{k}_2} \prod_{j \in M_D} \frac{1}{N_j + 1} \frac{1}{d_j^{k_{2j}}} \sin \frac{k_{1j} k_{2j} \pi}{N_j + 1} \\ \cdot \prod_{j \in M_N} \frac{1}{N_j} \left\{ \begin{array}{ll} \cos \frac{k_{1j}(2k_{2j}-1)\pi}{2N_j} & : \ k_{1j} < N_j \\ \frac{1}{2} & : \ k_{1j} = N_j \end{array} \right\} \cdot \prod_{j \in M_P} \frac{1}{N_j} e^{\frac{2k_{1j}k_{2j}\pi}{N_j} i} \mathfrak{f}_{\vec{k}_2} \Big\|^2 \end{aligned}$$

mit $d_j = \sqrt{1 - \frac{h_j r_j}{1 + h_j r_j \delta_j}}$ und $C_3^2 = C_2^2 2^{d + |M_D| + |M_N|} \prod\limits_{i=1}^d l_i$.
Durch Verallgemeinerung der Abelschen partiellen Summation

$$\sum_{k=1}^N A_k \xi_k = A_{N+1} \Xi_N + \sum_{k=1}^N (A_k - A_{k+1}) \Xi_k \quad \text{mit} \quad \Xi_k = \sum_{i=1}^k \xi_i, \quad A_k, \xi_k \in \mathbb{R}$$

auf $d$ Dimensionen erhält man daraus

$$\begin{aligned} \|D^\beta \mathfrak{F}(t)\|^2 \le C_3^2 \sum_{\vec{k}_1} \|D_{\vec{k}_1}^\beta\|^2 \cdot \Big\| & \sum_{\vec{k}_2} \prod_{i=1}^d \sigma_i(k_{1i}, k_{2i}) \mathfrak{f}_{[\vec{k}_2]} \\ & + \sum_{i=1}^{d-1} \sum_{\substack{j_1, \ldots, j_i = 1, \ldots, d \\ j_1 < j_2 < \ldots < j_i}} \prod_{l=1}^i \sigma_{j_l}(k_{1 j_l}, N_{j_l}) \sum_{\substack{k_{2j} \\ j \in \mathfrak{M}_{j_1 \ldots j_i}}} \prod_{l \in \mathfrak{M}_{j_1 \ldots j_i}} \sigma_l(k_{1l}, k_{2l}) \\ & \cdot \mathfrak{f}_{[k_{21} \ldots k_{2(j_1-1)}](N_{j_1}+1)[k_{2(j_1+1)} \ldots k_{2(j_2-1)}] \ldots [k_{2(j_i+1)} \ldots k_{2d}]} \\ & + \prod_{i=1}^d \sigma_i(k_{1i}, N_i) \mathfrak{f}_{(N_1+1) \ldots (N_d+1)} \Big\|^2 \end{aligned}$$

$$\leq C_3^2 \sum_{\vec{k}_1} \|D^\beta_{\vec{k}_1}\|^2 \cdot \prod_{i=1}^{d} \max_{k_{2i}=1}^{N_i} |\sigma_i(k_{1i}, k_{2i})|^2 \cdot \Big\| \sum_{\vec{k}_2} |\mathfrak{f}_{[\vec{k}_2]}|$$
$$+ \sum_{i=1}^{d-1} \sum_{\substack{j_1,\ldots,j_i=1,\ldots,d \\ j_1<j_2<\ldots<j_i}} \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\ldots j_i}}} |\mathfrak{f}_{[k_{21}\ldots k_{2(j_1-1)}](N_{j_1}+1)[k_{2(j_1+1)}\ldots k_{2(j_2-1)}]\ldots[k_{2(j_i+1)}\ldots k_{2d}]}|$$
$$+ |\mathfrak{f}_{(N_1+1)\ldots(N_d+1)}| \Big\|^2, \tag{5.46}$$

wobei der Betrag komponentenweise zu verstehen ist und

$$\mathfrak{M}_{j_1\ldots j_i} = \{1,\ldots,d\} \setminus \{j_1,\ldots,j_i\}$$

sowie

$$\sigma_j(k,l) = \begin{cases} \frac{1}{N_j+1} \sum_{m=1}^{l} \frac{1}{d_j^m} \sin\frac{km\pi}{N_j+1} & : \ j \in M_D \\ \frac{1}{N_j} \sum_{m=1}^{l} \begin{cases} \cos\frac{k(2m-1)\pi}{2N_j} & : \ k < N_j \\ \frac{1}{2} & : \ k = N_j \end{cases} & : \ j \in M_N \\ \frac{1}{N_j} \sum_{m=1}^{l} e^{\frac{2km\pi i}{N_j}} & : \ j \in M_P \end{cases}$$

gelten.
Ist

$$M' = \max_{\substack{\vec{x}\in\overline{\Omega}, t\in[t_0,t_e], \\ i=1,\ldots,n}} \left\{ |\mathfrak{f}_i(t,\vec{x})|, \left|\frac{\partial f_i}{\partial x_1}\right|, \ldots, \left|\frac{\partial f_i}{\partial x_d}\right|, \left|\frac{\partial^2 f_i}{\partial x_1 \partial x_2}\right|, \ldots, \left|\frac{\partial^2 f_i}{\partial x_{d-1}\partial x_d}\right|, \ldots, \left|\frac{\partial^d f_i}{\partial x_1 \ldots \partial x_d}\right| \right\},$$

so folgt analog zum Beweis von Satz 5.14

$$\|D^\beta \mathfrak{F}(t)\|^2 \leq C_4^2 \sum_{\vec{k}_1} \|D^\beta_{\vec{k}_1}\|^2 \cdot \prod_{i=1}^{d} \max_{k_{2i}=1}^{N_i} |\sigma_i(k_{1i}, k_{2i})|^2$$

mit $C_4 = C_3 2^d n M' \prod_{i=1}^{d} \max\{1, 2l_i\}$.
Wegen der aus (2.28) und der Voraussetzung (5.44) folgenden Abschätzung

$$\|D^\beta_{\vec{k}}\| \leq n^{-\frac{1}{2}} \|D^\beta_{\vec{k}}\|_M = \sqrt{n} \max_{i,j=1}^{n} |(D^\beta_{\vec{k}})_{ij}| \leq \sqrt{n} C_1 \left(1 + \sum_{i=1}^{d} |\lambda_{i,k_i}|^\beta\right) \tag{5.47}$$

und

$$\left(\sum_{i=1}^{d} a_i\right)^2 = \sum_{i=1}^{d} a_i^2 + 2 \sum_{1\leq i<j\leq d} a_i a_j \leq \sum_{i=1}^{d} a_i^2 + \sum_{1\leq i<j\leq d} (a_i^2 + a_j^2) = d \sum_{i=1}^{d} a_i^2 \tag{5.48}$$

erhält man daraus

$$\|D^\beta \mathfrak{F}(t)\|^2 \leq C_5^2 \sum_{\vec{k}_1} \left(1 + \sum_{i=1}^{d} |\lambda_{i,k_{1i}}|^{2\beta}\right) \cdot \prod_{j=1}^{d} \max_{k_{2j}=1}^{N_j} |\sigma_j(k_{1j}, k_{2j})|^2 \tag{5.49}$$

mit $C_5^2 = C_4^2 n C_1^2 (d+1)$.
Ist $j \in M_D$, so kann $|\sigma_j(k,l)|$ wie folgt abgeschätzt werden:

Es gilt

$$\begin{aligned}
\sigma_j(k,l) &= \frac{1}{N_j+1}\mathfrak{Im}\sum_{m=0}^{l}\frac{1}{d_j^m}e^{\frac{km\pi i}{N_j+1}} = \frac{1}{N_j+1}\mathfrak{Im}\frac{\frac{1}{d_j^{l+1}}e^{\frac{k(l+1)\pi i}{N_j+1}}-1}{\frac{1}{d_j}e^{\frac{k\pi i}{N_j+1}}-1}\\
&= \frac{1}{N_j+1}\mathfrak{Im}\frac{\left(\frac{1}{d_j^{l+1}}e^{\frac{k(l+1)\pi i}{N_j+1}}-1\right)\left(\frac{1}{d_j}e^{-\frac{k\pi i}{N_j+1}}-1\right)}{\left(\frac{1}{d_j}e^{\frac{k\pi i}{N_j+1}}-1\right)\left(\frac{1}{d_j}e^{-\frac{k\pi i}{N_j+1}}-1\right)}\\
&= \frac{1}{N_j+1}\,\frac{\frac{1}{d_j^{l+2}}\sin\frac{kl\pi}{N_j+1}+\frac{1}{d_j}\sin\frac{k\pi}{N_j+1}-\frac{1}{d_j^{l+1}}\sin\frac{k(l+1)\pi}{N_j+1}}{1+\frac{1}{d_j^2}-\frac{2}{d_j}\cos\frac{k\pi}{N_j+1}}\\
&= \frac{1}{N_j+1}\,\frac{\frac{1}{d_j^{l+1}}\sin\frac{kl\pi}{N_j+1}\left(\frac{1}{d_j}-\cos\frac{k\pi}{N_j+1}\right)+\frac{1}{d_j}\sin\frac{k\pi}{N_j+1}\left(1-\frac{1}{d_j^l}\cos\frac{kl\pi}{N_j+1}\right)}{\left(1-\frac{1}{d_j}\right)^2+\frac{2}{d_j}\left(1-\cos\frac{k\pi}{N_j+1}\right)}.
\end{aligned}$$

Daraus erhält man

$$\begin{aligned}
|\sigma_j(k,l)| &\le \frac{1}{N_j+1}\,\frac{1}{d_j^{l+1}}\left(\frac{|1-d_j|}{2(1-\cos\frac{k\pi}{N_j+1})}+\frac{d_j}{2}\right)+\frac{1}{N_j+1}\,\frac{1+\frac{1}{d_j^l}}{2}\,\frac{\sin\frac{k\pi}{N_j+1}}{1-\cos\frac{k\pi}{N_j+1}}\\
&= \frac{1}{N_j+1}\,\frac{1}{d_j^{l+1}}\,\frac{|1-d_j|}{4\sin^2\frac{k\pi}{2(N_j+1)}}+\frac{1}{2}\,\frac{1}{N_j+1}\,\frac{1}{d_j^l}+\frac{1}{N_j+1}\,\frac{1+\frac{1}{d_j^l}}{2}\cot\frac{k\pi}{2(N_j+1)}.
\end{aligned}\tag{5.50}$$

Mit (4.22) folgt $\frac{1}{d_j^l}\le 2e^{l_j|r_j|}$ für $l\ge 0$. Für $x\in(0,\frac{\pi}{2})$ gilt $x<\tan x$, also

$$\cot x < \frac{1}{x}\tag{5.51a}$$

und $x\cos x-\sin x<0$. $\frac{\sin x}{x}$ ist folglich in $(0,\frac{\pi}{2})$ monoton fallend, und es gilt

$$\frac{\sin x}{x} > \frac{2}{\pi}.\tag{5.51b}$$

Damit erhält man aus (5.50)

$$|\sigma_j(k,l)| \le e^{l_j|r_j|}\frac{(N_j+1)|1-d_j|}{2k^2}+\frac{1}{N_j+1}e^{l_j|r_j|}+\frac{1+2e^{l_j|r_j|}}{k\pi}.$$

Aus dem Mittelwertsatz ergibt sich

$$1-d_j = h_j\frac{-r_j}{2(1+\zeta r_j\delta_j)^2\sqrt{1-\frac{\zeta r_j}{1+\zeta r_j\delta_j}}}$$

mit $\zeta\in(0,h_j)$. Für hinreichend kleine $h_j$ ($h_j|r_j|<\frac{1}{2}$) folgt daraus

$$|1-d_j| < \frac{4\sqrt{3}l_j}{N_j+1}|r_j|$$

und schließlich

$$|\sigma_j(k,l)| \le \frac{C_{\sigma_j}}{k} \quad \text{für} \quad j\in M_D$$

mit $C_{\sigma_j} = (2\sqrt{3}l_j|r_j|+1)e^{l_j|r_j|} + \frac{1+2e^{l_j|r_j|}}{\pi}$.
Eine ähnliche Abschätzung erhält man im Fall $j \in M_N$: Für $k < N_j$ gilt

$$\begin{aligned}
\sigma_j(k,l) &= \frac{1}{N_j}\sum_{m=1}^{l}\cos\frac{k(2m-1)\pi}{2N_j} = \frac{1}{N_j}\mathfrak{Re}e^{\frac{k\pi i}{2N_j}}\sum_{m=0}^{l-1}e^{\frac{km\pi i}{N_j}} \\
&= \frac{1}{N_j}\mathfrak{Re}e^{\frac{k\pi i}{2N_j}}\,\frac{e^{\frac{kl\pi i}{N_j}}-1}{e^{\frac{k\pi i}{N_j}}-1} = \frac{1}{N_j}\mathfrak{Re}e^{\frac{k\pi i}{2N_j}}\,\frac{\left(e^{\frac{kl\pi i}{N_j}}-1\right)\left(e^{-\frac{k\pi i}{N_j}}-1\right)}{\left(e^{\frac{k\pi i}{N_j}}-1\right)\left(e^{-\frac{k\pi i}{N_j}}-1\right)} \\
&= \frac{1}{N_j}\mathfrak{Re}\frac{e^{\frac{k(2l-1)\pi i}{2N_j}} - e^{\frac{k(2l+1)\pi i}{2N_j}} - e^{-\frac{k\pi i}{2N_j}} + e^{\frac{k\pi i}{2N_j}}}{2(1-\cos\frac{k\pi}{N_j})} \\
&= \frac{1}{N_j}\,\frac{\cos\frac{k(2l-1)\pi}{2N_j} - \cos\frac{k(2l+1)\pi}{2N_j}}{2(1-\cos\frac{k\pi}{N_j})} = \frac{1}{N_j}\,\frac{\sin\frac{kl\pi}{N_j}}{2\sin\frac{k\pi}{2N_j}}
\end{aligned} \tag{5.52a}$$

und damit

$$|\sigma_j(k,l)| \le \frac{1}{2k} \text{ für } j \in M_N \text{ und } k < N_j.$$

Für $k = N_j$ gilt

$$|\sigma_j(N_j,l)| = \frac{1}{N_j}\sum_{m=1}^{l}\frac{1}{2} \le \frac{1}{2}. \tag{5.52b}$$

Für $j \in M_P$ schließlich gilt für $k < N_j$

$$\begin{aligned}
|\sigma_j(k,l)| &= \frac{1}{N_j}\left|\sum_{m=1}^{l}e^{\frac{2km\pi i}{N_j}}\right| = \frac{1}{N_j}\left|\sum_{m=0}^{l-1}e^{\frac{2km\pi i}{N_j}}\right| = \frac{1}{N_j}\frac{|e^{\frac{2kl\pi i}{N_j}}-1|}{|e^{\frac{2k\pi i}{N_j}}-1|} \\
&= \frac{1}{N_j}\sqrt{\frac{1-\cos\frac{2kl\pi}{N_j}}{1-\cos\frac{2k\pi}{N_j}}} \le \frac{1}{N_j}\,\frac{1}{\sin\frac{k\pi}{N_j}} \le \frac{1}{2\min\{k,N_j-k\}}
\end{aligned}$$

und für $k = N_j$

$$|\sigma_j(N_j,l)| = \frac{1}{N_j}\left|\sum_{m=1}^{l}e^{2m\pi i}\right| = \frac{l}{N_j} \le 1.$$

Als nächstes sollen Abschätzungen für die Beträge der Eigenwerte $|\lambda_{i,k_i}|$ hergeleitet werden. Ist $j \in M_D$, so erhält man aus der für die Eigenwerte geltenden Gleichung (4.16) für genügend kleine $h_i$ ($h_i|r_i| < \frac{1}{2}$, $h_i$ definiert gemäß (4.2)) durch Taylor-Entwicklung von $\sqrt{1-x}$ mit $x = \frac{h_i r_i}{1+h_i r_i \delta_i}$ und $\zeta \in (0, \frac{h_i r_i}{1+h_i r_i \delta_i})$ sowie Nutzung von $\sin x < x$ für $x > 0$

$$\begin{aligned}
|\lambda_{i,k_i}| &\le \left|\frac{r_i}{h_i} - 2\frac{1+h_i r_i\delta_i}{h_i^2}\left(\frac{1}{2}\,\frac{h_i r_i}{1+h_i r_i\delta_i} + \frac{1}{8}\,h_i^2\left(\frac{r_i}{1+h_i r_i\delta_i}\right)^2(1-\zeta)^{-\frac{3}{2}}\right)\right| \\
&\quad + \frac{8}{h_i^2}\cdot 2\left(\frac{k_i\pi}{2(N_i+1)}\right)^2 \\
&\le \sqrt{2}r_i^2 + k_i^2\frac{\pi^2}{l_i^2} \le \left(\sqrt{2}r_i^2 + \frac{\pi^2}{l_i^2}\right)k_i^2.
\end{aligned}$$

Für $i \in M_P$ ist nach (4.27) und (4.25) für genügend kleine $h_i$

$$|\lambda_{i,k_i}|^2 = \left[\left(\frac{2+2h_i r_i\delta_i}{h_i^2}\right)^2 - \frac{2r_i}{h_i}\,\frac{2+2h_i r_i\delta_i}{h_i^2}\right]\left(1-\cos\frac{2k_i\pi}{N_i}\right)^2 + \frac{r_i^2}{h_i^2}\left(2-2\cos\frac{2k_i\pi}{N_i}\right)$$

$$
\begin{aligned}
&= (2 + 2h_i r_i(\delta_i - 1))\,(2 + 2h_i r_i \delta_i)\,\frac{4}{h_i^4}\sin^4\frac{k_i\pi}{N_i} + \frac{4r_i^2}{h_i^2}\sin^2\frac{k_i\pi}{N_i} \\
&\le \frac{4\pi^4}{l^4}\min\{k_i, N_i - k_i\}^4 + \frac{r_i^2\pi^2}{l_i^2}\min\{k_i, N_i - k_i\}^2 \\
&\le \frac{\pi^2}{l_i^2}\left(\frac{4\pi^2}{l_i^2} + r_i^2\right)\min\{k_i, N_i - k_i\}^4,
\end{aligned}
$$

wobei wieder $\sin x < x$ für $x > 0$ ausgenutzt wurde.
Für $i \in M_N$ schließlich gelten nach (4.40) und (4.32) wegen $r_i = 0$ für $k_i = 1, \dots, N_i - 1$

$$
|\lambda_{i,k_i}| = \frac{4}{h_i^2}\sin^2\frac{k_i\pi}{2N_i} \le \frac{\pi^2}{4l_i^2}k_i^2 \tag{5.53}
$$

und $\lambda_{i,N_i} = 0$.
Durch Einsetzen der für $|\lambda_{i,k_i}|$ und $|\sigma_i(k_{1i}, k_{2i})|$ gefundenen Abschätzungen in (5.49) folgt

$$
\begin{aligned}
&\|D^\beta \mathfrak{F}(t)\|^2 \\
&\le C_6^2 \sum_{\vec{k}_1}\left(1 + \sum_{i\in M_D} k_{1i}^{4\beta} + \sum_{i\in M_P}\min\{k_{1i}, (N_i - k_{1i})\}^{4\beta} + \sum_{i\in M_N}\left\{\begin{array}{lcl} k_{1i}^{4\beta} & : & k_{1i} < N_i \\ 0 & : & k_{1i} = N_i \end{array}\right\}\right) \\
&\quad \cdot \prod_{j\in M_D}\frac{1}{k_{1j}^2}\prod_{j\in M_N}\left\{\begin{array}{lcl} \frac{1}{k_{1j}^2} & : & k_{1j} < N_j \\ 1 & : & k_{1j} = N_j \end{array}\right\}\prod_{j\in M_P}\left\{\begin{array}{lcl} \frac{1}{\min\{k_{1j}^2, (N_j - k_{1j})^2\}} & : & k_{1j} < N_j \\ 1 & : & k_{1j} = N_j \end{array}\right\} \\
&\le C_6^2\left\{\prod_{j\in M_D}\left(\sum_{k=1}^{N_j}\frac{1}{k^2}\right)\prod_{j\in M_N}\left(1 + \sum_{k=1}^{N_j - 1}\frac{1}{k^2}\right)\prod_{j\in M_P}\left(1 + 2\sum_{k=1}^{\lfloor\frac{N_j}{2}\rfloor}\frac{1}{k^2}\right)\right. \\
&\quad + \sum_{i\in M_D}\sum_{k=1}^{N_i} k^{4\beta-2}\prod_{j\in M_D\setminus\{i\}}\left(\sum_{k=1}^{N_j}\frac{1}{k^2}\right)\prod_{j\in M_N}\left(1 + \sum_{k=1}^{N_j-1}\frac{1}{k^2}\right)\prod_{j\in M_P}\left(1 + 2\sum_{k=1}^{\lfloor\frac{N_j}{2}\rfloor}\frac{1}{k^2}\right) \\
&\quad + \sum_{i\in M_N}\sum_{k=1}^{N_i - 1} k^{4\beta-2}\prod_{j\in M_N\setminus\{i\}}\left(1 + \sum_{k=1}^{N_j-1}\frac{1}{k^2}\right)\prod_{j\in M_D}\left(\sum_{k=1}^{N_j}\frac{1}{k^2}\right)\prod_{j\in M_P}\left(1 + 2\sum_{k=1}^{\lfloor\frac{N_j}{2}\rfloor}\frac{1}{k^2}\right) \\
&\quad \left. + \sum_{i\in M_P} 2\sum_{k=1}^{\lfloor\frac{N_j}{2}\rfloor} k^{4\beta-2}\prod_{j\in M_P\setminus\{i\}}\left(1 + 2\sum_{k=1}^{\lfloor\frac{N_j}{2}\rfloor}\frac{1}{k^2}\right)\prod_{j\in M_D}\left(\sum_{k=1}^{N_j}\frac{1}{k^2}\right)\prod_{j\in M_N}\left(1 + \sum_{k=1}^{N_j-1}\frac{1}{k^2}\right)\right\},
\end{aligned}
$$

wobei

$$
C_6^2 = C_5^2 \max_{i=1}^{d}\left\{\left(\sqrt{2}r_i^2 + \frac{\pi^2}{l_i^2}\right)^{2\beta}, \left[\frac{\pi^2}{l_i^2}\left(\frac{4\pi^2}{l_i^2} + r_i^2\right)\right]^\beta, \left(\frac{\pi^2}{4l_i^2}\right)^{2\beta}\right\}\prod_{j\in M_D} C_{\sigma_j}^2
$$

gesetzt werden kann und $\lfloor x \rfloor$ für eine positive reelle Zahl $x$ deren ganzzahligen Anteil bedeuten soll.
Aus $\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6} < 2$ folgt

$$
\begin{aligned}
\|D^\beta \mathfrak{F}(t)\|^2 &\le C_6^2 5^d\left(1 + \sum_{i\in M_D}\sum_{k=1}^{N_i} k^{4\beta-2} + \sum_{i\in M_N}\sum_{k=1}^{N_i - 1} k^{4\beta-2} + \sum_{i\in M_P}\sum_{k=1}^{\lfloor\frac{N_j}{2}\rfloor} k^{4\beta-2}\right) \\
&\le C_6^2 5^d\left(1 + \sum_{i=1}^{d}\sum_{k=1}^{N_i} k^{4\beta-2}\right).
\end{aligned} \tag{5.54}
$$

Für $N_i \to \infty$ sind diese Summenfolgen genau dann konvergent, wenn $4\beta - 2 < -1$, also $\beta < \frac{1}{4}$ gilt, womit die Behauptung bewiesen ist. □

**Bemerkung 5.20** Man kann zeigen, daß die obere Schranke scharf ist, d. h. glatte Funktionen $\mathfrak{f}$ existieren, für die $\|D^{\frac{1}{4}}\mathfrak{F}(t)\| \to \infty$ für $\vec{h} \to 0$ gilt. Da $\|D^{\frac{1}{4}-\varepsilon}U^{(l)}_{\vec{h}}\|$ in die Fehlerkonstante des globalen Diskretisierungsfehlers eingeht, kann diese für sehr kleine $\varepsilon > 0$ sehr groß werden: $\sum\limits_{k=1}^{N_i} k^{4\beta-2}$ kann wegen

$$\int_2^{N_i+1} x^{4\beta-2}\,dx < \sum_{k=1}^{N_i} k^{4\beta-2} < 1 + \int_1^{N_i} x^{4\beta-2}\,dx$$

für $\beta = \frac{1}{4} - \varepsilon$ durch

$$1 + \frac{1 - N_i^{4\beta-1}}{1-4\beta} < 1 + \frac{1}{1-4\beta} = \frac{1}{4\varepsilon} + 1$$

abgeschätzt werden. Aus (5.54) folgt damit für kleine $\varepsilon > 0$

$$\|D^{\frac{1}{4}-\varepsilon}U^{(l)}_{\vec{h}}\| \le \frac{C}{\sqrt{\varepsilon}}.$$

Für $\beta = \frac{1}{4}$ erhält man dagegen

$$\ln(N_i+1) - \ln 2 \le \sum_{k=1}^{N_i} k^{4\beta-2} < 1 + \ln N_i \le 1 + \ln \frac{2l_i}{h_i}$$

und damit unter den Voraussetzungen (a)-(e) des Satzes 5.11 für den globalen Diskretisierungsfehler

$$\|e_{m+1}\| = \mathcal{O}\left(\sqrt{\sum_{i=1}^{d} |\ln h_i|}\tau^{q+1.25}\right) + \mathcal{O}(\tau^p) + \sum_{i=1}^{d}\mathcal{O}(h_i^{p_i}).$$ □

In Konvergenzsatz 5.11 kann damit unter den Voraussetzungen von Satz 5.19

$$\alpha_r \le \alpha = -\frac{3}{4} - \varepsilon$$

mit $\varepsilon > 0$ beliebig klein gewählt werden.
Ist $M_D = M_P = \emptyset$, d. h., es werden nur Neumann-Randbedingungen vorgeschrieben, so kann Satz 5.19 verbessert werden. Der Einfachheit halber wird zusätzlich vorausgesetzt, daß die Ortsschrittweiten $h_i$ alle gleich einer Konstanten $h$ sind.

**Satz 5.21** Gelten zusätzlich zu den Voraussetzungen von Satz 5.19 $M_D = M_P = \emptyset$ und $h_i = h$, $i = 1, \ldots, d$, so gilt (5.45) sogar für $\beta < \frac{3}{4}$. □

In Konvergenzsatz 5.11 kann dann

$$\alpha_r \le \alpha = -\frac{1}{4} - \varepsilon$$

mit $\varepsilon > 0$ beliebig klein gewählt werden.
Zum Beweis von Satz 5.21 wird zunächst das folgende Lemma gezeigt:

**Lemma 5.22** Seien $M_D = M_P = \emptyset$ und $h_i = h$, $i = 1, \ldots, d$. Existieren auf $\overline{\Omega}' = \overline{\Omega}'_1 \times \ldots \times \overline{\Omega}'_d$ mit $\overline{\Omega}'_i$ aus (4.33) die $(d+j)$-ten Ableitungen von $\mathfrak{f}$ und sind stetig, so ist

$$\prod_{i=1}^{d} N_i \sum_{\substack{k_i=1,\ldots,N_i \\ i=1,\ldots,j}} |\mathfrak{f}_{[[k_1 \ldots k_j]][k_{j+1} \ldots k_d]}|$$

mit

$$\mathfrak{f}_{[[k_1 \ldots k_j]][k_{j+1} \ldots k_d]} = \sum_{\substack{\nu_i=0,1 \\ i=1,\ldots,d}} \sum_{\substack{\mu_l=0,1 \\ l=1,\ldots,j}} (-1)^{\sum_{i=1}^{d} \nu_i + \sum_{i=1}^{j} \mu_i}$$
$$\mathfrak{f}(x_{1k_1} + (\nu_1 + \mu_1)h, \ldots, x_{jk_j} + (\nu_j + \mu_j)h, x_{(j+1)k_{j+1}} + \nu_{j+1}h, \ldots, x_{dk_d} + \nu_d h)$$

beschränkt. □

**Beweis:** Für die Funktion

$$z(\vec{y}) = \sum_{\substack{\nu_i=0,1 \\ i=1,\ldots,d}} \sum_{\substack{\mu_l=0,1 \\ l=1,\ldots,j}} (-1)^{\sum_{i=1}^{d} \nu_i + \sum_{i=1}^{j} \mu_i}$$
$$\mathfrak{f}(x_{1k_1} + (\nu_1 + \mu_1)y_1, \ldots, x_{jk_j} + (\nu_j + \mu_j)y_j, x_{(j+1)k_{j+1}} + \nu_{j+1}y_{j+1}, \ldots, x_{dk_d} + \nu_d y_d)$$

gelten

$$z(\vec{y}) = 0, \quad \text{falls} \quad y_i = 0 \quad \text{für ein } i \in \{1, \ldots, d\}$$

und für $i \in \{1, \ldots, j\}$ auch

$$z_{y_i}(\vec{y}) = 0, \quad \text{falls} \quad y_i = 0.$$

Damit verschwinden an $\vec{y} = 0$ alle Ableitungen von $z(\vec{y})$ bis zur $j + d - 1$-ten Ordnung, und der Satz von Taylor in $d$ Dimensionen liefert

$$z(h \cdot \mathbb{1}_d) = \frac{h^{j+d}}{(j+d)!} (\mathbb{1}_d \cdot \nabla)^{j+d} z(\vec{\xi}) \quad \text{mit} \quad \xi_i \in (0, h), \quad i = 1, \ldots, d.$$

Mit

$$M' = \max_{\vec{x} \in \overline{\Omega}'} \left\{ \left| \frac{\partial^{d+j} \mathfrak{f}}{\partial x_1^{d+j}}(\vec{x}) \right|, \left| \frac{\partial^{d+j} \mathfrak{f}}{\partial x_1^{d+j-1} \partial x_2}(\vec{x}) \right|, \left| \frac{\partial^{d+j} \mathfrak{f}}{\partial x_1^{d+j-1} \partial x_3}(\vec{x}) \right|, \ldots, \left| \frac{\partial^{d+j} \mathfrak{f}}{\partial x_d^{d+j}}(\vec{x}) \right| \right\}$$

(Maximum aller Ableitungen der $(d+j)$-ten Ordnung über $\overline{\Omega}'$) folgt daraus

$$|\mathfrak{f}_{[[k_1 \ldots k_j]][k_{j+1} \ldots k_d]}| = |z(h \cdot \mathbb{1}_d)| \leq \frac{h^{j+d}}{(j+d)!} d^{j+d} 2^{j+d} M'$$

und damit wegen $N_i h_i \leq 2 l_i$

$$\prod_{i=1}^{d} N_i \sum_{\substack{k_i=1,\ldots,N_i \\ i=1,\ldots,j}} |\mathfrak{f}_{[[k_1 \ldots k_j]][k_{j+1} \ldots k_d]}| \leq \frac{d^{j+d}}{(j+d)!} 2^{2j+2d} M' \prod_{i=1}^{j} l_i^2 \prod_{i=j+1}^{d} l_i.$$

□

Mit diesem Lemma kann nun Satz 5.21 bewiesen werden:
**Beweis:** Aus (5.52) folgt

$$\sigma_j(k, N_j) = \begin{cases} 0 & : \ k < N_j \\ \frac{1}{2} & : \ k = N_j \end{cases} .$$

