Zur numerischen Behandlung räumlich mehrdimensionaler parabolischer Differentialgleichungen mit linear-impliziten Splitting-Methoden und linearer partieller differentiell-algebraischer Systeme

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 Frau Claudia Eichler-Liebenow

geb. am: 14.05.1971 in Mühlhausen/Thür.

Gutachter:

1. Prof. Dr. K. Strehmel
2. Prof. Dr. P. Rentrop
3. Prof. Dr. W. Lucht

Halle(Saale), den 18. Juni 1999

„... numerical experimentation is a part of the subject of numerical solution of partial differential equations.“, J. W. Thomas [Tho95]

An dieser Stelle möchte ich mich recht herzlich bei Prof. Dr. K. Strehmel für seine kontinuierliche Betreuung seit meinem Studium und für seine Bemühungen, mir die Durchführung meines Promotionsvorhabens zu ermöglichen, *bedanken*. Mein besonderer *Dank* gilt auch Prof. Dr. W. Lucht und Prof. Dr. R. Weiner für die zahlreichen Diskussionen und die gemeinsame Arbeit. Ebenfalls *danke* ich Dr. L. Boltze sowie Dr. M. Büttner, Dipl.-Math. R. Manitz, Dr. J. Wensch und allen anderen Kollegen des Institutes für Numerische Mathematik für deren Unterstützung und das freundliche Arbeitsklima.

Weiterhin *bedanke* ich mich bei der Graduiertenförderung des Landes Sachsen-Anhalt, die mir durch die Bereitstellung eines Stipendiums das Promotionsstudium lange Zeit ermöglichte. So konnte ich während eines Aufenthaltes am CWI (Center for Mathematics and Computer Science) in Amsterdam in der Forschungsgruppe von Prof. P. van der Houwen wertvolle Erfahrungen sammeln.

Meinen Eltern und ganz besonders Stefan *danke* ich für ihre Hilfe und Verständnis während meiner Promotionszeit.

Inhaltsverzeichnis

Einleitung

Die mathematische Modellierung zahlreicher Zusammenhänge aus Naturwissenschaft und Technik führt oftmals auf spezielle Klassen zeitabhängiger partieller Differentialgleichungen. Daher ist die Untersuchung ihrer analytischen Lösung und ihrer numerischen Behandlung ein wichtiger Bestandteil der Mathematik. Eine numerische Simulation ist ein nützliches Hilfsmittel für den Naturwissenschaftler und Ingenieur, wenn das numerische Verfahren sicher und effizient arbeitet und die Modellierung möglichst realitätsnah ist. Die zunehmende Komplexität von Modellierungen erfordert sowohl die Entwicklung effektiver und zuverlässiger numerischer Integrationsverfahren für spezielle Klassen partieller Differentialgleichungssysteme als auch die Einbeziehung neuer Probleme in die Betrachtungen.

Die vorliegende Arbeit unterteilt sich in zwei Schwerpunkte, in denen sich mit der numerischen Behandlung jeweils einer speziellen Klasse von *Anfangs-Randwertproblemen* (ARWP) partieller Differentialgleichungen beschäftigt wird. Beiden Klassen gemein ist, daß zeitlich veränderliche Prozesse in einem räumlichen Gebiet betrachtet werden. Solch ein Prozeß ist auch wesentlich von dem momentanen Zustand des betrachteten Gebietes zum Beginn des Prozesses und den möglichen Einflüssen des das Gebiet umgebenen Mediums bestimmt. Dies wird durch die Formulierung von Anfangs- und geeigneten Randbedingungen berücksichtigt.

Im ersten Kapitel werden einige mathematische Grundlagen vorbereitet, die in Kapitel 2 und 3 verwendet werden. Es wird das Prinzip der *Linienmethode* zur numerischen Behandlung von zeitabhängigen partiellen Differentialgleichungen vorgestellt und anschließend die vertikale Linienmethode näher erläutert. Hierbei wird auf eine Ortsdiskretisierung mittels finiter Differenzen, die Zeitintegration des semidiskreten Problems mit Einschrittverfahren und die Konvergenz der Gesamtdiskretisierung eingegangen. Weiterhin werden einige Grundlagen und Begriffe der Theorie *linearer differentiell-algebraischer Gleichungen* mit konstanten Koeffizientenmatrizen zusammengestellt.

In Kapitel 2 wird eine Klasse *linear-impliziter Splitting-Methoden* zur numerischen Lösung *räumlich mehrdimensionaler parabolischer Anfangs-Randwertprobleme* entwickelt. Diese erweisen sich für zahlreiche solcher Probleme als sehr effektiv. Hierzu gehören skalare parabolische Differentialgleichungen der Form

$$\frac{\partial u}{\partial t}(t,x) = \sum_{i=1}^{d} \left[a_i(t,x,u) \frac{\partial^2 u}{\partial x_i^2}(t,x) + b_i(t,x,u) \frac{\partial u}{\partial x_i}(t,x) \right] + g(t,x,u) \qquad (1)$$

mit $x = (x_1, \ldots, x_d) \in \Omega = (0,1)^d \subset \mathbb{R}^d$, $t \in \mathfrak{I} = (0, t_e)$, $0 < t_e < \infty$. Es sei $\bar{\mathfrak{I}} = [0, t_e]$ und $u : \bar{\mathfrak{I}} \times [0,1]^d \to \mathbb{R}$. Die Funktionen $a_i, b_i, g : \bar{\mathfrak{I}} \times [0,1]^d \times \mathbb{R} \to \mathbb{R}$ seien hinreichend glatt. Es gelte $a_i > 0$, und die b_i seien von moderater Größe gegenüber den a_i. Für die Funktion g gelte $\frac{\partial g}{\partial u} \leq 0$. Unter diesen Voraussetzungen ist (1) eine parabolische Differentialgleichung (siehe z.B. [Mei90]). Weiterhin seien Anfangs- und geeignete Randbedingungen vorgeschrieben. Anstelle des Intervalls $[0,1]$ für jede Ortskoordinate kann man auch Intervalle $[a,b] \subset \mathbb{R}$ wählen, da man das gestellte Problem durch geeignete Koordinatentransformationen stets auf das Intervall $[0,1]$ überführen kann. Unter einer Lösung (im klassischen Sinne) der betrachteten Differentialgleichung verstehen wir eine stetige Funktion u mit mindestens einer stetigen Ableitung nach der Variablen t und zwei stetigen Ableitungen nach den Variablen x_i auf $\bar{\mathfrak{I}} \times [0,1]^d$ $(i = 1(1)d)$, die die Differentialgleichung und die gegebenen Anfangs- und Randbedingungen erfüllt.

Die betrachtete Problemklasse (1) kann z.B. auch auf Systeme parabolischer Differentialgleichungen erweitert werden. Hierbei wird allerdings nur eine Kopplung im Quellterm zugelassen, d.h., wir betrachten ein semilineares System der Form

$$\frac{\partial u}{\partial t}(t,x) = D\, \Delta u(t,x) + g(t,x,u) \qquad (2)$$

mit $x = (x_1, \ldots, x_d) \in \Omega$, $t \in \mathfrak{I}$, $u : \bar{\mathfrak{I}} \times [0,1]^d \to \mathbb{R}^n$. Es sei $D = \mathrm{diag}(\delta_1(t,x), \ldots, \delta_n(t,x)) \in \mathbb{R}^{n \times n}$ eine Diagonalmatrix, $\delta_i > 0$ stetig mit stetigen partiellen Ableitungen erster Ordnung. Weiterhin seien Anfangs- und geeignete Randbedingungen vorgeschrieben.

Zur numerischen Behandlung von (1) bzw. (2) wird die vertikale Linienmethode verfolgt, wobei die partiellen Ableitungen bez. der Ortsvariablen mittels finiter Differenzen approximiert werden. Die numerische Lösung des erhaltenen gewöhnlichen Differentialgleichungssystems, des semidiskreten Problems, ist Gegenstand des zweiten Kapitels. Dieses semidiskrete System ist in Abhängigkeit von der Zahl der Ortsvariablen und der Feinheit des Ortsgitters sehr groß und steif, besitzt aber eine spezielle Struktur. *Splitting-Methoden* sind Integrationsmethoden, die diese Struktur ausnutzen und dadurch den Rechenaufwand gegenüber anderen Integrationsmethoden erheblich reduzieren. Es wird zunächst ein Überblick über

einige bekannte Splitting-Methoden gegeben und auf die Stabilität von linearen Operator-Splitting-Methoden eingegangen. Hierfür werden geeignete Stabilitätsbegriffe definiert. Anschließend wird eine neue Klasse von linear-impliziten Splitting-Methoden eingeführt und hinsichtlich ihrer Konsistenz- und Stabilitätseigenschaften untersucht. Es zeigt sich, daß diese linear-impliziten Splitting-Methoden gute Stabilitätseigenschaften mit guter Implementierbarkeit verbinden. Es werden spezielle Verfahren angegeben und ihre Effektivität anhand von Beispielen illustriert.

Die Betrachtungen in Kapitel 3 wenden sich einer speziellen Klasse von Systemen partieller Differentialgleichungen zu, die bei der Modellierung komplexer Anwendungsprobleme in zunehmenden Maße auftreten und deren analytische und numerische Lösung daher in den letzten Jahren wachsendes Interesse gefunden hat. Diese partiellen Differentialgleichungssysteme bestehen aus einer Kopplung von Gleichungen unterschiedlichen Typs. Es werden (zeitabhängige) partielle Differentialgleichungen (engl.: *PDEs, partial differential equations*) z.B. mit differentiell-algebraischen Gleichungen (engl.: *DAEs, differential algebraic equations*) oder gewöhnlichen Differentialgleichungen (engl.: *ODEs, ordinary differential equations*) oder algebraischen Gleichungen gekoppelt. Derartige Systeme werden daher auch als partielle differentiell-algebraische Gleichungen (engl.: *PDAEs, partial differential algebraic equations*) bezeichnet. Ziel des dritten Kapitels der vorliegenden Arbeit ist es, eine Charakterisierung der Eigenschaften linearer PDAEs der Form

$$A\,\frac{\partial u}{\partial t}(t,x) + B\,\frac{\partial^2 u}{\partial x^2}(t,x) + C\,u(t,x) = g(t,x), \quad (t,x) \in \mathfrak{I} \times \Omega \tag{3}$$

mit $A, B, C \in \mathbb{R}^{n\times n}$, $\Omega = (-l, l)$ $(0 < l < \infty)$, $\bar{\Omega} = [-l, l]$ und $u, g : \bar{\mathfrak{I}} \times \bar{\Omega} \to \mathbb{R}^n$ zu geben. Mindestens eine der beiden Matrizen A, B sei dabei singulär. Für eine eindeutige Lösbarkeit muß (3) durch Anfangs- und geeignete Randbedingungen ergänzt werden. In dem Kapitel wird ausgeführt, daß im Gegensatz zu PDEs mit regulären Matrizen A, B (z.B. parabolische Differentialgleichungssysteme) bei singulärem A und/oder B nicht für alle Komponenten von u Anfangs- und/oder Randbedingungen vorgegeben werden können. Sie müssen gewissen zusätzlichen Bedingungen genügen. Die lineare PDAE (3) wird der Laplace- und einer endlichen Fouriertransformation unterzogen. Durch die beiden Transformationen wird (3) auf parameterabhängige DAEs überführt. Auf dieser Grundlage werden in Analogie zur Theorie der DAEs ein differentieller Ortsindex und ein einheitlicher differentieller Zeitindex für lineare PDAEs eingeführt. Diese charakterisieren spezielle Eigenschaften der betrachteten Problemklasse sowohl bez. der analytischen Lösung als auch der numerischen Behandlung. Anhand von Beispielen werden ferner die Schwierigkeiten aufgezeigt, die mit einem 'nichteinheitlichen' Zeitindex einer linearen PDAE verbunden sind. Die

beiden eingeführten Indexe werden dann bei der numerischen Behandlung linearer PDAEs verwendet, um Konvergenzaussagen zu treffen. Hierzu wird wieder die Linienmethode benutzt, wobei partielle Ortsableitungen durch finite Differenzenapproximationen ersetzt werden. Für die Zeitintegration des semidiskreten Problems wird das implizite Euler-Verfahren bzw. die Trapezregel verwendet. Es wird der Gesamtdiskretisierungsfehler des hieraus bei äquidistanter Orts- und Zeitschrittweite resultierenden BTCS Schemas bzw. Crank-Nicolson Verfahrens untersucht. Unter gewissen Voraussetzungen werden Konvergenzaussagen in Abhängigkeit der beiden Indexe getroffen. Für schwach gekoppelte lineare PDAEs können diese Aussagen unter schwächeren bzw. einfacher zu überprüfenden Voraussetzungen gefunden werden. Einige numerische Testrechnungen illustrieren die Ergebnisse und die Besonderheiten linearer PDAEs bei der numerischen Behandlung.

Die einzelnen Kapitel dieser Arbeit unterteilen sich in Abschnitte, Unterabschnitte und Paragraphen. Die Nummerierung für Strukturen wie Definitionen, Sätze, Lemmata, Beispiele usw. bestehen aus 3 Ziffern: der Kapitel- und der Abschnittsnummer sowie einer innerhalb eines Abschnitts über alle diese Strukturen fortlaufenden Nummer. Die Formelnummern setzen sich ebenfalls aus der Kapitel-, der Abschnitts- und einer im Abschnitt fortlaufenden Nummer zusammen. Am Ende der Arbeit ist ein Literaturverzeichnis und eine Übersicht über häufig verwendete Abkürzungen und Bezeichnungen zu finden.

Einige Ergebnisse dieser Arbeit wurden in [EL98a] und [Luc99] veröffentlicht.

KAPITEL 1

Mathematische Grundlagen

1.1. Die Linienmethode

Zur numerischen Behandlung von zeitabhängigen partiellen Differentialgleichungen werden diese diskretisiert. Eine übliche Methode ist, die Orts- und die Zeitdiskretisierung als zwei aufeinanderfolgende Prozesse zu betrachten, obwohl diese im gesamten numerischen Verfahren nicht unabhängig voneinander sind. So unterscheidet man zwischen der (vertikalen) Linienmethode und der Rothe-Methode (oder auch horizontale Linienmethode) [Gro92]. Wird die Lösung eines zeitabhängigen Anfangs-Randwertproblems mit der *Rothe-Methode* approximiert, so wird das Ausgangsproblem durch Diskretisierung bezüglich der Zeitvariablen in eine Folge von elliptischen Randwertproblemen überführt. Bei der (*vertikalen*) *Linienmethode* wird der numerische Integrationsprozeß untergliedert in die Semidiskretisierung bezüglich der Ortsvariablen und anschließender Zeitintegration des resultierenden Anfangswertproblems gewöhnlicher Differentialgleichungen. Im folgenden beschränken wir uns auf die (vertikale) Linienmethode.

Die (vertikale) Linienmethode. In der *Semidiskretisierung bez. der Ortsvariablen* werden die in den ARWPn der partiellen Differentialgleichungen (1), (2) bzw. (3) auftretenden partiellen Ortsableitungen im allgemeinen entweder mittels „finiter Differenzen“ ([Mit80],[Tho95]), mittels „finiter Elemente“ ([Mit78]) oder nach der „Spektralmethode“ ([Got77]) approximiert. Man erhält ein System von Anfangswertproblemen von r gewöhnlichen Differentialgleichungen 1.Ordnung bzw. eine DAE, d.h. ein System der Form

$$\bar{A}\, w'(t) = f(t, w(t)), \quad w(0) = w_0, \quad w(t) \in \mathbb{R}^r, \tag{1.1.1}$$

mit $f : \bar{\mathfrak{I}} \times \mathbb{R}^r \longrightarrow \mathbb{R}^r$, $r \in \mathbb{N}$. Für die Problemklassen (1),(2) ist $\bar{A} = I_r$. Für die PDAE (3) ist $\bar{A}$ von A abhängig und singulär, wenn A singulär ist, und der Anfangswert w_0 muß dann gewissen zusätzlichen Bedingungen genügen (siehe Abschnitt 1.2). Insbesondere ist r abhängig von der Anzahl n der Gleichungen des zugrundeliegenden partiellen Differentialgleichungssystems, der Dimension d von Ω und der Art der Ortsdiskretisierung (z.B. der Feinheit des Ortsgitters bei Verwendung „finiter Differenzen“). Die Gleichung (1.1.1) wird auch als *semidiskretes System*

bezeichnet, da nur bezüglich der Ortsvariablen x diskretisiert, aber die Zeit t weiterhin stetig gelassen wurde. In der *Zeitintegration* wird anschließend das erhaltene Problem (1.1.1) durch eine numerische Integrationsmethode gelöst. Hierbei wird im folgenden nur die Verwendung von Einschrittverfahren betrachtet.

1.1.1. Ortsdiskretisierung

Im weiteren werden wir finite Differenzen zur Semidiskretisierung nutzen. Auf $\bar{\Omega}$ wird ein Ortsgitter Ω_h gelegt. Der Einfachheit wird ein Gitter gewählt, das äquidistant bez. einer Raumdimension ist. Für $\Omega = (a,b)^d$ ist

$$\begin{aligned}\Omega_h = \Big\{ x_{i_1,\dots,i_d} = (x_{i_1},\dots,x_{i_d})^\top : x_{i_k} = a + i_k\, h_k, \\ i_k = 0(1)M_k + 1,\ h_k = \tfrac{b-a}{M_k+1},\ k = 1(1)d \Big\}.\end{aligned} \tag{1.1.2}$$

Die partiellen Ortsableitungen werden nun durch geeignete Differenzenquotienten approximiert. Durch Einsetzen dieser Approximationen in die zu betrachtende partielle Differentialgleichung erhält man für jeden Gitterpunkt eine semidiskrete (zeitabhängige) Gleichung. Aufgrund der Abhängigkeiten zu den benachbarten Gitterpunkten in (1.1.8) und (1.1.9) bilden diese Gleichungen für alle Gitterpunkte ein großes System gewöhnlicher Differentialgleichungen der Form (1.1.1) der Dimension r. Es gilt $r \sim n\, M_1 \dots M_d$. Jede Komponente der zu bestimmenden Vektorfunktion $w(t)$ ist genau einem Gitterpunkt zugeordnet.