Durch erneute Anwendung der Abelschen partiellen Summation erhält man damit aus (5.46)

$$
\begin{aligned}
\|D^\beta \mathfrak{F}(t)\|^2 \le\ & C_3^2 \sum_{\vec{k}_1} \|D^\beta_{\vec{k}_1}\|^2 \cdot \Big\| \sum_{\vec{k}_2} \prod_{i=1}^{d} N_i \pi_i(k_{1i}, k_{2i}) \mathfrak{f}_{[[\vec{k}_2]]} \\
& + \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \prod_{l=1}^{i} N_{j_l} \pi_{j_l}(k_{1j_l}, N_{j_l}) \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\dots j_i}}} \prod_{l\in\mathfrak{M}_{j_1\dots j_i}} N_l \pi_l(k_{1l}, k_{2l}) \\
& \cdot \mathfrak{f}_{[[k_{21}\dots k_{2(j_1-1)}]][N_{j_1}+1][[k_{2(j_1+1)}\dots k_{2(j_2-1)}]]\dots[[k_{2(j_i+1)}\dots k_{2d}]]} \\
& + \prod_{i=1}^{d} N_i \pi_i(k_{1i}, N_i) \mathfrak{f}_{[(N_1+1)\dots(N_d+1)]} \\
& + \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \prod_{l=1}^{i} \sigma_{j_l}(k_{1j_l}, N_{j_l}) \Bigg( \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\dots j_i}}} \prod_{l\in\mathfrak{M}_{j_1\dots j_i}} N_l \pi_l(k_{1l}, k_{2l}) \\
& \cdot \mathfrak{f}_{[[k_{21}\dots k_{2(j_1-1)}]](N_{j_1}+1)[[k_{2(j_1+1)}\dots k_{2(j_2-1)}]]\dots[[k_{2(j_i+1)}\dots k_{2d}]]} \\
& + \sum_{i_2=1}^{|\mathfrak{M}_{j_1\dots j_i}|-1} \sum_{\substack{m_1,\dots,m_{i_2}=1,\dots,d \\ m_1<m_2<\dots<m_{i_2}}} \prod_{r=1}^{i_2} N_{m_r} \pi_{m_r}(k_{1m_r}, N_{m_r}) \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\dots j_i, m_1,\dots,m_{i_2}}}} \\
& \prod_{l\in\mathfrak{M}_{j_1\dots j_i, m_1,\dots,m_{i_2}}} N_l \pi_l(k_{1l}, k_{2l}) \cdot \mathfrak{f}_{[[k_{21}\dots k_{2(m_1-1)}]][N_{m_1}+1][[\dots]](N_{j_1}+1)\dots} \\
& + \prod_{i_2\in\mathfrak{M}_{j_1\dots j_i}} N_{i_2} \pi_{i_2}(k_{1i_2}, N_{i_2}) \mathfrak{f}_{[(N_1+1)\dots(N_{j_1-1}+1)](N_{j_1}+1)\dots} \Bigg) \\
& + \prod_{i=1}^{d} \sigma_i(k_{1i}, N_i) \mathfrak{f}_{(N_1+1)\dots(N_d+1)} \Big\|^2 \\
\le\ & C_3^2 \sum_{\vec{k}_1} \|D^\beta_{\vec{k}_1}\|^2 \cdot \Bigg[ \prod_{i=1}^{d} \max_{k_{2i}=1}^{N_i} |\pi_i(k_{1i}, k_{2i})| \cdot \Big\| \prod_{i=1}^{d} N_i \Big( \sum_{\vec{k}_2} |\mathfrak{f}_{[[\vec{k}_2]]}| \\
& + \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\dots j_i}}} |\mathfrak{f}_{[[k_{21}\dots k_{2(j_1-1)}]][N_{j_1}+1][[k_{2(j_1+1)}\dots k_{2(j_2-1)}]]\dots[[k_{2(j_i+1)}\dots k_{2d}]]}| \\
& + |\mathfrak{f}_{[(N_1+1)\dots(N_d+1)]}| \Big) \Big\| \\
& + \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \prod_{l=1}^{i} \left\{ \begin{array}{ll} 0 : & k_{1j_l} < N_{j_l} \\ \frac{1}{2} : & k_{1j_l} = N_{j_l} \end{array} \right\} \prod_{l\in\mathfrak{M}_{j_1\dots j_i}} \max_{k_{2l}=1}^{N_l} |\pi_l(k_{1l}, k_{2l})| \\
& \cdot \Big\| \prod_{l\in\mathfrak{M}_{j_1\dots j_i}} N_l \cdot \Big( \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\dots j_i}}} |\mathfrak{f}_{[[k_{21}\dots k_{2(j_1-1)}]](N_{j_1}+1)[[k_{2(j_1+1)}\dots k_{2(j_2-1)}]]\dots[[k_{2(j_i+1)}\dots k_{2d}]]}| \\
& + \sum_{i_2=1}^{|\mathfrak{M}_{j_1\dots j_i}|-1} \sum_{\substack{m_1,\dots,m_{i_2}=1,\dots,d \\ m_1<m_2<\dots<m_{i_2}}} \sum_{\substack{k_{2j} \\ j\in\mathfrak{M}_{j_1\dots j_i, m_1,\dots,m_{i_2}}}} |\mathfrak{f}_{[[k_{21}\dots k_{2(m_1-1)}]][N_{m_1}+1][[\dots]](N_{j_1}+1)\dots} \\
& + |\mathfrak{f}_{[(N_1+1)\dots(N_{j_1-1}+1)](N_{j_1}+1)\dots}| \Big) \Big\| + \prod_{i=1}^{d} \left\{ \begin{array}{ll} 0 : & k_{1i} < N_i \\ \frac{1}{2} : & k_{1i} = N_i \end{array} \right\} \|\mathfrak{f}_{(N_1+1)\dots(N_d+1)}\| \Bigg]^2
\end{aligned}
$$

mit

$$\pi_i(k,l) = \frac{1}{N_i} \sum_{m=1}^{l} \sigma_i(k,m). \tag{5.55}$$

Analog zum Beweis von Lemma 5.22 folgt daraus

$$\begin{aligned}\|D^\beta \mathfrak{F}(t)\|^2 \le C_4^2 \sum_{\vec{k}_1} \|D^\beta_{\vec{k}_1}\|^2 \cdot \Big[ &\prod_{i=1}^{d} \max_{k_{2i}=1}^{N_i} |\pi_i(k_{1i},k_{2i})| \\ &+ \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \prod_{l=1}^{i} \left\{ \begin{array}{ll} 0: & k_{1j_l} < N_{j_l} \\ 1: & k_{1j_l} = N_{j_l} \end{array} \right\} \prod_{l \in \mathfrak{M}_{j_1\dots j_i}} \max_{k_{2l}=1}^{N_l} |\pi_l(k_{1l},k_{2l})| + \prod_{i=1}^{d} \left\{ \begin{array}{ll} 0: & k_{1i} < N_i \\ 1: & k_{1i} = N_i \end{array} \right\} \Big]^2\end{aligned}$$

mit $C_4 = C_3 3^d \frac{d^{2d}}{(2d)!} 2^{4d} n M' \prod_{i=1}^{d} \max\{1, l_i^2\}$, wobei $M'$ das Maximum der Beträge der Komponenten der Funktion $\mathfrak{f}$ und ihrer Ableitungen bis zur $2d$-ten Ordnung über $\overline{\Omega}'$ ist.
Einsetzen von (5.47) und Nutzung von (5.48) und $\lambda_{i,N_i} = 0$ liefern

$$\begin{aligned}\|D^\beta \mathfrak{F}(t)\|^2 \le C_5^2 \Bigg[ &\sum_{\vec{k}_1} \left( 1 + \sum_{v=1}^{d} |\lambda_{v,k_1 v}|^{2\beta} \right) \prod_{i=1}^{d} \max_{k_{2i}=1}^{N_i} |\pi_i(k_{1i},k_{2i})|^2 \\ &+ \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \sum_{\substack{k_{1j} \\ j \in \mathfrak{M}_{j_1\dots j_i}}} \left( 1 + \sum_{v \in \mathfrak{M}_{j_1\dots j_i}} |\lambda_{v,k_1 v}|^{2\beta} \right) \prod_{l \in \mathfrak{M}_{j_1\dots j_i}} \max_{k_{2l}=1}^{N_l} |\pi_l(k_{1l},k_{2l})|^2 + 1 \Bigg]\end{aligned} \tag{5.56}$$

mit $C_5^2 = C_4^2 n C_1^2 (d+1) 2^d$.
Durch Einsetzen von (5.52) in (5.55) folgt

$$\pi_i(k,l) = \left\{ \begin{array}{ll} \frac{1}{2N_i^2 \sin\frac{k\pi}{2N_i}} \sum_{m=1}^{l} \sin\frac{km\pi}{N_i}: & k < N_i \\ \frac{1}{2N_i^2} \sum_{m=1}^{l} m \quad : & k = N_i \end{array} \right. .$$

Für $k < N_i$ gilt analog zu (5.50)

$$|\pi_i(k,l)| \le \frac{1}{2N_i \sin\frac{k\pi}{2N_i}} \left( \frac{1}{2N_i} + \frac{1}{N_i} \cot\frac{k\pi}{2N_i} \right),$$

und mit den Abschätzungen (5.51) folgt schließlich

$$|\pi_i(k,l)| \le \frac{1}{2k} \left( \frac{1}{2N_i} + \frac{2}{k\pi} \right) \le \frac{1}{k^2}, \quad k < N_i.$$

Für $k = N_i$ erhält man $\pi_i(N_i,l) = \frac{l(l+1)}{4N_i^2} \le \frac{1}{2}$. Einsetzen in (5.56) liefert unter Berücksichtigung von (5.53)

$$\begin{aligned}\|D^\beta \mathfrak{F}(t)\|^2 \le C_6^2 \Bigg[ &\sum_{\vec{k}_1} \left( 1 + \sum_{v=1}^{d} \left\{ \begin{array}{ll} k_{1v}^{4\beta}: & k_{1v} < N_v \\ 0 \quad : & k_{1v} = N_v \end{array} \right\} \right) \prod_{i=1}^{d} \left\{ \begin{array}{ll} \frac{1}{k_{1i}^4}: & k_{1i} < N_i \\ \frac{1}{4} \quad : & k_{1i} = N_i \end{array} \right\} \\ &+ \sum_{i=1}^{d-1} \sum_{\substack{j_1,\dots,j_i=1,\dots,d \\ j_1<j_2<\dots<j_i}} \sum_{\substack{k_{1j} \\ j \in \mathfrak{M}_{j_1\dots j_i}}}\end{aligned}$$

$$\left(1+\sum_{v\in\mathfrak{M}_{j_1\dots j_i}}\left\{\begin{array}{ll}k_{1v}^{4\beta} & :\ k_{1v}<N_v\\ 0 & :\ k_{1v}=N_v\end{array}\right\}\right)\prod_{l\in\mathfrak{M}_{j_1\dots j_i}}\left\{\begin{array}{ll}\frac{1}{k_{1l}^4} & :\ k_{1l}<N_l\\ \frac{1}{4} & :\ k_{1l}=N_l\end{array}\right\}+1\Bigg]$$

$$= C_6^2\Bigg[\sum_{v=1}^{d}\sum_{k=1}^{N_v-1}k^{4\beta-4}\prod_{\substack{i=1\\ i\neq v}}^{d}\left(\frac{1}{4}+\sum_{k=1}^{N_i-1}\frac{1}{k^4}\right)+\prod_{i=1}^{d}\left(\frac{1}{4}+\sum_{k=1}^{N_i-1}\frac{1}{k^4}\right)$$

$$+\sum_{i=1}^{d-1}\sum_{\substack{j_1,\dots,j_i=1,\dots,d\\ j_1<j_2<\dots<j_i}}\ \sum_{v\in\mathfrak{M}_{j_1\dots j_i}}\ \sum_{k_{1v}=1}^{N_v-1}k_{1v}^{4\beta-4}\prod_{l\in\mathfrak{M}_{j_1\dots j_i v}}\left(\frac{1}{4}+\sum_{k=1}^{N_l-1}\frac{1}{k^4}\right)$$

$$+\sum_{i=1}^{d-1}\sum_{\substack{j_1,\dots,j_i=1,\dots,d\\ j_1<j_2<\dots<j_i}}\ \prod_{l\in\mathfrak{M}_{j_1\dots j_i}}\left(\frac{1}{4}+\sum_{k=1}^{N_l-1}\frac{1}{k^4}\right)+1\Bigg]$$

mit $C_6 = C_5\max\left\{1,\left(\frac{\pi^2}{4l_i^2}\right)^{\beta}:\ i=1,\dots,d\right\}$.
Wegen $\sum\limits_{k=1}^{\infty}\frac{1}{k^4}=\frac{\pi^2}{90}<\frac{3}{4}$ folgt daraus

$$\|D^{\beta}\mathfrak{F}(t)\|^2 \le C_6^2 2^{d+1}\left(1+\sum_{i=1}^{d}\sum_{k=1}^{N_i-1}k^{4\beta-4}\right).$$

Für $N_i\to\infty$ sind diese Summenfolgen genau dann konvergent, wenn $4\beta-4<-1$, also $\beta<\frac{3}{4}$ gilt. □

Sind die $(q+1)$-ten Zeitableitungen der Dirichlet- und Neumann-Randbedingungen in jeder Raumdimension homogen, d. h.

$$B_i\frac{\partial^{q+1}u}{\partial t^{q+1}}=0 \text{ für } i\in M_D,\quad \frac{\partial}{\partial x_i}B_i\frac{\partial^{q+1}u}{\partial t^{q+1}}=0 \text{ für } i\in M_N,\quad \vec{x}\in\partial_i\Omega, \tag{5.57}$$

so gilt nach (4.46), (3.2e) und (3.2f) $\omega^{(l)}(t)\equiv 0$ für $l=q+1,\dots,p+1$, und mit der Gleichung (4.49) für den lokalen Ortsdiskretisierungsfehler erhält man unter der Voraussetzung, daß $D^{\alpha}$ und damit auch $D^{1+\alpha}$ existieren,

$$D^{1+\alpha}U_{\vec{h}}^{(l)}(t)=D^{\alpha}\left(MU_{\vec{h}}^{(l+1)}(t)-F^{(l)}(t)-\alpha_{\vec{h}}^{(l)}(t)\right).$$

Für $\alpha<\frac{1}{4}$ bzw. $\alpha<\frac{3}{4}$ sind unter den Voraussetzungen von Satz 5.19 bzw. Satz 5.21 die Vektornormen $\|D^{\alpha}MU_{\vec{h}}^{(l+1)}(t)\|$ und $\|D^{\alpha}F^{(l)}(t)\|$ für hinreichend kleine $\vec{h}$ beschränkt. $\|D^{\alpha}\alpha_{\vec{h}}^{(l)}(t)\|$ ist nach (4.50) und (4.51) unter der Voraussetzung, daß die exakte Lösung $u$ genügend glatt ist, beschränkt, falls $\|D^{\alpha}(I_N\otimes B_i)h_i^{p_i}\|$, $i=1,\dots,d$, beschränkt bleibt.
Insgesamt ergibt sich damit nach Diagonalisierung der folgende Satz:

**Satz 5.23** Seien im PDA-System (3.2) $r_i=0$ für $i\in M_N$ und $\beta<\frac{1}{4}$. Für $h_v\to 0$ $(N_v\to\infty)$, $v=1,\dots,d$, existiere $D_{\vec{k}}^{\beta}$ für alle $\vec{k}$ mit $k_v=1,\dots,N_v$, und es gelte mit einer von $\vec{h}$ unabhängigen Konstanten $C_1$

$$|(D_{\vec{k}}^{\beta})_{ij}|\le C_1\left(1+\sum_{v=1}^{d}|\lambda_{v,k_v}|^{\beta}\right),\quad i,j=1,\dots,n.$$

Dann ist $\|D^{\beta}U_{\vec{h}}^{(l)}(t)\|$ für $\vec{h}\to 0$ beschränkt, $l=q+1,\dots,p+1$.
Ist im Fall homogener Dirichlet- und homogener Neumann-Randbedingungen im Sinne von (5.57) zusätzlich

$$\|D_{\vec{k}}^{\beta}B_v h_v^{p_v}\|$$

beschränkt, so ist sogar $\|D^{1+\beta}U_{\vec{h}}^{(l)}(t)\|$ beschränkt für $\vec{h} \to 0$.
Gelten $M_D = M_P = \emptyset$ und $h_i = h$, $i = 1, \ldots, d$, so kann $\beta < \frac{3}{4}$ gewählt werden. □

Abschließend wird der Fall periodischer Randbedingungen betrachtet. Es gilt allgemein:

**Lemma 5.24** Die Funktion $w : \mathbb{R} \to \mathbb{R}$ habe periodische Randbedingungen,

$$w(x) = w(x + 2l), \quad x \in \mathbb{R}.$$

Ist $w \in C^{2m}$ mit $m \in \mathbb{N}$, so gilt für den zum Ortsgitter (4.3) mit $h = \frac{2l}{N}$ gehörigen Gittervektor $W_h$, daß mit $\frac{1}{h^2}P_h$ aus (4.26) $\| \left(\frac{1}{h^2}P_h\right)^m W_h\|$ für alle $h > 0$ beschränkt ist. □

**Beweis:** Der Beweis erfolgt durch vollständige Induktion:
Für $m = 0$ ist die Behauptung erfüllt.
Die Behauptung gelte für alle $m \leq i \in \mathbb{N}$. Mit der Taylor-Entwicklung (4.6) folgt

$$\begin{aligned}
\left(\frac{1}{h^2}P_h\right)^{i+1} W_h = \left(\frac{1}{h^2}P_h\right)^{i} \Bigg[ & r\sum_{j=0}^{i} \frac{h^{2j}}{(2j+1)!} W_h^{(2j+1)} + (2 + hr(2\delta - 1)) \sum_{j=1}^{i} \frac{h^{2j-2}}{(2j)!} W_h^{(2j)} \\
& + \frac{h^{2i}}{(2i+2)!} \Big\{ (1 + hr\delta) \left( w^{(2i+2)}(\zeta_{11}), \ldots, w^{(2i+2)}(\zeta_{1N}) \right)^\top \\
& + (1 + hr(\delta - 1)) \left( w^{(2i+2)}(\zeta_{21}), \ldots, w^{(2i+2)}(\zeta_{2N}) \right)^\top \Big\} \Bigg] \\
= r\sum_{j=0}^{i} \frac{P_h^j}{(2j+1)!} & \left(\frac{1}{h^2}P_h\right)^{i-j} W_h^{(2j+1)} \\
& + (2 + hr(2\delta - 1)) \sum_{j=1}^{i} \frac{P_h^{j-1}}{(2j)!} \left(\frac{1}{h^2}P_h\right)^{i+1-j} W_h^{(2j)} \\
& + \frac{P_h^i}{(2i+2)!} \Big\{ (1 + hr\delta) \left( w^{(2i+2)}(\zeta_{11}), \ldots, w^{(2i+2)}(\zeta_{1N}) \right)^\top \\
& + (1 + hr(\delta - 1)) \left( w^{(2i+2)}(\zeta_{21}), \ldots, w^{(2i+2)}(\zeta_{2N}) \right)^\top \Big\}
\end{aligned}$$

mit

$$\zeta_{1k} \in (x_k, x_{k+1}), \quad \zeta_{2k} \in (x_{k-1}, x_k), \quad k = 1, \ldots, N, \quad W_h^{(l)} = \left( w^{(l)}(x_1), \ldots, w^{(l)}(x_N) \right)^\top.$$

Da

$$\|P_h\| \leq \|P_h\|_\infty = |1 + hr(\delta - 1)| + |2 - hr(1 - 2\delta)| + |1 + hr\delta| \leq 4 + 3h|r|\delta$$

gilt und aus $w \in C^{2(i+1)}$ auch $w^{(2j+1)} \in C^{2(i-j)}$ und $w^{(2j)} \in C^{2(i+1-j)}$ folgen, erhält man daraus die Gültigkeit der Behauptung für $m = i + 1$. □

Eine analoge Aussage kann man auch für vektorwertige Funktionen mit periodischen Randbedingungen auf dem $d$-dimensionalen Ortsgitter (4.44) beweisen. Für periodische Randbedingungen gilt damit der folgende Satz:

**Satz 5.25** Ist die exakte Lösung für $m \in \mathbb{N}$ $2m$-mal stetig differenzierbar, so ist $\|D^m U_{\vec{h}}^{(l)}(t)\|$ für periodische Randbedingungen in jeder Raumdimension unabhängig von $\vec{h}$ beschränkt. □

Zusammengefaßt liefern die vorangegangenen Untersuchungen folgendes Resultat: Für bestimmte Beispiele kann man erwarten, daß das numerische Verfahren bei Aufgaben mit Dirichlet-Randbedingungen mit einer um $\frac{1}{2}$ niedrigeren Zeitordnung als bei Neumann-Randbedingungen konvergiert. Im Fall homogener Dirichlet- oder Neumann-Randbedingungen im Sinne

von (5.57) konvergiert das Verfahren mit einer um bis zu eins höheren Zeitordnung als bei inhomogenen Randbedingungen. Dieses Verhalten ist auch bei semilinearen skalaren parabolischen Anfangsrandwertproblemen bekannt, vgl. Ostermann/Roche [42]. Interessant ist hier, daß nicht die Randwerte selbst homogen sein müssen, sondern nur ihre $(q+1)$-ten Zeitableitungen, siehe Beispiel 6.14. Werden nur periodische Randbedingungen vorgeschrieben, so sind dagegen keine gebrochenen Konvergenzordnungen zu erwarten.
Die Zeitordnung des Verfahrens hängt bei PDA-Systemen im allgemeinen nicht nur von den Randbedingungen, sondern auch von ihrem differentiellen Zeitindex ab, wie im folgenden Abschnitt gezeigt wird.

### 5.2.3 Konvergenz in Abhängigkeit vom Zeitindex

Um den Einfluß des in Definition 4.3 eingeführten differentiellen Zeitindex $\nu_{dt}$ des linearen PDA-Systems auf die zeitliche Konvergenzordnung des Verfahrens (5.3) zu untersuchen, wird zunächst eine Weierstraß-Kronecker-Transformation (4.68) angewendet.
Seien mit einer regulären Matrix $T_{\mathfrak{A}}$

$$J_{\mathfrak{A}} = T_{\mathfrak{A}}^{-1}\mathfrak{A}T_{\mathfrak{A}} \tag{5.58}$$

die Jordansche Normalform von $\mathfrak{A}$ und $\mathfrak{A}$ regulär. Dann gilt mit (4.68)

$$\begin{aligned}\det G(A,\tau D_{\vec{k}}) &= \det\left(I_s\otimes A - \tau J_{\mathfrak{A}}\otimes D_{\vec{k}}\right) = \prod_{i=1}^{s}\det\left(A-\tau\lambda_{\mathfrak{A},i}D_{\vec{k}}\right)\\ &= \frac{1}{\left(\det(P_{\vec{k}})\det(Q_{\vec{k}})\right)^s}\prod_{i=1}^{s}(-\tau\lambda_{\mathfrak{A},i})^{\sum_{l=1}^{l_{\vec{k}}}m_{\vec{k}l}}\prod_{l=1}^{s_{\vec{k}}}\left(1-\tau\lambda_{\mathfrak{A},i}\kappa_{\vec{k}l}\right)^{n_{\vec{k}l}}\end{aligned}$$

Dabei sind die $\lambda_{\mathfrak{A},i}$, $i=1,\dots,s$, die Eigenwerte der Matrix $\mathfrak{A}$, es gilt $\lambda_{\mathfrak{A},i}\neq 0$. Es gilt das folgende Lemma:

**Lemma 5.26** Für $A$-stabile Runge-Kutta-Verfahren gilt $\mathfrak{Re}\lambda_{\mathfrak{A},j}\geq 0$. Ist für ein $\kappa\in\mathbb{C}$ $\mathfrak{Re}\kappa\leq 0$, so folgt $\tau\lambda_{\mathfrak{A},j}\kappa\neq 1$ für alle $\tau\geq 0$. □

**Beweis:** Aus der $A$-Stabilität des Verfahrens folgt, daß die Stabilitätsfunktion $R(z)=1+zb^\top(I_s-z\mathfrak{A})^{-1}\mathbb{1}_s$ für alle $z\in\mathbb{C}$ mit $\mathfrak{Re}z\leq 0$ definiert ist und damit für diese $z$ die Matrix $I_s-z\mathfrak{A}$ regulär ist. Wegen

$$\det(I_s-z\mathfrak{A}) = \det(I_s-zJ_{\mathfrak{A}}) = \prod_{j=1}^{s}(1-z\lambda_{\mathfrak{A},j})$$

folgt daraus $1-z\lambda_{\mathfrak{A},j}\neq 0$ für alle $z\in\mathbb{C}$ mit $\mathfrak{Re}z\leq 0$ und insbesondere für $z=\tau\kappa$. □

Für $A$-stabile Runge-Kutta-Verfahren mit regulärer Verfahrensmatrix ist damit für $\mathfrak{Re}\kappa_{\vec{k}l}\leq 0$, $l=1,\dots,s_{\vec{k}}$, die Matrix $G(A,\tau D_{\vec{k}})$ regulär. Entsprechend den Definitionsgleichungen (5.8) - (5.10) existieren dann $J(A,\tau D_{\vec{k}})$, $R(A,\tau D_{\vec{k}})$ und $L(A,\tau D_{\vec{k}})$, und aus (5.29), (5.30) und (5.33) erhält man

$$\begin{aligned}J(A,\tau D_{\vec{k}}) = Q_{\vec{k}}\Bigg(\operatorname{diag}\Bigg\{&\left(J(I_{n_{\vec{k}1}},\tau R_{\vec{k}1})\right)_j,\dots,\left(J(I_{n_{\vec{k}s_{\vec{k}}}},\tau R_{\vec{k}s_{\vec{k}}})\right)_j,\left(J(N_{m_{\vec{k}1}},\tau I_{m_{\vec{k}1}})\right)_j,\\ &\dots,\left(J(N_{m_{\vec{k}l_{\vec{k}}}},\tau I_{m_{\vec{k}l_{\vec{k}}}})\right)_j\Bigg\}\Bigg)_{j=1,\dots,s}(I_s\otimes P_{\vec{k}})\end{aligned} \tag{5.59}$$

und entsprechend

$$R(A,\tau D_{\vec{k}}) = Q_{\vec{k}}\mathrm{diag}\left\{\ldots,R(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1}),\ldots,R(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}}),\ldots\right\}Q_{\vec{k}}^{-1}, \tag{5.60}$$

$$L(A,\tau D_{\vec{k}}) = Q_{\vec{k}}\left(\mathrm{diag}\left\{\ldots,\left(L(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})\right)_j,\ldots,(L(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}}))_j,\ldots\right\}\right)_{j=1,\ldots,s}$$
$$\left(I_s\otimes Q_{\vec{k}}^{-1}\right).$$

Damit folgt aus (5.41) für den globalen Gesamtdiskretisierungsfehler $e_{m+1}$ die Darstellung

$$\begin{aligned}
e_{m+1} = \tau\sum_{v=1}^{d}h_v^{p_v}Q\sum_{i=0}^{m}\Bigg(&\mathrm{diag}_{\vec{k}}\left\{Q_{\vec{k}}\mathrm{diag}\left\{\ldots,\left(R(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})^iJ(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})\right)_j,\ldots,\right.\right.\\
&\left.\left.\left(R(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}})^iJ(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}})\right)_j,\ldots\right\}P_{\vec{k}}B_v\right\}_{j=1,\ldots,s}\left(I_s\otimes Q^{-1}\right)\Gamma_{m+1-i}^{(v)}\\
&+Q\sum_{i=0}^{m}\mathrm{diag}_{\vec{k}}\left\{Q_{\vec{k}}\mathrm{diag}\left\{\ldots,R(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})^i,\ldots,R(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}})^i,\ldots\right\}Q_{\vec{k}}^{-1}\right\}Q^{-1}\delta_{m+1-i}\\
&+Q\sum_{i=0}^{m}\mathrm{diag}_{\vec{k}}\left\{Q_{\vec{k}}\mathrm{diag}\left\{\ldots,\left(R(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})^iL(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})\right)_j,\ldots,\right.\right.\\
&\left.\left.\left(R(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}})^iL(N_{m_{\vec{k}j_2}},\tau I_{m_{\vec{k}j_2}})\right)_j,\ldots\right\}Q_{\vec{k}}^{-1}\right\}_{j=1,\ldots,s}(I_s\otimes Q^{-1})\mathcal{O}(\tau^{p+1})\\
&+Q\sum_{r=q+1}^{p}\frac{\tau^{r+1+\alpha_r}}{r!}\Bigg(\mathrm{diag}_{\vec{k}}\left\{W_{r\alpha_r}(A,\tau D_{\vec{k}})\right\}Q^{-1}D^{1+\alpha_r}U_{\vec{h}}^{(r)}(t_m)\\
&-\sum_{i=0}^{m-1}\int_{t_i}^{t_{i+1}}\mathrm{diag}_{\vec{k}}\left\{R(A,\tau D_{\vec{k}})^{m-i}W_{r\alpha_r}(A,\tau D_{\vec{k}})\right\}Q^{-1}D^{1+\alpha_r}U_{\vec{h}}^{(r+1)}(s)ds\\
&-\mathrm{diag}_{\vec{k}}\left\{R(A,\tau D_{\vec{k}})^{m+1}W_{r\alpha_r}(A,\tau D_{\vec{k}})\right\}Q^{-1}D^{1+\alpha_r}U_{\vec{h}}^{(r)}(t_0)\Bigg).
\end{aligned}$$

Im folgenden wird vorausgesetzt, daß

$$\|Q_{\vec{k}}\mathrm{diag}\{N_{n_{\vec{k}1}}^i,\mathfrak{o},\ldots,\mathfrak{o}\}Q_{\vec{k}}^{-1}\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N_{m_{\vec{k}l_{\vec{k}}}}^i\}Q_{\vec{k}}^{-1}\| \tag{5.61a}$$

und

$$\|Q_{\vec{k}}\mathrm{diag}\{N_{n_{\vec{k}1}}^i,\mathfrak{o},\ldots,\mathfrak{o}\}P_{\vec{k}}B_v\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N_{m_{\vec{k}l_{\vec{k}}}}^i\}P_{\vec{k}}B_v\| \tag{5.61b}$$

für $i=0,\ldots,\nu_{dt}-1$ und alle hinreichend kleinen $\vec{h}$ beschränkt sind.
Für die Ableitungen der Funktionen (vgl. (5.8) - (5.10))

$$J(z) = J(1,z) = b^\top(I_s-z\mathfrak{A})^{-1}, \tag{5.62a}$$

$$\tilde{J}(z) = J(z,\tau) = b^\top(I_sz-\tau\mathfrak{A})^{-1} = \frac{1}{z}J\left(\frac{\tau}{z}\right), \tag{5.62b}$$

$$R(z) = R(1,z) = 1+zJ(z)\mathbb{1}_s,$$

$$\tilde{R}(z) = R(z,\tau) = 1+\tau\tilde{J}(z)\mathbb{1}_s = R\left(\frac{\tau}{z}\right), \tag{5.62c}$$

$$L(z) = L(1,z) = zJ(z),$$

$$\tilde{L}(z) = L(z,\tau) = \tau\tilde{J}(z) \tag{5.62d}$$

gelten

$$J^{(k)}(z) = b^\top(I_s-z\mathfrak{A})^{-1-k}\mathfrak{A}^kk!,\quad J^{(k)}(0) = b^\top\mathfrak{A}^kk!,\quad k\geq 0, \tag{5.63a}$$

$$\tilde{J}^{(k)}(z) = b^\top (-1)^k k! (I_s z - \tau\mathfrak{A})^{-1-k}, \quad \tilde{J}^{(k)}(0) = -b^\top \mathfrak{A}^{-k-1} \frac{k!}{\tau^{1+k}}, \quad k \geq 0, \tag{5.63b}$$

$$R(0) = 1, \quad R^{(k)}(0) = kJ^{(k-1)}(0)\mathbb{1}_s = k! b^T \mathfrak{A}^{k-1} \mathbb{1}_s, \quad k \geq 1, \tag{5.63c}$$

$$\tilde{R}(0) = 1 - b^\top \mathfrak{A}^{-1} \mathbb{1}_s, \quad \tilde{R}^{(k)}(0) = -b^\top \mathfrak{A}^{-k-1} \mathbb{1}_s \frac{k!}{\tau^k}, \quad k \geq 1, \tag{5.63d}$$

$$L^{(k)}(z) = J^{(k)}(z) z + k J^{(k-1)}(z), \quad k \geq 0 \tag{5.63e}$$

$$= J^{(k)}(z) \mathfrak{A}^{-1}, \quad k \geq 1. \tag{5.63f}$$

Mit der Fehlerwachstumsfunktion

$$\varphi_R(x) = \sup_{\mathfrak{Re} z \leq x} |R(z)|$$

folgt aus Satz 2.13: Wenn

$$\varphi_R(\mu_2[R_{\vec{k}j_1}]) \leq 1 \tag{5.64}$$

gilt, so ist

$$\|R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})^i\| \leq \|R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})\|^i \leq 1$$

und damit

$$\sum_{i=0}^{m} \|R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})^i \tau^{p+1}\| = \mathcal{O}(\tau^p).$$

Es gilt nach (2.26) $\mu_2[R_{\vec{k}j_1}] = \frac{1}{2}\lambda_{\max}(R_{\vec{k}j_1} + \overline{R}_{\vec{k}j_1}^\top)$. Die Matrix $\frac{1}{2}(R_{\vec{k}j_1} + \overline{R}_{\vec{k}j_1}^\top)$ hat nach (4.68c) die Gestalt (A.1) mit $a = c = \frac{1}{2}$, $b = \mathfrak{Re}\kappa_{\vec{k}j_1}$. Damit folgt aus (A.15)

$$\mu_2[R_{\vec{k}j_1}] = \mathfrak{Re}\kappa_{\vec{k}j_1} + \cos\frac{\pi}{n_{\vec{k}j_1} + 1}.$$

(5.64) ist folglich für $A$-stabile Runge-Kutta-Verfahren erfüllt, wenn

$$\mathfrak{Re}\kappa_{\vec{k}j_1} \leq -\cos\frac{\pi}{n_{\vec{k}j_1} + 1}$$

gilt.
$\|L(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})\|$ und $\|J(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})\|$ sind nach (2.25) genau dann beschränkt, wenn $|L^{(k)}(\tau\kappa_{\vec{k}j_1})|$ und $|J^{(k)}(\tau\kappa_{\vec{k}j_1})|$ für $k = 0, \ldots, n_{\vec{k}j_1} - 1$ beschränkt sind, das ist nach (5.63e) und (5.63f) der Fall, wenn $|\tau\kappa_{\vec{k}j_1} J(\tau\kappa_{\vec{k}j_1})|$ und $|J^{(k)}(\tau\kappa_{\vec{k}j_1})|$ für $k = 0, \ldots, n_{\vec{k}j_1} - 1$ beschränkt sind, wofür nach (5.63a) die Beschränktheit von $\|\tau\kappa_{\vec{k}j_1}(I_s - \tau\kappa_{\vec{k}j_1}\mathfrak{A})^{-1}\|$ und $\|(I_s - \tau\kappa_{\vec{k}j_1}\mathfrak{A})^{-1}\|$ ausreichend ist. Wegen (5.58) und (2.25) gilt

$$\begin{aligned}\|\tau\kappa_{\vec{k}j_1}(I_s - \tau\kappa_{\vec{k}j_1}\mathfrak{A})^{-1}\| &\leq \|T_{\mathfrak{A}}\| \cdot \|T_{\mathfrak{A}}^{-1}\| \cdot \|\tau\kappa_{\vec{k}j_1}(I_s - \tau\kappa_{\vec{k}j_1} J_{\mathfrak{A}})^{-1}\| \\ &\leq \|T_{\mathfrak{A}}\| \cdot \|T_{\mathfrak{A}}^{-1}\| \cdot \sum_{l=1}^{s} \sum_{i=0}^{s-1} \left| \frac{\tau\kappa_{\vec{k}j_1}}{1 - \tau\kappa_{\vec{k}j_1}\lambda_{\mathfrak{A},l}} \right|^{i+1} \|N_s^i\|,\end{aligned}$$

wobei $N_s$ eine nilpotente Matrix gemäß (4.68c) ist.
Beschränktheit von $\|L(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})\|$ und $\|J(I_{n_{\vec{k}j_1}}, \tau R_{\tau\vec{k}j_1})\|$ liegt damit vor, wenn $\left| \frac{\tau\kappa_{\vec{k}j_1}}{1 - \tau\kappa_{\vec{k}j_1}\lambda_{\mathfrak{A},l}} \right|$ und $\frac{1}{|1 - \tau\kappa_{\vec{k}j_1}\lambda_{\mathfrak{A},l}|}$ beschränkt bleiben, was nach Lemma 5.26 wegen $\mathfrak{Re}\kappa_{\vec{k}j_1} \leq 0$ der Fall ist. Unter obigen Voraussetzungen gelten dann auch

$$\sum_{i=0}^{m} \|R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})^i L(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}) \tau^{p+1}\| = \mathcal{O}(\tau^p)$$

und

$$\sum_{i=0}^{m} \|R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})^i J(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})\tau^{p+1}\| = \mathcal{O}(\tau^p).$$

Mit den in (5.21) eingeführten $\tilde{c}^k$ erhält man aus (5.19) durch Taylor-Entwicklung

$$\Gamma_{m+1}^{(v)} = \sum_{k=0}^{\nu_{dt}-1} \frac{\tau^k}{k!}\tilde{c}^k \otimes \frac{\partial^k}{\partial t^k}\gamma^{(v)}(t_m) + \mathcal{O}(\tau^{\nu_{dt}})$$

und

$$\begin{aligned}
\Gamma_{m+1-i}^{(v)} &= \sum_{k=0}^{\nu_{dt}-1}\sum_{l=0}^{\nu_{dt}-1-k} \frac{\tau^{k+l}(-i)^l}{k!l!}\tilde{c}^k \otimes \frac{\partial^{k+l}}{\partial t^{k+l}}\gamma^{(v)}(t_m) + \mathcal{O}(\tau^{\nu_{dt}}) \\
&= \sum_{k=0}^{\nu_{dt}-1}\sum_{l=k}^{\nu_{dt}-1} \frac{\tau^l(-i)^{l-k}}{k!(l-k)!}\tilde{c}^k \otimes \frac{\partial^l}{\partial t^l}\gamma^{(v)}(t_m) + \mathcal{O}(\tau^{\nu_{dt}}) \\
&= \sum_{l=0}^{\nu_{dt}-1}\sum_{k=0}^{l} \frac{\tau^l(-i)^{l-k}}{k!(l-k)!}\tilde{c}^k \otimes \frac{\partial^l}{\partial t^l}\gamma^{(v)}(t_m) + \mathcal{O}(\tau^{\nu_{dt}}),
\end{aligned} \tag{5.65}$$

wobei $0^0 = 1$ gesetzt wird. Daraus folgt für den Term

$$\begin{aligned}
a = \tau Q \sum_{i=0}^{m} \Bigg( \mathrm{diag}_{\vec{k}} \Bigg\{ Q_{\vec{k}} \mathrm{diag} \Bigg\{ \mathfrak{o}, \ldots, \mathfrak{o}, \Big( R(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}})^i J(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}}) \Big)_j, \ldots, \\
\Big( R(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})^i J(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}}) \Big)_j \Bigg\} P_{\vec{k}} B_v \Bigg\} Q^{-1} \Bigg)_{j=1,\ldots,s} \Gamma_{m+1-i}^{(v)}
\end{aligned}$$

des Gesamtdiskretisierungsfehlers

$$\begin{aligned}
a = Q \sum_{j_2=1}^{l_{\vec{k}}} \sum_{j=0}^{\nu_{dt}-1} \frac{1}{j!} \sum_{l=0}^{\nu_{dt}-1} \mathrm{diag}_{\vec{k}} \Bigg\{ Q_{\vec{k}} \mathrm{diag} \Bigg\{ \mathfrak{o}, \ldots, \mathfrak{o}, \frac{N^j_{m_{\vec{k}j_2}}}{\tau^{j-l}} \sum_{k=0}^{l} \frac{1}{k!(l-k)!} \cdot \\
\cdot \sum_{i=0}^{m} (-i)^{l-k} \left( \tau^{j+1} \tilde{R}(z)^i \tilde{J}(z) \tilde{c}^k \right)^{(j)} (0), \mathfrak{o}, \ldots, \mathfrak{o} \Bigg\} P_{\vec{k}} B_v \Bigg\} Q^{-1} \frac{\partial^l}{\partial t^l} \gamma^{(v)}(t_m) \\
+ \sum_{j_2=1}^{l_{\vec{k}}} \sum_{j=0}^{\nu_{dt}-1} \frac{1}{j!} \sum_{i=0}^{m} \Bigg( \mathrm{diag}_{\vec{k}} \Bigg\{ Q_{\vec{k}} \mathrm{diag} \Big\{ \mathfrak{o}, \ldots, \mathfrak{o}, N^j_{m_{\vec{k}j_2}} \left( \tau^{j+1} \tilde{R}(z)^i \tilde{J}(z) \right)^{(j)}_l (0), \mathfrak{o}, \ldots \\
\ldots, \mathfrak{o} \Big\} P_{\vec{k}} B_v \Bigg\} Q^{-1} \Bigg)_{l=1,\ldots,s} \mathcal{O}(\tau^{\nu_{dt}-j}).
\end{aligned}$$

Bei hinreichend glatter exakter Lösung reicht es für die Beschränktheit von $a$ unter der Voraussetzung (5.61b) deshalb aus, wenn

$$\sum_{k=0}^{l} \frac{1}{k!(l-k)!} \sum_{i=0}^{m} (-i)^{l-k} \left( \tau^{j+1} \tilde{R}(z)^i \tilde{J}(z) \tilde{c}^k \right)^{(j)} (0) = 0 \tag{5.66a}$$

für $l = 0, \ldots, j-1,\ j = 1, \ldots, \nu_{dt} - 1$ gilt und

$$\tau^{l-j} \sum_{k=0}^{l} \frac{1}{k!(l-k)!} \sum_{i=0}^{m} (-i)^{l-k} \left( \tau^{j+1} \tilde{R}(z)^i \tilde{J}(z) \tilde{c}^k \right)^{(j)} (0) \tag{5.66b}$$

für $l = j, \ldots, \nu_{dt} - 1$, $j = 0, \ldots, \nu_{dt} - 1$ und

$$\tau^{\nu_{dt}-j} \sum_{i=0}^{m} \left| \left( \tau^{j+1} \tilde{R}(z)^i \tilde{J}(z) \right)^{(j)} (0) \right| \tag{5.66c}$$

für $j = 0, \ldots, \nu_{dt} - 1$ beschränkt bleiben (der Betrag ist hier wieder komponentenweise zu verstehen).
Für $\nu_{dt} = 1$ heißt dies wegen (5.63b) und (5.63d), daß

$$\sum_{i=0}^{m} \tau \tilde{R}(0)^i \tilde{J}(0) \mathbb{1}_s = \sum_{i=0}^{m} (1 - b^\top \mathfrak{A}^{-1} \mathbb{1}_s)^i (-b^\top \mathfrak{A}^{-1} \mathbb{1}_s)$$

und

$$\tau \sum_{i=0}^{m} |\tau \tilde{R}(0)^i \tilde{J}(0)| = \tau \sum_{i=0}^{m} |(1 - b^\top \mathfrak{A}^{-1} \mathbb{1}_s)^i (-b^\top \mathfrak{A}^{-1})|$$

für $\tau \to 0$ ($m \to \infty$) beschränkt bleiben und also

$$|1 - b^\top \mathfrak{A}^{-1} \mathbb{1}_s| = |\tilde{R}(0)| = \lim_{z \to -\infty} |R(z)| \leq 1$$

gelten muß, dies ist für A-stabile Runge-Kutta-Verfahren erfüllt.
Für $\nu_{dt} > 1$ folgt aus (5.66a) mit $j = 1$ wegen (5.62c)

$$\begin{aligned} 0 &= \sum_{i=0}^{m} \tau^2 \left( \tilde{R}(z)^i \tilde{J}(z) \mathbb{1}_s \right)' (0) = \sum_{i=0}^{m} \tau \left( \tilde{R}(z)^i (\tilde{R}(z) - 1) \right)' (0) \\ &= \tau \left( \tilde{R}(z)^{m+1} - 1 \right)' (0) = (m+1) \tilde{R}(0)^m \tau \tilde{R}'(0). \end{aligned}$$

Für $\nu_{dt} > 1$ und $\tilde{R}'(0) \neq 0$, d. h. nach (5.63d) $b^\top \mathfrak{A}^{-2} \mathbb{1}_s \neq 0$, kann (5.66a) deshalb nur erfüllt werden, wenn

$$\tilde{R}(0) = \lim_{z \to -\infty} R(z) = 0 \quad \text{bzw.} \quad b^\top \mathfrak{A}^{-1} \mathbb{1}_s = 1$$

gilt, dies ist für $L$-stabile Runge-Kutta-Verfahren erfüllt. In (5.66) kann dann die Summationsgrenze $m$ für $m \geq \nu_{dt} - 1$ durch $j$ ersetzt werden, und nach (5.63b) und (5.63d) gilt

$$\left( \tilde{R}(z)^i \tilde{J}(z) \right)^{(j)} (0) = \mathcal{O}(\frac{1}{\tau^{j+1}}).$$

Damit sind (5.66b) und (5.66c) für alle $\nu_{dt}$ beschränkt, und aus (5.66a) erhält man die Bedingung

$$\sum_{k=0}^{l} \frac{1}{k!(l-k)!} \sum_{i=0}^{j} (-i)^{l-k} \left( \tau^{j+1} \tilde{R}(z)^i \tilde{J}(z) \tilde{c}^k \right)^{(j)} (0) = 0 \tag{5.67}$$

für $l = 0, \ldots, j-1,\ j = 1, \ldots, \nu_{dt} - 1$.
Analog zu der Taylor-Entwicklung (5.65) für $\Gamma^{(v)}_{m+1-i}$ folgt für den gemäß (5.16a) definierten Residuenfehler unter Berücksichtigung der Konsistenzbedingung (2.16)

$$\delta_{m+1-i} = \sum_{l=p+1}^{p+\nu_{dt}-2} \sum_{k=p+1}^{l} \frac{\tau^l (-i)^{l-k} (1 - k b^\top \tilde{c}^{k-1})}{k!(l-k)!} U^{(l)}_{\vec{h}}(t_m) + \mathcal{O}(\tau^{p+\nu_{dt}-1}).$$

Damit gilt für $v \in \mathbb{N},\ v \leq p$

$$Q \sum_{i=0}^{m} \mathrm{diag}_{\vec{k}} \left\{ Q_{\vec{k}} \mathrm{diag} \left\{ \mathfrak{o}, \ldots, \mathfrak{o}, R(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}})^i, \ldots, R(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})^i \right\} Q^{-1}_{\vec{k}} \right\} Q^{-1} \delta_{m+1-i} \tag{5.68}$$

$$\begin{aligned} &= Q \sum_{j_2=1}^{l_{\vec{k}}} \tau^v \sum_{j=0}^{\nu_{dt}-1} \frac{1}{j!} \sum_{l=p+1}^{p+\nu_{dt}-2} \mathrm{diag}_{\vec{k}} \Big\{ Q_{\vec{k}} \mathrm{diag} \Big\{ \mathfrak{o}, \ldots, \mathfrak{o}, \frac{N^j_{m_{\vec{k}j_2}}}{\tau^{v+j-l}} \sum_{k=p+1}^{l} \frac{(1 - k b^\top \tilde{c}^{k-1})}{k!(l-k)!} \cdot \\ &\cdot \sum_{i=0}^{m} (-i)^{l-k} \left( \tau^j \tilde{R}(z)^i \right)^{(j)} (0), \mathfrak{o}, \ldots, \mathfrak{o} \Big\} Q^{-1}_{\vec{k}} \Big\} Q^{-1} U^{(l)}_{\vec{h}}(t_m) + \sum_{j_2=1}^{l_{\vec{k}}} \sum_{j=0}^{\nu_{dt}-1} \frac{\tau^{\nu_{dt}-j-1}}{j!} \cdot \\ &\cdot \sum_{i=0}^{m} \mathrm{diag}_{\vec{k}} \left\{ Q_{\vec{k}} \mathrm{diag} \left\{ \mathfrak{o}, \ldots, \mathfrak{o}, N^j_{m_{\vec{k}j_2}} \left( \tau^j \tilde{R}(z)^i \right)^{(j)} (0), \mathfrak{o}, \ldots, \mathfrak{o} \right\} Q^{-1}_{\vec{k}} \right\} Q^{-1} \mathcal{O}(\tau^p). \end{aligned}$$

Unter obigen Voraussetzungen ist (5.68) deshalb gleich $\mathcal{O}(\tau^v)$, falls

$$\sum_{k=p+1}^{l} \frac{(1-kb^\top \tilde{c}^{k-1})}{k!(l-k)!} \sum_{i=0}^{j} (-i)^{l-k} \left(\tau^j \tilde{R}(z)^i\right)^{(j)}(0) = 0 \tag{5.69}$$

gilt für $l = p+1, \ldots, v+j-1$, $j = 0, \ldots, \nu_{dt}-1$.
Analog erhält man

$$\begin{aligned} & Q \sum_{i=0}^{m} \operatorname{diag}_{\vec{k}} \Bigg\{ Q_{\vec{k}} \operatorname{diag} \Bigg\{ \mathfrak{o}, \ldots, \mathfrak{o}, \left( R(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}})^i L(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}}) \right)_j, \ldots, \\ & \left( R(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})^i L(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}}) \right)_j \Bigg\} \left( I_s \otimes Q_{\vec{k}}^{-1} \right) \Bigg\}_{j=1,\ldots,s} (I_s \otimes Q^{-1}) \mathcal{O}(\tau^{p+1}) \\ = \; & \mathcal{O}(\tau^v), \end{aligned}$$

falls

$$\sum_{k=p+1}^{l} \frac{1}{k!(l-k)!} \sum_{i=0}^{j} (-i)^{l-k} \left(\tau^j \tilde{R}(z)^i \tilde{L}(z)[\tilde{c}^k - k\mathfrak{A}\tilde{c}^{k-1}]\right)^{(j)}(0) = 0 \tag{5.70}$$

für $l = p+1, \ldots, v+j-1$, $j = 0, \ldots, \nu_{dt}-1$ gilt.
Sei

$$p^\star_{\nu_{dt}} = \max\{v \le p : (5.69) \text{ und } (5.70) \text{ sind für } l = p+1, \ldots, v+j-1,\ j = 0, \ldots, \nu_{dt}-1 \text{ erfüllt}\}.$$

Dann gelten

$$p^\star_{\nu_{dt}} = p \quad \text{für} \quad \nu_{dt} \le 2 \tag{5.71a}$$

und

$$p^\star_{\nu_{dt}} \ge p + 2 - \nu_{dt} \quad \text{für} \quad \nu_{dt} \ge 3. \tag{5.71b}$$

Im folgenden wird zunächst $\alpha_r \in \{-1, 0\}$ angenommen. Für $\alpha_r = -1$ erhält man durch Einsetzen der für $R(A, \tau D_{\vec{k}})$, $J(A, \tau D_{\vec{k}})$ und $D_{\vec{k}}$ geltenden Gleichungen (5.60), (5.59) und (4.68b) in (5.27)

$$\begin{aligned} & \operatorname{diag}\left\{\ldots, I_{n_{\vec{k}j_1}} - R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}), \ldots, I_{m_{\vec{k}j_2}} - R(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}), \ldots\right\} Q_{\vec{k}}^{-1} W_{r(-1)}(A, \tau D_{\vec{k}}) Q_{\vec{k}} \\ = \; & \operatorname{diag}\Big\{\ldots, J(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}) \left( \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes I_{n_{\vec{k}j_1}} \right) \tau R_{\vec{k}j_1}, \\ & \quad \ldots, \tau J(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}) \left( \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes I_{m_{\vec{k}j_2}} \right), \ldots \Big\} \end{aligned}$$

und für $\alpha_r = 0$

$$\begin{aligned} & \operatorname{diag}\left\{\ldots, I_{n_{\vec{k}j_1}} - R(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}), \ldots, I_{m_{\vec{k}j_2}} - R(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}), \ldots\right\} \\ & Q_{\vec{k}}^{-1} W_{r(-1)}(A, \tau D_{\vec{k}}) P_{\vec{k}}^{-1} \operatorname{diag}\left\{\ldots, R_{\vec{k}j_1}, \ldots, I_{m_{\vec{k}j_2}}, \ldots\right\} \\ = \; & \operatorname{diag}\Big\{\ldots, J(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}) \left( \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes I_{n_{\vec{k}j_1}} \right) R_{\vec{k}j_1}, \\ & \quad \ldots, J(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}) \left( \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] \otimes I_{m_{\vec{k}j_2}} \right), \ldots \Big\}. \end{aligned}$$

$W_{r\alpha_r}(A, \tau D_{\vec{k}})$ existiert daher genau dann, wenn für alle $j_1$ und $j_2$ $W_{r\alpha_r}(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})$ und $W_{r\alpha_r}(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}})$ existieren, in diesem Fall können

$$\begin{aligned} W_{r(-1)}(A, \tau D_{\vec{k}}) &= Q_{\vec{k}} \operatorname{diag}\left\{\ldots, W_{r(-1)}(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}), \ldots, W_{r(-1)}(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}), \ldots\right\} Q_{\vec{k}}^{-1}, \\ W_{r0}(A, \tau D_{\vec{k}}) &= Q_{\vec{k}} \operatorname{diag}\left\{\ldots, W_{r0}(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1}), \ldots, W_{r0}(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}), \ldots\right\} P_{\vec{k}} \end{aligned}$$

gesetzt werden.
Ist $(I_{m_{\vec{k}j_2}} - R(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}}))$ regulär, so existiert $W_{r\alpha_r}(N_{m_{\vec{k}j_2}}, \tau I_{m_{\vec{k}j_2}})$ nach Bemerkung 5.9 eindeutig. Nach der Formel (2.25) für die Matrixfunktion eines Jordan-Blocks ist dies genau dann der Fall, wenn $\tilde{R}(0) \neq 1$ gilt, d. h. $\lim\limits_{z\to-\infty} R(z) \neq 1$. Dann gilt

$$\begin{aligned}
&\operatorname{diag}\Big\{\mathfrak{o}, \dots, \mathfrak{o}, R(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}})^i W_{r\alpha_r}(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}}), \dots, \\
&\qquad\qquad R(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})^i W_{r\alpha_r}(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})\Big\} \\
&= \sum_{j_2=1}^{l_{\vec{k}}} \sum_{j=0}^{\nu_{dt}-1} \operatorname{diag}\left\{\mathfrak{o}, \dots, \mathfrak{o}, \frac{N_{m_{\vec{k}j_2}}^j}{j!} \left(\tilde{R}(z)^i \tilde{W}_{r\alpha_r}(z)\right)^{(j)}(0), \mathfrak{o}, \dots, \mathfrak{o}\right\}
\end{aligned}$$

mit

$$\tilde{W}_{r\alpha_r}(z) = W_{r\alpha_r}(z, \tau) = \frac{1}{\tau^{\alpha_r}} \left(1 - \tilde{R}(z)\right)^{-1} \tilde{J}(z) \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right].$$

Nach (5.63d) und (5.63b) gilt

$$\left(\tilde{R}(z)^i \tilde{W}_{r\alpha_r}(z)\right)^{(j)}(0) = \mathcal{O}\left(\frac{1}{\tau^{j+1+\alpha_r}}\right), \quad j \geq 0.$$

Daraus folgt

$$\begin{aligned}
&\Big\|\operatorname{diag}\Big\{\mathfrak{o}, \dots, \mathfrak{o}, R(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}})^i W_{r\alpha_r}(N_{m_{\vec{k}1}}, \tau I_{m_{\vec{k}1}}), \dots, \\
&\qquad\qquad R(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})^i W_{r\alpha_r}(N_{m_{\vec{k}l_{\vec{k}}}}, \tau I_{m_{\vec{k}l_{\vec{k}}}})\Big\}\Big\| = \mathcal{O}\left(\frac{1}{\tau^{\nu_{dt}+\alpha_r}}\right)
\end{aligned}$$

und, falls für $\alpha_r = -1$ die Matrixnormen in (5.61a) bzw. für $\alpha_r = 0$

$$\|Q_{\vec{k}} \operatorname{diag}\{N_{n_{\vec{k}1}}^i, \mathfrak{o}, \dots, \mathfrak{o}\} P_{\vec{k}}\|, \dots, \|Q_{\vec{k}} \operatorname{diag}\{\mathfrak{o}, \dots, \mathfrak{o}, N_{m_{\vec{k}l_{\vec{k}}}}^i\} P_{\vec{k}}\| \tag{5.72}$$

beschränkt sind (daraus folgt die Beschränktheit der Matrizen in (5.61b)) sowie $W_{r\alpha_r}(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})$ beschränkt ist, unter obigen Voraussetzungen schließlich

$$\begin{aligned}
&\tau^{r+1+\alpha_r} \|R(A, \tau D_{\vec{k}})^i W_{r\alpha_r}(A, \tau D_{\vec{k}})\| \\
&= \begin{cases} \mathcal{O}(\tau^{r+1+\alpha_r}) \ : \ \nu_{dt} = 0 \text{ oder } \tilde{W}_{r\alpha_r}^{(i)}(0) = 0 \text{ für } i = 0, \dots, \nu_{dt} - 1 \\ \mathcal{O}(\tau^{r+1-\nu_{dt}}) \ : \ \text{sonst} \end{cases}
\end{aligned}$$

für $r = q+1, \dots, p$. Nach der Leibnizschen Formel ist

$$\tilde{W}_{r\alpha_r}^{(k)}(z) = \frac{1}{\tau^{\alpha_r}} \sum_{i=0}^{k} \binom{k}{i} \left(\left(1 - \tilde{R}(z)\right)^{-1}\right)^{(k-i)} \tilde{J}^{(i)}(z) \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right].$$

Durch Induktion über $k$ folgt deshalb, daß die Bedingung

$$\tilde{W}_{r\alpha_r}^{(i)}(0) = 0 \quad \text{für} \quad i = 0, \dots, \nu_{dt} - 1$$

äquivalent ist zu

$$\tilde{J}^{(k)}(0) \left[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}\right] = 0, \quad k = 0, \dots, \nu_{dt} - 1.$$

Einsetzen von (5.63b) liefert

$$b^\top \mathfrak{A}^{-k-1} \tilde{c}^r = r b^\top \mathfrak{A}^{-k} \tilde{c}^{r-1}, \quad k = 0, \dots, \nu_{dt} - 1. \tag{5.73}$$

**Bemerkung 5.27** Für steifgenaue Verfahren mit einer Stufenordnung $q \geq 1$ gelten nach der Definitionsgleichung (2.20) der Steifgenauigkeit und den vereinfachenden Bedingungen $C(q)$ und $B(p)$ ((2.17) und (2.16)) mit $k = 1$

$$b^\top \mathfrak{A}^{-1} = e_s^\top \quad \text{und} \quad c_s = \sum_{j=1}^{s} a_{sj} = \sum_{j=1}^{s} b_j = 1.$$

Damit folgt

$$b^\top \mathfrak{A}^{-1} \tilde{c}^r = e_s^\top \tilde{c}^r = 1 \quad \text{für} \quad r = q+1, \ldots, p,$$

Bedingung (5.73) ist bei diesen Verfahren für $\nu_{dt} = 1$ stets erfüllt. □

Sei nun wieder $\alpha_r \geq -1$ beliebig. Ist $W_{r\alpha_r}(A, \tau D_{\vec{k}})$ beschränkt, so folgt unter den obigen Voraussetzungen wegen

$$\|R(A, \tau D_{\vec{k}})^i\| = \mathcal{O}\left(\frac{1}{\tau^{\nu_{dt}-1}}\right)$$

analog

$$\tau^{r+1+\alpha_r} \|R(A, \tau D_{\vec{k}})^i W_{r\alpha_r}(A, \tau D_{\vec{k}})\| = \mathcal{O}(\tau^{r+1+\alpha_r - \max\{0, \nu_{dt}-1\}}), \quad r = q+1, \ldots, p.$$

Zusammengefaßt erhält man für Runge-Kutta-Verfahren mit regulärer Verfahrensmatrix den folgenden Konvergenzsatz:

**Satz 5.28** Seien mit $\alpha_r \in \mathbb{R}$, $\alpha_r \geq -1$, für $\tau \to 0$ ($M_e \to \infty$), $h_v \to 0$ ($N_v \to \infty$), $v = 1, \ldots, d$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die folgenden Voraussetzungen erfüllt:

(a) Es existieren für die Matrizenbüschel $D_{\vec{k}} + \lambda A$ Weierstraß-Kronecker-Zerlegungen gemäß (4.68) mit beschränkten Normen

$$\|Q_{\vec{k}} \mathrm{diag}\{N^i_{n_{\vec{k}1}}, \mathfrak{o}, \ldots, \mathfrak{o}\} Q^{-1}_{\vec{k}}\|, \ldots, \|Q_{\vec{k}} \mathrm{diag}\{\mathfrak{o}, \ldots, \mathfrak{o}, N^i_{m_{\vec{k}l_{\vec{k}}}}\} Q^{-1}_{\vec{k}}\| \tag{5.74}$$

und

$$\|Q_{\vec{k}} \mathrm{diag}\{N^i_{n_{\vec{k}1}}, \mathfrak{o}, \ldots, \mathfrak{o}\} P_{\vec{k}} B_v\|, \ldots, \|Q_{\vec{k}} \mathrm{diag}\{\mathfrak{o}, \ldots, \mathfrak{o}, N^i_{m_{\vec{k}l_{\vec{k}}}}\} P_{\vec{k}} B_v\|$$

für $i = 0, \ldots, \nu_{dt} - 1$,

(b) $\mathfrak{Re}\kappa_{\vec{k}j_1} \leq -\cos\frac{\pi}{n_{\vec{k}j_1}+1} = \begin{cases} 0 & : n_{\vec{k}j_1} = 1 \\ -\frac{1}{2} & : n_{\vec{k}j_1} = 2 \\ -\frac{1}{2}\sqrt{2} & : n_{\vec{k}j_1} = 3 \\ \vdots & \end{cases}, \quad j_1 = 1, \ldots, l_{\vec{k}},$

(c) für $\nu_{dt} \geq 3$ sei (5.67) für $l = 0, \ldots, j-1,\ j = 1, \ldots, \nu_{dt} - 1$ erfüllt,

(d) im Fall $\alpha_r \notin \{-1, 0\}$ existieren $W_{r\alpha_r}(A, \tau D_{\vec{k}})$ gemäß (5.27) und seien beschränkt, im Fall $\alpha_r \in \{-1, 0\}$ existieren beschränkte Matrizen $W_{r\alpha_r}(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})$ für $j_1 = 1, \ldots, l_{\vec{k}}$, $r = q+1, \ldots, p$, für $\alpha_r = 0$ sind die Matrixnormen in (5.72) beschränkt,

(e) $\|D^{1+\alpha_r} U^{(r)}_{\vec{h}}(t)\|$ und $\|D^{1+\alpha_r} U^{(r+1)}_{\vec{h}}(t)\|$ sind für $r = q+1, \ldots, p$, $t \in [t_0, t_e]$ beschränkt.

Dann ist das Diskretisierungsverfahren (5.3) für $L$-stabile Runge-Kutta-Methoden (im Fall $\nu_{dt} = 0$ reicht $A$-Stabilität, im Fall $\nu_{dt} = 1$ $A$-Stabilität mit $\lim_{z \to -\infty} R(z) \neq 1$) mit regulärer Verfahrensmatrix für lineare Systeme mit genügend glatter exakter Lösung nach $\nu_{dt}$ Zeitschritten konvergent mit der Ordnung $(p_1, \ldots, p_d, p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes mit

$$p^\star = \min\{p^\star_{\nu_{dt}}, p_r : \ r = q+1, \ldots, p\},$$

wobei

$$p_r = \begin{cases} r+1+\alpha_r-\max\{0,\nu_{dt}-1\} & : \alpha_r \notin \{-1,0\} \\ r+1+\alpha_r & : \alpha_r \in \{-1,0\} \text{ und } b^\top \mathfrak{A}^{-i-1}\tilde{c}^r = r b^\top \mathfrak{A}^{-i}\tilde{c}^{r-1} \\ & \quad \text{für } i=0,\ldots,\nu_{dt}-1 \\ r+1-\nu_{dt} & : \text{sonst} \end{cases}$$

gilt. □

**Bemerkung 5.29** Die Ursache dafür, daß im allgemeinen Konvergenz erst nach $\nu_{dt}$ Zeitschritten vorliegt, ist, daß als Startvektor der auf das Ortsgitter $\Omega_{\vec{h}}$ eingeschränkte exakte Anfangswert $U_{\vec{h}}(t_0)$ verwendet wurde, der im Gegensatz zu $U(t_0)$ in (4.67) im allgemeinen kein konsistenter Anfangswert des MOL-DA-Systems (4.45) ist, vgl. Beispiel 6.9. □

**Bemerkung 5.30** Voraussetzung (a) von Konvergenzsatz 5.28 kann wie folgt abgeschwächt werden: Da mit $L(\infty) = \lim_{z\to\infty} L(z)$

$$\begin{aligned} & Q_{\vec{k}}\mathrm{diag}\left\{L(I_{n_{\vec{k}1}},\tau R_{\vec{k}1})_j,\ldots,L(I_{n_{\vec{k}s_{\vec{k}}}},\tau R_{\vec{k}s_{\vec{k}}})_j,\mathfrak{o},\ldots,\mathfrak{o}\right\}Q_{\vec{k}}^{-1} \\ = {} & L(\infty)_j Q_{\vec{k}}\mathrm{diag}\left\{I_{n_{\vec{k}1}},\ldots,I_{n_{\vec{k}s_{\vec{k}}}},\mathfrak{o},\ldots,\mathfrak{o}\right\}Q_{\vec{k}}^{-1} \\ & + Q_{\vec{k}}\mathrm{diag}\left\{L(I_{n_{\vec{k}1}},\tau R_{\vec{k}1})_j - L(\infty)_j,\ldots,L(I_{n_{\vec{k}s_{\vec{k}}}},\tau R_{\vec{k}s_{\vec{k}}})_j - L(\infty)_j,\mathfrak{o},\ldots,\mathfrak{o}\right\}Q_{\vec{k}}^{-1} \end{aligned}$$

gilt und $z(L(z)-L(\infty))$ und für $i \geq 1$ auch $zL^{(i)}(z)$ sowie für $L$-stabile Runge-Kutta-Verfahren $zR^{(i)}(z)$ für $i \geq 0$ beschränkt bleiben, sind $\|\tau\kappa_{\vec{k}j_1}\left(L(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1}) - L(\infty)\right)\|$ und $\|\tau\kappa_{\vec{k}j_1}R(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})\|$ beschränkt. Deshalb kann im Fall $p^\star \leq p-1$ anstelle der Beschränktheit der Matrixnormen in (5.74) diejenige von

$$\|Q_{\vec{k}}\mathrm{diag}\{I_{n_{\vec{k}1}},\ldots,I_{n_{\vec{k}s_{\vec{k}}}},\mathfrak{o},\ldots,\mathfrak{o}\}Q_{\vec{k}}^{-1}\|,$$

$$\left\|Q_{\vec{k}}\mathrm{diag}\left\{\frac{1}{\kappa_{\vec{k}1}}N^i_{n_{\vec{k}1}},\mathfrak{o},\ldots,\mathfrak{o}\right\}Q_{\vec{k}}^{-1}\right\|,\ldots,\left\|Q_{\vec{k}}\mathrm{diag}\left\{\mathfrak{o},\ldots,\mathfrak{o},\frac{1}{\kappa_{\vec{k}s_{\vec{k}}}}N^i_{n_{\vec{k}s_{\vec{k}}}},\mathfrak{o},\ldots,\mathfrak{o}\right\}Q_{\vec{k}}^{-1}\right\|$$

und

$$\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N^i_{m_{\vec{k}1}},\mathfrak{o},\ldots,\mathfrak{o}\}Q_{\vec{k}}^{-1}\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N^i_{m_{\vec{k}l_{\vec{k}}}}\}Q_{\vec{k}}^{-1}\|$$

für $i=0,\ldots,\nu_{dt}-1$ vorausgesetzt werden. □

Während Voraussetzung (e) von Konvergenzsatz 5.28 Gegenstand der Sätze 5.23 und 5.25 war, liefern die folgenden Lemmata Aussagen über die Erfüllung der Voraussetzungen (c) und (d).