Um die von x abhängige Lösung $u(t,x)$ mit dem bez. des Ortes diskreten Vektor $w(t)$ des $\mathbb{R}^r$ zu vergleichen, führen wir die Bezeichnung $u_h(t)$ als geeignete Repräsentation der exakten Lösung $u(t,x)$ im $\mathbb{R}^r$ ein. Bei einer Semidiskretisierung mittels finiter Differenzen ist $u_h(t)$ die Restriktion der exakten Lösung auf das Gitter. D.h., entspricht die k-te Komponente von $w(t)$ dem Punkt $x_{i_1,\dots,i_d} \in \Omega_h$, so ist die k-te Komponente von $u_h(t)$ gleich dem Wert $u(t, x_{i_1,\dots,i_d})$.

Bemerkung 1.1.1. Zur Beurteilung der Güte von Diskretisierungsverfahren werden im folgenden Normen von Fehlergrößen in einem Vektorraum betrachtet. Hierbei werden durch ein Skalarprodukt induzierte Normen des $\mathbb{R}^k$

$$\|y\| = \sqrt{\langle y, y\rangle} = \sqrt{y^\top R\, y}, \quad y \in \mathbb{R}^k, \tag{1.1.3}$$

mit einer Matrix $R \in \mathbb{R}^{k\times k}$ (symmetrisch, positiv definit) verwendet.

Die Euklidische Norm (2-Norm) ist mit $R = I_k$ gegeben durch

$$\|y\|_{2,k} = \sqrt{y^\top y} = \sqrt{\textstyle\sum_{i=1}^k y_i^2}, \quad y \in \mathbb{R}^k\,.$$

Sei M^d die Anzahl der inneren Ortsgitterpunkte eines äquidistanten Gitters und $w(t) \in \mathbb{R}^{nM^d}$ wie oben beschrieben. Dann wächst die Dimension von w mit feiner

werdendem Ortsgitter ($M \to \infty$). Es ist daher zweckmäßig, eine diskrete L_2-Norm des $\mathbb{R}^{nM^d}$ zu verwenden, die man mit $R = h^d I_{nM^d}$ und $h \sim M^{-1}$ erhält, d.h.

$$\|y\| = \sqrt{h^d\, y^\top y}, \quad y \in \mathbb{R}^{nM^d \times nM^d}. \tag{1.1.4}$$

Insbesondere gilt mit dieser Norm für eine stetige Funktion $u : \bar{\mathfrak{I}} \times \bar{\Omega} \to \mathbb{R}^n$ mit $\int_{\bar{\Omega}} u^\top(t,x)u(t,x)\,dx < \infty$ $(u \in L_2(\bar{\Omega}))$

$$\|u_h(t)\| \to \|u(t,x)\|_{L_2} \quad \text{für } h \to 0 \quad \forall t \in \bar{\mathfrak{I}},$$

wobei $\|u(t,x)\|_{L_2}^2 = \int_{\bar{\Omega}} u^\top(t,x)u(t,x)\,dx$ die L_2-Norm und $u_h(t)$ die Restriktion von u auf das Ortsgitter ist. D.h., die Normen sind aufeinander abgestimmt (vgl. [Sam84]).

Diese hier genannten Normen werden im folgenden stets verwendet, sofern dies nicht näher erläutert wird. Bezüglich der gewählten Vektornorm ist $\|A\|$ die zugeordnete Matrixnorm und $\mu[A]$ die induzierte logarithmische Matrixnorm für Matrizen $A \in \mathbb{R}^{k\times k}$. □

Um die Güte der Semidiskretisierung einschätzen zu können, werden der *lokale* und der *globale Ortsdiskretisierungsfehler* untersucht.

Definition 1.1.2. *Sei $u_h(t)$ die Restriktion der exakten Lösung $u(t,x)$ der betrachteten partiellen Differentialgleichung der Form (1), (2) oder (3) auf das Gitter Ω_h und $f(t,w)$ die rechte Seite des semidiskreten Problems (1.1.1). Dann heißt die Größe*

$$\alpha_h(t) := \bar{A}\, u_h'(t) - f(t, u_h(t))$$

lokaler Ortsdiskretisierungsfehler.

Definition 1.1.3. *Sei $\mathfrak{I}^* := [0, t^*]$, $t^* > 0$. Die Semidiskretisierung heißt* konsistent *bzw.* konsistent der Ordnung p^* *auf $\mathfrak{I}^*$, wenn*

$$\max_{t\in\mathfrak{I}^*} \|\alpha_h(t)\| \to 0 \qquad \textit{für} \quad h \to 0$$

$$\textit{bzw.} \qquad \max_{t\in\mathfrak{I}^*} \|\alpha_h(t)\| = \mathcal{O}(h^{p^*}) \qquad \textit{für} \quad h \to 0$$

gleichmäßig in t gilt, wobei $h = \max\{h_k : k = 1(1)d\}$.

Der lokale Ortsdiskretisierungsfehler stellt den Defekt der Gitterfunktion $u_h(t)$ gegenüber der Differentialgleichung (1.1.1) dar. Er gibt an wie 'gut' $u_h(t)$ die Differentialgleichung (1.1.1) erfüllt.

Definition 1.1.4. *Sei $w(t)$ die exakte Lösung von (1.1.1), dann heißt der Vektor*

$$\eta_h(t) \quad := \quad w(t) - u_h(t) \tag{1.1.5}$$

globaler Ortsdiskretisierungsfehler.

DEFINITION 1.1.5. *Sei* $\mathfrak{I}^* := [0, t^*]$, $t^* > 0$. *Die Semidiskretisierung heißt* konvergent *bzw. konvergent der Ordnung* p^* *auf* $\mathfrak{I}^*$, *wenn*

$$\max_{t\in\mathfrak{I}^*} \|\eta_h(t)\| \to 0 \qquad \textit{für} \quad h \to 0$$

$$\textit{bzw.} \qquad \max_{t\in\mathfrak{I}^*} \|\eta_h(t)\| = \mathcal{O}(h^{p^*}) \qquad \textit{für} \quad h \to 0.$$

Für die partiellen Ableitungen nach den Ortsvariablen werden nun Approximationen 2.Ordnung verwendet. D.h., es werden für $t \in [0, t_e]$ und die inneren Gitterpunkte $x = (x_1, \dots, x_d) \in \Omega$ die Beziehungen

$$\frac{\partial^2 u}{\partial x_k^2}(t,x) = \frac{1}{h_k^2}\Big(u(t, x - h_k\, e_k) - 2\, u(t,x) + u(t, x + h_k\, e_k)\Big) + \underbrace{h_k^2 \frac{1}{24}\left(\frac{\partial^4 u}{\partial x_k^4}(t, x - h_k\xi_1 e_k) + \frac{\partial^4 u}{\partial x_k^4}(t, x + h_k\xi_2 e_k)\right)}_{=\ \mathcal{O}(h_k^2)}, \tag{1.1.6}$$

$$\frac{\partial u}{\partial x_k}(t,x) = \frac{1}{2h_k}\Big(u(t, x + h_k\, e_k) - u(t, x - h_k\, e_k)\Big) + \underbrace{h_k^2 \frac{1}{2}\left(\frac{\partial^2 u}{\partial x_k^2}(t, x - h_k\bar{\xi}_1 e_k) - \frac{\partial^2 u}{\partial x_k^2}(t, x + h_k\bar{\xi}_2 e_k)\right)}_{=\ \mathcal{O}(h_k^2)}, \tag{1.1.7}$$

$k = 1(1)d$, $e_k = (e_{k1}, \dots, e_{kd})^\top \in \mathbb{R}^d$, $e_{kj} = 0$ für $(k \neq j)$ und $e_{kk} = 1$ benutzt, wobei $h_k > 0$ hinreichend klein sei und $\xi_1, \bar{\xi}_1, \xi_2, \bar{\xi}_2 \in (0,1)$ gilt. Es sei bemerkt, daß für jede Komponente von $\frac{\partial^4 u}{\partial x_k^4}$ bzw. $\frac{\partial^2 u}{\partial x_k^2}$ die ξ_i bzw. $\bar{\xi}_i$ $(i = 1, 2)$ unterschiedlich sein können. Unter Verwendung der Beziehungen (1.1.6) und (1.1.7) approximiert man in jedem Punkt $x_{i_1,\dots,i_d} \in \Omega_h \cap \Omega$ die partiellen Ortsableitungen durch

$$\frac{\partial^2 u}{\partial x_k^2}(t, x_{i_1,\dots,i_d}) \approx \frac{1}{h_k^2}\Big(u_{i_1,\dots,i_{k-1},\dots,i_d}(t) - 2\, u_{i_1,\dots,i_d}(t) + u_{i_1,\dots,i_{k+1},\dots,i_d}(t)\Big), \tag{1.1.8}$$

$$\frac{\partial u}{\partial x_k}(t, x_{i_1,\dots,i_d}) \approx \frac{1}{2h_k}\Big(u_{i_1,\dots,i_{k+1},\dots,i_d}(t) - u_{i_1,\dots,i_{k-1},\dots,i_d}(t)\Big), \tag{1.1.9}$$

wobei $u_{i_1,\dots,i_d}(t) = u(t, x_{i_1,\dots,i_d})$, $k = 1(1)d$. Bei Verwendung von (1.1.8) bzw. (1.1.9) zur Ortsdiskretisierung von (1), (2) bzw. (3) ist die Semidiskretisierung für hinreichend glattes u auf $\mathfrak{I}^* \times \Omega$ konsistent der Ordnung $p^* = 2$.

BEISPIEL 1.1.6. Verwendet man z.B. zur Diskretisierung des ARWPs der eindimensionalen Wärmeleitungsgleichung

$$u_t = u_{xx}, \quad x \in (0,1),\ t > 0,$$

mit $u(t,0) = u(t,1) = 0$ und $u(0,x) = u_0(x)$ $(x \in [0,1], t \geq 0)$ für die Ortsdiskretisierung mit äquidistanter Gitterweite $h = \frac{1}{M+1}$ die Approximationen (1.1.6) und wählt $w(t) = (w(t,x_1), \ldots, w(t,x_M))^\top \approx (u(t,x_1), \ldots, u(t,x_M))^\top =: u_h(t)$, so erhält man das semidiskrete Problem $(t > 0)$

$$w'(t) = A\,w(t),\ w(0) = u_h(0), \quad A = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{M\times M}. \tag{1.1.10}$$

□

1.1.2. Zeitintegration semidiskreter parabolischer ARWPe

Im folgenden beschränken wir uns auf das semidiskrete Problem, das aus der Semidiskretisierung parabolischer ARWPe (1) bzw. (2) entsteht[1]. D.h., wir betrachten (1.1.1) mit $\bar{A} = I_r$ für festes h:

$$w'(t) = f(t, w(t)), \quad w(0) = w_0. \tag{1.1.11}$$

Wir werden hierzu in diesem Abschnitt auf einige wichtige Grundlagen der Theorie der Einschrittverfahren für gewöhnliche Differentialgleichungen eingehen (vgl. z.B. [Hai93],[Hai96],[Str95]). Auf das Zeitintervall $[0, t_e]$, $t_e < \infty$, wird ein Punktgitter

$$\mathfrak{I}_\tau := \{0, t_1, \ldots, t_{m_{end}}\}, \qquad 0 = t_0 < t_1 < \ldots < t_{m_{end}} = t_e$$

mit den Schrittweiten $\tau_m = t_m - t_{m-1}$, $m = 1(1)m_{end}$, gelegt. Gesucht wird eine Näherungslösung für das Anfangswertproblem (1.1.11), d.h. eine Gitterfunktion

$$v_\tau : \mathfrak{I}_\tau \longrightarrow \mathbb{R}^r \quad \text{mit} \quad v_\tau(t_m) \approx w(t_m),\ m = 0(1)m_{end}.$$

Unter einem *Diskretisierungsverfahren* zur Approximation der Lösung des Anfangswertproblems wird eine Verfahrensvorschrift verstanden, die jedem $t_m \in \mathfrak{I}_\tau$ einen Vektor v_m zuordnet. Ein *Einschrittverfahren* ist ein Diskretisierungsverfahren zur Bestimmung einer Gitterfunktion $v_m = v_\tau(t_m) \approx w(t_m)$ der Gestalt

$$\begin{aligned} v_{m+1} &= v_m + \tau_m \Phi(t_m, v_m; \tau_m), \qquad m = 0(1)m_{end}-1 \\ v_0 &= w(0) \end{aligned} \tag{1.1.12}$$

mit $\Phi := \Phi(t, v; \tau)$: *Verfahrens- oder Inkrementfunktion* des Einschrittverfahrens. (Dabei ist die Darstellung (1.1.12) nur formal explizit und umfaßt auch implizite Methoden.)

Definition 1.1.7. *Sei* $\mathfrak{S} := \{(t,w) : 0 \leq t \leq t_e, w \in \mathbb{R}^r\}$, $f : \mathfrak{S} \to \mathbb{R}^r$ *und* $\langle .,. \rangle$ *ein Skalarprodukt im* $\mathbb{R}^r$ *mit der zugehörigen Norm* $\|y\| = \sqrt{\langle y, y\rangle}$ $(y \in \mathbb{R}^r)$.

[1] Das semidiskrete Problem der PDAE (3) wird in Kapitel 3 untersucht.

Dann genügt die Funktion f einer einseitigen Lipschitz-Bedingung*, wenn gilt*

$$\langle f(t,v) - f(t,w), v - w\rangle \leq l(t)\|v-w\|^2, \quad \textit{für alle} \quad (t,v),(t,w) \in \mathfrak{S},$$

wobei $l(t)$ einseitige Lipschitz-Konstante *von f auf $\mathfrak{S}$ heißt.*

Die Funktion f heißt in $\mathfrak{S}$ (bez. w) Lipschitz-stetig*, wenn es eine Konstante $L(t) > 0$ gibt, so daß*

$$\|f(t,v) - f(t,w)\| \leq L(t)\,\|v-w\| \quad \textit{für alle} \quad (t,v),(t,w) \in \mathfrak{S}$$

gilt. Die Konstante $L(t)$ heißt Lipschitz-Konstante *von f auf $\mathfrak{S}$.*

Es bezeichne $\mathfrak{Lip}(\mathfrak{S}) := \{f \text{ mit } f : \mathfrak{S} \to \mathbb{R}^r,\ f \text{ Lipschitz-stetig in } \mathfrak{S}\}$. (Für die Lipschitz-Stetigkeit ist hinreichend, daß die Funktion auf $\mathfrak{S}$ stetig differenzierbar ist.)

Voraussetzung 1.1.8. Für hinreichend kleine Schrittweiten $\tau > 0$ ordne die Abbildung Φ jeder Funktion $f \in \mathfrak{Lip}(\mathfrak{S})$ eine Funktion $\Phi(t,w;\tau) \in \mathfrak{Lip}(\mathfrak{S})$ zu. Ferner sei Φ in τ stetig.

Definition 1.1.9. *Sei $\tilde{v}_{m+1}$ das Resultat eines Schrittes von (1.1.12) mit dem Startvektor auf der exakten Lösungskurve, d.h. $\tilde{v}_{m+1} := w(t_m)+\tau_m\Phi(t_m, w(t_m);\tau_m)$. Dann heißt*

$$le_\tau(t_{m+1}) = le_\tau(t_m+\tau_m) \ := \ w(t_{m+1}) - \tilde{v}_{m+1}$$

lokaler Zeitdiskretisierungsfehler *des Einschrittverfahrens an der Stelle $t = t_m+\tau_m$.*

Diese Fehlergröße dient zur qualitativen Beurteilung der Verfahrensfunktion Φ. Sie gibt an, ob diese für jede Funktion $f \in \mathfrak{Lip}(\mathfrak{S})$ eine lokale Approximation der Differentialgleichung (1.1.11) liefert.

Definition 1.1.10. *Sei $w(t)$ die Lösung des Anfangswertproblems (1.1.11) auf $[0,t_e]$. Dann heißt ein Einschrittverfahren* konsistent *(mit dem Anfangswertproblem) bez. einer Norm $\|.\|$, wenn für jede Funktion $f \in \mathfrak{Lip}(\mathfrak{S})$ die Beziehung*

$$\max_{m=0(1)m_{end}-1} \|f(t_m, w(t_m)) - \Phi(t_m, w(t_m);\tau_m)\| \to 0 \tag{1.1.13}$$

für $\tau_{max} \to 0$ gilt, wobei $\tau_{max} = \max\{\tau_m : m = 0(1)m_{end}-1\}$.

Lemma 1.1.11. [Str95] *Die Bedingung (1.1.13) ist genau dann erfüllt, wenn*

$$\max_{m=0(1)m_{end}-1} \frac{\|le_\tau(t_m+\tau_m)\|}{\tau_m} \to 0 \quad \textit{für} \quad \tau_{max} \to 0 \tag{1.1.14}$$

gilt.