**Lemma 5.31** Ist $\nu_{dt}=3$ oder $\nu_{dt}=4$, dann ist für die betrachteten Runge-Kutta-Verfahren Voraussetzung (c) in Satz 5.28 erfüllt, wenn $q \geq \nu_{dt}-2$ ist. □

**Beweis:** Aus der vereinfachenden Bedingung $C(q)$ (siehe (2.17)) folgt für $k=1,\ldots,q$

$$b^\top \mathfrak{A}^{-l}\tilde{c}^k = k b^\top \mathfrak{A}^{-l+1}\tilde{c}^{k-1} = \ldots = k! b^\top \mathfrak{A}^{-l+k}\mathbb{1}_s \tag{5.75a}$$

und mit der vereinfachenden Bedingung $B(p)$ (siehe (2.16)) für $k=1,\ldots,q$, $l \leq k$ und $k-l \leq p-1$

$$b^\top \mathfrak{A}^{-l}\tilde{c}^k = \frac{k!}{(k-l)!}b^\top \tilde{c}^{k-l} = \frac{k!}{(k-l+1)!}. \tag{5.75b}$$

Damit erhält man für $j = 2$ aus (5.67) mit (5.63b), (5.63d) und den für $L$-stabile Verfahren geltenden Beziehungen $b^\top \mathfrak{A}^{-1} \mathbb{1}_s = 1$ bzw. $\tilde{R}(0) = 0$ und $\tilde{J}(0)\mathbb{1}_s = -\frac{1}{\tau}$

$$
\begin{aligned}
&\sum_{i=0}^{2} \left(\tau^3 \tilde{R}(z)^i \tilde{J}(z)\mathbb{1}_s\right)''(0) \\
&= \tau^3 \left(\tilde{J}''(0)\mathbb{1}_s - \frac{1}{\tau}\tilde{R}''(0) + 2\tilde{R}'(0)\tilde{J}'(0)\mathbb{1}_s - \frac{2}{\tau}\tilde{R}'(0)^2\right) \\
&= 2\left(-b^\top \mathfrak{A}^{-3}\mathbb{1}_s + b^\top \mathfrak{A}^{-3}\mathbb{1}_s + (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-2}\mathbb{1}_s) - (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2\right) \\
&= 0,
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_{i=0}^{2} i\left(\tau^3 \tilde{R}(z)^i \tilde{J}(z)\mathbb{1}_s\right)''(0) + \sum_{i=0}^{2}\left(\tau^3 \tilde{R}(z)^i \tilde{J}(z)\tilde{c}\right)''(0) \\
&= \tau^3\left(\frac{1}{\tau}\tilde{R}''(0) - 2\tilde{R}'(0)\tilde{J}'(0)\mathbb{1}_s + \frac{4}{\tau}\tilde{R}'(0)^2\right. \\
&\qquad \left. + \tilde{J}''(0)\tilde{c} + \tilde{R}''(0)\tilde{J}(0)\tilde{c} + 2\tilde{R}'(0)\tilde{J}'(0)\tilde{c} + 2\tilde{R}'(0)^2\tilde{J}(0)\tilde{c}\right) \\
&= 2\left(-b^\top \mathfrak{A}^{-3}\mathbb{1}_s - (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-2}\mathbb{1}_s) + 2(b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2\right. \\
&\qquad \left. - b^\top \mathfrak{A}^{-3}\tilde{c} + (b^\top \mathfrak{A}^{-3}\mathbb{1}_s)(b^\top \mathfrak{A}^{-1}\tilde{c}) + (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-2}\tilde{c}) - (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top \mathfrak{A}^{-1}\tilde{c})\right) \\
&= 2\left(-b^\top \mathfrak{A}^{-3}\mathbb{1}_s + (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2\right. \\
&\qquad \left. - b^\top \mathfrak{A}^{-2}\mathbb{1}_s + (b^\top \mathfrak{A}^{-3}\mathbb{1}_s) + (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-1}\mathbb{1}_s) - (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2\right) \\
&= 0
\end{aligned}
$$

und damit für $L$-stabile Runge-Kutta-Verfahren mit einer Stufenordnung $q \geq 1$ keine weiteren Bedingungen.
Für $j = 3$ ergeben sich analog

$$
\begin{aligned}
&\sum_{i=0}^{3} \left(\tau^4 \tilde{R}(z)^i \tilde{J}(z)\mathbb{1}_s\right)'''(0) \\
&= \tau^4 \left(\tilde{J}'''(0)\mathbb{1}_s - \frac{1}{\tau}\tilde{R}'''(0) + 3\tilde{R}''(0)\tilde{J}'(0)\mathbb{1}_s + 3\tilde{R}'(0)\tilde{J}''(0)\mathbb{1}_s - \frac{6}{\tau}\tilde{R}'(0)\tilde{R}''(0)\right. \\
&\qquad \left. + 6\tilde{R}'(0)^2\tilde{J}'(0)\mathbb{1}_s - \frac{6}{\tau}\tilde{R}'(0)^3\right) \\
&= 6\left(-b^\top \mathfrak{A}^{-4}\mathbb{1}_s + b^\top \mathfrak{A}^{-4}\mathbb{1}_s + (b^\top \mathfrak{A}^{-3}\mathbb{1}_s)(b^\top \mathfrak{A}^{-2}\mathbb{1}_s) + (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-3}\mathbb{1}_s)\right. \\
&\quad \left. - 2(b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-3}\mathbb{1}_s) - (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top \mathfrak{A}^{-2}\mathbb{1}_s) + (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^3\right) \\
&= 0,
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_{i=0}^{3} i\left(\tau^4 \tilde{R}(z)^i \tilde{J}(z)\mathbb{1}_s\right)'''(0) + \sum_{i=0}^{3}\left(\tau^4 \tilde{R}(z)^i \tilde{J}(z)\tilde{c}\right)'''(0) \\
&= \tau^4\left(\frac{1}{\tau}\tilde{R}'''(0) - 3\tilde{R}''(0)\tilde{J}'(0)\mathbb{1}_s - 3\tilde{R}'(0)\tilde{J}''(0)\mathbb{1}_s + \frac{12}{\tau}\tilde{R}'(0)\tilde{R}''(0) - 12\tilde{R}'(0)^2\tilde{J}'(0)\mathbb{1}_s\right. \\
&\qquad + \frac{18}{\tau}\tilde{R}'(0)^3 + \tilde{J}'''(0)\tilde{c} + \tilde{R}'''(0)\tilde{J}(0)\tilde{c} + 3\tilde{R}''(0)\tilde{J}'(0)\tilde{c} + 3\tilde{R}'(0)\tilde{J}''(0)\tilde{c} \\
&\qquad \left. + 6\tilde{R}'(0)\tilde{R}''(0)\tilde{J}(0)\tilde{c} + 6\tilde{R}'(0)^2\tilde{J}'(0)\tilde{c} + 6\tilde{R}'(0)^3\tilde{J}(0)\tilde{c}\right) \\
&= 6\left(-b^\top \mathfrak{A}^{-4}\mathbb{1}_s - (b^\top \mathfrak{A}^{-3}\mathbb{1}_s)(b^\top \mathfrak{A}^{-2}\mathbb{1}_s) - (b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-3}\mathbb{1}_s) + 4(b^\top \mathfrak{A}^{-2}\mathbb{1}_s)(b^\top \mathfrak{A}^{-3}\mathbb{1}_s)\right. \\
&\qquad \left. + 2(b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top \mathfrak{A}^{-2}\mathbb{1}_s) - 3(b^\top \mathfrak{A}^{-2}\mathbb{1}_s)^3 - b^\top \mathfrak{A}^{-4}\tilde{c} + (b^\top \mathfrak{A}^{-4}\mathbb{1}_s)(b^\top \mathfrak{A}^{-1}\tilde{c})\right.
\end{aligned}
$$

$$
\begin{aligned}
&+(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\tilde{c})+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\tilde{c})-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\tilde{c})\\
&-(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top\mathfrak{A}^{-2}\tilde{c})+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3(b^\top\mathfrak{A}^{-2}\tilde{c})\Big)\\
=\;&6\Big(-b^\top\mathfrak{A}^{-4}\mathbb{1}_s+2(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)-(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3-b^\top\mathfrak{A}^{-3}\mathbb{1}_s+b^\top\mathfrak{A}^{-4}\mathbb{1}_s\\
&+(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)\\
&-(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)\Big)\\
=\;&0,
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=0}^{3}i^2\left(\tau^4\tilde{R}(z)^i\tilde{J}(z)\mathbb{1}_s\right)'''(0)-\sum_{i=0}^{3}i\left(\tau^4\tilde{R}(z)^i\tilde{J}(z)\tilde{c}\right)'''(0)+\frac{1}{2}\sum_{i=0}^{3}\left(\tau^4\tilde{R}(z)^i\tilde{J}(z)\tilde{c}^2\right)'''(0)\\
=\;&\frac{\tau^4}{2}\Bigg(-\frac{1}{\tau}\tilde{R}'''(0)+3\tilde{R}''(0)\tilde{J}'(0)\mathbb{1}_s+3\tilde{R}'(0)\tilde{J}''(0)\mathbb{1}_s-\frac{24}{\tau}\tilde{R}'(0)\tilde{R}''(0)+24\tilde{R}'(0)^2\tilde{J}'(0)\mathbb{1}_s\\
&-\frac{54}{\tau}\tilde{R}'(0)^3-2\tilde{R}'''(0)\tilde{J}(0)\tilde{c}-6\tilde{R}''(0)\tilde{J}'(0)\tilde{c}-6\tilde{R}'(0)\tilde{J}''(0)\tilde{c}-24\tilde{R}'(0)\tilde{R}''(0)\tilde{J}(0)\tilde{c}\\
&-24\tilde{R}'(0)^2\tilde{J}'(0)\tilde{c}-36\tilde{R}'(0)^3\tilde{J}(0)\tilde{c}+\tilde{J}'''(0)\tilde{c}^2+\tilde{R}'''(0)\tilde{J}(0)\tilde{c}^2+3\tilde{R}''(0)\tilde{J}'(0)\tilde{c}^2\\
&+3\tilde{R}'(0)\tilde{J}''(0)\tilde{c}^2+6\tilde{R}'(0)\tilde{R}''(0)\tilde{J}(0)\tilde{c}^2+6\tilde{R}'(0)^2\tilde{J}'(0)\tilde{c}^2+6\tilde{R}'(0)^3\tilde{J}(0)\tilde{c}^2\Bigg)\\
=\;&3\Big(b^\top\mathfrak{A}^{-4}\mathbb{1}_s+(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)-8(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)\\
&-4(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)+9(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3-2(b^\top\mathfrak{A}^{-4}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\tilde{c})\\
&-2(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\tilde{c})-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\tilde{c})+8(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\tilde{c})\\
&+4(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top\mathfrak{A}^{-2}\tilde{c})-6(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3(b^\top\mathfrak{A}^{-1}\tilde{c})-b^\top\mathfrak{A}^{-4}\tilde{c}^2+(b^\top\mathfrak{A}^{-4}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\tilde{c}^2)\\
&+(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\tilde{c}^2)+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\tilde{c}^2)-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\tilde{c}^2)\\
&-(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top\mathfrak{A}^{-2}\tilde{c}^2)+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3(b^\top\mathfrak{A}^{-1}\tilde{c}^2)\Big)\\
=\;&3\Big(b^\top\mathfrak{A}^{-4}\mathbb{1}_s-6(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)-4(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3+9(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3-2(b^\top\mathfrak{A}^{-4}\mathbb{1}_s)\\
&-2(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)+8(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)\\
&+4(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)-6(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3-2b^\top\mathfrak{A}^{-2}\mathbb{1}_s+b^\top\mathfrak{A}^{-4}\mathbb{1}_s+2b^\top\mathfrak{A}^{-3}\mathbb{1}_s\\
&+2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-1}\mathbb{1}_s)-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)(b^\top\mathfrak{A}^{-3}\mathbb{1}_s)-2(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^2+(b^\top\mathfrak{A}^{-2}\mathbb{1}_s)^3\Big)\\
=\;&0,
\end{aligned}
$$

(5.67) ist für $q\geq 2$ erfüllt. □

Unter den Voraussetzungen des folgenden Lemmas ist für $\alpha_r\in\{-1,0\}$ die Voraussetzung (d) in Satz 5.28 erfüllt:

**Lemma 5.32** Das betrachtete Runge-Kutta-Verfahren sei $A$-stabil mit regulärer Verfahrensmatrix, und es gelte für die Stabilitätsfunktion $R(it)\neq 1$ für $t\in\mathbb{R}\setminus\{0\}$ sowie $\mathfrak{Re}\kappa_{\vec{k}j_1}\leq 0$ (folgt aus Voraussetzung (b) von Satz 5.28) und $\lim\limits_{z\to-\infty}R(z)\neq 1$.
Sind dann die Matrixnormen in (5.72) beschränkt, so existieren für $r=q+1,\ldots,p-1$ beschränkte Matrizen $W_{r0}(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})$.
Sind die Matrixnormen in (5.74)beschränkt (siehe Voraussetzung (a) von Satz 5.28), so existieren für $r=q+1,\ldots,p$ beschränkte Matrizen $W_{r(-1)}(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})$. □

**Beweis:** Sei $\alpha_r\in\{-1,0\}$. Aus der Definition 2.12 der Matrixfunktionen und der Formel für die Matrixfunktion eines Jordan-Blocks (2.25) folgt, daß die Matrizen $W_{r\alpha_r}(I_{n_{\vec{k}j_1}},\tau R_{\vec{k}j_1})$ existieren und beschränkt sind, falls die (gebrochen rationale) Funktion

$$
W_{r\alpha_r}(z)=W_{r\alpha_r}(1,z)=\frac{J(z)[\tilde{c}^r-r\mathfrak{A}\tilde{c}^{r-1}]}{1-R(z)}z^{-\alpha_r}\tag{5.76}
$$

für $z \in \mathbb{C}^-$ existiert und beschränkt ist.
Da $R(z)$ $A$-verträglich ist, ist $R(z)$ analytisch für $\mathfrak{Re}z < 0$, und es gilt $|R(z)| \leq 1$ für $\mathfrak{Re}z = 0$. Aus dem Maximumprinzip für den Betrag analytischer Funktionen folgt deshalb, daß $|R(z)| < 1$ und damit $R(z) \neq 1$ für $\mathfrak{Re}z < 0$ gilt. $R(it) \neq 1$ für $t \in R \setminus \{0\}$ gilt nach Voraussetzung. Damit ist $W_{r\alpha_r}(z)$ für $z \in \mathbb{C}^- \setminus \{0\}$ gemäß (5.76) definiert. An $z = 0$ gilt jedoch $R(0) = 1$, und damit verschwindet der Nenner. Aus den Runge-Kutta-Ordnungsbedingungen für die elementaren Differentiale $f^{(r-1)}f \ldots f$ und $f'f^{(r-2)}f \ldots f$ folgt $b^\top \tilde{c}^r = rb^\top \mathfrak{A}\tilde{c}^{r-1}$ und damit $J(0)[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}] = 0$ für $r = q+1, \ldots, p-1$. Ist $\alpha_r = 0$, so verschwindet deshalb für $r = q+1, \ldots, p-1$ auch der Zähler, ist $\alpha_r = -1$, so gilt dies sogar für $r = q+1, \ldots, p$. Nach der Bernoulli-l'Hospitalschen Regel gilt deshalb

$$\lim_{z\to 0} W_{r\alpha_r}(z) = \lim_{z\to 0} \frac{J'(z)[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}]z^{-\alpha_r} - \alpha_r J(z)[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}]z^{-\alpha_r-1}}{-R'(z)},$$

falls der Ausdruck auf der rechten Seite existiert. Wegen $R'(0) = 1$ kann deshalb die Lücke im Definitionsbereich von $W_{r\alpha_r}(z)$ durch Definition von

$$W_{r\alpha_r}(0) = \begin{cases} -J(0)[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}]: & \alpha_r = -1,\ r = q+1, \ldots, p \\ -J'(0)[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}]: & \alpha_r = 0,\ r = q+1, \ldots, p-1 \end{cases}$$

glatt geschlossen werden.
Ist $(I_s - z\mathfrak{A})_{Adj}$ die adjungierte Matrix von $(I_s - z\mathfrak{A})$, so gilt

$$J(z) = \frac{1}{\det(I_s - z\mathfrak{A})} b^\top (I_s - z\mathfrak{A})_{Adj},$$

und mit der Stetterschen Darstellung (2.21) von $R(z)$ folgt aus (5.76)

$$W_{r\alpha_r}(z) = \frac{b^\top (I_s - z\mathfrak{A})_{Adj}[\tilde{c}^r - r\mathfrak{A}\tilde{c}^{r-1}]z^{-\alpha_r}}{\det(I_s - z\mathfrak{A}) - \det(I_s - z\mathfrak{A} + z\mathbb{1}_s b^\top)}. \tag{5.77}$$

Da $\mathfrak{A}$ regulär ist und $\lim_{z\to-\infty} R(z) \neq 1$ gilt, hat das Nennerpolynom den Grad $s$. Das Zählerpolynom hat höchstens den Grad $s$. Damit ist $W_{r\alpha_r}(z)$ für $z \in \mathbb{C}^-$ beschränkt, und es folgt insgesamt die Behauptung. □

Zum Beispiel für die Radau-IA-, Radau-IIA- und Lobatto-IIIC-Verfahren ($L$-stabil mit regulärer Verfahrensmatrix) sowie das einstufige Gauß-Verfahren sind die Voraussetzungen von Lemma 5.32 an das Verfahren erfüllt:

**Lemma 5.33** Für die Stabilitätsfunktionen der Radau-IA-, Radau-IIA- und Lobatto-IIIC-Verfahren sowie des einstufigen Gauß-Verfahrens gilt $R(it) \neq 1$ für $t \in \mathbb{R} \setminus \{0\}$. □

**Beweis:** Die Stabilitätsfunktion des einstufigen Gauß-Verfahrens ist die Padé-Approximation vom Index $(1,1)$

$$R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}},$$

und die Behauptung folgt für dieses Verfahren durch direktes Ausrechnen.
Die Stabilitätsfunktionen $R(z) = \frac{P_{jk}(z)}{Q_{jk}(z)}$ ($j = \mathrm{grad}(P_{jk})$, $k = \mathrm{grad}(Q_{jk})$) der $s$-stufigen Radau-IA- und -IIA- bzw. Lobatto-IIIC-Verfahren sind die Padé-Approximationen vom Index $(j,k) = (s, s-1)$ bzw. $(s, s-2)$, und es gilt für die Konsistenzordnung $p = j + k$.
Aus

$$|e^z| - \frac{|P_{jk}(z)|}{|Q_{jk}(z)|} = \mathcal{O}(z^{p+1}) \quad \text{für} \quad z \to 0$$

folgt für $z = it$ wegen $|e^{it}| = 1$ für $t \in \mathbb{R}$

$$|Q_{jk}(iy)| - |P_{jk}(iy)| = \mathcal{O}(y^{p+1}) \quad \text{für} \quad y \to 0$$

und damit für das sogenannte E-Polynom

$$\begin{aligned} E_{jk}(y) &= (|Q_{jk}(iy)| - |P_{jk}(iy)|)(|Q_{jk}(iy)| + |P_{jk}(iy)|) \\ &= |Q_{jk}(iy)|^2 - |P_{jk}(iy)|^2 = \mathcal{O}(y^{j+k+1}) \quad \text{für} \quad y \to 0. \end{aligned}$$

Da das E-Polynom in $y$ gerade ist und für die betrachteten Verfahren höchstens den Grad $2s$ hat, folgt $E_{sk}(y) = c_k y^{2s}$ für $k = s-1$ und $k = s-2$. Aus $\text{grad}(Q_{sk}) > \text{grad}(P_{sk})$ folgt $c_k > 0$ und damit schließlich $E_{sk}(y) > 0$ für $y \neq 0$, was äquivalent ist zu $|R(it)| < 1$ für $t \in \mathbb{R} \setminus \{0\}$. □

### 5.2.4 Konvergenzuntersuchungen für semilineare PDA-Systeme

In diesem Abschnitt sollen die Konvergenzuntersuchungen für lineare Systeme aus Abschnitt 5.2.2 auf Systeme mit nichtlinearer rechter Seite erweitert werden,

$$f = f(t, \vec{x}, u).$$

Dazu wird zunächst eine Abschätzung für den globalen Diskretisierungsfehler der Stufen des Runge-Kutta-Verfahrens (globaler Stufenfehler) im linearen Fall angegeben, die für die Konvergenzuntersuchung im nichtlinearen Fall benötigt wird.

**Definition 5.34** Mit

$$E_{m+1}^{(i)} = U_{\vec{h}}(t_m + c_i\tau) - U_{m+1}^{(i)}$$

wird der globale Stufenfehler der $i$-ten Stufe des Runge-Kutta-Verfahrens bezeichnet. □

Mit

$$E_{m+1} = \left( {E_{m+1}^{(1)}}^\top, \ldots, {E_{m+1}^{(s)}}^\top \right)^\top$$

und der Definitionsgleichung des globalen Gesamtdiskretisierungsfehlers (5.15) sowie den Störungen (5.17) folgt aus dem Störungslemma 5.3

$$\begin{aligned} E_{m+1} = {} & H(M, \tau D) e_m + G(M, \tau D)^{-1} (I_s \otimes M) \Delta_{m+1} \\ & + \tau G(M, \tau D)^{-1} (\mathfrak{A} \otimes I_{Nn}) \sum_{v=1}^{d} h_v^{p_v} (I_{sN} \otimes B_v) \Gamma_{m+1}^{(v)}. \end{aligned}$$

Daraus erhält man mit der Gleichung (5.23) des Verbundvektors der Residuenfehler der Stufen und der Gleichung für den globalen Gesamtdiskretisierungsfehler (5.28) die folgende Darstellung für den globalen Stufenfehler:

$$\begin{aligned} E_{m+1} = {} & G(M, \tau D)^{-1} (I_s \otimes M) \left( \sum_{r=q+1}^{p} \frac{\tau^r}{r!} \left[ \tilde{c}^r - r \mathfrak{A} \tilde{c}^{r-1} \right] \otimes U_{\vec{h}}^{(r)}(t_m) + \mathcal{O}(\tau^{p+1}) \right) \\ & + \tau \sum_{v=1}^{d} h_v^{p_v} \sum_{i=0}^{m} H(M, \tau D) R(M, \tau D)^i J(M, \tau D) \left( I_{sN} \otimes B_v \right) \Gamma_{m+1-i}^{(v)} \\ & + \sum_{i=0}^{m} H(M, \tau D) R(M, \tau D)^i \delta_{m+1-i} + \sum_{i=0}^{m} H(M, \tau D) R(M, \tau D)^i L(M, \tau D) \mathcal{O}(\tau^{p+1}) \\ & + \sum_{r=q+1}^{p} \frac{\tau^{r+1+\alpha_r}}{r!} \Bigg( - \sum_{i=0}^{m-1} \int_{t_i}^{t_{i+1}} H(M, \tau D) R(M, \tau D)^{m-i} W_{r\alpha_r}(M, \tau D) D^{1+\alpha_r} U_{\vec{h}}^{(r+1)}(s) ds \end{aligned}$$

$$
\begin{aligned}
&+H(M,\tau D)W_{r\alpha_r}(M,\tau D)D^{1+\alpha_r}U_{\vec{h}}^{(r)}(t_m)\\
&-H(M,\tau D)R(M,\tau D)^{m+1}W_{r\alpha_r}(M,\tau D)D^{1+\alpha_r}U_{\vec{h}}^{(r)}(t_0)\Bigg)\\
&+\tau G(M,\tau D)^{-1}(\mathfrak{A}\otimes I_{Nn})\sum_{v=1}^{d}h_v^{p_v}(I_{sN}\otimes B_v)\Gamma_{m+1}^{(v)}.
\end{aligned}
$$

Aus (4.58) folgen mit (5.34) und $H(A,\tau D_{\vec{k}}) = G(A,\tau D_{\vec{k}})^{-1}(\mathbb{1}_s\otimes A)$

$$
(I_s\otimes Q^{-1})\left(G(M,\tau D)^{-1}(I_s\otimes M)\right)(I_s\otimes Q) = \left(\mathrm{diag}_{\vec{k}}\left\{\left(G(A,\tau D_{\vec{k}})^{-1}(I_s\otimes A)\right)_{ij}\right\}\right)_{i,j=1,\ldots,s},
$$

und

$$
(I_s\otimes Q^{-1})H(M,\tau D)Q = \left(\mathrm{diag}_{\vec{k}}\left\{\left(H(A,\tau D_{\vec{k}})\right)_i\right\}\right)_{i=1,\ldots,s}. \tag{5.78}
$$

Mit $\alpha_r = -1$ ergibt sich damit analog zu Satz 5.11 das folgende Lemma:

**Lemma 5.35** Sei $q^\star \in [1,\min(p,q+1)]$ die größte ganze Zahl, so daß mit $\bar{M}\in\mathbb{N}$ für $\tau\to 0$ $(M_e\to\infty)$, $h_v\to 0$ $(N_v\to\infty)$, $v=1,\ldots,d$ und alle $\vec{k}$ mit $k_v = 1,\ldots,N_v$ die folgenden Voraussetzungen erfüllt sind:

(a) $G(A,\tau D_{\vec{k}})$ ist regulär,

(b) $W_{r(-1)}(A,\tau D_{\vec{k}})$ existiert gemäß (5.27), $r = q+1,\ldots,p$,

(c) $\|G(A,\tau D_{\vec{k}})^{-1}(I_s\otimes A)\tau^{q+1-q^\star}\|$ und $\|\tau G(A,\tau D_{\vec{k}})^{-1}(\mathfrak{A}\otimes B_v)\|$ sind beschränkt,

(d) für $i=0,\ldots,\bar{M}$ sind $\|\tau H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i J(A,\tau D_{\vec{k}})\,(I_s\otimes B_v)\,\|$, $\|H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i\tau^{p+1-q^\star}\|$ und $\|H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i L(A,\tau D_{\vec{k}})\tau^{p+1-q^\star}\|$ beschränkt,

(e) für $i=\bar{M}+1,\ldots,M_e-1$ sind $\|H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i J(A,\tau D_{\vec{k}})\,(I_s\otimes B_v)\,\|$, $\|H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i\tau^{p-q^\star}\|$ und $\|H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i L(A,\tau D_{\vec{k}})\tau^{p-q^\star}\|$ beschränkt,

(f) $\|\tau^{r-q^\star}H(A,\tau D_{\vec{k}})R(A,\tau D_{\vec{k}})^i W_{r(-1)}(A,\tau D_{\vec{k}})\|$ ist für $i=0,\ldots,M_e$ und $r=q+1,\ldots,p$ beschränkt.

Dann gilt für lineare Systeme für den Vektor der Diskretisierungsfehler der Stufen des Runge-Kutta-Verfahrens bei genügend glatter exakter Lösung

$$
E_{m+1} = \mathcal{O}\left(\tau^{q^\star}\right) + \mathcal{O}\left(\sum_{v=1}^{d}h_v^{p_v}\right).
$$

□

Im folgenden wird zunächst vorausgesetzt, daß $f$ komponentenweise eine globale Lipschitz-Bedingung erfüllt: Für $i=1,\ldots,n$ existiere $l_{f_i}$ mit

$$
|f_i(t,\vec{x},u_1) - f_i(t,\vec{x},u_2)| \le l_{f_i}\|u_1-u_2\|_2 \tag{5.79}
$$

für alle $u_1,u_2\in\mathbb{R}^n$, $t\in[t_0,t_e]$, $\vec{x}\in\Omega$.
Aus den Gleichungen (5.3b) für die Stufenwerte $S_{m+1}$ und (5.3c) für die Steigungswerte $K_{m+1}$ folgt mit der in (5.7) definierten Matrix $G(M,\tau D)$

$$
G(M,\tau D)K_{m+1} = (\mathbb{1}_s\otimes D)U_m + \bar{F}(t_{m+1},S_{m+1}) \tag{5.80}
$$

und damit aus der Verfahrensgleichung (5.3a) und den in (5.8) und (5.9) definierten Matrizen $J(M,\tau D)$ und $R(M,\tau D)$

$$\begin{aligned} U_{m+1} &= U_m + \tau(b^\top \otimes I_{Nn})G(M,\tau D)^{-1}\left((\mathbb{1}_s \otimes D)U_m + \bar{F}(t_{m+1},S_{m+1})\right) \\ &= R(M,\tau D)U_m + \tau J(M,\tau D)\bar{F}(t_{m+1},S_{m+1}). \end{aligned}$$

Ausführen der Rekursion liefert

$$U_{m+1} = R(M,\tau D)^{m+1}U_0 + \tau\sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D)\bar{F}(t_{m+1-i},S_{m+1-i}). \tag{5.81}$$

Durch Einsetzen von (5.80) in (5.3b) folgt mit (5.7)

$$\begin{aligned} S_{m+1} &= \mathbb{1}_s \otimes U_m + \tau\left(\mathfrak{A}\otimes I_{Nn}\right)G(M,\tau D)^{-1}\left((\mathbb{1}_s\otimes D)U_m + \bar{F}(t_{m+1},S_{m+1})\right) \\ &= G(M,\tau D)^{-1}\left(G(M,\tau D) + \tau\mathfrak{A}\otimes D\right)\left(\mathbb{1}_s\otimes I_{Nn}\right)U_m \\ &\quad +\tau G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\bar{F}(t_{m+1},S_{m+1}) \\ &= G(M,\tau D)^{-1}\left(\mathbb{1}_s\otimes M\right)U_m + \tau G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\bar{F}(t_{m+1},S_{m+1}) \end{aligned} \tag{5.82}$$

und mit der in (5.11) definierten Matrix $H(M,\tau D)$

$$S_{m+1} = H(M,\tau D)U_m + \tau G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\bar{F}(t_{m+1},S_{m+1}).$$

Durch Einsetzen von (5.81) erhält man

$$\begin{aligned} S_{m+1} &= H(M,\tau D)R(M,\tau D)^m U_0 + \tau G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\bar{F}(t_{m+1},S_{m+1}) \\ &\quad +\tau\sum_{i=0}^{m-1} H(M,\tau D)R(M,\tau D)^i J(M,\tau D)\bar{F}(t_{m-i},S_{m-i}). \end{aligned} \tag{5.83}$$

Das System (5.3) mit vorgegebenem Startvektor $U_0$ hat genau dann eine eindeutige Lösung, wenn (5.83) eindeutig lösbar ist. Es soll deshalb nun unter Verwendung des Banachschen Fixpunktsatzes eine hinreichende Bedingung dafür hergeleitet werden.
Aus der Lipschitz-Bedingung (5.79) folgt für alle Vektoren $Z_1,\ Z_2 \in \mathbb{R}^{sNn}$

$$\|\left(I_{sN}\otimes\vec{e}_i^{\top}\right)\left(\bar{F}(t_{m+1},Z_1) - \bar{F}(t_{m+1},Z_2)\right)\| \le l_{f_i}\|Z_1 - Z_2\|, \tag{5.84}$$

wobei $\vec{e}_i \in \mathbb{R}^n$ der $i$-te Einheitsvektor sei. Für die Abbildung

$$\begin{aligned} T:\ Z\in\mathbb{R}^{sNn} \to\ & H(M,\tau D)R(M,\tau D)^m U_0 + \tau G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\bar{F}(t_{m+1},Z) \\ & +\tau\sum_{i=0}^{m-1} H(M,\tau D)R(M,\tau D)^i J(M,\tau D)\bar{F}(t_{m-i},S_{m-i}) \end{aligned} \tag{5.85}$$

gilt deshalb

$$\begin{aligned} \|TZ_1 - TZ_2\| &= \tau\|G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\left(\bar{F}(t_{m+1},Z_1) - \bar{F}(t_{m+1},Z_2)\right)\| \\ &= \tau\|\sum_{i=1}^{n} G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes I_{Nn}\right)\left(I_{sN}\otimes\vec{e}_i\right) \\ &\qquad\qquad \cdot\left(I_{sN}\otimes\vec{e}_i^{\top}\right)\left(\bar{F}(t_{m+1},Z_1) - \bar{F}(t_{m+1},Z_2)\right)\| \\ &\le \tau\sum_{i=1}^{n} l_{f_i}\|G(M,\tau D)^{-1}\left(\mathfrak{A}\otimes\vec{e}_i\right)\|\ \|Z_1 - Z_2\|. \end{aligned}$$

$T$ bildet den Banachraum $\left(\mathbb{R}^{sNn},\|\cdot\|\right)$ in sich ab. Ist $T$ kontraktiv, so folgt deshalb aus dem Banachschen Fixpunktsatz die Existenz einer eindeutigen Lösung von (5.83) bzw. (5.3). Mit (5.34), (4.58) und der für die Normen geltenden Beziehung (4.66) erhält man deshalb den folgenden Existenzsatz:

**Satz 5.36** Sind mit einem $\kappa_1 < 1$ für $\tau \to 0$ ($M_e \to \infty$), $h_v \to 0$ ($N_v \to \infty$), $v = 1, \ldots, d$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die Matrizen $G(A, \tau D_{\vec{k}})$ regulär und gilt

$$\tau \prod_{i=1}^{d} \|S_{P_i}\| \, \|S_{P_i}^{-1}\| \sum_{j=1}^{n} l_{f_j} \max_{\vec{k}} \|G(A, \tau D_{\vec{k}})^{-1} \, (\mathfrak{A} \otimes \vec{e}_j) \, \| \leq \kappa_1, \tag{5.86}$$

so besitzt das Verfahren (5.3) für hinreichend kleine $\tau$ und $\vec{h}$ eine eindeutige Lösung. □

Die exakte Lösung $u(t, \vec{x})$ des Ausgangssystems (3.2) erfüllt das lineare PDA-System

$$A \, v_t(t, \vec{x}) + \sum_{i=1}^{d} B_i \, (v_{x_i x_i}(t, \vec{x}) + r_i v_{x_i}(t, \vec{x})) + C \, v(t, \vec{x}) = \Phi(t, \vec{x})$$

mit $\Phi(t, \vec{x}) = f(t, \vec{x}, u(t, \vec{x}))$. Das zugehörige lineare MOL-DA-System ist analog zu (4.45) durch

$$M\dot{V}(t) = DV(t) + \tilde{F}(t, U_{\vec{h}}(t)), \tag{5.87a}$$

$$V(t_0) = U(t_0) \tag{5.87b}$$

gegeben. Dabei ist $U_{\vec{h}}(t)$ die Einschränkung von $u(t, \vec{x})$ auf das Ortsgitter $\Omega_{\vec{h}}$. Für die numerische Lösung $V_m$ dieses Systems mit dem Verfahren (5.3) und die Runge-Kutta-Stufen $S_m^V$ gelten nach (5.81) und (5.83)

$$V_{m+1} = R(M, \tau D)^{m+1} U_0 + \tau \sum_{i=0}^{m} R(M, \tau D)^i J(M, \tau D) \bar{F}(t_{m+1-i}, S_{m+1-i}^{U_{\vec{h}}}), \tag{5.88}$$

$$\begin{aligned} S_{m+1}^V = {} & H(M, \tau D) R(M, \tau D)^m U_0 + \tau G(M, \tau D)^{-1} \, (\mathfrak{A} \otimes I_{Nn}) \, \bar{F}(t_{m+1}, S_{m+1}^{U_{\vec{h}}}) \\ & + \tau \sum_{i=0}^{m-1} H(M, \tau D) R(M, \tau D)^i J(M, \tau D) \bar{F}(t_{m-i}, S_{m-i}^{U_{\vec{h}}}) \end{aligned} \tag{5.89}$$

mit

$$S_{m+1}^{U_{\vec{h}}} = \left( U_{\vec{h}}(t_m + c_1 \tau)^\top, \ldots, U_{\vec{h}}(t_m + c_s \tau)^\top \right)^\top.$$

Im folgenden bezeichne der Vektor

$$E_{m+1}^V = S_{m+1}^{U_{\vec{h}}} - S_{m+1}^V \tag{5.90}$$

den Diskretisierungsfehler der Stufen bezüglich der exakten Stufenwerte $S_{m+1}^{U_{\vec{h}}}$ des semilinearen PDA-Systems und der Runge-Kutta-Stufen $S_{m+1}^V$ des linearen DA-Systems (5.87). Ferner sei

$$e_{m+1}^V = U_{\vec{h}}(t_{m+1}) - V_{m+1}. \tag{5.91}$$

Durch Einsetzen von (5.89) in (5.90) folgt

$$\begin{aligned} S_{m+1}^{U_{\vec{h}}} = {} & H(M, \tau D) R(M, \tau D)^m U_0 + \tau G(M, \tau D)^{-1} \, (\mathfrak{A} \otimes I_{Nn}) \, \bar{F}(t_{m+1}, S_{m+1}^{U_{\vec{h}}}) \\ & + \tau \sum_{i=0}^{m-1} H(M, \tau D) R(M, \tau D)^i J(M, \tau D) \bar{F}(t_{m-i}, S_{m-i}^{U_{\vec{h}}}) + E_{m+1}^V \end{aligned}$$

und durch Subtraktion von (5.83)

$$\begin{aligned} & S_{m+1}^{U_{\vec{h}}} - S_{m+1} \\ = {} & \tau G(M, \tau D)^{-1} \, (\mathfrak{A} \otimes I_{Nn}) \left( \bar{F}(t_{m+1}, S_{m+1}^{U_{\vec{h}}}) - \bar{F}(t_{m+1}, S_{m+1}) \right) \\ & + \tau \sum_{i=0}^{m-1} H(M, \tau D) R(M, \tau D)^i J(M, \tau D) \left( \bar{F}(t_{m-i}, S_{m-i}^{U_{\vec{h}}}) - \bar{F}(t_{m-i}, S_{m-i}) \right) \\ & + E_{m+1}^V. \end{aligned}$$

Mit der Lipschitz-Bedingung (5.84) ergibt sich

$$\begin{aligned}\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}\| \le & \sum_{j=0}^{n} \tau l_{f_j} \|G(M,\tau D)^{-1} \left(\mathfrak{A} \otimes \vec{e}_j\right)\| \|S_{m+1}^{U_{\vec{h}}} - S_{m+1}\| \\ & + \sum_{j=0}^{n} l_{f_j} \max_{i=0}^{m-1} \|H(M,\tau D) R(M,\tau D)^i J(M,\tau D) \left(I_{sN} \otimes \vec{e}_j\right)\| \\ & \cdot \tau \sum_{i=0}^{m-1} \|S_{m-i}^{U_{\vec{h}}} - S_{m-i}\| + \|E_{m+1}^V\|.\end{aligned}$$

Daraus folgt: Ist mit einem $M_0 \in \mathbb{N}$

$$\kappa_2 = \sup_{M_e, N_v \ge M_0, v=1,\ldots,d} \left\{ \sum_{j=0}^{n} l_{f_j} \|H(M,\tau D) R(M,\tau D)^{i_1} J(M,\tau D) \left(I_{sN} \otimes \vec{e}_j\right)\| : i_1 = 0, \ldots, M_e - 2, \tau = \frac{t_e - t_0}{M_e} \right\} < \infty, \tag{5.92}$$

so erhält man unter der Voraussetzung (5.86)

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}\| \le \frac{\kappa_2}{1-\kappa_1} \tau \sum_{i=0}^{m-1} \|S_{i+1}^{U_{\vec{h}}} - S_{i+1}\| + \frac{\|E_{m+1}^V\|}{1-\kappa_1}. \tag{5.93}$$

Nun wird das folgende diskrete Gronwall-Lemma angewendet (vgl. Werner/Arndt [58]):

**Lemma 5.37** Seien $\sigma_j$ und $\xi_j$, $j = 0, 1, \ldots, m$, nichtnegative Zahlen mit $\xi_0 \le \xi_1 \le \ldots \le \xi_m$. Mit $\delta \ge 0$, $(\tau_0, \tau_1, \ldots, \tau_{m-1}) \in \mathbb{R}_+^m$, $t_{j+1} = t_j + \tau_j$ gelte die Abschätzung

$$\sigma_0 \le \xi_0 \quad \text{und} \quad \sigma_{j+1} \le \delta \sum_{i=0}^{j} \tau_i \sigma_i + \xi_{j+1}, \quad j = 0, 1, \ldots, m-1.$$