Für praktische Belange ist die Güte der Approximation $\tilde{v}_{m+1}$ wichtig. Ein Maß für die lokale Genauigkeit ist die Konsistenzordnung.

DEFINITION 1.1.12. *Ein Einschrittverfahren (1.1.12) besitzt die* Konsistenzordnung p *(*klassische Konsistenzordnung p*) bez. einer Norm $\|.\|$, wenn p die größte positive ganze Zahl ist, so daß für jede genügend oft stetig differenzierbare Lösung $w(t)$ von (1.1.11) gilt*

$$\max_{t\in\mathfrak{J}_\tau\setminus\{t_{m_{end}}\}} \|le_\tau(t+\tau)\| \le C\tau^{p+1} \qquad \textit{für} \qquad \tau\in(0,T] \tag{1.1.15}$$

mit einer von τ unabhängigen Konstanten C, $T>0$ hinreichend klein.

C ist abhängig von Schranken für f und partiellen Ableitungen von f bis zur Ordnung p. Insbesondere ist C auch abhängig von der Lipschitz-Konstanten von f. Voraussetzung für die Konsistenzordnung p eines Einschrittverfahrens ist, daß $w(t)$ mindestens $(p+1)$-mal stetig differenzierbar in $[0,t_e]$ ist.

DEFINITION 1.1.13. *Ein Einschrittverfahren heißt* konvergent *für die Anfangswertaufgabe (1.1.11) auf $[0,t_e]$ bez. einer Norm $\|.\|$, wenn für jede Folge von Gittern $\mathfrak{J}_\tau$ mit $\tau_{max}\to 0$ für den* globalen Zeitdiskretisierungsfehler

$$e_\tau(t) := w(t) - w_\tau(t)$$

die Beziehung $\max_{t\in\mathfrak{J}_\tau}\|e_\tau(t)\|\to 0$ für $\tau_{max}\to 0$ gilt.

Konvergenz bedeutet, daß das Einschrittverfahren für feiner werdende Diskretisierungen die analytische Lösung $w(t)$ in exakter Arithmetik beliebig genau approximiert. Das Einschrittverfahren besitzt die *Konvergenzordnung* p^*, wenn p^* die größte positive ganze Zahl ist, so daß für alle Schrittweiten mit $\tau_{max}\in(0,T]$ gilt

$$\max_{t\in\mathfrak{J}_\tau}\|e_\tau(t)\| \le C\,\tau_{max}^{p^*}\ , \tag{1.1.16}$$

wobei die Konstante C unabhängig von τ_{max} ist.

Systeme gewöhnlicher Differentialgleichungen, die durch sehr große Lipschitz-Konstanten gekennzeichnet sind, während sie i.allg. eine einseitige Lipschitz-Konstante von moderater Größe besitzen, werden als *steife Systeme* bezeichnet ([Str95]). Sie stellen hohe Anforderung an die Stabilität numerischer Verfahren. Insbesondere sind Systeme, die wir aus der Semidiskretisierung parabolischer ARWPe (wie in Abschnitt 1.1.1 beschrieben) erhalten, steife Differentialgleichungen:

BEISPIEL 1.1.14. Die Matrix A in Beispiel 1.1.6 hat die Eigenwerte ([Tho95]) $\lambda_k = -\frac{4}{h^2}\sin^2\left(\frac{k\pi h}{2}\right) < 0$ $(k=1(1)M)$. D.h., $\lambda_1\approx -\pi^2$ und $\lambda_M\approx -\frac{\pi^2}{h^2}\to-\infty$ für $h\to 0$. Die Lipschitz-Konstante der rechten Seite von (1.1.10) ist durch $\|A\|$ und die einseitige Lipschitz-Konstante durch $\mu[A]$ gegeben (vgl. [Dek84]). Da A symmetrisch ist, gilt in der Spektralnorm $\|A\|_{2,M} = \max_{k=1(1)M)}\{|\lambda_k|\} = -\lambda_M$ und in der der 2-Norm zugeordneten logarithmischen Matrixnorm $\mu[A] = \max_{k=1(1)M}\{\lambda_k\} = \lambda_1 \approx -\pi < 0$. D.h., (1.1.10) ist steif. □

1.1.3. Konsistenz und Konvergenz der Gesamtdiskretisierung

In Anlehnung an Verwer/Sanz-Serna, die in [Ver84] die Konvergenz von auf der Linienmethode basierenden Approximationen partieller Differentialgleichungen untersuchen (siehe auch [Sha94]), wird in diesem Abschnitt auf die Konvergenz der Gesamtdiskretisierung, die sich aus der Orts- und der Zeitdiskretisierung zusammensetzt, eingegangen.

Hierzu betrachten wir den Gesamtdiskretisierungsfehler:

DEFINITION 1.1.15. *Der* Gesamtdiskretisierungsfehler *der numerischen Lösung v_{m+1} eines Diskretisierungsverfahrens zur Lösung von (1),(2) oder (3) gegenüber der exakten Lösung zum Zeitpunkt t_{m+1} ist definiert durch*

$$\varepsilon_h(t_{m+1}) := u_h(t_{m+1}) - v_{m+1}. \tag{1.1.17}$$

Der Einfachheit halber betrachten wir äquidistante Schrittweiten h und τ. Wir halten $t_{m+1} \in \mathfrak{I}$ fest und nehmen im Grenzprozeß an, daß $\tau \to 0$ und $m \to \infty$ so erfolgen, daß $(m+1)\tau = t_{m+1}$.

DEFINITION 1.1.16. *Eine Gesamtdiskretisierung heißt bez. einer Norm $\|.\|$* konvergent der Ordnung (p, q)*, wenn die Ordnungsbedingung*

$$\|\varepsilon_h(t_{m+1})\| = \mathcal{O}(h^p) + \mathcal{O}(\tau^q) \quad \textit{für} \quad \tau,\, h \to 0,\ m = 0, 1, ... \tag{1.1.18}$$

erfüllt ist.

Es ist $\varepsilon_h(t_{m+1}) = u_h(t_{m+1}) - w(t_{m+1}) + w(t_{m+1}) - v_{m+1}$ und

$$\|\varepsilon_h(t_{m+1})\| \le \|\eta_h(t_{m+1})\| + \|e_\tau(t_{m+1})\|. \tag{1.1.19}$$

Der Gesamtdiskretisierungsfehler verschwindet aber nicht notwendig, wenn η_h für $h \to 0$ und e_τ für $\tau \to 0$ verschwinden. Dies resultiert daraus, daß i.allg. in die Abschätzungen für $\|e_\tau(t_{m+1})\|$ die Lipschitz-Konstante des gewöhnliches Systems einfließt. Diese wird für semidiskrete Probleme, die z.B. aus einer PDE der Form (1) resultieren, für $h \to 0$ beliebig groß (vgl. Beispiel 1.1.14).

Für die Konvergenz der Gesamtdiskretisierung muß daher gefordert werden, daß das Verfahren für die Zeitintegration bez. des Ortsgitters unabhängig vom Verhältnis von Orts- und Zeitschrittweite konvergiert. D.h., für $v_{m+1} \to u_h(t_{m+1})$ mit $h, \tau \to 0$ muß $v_{m+1} \to w(t_{m+1})$ für $\tau \to 0$ unabhängig von h (*gleichmäßig*) sein ([Dek84], [Auz90]).

DEFINITION 1.1.17. *Gilt die Beziehung (1.1.18) nur unter einer Bedingung zwischen τ und h, so heißt die Gesamtdiskretisierung* bedingt konvergent der Ordnung (p, q).

Beispiel 1.1.18. Wendet man auf das semidiskrete Problem (1.1.10) das explizite Euler-Verfahren $v_{m+1} = v_m + \tau f(t_m, v_m)$ an, so gilt $\|\eta_h(t_{m+1})\| = \mathcal{O}(h^2)$ für $h \to 0$ und $\|e_\tau(t_{m+1})\| = \mathcal{O}(\tau)$ für h fest und $\tau \to 0$. Allerdings erhält man $\|e_\tau(t_{m+1})\| = \mathcal{O}(\tau)$ für $\tau, h \to 0$ nur unter der Bedingung $\frac{\tau}{h^2} \leq \frac{1}{2}$ (vgl. z.B. [Ric67]). Dies bedeutet für $h \to 0$ eine starke Schrittweiteneinschränkung an τ, und die Gesamtdiskretisierung ist in diesem Beispiel nur bedingt konvergent. □

Für steife gewöhnliche Differentialgleichungssysteme wurde das Konzept der sogenannten *B-Konsistenz* entwickelt ([Fra81],[Dek84]), auf die in Abschnitt 2.5.3 eingegangen wird. Es liefert bez. der Steifheit gleichmäßige Fehlerabschätzungen. Die Übertragung dieses Prinzips auf die Konvergenzanalyse der Linienmethode bedeutet, daß Fehlerschranken gesucht werden, die unabhängig vom Ortsdiskretisierungsparameter h sind.

Verwer/Sanz-Serna weisen in ihrer Arbeit [Ver84] darauf hin, daß es bei Verwendung von (1.1.19) i.allg. kompliziert sein kann, die in h gleichmäßige Konvergenz von $e_\tau(t_{m+1})$ bez. τ nachzuweisen. Sie stellen fest, daß es für die Konvergenzanalyse meist günstiger ist, den Ausdruck

$$\varepsilon_h(t_{m+1}) = le_h(t_{m+1}) + \hat{v}_{m+1} - v_{m+1}$$

auszuwerten, wobei $le_h(t_{m+1})$ der lokale Gesamtdiskretisierungsfehler ist:

Definition 1.1.19. *Der* lokale Gesamtdiskretisierungsfehler *eines Diskretisierungsverfahrens zur Lösung von (1),(2) oder (3) ist definiert durch*

$$le_h(t_{m+1}) := u_h(t_{m+1}) - \hat{v}_{m+1}, \tag{1.1.20}$$

wobei $\hat{v}_{m+1}$ das Resultat eines Zeitschrittes mit dem Einschrittverfahren mit dem Startvektor auf der exakten Lösungskurve ist, d.h. wenn $v_m = u_h(t_m)$.

Bemerkung 1.1.20. Die Untersuchung der klassischen Konvergenzordnung ist im Zusammenhang mit der Linienmethode weiterhin eine übliche Vorgehensweise zur Herleitung effektiver Diskretisierungsverfahren, da sie die Güte des Diskretisierungsverfahrens bez. der Zeitdiskretisierung charakterisiert. Eine gute klassische Konsistenzordnung ist somit eine notwendige Eigenschaft für die Güte der Gesamtdiskretisierung. □

1.2. Indexe für DAEs

Im folgenden Abschnitt wollen wir für unsere Untersuchungen relevante Begriffe und Ergebnisse aus der Theorie der DAEs (auch als Algebro-Differentialgleichungen, differential-algebraische oder differentiell-algebraische Gleichungen bezeichnet) kurz

vorstellen (vgl. z.B. [Pet82],[Gri86],[Hai96], [Bre89], [Deu94], [Ren96],[ES98]). Eine DAE hat die allgemeine Form

$$F(t, w(t), w'(t)) = 0, \quad \text{mit } t \in \mathfrak{I},\ w, w' : \bar{\mathfrak{I}} \to \mathbb{R}^r, \tag{1.2.1}$$

wobei die Funktion $F \in \mathbb{R}^r$ stetig ist und eine stetige partielle Ableitung $\frac{\partial F}{\partial w'}$ hat. Ist $\frac{\partial F}{\partial w'}$ in einer Umgebung der Lösung $w(t)$ regulär, so ist (1.2.1) nach $w'(t)$ lokal auflösbar, d.h., (1.2.1) stellt dann eine implizite ODE dar. Dies ist nicht der Fall, wenn die Jacobi-Matrix $\frac{\partial F}{\partial w'}$ singulär ist. Sei $\frac{\partial F}{\partial w'}$ singulär im gesamten betrachten Gebiet. Dann enthält (1.2.1) auch *algebraische* Gleichungen und man nennt (1.2.1) eine DAE.

Da wir in Kapitel 3 der vorliegenden Arbeit lineare PDAEs untersuchen, wollen wir in diesem Abschnitt einige Resultate für lineare DAEs mit konstanten Koeffizienten der Form

$$X\, w'(t) + Y\, w(t) = g(t), \qquad w(0) = w_0, \qquad t \in \bar{\mathfrak{I}},\ X, Y \in \mathbb{R}^{r\times r} \tag{1.2.2}$$

zusammenstellen (vgl. z.B. [Gri86],[Bre89]). $g(t) \in \mathbb{R}^r$ ist eine gegebene Vektorfunktion. Wenn die Matrix X regulär ist, ist (1.2.2) eine ODE. Sei im folgenden X singulär. Das Lösungsverhalten einer linearen DAE (1.2.2) ist bestimmt durch das Matrixbüschel (X, Y).

Definition 1.2.1. *Ein* Matrixbüschel *(auch Matrizenbüschel, engl.* matrix pencil*)* (X, Y) *zweier Matrizen* $X, Y \in \mathbb{C}^{r\times r}$ *ist eine Familie* $(X, Y) = \{\lambda X + Y\}_{\lambda\in\mathbb{C}}$.

Definition 1.2.2. *Der* Index $\operatorname{ind}(X)$ einer Matrix X *ist die kleinste nichtnegative Zahl* k*, für die* $\operatorname{rang}(X^k) = \operatorname{rang}(X^{k+1})$ *gilt.*

Definition 1.2.3. *Man nennt das Matrixbüschel* regulär*, wenn das Polynom* $d(\lambda) := \det(\lambda X + Y)$ $(\lambda \in \mathbb{C})$ *nicht identisch verschwindet, und* singulär*, wenn* $d(\lambda) = 0\ \forall\ \lambda \in \mathbb{C}$.

Definition 1.2.4. *Sei* (X, Y) *ein reguläres Matrixbüschel und* $\lambda \in \mathbb{C}$ *mit* $\lambda X + Y$ *regulär. Der* Index $\operatorname{ind}(X, Y)$ eines regulären Matrixbüschels (X, Y) *ist definiert als* $\operatorname{ind}(X, Y) := \operatorname{ind}([\lambda X + Y]^{-1} X)$.

Der Index $\operatorname{ind}(X, Y)$ in Definition 1.2.4 ist unabhängig vom gewählten λ.

Lemma 1.2.5. [Gri86] *Sei* (X, Y) *ein reguläres Matrixbüschel mit* $X, Y \in \mathbb{C}^{r\times r}$ *und* $T \in \mathbb{C}^{r\times r}$ *eine nichtsinguläre Matrix. Dann gilt*

$$\operatorname{ind}(T\,X, T\,Y) = \operatorname{ind}(X\,T, Y\,T) = \operatorname{ind}(X, Y).$$

Für singuläre Matrixbüschel hat die DAE (1.2.2) keine oder unendlich viele Lösungen für einen gegebenen Anfangswert. Wir wollen daher im weiteren nur reguläre Matrixbüschel (X, Y) zugrunde legen. Dann kann man die Matrizen X, Y der

Weierstraß-Kronecker-Transformation unterziehen (vgl. [Wei68],[Kro90],[Gan86]). Für reguläre Matrixbüschel (X, Y) existieren reguläre Matrizen $S, T \in \mathbb{C}^{r\times r}$, so daß

$$S\,X\,T \;=\; \begin{pmatrix} I_{r_1} & 0 \\ 0 & N \end{pmatrix},\; S\,Y\,T \;=\; \begin{pmatrix} R & 0 \\ 0 & I_{r_2} \end{pmatrix},\; N_j = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{pmatrix} \in \mathbb{N}^{r_{2j}\times r_{2j}}$$

gilt, wobei $N = \text{blockdiag}(N_1, \ldots, N_k) \in \mathbb{N}^{r_2\times r_2}$ mit $r_2 = r_{21}+\ldots+r_{2k}$, $r = r_1+r_2$, und $R \in \mathbb{C}^{r_1\times r_1}$ sind. Die Matrizen $I_{r_j} \in \mathbb{N}^{r_j\times r_j}$, $j = 1, 2$, sind Einheitsmatrizen. Die Jordan-Kettenmatrix N ist eine *nilpotente* Matrix, d.h., es gibt eine Zahl $\nu \in \mathbb{N}^+$, so daß $N^\nu = 0$ gilt. Für das kleinste ν, das dies erfüllt, gilt $\text{ind}(N) = \nu$, und ν wird *Nilpotenz* oder *Riesz-Index* von N genannt. Weiterhin kann man zeigen, daß $\text{ind}(X, Y) = \text{ind}(N) = \nu$, d.h., der Index des Matrixbüschels (X, Y) ist gleich der Nilpotenz ν der Matrix N. Seien $T(x^\top, y^\top)^\top := w$ und $S(s_1^\top, s_2^\top)^\top := g$. Dann können wir (1.2.2) in das entkoppelte System

$$x'(t) + R\,x(t) = s_1(t), \qquad N\,y'(t) + y(t) = s_2(t) \tag{1.2.3}$$

transformieren. Die Gleichung für $x(t)$ ist eine ODE. Für $y(t)$ gilt:

$$\begin{aligned} y(t) &= s_2(t) - N\,y'(t) \;=\; s_2(t) - N\,s_2'(t) + N^2\,y''(t) \;=\; \ldots \\ &= s_2(t) - N\,s_2'(t) + \ldots (-1)^{\nu-1} N^{\nu-1} s^{(\nu-1)}(t) + (-1)^\nu \underbrace{N^\nu}_{=0} y^{(\nu)}(t) \\ &= \sum_{i=0}^{\nu-1} (-N)^i \frac{d^i}{d\,t^i} s_2(t). \end{aligned} \tag{1.2.4}$$

Diese Lösungsdarstellung verdeutlicht wesentliche Unterschiede der Lösung einer linearen DAE zur Lösung einer linearen ODE: Die Gleichung für $y(t)$ in (1.2.3) ist einer algebraischen Gleichung äquivalent, und $y(t)$ ist vollständig durch die Komponenten der rechten Seite $s_2(t)$ und ihre Ableitungen bestimmt. Dies bedeutet insbesondere daß das Differentialgleichungssystem (1.2.3) für $\nu > 1$ im Gegensatz zu ODEs nur für hinreichend oft stetig differenzierbare Funktionen $s_2(t)$ eine klassische Lösung hat. Es ergeben sich unterschiedlich starke Forderungen an die Differenzierbarkeit der einzelnen Komponenten der rechten Seite der DAE. Der Index ν des Matrixbüschels (X, Y) gibt folglich an, daß mindestens eine Komponente der rechten Seite $f(t)$ der DAE (1.2.2) $\nu - 1$ mal stetig differenzierbar sein muß. Weiterhin müssen die Anfangswerte w_0 des Anfangswertproblems (1.2.2) sogenannten *Konsistenzbedingungen* genügen und können daher nicht beliebig vorgegeben werden.

Mit

$$\begin{aligned} x &= \bar{T}_1 w \\ y &= \bar{T}_2 w \end{aligned} \quad \text{wenn} \quad T^{-1} = \begin{pmatrix} \bar{T}_1 \\ \bar{T}_2 \end{pmatrix} \quad \text{mit} \quad \bar{T}_1 \in \mathbb{C}^{r_1 \times r}, \bar{T}_2 \in \mathbb{C}^{r_2 \times r}\,.$$

lauten die Konsistenzbedingungen an w_0:

$$\bar{T}_2\, w_0 = \sum_{i=0}^{\nu-1} (-N)^i s_2^{(i)}(0) = y(0).$$

Anfangswerte, die dieser Bedingung genügen, werden *konsistente Anfangswerte* der DAE (1.2.2) genannt. Die Untersuchung nichtlinearer DAEs (1.2.1) ist komplizierter. Sie werden durch den von Gear 1988 ([Gea88],[Gea90]) eingeführten *Differentiationsindex* charakterisiert. Im folgendem setzen wir voraus, daß F genügend oft stetig differenzierbar ist.

Definition 1.2.6. *Die differentiell-algebraische Gleichung (1.2.1) hat den Differentiationsindex* $\mathrm{di} = \nu$*, wenn* ν *die minimale Anzahl von Differentiationen*

$$\begin{aligned} F(t, w, w') &= 0 \\ \frac{d}{dt} F(t, w, w') &= F_t + F_w w' + F_{w'} w'' \;=\; 0 \\ &\vdots \\ \frac{d^\nu}{dt^\nu} F(t, w, w') &= \frac{\partial^\nu}{\partial t^\nu} F + \ldots + F_{w'} w^{(\nu+1)} \;=\; 0 \end{aligned}$$

ist, so daß die Gleichung (1.2.1) durch algebraische Umformungen in ein explizites gewöhnliches Differentialgleichungssystem $w'(t) = f(t, w)$ *überführt werden kann. Dieses System heißt das der DAE (1.2.1)* zugrundeliegende Differentialgleichungssystem*, welches nicht unabhängig von den sogenannten* Zwangsbedingungen *(engl: constrains) betrachtet werden darf (*[Mär98]*).*

Der Differentiationsindex macht Aussagen über die Struktur der DAE und charakterisiert den algebraischen Teil der DAE. Er verdeutlicht den Unterschied von DAEs zu gewöhnlichen Differentialgleichungssystemen, indem die Überführbarkeit der einen Klasse in die andere betrachtet wird. (Eine explizite ODE hat folglich den Differentiationsindex $\mathrm{di} = 0$.)

Bemerkung 1.2.7. Der Differentiationsindex di der linearen DAE (1.2.2) stimmt mit dem Index $\nu = \mathrm{ind}(X, Y)$ des Matrixbüschels (X, Y) überein, d.h., im lineare Fall gilt $\mathrm{di} = \nu$. □

SATZ 1.2.8. [Gri86] *Die lineare DAE mit konstanten Koeffizienten der Form*

$$\begin{aligned} X_1 w'(t) + Y_1 w(t) &= g_1 \\ Y_2 w(t) &= g_2 \end{aligned} \tag{1.2.5}$$

($X_1, Y_1 \in \mathbb{R}^{n_1 \times n}, Y_2 \in \mathbb{R}^{(n-n_1)\times n}$) hat den Index 1, genau dann wenn $\binom{X_1}{Y_2} \in \mathbb{R}^{n\times n}$ eine nichtsinguläre Matrix ist.

Sei $X_1 = (I_{n_1}\ 0) \in \mathbb{R}^{n_1 \times n}$, $Y_1 = (Y_{11}\ Y_{12})$, $Y_2 = (Y_{21}\ Y_{22})$ ($Y_{ij} \in \mathbb{R}^{n_i \times n_j}$, $i, j = 1, 2$). Da I_{n_1} regulär, ist $\binom{X_1}{Y_2}$ genau dann regulär, wenn Y_{22} regulär ist. Hieraus ergibt sich

FOLGERUNG 1.2.9. *Die lineare, semi-explizite DAE*

$$\begin{aligned} w_1'(t) + Y_{11} w_1(t) + Y_{12} w_2(t) &= g_1(t) \\ Y_{21} w_1(t) + Y_{22} w_2(t) &= g_2(t) \end{aligned} \tag{1.2.6}$$

($w_i \in \mathbb{R}^{n_i}$, $Y_{ij} \in \mathbb{R}^{n_i \times n_j}$, $i, j = 1, 2$) hat den Index 1, genau dann wenn Y_{22} regulär ist.

BEMERKUNG 1.2.10. Auf andere Indexbegriffe für DAEs, w.z.B. auf den von Hairer/Lubich/Roche 1989 ([Hai89]) eingeführten *Störungsindex*, wollen wir nicht eingehen. □

KAPITEL 2

Linear-implizite Splitting-Methoden für räumlich mehrdimensionale parabolische Differentialgleichungen

2.1. Einführung

In diesem Kapitel werden wir eine Klasse von Verfahren herleiten, die zur numerischen Lösung räumlich mehrdimensionaler parabolischer Anfangs-Randwertprobleme geeignet sind, wenn der Ortsraum ein Parallelepiped ist (vgl. (1) und (2)). Wir bezeichnen diese Verfahren als *linear-implizite Splitting-Methoden*. D.h., wir ordnen sie der Klasse der Splitting-Methoden zu. Die Bezeichnung *Splitting-Methoden* findet man in der mathematischen Literatur in unterschiedlichen Zusammenhängen. Allen gemein ist das Ziel, durch geeignetes Aufsplitten komplexer Aufgabenstellungen mehrere einfacher zu lösende Probleme zu erhalten.

Ein Überblick zu *Splitting-Methoden zur numerischen Lösung von Anfangs- oder Anfangs-Randwertproblemen von Differentialgleichungen* wird im folgenden Abschnitt gegeben. Anschließend werden wir auf die Stabilität von Splitting-Methoden eingehen und verschiedene Stabilitätsbegriffe einführen. Ausgehend von diesen Stabilitätsbetrachtungen werden in Abschnitt 2.4 linear-implizite Splitting-Methoden definiert und anschließend deren Eigenschaften untersucht. In Abschnitt 2.6 werden die eingeführten Methoden anhand von Beispielen numerisch getestet.

2.2. Überblick über Splitting-Methoden

Seit vielen Jahren haben sich Splitting-Methoden für die numerische Lösung von zeitabhängigen, mehrdimensionalen partiellen Differentialgleichungen bewährt ([Hou79],[Mar90],[Mit80]). Eine typische Vorgehensweise beim Splitten ist zum Beispiel, ein räumlich mehrdimensionales Problem derart zu zerlegen, daß man nur eine Folge von eindimensionalen Berechnungen durchzuführen hat. Dies führte zur Entwicklung einer großen Anzahl von Splitting-Methoden (splitting methods), die in der Literatur auch als Teilschrittverfahren (fractional step methods) bezeichnet werden. Die bekanntesten sind die sogenannten alternierenden Richtungsverfahren (alternating direction implicit methods, *ADI methods*), die lokal eindimensionalen Methoden (locally one-dimensional methods, *LOD methods*) und die Hopscotch-Typ Methoden (*hopscotch methods*).

Zur numerischen Behandlung des Anfangs-Randwertproblems partieller Differentialgleichungen wird der bereits in Kapitel 1.1 erläuterten Linienmethode gefolgt. *(Lineare) Splitting-Methoden* sind Integrationsmethoden, die die Eigenschaft ausnutzen, daß die rechte Seite $f = f(t, w)$ des semidiskreten gewöhnlichen Differentialgleichungssystems

$$\begin{aligned} \frac{d\,w}{d\,t}(t) &= f(t, w(t)), \qquad f : [0, t_e] \times \mathbb{R}^r \longrightarrow \mathbb{R}^r, \\ w(0) &= w_0 \in \mathbb{R}^r,\ t \in [0, t_e], \end{aligned} \tag{2.2.1}$$

einer *linearen Splitting-Relation*

$$f(t, w) = \sum_{i=1}^{k} f_i(t, w), \quad f_i : [0, t_e] \times \mathbb{R}^r \longrightarrow \mathbb{R}^r \tag{2.2.2}$$

genügt. Die Funktionen f_i nennt man *Splitting-Funktionen*. Die Funktionen f und f_i hängen von der originalen partiellen Differentialgleichung und dem Typ der Semidiskretisierung ab. Hierbei haben die Splitting-Funktionen f_i i.allg. eine Jacobi-Matrix mit einfacher Struktur oder anderen besonderen Eigenschaften (vgl. Abschnitt 2.2.2 und 2.2.3).

Lineare Splitting-Methoden unterscheiden sich durch die unterschiedliche Wahl der *Splitting-Funktionen* f_i und deren Verwendung in den Verfahrensvorschriften zur Zeitintegration, den sogenannten *Splitting-Formeln*.

Bei der Formulierung der verschiedenen Splitting-Methoden betrachten wir zunächst Splitting-Formeln und gehen später auch auf die spezielle Wahl von Splitting-Funktionen ein.

2.2.1. Lineare Splitting-Formeln

In diesem Abschnitt werden wir einige bekannte lineare Splitting-Formeln und ihre Eigenschaften vorstellen. In [Hou79] wurden autonome Probleme (2.2.1) betrachtet und eine allgemeine Klasse linearer Splitting-Formeln eingeführt. Für die numerische Integration von nichtautonomen Problemen ist das Analogon durch folgende Definition gegeben.

Definition 2.2.1. *Eine s-stufige Einschrittintegrationsformel der Form*

$$\begin{aligned} v_{m+1}^{(0)} &= v_m \\ v_{m+1}^{(j)} &= v_m \ + \ \tau_m \ \sum_{l=0}^{j} \sum_{i=1}^{k} \lambda_{jli}\ f_i(t_m + c_l \tau_m, v_{m+1}^{(l)})\ ,\ j = 1(1)s\ , \\ v_{m+1} &= v_{m+1}^{(s)} \end{aligned} \tag{2.2.3}$$

ist eine lineare Splitting-Formel, *wobei* v_m *die numerische Approximation der Lösung von (2.2.1) zum Zeitpunkt* $t = t_m$ *ist und* $\tau_m = t_{m+1} - t_m$ *die Zeitschrittweite bezeichnet. Die Parameter* λ_{jli}, c_l *sind geeignet gewählte reelle Zahlen, die die Integrationsformel vollständig festlegen.*

Abhängig von diesen Parametern erhält man unterschiedliche (klassische) Konsistenz- und Stabilitätseigenschaften. Zweckmäßig ist es, die Parameter zur Ausnutzung der Splitting-Relation zu verwenden, um ein für die numerische Behandlung geeignetes und effektives Verfahren zu erhalten.

Bemerkung 2.2.2.

- Eine Stufe der Splitting-Formel, die implizit ist, soll dies nur bezüglich einer der Splitting-Funktionen sein, damit die aufzulösenden impliziten Beziehungen auf die Lösung von skalaren Gleichungen oder von Gleichungssystemen mit tridiagonaler Koeffizientenmatrix führen. D.h., wenn $\lambda_{jjl} \neq 0$ für $l \in \{1, \ldots, k\}$ gilt, dann soll $\lambda_{jji} = 0$ für alle $i \neq l$ gelten.
- Für $\lambda_{jji} = 0$, $i = 1(1)k$, ist das Schema (2.2.3) explizit. Dieser Fall tritt aber in der Theorie der Splitting-Methoden selten auf.
- Für $k = 1$ liegt kein Splitting vor, und das Schema (2.2.3) geht über in ein s-stufiges diagonal-implizites Runge-Kutta-Verfahren ([Dek84]). □

In [Lie94] (für autonome Formeln in [Hou79]) findet man folgendes Lemma:

Lemma 2.2.3. *Die lineare Splitting-Formel (2.2.3) ist von 2.Ordnung konsistent mit dem gewöhnlichen System (2.2.1), wenn die* λ_{jli} *die folgenden Bedingungen erfüllen:*

$$\sum_{l=0}^{s} \lambda_{sli} = 1 \qquad \forall\ i = 1(1)k\ , \tag{2.2.4}$$

$$\sum_{l=1}^{s} \lambda_{sli}\ c_l = \frac{1}{2} \qquad \forall\ i = 1(1)k\ , \tag{2.2.5}$$

$$\sum_{l=1}^{s} \sum_{\nu=0}^{l} \lambda_{sli}\ \lambda_{l\nu\mu} = \frac{1}{2} \qquad \forall\ i, \mu = 1(1)k\ . \tag{2.2.6}$$

Ist zusätzlich die Knotenbedingung

$$\sum_{l=0}^{j} \lambda_{jli} = c_j \qquad \forall j = 1(1)s,\ \forall i = 1(1)k \tag{2.2.7}$$

erfüllt, so sind die Stufenwerte $v_{m+1}^{(j)}$ *konsistente Approximationen 1. Ordnung an* $w(t_m + c_j\tau_m)$.

In den folgenden beiden Abschnitten werden spezielle Splitting-Methoden angegeben. Wir werden verdeutlichen, wie die geeignete Wahl der Splitting-Funktionen zu numerisch effektiven Verfahren führt, und unterscheiden dabei *Operator*- und *Gebiets*-Splitting.

2.2.2. Operator-Splitting-Methoden

Operator-Splitting-Methoden sind dadurch gekennzeichnet, daß eine Splitting-Funktion $f_i(t,w),\ i=1(1)k$, einem semidiskretisierten Operator G_i der betrachteten partiellen Differentialgleichung

$$u_t = G\,[u] = \sum_{i=1}^{k} G_i\,[u]$$

zugeordnet werden kann. Hierbei ist G_i ein Operator, der von t, x, u und Ortsableitungen von u abhängen kann. Für die numerische Integration von (2.2.1) können dann spezielle Eigenschaften der f_i, w.z.B. tridiagonale Jacobi-Matrizen oder steife und nichtsteife Anteile, berücksichtigt werden. Die am häufigsten verwendeten und auch von uns ausschließlich betrachteten Operator-Splitting-Methoden beruhen auf der Zuordnung je einer Raumdimension zu einem f_i.

Wir wollen unseren Betrachtungen parabolische Differentialgleichungen zugrunde legen, die sich durch

$$\frac{\partial u}{\partial t}\,(t,x) = \sum_{i=1}^{d} G_i\left(t, x_1, \dots, x_d, u, \frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i^2}\right), \tag{2.2.8}$$

mit Operatoren G_i, $t \in [0, t_e]$ und $x = (x_1, \dots, x_d) \in \Omega := (0,1)^d \subset \mathbb{R}^d$ darstellen lassen. Die Kopplung der Ortsableitungen in den Operatoren G_i sei linear, w.z.B. in den partiellen Differentialgleichungen der Form (1) oder (2). Weiterhin seien Anfangs- und homogene Dirichlet-Randbedingungen

$$u(0, x_1, \dots, x_d) = u_0(x_1, \dots, x_d) \qquad \forall x = (x_1, \dots, x_d) \in \Omega, \tag{2.2.9}$$

$$u(t, x_1, \dots, x_d) = 0 \qquad \forall t \in [0, t_e],\ \forall x = (x_1, \dots, x_d) \in \partial\Omega \tag{2.2.10}$$

vorgeschrieben, die untereinander einer Verträglichkeitsbedingung genügen, d.h. $u_0(x_1, \dots, x_d) = 0\ \forall x = (x_1, \dots, x_d) \in \partial\Omega$.

Wir erhalten nach der Semidiskretisierung bez. der Ortsvariablen mittels finiter Differenzen ein Anfangswertproblem der Form (2.2.1), (2.2.2) mit $k = r$, d.h.

$$\begin{aligned} \frac{d\,w}{d\,t}(t) &= f(t, w(t)) := \sum_{i=1}^{d} f_i(t, w(t)), \\ w(0) &= w_0 \in \mathbb{R}^r,\ t \in [0, t_e], \end{aligned} \tag{2.2.11}$$

$f, f_i : [0, t_e] \times \mathbb{R}^r \longrightarrow \mathbb{R}^r$, wobei f_i dem semidiskretisierten Operator G_i $(i = 1(1)d)$ entspreche und r proportional zur Zahl der inneren Ortsgitterpunkte ist (für äquidistante Ortsgitter mit M inneren Punkten je Dimension ist $r = nM^d$).