Dann gilt auch

$$\sigma_j \le \xi_j e^{\delta(t_j - t_0)}, \quad j = 0, 1, \ldots, m. \qquad \square$$

Damit folgt aus (5.93)

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}\| \le \|E_{m+1}^V\| \frac{1}{1-\kappa_1} e^{\frac{\kappa_2}{1-\kappa_1}\tau(m+1)} \le \|E_{m+1}^V\| \frac{e^{\frac{\kappa_2}{1-\kappa_1}(t_e - t_0)}}{1-\kappa_1}.$$

Mit Lemma 5.35 für den globalen Stufenfehler erhält man

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}\| = \mathcal{O}\left(\tau^{q^\star}\right) + \mathcal{O}\left(\sum_{v=1}^{d} h_v^{p_v}\right). \tag{5.94}$$

Durch Einsetzen von (5.88) in (5.91) folgt

$$U_{\vec{h}}(t_{m+1}) = R(M,\tau D)^{m+1} U_0 + \tau \sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D) \bar{F}(t_{m+1-i}, S_{m+1-i}^{U_{\vec{h}}}) + e_{m+1}^V$$

und durch Subtraktion von (5.81)

$$\begin{aligned}U_{\vec{h}}(t_{m+1}) - U_{m+1} = & \tau \sum_{i=0}^{m} R(M,\tau D)^i J(M,\tau D) \left(\bar{F}(t_{m+1-i}, S_{m+1-i}^{U_{\vec{h}}}) - \bar{F}(t_{m+1-i}, S_{m+1-i})\right) \\ & + e_{m+1}^V.\end{aligned}$$

Mit der Lipschitz-Bedingung (5.84) ergibt sich daraus

$$\|U_{\vec{h}}(t_{m+1}) - U_{m+1}\| \leq \sum_{j=1}^{n} \tau l_{f_j} \sum_{i=0}^{m} \|R(M,\tau D)^i J(M,\tau D)\,(I_{sN} \otimes \vec{e}_j)\,\| \; \|S_{m+1-i}^{U_{\vec{h}}} - S_{m+1-i}\| + \|e_{m+1}^V\|. \tag{5.95}$$

Insgesamt folgt mit (5.94) und durch Diagonalisierung der in (5.92) und (5.95) auftretenden Matrizen der folgende Konvergenzsatz:

**Satz 5.38** Sei $f$ komponentenweise global Lipschitz-stetig mit den Konstanten $l_{f_j}$, $j = 1, \ldots, n$. Gilt dann zusätzlich zu den Voraussetzungen von Konvergenzsatz 5.11, Lemma 5.35 (Konvergenz der Stufenwerte) und Satz 5.36 (Existenz der Runge-Kutta-Lösung) für $j = 1, \ldots, n$, daß

$$l_{f_j}\|\tau R(A,\tau D_{\vec{k}})^i J(A,\tau D_{\vec{k}})\,(I_s \otimes \vec{e}_j)\,\| \quad \text{für} \quad i = 0, \ldots, \bar{M},$$

$$l_{f_j}\|R(A,\tau D_{\vec{k}})^i J(A,\tau D_{\vec{k}})\,(I_s \otimes \vec{e}_j)\,\| \quad \text{für} \quad i = \bar{M}+1, \ldots, M_e - 1$$

sowie

$$l_{f_j}\|H(A,\tau D_{\vec{k}}) R(A,\tau D_{\vec{k}})^i J(A,\tau D_{\vec{k}})\,(I_s \otimes \vec{e}_j)\,\| \quad \text{für} \quad i = 0, \ldots, M_e - 1$$

sämtlich beschränkt sind, so ist das Diskretisierungsverfahren (5.3) konvergent mit der Ordnung $(p_1, \ldots, p_d, \min\{p^\star, q^\star\})$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

Satz 5.38 ist nur für global Lipschitz-stetige Funktionen $f$ anwendbar. Diese Voraussetzung ist jedoch zum Beispiel für Beispiel 3.2 (Pharmakokinetik in der Leber) und Beispiel 3.3 (Pulververbrennung) nicht erfüllt. Ist $f$ nur lokal Lipschitz-stetig in einer Umgebung der exakten Lösung $u$ mit Lipschitz-Konstanten $l_{f_j}(\gamma)$,

$$|f_j(t,\vec{x},y^1) - f_j(t,\vec{x},y^2)| \leq l_{f_j}(\gamma)\|y^1 - y^2\|_2 \tag{5.96}$$

für $y^1, y^2 \in \mathbb{R}^n$ mit $\|y^1 - u\|_\infty \leq \gamma$, $\|y^2 - u\|_\infty \leq \gamma$, $t \in [t_0, t_e]$, $\vec{x} \in \Omega$, $j = 1, \ldots, n$ und $\gamma \in \mathbb{R}$, $\gamma > 0$, dann kann die bedingte Konvergenz des Runge-Kutta-Verfahrens wie folgt gezeigt werden:
Es sei

$$f^\gamma(t,\vec{x},y) = f(t,\vec{x},v(\gamma,y)) \quad \text{mit} \quad v : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$$

und

$$v_j(\gamma,y) = \begin{cases} y_j & : \; |y_j - u_j| \leq \gamma \\ u_j + \gamma & : \; y_j > u_j + \gamma \\ u_j - \gamma & : \; y_j < u_j - \gamma \end{cases}.$$

Dann ist $f^\gamma$ komponentenweise (global) Lipschitz-stetig mit Lipschitz-Konstanten $l_{f_j}(\gamma)$.
Es wird nun das PDA-System betrachtet, das entsteht, wenn $f$ durch $f^\gamma$ ersetzt wird. Aufgrund der Definition von $f^\gamma$ gilt $f^\gamma(t,\vec{x},u) = f(t,\vec{x},u)$, $u$ ist damit auch exakte Lösung des geänderten PDA-Systems. Für die Stufenwerte $S_{m+1}^\gamma$ des zugehörigen Diskretisierungsverfahrens gilt unter den Voraussetzungen des Satzes 5.38

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}^\gamma\| = \mathcal{O}\left(\tau^{q^\star}\right) + \mathcal{O}\left(\sum_{v=1}^{d} h_v^{p_v}\right).$$

Gehen $\tau,\ h_i$, $i = 1, \ldots, d$, so gegen Null, daß

$$\tau^{2q^\star} \Big/ \prod_{i=1}^{d} h_i \to 0 \quad \text{und} \quad h_j^{2p_j} \Big/ \prod_{i=1}^{d} h_i \to 0 \quad \text{für} \quad j = 1, \ldots, d$$

gelten, so folgt wegen

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}^{\gamma}\|_\infty \sqrt{\prod_{i=1}^{d} h_i} \le \|S_{m+1}^{U_{\vec{h}}} - S_{m+1}^{\gamma}\|$$

deshalb auch

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}^{\gamma}\|_\infty \to 0$$

und damit für hinreichend kleine $\tau$ und $h_i$

$$\|S_{m+1}^{U_{\vec{h}}} - S_{m+1}^{\gamma}\|_\infty \le \gamma.$$

$S_{m+1}^{\gamma}$ erfüllt dann aufgrund der Definition von $f^\gamma$ Gleichung (5.83), wegen (5.80) und der Verfahrensgleichung (5.3a) liefert die Diskretisierung mit $f^\gamma$ also die gleichen Näherungswerte wie diejenige mit $f$. Daraus folgt:

**Satz 5.39** Sei $f$ gemäß (5.96) komponentenweise lokal Lipschitz-stetig in einer Umgebung der exakten Lösung mit Lipschitz-Konstanten $l_{f_j}(\gamma)$, $j = 1, \ldots, n$. Mit $l_{f_j}(\gamma)$ seien die Voraussetzungen des Satzes 5.38 erfüllt. Gehen $\tau,\ h_i$, $i = 1, \ldots, d$, so gegen Null, daß

$$\frac{\tau^{2q^\star}}{\prod_{i=1}^{d} h_i} \to 0, \quad \frac{h_j^{2p_j}}{\prod_{i=1}^{d} h_i} \to 0, \quad j = 1, \ldots, d, \tag{5.97}$$

so ist das Diskretisierungsverfahren (5.3) bedingt konvergent mit der Ordnung $(p_1, \ldots, p_d, \min\{p^\star, q^\star\})$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

**Bemerkung 5.40** Bedingung (5.97) ist im Fall $p_i = 2$, $i = 1, \ldots, d$, nur für $d \le 3$ erfüllbar. Falls ein $p_i = 1$ ist, so muß $d \le 2$ gelten. □

### 5.2.5 Konvergenz bei Anwendung steifgenauer Runge-Kutta-Verfahren mit singulärer Verfahrensmatrix

Ist die Verfahrensmatrix $\mathfrak{A}$ des Runge-Kutta-Verfahrens singulär, so ist das Runge-Kutta-System (5.1) schon im linearen Fall bei singulärer Matrix $A$ nicht eindeutig lösbar (vgl. Bemerkung 5.2). In diesem Abschnitt soll der Spezialfall betrachtet werden, daß die Verfahrensmatrix $\mathfrak{A}$ des Runge-Kutta-Verfahrens die Gestalt

$$\mathfrak{A} = \begin{pmatrix} 0 & \mathfrak{o} \\ \breve{\mathfrak{a}} & \breve{\mathfrak{A}} \end{pmatrix} \in \mathbb{R}^{s,s} \quad \text{mit} \quad \breve{\mathfrak{a}} \in \mathbb{R}^{s-1},\ \breve{\mathfrak{A}} \in \mathbb{R}^{s-1,s-1}$$

mit regulärer Matrix $\breve{\mathfrak{A}}$ hat, das Runge-Kutta-Verfahren steifgenau ist, also (2.20) erfüllt, und eine Stufenordnung $q \ge 1$ hat. Beispiele für solche Methoden sind die Lobatto-IIIA-Verfahren. Im allgemeinen existieren bei singulärem $A$ keine eindeutigen Steigungswerte $K_{m+1}^{(i)}$, die Stufenwerte $U_{m+1}^{(i)}$ und die Lösung $U_m$ können jedoch trotzdem eindeutig bestimmbar sein.
Da die Elemente der ersten Zeile der Verfahrensmatrix verschwinden ($a_{1i} = 0$, $i = 1, \ldots, s$), gilt

$$U_{m+1}^{(1)} = U_m. \tag{5.98a}$$

Aus der Steifgenauigkeit folgt

$$U_{m+1}^{(s)} = U_{m+1}. \tag{5.98b}$$

Zunächst soll wieder ein gestörtes Diskretisierungsverfahren (5.4) betrachtet werden. Die Beziehungen (5.98) sollen auch für das gestörte Diskretisierungsverfahren gelten, das heißt, für den Störungsvektor

$$\Theta_{m+1} = \left( {\Theta_{m+1}^{(1)}}^\top, \ldots, {\Theta_{m+1}^{(s)}}^\top \right)^\top$$

zum Zeitpunkt $t_{m+1}$ werden

$$\Theta_{m+1}^{(1)} = 0 \tag{5.99a}$$

und

$$\Theta_{m+1}^{(s)} = \theta_{m+1} \tag{5.99b}$$

gesetzt.
Definiert man für $\bar{N} \in \mathbb{N}$ und Matrizen $O, K \in \mathbb{R}^{\bar{N},\bar{N}}$

$$\breve{G}(O,K) = I_{s-1} \otimes O - \tau \breve{\mathfrak{A}} \otimes K \tag{5.100}$$

und

$$\breve{\Theta}_{m+1} = \left( {\Theta_{m+1}^{(2)}}^\top, \ldots, {\Theta_{m+1}^{(s)}}^\top \right)^\top,$$
$$\breve{S}_{m+1} = \left( {U_{m+1}^{(2)}}^\top, \ldots, {U_{m+1}^{(s)}}^\top \right)^\top$$

sowie

$$\breve{\hat{S}}_{m+1} = \left( {\hat{U}_{m+1}^{(2)}}^\top, \ldots, {\hat{U}_{m+1}^{(s)}}^\top \right)^\top,$$

wobei $\hat{U}_{m+1}^{(1)}, \ldots, \hat{U}_{m+1}^{(s)}$ die gestörten Stufenwerte seien, so ergibt sich aus (5.13) unter Berücksichtigung von (5.98a) und der aus (5.99a) und (5.4b) folgenden Beziehung $\hat{U}_{m+1}^{(1)} = \hat{U}_m$

$$\begin{aligned}
\breve{G}(M,\tau D)\left(\breve{\hat{S}}_{m+1} - \breve{S}_{m+1}\right) &= \tau(\breve{\mathfrak{a}} \otimes D)\left(\hat{U}_{m+1}^{(1)} - U_{m+1}^{(1)}\right) + \tau\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\rho_{m+1} \\
&\quad + (I_{s-1} \otimes M)\left(\mathbb{1}_{s-1} \otimes \left(\hat{U}_m - U_m\right) + \breve{\Theta}_{m+1}\right) \\
&= (\mathbb{1}_{s-1} \otimes M + \tau\breve{\mathfrak{a}} \otimes D)\left(\hat{U}_m - U_m\right) + (I_{s-1} \otimes M)\breve{\Theta}_{m+1} \\
&\quad + \tau\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\rho_{m+1}.
\end{aligned}$$

Unter der Voraussetzung, daß $\breve{G}(M,\tau D)$ regulär ist, sind die Stufenwerte $S_{m+1}$ und $\hat{S}_{m+1}$ der Systeme (5.3) und (5.4) im linearen Fall eindeutig bestimmt, und es folgt

$$\begin{aligned}
\breve{\hat{S}}_{m+1} - \breve{S}_{m+1} &= \breve{G}(M,\tau D)^{-1}\left(\mathbb{1}_{s-1} \otimes M + \tau\breve{\mathfrak{a}} \otimes D\right)\left(\hat{U}_m - U_m\right) \\
&\quad + \breve{G}(M,\tau D)^{-1}(I_{s-1} \otimes M)\breve{\Theta}_{m+1} + \tau\breve{G}(M,\tau D)^{-1}\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\rho_{m+1} \qquad (5.101a) \\
&= \breve{G}(M,\tau D)^{-1}\left(\mathbb{1}_{s-1} \otimes M + \tau\breve{\mathfrak{a}} \otimes D\right)\left(\hat{U}_m - U_m\right) + \breve{\Theta}_{m+1} \\
&\quad + \tau\breve{G}(M,\tau D)^{-1}\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\left((I_s \otimes D)\Theta_{m+1} + \rho_{m+1}\right). \qquad (5.101b)
\end{aligned}$$

Mit dem Einheitsvektor $\vec{e}_{s-1} = (0, \ldots, 0, 1)^\top \in \mathbb{R}^{s-1}$, der Definition

$$\breve{J}(O,K) = \left(\vec{e}_{s-1}^{\,\top} \otimes I_{\bar{N}}\right)\breve{G}(O,K)^{-1}\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{\bar{N}}\right) \tag{5.102}$$

sowie analog (5.9) definiertem $\breve{R}$ und der in (5.100) eingeführten Matrix $\breve{G}$ folgt

$$\begin{aligned}
\breve{R}(M,\tau D) &= I_{Nn} + \tau\breve{J}(M,\tau D)(\mathbb{1}_s \otimes D) \\
&= I_{Nn} + \tau\left(\vec{e}_{s-1}^{\,\top} \otimes I_{Nn}\right)\breve{G}(M,\tau D)^{-1}\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)(\mathbb{1}_s \otimes D) \\
&= I_{Nn} + \left(\vec{e}_{s-1}^{\,\top} \otimes I_{Nn}\right)\breve{G}(M,\tau D)^{-1}\left(\tau\breve{\mathfrak{a}} \otimes D + \tau(\breve{\mathfrak{A}} \otimes D)(\mathbb{1}_{s-1} \otimes I_{Nn})\right) \\
&= I_{Nn} + \left(\vec{e}_{s-1}^{\,\top} \otimes I_{Nn}\right)\breve{G}(M,\tau D)^{-1} \\
&\quad \cdot \left(\tau\breve{\mathfrak{a}} \otimes D + (I_{s-1} \otimes M - \breve{G}(M,\tau D))(\mathbb{1}_{s-1} \otimes I_{Nn})\right) \\
&= \left(\vec{e}_{s-1}^{\,\top} \otimes I_{Nn}\right)\breve{G}(M,\tau D)^{-1}\left(\mathbb{1}_{s-1} \otimes M + \tau\breve{\mathfrak{a}} \otimes D\right). \qquad (5.103)
\end{aligned}$$

Aus den Runge-Kutta-Gleichungen (5.4a) und (5.4b) folgt wegen der Steifgenauigkeit und (5.99b) $\hat{U}_{m+1} = \hat{U}^{(s)}_{m+1}$ und mit (5.98b)

$$\hat{U}_{m+1} - U_{m+1} = \hat{U}^{(s)}_{m+1} - U^{(s)}_{m+1} = (\vec{e}_{s-1}^{\top} \otimes I_{Nn}) \left( \check{\hat{S}}_{m+1} - \check{S}_{m+1} \right).$$

Einsetzen von (5.101b) liefert schließlich mit (5.99b) und analog (5.10) definiertem $\check{L}$

$$\hat{U}_{m+1} - U_{m+1} = \check{R}(M, \tau D) \left( \hat{U}_m - U_m \right) + \tau \check{J}(M, \tau D) \rho_{m+1} + \theta_{m+1} + \check{L}(M, \tau D) \Theta_{m+1},$$

eine zu (5.14) analoge Gleichung. Mit der Definition

$$\check{H}(O, K) = \check{G}(O, K)^{-1} \left( \mathbb{1}_{s-1} \otimes O + \tau \check{\mathfrak{a}} \otimes K \right) \tag{5.104}$$

erhält man aus (5.101a)

$$\begin{aligned} \check{\hat{S}}_{m+1} - \check{S}_{m+1} = {} & \check{H}(M, \tau D) \left( \hat{U}_m - U_m \right) + \check{G}(M, \tau D)^{-1} (I_{s-1} \otimes M) \check{\Theta}_{m+1} \\ & + \tau \check{G}(M, \tau D)^{-1} \left( (\check{\mathfrak{a}}, \check{\mathfrak{A}}) \otimes I_{Nn} \right) \rho_{m+1}. \end{aligned}$$

Damit folgt die Gültigkeit des Störungslemmas 5.3 mit den entsprechenden $\check{}$-Größen auch für die in diesem Abschnitt betrachteten Runge-Kutta-Verfahren, wobei der Term $\mathfrak{A} \otimes I_{Nn}$ durch $(\check{\mathfrak{a}}, \check{\mathfrak{A}}) \otimes I_{Nn}$ ersetzt werden muß.
Aus der vereinfachenden Bedingung $C(q)$ (siehe (2.17)) folgt für $k = 1$

$$\sum_{j=1}^{s} a_{ij} = c_i, \quad i = 1, \ldots, s, \tag{5.105}$$

und damit $c_1 = 0$. Für den in (5.16b) definierten Residuenfehler des ersten Stufenwertes folgt deshalb

$$\Delta^{(1)}_m = 0.$$

Für steifgenaue Runge-Kutta-Verfahren (d. h., (2.20) ist erfüllt) gilt aufgrund der Konsistenzbedingung (2.16) mit $k = 1$ und wegen (5.105)

$$1 = \sum_{i=1}^{s} b_i = \sum_{i=1}^{s} a_{si} = c_s.$$

Damit folgt für die in (5.16) definierten Residuenfehler des Verfahrens und der Stufenwerte

$$\delta_{m+1} = \Delta^{(s)}_{m+1}.$$

Die Voraussetzungen (5.99) sind für die Wahl (5.17) der Störungen damit erfüllt. Setzt man

$$\check{\Delta}_m = \left( {\Delta^{(2)}_m}^{\top}, \ldots, {\Delta^{(s)}_m}^{\top} \right)^{\top},$$

so gilt deshalb auch eine zur Fehlergleichung (5.28) analoge Formel mit den entsprechenden $\check{}$-Größen für die in diesem Abschnitt betrachteten Runge-Kutta-Verfahren.
Mit (5.29) gilt für die in (5.100) und (5.102) definierten Matrizen $\check{G}(O, K)$ und $\check{J}(O, K)$

$$(I_{s-1} \otimes P) \check{G}(O, K) (I_{s-1} \otimes S) = \check{G}(\mathrm{diag}_k\{A_k\}, \mathrm{diag}_k\{C_k\}),$$

$$\begin{aligned} S^{-1} \check{J}(O, K)(I_s \otimes P^{-1}) &= (\vec{e}_{s-1}^{\top} \otimes I_{\bar{N}})(I_{s-1} \otimes S^{-1}) \check{G}(O, K)^{-1} (I_{s-1} \otimes P^{-1}) \left( (\check{\mathfrak{a}}, \check{\mathfrak{A}}) \otimes I_{\bar{N}} \right) \\ &= (\vec{e}_{s-1}^{\top} \otimes I_{\bar{N}}) \check{G}(\mathrm{diag}_k\{A_k\}, \mathrm{diag}_k\{C_k\})^{-1} \left( (\check{\mathfrak{a}}, \check{\mathfrak{A}}) \otimes I_{\bar{N}} \right) \\ &= \check{J}(\mathrm{diag}_k\{A_k\}, \mathrm{diag}_k\{C_k\}). \end{aligned}$$

Analog zu (5.33a) gilt

$$\breve{G}(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\})^{-1} = \left(\mathrm{diag}_k\left\{\left(\breve{G}(A_k,C_k)^{-1}\right)_{ij}\right\}\right)_{i,j=1,\ldots,s-1},$$

und es folgt

$$\begin{aligned}\breve{J}(\mathrm{diag}_k\{A_k\},\mathrm{diag}_k\{C_k\}) &= \left(\vec{e}_{s-1}^{\top}\otimes I_{\bar{N}}\right)\left(\mathrm{diag}_k\left\{\left(\breve{G}(A_k,C_k)^{-1}\right)_{ij}\right\}\right)_{i,j=1,\ldots,s-1}\\ &\quad\left((\breve{\mathfrak{a}},\breve{\mathfrak{A}})\otimes I_{\bar{N}}\right)\\ &= \left(\mathrm{diag}_k\left\{\left(\left(\vec{e}_{s-1}^{\top}\otimes I_{n_k}\right)\breve{G}(A_k,C_k)^{-1}\left((\breve{\mathfrak{a}},\breve{\mathfrak{A}})\otimes I_{n_k}\right)\right)_j\right\}\right)_{j=1,\ldots,s}\\ &= \left(\mathrm{diag}_k\left\{\left(\breve{J}(A_k,C_k)\right)_j\right\}\right)_{j=1,\ldots,s}.\end{aligned}$$

Damit erhält man analog zu (5.30c), (5.30d), (5.33c), (5.33d) und (5.78)

$$\begin{aligned}S^{-1}\breve{R}(O,K)S &= \mathrm{diag}_k\left\{\breve{R}(A_k,C_k)\right\},\\ S^{-1}\breve{L}(O,K)(I_s\otimes S) &= \left(\mathrm{diag}_k\left\{\left(\breve{L}(A_k,C_k)\right)_j\right\}\right)_{j=1,\ldots,s},\\ (I_{s-1}\otimes S^{-1})\breve{H}(O,K)(I_s\otimes S) &= \left(\mathrm{diag}_k\left\{\left(\breve{H}(A_k,C_k)\right)_i\right\}\right)_{i=1,\ldots,s-1},\end{aligned}$$

und es folgt die Gültigkeit von Konvergenzsatz 5.11 mit den entsprechenden $\breve{}$-Größen:

**Satz 5.41** Sei $p^\star \in [1,\min(p,r+1+\alpha_r:\ r=q+1,\ldots,p)]$ mit $\alpha_r\in\mathbb{R}$, $\alpha_r\geq -1$, die größte reelle Zahl, so daß mit $\bar{M}\in\mathbb{N}$ für $\tau\to 0$ ($M_e\to\infty$), $h_v\to 0$ ($N_v\to\infty$), $v=1,\ldots,d$, $r=q+1,\ldots,p$ und alle $\vec{k}$ mit $k_v=1,\ldots,N_v$ die folgenden Voraussetzungen erfüllt sind:

(a) $\breve{G}(A,\tau D_{\vec{k}})$ ist regulär,

(b) $\breve{W}_{r\alpha}(A,\tau D_{\vec{k}})$ existiert analog zu (5.27),

(c) für $i=0,\ldots,\bar{M}$ sind $\|\tau\breve{R}(A,\tau D_{\vec{k}})^i\breve{J}(A,\tau D_{\vec{k}})\,(I_s\otimes B_v)\,\|$, $\|\breve{R}(A,\tau D_{\vec{k}})^i\tau^{p+1-p^\star}\|$ und $\|\breve{R}(A,\tau D_{\vec{k}})^i\breve{L}(A,\tau D_{\vec{k}})\tau^{p+1-p^\star}\|$ beschränkt,

(d) für $i=\bar{M}+1,\ldots,M_e-1$ sind $\|\breve{R}(A,\tau D_{\vec{k}})^i\breve{J}(A,\tau D_{\vec{k}})\,(I_s\otimes B_v)\,\|$, $\|\breve{R}(A,\tau D_{\vec{k}})^i\tau^{p-p^\star}\|$ und $\|\breve{R}(A,\tau D_{\vec{k}})^i\breve{L}(A,\tau D_{\vec{k}})\tau^{p-p^\star}\|$ beschränkt,

(e) $\|\tau^{r+1+\alpha_r-p^\star}\breve{R}(A,\tau D_{\vec{k}})^i\breve{W}_{r\alpha}(A,\tau D_{\vec{k}})\|$ ist für $i=0,\ldots,M_e$ beschränkt,

(f) $\|D^{1+\alpha_r}U_{\vec{h}}^{(r)}(t)\|$ und $\|D^{1+\alpha_r}U_{\vec{h}}^{(r+1)}(t)\|$ sind für $t\in[t_0,t_e]$ beschränkt.

Dann ist das Diskretisierungsverfahren (5.3) für lineare Systeme mit genügend glatter Lösung konvergent mit der Ordnung $(p_1,\ldots,p_d,p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

Setzt man

$$\breve{E}_{m+1} = \left({E_{m+1}^{(2)}}^{\top},\ldots,{E_{m+1}^{(s)}}^{\top}\right)^{\top},$$

so läßt sich auch Lemma 5.35 übertragen, wobei der Term $\|\tau G(A,\tau D_{\vec{k}})^{-1}(\mathfrak{A}\otimes B_v)\|$ durch $\|\tau\breve{G}(A,\tau D_{\vec{k}})^{-1}((\breve{\mathfrak{a}},\breve{\mathfrak{A}})\otimes B_v)\|$ ersetzt werden muß :

**Lemma 5.42** Sei $q^\star \in [1, \min(p, q+1)]$ die größte ganze Zahl, so daß mit $\bar{M} \in \mathbb{N}$ für $\tau \to 0$ ($M_e \to \infty$), $h_v \to 0$ ($N_v \to \infty$), $v = 1, \ldots, d$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die folgenden Voraussetzungen erfüllt sind:

(a) $\breve{G}(A, \tau D_{\vec{k}})$ ist regulär,

(b) $\breve{W}_{r(-1)}(A, \tau D_{\vec{k}})$ existiert analog zu (5.27), $r = q+1, \ldots, p$,

(c) $\breve{G}(A, \tau D_{\vec{k}})^{-1}(I_{s-1} \otimes A)\tau^{q+1-q^\star}\|$ und $\|\tau\breve{G}(A, \tau D_{\vec{k}})^{-1}((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes B_v)\|$ sind beschränkt,

(d) für $i = 0, \ldots, \bar{M}$ sind $\|\tau\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i \breve{J}(A, \tau D_{\vec{k}})\,(I_s \otimes B_v)\,\|$, $\|\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i\tau^{p+1-q^\star}\|$ und $\|\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i \breve{L}(A, \tau D_{\vec{k}})\tau^{p+1-q^\star}\|$ beschränkt,

(e) für $i = \bar{M}+1, \ldots, M_e - 1$ sind $\|\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i \breve{J}(A, \tau D_{\vec{k}})\,(I_s \otimes B_v)\,\|$, $\|\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i\tau^{p-q^\star}\|$ und $\|\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i \breve{L}(A, \tau D_{\vec{k}})\tau^{p-q^\star}\|$ beschränkt,

(f) $\|\tau^{r-q^\star}\breve{H}(A, \tau D_{\vec{k}})\breve{R}(A, \tau D_{\vec{k}})^i\breve{W}_{r(-1)}(A, \tau D_{\vec{k}})\|$ ist für $i = 0, \ldots, M_e$ und $r = q+1, \ldots, p$ beschränkt.

Dann gilt für lineare Systeme für den Vektor der Diskretisierungsfehler der Stufen des Runge-Kutta-Verfahrens bei genügend glatter exakter Lösung

$$\breve{E}_{m+1} = \mathcal{O}\left(\tau^{q^\star}\right) + \mathcal{O}\left(\sum_{v=1}^{d} h_v^{p_v}\right).$$

□

Aus (5.82) folgt mit (5.98a)

$$\begin{aligned}\breve{G}(M, \tau D)\breve{S}_{m+1} &= \tau(\breve{\mathfrak{a}} \otimes D)U^{(1)}_{m+1} + (\mathbb{1}_{s-1} \otimes M)\,U_m + \tau\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\bar{F}(t_{m+1}, U^{(1)}_{m+1}, \breve{S}_{m+1}) \\ &= (\mathbb{1}_{s-1} \otimes M + \tau\breve{\mathfrak{a}} \otimes D)\,U_m + \tau\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\bar{F}(t_{m+1}, U_m, \breve{S}_{m+1})\end{aligned}$$

und mit der in (5.104) definierten Matrix $\breve{H}(M, \tau D)$ bzw. (5.98b) und den Gleichungen (5.103) und (5.102) für $\breve{R}$ und $\breve{J}$

$$\begin{aligned}\breve{S}_{m+1} &= \breve{H}(M, \tau D)U_m + \tau\breve{G}(M, \tau D)^{-1}\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\bar{F}(t_{m+1}, U_m, \breve{S}_{m+1}), \\ U_{m+1} &= \left(\vec{e}_{s-1}^{\top} \otimes I_{Nn}\right)\breve{S}_{m+1} \\ &= \breve{R}(M, \tau D)U_m + \tau\breve{J}(M, \tau D)\bar{F}(t_{m+1}, U_m, \breve{S}_{m+1}).\end{aligned}$$

Durch Ausführen der Rekursion erhält man daraus die zu (5.81) und (5.83) analogen Gleichungen

$$\begin{aligned}U_{m+1} &= \breve{R}(M, \tau D)^{m+1}U_0 + \tau\sum_{i=0}^{m}\breve{R}(M, \tau D)^i\breve{J}(M, \tau D)\bar{F}(t_{m+1-i}, U_{m-i}, \breve{S}_{m+1-i}), \\ \breve{S}_{m+1} &= \breve{H}(M, \tau D)\breve{R}(M, \tau D)^m U_0 + \tau\breve{G}(M, \tau D)^{-1}\left((\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \otimes I_{Nn}\right)\bar{F}(t_{m+1}, U_m, \breve{S}_{m+1}) \\ &\quad + \tau\sum_{i=0}^{m-1}\breve{H}(M, \tau D)\breve{R}(M, \tau D)^i\breve{J}(M, \tau D)\bar{F}(t_{m-i}, U_{m-1-i}, \breve{S}_{m-i}).\end{aligned}$$

Aus der Lipschitz-Bedingung (5.84) folgt für $Z_1, Z_2 \in \mathbb{R}^{(s-1)Nn}$

$$\|\left(I_{(s-1)N} \otimes \vec{e}_i^{\top}\right)\left(\bar{F}(t_{m+1}, Z_1) - \bar{F}(t_{m+1}, Z_2)\right)\| \le l_{f_i}\|Z_1 - Z_2\|.$$

Durch Betrachtung von

$$\breve{T}:\ Z\in\mathbb{R}^{(s-1)Nn}\to \breve{H}(M,\tau D)\breve{R}(M,\tau D)^m U_0+\tau\breve{G}(M,\tau D)^{-1}\left((\breve{\mathfrak{a}},\breve{\mathfrak{A}})\otimes I_{Nn}\right)\bar{F}(t_{m+1},U_m,Z)$$
$$+\tau\sum_{i=0}^{m-1}\breve{H}(M,\tau D)\breve{R}(M,\tau D)^i\breve{J}(M,\tau D)\bar{F}(t_{m-i},U_{m-1-i},\breve{S}_{m-i})$$

anstelle von (5.85) erhält man deshalb analog zu Satz 5.36 den folgenden Existenzsatz:

**Satz 5.43** Sind mit einem $\kappa_1<1$ für $\tau\to 0$ ($M_e\to\infty$), $h_v\to 0$ ($N_v\to\infty$), $v=1,\dots,d$ und alle $\vec{k}$ mit $k_v=1,\dots,N_v$ die Matrizen $\breve{G}(A,\tau D_{\vec{k}})$ regulär und gilt

$$\tau\prod_{i=1}^{d}\|S_{P_i}\|\ \|S_{P_i}^{-1}\|\sum_{j=1}^{n}l_{f_j}\max_{\vec{k}}\|\breve{G}(A,\tau D_{\vec{k}})^{-1}\left((\breve{\mathfrak{a}},\breve{\mathfrak{A}})\otimes\vec{e}_j\right)\|\le\kappa_1,$$

so besitzt das Verfahren (5.3) für hinreichend kleine $\tau$ und $\vec{h}$ eine eindeutige Lösung. □

Auch die Gleichungen (5.88) - (5.95) lassen sich übertragen, man erhält schließlich analog zu Satz 5.38 den folgenden Konvergenzsatz:

**Satz 5.44** Sei $f$ komponentenweise global Lipschitz-stetig mit den Konstanten $l_{f_j}$, $j=1,\dots,n$. Gilt dann zusätzlich zu den Voraussetzungen von Konvergenzsatz 5.41, Lemma 5.42 (Konvergenz der Stufenwerte) und Satz 5.43 für $j=1,\dots,n$, daß

$$l_{f_j}\|\tau\breve{R}(A,\tau D_{\vec{k}})^i\breve{J}(A,\tau D_{\vec{k}})\,(I_s\otimes\vec{e}_j)\,\|\quad\text{für}\quad i=0,\dots,\bar{M},$$

$$l_{f_j}\|\breve{R}(A,\tau D_{\vec{k}})^i\breve{J}(A,\tau D_{\vec{k}})\,(I_s\otimes\vec{e}_j)\,\|\quad\text{für}\quad i=\bar{M}+1,\dots,M_e-1$$

sowie

$$l_{f_j}\|\breve{H}(A,\tau D_{\vec{k}})\breve{R}(A,\tau D_{\vec{k}})^i\breve{J}(A,\tau D_{\vec{k}})\,(I_s\otimes\vec{e}_j)\,\|\quad\text{für}\quad i=0,\dots,M_e-1$$

sämtlich beschränkt sind, so ist das Diskretisierungsverfahren (5.3) konvergent mit der Ordnung $(p_1,\dots,p_d,\min\{p^\star,q^\star\})$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

Im linearen Fall lassen sich auch für die hier betrachteten Verfahren Aussagen über die Konvergenz in Abhängigkeit vom Zeitindex treffen:
Wegen

$$b^\top(I_s-z\mathfrak{A})^{-1}=\vec{e}_s^{\,\top}(I_s-z\mathfrak{A})^{-1}\mathfrak{A}=\vec{e}_{s-1}^{\,\top}(I_{s-1}-z\breve{\mathfrak{A}})^{-1}(\breve{\mathfrak{a}},\breve{\mathfrak{A}})$$

stimmt die entsprechend (5.102) definierte Funktion $\breve{J}(z)=\breve{J}(1,z)$ mit der in (5.62a) definierten überein und damit dann auch $R(z)$ und $L(z)$. Aus $\tilde{J}(z)=\frac{1}{z}J(\frac{\tau}{z})$ folgt, daß auch

$$\breve{\tilde{J}}(z)=\breve{J}(z,\tau)=\vec{e}_{s-1}^{\,\top}(zI_{s-1}-\tau\breve{\mathfrak{A}})^{-1}(\breve{\mathfrak{a}},\breve{\mathfrak{A}})$$

für $z\neq 0$ mit der in (5.62b) definierten Funktion $\tilde{J}(z)$ übereinstimmt, für $z=0$ ist sie gleich dem entsprechenden Grenzwert. Konvergenzsatz 5.28 kann damit direkt übertragen werden:

**Satz 5.45** Seien mit $\alpha_r\in\mathbb{R}$, $\alpha_r\ge -1$, für $\tau\to 0$ ($M_e\to\infty$), $h_v\to 0$ ($N_v\to\infty$), $v=1,\dots,d$ und alle $\vec{k}$ mit $k_v=1,\dots,N_v$ die folgenden Voraussetzungen erfüllt:

(a) Es existieren für die Matrizenbüschel $D_{\vec{k}}+\lambda A$ Weierstraß-Kronecker-Zerlegungen gemäß (4.68) mit beschränkten Normen

$$\|Q_{\vec{k}}\mathrm{diag}\{N^i_{n_{\vec{k}1}},\mathfrak{o},\dots,\mathfrak{o}\}Q_{\vec{k}}^{-1}\|,\dots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\dots,\mathfrak{o},N^i_{m_{\vec{k}l_{\vec{k}}}}\}Q_{\vec{k}}^{-1}\|$$

und

$$\|Q_{\vec{k}}\mathrm{diag}\{N^i_{n_{\vec{k}1}},\mathfrak{o},\dots,\mathfrak{o}\}P_{\vec{k}}B_v\|,\dots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\dots,\mathfrak{o},N^i_{m_{\vec{k}l_{\vec{k}}}}\}P_{\vec{k}}B_v\|$$

für $i=0,\dots,\nu_{dt}-1$,

(b) $\mathfrak{Re}\kappa_{\vec{k}j_1} \leq -\cos\frac{\pi}{n_{\vec{k}j_1}+1} = \begin{cases} 0 & : n_{\vec{k}j_1} = 1 \\ -\frac{1}{2} & : n_{\vec{k}j_1} = 2 \\ -\frac{1}{2}\sqrt{2} & : n_{\vec{k}j_1} = 3 \\ \vdots & \end{cases}, \quad j_1 = 1, \ldots, l_{\vec{k}},$

(c) für $\nu_{dt} \geq 3$ sei (5.67) für $l = 0, \ldots, j-1,\ j = 1, \ldots, \nu_{dt}-1$ erfüllt,

(d) im Fall $\alpha_r \notin \{-1, 0\}$ existieren $\breve{W}_{r\alpha_r}(A, \tau D_{\vec{k}})$ analog zu (5.27) und seien beschränkt, im Fall $\alpha_r \in \{-1, 0\}$ existieren beschränkte Matrizen $\breve{W}_{r\alpha_r}(I_{n_{\vec{k}j_1}}, \tau R_{\vec{k}j_1})$ für $j_1 = 1, \ldots, l_{\vec{k}}$, $r = q+1, \ldots, p$, für $\alpha_r = 0$ sind die Matrixnormen in (5.72) beschränkt,

(e) $\|D^{1+\alpha_r} U_{\vec{h}}^{(r)}(t)\|$ und $\|D^{1+\alpha_r} U_{\vec{h}}^{(r+1)}(t)\|$ sind für $r = q+1, \ldots, p$, $t \in [t_0, t_e]$ beschränkt.