Da das semidiskrete Problem (2.2.11) i.allg. ein steifes System ist, ist die Verwendung von impliziten Methoden zur numerischen Integration zweckmäßig. Bei Verwendung von impliziten ODE-Solvern, bei denen zur Lösung der impliziten Beziehungen die volle Jacobi-Matrix $\frac{\partial f}{\partial w}$ verwendet wird, ist der Rechenaufwand je nach Feinheit der Ortsdiskretisierung sehr hoch. Daher sollte man die spezielle Struktur von (2.2.11) ausnutzen, weil die Jacobi-Matrizen $\frac{\partial f_i}{\partial w}, i = 1(1)d$, der Funktionen f_i bei geeigneter Ortsdiskretisierung stets auf tridiagonale Gestalt transformiert werden können.

Bemerkung 2.2.4. Neben den Splitting-Methoden gibt es implizite Integrationsmethoden zur Lösung von semidiskreten ODEs der Form (2.2.1), bei denen zur Lösung der auftretenden Gleichungssysteme in ähnlicher Weise die spezielle Struktur der rechten Seite f ausgenutzt wird (z.B. Verfahren, bei der die auftretenden Gleichungssysteme durch geeignete Faktorisierungen [Hou97], [EL98b] gelöst werden). □

Die Ausnutzung der Splitting-Relation (2.2.2) durch Splitting-Formeln und der damit verbundenen Reduzierung des Rechenaufwandes wollen wir zunächst anhand des zweidimensionalen Falls ($d = 2$) erläutern. Wir betrachten nun

$$u_t(t, x, y) = G_1(t, x, y, u, u_x, u_{xx}) + G_2(t, x, y, u, u_y, u_{yy}). \tag{2.2.12}$$

Dies entspricht Gleichung (2.2.8) mit $x = x_1, y = x_2$. Mit der 2-stufigen linearen Splitting-Formel

$$\begin{aligned} v_{m+1}^{(1)} &= v_m + \tau_m \left[\left(\lambda - \frac{1}{2} \right) f_1\big(t_m, v_m\big) \right. \\ &\qquad \left. + \frac{1}{2} f_1\big(t_m + \lambda\tau_m, v_{m+1}^{(1)}\big) + \lambda f_2\big(t_m, v_m\big) \right] \\ v_{m+1} &= v_m + \tau_m \left[\left(\frac{2\lambda - 1}{2\lambda} \right) f_1\big(t_m, v_m\big) + \frac{1}{2\lambda} f_1\big(t + \lambda\tau_m, v_{m+1}^{(1)}\big) \right. \\ &\qquad \left. + \frac{1}{2} f_2\big(t_m, v_m\big) + \frac{1}{2} f_2\big(t_m + \tau_m, v_{m+1}\big) \right], \end{aligned} \tag{2.2.13}$$

die bereits von zweiter Ordnung konsistent ist (vgl. Lemma 2.2.3), lassen sich durch Wahl des freien Parameters $\lambda > 0$ und durch die Spezifizierung der Splitting-Funktionen f_i verschiedene bekannte Splitting-Verfahren generieren.

Nach Einsetzen von $\lambda = \frac{1}{2}$ in (2.2.13) erhalten wir die bekannte *alternating direction implicit* (ADI)-Methode von Peaceman-Rachford ([Pea55],[Hou79],[Hun89]):

DEFINITION 2.2.5 (Peaceman-Rachford-Verfahren). *Bei Verwendung von Splitting-Funktionen f_1 und f_2, die den semidiskretisierten Operatoren G_1 und G_2 in (2.2.8) entsprechen, erhält man mit der lineare Splitting-Formel*

$$\begin{aligned} v_{m+1}^{(1)} &= v_m \quad + \tfrac{1}{2}\tau_m \left(f_1(t_m + \tfrac{\tau_m}{2}, v_{m+1}^{(1)}) + f_2(t_m, v_m) \right), \\ v_{m+1} &= v_{m+1}^{(1)} + \tfrac{1}{2}\tau_m \left(f_1(t_m + \tfrac{\tau_m}{2}, v_{m+1}^{(1)}) + f_2(t_{m+1}, v_{m+1}) \right). \end{aligned} \tag{2.2.14}$$

das Peaceman-Rachford-Verfahren.

Die Näherung $v_{m+1}^{(1)}$ ist eine von 1.Ordnung konsistente Approximation der Lösung zum Zeitpunkt $t = t_m + \frac{1}{2}\tau_m$.

BEMERKUNG 2.2.6. Das Peaceman-Rachford-Verfahren ist das historisch älteste Verfahren vom ADI-Typ. Wählen wir $\lambda = 1$ in (2.2.13) so erhalten wir das ADI-Verfahren von Douglas und Rachford [Dou56],[Hou79]. □

Wegen der speziellen Struktur der G_i und der Semidiskretisierung mittels finiter Differenzen 2.Ordnung (vgl. Abschnitt 1.1.1) nennt man $f = f_1 + f_2$ auch *5-Punkte-gekoppelt*. Bei der Semidiskretisierung wurde auf Ω ein Punktgitter gelegt. Jede Komponente von w und f ist daher mit einem inneren Gitterpunkt assoziiert. Sei die i-te Komponente von $f(t, w)$ dem Gitterpunkt X zugeordnet. Dann ist diese Komponente nur abhängig von t und von den Komponenten von w, die X und den nächsten auf den Gitterlinien liegenden Nachbarn von X zugeordnet sind. D.h., jede Komponente von $f(t, w)$ ist abhängig von einem *5-Punkte-Differenzenstern* (siehe Abbildung 2.1). Da im Operator G_1 die partiellen Ableitungen nach y bzw.

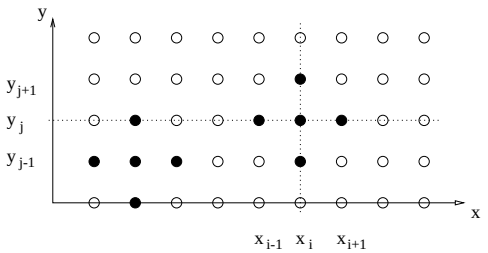

ABBILDUNG 2.1. 5-Punkte-Differenzenstern

in G_2 die nach x fehlen, sind die f_i als Semidiskretisierungen der Operatoren G_i jeweils 3-Punkte-gekoppelt. Dies bedeutet, daß sich die Jacobi-Matrizen der f_i stets durch Permutation auf tridiagonale Gestalt überführen lassen. Bei Verwendung von tridiagonalen Jacobi-Matrizen wiederum verringert sich der Rechenaufand für implizite Integrationsmethoden wesentlich. Hieraus resultiert der Vorteil der Splitting-Methoden gegenüber anderen numerischen Verfahren.

Das Peaceman-Rachford-Verfahren ist bekanntlich nur für *Zweiterm-Splitting-Funktionen* ($d = 2$) geeignet (siehe [Mit80]). Die für *Multiterm-Splitting-Funktionen* ($d \geq 2$) geeignete *Methode von Douglas* [Dou62],[Hun98a], die auch als Methode der stabilisierenden Korrekturen [Mar90] (*method of stabilizing corrections*) bezeichnet wird, ist durch

$$\begin{aligned} v_{m+1}^{(0)} &= v_m \quad + \tau_m f(t_m, v_m), \\ v_{m+1}^{(j)} &= v_{m+1}^{(j-1)} + \tfrac{1}{2}\tau_m \left(f_j(t_{m+1}, v_{m+1}^{(j)}) - f_j(t_m, v_m) \right), \ j = 1(1)d, \\ v_{m+1} &= v_{m+1}^{(d)} \end{aligned} \tag{2.2.15}$$

gegeben. Diese Methode besitzt die klassische Konsistenzordnung 2 und hat gegenüber anderen Splitting-Methoden den Vorteil, daß alle Stufenwerte $v_{m+1}^{(j)}$ konsistente Approximationen an die exakte Lösung zum Zeitpunkt t_{m+1} sind.

Wir wollen nun zwei lokal eindimensionalen Methoden vorstellen. Die von Yanenko [Yan71] entwickelte und sogenannte *lokal eindimensionale Methode von Yanenko* zur Lösung von (2.2.8) ist durch folgende r-stufige lineare Splitting-Formel gegeben:

$$\begin{aligned} v_{m+1}^{(0)} &= v_m, \\ v_{m+1}^{(j)} &= v_{m+1}^{(j-1)} + \tau_m \Big[(1-\alpha) f_j(t_m + c_{j-1}\tau_m, v_{m+1}^{(j-1)}) \\ &\qquad + \alpha f_j(t_m + c_j \tau_m, v_{m+1}^{(j)}) \Big], \ j = 1(1)d, \\ v_{m+1} &= v_{m+1}^{(d)}, \end{aligned} \tag{2.2.16}$$

wobei α ein noch freier Parameter ist. Für beliebige α ist die Formel von erster Ordnung konsistent mit (2.2.11). Die Methode wird lokal eindimensional genannt, da in der j-ten Stufe nur der semidiskretisierte eindimensionale Operator G_j benutzt wird. In Anwendungen ist der freie Parameter α im allgemeinen gleich $\frac{1}{2}$ oder 1. Erneut besitzen die zu lösenden Gleichungssysteme in jeder Stufe tridiagonale Jacobi-Matrizen.

Die von zweiter Ordnung konsistente *Trapez-Splitting-Methode* TRAPSP (trapezoidal splitting method) ist durch folgendes Schema gegeben ([Mat86],[Hun98b]):

$$\begin{aligned} v_{m+1}^{(0)} &= v_m, \\ v_{m+1}^{(i)} &= v_{m+1}^{(i-1)} \quad + \tfrac{1}{2}\tau_m f_i(t_m, v_{m+1}^{(i-1)}), & i &= 1(1)d, \\ v_{m+1}^{(d+j)} &= v_{m+1}^{(d+j-1)} + \tfrac{1}{2}\tau_m f_{d-j+1}(t_{m+1}, v_{m+1}^{(d+j)}), & j &= 1(1)d, \\ v_{m+1} &= v_{m+1}^{(2d)}. \end{aligned} \tag{2.2.17}$$

Diese Methode kann ebenfalls für beliebige $r \geq 1$ angewandt werden.

2.2.3. Gebiets-Splitting-Methoden

Zur Vervollständigung des hier gegebenen Überblicks über Splitting-Methoden wird jetzt das Prinzip von Gebiets-Splitting-Methoden vorgestellt. Darüberhinaus werden wir uns mit dieser Art von Splitting-Methoden nicht weiter beschäftigen. Gebiets-Splitting-Methoden eignen sich für partielle Differentialgleichungen mit nichtlinearen Kopplungen oder gemischten partiellen Ortsableitungen. Für diese Gleichungen kann man keine Aufteilung (2.2.8) in Operatoren, die einer Raumdimension zugeordnet werden können, vornehmen. Allerdings wird erneut eine Zuordnung einer Komponente von $f(t,w)$ des semidiskreten Problems (2.2.1) zu einem Gitterpunkt getroffen und Splitting-Funktionen $f_i(t,w)$ gebildet. Wir wollen dies an einem Beispiel vorführen.

Die *Odd-Even Hopscotch Methode* ist geeignet für Probleme (2.2.1), die aus der Semidiskretisierung einer räumlich zweidimensionalen ($d = 2$) nichtlinearen parabolischen Differentialgleichung der Form

$$u_t(t,x,y) = G(t,x,y,u,u_x,u_y,u_{xx},u_{yy}) \tag{2.2.18}$$

mit entsprechenden Anfangs- und Randbedingungen resultieren. Hierbei ist die Kopplung voll nichtlinear. Für die weiteren Betrachtungen werden Vektorfunktionen $f_\circ, f_\bullet, f_+$ und $f_\times$ eingeführt, so daß

$$f(t,w) = f_\circ(t,w) \;+\; f_\bullet(t,w) \;+\; f_+(t,w) \;+\; f_\times(t,w)\ . \tag{2.2.19}$$

Zur Definition dieser Vektorfunktionen teilt man die Gitterpunktmenge $\mathfrak{J}_\Omega$ in 4 disjunkte Teilmengen $\mathfrak{J}_{\Omega_\circ}, \mathfrak{J}_{\Omega_\bullet}, \mathfrak{J}_{\Omega_+}$ und $\mathfrak{J}_{\Omega_\times}$, wie in Figur 2.2 durch die entsprechenden Symbole angedeutet. Sei $\wedge \in \{\circ, \bullet, +, \times\}$ und bezeichne $f_\wedge^{[i]}(t,w)$ die i-te Kompo-

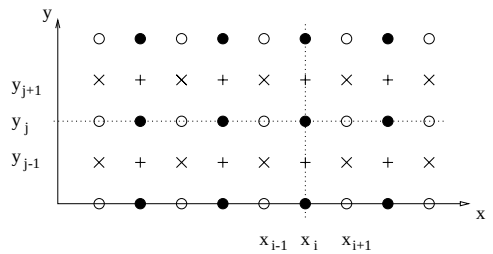

Abbildung 2.2. Vier Gitterpunktmengen

nente von $f_\wedge(t,w)$, dann werden die Funktionen $f_\wedge(t,w)$ definiert durch

$$f_\wedge^{[i]}(t,w) = \begin{cases} f^{[i]}(t,w)\ \text{, wenn der zugeordnete Gitterpunkt } \in \mathfrak{J}_{\Omega_\wedge} \\ 0 \qquad\qquad \text{, sonst.} \end{cases} \tag{2.2.20}$$

Als Splitting-Funktionen f_1 und f_2 wählen wir nun

$$f_1(t,w) = f_\circ(t,w) + f_+(t,w)\ ,$$
$$f_2(t,w) = f_\bullet(t,w) + f_\times(t,w)\ .$$

Nach Einsetzen dieser Funktionen in (2.2.13) mit $\lambda = \frac{1}{2}$ erhalten wir die *Odd-Even Hopscotch Methode*.

Wir berechnen zuerst die Komponenten von $v_{m+1}^{(1)}$, die den Gitterpunkten $\in \mathfrak{J}_{\Omega_\bullet} \cup \mathfrak{J}_{\Omega_\times}$ zugeordnet sind, und anschließend die für die Punkte $\in \mathfrak{J}_{\Omega_\circ} \cup \mathfrak{J}_{\Omega_+}$. Dabei sind nur skalare Gleichungen zu lösen. Ebenso geht man mit vertauschter Reihenfolge bei der Berechnung von v_{m+1} vor.

In unseren folgenden Betrachtungen werden wir uns, wenn wir von Splitting-Methoden sprechen, auf Operator-Splitting-Methoden beziehen, wie in Abschnitt 2.2.2 beschrieben.

2.3. Lineare Stabilität von Operator-Splitting-Methoden

In Analogie zur Stabilitätsuntersuchung für gewöhnliche Differentialgleichungen ([Dah59], [Cro79], [Dor93], [Hai96] u.a.; siehe auch [Dah85]) legen wir unseren Betrachtungen Probleme der Form (2.2.1) zugrunde, wobei die Funktion f einer Klasse von linearen Funktionen angehöre, die wir in diesem Abschnitt einführen werden. Die skalare Stabilitätstestgleichung $w' = \lambda\, w$, die für die Untersuchung linearer Stabilität von ODE-Solvern i.allg. verwendet wird, ist für die Stabilitätsuntersuchung von Splitting-Methoden nicht ausreichend. Aus der Semidiskretisierung von PDEs erhält man stets ein System gewöhnlicher Differentialgleichungen. Daher ist es sinnvoll, als Testproblem lineare Systeme gewöhnlicher Differentialgleichungen der Form

$$w'(t) = f(t,w) \ = \ A\, w(t) \ = \ \sum_{i=1}^{d} A_i\, w(t), \quad w(0) = w_0, \tag{2.3.1}$$

$w \in \mathbb{R}^r,\ A_i \in \mathbb{R}^{r\times r}$, zu betrachten. (Ein solches System erhält man z.B. bei der Semidiskretisierung der d-dimensionalen Wärmeleitungsgleichung mit homogenen Dirichlet-Randbedingungen.) Das semidiskrete Problem (2.3.1) einer parabolischen Differentialgleichung ist im allgemeinen groß und steif, wenn die Ortsgitterweiten klein sind. Dies kann durch die Voraussetzung $\mu[A_i] \le 0$ für $i = 1(1)d$ an die Matrizen A_i charakterisiert werden (siehe Beispiel 1.1.14). Für die exakte Lösung von (2.3.1) gilt

$$w(t) = e^{At}\, w_0 \qquad \Longrightarrow \qquad w(t+\tau) = e^{A\tau} w(t).$$

Es gilt $\|e^{A\tau}\| \leq e^{\mu[A]\tau}$ (vgl. [Dek84]). Wegen $\mu[A] \leq \sum_{i=1}^{d} \mu[A_i] \leq 0$ ist $\|e^{A\tau}\| \leq 1$, woraus

$$\|w(t+\tau)\| \leq \|w(t)\| \qquad \forall\ \tau > 0, \tag{2.3.2}$$

$$\lim_{\tau\mu[A_i]\to-\infty} w(t+\tau) = 0 \qquad \forall\ i = 1(1)d \tag{2.3.3}$$

folgt.

Wendet man eine lineare Splitting-Formel (2.2.3) auf das Testproblem (2.3.1) an, so erhält man mit $\tau = \tau_m$ die Beziehung

$$v_{m+1} = R(\tau A_1, \ldots, \tau A_d)\, v_m\ . \tag{2.3.4}$$

Hierbei heißt die Matrixfunktion R *Stabilitätsmatrix* und die zugeordnete rationale Funktion *Stabilitätsfunktion* der Splitting-Formel. (R entspricht der Stabilitätsfunktion von Standard-Einschrittverfahren für ODEs.) Aus (2.2.3) erhalten wir die folgenden Beziehungen[1]

$$\begin{aligned}
&R^{(0)}(\tau A_1, \ldots, \tau A_d) = I, \qquad I \in \mathbb{R}^{r\times r} \text{ Einheitsmatrix},\\
&R^{(j)}(\tau A_1, \ldots, \tau A_d) =\\
&\quad \Big(I - \tau\lambda_{jjk} A_k\Big)^{-1} \left(I + \sum_{l=0}^{j-1} \sum_{i=1}^{d} \tau\lambda_{jli} A_i\, R^{(l)}(\tau A_1, \ldots, \tau A_d)\right),\ j = 1(1)s,\\
&\quad R(\tau A_1, \ldots, \tau A_d) = R^{(s)}(\tau A_1, \ldots, \tau A_d).
\end{aligned}$$

Wünschenswert ist es, daß sich die numerischen Lösung von (2.3.1) qualitativ wie die exakte Lösung verhält. Das Analogon von Eigenschaft (2.3.2) für die numerische Lösung ist $\|v_{m+1}\| \leq \|v_m\|$. Daher fordern wir

$$\|R(\tau A_1, \ldots, \tau A_d)\| \leq 1 \qquad \forall\ \tau > 0\ , \tag{2.3.5}$$

wobei die Matrixnorm der in (2.3.2) verwendeten Vektornorm zugeordnet ist. Das Analogon von (2.3.3) für die numerische Lösung fordert von der Stabilitätsmatrix $R(.)$, daß sie

$$\lim_{\tau\mu[A_i]\to-\infty} \|R(\tau A_1, \ldots, \tau A_d)\| = 0 \qquad \forall\ i = 1(1)d \tag{2.3.6}$$

erfüllt. Die Beziehungen (2.3.5) und (2.3.6) beschreiben Stabilitätseigenschaften einer Splitting-Formel.

[1] Beweis erfolgt durch vollständige Induktion (vgl. [Lie94])

Wir führen nun zwei Klassen linearer Vektorfunktionen ein:

$$\tilde{\mathfrak{A}} := \left\{ f \;:\; f(t,w) = \sum_{i=1}^{d} f_i(t,w) = \sum_{i=1}^{d} A_i\, w,\ w \in \mathbb{R}^r, \right. \\ \left. A_i \in \mathbb{R}^{r\times r} \ \text{ und } \ \mu[A_i] \le 0 \right\}, \tag{2.3.7}$$

$$\tilde{\mathfrak{A}}_C := \left\{ f \;:\; f(t,w) = \sum_{i=1}^{d} f_i(t,w) = \sum_{i=1}^{d} A_i\, w,\ w \in \mathbb{R}^r, \right. \\ \left. A_i \in \mathbb{R}^{r\times r},\ \mu[A_i] \le 0 \text{ und } A_i \text{ vertauschbar}^2 \right\} \subset \tilde{\mathfrak{A}}. \tag{2.3.8}$$

Definition 2.3.1. *Eine auf die Klasse $\tilde{\mathfrak{A}}$ bzw. $\tilde{\mathfrak{A}}_C$ angewendete Splitting-Formel heißt* $\tilde{A}$-stabil *bzw. $\tilde{A}_C$-stabil, wenn ihre Stabilitätsmatrix R die Eigenschaft (2.3.5) hat.*

Definition 2.3.2. *Eine auf die Klasse $\tilde{\mathfrak{A}}$ bzw. $\tilde{\mathfrak{A}}_C$ angewendete Splitting-Formel heißt* $\tilde{L}$-stabil *bzw. $\tilde{L}_C$-stabil, wenn sie $\tilde{A}$-stabil bzw. $\tilde{A}_C$-stabil ist und ihre Stabilitätsmatrix R die Eigenschaft (2.3.6) hat.*

Bemerkung 2.3.3. Die Stabilitätsuntersuchungen für Splitting-Methoden in [Hou79] und [War79] beziehen sich auf die Untersuchung der Eigenschaft (2.3.5), und die dort verwendeten Methoden setzen die Vertauschbarkeit der Matrizen A_i voraus. D.h., es wird die Klasse $\tilde{\mathfrak{A}}_C$ betrachtet. □

Weiterhin werden wir für unsere Untersuchungen ein Lemma benutzen, das auf einem Ergebnis von J. von Neumann [Neu51] basiert (vgl. [Hai82]).

Lemma 2.3.4. *Sei die Norm $\|.\|$ durch ein Skalarprodukt induziert und sei $A \in \mathbb{R}^{r\times r}$ eine gegeben Matrix mit $\mu[A] \le \nu$. Sei $R(z)$ eine in $S(\nu) := \{z \in \mathbb{C} : \mathrm{Re}(z) \le \nu\}$ analytische rationale Funktion. Dann existiert $R(A)$ und in der zugeordneten Matrixnorm gilt $\|R(A)\| \le \overline{R}(\nu)$, wobei $\overline{R}(\nu) := \sup\{\, |R(z)| \;:\; z \in S(\nu)\,\}$.*

Im folgenden setzen wir voraus:

Voraussetzung 2.3.5. Die Stabilitätsmatrix $R(\tau A_1, \ldots, \tau A_d)$ kann in Faktoren $R_{(ki)}(\tau A_i)$ zerlegt werden, die jeweils nur von einem der Matrixargumente τA_i abhängen, d.h.

$$R(\tau A_1, \ldots, \tau A_d) = \prod_k \prod_{i\in\{1,\ldots,d\}} R_{(ki)}(\tau A_i). \tag{2.3.9}$$

[2] engl: commuting

BEMERKUNG 2.3.6. Stabilitätsuntersuchungen für Verfahren mit Stabilitätsmatrizen, die diese Eigenschaft nicht besitzen, sind für vertauschbare Operatoren u.a. in [Jor94] zu finden. □

Unter der Voraussetzung 2.3.5 können wir mit Lemma 2.3.4 die Norm der Stabilitätsmatrix (2.3.9) einer lineare Splitting-Formel abschätzen. Wenden wir eine lineare Splitting-Formel auf die Klasse $\tilde{\mathfrak{A}}_C$ an, so erhalten wir mit der Bezeichnung $R_{(i)}(\tau A_i) := \prod_k R_{(ki)}(\tau A_i)$

$$\|R(\tau\, A_1, \dots, \tau\, A_d)\| \;\le\; \prod_{i=1}^{d} \|R_{(i)}(\tau\, A_i)\| \;\le\; \prod_{i=1}^{d} \sup\{|R_{(i)}(z)| \;:\; \mathrm{Re}(z) \le 0\}.$$

$\tilde{A}_C$-Stabilität ergibt sich hieraus, wenn $|R_{(i)}(z)| \le 1\ \forall z \in \mathbb{C}^-,\ \forall i = 1(1)d$.

Wir betrachten nun die Klasse $\tilde{\mathfrak{A}}$. In diesem Fall kann es sein, daß die Matrizen A_i nicht vertauschbar sind. Es wird wie folgt abgeschätzt:

$$\begin{aligned}\|R(\tau\, A_1, \dots, \tau\, A_d)\| &\le \prod_{k} \prod_{i\in\{1,\dots,d\}} \|R_{(ki)}(\tau\, A_i)\| \\ &\le \prod_{k} \prod_{i\in\{1,\dots,d\}} \sup\{|R_{(ki)}(z)| \;:\; \mathrm{Re}(z) \le 0\}.\end{aligned}$$

Daher können wir i.allg. nur dann $\tilde{A}$-Stabilität nachweisen, wenn jedes der $|R_{(ki)}(z)|$ durch 1 beschränkt ist für alle $z \in \mathbb{C}^-$.

In den folgenden Beispielen wird für einige Splitting-Formeln die Faktorisierung der Stabilitätsmatrix gemäß Voraussetzung 2.3.5 und die Stabilitätseigenschaft angegeben.

BEISPIEL 2.3.7. Die Stabilitätsmatrix für das Peaceman-Rachford-Verfahren (2.2.14) ist gegeben durch

$$R(\tau A_1, \tau A_2) \;=\; \underbrace{(I - \tfrac{\tau}{2}A_2)^{-1}}_{=:R_{(12)}(\tau\, A_2)}\, \underbrace{(I + \tfrac{\tau}{2}A_1)}_{=:R_{(11)}(\tau\, A_1)}\, \underbrace{(I - \tfrac{\tau}{2}A_1)^{-1}}_{=:R_{(21)}(\tau\, A_1)}\, \underbrace{(I + \tfrac{\tau}{2}A_2)}_{=:R_{(22)}(\tau\, A_2)}.$$

Obwohl $R(.)$ die Voraussetzung 2.3.5 erfüllt, sind die Faktoren $|R_{(11)}(z)|$ und $|R_{(22)}(z)|$ nicht beschränkt. Wenden wir diese Splitting-Formel allerdings auf die Klasse $\tilde{\mathfrak{A}}_C$, so können wir Faktoren vertauschen und erhalten

$$R(\tau A_1, \tau A_2) \;=\; \underbrace{(I - \tfrac{\tau}{2}A_2)^{-1}(I + \tfrac{\tau}{2}A_2)}_{=:R_{(2)}(\tau\, A_2)}\, \underbrace{(I - \tfrac{\tau}{2}A_1)^{-1}(I + \tfrac{\tau}{2}A_1)}_{=:R_{(1)}(\tau\, A_1)}.$$

Da $|R_{(1)}(z)|, |R_{(2)}(z)| \le 1\ \forall z \in \mathbb{C}^-$ und $|R_{(1)}(-\infty)| = |R_{(2)}(-\infty)| = 1$, ist das Schema (2.2.14) $\tilde{A}_C$-stabil. □

Beispiel 2.3.8. Für die Stabilitätsmatrix der lokal eindimensionalen Methode von Yanenko (2.2.16) gilt

$$R(\tau A_1, \dots, \tau A_d) = \prod_{i=d}^{1} (I - \tau\alpha A_i)^{-1}(I + \tau(1-\alpha)A_i).$$

Die Norm dieser Matrix kann leicht abgeschätzt werden, und man erhält $\tilde{A}$-Stabilität für $\alpha \in [\frac{1}{2}, \infty)$ und $\tilde{L}$-Stabilität für $\alpha = 1$. □

Beispiel 2.3.9. Die Trapez-Splitting-Methode (2.2.17) ist $\tilde{A}_C$-stabil, weil ihre Stabilitätsmatrix durch

$$R(\tau A_1, \dots, \tau A_d) = \prod_{i=1}^{d} (I - \tfrac{\tau}{2}A_i)^{-1} \prod_{i=d}^{1} (I + \tfrac{\tau}{2}A_i)$$

gegeben ist und für vertauschbare Matrizen A_i wie folgt geschrieben werden kann:

$$R(\tau A_1, \dots, \tau A_d) = \prod_{i=1}^{d} (I - \tfrac{\tau}{2}A_i)^{-1}(I + \tfrac{\tau}{2}A_i).$$

□

Bemerkung 2.3.10. Im Gegensatz zu den oben genannten Methoden erfüllt die Stabilitätsmatrix

$$R(\tau A_1, \dots, \tau A_d) = I + \left(\prod_{i=1}^{d} \left(I - \tfrac{\tau}{2}A_i\right) \right)^{-1} \sum_{i=1}^{d} \tau A_i$$

der Splitting-Methode von Douglas (2.2.15) die Voraussetzung 2.3.5 nicht. Ausführliche Stabilitätsuntersuchungen für diese Methode sind für vertauschbare Matrizen A_i in [Hun98a] zu finden. Dort wird gezeigt, daß diese Methode $\tilde{A}_C$-stabil nur für $d \leq 2$ ist und für $d \geq 3$ Stabilität (im Sinne von Gleichung (2.3.5)) nur unter starken Einschränkungen der Schrittweite τ erhalten wird. □

Zusammenfassend stellen wir fest, daß es Splitting-Methoden zur Lösung von Systemen (2.2.11) gibt (die aus der Semidiskretisierung einer PDE entstanden sind), die entweder von erster Ordnung konsistent und $\tilde{L}$-stabil oder von Konsistenzordnung zwei, aber nur $\tilde{A}_C$-stabil sind. In allen diesen Methoden sind voll implizite Gleichungen zu lösen.

2.4. Definition linear-impliziter Splitting Methoden

Ein Ziel dieses Kapitels ist es, Splitting-Formeln zu finden, die von klassischer Konsistenzordnung zwei, $\tilde{L}$-stabil und bei denen nur lineare Gleichungssysteme mit einfacher Koeffizientenmatrix zu lösen sind. Beispiele von Diskretisierungsmethoden zur Lösung von AWPn gewöhnlicher Differentialgleichungssysteme, bei denen

nur *lineare* Gleichungen gelöst werden müssen, sind die *linear-impliziten Runge-Kutta-Methoden*, w.z.B. die Rosenbrock-Methoden oder die adaptiven Runge-Kutta-Methoden (siehe [Str92]). Diese Methoden haben gute Stabilitätseigenschaften und sind einfach zu implementieren. Sie beziehen eine Approximation an die Jacobi-Matrix direkt in die Verfahrensvorschrift ein, aber sie nutzen nicht die spezielle Splitting-Relation (2.2.2) der rechten Seite aus.

Daher wollen wir eine Klasse von linear-impliziten Splitting-Formeln einführen, die eine ähnliche Struktur wie die *adaptiven Runge-Kutta-Verfahren* haben.

DEFINITION 2.4.1. *Eine* $2d$*-stufige,* lokal eindimensionale, linear-implizite Splitting-Formel *zur Lösung von (2.2.11) ist durch die folgende Vorschrift gegeben:*

$$
\begin{aligned}
v_{m+1}^{(0)} &= v_m, \\
v_{m+1}^{(s)} &= R_0^{(s)}(a_s\tau T_s)\, v_{m+1}^{(s-1)} + \tau\, R_1^{(s)}(a_s\tau T_s) \sum_{j=0}^{s-1} b_{sj} K_{sj}, \quad s = 1(1)d, \\
v_{m+1}^{(\tilde{s})} &= R_0^{(\tilde{s})}(a_{\tilde{s}}\tau T_{2d-\tilde{s}+1})\, v_{m+1}^{(\tilde{s}-1)} \\
&\qquad + \tau\, R_1^{(\tilde{s})}(a_{\tilde{s}}\tau T_{2d+1-\tilde{s}}) \sum_{j=0}^{\tilde{s}-1} b_{\tilde{s}j} K_{2d-\tilde{s}+1,j}, \quad \tilde{s} = d{+}1(1)2d, \\
v_{m+1} &= v_{m+1}^{(2d)} \\
\textit{mit} \quad K_{ij} &= f_i(t_m{+}c_j\tau, v_{m+1}^{(j)}) - T_i\, v_{m+1}^{(j)},\ i{=}1(1)d,\ j{=}0(1)2d{-}1.
\end{aligned}
\tag{2.4.1}
$$

Hierbei sind die T_s $(s = 1(1)d)$ beliebige, konstante Matrizen. Allerdings nimmt man aus Stabilitätsgründen i. allg. an, daß sie Approximationen der Jacobi-Matrizen $\frac{\partial f_s}{\partial w}$ sind. Die skalaren Größen a_s, c_s und b_{ij} sind Parameter der Methode. Die Matrixfunktionen $R_0^{(*)}(.)$, $R_1^{(*)}(.)$ sollen rationalen Funktionen $R_0^{(*)}(z)$, $z \in \mathbb{C}$, zugeordnet sein, wobei $R_0^{(*)}(z)$ eine Approximation an die Exponentialfunktion e^z (für $z \to 0$) von mindestens erster Ordnung und

$$R_1^{(*)}(z) = \frac{R_0^{(*)}(z) - 1}{z}$$

sei. Weiterhin nehmen wir zumindest für die letzten d Stufen $(\tilde{s} = r{+}1(1)2d)$ an, daß

$$R_0^{(\tilde{s})}(z) \text{ analytisch für } \mathrm{Re}(z) \le 0 \text{ ist und } |R_0^{(\tilde{s})}(-\infty)| < \infty \text{ gilt.} \tag{2.4.2}$$

Dies bedeutet, daß die letzten d Stufen linear-implizit sind.

2.5. Eigenschaften linear-impliziter Splitting-Methoden

2.5.1. Klassische Konsistenzbetrachtungen und Ordnungsbedingungen

In diesem Abschnitt werden wir Ordnungsbedingungen für die (klassische) Konsistenzordnung bez. der Zeit für eine linear-implizite Splitting-Formel (2.4.1) herleiten.

In der linear-impliziten Splitting-Formel (2.4.1) werden Approximationen an Exponentialfunktionen von Matrizen verwendet. Man spricht daher auch von *exponentiellem Splitting*. Die Untersuchungen in [She89] und [She93] des Fehlers der numerischen Lösung, der durch das Splitting verursacht wird, zeigen, daß die maximale Approximationsordnung von exponentiellem Splitting zwei ist. Dies hat zur Folge, daß die klassische Konsistenzordnung maximal von zweiter Ordnung sein kann.

Zur Bestimmung der Ordnungsbedingungen verwenden wir die Taylor-Entwicklungen der numerischen und der exakten Lösung zum Zeitpunkt t_m und fordern Konsistenzordnung zwei (siehe Abschnitt 1.1.2), d.h.

$$\tilde{v}_{m+1} - w(t_{m+1}) \;=\; \mathcal{O}(\tau^3), \qquad \tau \to 0.$$

Aus dieser Forderung erhalten wir durch Koeffizientenvergleich Ordnungsbedingungen an die Parameter der Methode (2.4.1).