Dann ist das Diskretisierungsverfahren (5.3) für $L$-stabile Runge-Kutta-Methoden (im Fall $\nu_{dt} = 0$ reicht $A$-Stabilität, im Fall $\nu_{dt} = 1$ $A$-Stabilität mit $\lim\limits_{z \to -\infty} R(z) \neq 1$) mit regulärer Verfahrensmatrix für lineare Systeme mit genügend glatter exakter Lösung nach $\nu_{dt}$ Zeitschritten konvergent mit der Ordnung $(p_1, \ldots, p_d, p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes mit

$$p^\star = \min\{p^\star_{\nu_{dt}}, p_r :\ r = q+1, \ldots, p\},$$

wobei

$$p_r = \begin{cases} r+1+\alpha_r - \max\{0, \nu_{dt}-1\} & : \alpha_r \notin \{-1, 0\} \\ r+1+\alpha_r & : \alpha_r \in \{-1, 0\} \text{ und für } i = 0, \ldots, \nu_{dt}-1 \text{ gilt} \\ & \quad \vec{e}_{s-1}^{\top} \breve{\mathfrak{A}}^{-i-1}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \breve{c}^r = r \vec{e}_{s-1}^{\top} \breve{\mathfrak{A}}^{-i}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \breve{c}^{r-1} \\ r+1-\nu_{dt} & : \text{sonst} \end{cases}$$

gilt. □

Aus der vereinfachenden Bedingung $C(q)$ (siehe (2.17)) folgt für $k = 1, \ldots, q$

$$\breve{\mathfrak{A}}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \tilde{c}^{k-1} = \frac{1}{k} (\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \tilde{c}^k$$

und damit anstelle von (5.75a)

$$\vec{e}_{s-1}^{\top} \breve{\mathfrak{A}}^{-l}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \tilde{c}^k = k \vec{e}_{s-1}^{\top} \breve{\mathfrak{A}}^{-l+1}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \tilde{c}^{k-1} = \ldots = k! \vec{e}_{s-1}^{\top} \breve{\mathfrak{A}}^{-l+k}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \mathbb{1}_s.$$

Mit der vereinfachenden Bedingung $B(p)$ (vgl. (2.16)) erhält man für $k = 1, \ldots, q$, $l \leq k$ und $k - l \leq p - 1$ anstelle von (5.75b)

$$\vec{e}_{s-1}^{\top} \breve{\mathfrak{A}}^{-l}(\breve{\mathfrak{a}}, \breve{\mathfrak{A}}) \breve{c}^{k-1} = \frac{k!}{(k-l+1)!}.$$

Damit folgt für die hier betrachteten Verfahren analog die Gültigkeit von Lemma 5.31.
Auch Lemma 5.32 kann im Fall $\alpha_r = 0$ übertragen werden, wobei anstelle der Regularität von $\mathfrak{A}$ diejenige von $\breve{\mathfrak{A}}$ vorausgesetzt werden muß und im Beweis die (5.77) entsprechende rationale Funktion nun mindestens den Nennergrad $s-1$ und einen Zählergrad von höchstens $s-1$ hat.
Die Stabilitätsfunktionen der $s$-stufigen Lobatto-IIIA-Verfahren sind die Padé-Approximationen vom Index $(s-1, s-1)$. Die Bedingung $R(it) \neq 1$ für $t \in \mathbb{R} \setminus \{0\}$ ist deshalb zum Beispiel für das dreistufige Lobatto-IIIA-Verfahren erfüllt.

# Kapitel 6

# Anwendung spezieller Runge-Kutta-Verfahren

Die bisher entwickelte Theorie soll nun am Beispiel zweier spezieller Runge-Kutta-Verfahren, des impliziten Euler-Verfahrens und des dreistufigen Radau-IIA-Verfahrens, angewendet werden.

## 6.1 Das implizite Euler-Verfahren

Für das implizite Euler-Verfahren gelten

$$s = 1, \quad \mathfrak{A} = (1), \quad b = c = (1), \quad p = q = 1.$$

Die Anwendung des impliziten Euler-Verfahrens mit der konstanten Schrittweite $\tau$ auf das DA-System (4.45) liefert das BTCS-Verfahren (backward in time, centered in space)

$$G(M, \tau D)\, U_{m+1} = MU_m + \tau \tilde{F}(t_{m+1}, U_{m+1}),$$

mit $U_0 = U(t_0)$, $U_m \approx U(t_m)$ und

$$G(M, \tau D) = M - \tau D.$$

Aus Satz 5.11 folgt mit

$$\begin{aligned}
G(A, \tau D_{\vec{k}}) &= A - \tau D_{\vec{k}} = A + \tau \left( \sum_{i=1}^{d} \lambda_{i,k_i} B_i + C \right), \\
J(A, \tau D_{\vec{k}}) &= G(A, \tau D_{\vec{k}})^{-1}, \\
R(A, \tau D_{\vec{k}}) &= I_n + \tau G(A, \tau D_{\vec{k}})^{-1} D_{\vec{k}} = G(A, \tau D_{\vec{k}})^{-1} \left( G(A, \tau D_{\vec{k}}) + \tau D_{\vec{k}} \right) \\
&= G(A, \tau D_{\vec{k}})^{-1} A, \\
L(A, \tau D_{\vec{k}}) &= \tau G(A, \tau D_{\vec{k}})^{-1} D_{\vec{k}} = G(A, \tau D_{\vec{k}})^{-1} A - I_n
\end{aligned}$$

**Satz 6.1** Sind mit $\bar{M} \in \mathbb{N}$ für $\tau \to 0$ ($M_e \to \infty$), $h_v \to 0$ ($N_v \to \infty$), $v = 1, \ldots, d$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die Matrizen $G(A, \tau D_{\vec{k}})$ regulär und sind

$$\| \tau \left( G(A, \tau D_{\vec{k}})^{-1} A \right)^{i_1} G(A, \tau D_{\vec{k}})^{-1} B_v \|, \quad \| \left( G(A, \tau D_{\vec{k}})^{-1} A \right)^{i_1+1} \tau \| \quad \text{für} \quad i_1 = 0, \ldots, \bar{M}$$

sowie

$$\| \left( G(A, \tau D_{\vec{k}})^{-1} A \right)^{i_2} G(A, \tau D_{\vec{k}})^{-1} B_v \|, \quad \| \left( G(A, \tau D_{\vec{k}})^{-1} A \right)^{i_2+1} \| \quad \text{für} \quad i_2 = \bar{M}+1, \ldots, M_e - 1$$

sämtlich beschränkt, so ist das BTCS-Verfahren für lineare Systeme konvergent mit der Ordnung $(p_1, \ldots, p_d, 1)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes, wenn die zweiten Ableitungen der exakten Lösung nach der Zeit und für $p_i = 1$ die dritten und für $p_i = 2$ die vierten Ableitungen der exakten Lösung nach $x_i$ beschränkt sind, $i = 1, \ldots, d$. □

**Bemerkung 6.2** Ist $A$ regulär, hat das System also den Zeitindex 0, so folgt schon aus der Beschränktheit von $\|\tau \left(G(A, \tau D_{\vec{k}})^{-1} A\right)^{i_1+1}\|$, $\|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^{i_2+1}\|$ für alle $\vec{k}$ wegen

$$\|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^{i} G(A, \tau D_{\vec{k}})^{-1} B_v\| \leq \|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^{i+1}\| \, \|A^{-1} B_v\|$$

die Beschränktheit der übrigen Matrizen in Satz 6.1, d. h., man erhält das obige Konvergenzresultat. □

**Beispiel 6.3** Gegeben seien das System aus Beispiel 4.13 mit differentiellem Zeitindex 2 sowie konsistente Anfangs- und Randwerte. Die Matrix

$$G(A, \tau D_k) = \begin{pmatrix} 1+\tau & \tau(1-\lambda_k) & 0 \\ 0 & 1 & \tau(1-\lambda_k) \\ 0 & \tau & 0 \end{pmatrix}$$

ist für $\tau > 0$ und $\lambda_k \leq 0$ regulär, es gelten

$$G(A, \tau D_k)^{-1} = \frac{1}{\tau^2(1+\tau)(\lambda_k - 1)} \begin{pmatrix} \tau^2(\lambda_k - 1) & 0 & \tau^2(1-\lambda_k)^2 \\ 0 & 0 & \tau(\lambda_k - 1)(1+\tau) \\ 0 & -\tau(1+\tau) & 1+\tau \end{pmatrix},$$

$$G(A, \tau D_k)^{-1} A = \frac{1}{\tau(1+\tau)(\lambda_k - 1)} \begin{pmatrix} \tau(\lambda_k - 1) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -(1+\tau) & 0 \end{pmatrix},$$

$$G(A, \tau D_k)^{-1} B = \frac{1}{\tau(1+\tau)(\lambda_k - 1)} \begin{pmatrix} 0 & -\tau(\lambda_k - 1) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1+\tau \end{pmatrix}$$

und für $i > 1$

$$\left(G(A, \tau D_k)^{-1} A\right)^{i} = \frac{1}{(1+\tau)^i} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

sowie für $i \geq 1$

$$\left(G(A, \tau D_k)^{-1} A\right)^{i} G(A, \tau D_k)^{-1} B = \frac{1}{(1+\tau)^{i+1}} \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Nach obigem Satz ist das BTCS-Verfahren bei genügend glatter exakter Lösung unbedingt konvergent mit der Ordnung (2,1). Dies wird auch im numerischen Experiment bestätigt: Werden Anfangs- und Dirichlet-Randwerte und die rechte Seite zum Beispiel so gewählt, daß

$$u(t, x) = \begin{pmatrix} e^x(1+t) \\ e^{2x} \sin(t) \\ e^{3x} \cos(t) \end{pmatrix}$$

in $[0, 1] \times [0, 1]$ die exakte Lösung ist, so erhält man als numerisch bestimmte Zeitordnung bezüglich der diskreten $L_2$-Norm 1:

| $0.1\tau^{-1}$ | $2^3$ | $2^4$ | $2^5$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^8$ | 0.99 | 0.99 | 1.00 |
| $2^9$ | 0.99 | 0.99 | 1.00 |

und als numerisch bestimmte Ortsordnung 2:

| $0.1\tau^{-1}$ | $2^3$ | $2^4$ | $2^5$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^8$ | 2.00 | 2.00 | 1.99 |
| $2^9$ | 1.99 | 2.00 | 2.00. |

□

**Bemerkung 6.4** Sei $e_{\tau,\vec{h}} = \|e_{M_e}\| = \|U_{\vec{h}}(t_{M_e}) - U_{M_e}\|$ die diskrete $L_2$-Norm des mit der Zeitschrittweite $\tau$ und den Ortsschrittweiten $\vec{h} = (h_1, \ldots, h_d)$ erhaltenen globalen Diskretisierungsfehlers zum Zeitpunkt $t_e$. Dann kann die numerische Konvergenzordnung bezüglich der Zeit $p_{num}$ bestimmt werden gemäß

$$p_{num\,\frac{\tau}{4},\vec{h}} = \mathrm{ld}\frac{e_{\tau,\vec{h}} - e_{\frac{\tau}{2},\vec{h}}}{e_{\frac{\tau}{2},\vec{h}} - e_{\frac{\tau}{4},\vec{h}}},$$

wobei $\mathrm{ld}(x)$ den Logarithmus von $x$ zur Basis 2 bezeichnet. Unter der Voraussetzung, daß $e_{\tau,\vec{h}} = C_1\tau^p + C_2$ gilt, wobei $C_1$ und $C_2$ von $\vec{h}$ abhängen, folgt

$$p_{num\,\frac{\tau}{4},\vec{h}} = \mathrm{ld}\ \frac{\tau^p - \left(\frac{\tau}{2}\right)^p}{\left(\frac{\tau}{2}\right)^p - \left(\frac{\tau}{4}\right)^p} = \mathrm{ld}\ 2^p = p.$$

Die numerische Konvergenzordnung bezüglich des Ortes wird analog bestimmt. □

Auch Konvergenzsatz 5.28 kann für das BTCS-Verfahren vereinfacht werden. Dazu wird das folgende Lemma verwendet:

**Lemma 6.5** Für das implizite Euler-Verfahren gelten für beliebigen differentiellen Zeitindex $\nu_{dt}$:

(a) Bedingung (5.67) ist erfüllt,

(b) $p^\star_{\nu_{dt}} = 1$.

□

**Beweis:**

(a) Für das implizite Euler-Verfahren gelten nach (5.62b) und (5.62c)

$$\tilde{J}(z) = \frac{1}{z-\tau}, \quad \tilde{R}(z) = 1 + \frac{\tau}{z-\tau} = \frac{z}{z-\tau}.$$

Daraus folgt mit der binomischen Reihe

$$\begin{aligned}
(\tau^{j+1}\tilde{R}(z)^i\tilde{J}(z)\tilde{c}^k)^{(j)}(0) &= \tau^{j+1}\left(\frac{d^j}{dz^j}\ \frac{z^i}{(z-\tau)^{i+1}}\right)(0) \\
&= (-1)^{i+1}\tau^{j-i}\left(\frac{d^j}{dz^j}\ \frac{z^i}{(1-\frac{z}{\tau})^{i+1}}\right)(0) \\
&= (-1)^{i+1}\tau^j\sum_{l=0}^{\infty}\binom{i+l}{l}\frac{d^j}{dz^j}\left(\frac{z}{\tau}\right)^{l+i}(0) \\
&= (-1)^{i+1}j!\binom{j}{i}
\end{aligned}$$

und damit, daß die Behauptung (a) äquivalent ist zu

$$\sum_{k=0}^{l} \frac{j!}{k!(l-k)!} \sum_{i=0}^{j} (-i)^{l-k} (-1)^{i+1} \binom{j}{i} = 0 \quad \text{für} \quad l = 0, \ldots, j-1$$

für alle $j \geq 1$. Dies ist genau dann erfüllt, wenn für $j \geq 1$

$$\sum_{i=0}^{j} i^k (-1)^i \binom{j}{i} = 0 \quad \text{für} \quad k = 0, \ldots, j-1$$

gilt, was mittels vollständiger Induktion über $k$ bewiesen werden soll:

- Ind.-Anf.: $\sum_{i=0}^{j} (-1)^i \binom{j}{i} = 0$ für $j \geq 1$ $(k = 0)$.
- Ind.-Vor.: $\sum_{i=0}^{j} i^l (-1)^i \binom{j}{i} = 0$ für $j > k$, $0 \leq l \leq k$.
- Ind.-Beh.: $\sum_{i=0}^{j} i^{k+1} (-1)^i \binom{j}{i} = 0$ für $j > k+1$.
- Ind.-Bew.:

$$\begin{aligned} \sum_{i=0}^{j} i^{k+1} (-1)^i \binom{j}{i} &= \sum_{i=1}^{j} i^k (-1)^i \binom{j-1}{i-1} j = -j \sum_{i=0}^{j-1} (i+1)^k (-1)^i \binom{j-1}{i} \\ &= -j \sum_{l=0}^{k} \binom{k}{l} \sum_{i=0}^{j-1} i^l (-1)^i \binom{j-1}{i} = 0 \end{aligned}$$

  nach Induktionsvoraussetzung, da $j - 1 > k \geq l$ gilt.

(b) Damit $p^{\star}_{\nu_{dt}} = 1$ gilt, müssen (5.69) und (5.70) für alle $j \geq 2$ und $l = 2, \ldots, j$ erfüllt sein. Nach (5.62d) gilt $\tilde{L}(z) = \tau \tilde{J}(z)$. Damit folgt die Gültigkeit von (5.70) wie im Beweis von (a).
Wegen

$$\begin{aligned} \left(\tau^j \tilde{R}(z)^i\right)^{(j)}(0) &= \tau^j \left(\frac{d^j}{dz^j} \frac{z^i}{(z-\tau)^i}\right)(0) = (-1)^i \tau^{j-i} \left(\frac{d^j}{dz^j} \frac{z^i}{\left(1-\frac{z}{\tau}\right)^i}\right)(0) \\ &= (-1)^i \tau^j \sum_{l=0}^{\infty} \binom{i+l-1}{l} \frac{d^j}{dz^j} \left(\frac{z}{\tau}\right)^{l+i}(0) \\ &= \begin{cases} 0 & : \ i = 0 \\ (-1)^i j! \binom{j-1}{i-1} & : \ i \geq 1 \end{cases} \end{aligned}$$

folgt wie im obigen Induktionsbeweis

$$\sum_{i=0}^{j} i^k \left(\tau^j \tilde{R}(z)^i\right)^{(j)}(0) = \sum_{i=1}^{j} i^k (-1)^i j! \binom{j-1}{i-1} = 0$$

für $k = 0, \ldots, j-2$, $j \geq 2$ und damit auch (5.69) für $l = 2, \ldots, j$.

□

Mit diesem Lemma erhält man aus Satz 5.28 den folgenden Konvergenzsatz für das BTCS-Verfahren:

**Satz 6.6** Seien für $\tau \to 0$ ($M_e \to \infty$), $h_v \to 0$ ($N_v \to \infty$), $v = 1, \ldots, d$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die folgenden Voraussetzungen erfüllt:

(a) Es existieren für die Matrizenbüschel $D_{\vec{k}} + \lambda A$ Weierstraß-Kronecker-Zerlegungen gemäß (4.68) mit beschränkten Normen

$$\|Q_{\vec{k}} \mathrm{diag}\{N^i_{n_{\vec{k}1}}, \mathfrak{o}, \ldots, \mathfrak{o}\} Q^{-1}_{\vec{k}}\|, \ldots, \|Q_{\vec{k}} \mathrm{diag}\{\mathfrak{o}, \ldots, \mathfrak{o}, N^i_{m_{\vec{k}l_{\vec{k}}}}\} Q^{-1}_{\vec{k}}\|$$

und

$$\|Q_{\vec{k}} \mathrm{diag}\{N^i_{n_{\vec{k}1}}, \mathfrak{o}, \ldots, \mathfrak{o}\} P_{\vec{k}} B_v\|, \ldots, \|Q_{\vec{k}} \mathrm{diag}\{\mathfrak{o}, \ldots, \mathfrak{o}, N^i_{m_{\vec{k}l_{\vec{k}}}}\} P_{\vec{k}} B_v\|$$

für $i = 0, \ldots, \nu_{dt} - 1$,

(b) $\mathfrak{Re}\kappa_{\vec{k}j_1} \le -\cos\frac{\pi}{n_{\vec{k}j_1}+1} = \begin{cases} 0 & : n_{\vec{k}j_1} = 1 \\ -\frac{1}{2} & : n_{\vec{k}j_1} = 2 \\ -\frac{1}{2}\sqrt{2} & : n_{\vec{k}j_1} = 3 \\ \ldots & \end{cases}, \quad j_1 = 1, \ldots, l_{\vec{k}}.$

Dann ist das BTCS-Verfahren für lineare Systeme mit genügend glatter exakter Lösung nach $\nu_{dt}$ Zeitschritten konvergent mit der Ordnung $(p_1, \ldots, p_d, 1)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

**Bemerkung 6.7** Die Konvergenz für beliebigen Zeitindex ist an die Voraussetzung konstanter Zeitschrittweite gebunden. Im Fall variabler Schrittweite erhält man für Systeme mit einem Zeitindex größer als 2 auch bei Erfüllung der Voraussetzungen (a) und (b) von Satz 6.6 im allgemeinen keine Konvergenz. □

**Beispiel 6.8** Betrachtet werde wie in Beispiel 6.3 wieder das System aus Beispiel 4.13 mit differentiellem Zeitindex 2. Für die Weierstraß-Kronecker-Transformation aus Beispiel 4.15 gelten neben den dort gezeigten Eigenschaften zusätzlich

$$Q_k \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} Q_k^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{und} \quad Q_k \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} Q_k^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{\lambda_k - 1} & 0 \end{pmatrix}.$$

Damit sind auch die Voraussetzungen von Konvergenzsatz 6.6 erfüllt. □

**Beispiel 6.9** Betrachtet werde das lineare PDA-System

$$\underbrace{\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}}_{=A} u_t + \underbrace{\begin{pmatrix} 0 & 0 & -1 \\ 0 & -1 & -1 \\ 0 & 0 & 0 \end{pmatrix}}_{=B} u_{xx} + \underbrace{\begin{pmatrix} -1 & -1 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}}_{=C} u = f$$

mit $x \in [-0.5, 0.5]$, $t \in [0.0, 1.0]$. Die rechte Seite, Anfangs- und Dirichlet-Randbedingungen werden so gewählt, daß

$$u(t,x) = \begin{pmatrix} e^x \sin(t) \\ e^{2x} \cos(t) \\ e^{3x+t} \end{pmatrix}$$

die exakte Lösung ist.
Aus (4.59) folgt

$$D_k = -\lambda_k B - C = \begin{pmatrix} 1 & 1 & 1+\lambda_k \\ 0 & 1+\lambda_k & \lambda_k \\ 0 & 0 & 1 \end{pmatrix}.$$

Mit

$$P_k = \begin{pmatrix} \lambda_k + 1 & -1 & -\lambda_k - \lambda_k^2 - 1 \\ 0 & 1 & -\lambda_k - 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{und} \quad Q_k = \begin{pmatrix} \frac{1}{\lambda_k+1} & 0 & 0 \\ 0 & \frac{1}{\lambda_k+1} & \frac{1}{\lambda_k+1} \\ 0 & 0 & 1 \end{pmatrix}$$

erhält man die Weierstraß-Kronecker-Zerlegung

$$P_k A Q_k = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad P_k D_k Q_k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Das PDA-System hat also den differentiellen Zeitindex 3. Wegen

$$Q_k^{-1} = \begin{pmatrix} \lambda_k + 1 & 0 & 0 \\ 0 & \lambda_k + 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}$$

gelten

$$Q_k \begin{pmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{pmatrix} Q_k^{-1} = \begin{pmatrix} a & b & \frac{c-b}{\lambda_k+1} \\ 0 & a & \frac{b}{\lambda_k+1} \\ 0 & 0 & a \end{pmatrix}$$

und

$$Q_k \begin{pmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{pmatrix} P_k B = \begin{pmatrix} 0 & \frac{a-b}{\lambda_k+1} & -\frac{a\lambda_k+b}{\lambda_k+1} \\ 0 & -\frac{a}{\lambda_k+1} & -\frac{a}{\lambda_k+1} \\ 0 & 0 & 0 \end{pmatrix}.$$

Da aus der Gleichung für die Eigenwerte (4.16) mit der Gleichung (4.2) für $h$ und der Abschätzung (5.51b)

$$\lambda_j = -\frac{4}{h^2} \sin^2 \frac{j\pi}{2(N+1)} = -\frac{(N+1)^2}{l^2} \sin^2 \frac{j\pi}{2(N+1)} \leq -\frac{j^2}{l^2} \leq -4 \tag{6.1}$$

folgt, sind die Voraussetzungen von Konvergenzsatz 6.6 damit erfüllt, und das BTCS-Verfahren konvergiert (nach 3 Zeitschritten).
Numerisch erhält man im Fall konstanter Zeitschrittweiten eine Konvergenzordnung bezüglich der Zeit von 1:

| $0.1\tau^{-1}$ | $2^4$ | $2^5$ | $2^6$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^8$ | 1.01 | 1.00 | 1.00 |
| $2^9$ | 1.01 | 1.00 | 1.00 |

und eine Konvergenzordnung bezüglich des Ortes von 2:

| $0.1\tau^{-1}$ | $2^4$ | $2^5$ | $2^6$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^8$ | 1.95 | 1.96 | 1.95 |
| $2^9$ | 1.91 | 1.95 | 1.97. |

In der folgenden Tabelle ist der Fehler nach dem jeweils ersten Zeitschritt aufgetragen, und man sieht, daß er für kleiner werdende $\tau$ nicht gegen Null geht:

| $0.1\tau^{-1}$ | $2^5$ | $2^6$ | $2^7$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^7$ | 0.062 | 0.063 | 0.064 |
| $2^8$ | 0.062 | 0.063 | 0.063 |
| $2^9$ | 0.062 | 0.063 | 0.063. |

Auch im Fall variabler Schrittweiten konvergiert das BTCS-Verfahren nicht, wie numerische Rechnungen bestätigen. □

Im nichtlinearen Fall folgt für global Lipschitz-stetige rechte Seiten $f$ mit

$$H(A, \tau D_{\vec{k}}) = G(A, \tau D_{\vec{k}})^{-1} A$$

aus Satz 5.38 der folgende Konvergenzsatz:

**Satz 6.10** Sei $f$ komponentenweise global Lipschitz-stetig mit den Konstanten $l_{f_j}$, $j = 1, \ldots, n$. Sind mit $\bar{M} \in \mathbb{N}$ für $\tau \to 0$ $(M_e \to \infty)$, $h_v \to 0$ $(N_v \to \infty)$, $v = 1, \ldots, d$ und alle $\vec{k}$ mit $k_v = 1, \ldots, N_v$ die Matrizen $G(A, \tau D_{\vec{k}})$ regulär und sind

$$\|\tau \left(G(A, \tau D_{\vec{k}})^{-1} A\right)^i G(A, \tau D_{\vec{k}})^{-1} B_v\|, \quad \|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^{i+1} \tau\| \quad \text{für} \quad i = 0, \ldots, \bar{M}$$

und

$$\|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^i G(A, \tau D_{\vec{k}})^{-1} B_v\|, \quad \|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^{i+1}\| \quad \text{für} \quad i = \bar{M}+1, \ldots, M_e$$

sowie

$$l_{f_j} \|\tau G(A, \tau D_{\vec{k}})^{-1} A G(A, \tau D_{\vec{k}})^{-1} \vec{e}_j\|, \quad l_{f_j} \|\left(G(A, \tau D_{\vec{k}})^{-1} A\right)^i G(A, \tau D_{\vec{k}})^{-1} \vec{e}_j\|$$

für $j = 1, \ldots, n$ und $i = 1, \ldots, M_e$ sämtlich beschränkt und existiert $\kappa_1 < 1$, so daß

$$\tau \prod_{i=1}^{d} \|S_{P_i}\| \, \|S_{P_i}^{-1}\| \sum_{j=1}^{n} l_{f_j} \max_{\vec{k}} \|G(A, \tau D_{\vec{k}})^{-1} \vec{e}_j\| \leq \kappa_1$$

gilt, so ist das BTCS-Verfahren für genügend glatte exakte Lösungen konvergent mit der Ordnung $(p_1, \ldots, p_d, 1)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes. □

Die Anwendung dieses Satzes verdeutlichen die folgenden zwei Beispiele:

**Beispiel 6.11** Betrachtet wird das System aus Beispiel 3.2 (Pharmakokinetik in der Leber),

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{=A} u_t + \underbrace{\begin{pmatrix} -\frac{D}{V^2} & 0 \\ 0 & 0 \end{pmatrix}}_{=B} (u_{xx} - \frac{VQ}{D} u_x) = \begin{pmatrix} -k_{12} u_1 + \varepsilon k_{21} u_2 \\ \frac{1}{\varepsilon} k_{12} u_1 - k_{21} u_2 - \frac{V_{\max} u_2}{K_m + u_2} \end{pmatrix}.$$

Es gilt

$$G(A, \tau D_k)^{-1} A = G(A, \tau D_k)^{-1} = (A + \tau \lambda_k B)^{-1} = \begin{pmatrix} \frac{1}{1 - \tau \lambda_k \frac{D}{V^2}} & 0 \\ 0 & 1 \end{pmatrix},$$

daraus folgt wegen $\tau$, $D > 0$ und da für genügend kleine $h$ analog zu (6.1) $\lambda_k \leq 0$ ist,

$$\|G(A, \tau D_k)^{-1} A\| = 1.$$

Da für die exakte Lösung $u_2 \geq 0$ gilt, folgt aus Satz 5.39 und Satz 6.10 mit

$$l_{f_1} = \sqrt{k_{12}^2 + \varepsilon^2 k_{21}^2} \quad \text{und} \quad l_{f_2}(\gamma) = \sqrt{\frac{k_{12}^2}{\varepsilon^2} + \left(k_{21} + \frac{V_{\max} K_m}{(K_m - \gamma)^2}\right)^2}$$

und den Schranken (4.23) an $\|S_P\|$ und $\|S_P^{-1}\|$, daß das BTCS-Verfahren für dieses Beispiel zumindest unter der Bedingung $\frac{\tau^2}{h} \to 0$ mit der Ordnung $(1, 1)$ gegen die exakte Lösung konvergiert. Zum Beispiel erhält man für das System (vgl. [17])

$$u_t = \begin{pmatrix} 0.96 & 0.00 \\ 0.00 & 0.00 \end{pmatrix} (u_{xx} - 10.4439 u_x) + \begin{pmatrix} -3.45 & 3.45 \\ 10.70 & -10.70 \end{pmatrix} u + \begin{pmatrix} 0 \\ \frac{-1520 u_2}{41.3 + u_2} \end{pmatrix}$$

auf $[0, 1] \times [-\frac{1}{2}, \frac{1}{2}]$ mit Anfangswert $u(0, x) = 0$ und Randwerten $u_1(t, -\frac{1}{2}) = 20 t e^{-10t}$, $u_1(t, \frac{1}{2}) = 0$ für $\delta = 0$ (die erste Ortsableitung wird durch den rückwärtsgenommenen Differenzenquotienten approximiert) 1 als numerisch bestimmte Zeitordnung:

| $0.1\tau^{-1}$ | $2^{10}$ | $2^{11}$ | $2^{12}$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^4$ | 1.00 | 1.00 | 1.00 |
| $2^5$ | 0.99 | 0.99 | 1.00 |
| $2^6$ | 0.96 | 0.99 | 0.99 |

und auch 1 als numerisch bestimmte Ortsordnung:

| $0.1\tau^{-1}$ | $2^{10}$ | $2^{11}$ | $2^{12}$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^4$ | 1.09 | 1.09 | 1.09 |
| $2^5$ | 1.04 | 1.04 | 1.04 |
| $2^6$ | 1.02 | 1.02 | 1.02. |

Setzt man dagegen $\delta = \frac{1}{2}$ (die erste Ortsableitung wird durch den zentralen Differenzenquotienten approximiert), so erhält man wieder 1 als numerisch bestimmte Zeitordnung:

| $0.1\tau^{-1}$ | $2^6$ | $2^7$ | $2^8$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^7$ | 1.03 | 1.02 | 1.01 |
| $2^8$ | 1.03 | 1.02 | 1.01, |

aber nun 2 als numerisch bestimmte Ortsordnung:

| $0.1\tau^{-1}$ | $2^6$ | $2^7$ | $2^8$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^7$ | 1.99 | 1.99 | 1.98 |
| $2^8$ | 1.99 | 2.00 | 2.00. |

□

**Beispiel 6.12** Das System aus Beispiel 3.3 (Pulververbrennung) läßt sich schreiben als

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{=A} u_t + \sum_{i=1}^{d} \underbrace{\begin{pmatrix} -k & 0 \\ 0 & 0 \end{pmatrix}}_{=B_i} u_{x_i x_i} = K_0 u_2 e^{-\frac{E}{u_1}} \begin{pmatrix} Q \\ -1 \end{pmatrix}$$

mit $\vec{x} \in (-1,1)^d$, $t \in (0,1)$. Als Anfangsbedingung wird

$$u(t,\vec{x}) = \begin{pmatrix} 500 \\ Y_0 \prod_{i=1}^{d} (1 - x_i^2) \end{pmatrix} \quad \text{mit} \quad Y_0 > 0$$

gewählt. Am Rand werden für $u_1$ homogene Neumann-Randbedingungen vorgeschrieben. Analog zu Beispiel 6.11 folgt

$$G(A, \tau D_{\vec{k}})^{-1} = G(A, \tau D_{\vec{k}})^{-1} A = \begin{pmatrix} \frac{1}{1 - \tau k \sum_{i=1}^{d} \lambda_{i,k_i}} & 0 \\ 0 & 1 \end{pmatrix}.$$

Aus (4.40) folgt $\lambda_{i,k_i} \leq 0$ und damit wegen $\tau$, $k > 0$

$$\| \left( G(A, \tau D_{\vec{k}})^{-1} A \right)^i \| = 1.$$

Da für die exakte Lösung $u_1 \geq 1$ und $0 \leq u_2 \leq Y_0$ gelten, folgt mit

$$\frac{E}{u_1^2} e^{-\frac{E}{u_1}} \leq \frac{4}{Ee^2},$$

daß als Lipschitz-Konstanten

$$l_{f_2}(\gamma) = K_0\sqrt{1+\frac{16(Y_0+\gamma)^2}{E^2e^4}}, \quad l_{f_1}(\gamma) = Ql_{f_2}(\gamma)$$

gewählt werden können. Mit (4.43) erhält man aus den Sätzen 5.39 und 6.10, daß das BTCS-Verfahren für dieses Beispiel zumindest unter der Bedingung $\frac{\tau^2}{h} \to 0$ mit der Ordnung $(1,1)$ gegen die exakte Lösung konvergiert. Dies wird auch numerisch bestätigt, für $d=1$, $Y_0=1$, $k=0.0001$, $K_0=5800$, $E=11000$ und $Q=2700$ (vgl. [30]) erhält man als numerisch bestimmte Zeitordnung 1:

| $0.1\tau^{-1}$ | $2^5$ | $2^6$ | $2^7$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 1.00 | 1.00 | 1.00 |
| $2^4$ | 0.98 | 0.99 | 1.00 |
| $2^5$ | 1.00 | 1.00 | 1.00 |

und als numerisch bestimmte Ortsordnung ebenfalls 1:

| $0.1\tau^{-1}$ | $2^5$ | $2^6$ | $2^7$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 0.98 | 0.98 | 0.98 |
| $2^4$ | 0.99 | 0.99 | 0.99 |
| $2^5$ | 1.00 | 1.00 | 1.00. |

□

## 6.2 Das dreistufige Radau-IIA-Verfahren

Das dreistufige Radau-IIA-Verfahren ist festgelegt durch die Verfahrensmatrix

$$\mathfrak{A} = \begin{pmatrix} \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ \frac{16-6\sqrt{6}}{36} & \frac{16+6\sqrt{6}}{36} & \frac{1}{9} \end{pmatrix},$$

den Wichtungsvektor

$$b = (\tfrac{16-6\sqrt{6}}{36}, \tfrac{16+6\sqrt{6}}{36}, \tfrac{1}{9})^\top$$

und den Knotenvektor

$$c = (\tfrac{4-\sqrt{6}}{10}, \tfrac{4+\sqrt{6}}{10}, 1)^\top.$$

Es hat die Konsistenzordnung $p=5$ und die Stufenordnung $q=3$.
Für dieses Verfahren gelten nach (5.71) $p^\star_{\nu_{dt}} = 5$ für $\nu_{dt} \in \{0,1,2\}$ und $p^\star_{\nu_{dt}} \le 7-\nu_{dt}$ für $\nu_{dt} \ge 3$. Nach Lemma 5.31 ist Voraussetzung (c) in Konvergenzsatz 5.28 für $\nu_{dt} \le 4$ erfüllt. Wählt man für $\nu_{dt} \in \{0,1\}$

$$\alpha_4 = \begin{cases} 0 & : \text{ alle Randbedingungen sind homogen im Sinne von (5.57) oder periodisch} \\ -\frac{1}{4}+\varepsilon & : M_D = M_P = \emptyset \text{ und nicht alle Randbedingungen sind homogen} \\ -\frac{3}{4}+\varepsilon & : \text{sonst} \end{cases}$$

und $\alpha_5 = -1$ und für $\nu_{dt} \in \{2,3,4\}$ $\alpha_r = -1$, so erhält man aus Satz 5.28 mit den Sätzen 5.23 und 5.25 und den Lemmata 5.32 und 5.33 zusammengefaßt den folgenden Konvergenzsatz:

**Satz 6.13** Seien für ein $\varepsilon > 0$ und $\tau \to 0$ $(M_e \to \infty)$, $h_v \to 0$ $(N_v \to \infty)$, $v = 1,\dots,d$ und alle $\vec{k}$ mit $k_v = 1,\dots,N_v$ die folgenden Voraussetzungen erfüllt:

(a) Es existieren für die Matrizenbüschel $D_{\vec{k}}+\lambda A$ Weierstraß-Kronecker-Zerlegungen gemäß (4.68) mit beschränkten Normen

$$\|Q_{\vec{k}}\mathrm{diag}\{N^i_{n_{\vec{k}1}},\mathfrak{o},\ldots,\mathfrak{o}\}Q^{-1}_{\vec{k}}\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N^i_{m_{\vec{k}l_{\vec{k}}}}\}Q^{-1}_{\vec{k}}\|$$

und

$$\|Q_{\vec{k}}\mathrm{diag}\{N^i_{n_{\vec{k}1}},\mathfrak{o},\ldots,\mathfrak{o}\}P_{\vec{k}}B_v\|,\ldots,\|Q_{\vec{k}}\mathrm{diag}\{\mathfrak{o},\ldots,\mathfrak{o},N^i_{m_{\vec{k}l_{\vec{k}}}}\}P_{\vec{k}}B_v\|$$

für $i=0,\ldots,\nu_{dt}-1$,

(b) $\mathfrak{Re}\kappa_{\vec{k}j_1} \le -\cos\frac{\pi}{n_{\vec{k}j_1}+1} = \begin{cases} 0 & : n_{\vec{k}j_1}=1 \\ -\frac{1}{2} & : n_{\vec{k}j_1}=2 \\ -\frac{1}{2}\sqrt{2} & : n_{\vec{k}j_1}=3 \\ \vdots & \end{cases}, \quad j_1=1,\ldots,l_{\vec{k}},$

(c) falls $\nu_{dt}\in\{0,1\}$, so seien $r_i=0$ für $i\in M_N$, es existiere $D^\beta_{\vec{k}}$, und es gelte mit einer von $\vec{h}$ unabhängigen Konstanten $C_1$ für $0\le\beta\le\frac{3}{4}$

$$|(D^\beta_{\vec{k}})_{ij}| \le C_1\left(1+\sum_{v=1}^{d}|\lambda_{v,k_v}|^\beta\right), \quad i,j=1,\ldots,n,$$

sowie

(i) falls alle Randbedingungen homogen im Sinne von (5.57) oder periodisch sind, so sind die Matrixnormen in (5.72) beschränkt,

(ii) falls $M_D=M_P=\emptyset$, so existieren beschränkte Matrizen $W_{4(-\frac{1}{4}+\varepsilon)}(A,\tau D_{\vec{k}})$,

(iii) andernfalls existieren beschränkte Matrizen $W_{4(-\frac{3}{4}+\varepsilon)}(A,\tau D_{\vec{k}})$.

Dann ist das Diskretisierungsverfahren (5.3) für lineare Systeme mit genügend glatter exakter Lösung nach $\nu_{dt}$ Zeitschritten (im Fall (c)(ii) unter der Bedingung $h_i=h$, $i=1,\ldots,d$) konvergent mit der Ordnung $(p_1,\ldots,p_d,p^\star)$ in der Maximumnorm bezüglich der Zeit und in der diskreten $L_2$-Norm bezüglich des Ortes mit

$$p^\star = \begin{cases} 5 & : \nu_{dt}\in\{0,1\} \text{ und alle Randbedingungen sind homogen im Sinne von (5.57) oder periodisch} \\ 4.75-\varepsilon & : \nu_{dt}\in\{0,1\},\ M_D=M_P=\emptyset \text{ und nicht alle Randbedingungen sind homogen} \\ 4.25-\varepsilon & : \nu_{dt}\in\{0,1\},\ M_D\ne\emptyset \text{ und nicht alle Randbedingungen sind homogen} \\ 3 & : \nu_{dt}=2 \\ 2 & : \nu_{dt}=3 \\ 1 & : \nu_{dt}=4 \end{cases}.$$

□

Das folgende Beispiel verdeutlicht die Abhängigkeit der Konvergenzordnung bezüglich der Zeit von der Art der Randbedingungen:

**Beispiel 6.14** Betrachtet werde das Differentialgleichungssystem

$$\underbrace{\begin{pmatrix} 0 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \end{pmatrix}}_{=A} u_t + u_{xx} - u = f(t,x)$$

mit $x \in [-1,1]$ und $t \in [0,1]$. Es gilt nach (4.59) $D_k = (1-\lambda_k)I_3$ mit $\lambda_k \le 0$ wegen (4.16), (4.27) bzw. (4.40).
Mit

$$P_k = \begin{pmatrix} -\frac{1}{2} - \frac{1}{2}\sqrt{3}i & 1 & 0 \\ -\frac{1}{2} + \frac{1}{2}\sqrt{3}i & 1 & 0 \\ 0 & \frac{1}{\lambda_k - 1} & -\frac{1}{\lambda_k - 1} \end{pmatrix} \quad \text{und} \quad Q_k = \begin{pmatrix} \frac{1}{2} - \frac{\sqrt{3}}{6}i & \frac{1}{2} + \frac{\sqrt{3}}{6}i & 0 \\ -\frac{\sqrt{3}}{3}i & \frac{\sqrt{3}}{3}i & 0 \\ -\frac{\sqrt{3}}{3}i & \frac{\sqrt{3}}{3}i & 1 \end{pmatrix}$$

erhält man die Weierstraß-Kronecker-Zerlegung

$$P_k A Q_k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad P_k D_k Q_k = \begin{pmatrix} \kappa_{k,1} & 0 & 0 \\ 0 & \kappa_{k,2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

mit

$$\kappa_{k,1} = \frac{1}{2}(\lambda_k - 1)(1 + \sqrt{3}i), \quad \kappa_{k,2} = \frac{1}{2}(\lambda_k - 1)(1 - \sqrt{3}i).$$

Das PDA-System hat also den differentiellen Zeitindex 1. Da die Elemente von $P_k$ für alle $\lambda_k \le 0$ beschränkt sind und $Q_k$ und damit auch $Q_k^{-1}$ unabhängig von $k$ sind, ist die Voraussetzung (a) des Satzes 6.13 erfüllt, und auch die Matrixnormen in (5.72) sind beschränkt.
Wegen $\mathfrak{Re}\kappa_{1/2} \le -\frac{1}{2}$ ist auch die Voraussetzung (b) erfüllt.
Für $0 < y$, $0 < \beta \le 1$ und

$$g(\beta) := 1 + y^\beta - (1+y)^\beta$$

gilt wegen $g(1) = 0$ und

$$g'(\beta) = y^\beta \ln y - (1+y)^\beta \ln(1+y) \le 0$$

auch

$$g(\beta) \ge 0$$

und damit

$$(1-\lambda_k)^\beta \le 1 + (-\lambda_k)^\beta \quad \text{für} \quad \lambda_k \le 0, \quad 0 < \beta \le 1.$$

Aus den Lemmata 5.32 und 5.33 folgt, daß beschränkte Matrizen $W_{4(-1)}(A, \tau D_k)$ und $W_{40}(A, \tau D_k)$ existieren. Da $D_k$ regulär ist und nach (5.27)

$$W_{4\alpha}(A, \tau D_k) = W_{4(-1)}(A, \tau D_k)(\tau D_k)^{-1-\alpha} = W_{40}(A, \tau D_k)(\tau D_k)^{-\alpha}$$

gilt, folgt

$$\|W_{4\alpha}(A, \tau D_k)\| \le \min\left\{ \|W_{4(-1)}(A, \tau D_k)\| \frac{1}{(\tau(1-\lambda_k))^{1+\alpha}}, \|W_{40}(A, \tau D_k)\| (\tau(1-\lambda_k))^{-\alpha} \right\},$$

und damit sind die Matrizen $W_{4\alpha}(A, \tau D_k)$ für $\alpha \in \{-\frac{3}{4} + \varepsilon, -\frac{1}{4} + \varepsilon\}$ beschränkt, und auch die Voraussetzung (c) des Satzes 6.13 ist für dieses Beispiel erfüllt.
Damit wird folgende Konvergenzordnung $p^\star$ bezüglich der Zeit vorhergesagt:

$$p^\star = \begin{cases} 5 & : \text{die Randbedingungen sind homogen im Sinne von (5.57) oder periodisch} \\ 4.75 - \varepsilon & : \text{inhomogene Neumann-Randbedingungen} \\ 4.25 - \varepsilon & : \text{inhomogene Dirichlet-Randbedingungen} \end{cases}$$

mit $\varepsilon > 0$ beliebig klein. Dies wird auch numerisch bestätigt:

(1) Wählt man die rechte Seite so, daß

$$u(t,x) = x^2 (e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top$$

die exakte Lösung ist, so liegen inhomogene Dirichlet-Randwerte

$$u(t,-1) = u(t,1) = (e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top$$

vor, und die numerisch bestimmte Zeitordnung beträgt 4.25:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 4.24 | 4.25 | 4.25 |
| $2^4$ | 4.24 | 4.25 | 4.25 |
| $2^5$ | 4.24 | 4.24 | 4.25. |

(2) Wird die rechte Seite dagegen so gewählt, daß

$$u(t,x) = (x^2-1)(e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top$$

exakte Lösung mit homogenen Dirichlet-Randwerten ist, so erhält man eine numerisch bestimmte Konvergenzordnung bezüglich der Zeit von 5:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 4.97 | 4.99 | 4.99 |
| $2^4$ | 4.97 | 4.99 | 4.99 |
| $2^5$ | 4.97 | 4.98 | 5.00. |

(3) Die rechte Seite wird nun so gewählt, daß

$$u(t,x) = (x^2-1)(e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top + t^3(1,1,1)^\top$$

exakte Lösung ist. Die Dirichlet-Randwerte $u(t,-1) = u(t,1) = t^3(1,1,1)^\top$ sind nun inhomogen, aber die Zeitableitungen vierter Ordnung (das dreistufige Radau-IIA-Verfahren hat die Stufenordnung 3) von $u(t,\pm 1)$ verschwinden. Wie erwartet ist die numerisch bestimmte Konvergenzordnung bezüglich der Zeit deshalb erneut gleich 5:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 4.97 | 4.99 | 5.05 |
| $2^4$ | 4.97 | 4.99 | 5.05 |
| $2^5$ | 4.97 | 4.99 | 5.05. |

(4) Wählt man dagegen

$$u(t,x) = (x^2-1)(e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top + t^4(1,1,1)^\top$$

als exakte Lösung, so verschwinden auch die Zeitableitungen vierter Ordnung von $u(t,\pm 1)$ nicht mehr, und man kann eine Ordnungsreduktion beobachten:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 4.26 | 4.26 | 4.26 |
| $2^4$ | 4.26 | 4.26 | 4.26 |
| $2^5$ | 4.26 | 4.26 | 4.25. |

(5) Wählt man die Lösung wie in (1), aber gibt nun (inhomogene) Neumann-Randbedingungen $u_x(t,\pm 1) = \pm 2(e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top$ vor, so erhält man eine numerisch bestimmte Zeitordnung von 4.75:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 4.73 | 4.74 | 4.74 |
| $2^4$ | 4.73 | 4.74 | 4.74 |
| $2^5$ | 4.73 | 4.74 | 4.74. |

(6) Wählt man

$$u(t,x) = x(\frac{x^2}{3} - 1)(e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top + xt^3(1,1,1)^\top$$

als exakte Lösung, so verschwinden die vierten Zeitableitungen der Neumann-Randbedingungen $u_x(t,\pm 1) = t^3(1,1,1)^\top$, und man erhält numerisch wieder eine Konvergenzordnung bezüglich der Zeit von 5:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^3$ | 4.98 | 4.99 | 5.04 |
| $2^4$ | 4.98 | 4.99 | 5.04 |
| $2^5$ | 4.98 | 4.99 | 5.04. |

(7) Wählt man schließlich die rechte Seite und die Anfangswerte so, daß

$$u(t,x) = \sin(\pi x)(e^{-t}, e^{-\frac{1}{2}t}, \sin t)^\top$$

exakte Lösung mit periodischen Randbedingungen ist, so erhält man numerisch ebenfalls die maximal mögliche Zeitordnung 5:

| $0.1\tau^{-1}$ | $2^2$ | $2^3$ | $2^4$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^0$ | 4.88 | 4.94 | 4.97 |
| $2^1$ | 4.89 | 4.95 | 4.97 |
| $2^2$ | 4.91 | 4.96 | 4.98. |

□

Die folgenden zwei Beispiele sind von höherem Zeitindex:

**Beispiel 6.15** Betrachtet wird das eine supraleitende Magnetspule beschreibende lineare PDA-System (3.1) aus Beispiel 3.6 mit den in Beispiel 3.9 bestimmten Anfangs- und Randwerten. Es gelten

$$D_k = \begin{pmatrix} \lambda_k & -1 & 0 & 0 \\ 0 & \lambda_k & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

und

$$P_k A Q_k = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad P_k D_k Q_k = \begin{pmatrix} \kappa_{k1} & 0 & 0 & 0 \\ 0 & \kappa_{k2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

mit $\kappa_{k1} = -\frac{i\lambda_k}{\sqrt{1-\lambda_k}}$, $\kappa_{k2} = \frac{i\lambda_k}{\sqrt{1-\lambda_k}}$ und

$$P_k = \begin{pmatrix} -\frac{i}{\sqrt{1-\lambda_k}} & -\frac{i\sqrt{1-\lambda_k}}{\lambda_k} & 1 & -1 \\ \frac{i}{\sqrt{1-\lambda_k}} & \frac{i\sqrt{1-\lambda_k}}{\lambda_k} & 1 & -1 \\ 0 & 0 & 1 & -\frac{1}{\lambda_k} \\ \frac{1}{\lambda_k} & 0 & 0 & 0 \end{pmatrix}, \; Q_k = \begin{pmatrix} \frac{1}{2(1-\lambda_k)} & \frac{1}{2(1-\lambda_k)} & 0 & -\frac{\lambda_k}{1-\lambda_k} \\ \frac{\lambda_k}{2(1-\lambda_k)} & \frac{\lambda_k}{2(1-\lambda_k)} & 0 & -\frac{\lambda_k}{1-\lambda_k} \\ -\frac{i\lambda_k}{2(1-\lambda_k)^{\frac{3}{2}}} & \frac{i\lambda_k}{2(1-\lambda_k)^{\frac{3}{2}}} & -\frac{\lambda_k}{1-\lambda_k} & 0 \\ -\frac{i\lambda_k^2}{2(1-\lambda_k)^{\frac{3}{2}}} & \frac{i\lambda_k^2}{2(1-\lambda_k)^{\frac{3}{2}}} & -\frac{\lambda_k}{1-\lambda_k} & 0 \end{pmatrix}.$$

Das PDA-System hat folglich den differentiellen Zeitindex 2. Analog zu (6.1) folgt $\lambda_k \le -\frac{4}{l^2}$ und damit $\mathfrak{Re}\kappa_{k1} = \mathfrak{Re}\kappa_{k2} = 0$, Voraussetzung (b) von Satz 6.13 ist erfüllt. Da

$$Q_k \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} P_k B = \frac{1}{2} \, \frac{\lambda_k + i\sqrt{1-\lambda_k}}{\lambda_k - 1 + i\lambda_k\sqrt{1-\lambda_k}} \begin{pmatrix} \frac{1}{1-\lambda_k} & \frac{1}{\lambda_k} & 0 & 0 \\ \frac{\lambda_k}{1-\lambda_k} & 1 & 0 & 0 \\ \frac{i\lambda_k}{(1-\lambda_k)^{\frac{3}{2}}} & \frac{i}{\sqrt{1-\lambda_k}} & 0 & 0 \\ \frac{i\lambda_k^2}{(1-\lambda_k)^{\frac{3}{2}}} & \frac{i\lambda_k}{\sqrt{1-\lambda_k}} & 0 & 0 \end{pmatrix},$$

$$Q_k \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} P_k B = \frac{1}{2} \frac{\lambda_k - i\sqrt{1-\lambda_k}}{\lambda_k - 1 - i\lambda_k\sqrt{1-\lambda_k}} \begin{pmatrix} \frac{1}{1-\lambda_k} & \frac{1}{\lambda_k} & 0 & 0 \\ \frac{\lambda_k}{1-\lambda_k} & 1 & 0 & 0 \\ -\frac{i\lambda_k}{(1-\lambda_k)^{\frac{3}{2}}} & -\frac{i}{\sqrt{1-\lambda_k}} & 0 & 0 \\ -\frac{i\lambda_k^2}{(1-\lambda_k)^{\frac{3}{2}}} & -\frac{i\lambda_k}{\sqrt{1-\lambda_k}} & 0 & 0 \end{pmatrix},$$

$$Q_k \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & 0 & a \end{pmatrix} P_k B = \begin{pmatrix} \frac{a}{1-\lambda_k} & 0 & 0 & 0 \\ \frac{a}{1-\lambda_k} & 0 & 0 & 0 \\ \frac{b}{1-\lambda_k} & 0 & 0 & 0 \\ \frac{b}{1-\lambda_k} & 0 & 0 & 0 \end{pmatrix}$$

und die Elemente von $\frac{1}{\kappa_{k1/2}} Q_k$ und

$$Q_k^{-1} = \begin{pmatrix} 1 & -1 & -\frac{i\sqrt{1-\lambda_k}}{\lambda_k} & \frac{i\sqrt{1-\lambda_k}}{\lambda_k} \\ 1 & -1 & \frac{i\sqrt{1-\lambda_k}}{\lambda_k} & -\frac{i\sqrt{1-\lambda_k}}{\lambda_k} \\ 0 & 0 & 1 & -\frac{1}{\lambda_k} \\ 1 & -\frac{1}{\lambda_k} & 0 & 0 \end{pmatrix}$$

sowie

$$Q_k \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} Q_k^{-1} = \begin{pmatrix} \frac{1}{1-\lambda_k} & -\frac{1}{1-\lambda_k} & 0 & 0 \\ \frac{\lambda_k}{1-\lambda_k} & -\frac{\lambda_k}{1-\lambda_k} & 0 & 0 \\ 0 & 0 & \frac{1}{1-\lambda_k} & -\frac{1}{1-\lambda_k} \\ 0 & 0 & \frac{\lambda_k}{1-\lambda_k} & -\frac{\lambda_k}{1-\lambda_k} \end{pmatrix}$$

und

$$Q_k \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & 0 & a \end{pmatrix} Q_k^{-1} = \begin{pmatrix} -\frac{a\lambda_k}{1-\lambda_k} & \frac{a}{1-\lambda_k} & 0 & 0 \\ -\frac{a\lambda_k}{1-\lambda_k} & \frac{a}{1-\lambda_k} & 0 & 0 \\ -\frac{b\lambda_k}{1-\lambda_k} & \frac{b}{1-\lambda_k} & -\frac{a\lambda_k}{1-\lambda_k} & \frac{a}{1-\lambda_k} \\ -\frac{b\lambda_k}{1-\lambda_k} & \frac{b}{1-\lambda_k} & -\frac{a\lambda_k}{1-\lambda_k} & \frac{a}{1-\lambda_k} \end{pmatrix}$$

alle beschränkt sind, ist nach Bemerkung 5.30 auch Voraussetzung (a) erfüllt, und es wird die Konvergenz des Verfahrens mit der Ordnung 3 bezüglich der Zeit vorausgesagt. Dies wird auch numerisch bestätigt:

| $0.1\tau^{-1}$ | $2^4$ | $2^5$ | $2^6$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^1$ | 3.00 | 3.00 | 3.00 |
| $2^2$ | 3.00 | 3.00 | 3.00 |
| $2^3$ | 3.00 | 3.00 | 3.00. |

□

**Beispiel 6.16** Betrachtet werde das System aus Beispiel 6.9 mit differentiellem Zeitindex 3. Wie dort gezeigt wurde, sind die Voraussetzungen (a) und (b) des Satzes 6.13 erfüllt, der damit eine Konvergenzordnung bezüglich der Zeit von 2 vorhersagt, was auch numerisch bestätigt wird:

| $0.1\tau^{-1}$ | $2^4$ | $2^5$ | $2^6$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^1$ | 1.99 | 1.99 | 1.99 |
| $2^2$ | 1.98 | 1.99 | 1.99 |
| $2^3$ | 1.98 | 1.99 | 1.99. |

Die Konvergenzordnung bezüglich des Ortes beträgt wie erwartet ebenfalls 2:

| $0.1\tau^{-1}$ | $2^4$ | $2^5$ | $2^6$ |
|---|---|---|---|
| $0.1h^{-1}$ | | | |
| $2^1$ | 1.95 | 1.95 | 1.95 |
| $2^2$ | 1.97 | 1.97 | 1.97 |
| $2^3$ | 1.99 | 1.99 | 1.99. |

□

# Zusammenfassung und weiterführende Bemerkungen

In der vorliegenden Arbeit wurde eine Klasse linearer PDA-Systeme mit konstanten Koeffizienten mittels der Linienmethode numerisch gelöst, indem auf das mittels finiter Differenzen bezüglich des Ortes diskretisierte System Runge-Kutta-Verfahren angewendet wurden. Es wurden Konvergenzresultate in Abhängigkeit von den Randwerten und dem differentiellen Zeitindex angegeben. Das von den DA-Systemen bekannte Indexkonzept hat sich damit auch hier bewährt. In der diskreten $L_2$-Norm ist die Konvergenzordnung bezüglich der Zeit bei Dirichlet- und Neumann-Randbedingungen im allgemeinen nichtganzzahlig und geringer als die für DA-Systeme vom gleichen Index erwartete Ordnung sowie bei Dirichlet- geringer als bei Neumann-Randbedingungen. Die Konvergenzordnung ist bei homogenen Randwerten um bis zu eins höher als bei inhomogenen Randwerten, wobei bemerkenswert ist, daß es nicht auf die Homogenität der Randwerte selbst, sondern auf diejenige ihrer $(q+1)$-ten Zeitableitungen ankommt, wobei $q$ die Stufenordnung des betrachteten Runge-Kutta-Verfahrens ist.
Aufbauend auf diesen Resultaten konnten auch für semilineare PDA-Systeme mit konstanten Koeffizienten Konvergenzsätze hergeleitet werden.
Die theoretischen Untersuchungen dieser Arbeit bilden die Grundlage für die Untersuchung weiterer Verfahrensklassen bezüglich der Zeitintegration, wie z. B. der linear-impliziten Runge-Kutta-Verfahren und der BDF-Methoden, die sich bei der numerischen Behandlung steifer Differentialgleichungen und differentiell-algebraischer Systeme ebenfalls als besonders effektiv erwiesen haben.

# Anhang A

# Bestimmung der Eigenwerte und Eigenvektoren spezieller Matrizen

**A1** Zunächst sollen die Eigenwerte $\lambda_j \in \mathbb{C}$ und Eigenvektoren $v_j = (v_{j1}, \ldots, v_{jN})^\top$, $j = 1, \ldots, N$, der tridiagonalen Matrix

$$T_1 = \begin{pmatrix} b & c & & \\ a & b & c & \\ & \ddots & \ddots & \\ & & a & b \end{pmatrix} \in \mathbb{R}^{N,N} \tag{A.1}$$

mit $a, b, c \in \mathbb{R}$, $a, c \neq 0$ bestimmt werden. Aus $T_1 v_j = \lambda_j v_j$ erhält man die Differenzengleichung

$$a v_{j(k-1)} + (b - \lambda_j) v_{jk} + c v_{j(k+1)} = 0, \quad k = 1, \ldots, N, \tag{A.2}$$

mit

$$v_{j0} = 0, \tag{A.3a}$$
$$v_{j(N+1)} = 0, \tag{A.3b}$$

für die nichttriviale Lösungen gesucht werden.
Hat die zugehörige charakteristische Gleichung

$$a + (b - \lambda_j) x_j + c x_j^2 = 0 \tag{A.4}$$

für $j = j'$ eine doppelte Nullstelle, so folgt für die allgemeine Lösung von (A.2) nach Samarskij [45]

$$v_{j'k} = \left(c_{1j'} + k c_{2j'}\right) x_{j'}^k, \tag{A.5}$$

mit Konstanten $c_{1j'}$ und $c_{2j'}$. Aus (A.3a) folgt dann $c_{1j'} = 0$ und damit aus (A.3b) $v_{j'k} = 0$, man erhält nur die triviale Lösung.
Im folgenden wird deshalb vorausgesetzt, daß die charakteristische Gleichung (A.4) zwei verschiedene Lösungen $x_{1j}$ und $x_{2j}$ hat, d. h.

$$(b - \lambda_j)^2 \neq 4ac. \tag{A.6}$$

Dann ergibt sich die allgemeine Lösung der Differenzengleichung (A.2) nach [45] zu

$$v_{jk} = c_{1j} x_{1j}^k + c_{2j} x_{2j}^k, \tag{A.7}$$

wobei $c_{1j}$ und $c_{2j}$ zunächst beliebige Konstanten sind.
Aus (A.3a) folgt

$$c_{1j} + c_{2,j} = 0 \tag{A.8}$$

und damit aus (A.3b)

$$x_{1j}^{N+1} = x_{2j}^{N+1}. \tag{A.9}$$

Nach dem Satz von Vieta folgt aus der charakteristischen Gleichung (A.4)

$$x_{1j}x_{2j} = \frac{a}{c} \tag{A.10}$$

und damit aus (A.9)

$$x_{1j}^{2(N+1)} = \left(\frac{a}{c}\right)^{N+1} = x_{2j}^{2(N+1)}. \tag{A.11}$$

Als Lösung von (A.10) und (A.11) erhält man die $N$ Paare verschiedener Wurzeln

$$x_{1j} = \sqrt{\frac{a}{c}}\left(\cos\frac{j\pi}{N+1} + i\sin\frac{j\pi}{N+1}\right), \tag{A.12a}$$

$$x_{2j} = \sqrt{\frac{a}{c}}\left(\cos\frac{j\pi}{N+1} - i\sin\frac{j\pi}{N+1}\right) \tag{A.12b}$$

mit $j = 1, \ldots, N$ (für $j = N+1$ wäre $x_{1j} = x_{2j}$). Für die Eigenvektoren $v_j$ von $T_1$ erhält man mit (A.7) und (A.8) damit

$$v_{jk} = c_j\sqrt{\frac{a}{c}}^{k}\sin\frac{jk\pi}{N+1}, \quad k = 1, \ldots, N, \quad c_j \in \mathbb{C}\setminus\{0\}, \tag{A.13}$$

es gilt wegen

$$\sin\frac{(k\pm1)j\pi}{N+1} = \sin\frac{jk\pi}{N+1}\cos\frac{j\pi}{N+1} \pm \cos\frac{jk\pi}{N+1}\sin\frac{j\pi}{N+1} \tag{A.14}$$

$$\begin{aligned} av_{j(k-1)} + bv_{jk} + cv_{j(k+1)} &= \sqrt{ac}\cos\frac{j\pi}{N+1}v_{jk} - c_j a\sqrt{\frac{a}{c}}^{k-1}\sin\frac{j\pi}{N+1}\cos\frac{jk\pi}{N+1} + bv_{jk} \\ &\quad + \sqrt{ac}\cos\frac{j\pi}{N+1}v_{jk} + c_j c\sqrt{\frac{a}{c}}^{k+1}\sin\frac{j\pi}{N+1}\cos\frac{jk\pi}{N+1} \\ &= \left(b + 2\sqrt{ac}\cos\frac{j\pi}{N+1}\right)v_{jk}, \quad k = 1, \ldots, N. \end{aligned}$$

Für die Eigenwerte $\lambda_j$ von $T_1$ folgt daraus

$$\lambda_j = b + 2\sqrt{ac}\cos\frac{j\pi}{N+1}, \quad j = 1, \ldots, N, \tag{A.15}$$

sie erfüllen Bedingung (A.6) und sind (wegen $ac \neq 0$) sämtlich voneinander verschieden. Dieses Ergebnis findet man auch in Thomas [55].