Die rationalen Funktionen $R_0^{(s)}(z)$ sind für $z \to 0$ Approximationen an e^z von mindestens erster Ordnung, d.h.

$$R_0^{(s)}(z) = 1 + z + \delta_s\, z^2 + \mathcal{O}(z^3), \qquad z \to 0,$$

wobei δ_s zunächst ein Parameter sei, den wir später spezifizieren. Weiterhin unterscheiden wir zwei Fälle für die Wahl der Matrizen T_s, nämlich $T_s = \frac{\partial f_s}{\partial w} + \mathcal{O}(\tau)$ und T_s beliebig.

1. Für den Fall $T_s = \frac{\partial f_s}{\partial w} + \mathcal{O}(\tau)$ erhalten wir für $s = 1(1)d$ unter der Voraussetzung $a_s = \sum_{j=0}^{s-1} b_{sj}$ die folgenden Ordnungsbedingungen für Ordnung zwei:

$$\begin{aligned}
&1 = a_s + a_{2d-s+1} && (\text{für} \quad f_s)\,,\\
&\frac{1}{2} = \sum_{j=0}^{s-1} b_{sj}\, c_j \;+\; \sum_{j=0}^{2d-s} b_{2d-s+1,j}\, c_j && \left(\text{für} \quad \tfrac{\partial f_s}{\partial t}\right),\\
&\frac{1}{2} = \delta_s\, a_s^2 + \delta_{2d-s+1}\, a_{2d-s+1}^2 + a_s\, a_{2d-s+1} && \left(\text{für} \quad \tfrac{\partial f_s}{\partial w} f_s\right),\\
&\frac{1}{2} = a_i\ , \quad i = 1(1)d, && \left(\text{für} \quad \tfrac{\partial f_s}{\partial w} f_i,\ i \neq s\right).
\end{aligned}$$

Durch Umformungen und Vereinfachungen eribt sich

$$\frac{1}{2} = a_s = \sum_{j=0}^{s-1} b_{sj} \ , \quad s = 1(1)2d,$$

$$1 = \delta_s + \delta_{2d-s+1}, \quad \frac{1}{2} = \sum_{j=0}^{s-1} (b_{sj} + b_{2d-s+1,j})\, c_j \ + \ \sum_{j=s}^{2d-s} b_{2d-s+1,j}\, c_j, \ s = 1(1)d.$$

2. Für beliebige Matrizen T_s erhalten wir, ebenfalls mit $a_s = \sum_{j=0}^{s-1} b_{sj}$, für $s = 1(1)d$ und $j \in \{1, \dots, d\}$ die Bedingungen

$$\begin{aligned}
& 1 = a_s + a_{2d-s+1} && (\text{für} \quad f_s)\,, \\
& \frac{1}{2} = \sum_{j=0}^{s-1} b_{sj}\, c_j \ + \ \sum_{j=0}^{2d-s} b_{2d-s+1,j}\, c_j && \left(\text{für} \quad \tfrac{\partial f_s}{\partial t}\right), \\
i < s: \quad & \frac{1}{2} = a_i \left(\sum_{j=i}^{s-1} b_{sj} + \sum_{j=i}^{2d-s} b_{2d-s+1,j} \right) && \left(\text{für} \quad \tfrac{\partial f_s}{\partial w} f_i\right), \\
& \frac{1}{2} = a_s \sum_{j=s}^{2d-s} b_{2d-s+1,j} && \left(\text{für} \quad \tfrac{\partial f_s}{\partial w} f_s\right), \\
i > s: \quad & \frac{1}{2} = a_i \sum_{j=i}^{2d-s} b_{2d-s+1,j} + (1 - a_s) \sum_{j=2d-i+1}^{2d-s} b_{2d-s+1,j} && \left(\text{für} \quad \tfrac{\partial f_s}{\partial w} f_i\right), \\
s < r: \quad & \frac{1}{2} = a_s \ , \qquad 1 \ = \ \delta_s + \delta_{2d-s+1} \ , && (\text{für} \quad T_s\, f_i, \ i \neq s)\,, \\
s = r: \quad & \frac{1}{2} = \delta_d\, a_d^2 + \delta_{d+1}\, a_{d+1}^2 + a_d\, a_{d+1} && (\text{für} \quad T_d\, f_d)\,.
\end{aligned}$$

In beiden Fällen kann das System von Ordnungsbedingungen gelöst werden, wobei einige Parameter noch immer frei wählbar sind. Diese können so gewählt werden, daß der Rechenaufwand bez. der Zahl der erforderlichen Funktionsaufrufe und der LU-Zerlegungen reduziert wird. So kann man z.B. LU-Zerlegungen wieder verwenden, wenn $R_0^{(s)}(a_s \tau T_s) = R_0^{(2d-s+1)}(a_{2d-s+1} \tau T_s)$.

Beispiel 2.5.1. Sei T_s beliebig und sei $d = 3$, $a_3 = a_4 = \frac{1}{2}$. Dann können die Ordnungsbedingungen für Ordnung zwei vereinfacht werden. Diese sind in Tabelle 2.1 angeben. Ähnliche Tabellen erhalten wir für beliebige d. □

$$a_1 = a_2 = a_3 = a_4 = a_5 = a_6 = \tfrac{1}{2},$$

s	b_{s0}	b_{s1}	b_{s2}	b_{s3}	b_{s4}	b_{s5}
1	$\frac{1}{2}$					
2	$\frac{1}{2} + b_{51}$	$-b_{51}$				
3	$\frac{1}{2} + b_{41} + b_{42}$	$-b_{41}$	$-b_{42}$			
4	$-(\frac{1}{2} + b_{41} + b_{42})$	b_{41}	b_{42}	1		
5	$-\frac{1}{2} - b_{51}$	b_{51}	b_{54}	$1 - 2b_{54}$	b_{54}	
6	$-\frac{1}{2}$	b_{65}	b_{64}	$1 - 2(b_{64} + b_{65})$	b_{64}	b_{65}

$$\delta_s + \delta_{7-s} = 1, \qquad s = 1, 2, 3,$$

$$c_3 = \frac{1}{2}, \quad 0 = b_{54}(c_2 + c_5 - 1), \quad 0 = b_{64}(c_2 + c_4 - 1) + b_{65}(c_1 + c_5 - 1).$$

TABELLE 2.1. Beispiel 2.5.1: Ordnungsbedingungen für klassische Konsistenz der Ordnung zwei einer linear-impliziten Splitting-Formel (2.4.1) mit beliebigen Matrizen T_s für $d = 3$.

2.5.2. Lineare Stabilität

Wenden wir eine linear-implizite Splitting-Formel (2.4.1) auf die Klasse $\tilde{\mathfrak{A}}$ an und wählen wir $T_s = A_s$ $(s = 1(1)d)$, so erhalten wir die Stabilitätsmatrix

$$R(\tau\, A_1, \ldots, \tau\, A_d) = \prod_{s=2d}^{d+1} R_0^{(s)}(a_s \tau A_{2d-s+1}) \prod_{s=d}^{1} R_0^{(s)}(a_s \tau A_s). \tag{2.5.1}$$

Diese Matrix genügt Voraussetzung 2.3.5, und durch die spezielle Wahl der Funktionen $R_0^{(s)}$ können wir direkt Einfluß auf die Stabilitäteigenschaften der linear-impliziten Splitting-Formel nehmen. Wählen wir z.B.

$$R_0^{(s)}(z) = \frac{1}{1 - z}, \qquad \forall\ s = 1(1)2d,$$

so erhalten wir $\tilde{L}$-Stabilität, da

$$|R_0^{(s)}(z)| < 1, \quad \forall\ z \in \mathbb{C},\ \mathrm{Re}(z) < 0, \quad \text{und} \quad R_0^{(s)}(-\infty) = 0. \tag{2.5.2}$$

Da $R_0^{(s)}(z)$ nur eine Approximation 1. Ordnung an e^z ist, liefern die Konsistenzbetrachtungen des vorangegangenen Abschnitts, daß für diese Wahl $p = 2$ nicht erreicht wird. In Abschnitt 2.5.4 wird gezeigt, daß es aber geeignete Matrixfunktionen gibt, so daß wir $\tilde{L}$-Stabilität und Konsistenzordnung zwei erhalten.

2.5.3. B-Konsistenz für lineare Probleme

In der numerischen Integration von steifen AWPn gewöhnlicher Differentialgleichungen kommt es oft vor, daß die erreichte Genauigkeit der numerischen Lösung eine geringere Ordnung hat als man nach der klassischen Konsistenzbetrachtung erwartet. Man spricht dann von einer Ordnungsreduktion ([Dek84],[SS86]).

Bei den Untersuchungen in Abschnitt 2.5.1, wurde die Steifheit des Problems (2.2.11) nicht berücksichtigt. In die $\mathcal{O}$-Terme der Abschätzung zur Bestimmung der klassischen Konsistenzordnung gehen die partiellen Ableitungen von f aus (2.2.11) ein. Auch wenn die Lösung $w(t)$ glatt im Sinne von

$$\left\| \frac{d^{\,k}\, w(t)}{d\; t^k} \right\| \le C_k \ , \ t \in [0,T], \ k = 1,2,\dots \ , \tag{2.5.3}$$

mit Konstanten C_k ist, die unabhängig von allen Größen sind, die die Steifheit des Problems (2.2.11) beeinflussen, kann $\|\frac{\partial f}{\partial w}\|$ für steife Probleme sehr groß werden. Somit sind die Aussagen der klassischen Konsistenzordnung nicht ausreichend, und man sucht Fehlerschranken, die unabhängig von der Steifheit des Problems sind.

Für diesen Zweck haben Frank, Schneid und Ueberhuber ([Fra81]) das Konzept der *B-Konsistenz* (und *B-Konvergenz*) eingeführt. Wir definieren die Klasse $\mathfrak{F}_\nu$ als die Probleme (2.2.11), deren rechte Seite die einseitigen Lipschitz-Konstante ν besitzt.

Definition 2.5.2. *Ein Einschrittverfahren heißt* B-konsistent *der Ordnung q auf der Klasse $\mathfrak{F}_\nu$, wenn für den lokalen Fehler*

$$\|\tilde{v}_{m+1} - w(t_{m+1})\| \le C\,\tau^{q+1} \ , \qquad \forall\ \tau \in (0,\tau^*], \tag{2.5.4}$$

gilt, wobei die reellen Größen C und $\tau^* > 0$ unabhängig von der Steifheit von $\mathfrak{F}_\nu$ sind.

Wir nehmen an, daß die im folgenden betrachteten Probleme der folgenden Voraussetzung genügen:

Voraussetzung 2.5.3. Die Lösung $w(t)$ von (2.2.11) erfülle (2.5.3) und

$$\left\| \frac{d^{\,k}}{d\,t^k} f_i(t,w(t)) \right\| \le L_{ik} \ , \qquad t \in [0,T], \ i = 1(1)d, \ k = 1,2,\dots \ , \tag{2.5.5}$$

mit von der Steifheit unabhängigen Konstanten L_{ik}.

Wir betrachten die Klasse linearer Probleme (2.3.1) mit $\mu[A_i] \le 0$ $(i = 1(1)d)$, d.h. die Klasse $\mathfrak{A}$. Voraussetzung 2.5.3 sichert, daß $\left\| A_i \frac{d^{\,k}}{dt^k} w(t) \right\| \le L_{ik}$ für $t \in [0,T]$, $i = 1(1)d$, $k = 1,2,\dots$. Die Norm des lokalen Fehlers einer linear-impliziten

Splitting-Formel (2.4.1) angewendet auf diese Klasse ist mittels Taylor-Entwicklung von $w(t_{m+1})$ gegeben durch

$$\begin{aligned}&\|\tilde{v}_{m+1} - w(t_{m+1})\| \\ &\leq \left\| R(\tau A_1, \dots, \tau A_d) w(t_m) - \left(I + \tau \sum_{i=1}^{d} A_i \right) w(t_m) \right\| + \frac{\tau^2}{2} C_2,\end{aligned}$$

mit C_2 aus (2.5.3). Um B-Konsistenzordnung eins zu erhalten müssen wir die $R_0^{(i)}(.)$, $i = 1(1)2d$, (siehe (2.5.1)) so wählen, daß

$$\left\| \left[\prod_{s=2d}^{d+1} R_0^{(s)}(a_s \tau A_{2d-s+1}) \prod_{s=d}^{1} R_0^{(s)}(a_s \tau A_s) \; - \; \left(I + \tau \sum_{i=1}^{d} A_i \right) \right] w(t_m) \right\| \leq \tau^2 \, \bar{C},$$

wobei $\bar{C}$ eine von der Steifheit des Problems unabhängige Konstante ist.

Folgerung 2.5.4. *Unter der Voraussetzung 2.5.3 ist eine linear-implizite Splitting-Formel (2.4.1) für $d = 2$ B-konsistent der Ordnung eins auf der Klasse $\tilde{\mathfrak{A}}$, wenn ihre Stabilitätsmatrix die Form*

$$R(\tau A_1, \tau A_2) = \left(I - \tfrac{\tau}{2} A_1\right)^{-1} \left(I - \tfrac{\tau}{2} A_2\right)^{-1} \left(I + \tfrac{\tau}{2} A_2\right) \left(I + \tfrac{\tau}{2} A_1\right) \tag{2.5.6}$$

oder

$$R(\tau A_1, \tau A_2) = \left(I - \tfrac{\tau}{4} A_1\right)^{-2} \left(I - \tfrac{\tau}{4} A_2\right)^{-2} \left(I + \tfrac{\tau}{4} A_2\right)^{2} \left(I + \tfrac{\tau}{4} A_1\right)^{2} \tag{2.5.7}$$

besitzt.

Beweis. Für R aus (2.5.6) gilt:

$$\begin{aligned}\big[R(\tau A_1, \tau A_2) - (I + \tau(A_1 + A_2))\big] w(t_m) &= \left(I - \tfrac{\tau}{2} A_1\right)^{-1} \left(I - \tfrac{\tau}{2} A_2\right)^{-1} \\ \underbrace{\Big[\left(I + \tfrac{\tau}{2} A_2\right) \left(I + \tfrac{\tau}{2} A_1\right) - \left(I - \tfrac{\tau}{2} A_2\right) \left(I - \tfrac{\tau}{2} A_1\right) (I + \tau(A_1 + A_2)) \Big] w(t_m)}& \\ = \Big[-\tfrac{\tau^2}{2} \big(A_1^2 + A_1 A_2 + A_2 A_1 + A_2^2\big) - \tfrac{\tau^3}{4} A_2 A_1 (A_1 + A_2) \Big] w(t_m)& \\ = -\tfrac{\tau^2}{2} w''(t_m) - \tfrac{\tau^3}{4} A_2 A_1 w'(t_m)&\end{aligned}$$

Nach Lemma 2.3.4 ist $\| \left(I - \frac{\tau}{2} A_1\right)^{-1} \| \leq 1$ und $\| \left(I - \frac{\tau}{2} A_2\right)^{-1} \frac{\tau}{2} A_2 \| \leq 1$, da $\mu[A_1], \mu[A_2] \leq 0$. Wegen $\|A_1 w'(t_m)\| \leq L_{11}$, folgt

$$\|(R - (I + \tau A) w(t_m)\| \leq \tfrac{\tau^2}{2} (C_2 + L_{11}),$$

wobei C_2 und L_1 nach Voraussetzung von der Steifheit unabhängig sind. Für R aus (2.5.7) man kann die Behauptung analog zeigen. □

BEMERKUNG 2.5.5.

1. Splitting-Methoden, deren Stabilitätsmatrizen die Form (2.5.6) oder (2.5.7) haben, sind $\tilde{A}_C$-stabil.
2. Die Aussagen von Folgerung 2.5.4 können auch auf Systeme der Form

$$w'(t) = A_1\, w(t) + g_1(t) + A_2\, w(t) + g_2(t) \tag{2.5.8}$$

erweitert werden, wobei $w, g_i \in \mathbb{R}^r$, $A_i \in \mathbb{R}^{r\times r}, \mu[A_i] \leq 0, i = 1, 2$. Hierbei können die Funktionen $g_i(t)$ z.B. die Randwerte der zugrundeliegenden PDE mit inhomogenen Randbedingungen enthalten (siehe [Hun89], [Hun92], [Hun98b]).
3. Wir konnten nur Splitting-Methoden finden, die $\tilde{L}$-stabil und B-konsistent einer Ordnung größer Null sind, unter der strengen Voraussetzung, daß $A_2\, A_1\, w(t)$ gleichmäßig beschränkt ist. In [Hun98b] wird erwähnt, daß dies z.B. für periodische Randbedingungen gilt.□

2.5.4. Beispiele linear-impliziter Splitting-Methoden

Wir werden jetzt spezielle Methoden angeben, die wir für unsere numerischen Tests in Abschnitt 2.6 verwenden. Eine linear-implizite Splitting-Formel ist nicht nur durch die Parameter a_s, c_s und b_{ij} in (2.4.1) sondern auch durch die spezielle Wahl der Matrixfunktionen $R_0^{(s)}(.)$ bestimmt. Wir geben zunächst einige Matrixfunktionen an, die Approximationen mindestens erster Ordnung an e^z sind.

$$\text{F1:}\quad R_0(a_s\tau T_s) = (I - \gamma\, a_s\tau T_s)^{-2}\,(I + (1-2\gamma)a_s\tau T_s)\,,\quad \gamma = 1 - \tfrac{\sqrt{2}}{2},$$
$$\Longrightarrow\quad R_1(a_s\tau T_s) = (I - \gamma\, a_s\tau T_s)^{-2}\,\left(I - \gamma^2 a_s\tau T_s\right), \tag{2.5.9}$$

$$\text{F2:}\quad R_0(a_s\tau T_s) = \left(I - \tfrac{a_s\tau}{2}T_s\right)^{-1}\left(I + \tfrac{a_s\tau}{2}T_s\right)$$
$$\Longrightarrow\quad R_1(a_s\tau T_s) = \left(I - \tfrac{a_s\tau}{2}T_s\right)^{-1}, \tag{2.5.10}$$

$$\text{F3:}\quad R_0(a_s\tau T_s) = I + a_s\tau T_s$$
$$\Longrightarrow\quad R_1(a_s\tau T_s) = I, \tag{2.5.11}$$

$$\text{F4:}\quad R_0(a_s\tau T_s) = (I - a_s\tau T_s)^{-1}$$
$$\Longrightarrow\quad R_1(a_s\tau T_s) = (I - a_s\tau T_s)^{-1}\,. \tag{2.5.12}$$

Durch Kombinationen von Matrixfunktionen mit den Splitting-Formeln der folgenden Beispiele erhalten wir spezielle Splitting-Methoden.