**A2** Nun sollen Eigenwerte und Eigenvektoren der tridiagonalen Matrix

$$T_2 = \begin{pmatrix} a+b & c & & \\ a & b & c & \\ & \ddots & \ddots & \\ & & a & b+c \end{pmatrix} \in \mathbb{R}^{N,N}$$

bestimmt werden. Es gelten wieder die Differenzengleichung (A.2), die charakteristische Gleichung (A.4) und die Lösungsdarstellungen (A.5) und (A.7). Anstelle von (A.3) hat man

$$v_{j0} = v_{j1} \tag{A.16a}$$

$$v_{jN} = v_{j(N+1)} \tag{A.16b}$$

In dem Fall, daß die charakteristische Gleichung für $j = j'$ eine doppelte Nullstelle hat, folgen aus (A.5) und (A.16)

$$\begin{aligned} c_{1j'} &= (c_{1j'} + c_{2j'})x_{j'}, \\ c_{1j'} + Nc_{2j'} &= (c_{1j'} + Nc_{2j'} + c_{2j'})x_{j'} \end{aligned}$$

und daraus nacheinander $Nc_{2j'} = Nc_{2j'}x_{j'}$, $c_{2j'} = 0$, $x_{1j'} = 1$ (da $c_{1j'}$ und $c_{2j'}$ nicht gleichzeitig verschwinden dürfen) und damit $v_{j'k} = c_{j'} \neq 0$. Aus der Differenzengleichung (A.2) erhält man $\lambda_{j'} = a + b + c$.
Hat die charakteristische Gleichung zwei verschiedene Nullstellen, so folgen aus (A.7) und (A.16)

$$\begin{aligned} c_{1j}(1 - x_{1j}) + c_{2j}(1 - x_{2j}) &= 0, \\ c_{1j}(x_{1j}^N - x_{1j}^{N+1}) + c_{2j}(x_{2j}^N - x_{2j}^{N+1}) &= 0. \end{aligned} \tag{A.17}$$

Damit dieses System für $c_{1j},\ c_{2j}$ eine nichttriviale Lösung hat, muß

$$(1 - x_{1j})(x_{2j}^N - x_{2j}^{N+1}) - (1 - x_{2j})(x_{1j}^N - x_{1j}^{N+1}) = (1 - x_{1j})(1 - x_{2j})(x_{2j}^N - x_{1j}^N) = 0$$

gelten. Sind $x_{1j} \neq 1$ und $x_{2j} \neq 1$, so folgt $x_{2j}^N = x_{1j}^N$ und daraus anstelle von (A.12)

$$\begin{aligned} x_{1j} &= \sqrt{\frac{a}{c}} \left( \cos \frac{j\pi}{N} + i \sin \frac{j\pi}{N} \right), \\ x_{2j} &= \sqrt{\frac{a}{c}} \left( \cos \frac{j\pi}{N} - i \sin \frac{j\pi}{N} \right) \end{aligned}$$

für $j = 1, \ldots, N-1$ (für $j = N$ wäre $x_{1j} = x_{2j}$).
Mit (A.7) und (A.17) erhält man für die Eigenvektoren

$$\begin{aligned} v_{jk} &= \frac{c_{1j}}{1 - x_{2j}} \left( (1 - x_{2j})x_{1j}^k - (1 - x_{1j})x_{2j}^k \right) \\ &= \frac{c_{1j}}{1 - x_{2j}} \left( x_{1j}^k - x_{2j}^k - x_{1j}x_{2j} \left( x_{1j}^{k-1} - x_{2j}^{k-1} \right) \right) \end{aligned}$$

und damit

$$v_{jk} = c'_j \sqrt{\frac{a}{c}}^k \left( \sin \frac{jk\pi}{N} - \sqrt{\frac{a}{c}} \sin \frac{j(k-1)\pi}{N} \right), \quad j = 1, \ldots, N-1, \quad k = 1, \ldots, N. \tag{A.18a}$$

Ist (z. B. für $j = N$) $x_{1N} = 1$ oder $x_{2N} = 1$, so folgt aus (A.17) und (A.7)

$$v_{Nk} = c_N \tag{A.18b}$$

und damit wie oben $\lambda_N = a + b + c$.
Mit dem Additionstheorem (A.14) folgt

$$v_{j(k\pm1)} = \left( \sqrt{\frac{a}{c}} \right)^{\pm1} \cos \frac{j\pi}{N} v_{jk} \pm c'_j \left( \sqrt{\frac{a}{c}} \right)^{k\pm1} \sin \frac{j\pi}{N} \left( \cos \frac{jk\pi}{N} - \sqrt{\frac{a}{c}} \cos \frac{j(k-1)\pi}{N} \right),$$

und daraus erhält man

$$\begin{aligned} (a + b)v_{j1} + cv_{j2} &= (a + b)v_{j1} + \sqrt{ac} \cos \frac{j\pi}{N} v_{j1} + ac'_j \sin \frac{j\pi}{N} \left( \cos \frac{j\pi}{N} - \sqrt{\frac{a}{c}} \right) \\ &= \left( b + 2\sqrt{ac} \cos \frac{j\pi}{N} \right) v_{j1}, \end{aligned}$$

$$
\begin{aligned}
& av_{j(k-1)} + bv_{jk} + cv_{j(k+1)} \\
= & \sqrt{ac}\cos\frac{j\pi}{N}v_{jk} - c'_j\sqrt{ac}\sqrt{\frac{a}{c}}^k \sin\frac{j\pi}{N}\left(\cos\frac{jk\pi}{N} - \sqrt{\frac{a}{c}}\cos\frac{j(k-1)\pi}{N}\right) \\
& + bv_{jk} + \sqrt{ac}\cos\frac{j\pi}{N}v_{jk} + c'_j\sqrt{ac}\sqrt{\frac{a}{c}}^k \sin\frac{j\pi}{N}\left(\cos\frac{jk\pi}{N} - \sqrt{\frac{a}{c}}\cos\frac{j(k-1)\pi}{N}\right) \\
= & \left(b + 2\sqrt{ac}\cos\frac{j\pi}{N}\right)v_{jk},
\end{aligned}
$$

$$
\begin{aligned}
& av_{j(N-1)} + (b+c)v_{jN} \\
= & \sqrt{ac}\cos\frac{j\pi}{N}v_{jN} - c'_j\sqrt{ac}\sqrt{\frac{a}{c}}^N \sin\frac{j\pi}{N}\left(\cos j\pi - \sqrt{\frac{a}{c}}\cos\frac{j(N-1)\pi}{N}\right) \\
& + (b+c)v_{jN} \\
= & \left(\sqrt{ac}\cos\frac{j\pi}{N} + b + c\right)v_{jN} - c'_j\sqrt{\frac{a}{c}}^{N+1} \sin\frac{j\pi}{N}\cos j\pi\left(c - \sqrt{ac}\cos\frac{j\pi}{N}\right) \\
= & \left(b + 2\sqrt{ac}\cos\frac{j\pi}{N}\right)v_{jN}.
\end{aligned}
$$

Für die Eigenwerte $\lambda_j$ von $T_2$ gilt also

$$
\begin{aligned}
\lambda_j &= b + 2\sqrt{ac}\cos\frac{j\pi}{N}, \quad j = 1, \ldots, N-1, \\
\lambda_N &= a + b + c.
\end{aligned}
\tag{A.19}
$$

**A3** Für periodische Toeplitz-Matrizen erhält man durch analoges Vorgehen folgendes Lemma:

**Lemma A.1** Die $N \times N$-Matrix

$$
\begin{pmatrix}
b & c & & a \\
a & b & c & \\
& \ddots & \ddots & \\
c & & a & b
\end{pmatrix}
$$

hat die Eigenwerte

$$
\lambda_j = b + (a+c)\cos\frac{2j\pi}{N} + i(c-a)\sin\frac{2j\pi}{N}, \quad j = 1, \ldots, N,
$$

zu den orthonormierten Eigenvektoren

$$
v_j = (v_{j1}, \ldots, v_{jN})^\top
$$

mit

$$
v_{jk} = \frac{1}{\sqrt{N}}e^{\frac{2jk\pi}{N}i}.
$$

□

**Beweis:** Mit

$$
v_{j0} = \frac{1}{\sqrt{N}} = v_{jN}
$$

und

$$
v_{j(N-1)} = \frac{1}{\sqrt{N}}e^{\frac{2j(N-1)\pi}{N}i} = \frac{1}{\sqrt{N}}e^{-\frac{2j\pi}{N}i} = v_{j(-1)}
$$

erhält man

$$\begin{aligned} av_{j(m-1)} + bv_{jm} + cv_{j(m+1)} &= \left(ae^{-\frac{2j\pi}{N}i} + b + ce^{\frac{2j\pi}{N}i}\right) \frac{1}{\sqrt{N}} e^{\frac{2jm\pi}{N}i} \\ &= \left(b + (a+c)\cos\frac{2j\pi}{N} + i(c-a)\sin\frac{2j\pi}{N}\right) v_{jm}. \end{aligned}$$

Da für $1 \leq j, j' \leq N$

$$\bar{v}_j^\top v_{j'} = \frac{1}{N}\sum_{k=1}^{N} e^{-\frac{2jk\pi}{N}i} e^{\frac{2j'k\pi}{N}i} = \frac{1}{N}\sum_{k=0}^{N-1} e^{-\frac{2(j-j')k\pi}{N}i} = \begin{cases} \frac{1}{N}\sum\limits_{k=0}^{N-1} 1 & = 1: \quad j = j' \\ \frac{1}{N}\frac{1-e^{-2(j-j')\pi i}}{1-e^{-2i(j-j')\pi/N}} & = 0: \quad j \neq j' \end{cases}$$

gilt, sind die Eigenvektoren orthonormiert. □

# Anhang B

# Konvergenz für PDA-Systeme mit variablen Koeffizienten

In Verallgemeinerung der Wärmeleitungsgleichung mit variablen Koeffizienten,

$$u_t = (p(x)u_x)_x - q(x)u, \quad p(x) > 0,$$

können als zweite Aufgabenklasse Systeme betrachtet werden, bei denen (3.2a) ersetzt wird durch

$$A\, u_t(t,\vec{x}) + \sum_{i=1}^{d} B_i\, a_i(x_i) \left( \left(p_i(x_i)u_{x_i}(t,\vec{x})\right)_{x_i} + q_i(x_i)u(t,\vec{x}) \right) + C\, u(t,\vec{x}) = f(t,\vec{x},u) \tag{B.1a}$$

mit

$$a_i(x_i) > 0, \quad p_i(x_i) \geq 0, \quad x_i \in [-l_i, l_i], \quad i = 1,\ldots,d. \tag{B.2}$$

Im Fall periodischer Randbedingungen (3.2c) wird zusätzlich die Periodizität von $p_i$ vorausgesetzt,

$$p_i(x) = p_i(x + 2l_i), \quad x \in \mathbb{R}, \quad i \in M_p. \tag{B.3}$$

Setzt man

$$a_i(x) = e^{-r_i x}, \quad p_i(x) = e^{r_i x}, \quad q_i(x) = 0, \tag{B.4}$$

so folgt

$$a_i(x_i) \left( \left(p_i(x_i)u_{x_i}(t,\vec{x})\right)_{x_i} + q_i(x_i)u(t,\vec{x}) \right) = u_{x_i x_i} + r_i u_{x_i}.$$

Das System (3.2) ist deshalb ein Spezialfall von (B.1), die Diskretisierung wird jedoch direkt, d. h. ohne Berücksichtigung der Transformation (B.4), vorgenommen (vgl. Bemerkung (B.1)). Die Beispiele 3.1, 3.3 und 3.4 mit ortsabhängigen Diffusionskonstanten $D_i(x_i)$ bzw. Temperaturleitfähigkeiten $k_i(x_i)$ sind Beispiele, die vom Typ (B.1), aber nicht vom Typ (3.2) sind. Probleme in rotationssymmetrischen Zylinder- bzw. Kugelkoordinaten können ebenfalls auf Systeme vom Typ (B.1) führen, weil dann der Laplace-Operator die Gestalt $\Delta u = \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial u}{\partial\rho}\right)$ bzw. $\Delta u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right)$ annimmt.
Zunächst soll wieder der Fall $d = 1$ betrachtet werden.

## B.1 Räumlich eindimensionales PDA-System

### B.1.1 Dirichlet-Randbedingungen

Auf dem äquidistanten Ortsgitter $\Omega_h$ aus (4.3) ergibt sich mit den Approximationen

$$a(x_k)\left(p(x_k)u_x(t,xk)\right)_x \approx \frac{1}{h^2}a(x_k)\Bigg(p\left(x_k+\frac{h}{2}\right)u(t,x_{k+1}) \\ -\left(p\left(x_k+\frac{h}{2}\right)+p\left(x_k-\frac{h}{2}\right)\right)u(t,x_k)+p\left(x_k-\frac{h}{2}\right)u(t,x_{k-1})\Bigg)$$

aus (B.1) die semidiskrete Gleichung

$$A\,\dot{u}_k(t)+\frac{1}{h^2}B\,\Big(a_k p_{k+\frac{1}{2}}u_{k+1}(t)\;-a_k\left(p_{k+\frac{1}{2}}+p_{k-\frac{1}{2}}-h^2q_k\right)u_k(t) \\ +a_k p_{k-\frac{1}{2}}u_{k-1}(t)\Big)+C\,u_k(t)\;=\;f_k(t,u_k),$$

wobei $a_k = a(x_k)$, $p_{k\pm\frac{1}{2}} = p(x_k \pm \frac{h}{2})$ und $q_k = q(x_k)$ gesetzt werden.
Durch Taylor-Entwicklung im Punkt $x_k$ erhält man

$$\frac{1}{h^2}\left[p\left(x_k+\frac{h}{2}\right)\left(u(t,x_{k+1})-u(t,x_k)\right)\;-p\left(x_k-\frac{h}{2}\right)\left(u(t,x_k)-u(t,x_{k-1})\right)\right] \\ = p(x_k)u_{xx}(t,x_k)+p'(x_k)u_x(t,x_k)+h^2\gamma_k(t), \tag{B.5}$$

wobei $\gamma_k(t)$ Ableitungen nach $x$ von $p(x)$ bis zur dritten und von $u(t,x)$ bis zur vierten Ordnung enthält, sind diese Ableitungen auf $[0,t_e]\times\overline{\Omega}$ beschränkt, so existiert wieder $K$ mit (4.11).
Es gilt wieder Gleichung (4.12), wobei aber nun anstelle von (4.13)

$$\omega(t) = \left(\frac{1}{h^2}a_1 p_{\frac{1}{2}}\psi^\top(t,-l),0,\ldots,0,\frac{1}{h^2}a_N p_{N+\frac{1}{2}}\psi^\top(t,l)\right)^\top \tag{B.6}$$

und anstelle von (4.14)

$$P = \begin{pmatrix} -s_1 & a_1p_{\frac{3}{2}} & & \\ a_2p_{\frac{3}{2}} & -s_2 & a_2p_{\frac{5}{2}} & \\ & & \cdots & \\ & & a_Np_{N-\frac{1}{2}} & -s_N \end{pmatrix} \tag{B.7}$$

mit der Abkürzung

$$s_k = a_k\left(p_{k+\frac{1}{2}}+p_{k-\frac{1}{2}}-h^2q_k\right)$$

gelten.
Weil nach (B.2) $a(x)p(x) \geq 0$ ist, folgt aus dem Satz von Gerschgorin (vgl. [52]) für die Eigenwerte von $\frac{1}{h^2}P$

$$\lambda_i \leq h^2 \max_{x\in[-l,l]} a(x)q(x), \quad i=1,\ldots,N. \tag{B.8}$$

Seien $\tilde{P}$ die Matrix $P$ aus (B.7) für $a(x)\equiv 1$ und

$$D_S = \mathrm{diag}\{a(x_1),\ldots,a(x_N)\}.$$

Dann ist $\tilde{P}$ symmetrisch, und es gilt

$$P = D_S\tilde{P}.$$

Da $a(x) > 0$ gilt, existiert $\sqrt{D_S}$ und ist regulär. Daher ist die symmetrische Matrix $\sqrt{D_S}\tilde{P}\sqrt{D_S}$ ähnlich zu $P$, und es existiert eine Orthogonalmatrix $Q_S$ mit

$$Q_S^\top\sqrt{D_S}\frac{1}{h^2}\tilde{P}\sqrt{D_S}Q_S = \mathrm{diag}\{\lambda_1,\ldots,\lambda_N\}.$$

Mit

$$S_P = Q_S^\top \left(\sqrt{D_S}\right)^{-1} \tag{B.9}$$

gilt also

$$S_P \frac{1}{h^2} P S_P^{-1} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_N\}.$$

Für die euklidischen Matrixnormen von $S_P$ und $S_P^{-1}$ erhält man $\|S_P\| = \| \left(\sqrt{D_S}\right)^{-1} \|$ und $\|S_P^{-1}\| = \|\sqrt{D_S}\|$. Daraus folgen

$$\min_{x\in[-l,l]} \frac{1}{\sqrt{a(x)}} \le \|S_P\| \le \max_{x\in[-l,l]} \frac{1}{\sqrt{a(x)}}, \tag{B.10a}$$

$$\min_{x\in[-l,l]} \sqrt{a(x)} \le \|S_P^{-1}\| \le \max_{x\in[-l,l]} \sqrt{a(x)}. \tag{B.10b}$$

**Bemerkung B.1** Setzt man $a(x)$, $p(x)$ und $q(x)$ gemäß (B.4), so erhält man nach (B.7) anstelle von (4.14)

$$P = \begin{pmatrix} -\left(e^{\frac{1}{2}rh} + e^{-\frac{1}{2}rh}\right) & e^{\frac{1}{2}rh} & & \\ e^{-\frac{1}{2}rh} & -\left(e^{\frac{1}{2}rh} + e^{-\frac{1}{2}rh}\right) & e^{\frac{1}{2}rh} & \\ & & \cdots & \\ & & e^{-\frac{1}{2}rh} & -\left(e^{\frac{1}{2}rh} + e^{-\frac{1}{2}rh}\right) \end{pmatrix},$$

und aus (B.10) folgen anstelle von (4.23)

$$e^{-\frac{|r|l}{2}} \le \|S_P\| \le e^{\frac{|r|l}{2}}, \quad e^{-\frac{|r|l}{2}} \le \|S_P^{-1}\| \le e^{\frac{|r|l}{2}}.$$

□

### B.1.2 Periodische Randbedingungen

Diskretisiert man (B.1) mit periodischen Randbedingungen anstelle von Dirichlet-Randbedingungen, so gelten anstelle von (B.6) $\omega(t) \equiv 0$ und damit $F(t,U) \equiv \tilde{F}(t,U)$ und anstelle von (B.7) unter Berücksichtigung von (B.3) ($p_{N+\frac{1}{2}} = p_{\frac{1}{2}}$)

$$P = \begin{pmatrix} -s_1 & a_1 p_{\frac{3}{2}} & & a_1 p_{\frac{1}{2}} \\ a_2 p_{\frac{3}{2}} & -s_2 & a_2 p_{\frac{5}{2}} & \\ & & \cdots & \\ a_N p_{N+\frac{1}{2}} & & a_N p_{N-\frac{1}{2}} & -s_N \end{pmatrix},$$

die Gleichungen (B.8) - (B.10) bleiben gültig mit $h = \frac{2l}{N}$.

### B.1.3 Neumann-Randbedingungen

Auf dem äquidistanten Offsetgitter $\Omega_h$ aus (4.31) gilt für $k = 2, \ldots, N-1$ wieder Gleichung (B.5). Für $k = 1$ folgt für die Approximation (4.34a) der Randbedingung

$$\begin{aligned} &\frac{1}{h^2}\left[Bp\left(x_1 + \frac{h}{2}\right)(u(t,x_2) - u(t,x_1)) \; -p\left(x_1 - \frac{h}{2}\right) h\chi(t,-l)\right] \\ &= \frac{B}{h^2}\left[\left(p(x_1) + \frac{h}{2}p'(x_1) + \frac{h^2}{8}p''(x_1) + \ldots\right)\right. \\ &\quad \left(hu_x(t,x_1) + \frac{h^2}{2}u_{xx}(t,x_1) + \frac{h^3}{6}u_{xxx}(t,x_1) + \ldots\right) \\ &\quad \left. - \left(p(x_1) - \frac{h}{2}p'(x_1) + \frac{h^2}{8}p''(x_1) \mp \ldots\right)\right. \end{aligned}$$

$$\left(hu_x(t,x_1) - \frac{h^2}{2}u_{xx}(t,x_1) + \frac{h^3}{8}u_{xxx}(t,x_1) \mp \dots\right)\Bigg]$$
$$= B\left[p(x_1)u_{xx}(t,x_1) + p'(x_1)u_x(t,x_1) + h\gamma_1(t)\right]$$

und für $k = N$ analog

$$\frac{1}{h^2}\left[p\left(x_N + \frac{h}{2}\right)h\chi(t,l) \; -Bp\left(x_N - \frac{h}{2}\right)(u(t,x_N) - u(t,x_{N-1}))\right]$$
$$= B\left[p(x_k)u_{xx}(t,x_N) + p'(x_k)u_x(t,x_N) + h\gamma_N(t)\right].$$

Anstelle von (B.6) gilt nun

$$\omega(t) = \left(-\frac{a_1 p_{\frac{1}{2}}}{h}\chi^\top(t,-l), 0, \dots, 0, \frac{a_N p_{N+\frac{1}{2}}}{h}\chi^\top(t,l)\right)^\top$$

und anstelle von (B.7)

$$P = \begin{pmatrix} -s_1 + a_1 p_{\frac{1}{2}} & a_1 p_{\frac{3}{2}} & & & \\ a_2 p_{\frac{3}{2}} & -s_2 & a_2 p_{\frac{5}{2}} & & \\ & & \dots & & \\ & & a_N p_{N-\frac{1}{2}} & -s_N + a_N p_{N+\frac{1}{2}} & \end{pmatrix}.$$

Die Gleichungen (B.8) - (B.10) des Abschnitts B.1.1 bleiben mit $h = \frac{2l}{N}$ wieder erhalten.

## B.2 Verallgemeinerung auf räumlich mehrdimensionales PDA-System

Für $d$ Dimensionen erhält man auf dem Ortsgitter (4.44) wieder das DA-System (4.45), wobei $r_{\vec{k}}(t)$ jetzt durch

$$r_{\vec{k}}(t) = \sum_{i\in M_D}\left\{\begin{array}{rl} \frac{a_{i1}p_{i\frac{1}{2}}}{h_i^2}\psi_i(t,\vec{x}_{k_1,\dots,k_{i-1},0,k_{i+1},\dots,k_d}): & k_i = 1 \\ \frac{a_{iN_i}p_{i(N_i+\frac{1}{2})}}{h_i^2}\psi_i(t,\vec{x}_{k_1,\dots,k_{i-1},N_i+1,k_{i+1},\dots,k_d}): & k_i = N_i \\ 0: & \text{sonst}\end{array}\right\}$$
$$+ \sum_{i\in M_N}\left\{\begin{array}{rl} -\frac{a_{i1}p_{i\frac{1}{2}}}{h_i}\chi_i(t,x_{1,k_1},\dots,x_{i-1,k_{i-1}},-l_i,x_{i+1,k_{i+1}},\dots): & k_i = 1 \\ \frac{a_{iN}p_{i(N+\frac{1}{2})}}{h_i}\chi_i(t,x_{1,k_1},\dots,x_{i-1,k_{i-1}},l_i,x_{i+1,k_{i+1}},\dots): & k_i = N_i \\ 0: & \text{sonst}\end{array}\right\}$$

definiert ist, die restlichen Gleichungen des Abschnitts 4.1.2 bleiben mit den entsprechenden Größen erhalten. Für den gemäß (4.49) definierten lokalen Ortsdiskretisierungsfehler folgt wieder (4.50) mit $p_i = 2$ für periodische und Dirichlet-Randbedingungen ($i \in M_P \cup M_D$) und $p_i = 1$ für Neumann-Randbedingungen ($i \in M_N$).
Damit gelten alle Konvergenzsätze aus Abschnitt 4.2 und Kapitel 5 auch für die hier betrachteten PDA-Systeme (B.1). Bei den Untersuchungen zur Beschränktheit von $\|D^{1+\alpha}U_{\vec{h}}^{(l)}(t)\|$ für $\alpha \notin \{-1, 0\}$ wurde jedoch die spezielle Gestalt der Eigenwerte und Eigenvektoren der diskretisierten Ortsdifferentialoperatoren $\frac{1}{h_i^2}P_i$ ausgenutzt, so daß die betreffenden Aussagen nicht direkt übertragen werden können.

# Literaturverzeichnis


[1] H. Altenbach, P. Deuring, K. Naumenko: A system of ordinary and partial differential equations describing creep behaviour of thin-walled shells. Zeitschrift für Analysis und ihre Anwendungen 18 (1999), Nr. 4, S. 1003-1030.

[2] M. Arnold: A perturbation analysis for the dynamical simulation of mechanical multibody systems. Applied Numerical Mathematics 18 (1995), S. 37-56.

[3] K. E. Brenan, S. L. Campbell and L. R. Petzold: Numerical solution of initial-value problems in differential-algebraic equations. North-Holland Publ. Co., Amsterdam 1989.

[4] P. Brenner, M. Crouzeix, V. Thomée: Single step methods for inhomogeneous linear differential equations in Banach space. R.A.I.R.O. Analyse numérique 16 (1982), Nr. 1, S. 5-26.

[5] P. N. Brown, A. C. Hindmarsh, L. R. Petzold: Using Krylov methods in the solution of large-scale differential-algebraic systems. SIAM Journal on Scientific Computing 15 (1994), Nr. 6, S. 1467-1488.

[6] P. N. Brown, A. C. Hindmarsh, L. R. Petzold: Consistent initial condition calculation for differential-algebraic systems. SIAM Journal on Scientific Computing 19 (1998), Nr. 5, S. 1495-1512.

[7] J. C. Butcher: Implicit Runge-Kutta-processes. Mathematics of Computation 18 (1964), Nr. 85, S. 50-64.

[8] M. P. Calvo, C. Palencia: Avoiding the order reduction of Runge-Kutta methods for linear initial boundary value problems. Mathematics of Computation 71 (2001), Nr. 240, S. 1529-1543.

[9] S. L. Campbell: Singular systems of differential equations. Pitman, London 1980.

[10] S. L. Campbell, C. D. Meyer Jr., N. J. Rose: Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients. SIAM Journal on Applied Mathematics 31 (1976), Nr. 3, S. 411-425.

[11] S. L. Campbell, W. Marszalek: ODE/DAE integrators and MOL problems. Zeitschrift für angewandte Mathematik und Mechanik 76 (1996), Suppl. 1, S. 251-254.

[12] S. L. Campbell, W. Marszalek: The index of an infinite dimensional implicit system. Mathematical and Computer Modelling of Dynamical Systems 5 (1999), Nr. 1, S. 18-42.

[13] G. G. Dahlquist: A special stability problem for linear multistep methods. BIT 3 (1963), S. 27-43.

[14] K. Dekker, J. G. Verwer: Stability of Runge-Kutta methods for stiff nonlinear differential equations. CWI Monograph, North-Holland, 1984.

[15] P. Deuflhard, F. Bornemann: Numerische Mathematik II. Gewöhnliche Differentialgleichungen. Walter de Gruyter, Berlin 2002.

[16] C. Eichler-Liebenow: Zur numerischen Behandlung räumlich mehrdimensionaler parabolischer Differentialgleichungen mit linear-impliziten Splitting-Methoden und linearer partieller differentiell-algebraischer Systeme. Dissertation, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale) 1999.

[17] K. Fukuoka, K. Yamaoka, M. Higashimori, T. Nakagawa: Analysis program based on finite element method, MULTI (FEM), for evaluation of dose-dependent local disposition of drug in liver. Journal of Pharmaceutical Sciences 88 (1999), Nr. 5, S. 538-543.

[18] C. W. Gear: Differential-algebraic index transformation. SIAM Journal on Scientific and Statistical Computing 9 (1988), S. 39-47.

[19] C. W. Gear: Differential-algebraic equations, indices, and integral algebraic equations. SIAM Journal on Numerical Analysis 27 (1990), Nr. 6, S. 1527-1534.

[20] G. H. Golub, C. F. van Loan: Matrix computations. Third Edition. The John Hopkins University Press, Baltimore 1996.

[21] E. Griepentrog, R. März: Differential-algebraic equations and their numerical treatment. Teubner Texte zur Mathematik, Band 88, Leipzig 1986.

[22] Ch. Großmann, H. G. Roos: Numerik partieller Differentialgleichungen. B.G. Teubner Verlagsgesellschaft, Stuttgart 1994.

[23] M. Günther, Y. Wagner: Index concepts for linear mixed systems of differential-algebraic and hyperbolic-type equations. SIAM Journal on Scientific Computing 22 (2000), Nr. 5, S. 1610-1629.

[24] E. Hairer, Ch. Lubich, M. Roche: The numerical solution of differential-algebraic systems by Runge-Kutta methods. Springer-Verlag, Berlin 1989.

[25] E. Hairer, S. P. Nørsett, G. Wanner: Solving ordinary differential equations I. Nonstiff problems. Springer-Verlag, Berlin 1993.

[26] E. Hairer, G. Wanner: Solving ordinary differential equations II. Stiff and differential-algebraic problems. Springer-Verlag, Berlin 2002.

[27] M. Hoschek: Einschrittverfahren zur numerischen Simulation elektrischer Schaltungen. VDI Verlag GmbH, Düsseldorf 1999.

[28] P. Knabner: Mathematische Modelle für Transport und Sorption gelöster Stoffe in porösen Medien. Methoden und Verfahren der mathematischen Physik, Band 36. Verlag Peter Lang, Frankfurt(Main) 1991.

[29] P. Kunkel, V. Mehrmann: Canonical forms for linear differential-algebraic equations with variable coefficients. Journal of Computational and Applied Mathematics 56 (1994), S. 225-251.

[30] J. Lang: Adaptive multilevel solution of nonlinear parabolic PDE systems. Theory, algorithm, and applications. Springer-Verlag, Berlin 2001.

[31] B. Leimkuhler, L. R. Petzold, C. W. Gear: Approximation methods for the consistent initialization of differential-algebraic equations. SIAM Journal on Numerical Analysis 28 (1991), Nr. 1, S. 205-226.

[32] P. Lin: A sequential regularization method for time-dependent incompressible Navier-Stokes equations. SIAM Journal on Numerical Analysis 34 (1997), Nr. 3, S. 1051-1071.

[33] Ch. Lubich, A.Ostermann: Runge-Kutta methods for parabolic equations and convolution quadrature. Mathematics of Computation 60 (1993), Nr. 201, S. 105-131.

[34] W. Lucht, K. Debrabant: Models of quasi-linear PDAEs with convection. Report No. 24 (2000), Martin-Luther-Universität Halle-Wittenberg, Fachbereich Mathematik und Informatik, 2000.

[35] W. Lucht, K. Debrabant: On quasi-linear PDAEs with convection: Applications, indices, numerical solution. Applied Numerical Mathematics 42 (2002), S. 297-314.

[36] W. Lucht, K. Strehmel: Discretization based indices for semilinear partial differential algebraic equations. Applied Numerical Mathematics 28 (1998), S. 371-386.

[37] W. Lucht, K. Strehmel and C. Eichler-Liebenow: Indexes and special discretization methods for linear partial differential algebraic equations. BIT numerical mathematics 39 (1999), Nr. 3, S. 484-512.

[38] R. März: EXTRA-ordinary differential equations. Attempts to an analysis of differential-algebraic systems. Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin 1997. - ISSN 0863-0976

[39] H. v. Mangoldt, L. Knopp: Höhere Mathematik, Band 1 bis 4. Hirzel Verlag, Stuttgart 1990.

[40] W. Marszalek: Analysis of partial differential algebraic equations. Dissertation, North Carolina State University, Rayleigh 1997.

[41] W. Marszalek, Z. W. Trzaska: Analysis of implicit hyperbolic multivariable systems. Applied Mathematical Modelling 19 (1995), Nr. 7, S. 400-410.

[42] A. Ostermann, M. Roche: Runge-Kutta methods for partial differential equations and fractional order of convergence. Mathematics of Computation 59 (1992), Nr. 200, S. 403-420.

[43] C. C. Pantelides: The consistent initialization of differential-algebraic systems. SIAM Journal on Scientific and Statistical Computing 9 (1988), Nr. 2, S. 213-231.

[44] P. J. Rabier, W. C. Rheinboldt: Theoretical and numerical analysis of differential-algebraic equations. Handbook of Numerical Analysis, Vol. VIII, Elsevier Science, Amsterdam 2002.

[45] A. A. Samarskij: Theorie der Differenzenverfahren. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig 1984.

[46] J. M. Sanz-Serna, J. G. Verwer, W. H. Hundsdorfer: Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations. Numerische Mathematik 50 (1986), S. 405-418.

[47] B. Simeon: An extended descriptor form for the numerical integration of multibody systems. Applied Numerical Mathematics 13 (1993), S. 209-220.

[48] B. Simeon, M. Arnold: Coupling DAEs and PDAEs for simulating the interaction of pantograph and catenary. Mathematical and Computer Modelling of Dynamical Systems 6 (2000), Nr. 2, S. 129-144.

[49] B. Simeon, C. Führer, P. Rentrop: Differential-algebraic equations in vehicle system dynamics. Surveys on Mathematics for Industry 1 (1991), Nr. 1, S. 1-37.

[50] B. Simeon, C. Führer, P. Rentrop: The Drazin inverse in multibody system dynamics. Numerische Mathematik 64 (1993), S. 521-539.

[51] H. J. Stetter: Analysis of discretization methods for ordinary differential equations. Springer-Verlag, Berlin 1973.

[52] J. Stoer, R. Bulirsch: Einführung in die Numerische Mathematik II. Springer-Verlag, Berlin 1990.

[53] K. Strehmel, R. Weiner: Linear-implizite Runge-Kutta-Methoden und ihre Anwendung. B. G. Teubner Verlagsgesellschaft, Leipzig 1992.

[54] K. Strehmel, R. Weiner: Numerik gewöhnlicher Differentialgleichungen. B. G. Teubner Verlagsgesellschaft, Stuttgart 1995.

[55] J. W. Thomas: Numerical partial differential equations: Finite difference methods. Springer Verlag, New York 1995.

[56] J. G. Verwer: Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines. Numerical analysis (Dundee, 1985), Harlow 1986, S. 220-237.

[57] W. Walter: Differential and integral inequalities. Springer-Verlag, Berlin 1970.

[58] H. Werner, H. Arndt: Gewöhnliche Differentialgleichungen: Eine Einführung in Theorie und Praxis. Springer Verlag, Berlin 1986.

[59] M. W. Zieße, H. G. Bock, J. V. Galitzendörfer, J. P. Schlöder: Parameter estimation in multispecies transport reaction systems using parallel algorithms. In: J. Gottlieb, P. Du Chateau (eds.): Parameter Identification and Inverse Problems in Hydrology, Geology and Ecology. Kluwer Academic Publishers, 1996, S. 273-282.

# Lebenslauf

**Kristian Debrabant**
geboren am 12.12.1975 in Halberstadt

Schulausbildung

| | |
|---|---|
| 1982-1990 | Polytechnische Oberschule „Geschwister Scholl“ in Lutherstadt Eisleben |
| 1990-1994 | Mathematisch-naturwissenschaftliches Gymnasium „Georg Cantor“ in Halle |
| 24.06.1994 | Abitur |

Studium

| | |
|---|---|
| 1994-2001 | Studium an der Martin-Luther-Universität Halle-Wittenberg,<br>Stipendiat der Studienstiftung des deutschen Volkes,<br>diverse Tätigkeiten als wissenschaftliche Hilfskraft in Mathematik, Physik, Informatik, Zahnmedizin |
| 02.11.1999 | Vordiplom in Informatik |
| 26.01.2000 | Diplom in Mathematik<br>Diplomarbeit bei Herrn Prof. Dr. K. Strehmel zum Thema „Theoretische und numerische Untersuchungen zu partiellen differentiell-algebraischen Systemen“ |
| 2000 | Preis der Georg-Cantor-Vereinigung 2000 |
| 05.07.2001 | Diplom in Physik<br>Diplomarbeit bei Herrn Prof. Dr. J. Louis zum Thema „Stringkompaktifizierung mit Termen höherer Ordnung“ |
| 2000-2002 | Promotionsstudium am Institut für Numerische Mathematik der Martin-Luther-Universität Halle-Wittenberg, Stipendiat der Graduiertenförderung des Landes Sachsen-Anhalt |
| seit Okt. 2002 | Wissenschaftlicher Mitarbeiter am Fachbereich Mathematik der Technischen Universität Darmstadt, Arbeitsgruppe Numerik partieller Differentialgleichungen |

## Selbständigkeitserklärung

Hiermit erkläre ich an Eides Statt, daß ich die vorliegende Arbeit selbständig und ohne unzulässige fremde Hilfe verfaßt, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Halle (Saale) im April 2004

Kristian Debrabant