BEISPIEL 2.5.6. Mit *LISM1* bezeichnen wir folgende Formel. Unter der Voraussetzung, daß T_s ein Approximation an die Jacobi-Matrix $\frac{\partial f_s}{\partial w}$ ist, ist diese von klassischer Konsistenzordnung zwei. Sei $s = 1(1)d$, $\tilde{s} = 2d - s + 1$, $v_{m+1}^{(0)} = v_m$,

$v_{m+1} = v_{m+1}^{(2d)}$:

LISM1:

$$\begin{aligned} v_{m+1}^{(s)} &= R_0\big(\tfrac{\tau}{2}T_s\big)v_{m+1}^{(s-1)} + \tfrac{\tau}{2}R_1\big(\tfrac{\tau}{2}T_s\big)\Big(f_s(t_m, v_m) - T_s v_m\Big), \\ v_{m+1}^{(\tilde{s})} &= R_0(\tfrac{\tau}{2}T_s)v_{m+1}^{(\tilde{s}-1)} + \tfrac{\tau}{2}R_1(\tfrac{\tau}{2}T_s)\Big(f_s(t_m+\tau, v_{m+1}^{(d)}) - T_s v_{m+1}^{(d)}\Big). \end{aligned} \tag{2.5.13}$$

□

Beispiel 2.5.7. *LISM2* ist eine spezielle Formel aus Beispiel 2.5.1, in dem T_s beliebig sein kann. Sei $s = 1(1)d$, $\tilde{s} = 2d - s + 1$, $v_{m+1}^{(0)} = v_m$, $v_{m+1} = v_{m+1}^{(2d)}$.

$$\begin{aligned} \text{LISM2:}\quad v_{m+1}^{(s)} &= R_0\big(\tfrac{\tau}{2}T_s\big)v_{m+1}^{(s-1)} + \tfrac{\tau}{2}R_1\big(\tfrac{\tau}{2}T_s\big)\,K_{s,s-1}, \\ v_{m+1}^{(\tilde{s})} &= R_0(\tfrac{\tau}{2}T_s)v_{m+1}^{(\tilde{s}-1)} + \tau R_1(\tfrac{\tau}{2}T_s)\left(-\tfrac{1}{2}K_{s,s-1} + K_{s,r}\right), \\ K_{s,s-1} &= f_s(t_m + \tfrac{\tau}{2}, v_{m+1}^{(s-1)}) - T_s v_{m+1}^{(s-1)}, \\ K_{s,r} &= f_s(t_m + \tfrac{\tau}{2}, v_{m+1}^{(d)}) - T_s\, v_{m+1}^{(d)}. \end{aligned} \tag{2.5.14}$$

Wir wollen bemerken, daß bereits berechnete Werte $K_{s,s-1}$ in den Stufen $2d-s+1$ wieder verwendet werden können. □

Wenn wir die Methoden LISM1 oder LISM2 mit der rationalen Funktion F1 kombinieren, erhalten wir die $\tilde{L}$-stabilen Methoden *LISM1F1* oder *LISM2F1*. Kombiniert mit F2 erhalten wir die $\tilde{A}$-stabilen Methoden *LISM1F2* und *LISM2F2*.

Bemerkung 2.5.8. Die Methoden LISM1 und LISM2 geben auch Möglichkeiten, Funktionsaufrufe, Jacobi-Matrixberechnungen, Matrix-mal-Vektor-Multiplikationen oder LU-Zerlegungen parallel durchzuführen. □

Beispiel 2.5.9. *LISM3F3F4* oder *LTRAP* bezeichne die folgende Formel: Seien $s = 1(1)d$, $\tilde{s} = 2d - s + 1$, $v_{m+1}^{(0)} = v_m$, $v_{m+1} = v_{m+1}^{(2d)}$. Für $R_0^{(s)}$ verwenden wir die Funktion F3, d.h., die ersten d Stufen sind explizit. Sei $T_s = \frac{\partial f_s}{\partial w}(t_m+\tau, v_{m+1}^{\tilde{s}-1}) + \mathcal{O}(\tau)$ und $R_0^{(\tilde{s})}$ die Funktion F4. Folglich sind nur die letzten d Stufen implizit.

LTRAP:

$$\begin{aligned} v_{m+1}^{(s)} &= v_{m+1}^{(s-1)} + \tfrac{\tau}{2}f_s(t_m, v_{m+1}^{(s-1)}), \qquad s = 1(1)d,\ \tilde{s} = 2d - s + 1, \\ v_{m+1}^{(\tilde{s})} &= v_{m+1}^{(\tilde{s}-1)} + \tfrac{\tau}{2}\left(I - \tfrac{\tau}{2}T_s\right)^{-1} f_s(t_m + \tau, v_{m+1}^{(\tilde{s}-1)}). \end{aligned} \tag{2.5.15}$$

Diese Methode entspricht der Trapez-Splitting-Methode (2.2.17), wenn dort die impliziten Beziehungen mit einer Newton-Iteration mit startend mit dem Stufenwert der vorherigen Stufe gelöst werden. □

BEISPIEL 2.5.10. Für $d = 2$ sind die Methoden LISM1F2, LISM2F2, LTRAP und die Trapez-Splitting-Methode (2.2.17) sowohl B-konsistent erster Ordnung als auch $\tilde{A}_C$-stabil auf der Klasse $\mathfrak{A}_C$. □

2.6. Numerische Beispiele

In diesem Abschnitt werden einige numerische Ergebnisse für die in Abschnitt 2.4 bzw. 2.5.4 eingeführten linear-impliziten Splitting-Methoden und die Trapez-Splitting-Methode (2.2.17) vorgestellt. Die impliziten Beziehungen in TRAPSP werden durch vereinfachte Newton-Iteration gelöst. Die Methoden wurden unter Verwendung einer Schrittweitensteuerung für die Zeitintegration mittels Richardson-Extrapolation (siehe z.B. [Str95]) implementiert.

Am Ende des numerischen Zeitintegrationsprozesses wird die numerische Lösung mit der exakten Lösung $u_h(t)$ der PDE in den Gitterpunkten des fest gewählten, äquidistanten Ortsgitters mit Gitterweite $h = \frac{1}{M+1}$ verglichen. In den Abbildungen der nachfolgenden Beispiele wird die Rechenzeit in Abhängigkeit vom Logarithmus des relativen Fehlers

$$ERR = \sqrt{\frac{1}{r}\sum_{i=1}^{r}\left(\frac{v_i - u_h(t_e)_i}{1 + |u_h(t_e)_i|}\right)^2}$$

zum Endzeitpunkt t_e dargestellt, wobei v die numerische Lösung in t_e ist. In den Beispielen wurde der Quellterm, eingeschränkt auf das Ortsgitter, zu gleichen Teilen auf die Splitting-Funktionen aufgeteilt, und mit verschiedenen Toleranzen für die Richardson-Extrapolation gerechnet. Die Verfahren wurden in C implementiert. Die Rechnungen wurden auf einer Workstation (SUN SPARC ULTRA 1, 128 MB) durchgeführt.

BEISPIEL 2.6.1. Wir betrachten das zweidimensionale Problem

$$u_t \;=\; \tfrac{1}{2}x(1-x)u_{xx} + \tfrac{1}{2}(1+\alpha\, x)\, y(1-y)u_{yy} - (1-x)\alpha u$$

auf $\Omega = [0,1]^2$ und $t \in [0,1]$. Die Anfangsbedingung und Dirichlet-Randbedingungen seien so gewählt, daß wir als exakte Lösung

$$u(t,x,y) \;=\; e^{-(2+\alpha)t}x(1-x)y(1-y)$$

erhalten. Da die exakte Lösung der PDE ein Polynom in den Ortsvariablen vom Grad kleiner 4 ist, wird durch die Semidiskretisierung kein Ortsfehler in die ODE (2.2.11) eingeführt. Somit ist die exakte Lösung der ODE gleich der exakten Lösung der PDE in den Gitterpunkten. Die semidiskrete ODE ist ein lineares System der Form (2.3.1). Wir betrachten zwei Fälle, $\alpha = 0$ und $\alpha = 100$.

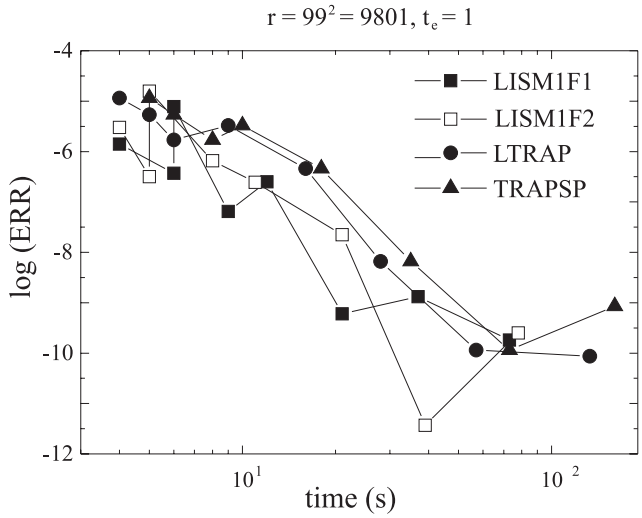

ABBILDUNG 2.3. Beispiel 2.6.1, $\alpha = 0$, $r = 99^2 = 9801$, $t_e = 1$

Für $\alpha = 0$ sind die Matrizen A_1 und A_2 vertauschbar, da wir ein äquidistantes Ortsgitter verwendet haben. In Abbildung 2.3 sehen wir, daß alle dargestellten Methoden vergleichbare Eigenschaften haben. Die $\tilde{L}$- und $\tilde{A}$-stabilen Methoden LISM1F1 und LISM1F2 sind für dieses lineare Beispiel bei gleicher Genauigkeit schneller als die $\tilde{A}_C$-stabilen Methoden TRAPSP und LTRAP.

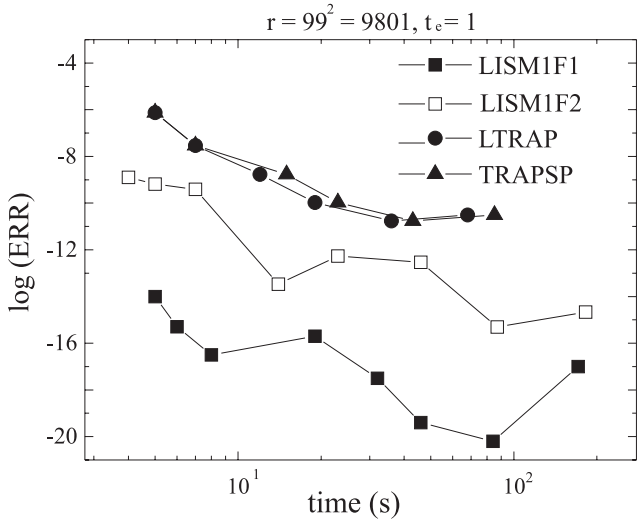

ABBILDUNG 2.4. Beispiel 2.6.1, $\alpha = 100$, $r = 99^2 = 9801$, $t_e = 1$

Für $\alpha = 100$ (siehe Abbildung 2.4) sind die Matrizen A_1 und A_2 nicht vertauschbar. Die Voraussetzung für die Stabilität von LTRAP und TRAPSP sind

nicht erfüllt. Die Methoden LISM1F1 und LISM1F2 benötigen diese Voraussetzungen nicht. Dies scheint der Grund für die besseren Ergebnisse der $\tilde{A}$-stabilen Methode LISM1F2 zu sein. Weiterhin zahlt sich für dieses Beispiel die $\tilde{L}$-Stabilität der Methode LISM1F1 aus. □

Beispiel 2.6.2. Wir betrachten die zweidimensionale, nichtlineare Diffusionsgleichung

$$u_t \;=\; e^u(u_{xx}+u_{yy}) + u\,\left(2\pi^2\,e^u - 1\right), \tag{2.6.1}$$

mit $\Omega = [0,1]^2$ und $t \in [0,10]$. Die Anfangsbedingung und Dirichlet-Randbedingungen seien so gewählt, daß wir als exakte Lösung

$$u(t,x,y) \;=\; \sin(\pi\,x)\sin(\pi\,y)\,e^{-t} \tag{2.6.2}$$

erhalten. Im Gegensatz zur in Abschnitt 1.1 beschriebenen Semidiskretisierung

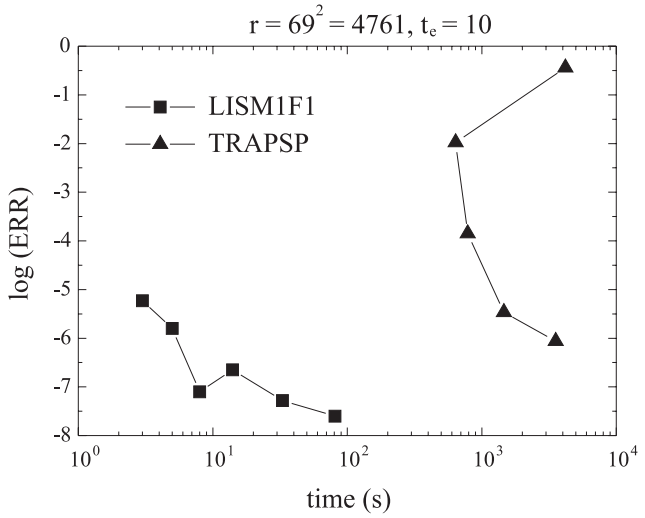

Abbildung 2.5. Beispiel 2.6.2, $r = 69^2 = 4761$, $t_e = 10$

berechnen wir für unsere Tests den analytischen Ausdruck des Ortsfehlers, der durch die Semidiskretisierung entsteht. Sei $x_i = i\,h, y_j = j\,h$ $(i,j = 1(1)M)$ und $u_{i,j}(t) = u(t,x_i,y_j)$. Es ist:

$$\begin{aligned} u_{xx}(t,x_i,y_j) &= -\pi^2\sin(\pi\,x_i)\sin(\pi\,y_j)\,e^{-t}, \\ &\sin(\pi\,(x_i \pm h)) = \sin(\pi\,x_i)\cos(\pi\,h) \pm \cos(\pi\,x_i)\sin(\pi\,h), \\ &\cos(\pi\,h) - 1 = -2\sin^2\left(\tfrac{\pi\,h}{2}\right) \\ \Longrightarrow\quad r_{i,j}^{(x)} &:= u_{xx}(t,x_i,y_j) - \tfrac{1}{h^2}\left(u_{i-1,j}(t) - 2u_{i,j}(t) + u_{i+1,j}(t)\right) \\ &= \left(-\pi^2\sin(\pi\,x_i) - \tfrac{2}{h^2}(\cos(\pi\,h)-1)\sin(\pi\,x_i)\right)\sin(\pi\,y_j)\,e^{-t} \end{aligned}$$

$$
\begin{aligned}
&= \left(-\pi^2 + \tfrac{4}{h^2}\sin^2\left(\tfrac{\pi\,h}{2}\right)\right)\sin(\pi\,x_i)\sin(\pi\,y_j)\,e^{-t} \\
&= \left(-\pi^2 + \tfrac{4}{h^2}\sin^2\left(\tfrac{\pi\,h}{2}\right)\right)u_{i,j}(t) \\
\Longrightarrow \quad r_{i,j}^{(y)} &:= u_{yy}(t,x_i,y_j) - \tfrac{1}{h^2}\left(u_{i,j-1}(t) - 2u_{i,j}(t) + u_{i,j+1}(t)\right) = r_{i,j}^{(x)}
\end{aligned}
$$

Somit ist durch Addition von $2e^{u_{i,j}(t)}\left(-\pi^2 + \frac{4}{h^2}\sin^2\left(\frac{\pi\,h}{2}\right)\right)u_{i,j}(t)$ für unsere Tests zu der dem Gitterpunkt (x_i, y_j) zugeordneten Komponente der rechten Seite der ODE (2.2.11) die exakte Lösung der ODE gleich der exakten Lösung der PDE (2.2.8), eingeschränkt auf das Gitter. D.h., ERR hängt nicht von einem Ortsfehler ab. Andererseits würde der Ortsfehler den Gesamtfehler bei kleinen Toleranzen für die Schrittweitensteuerung dominieren. Abbildung 2.5 zeigt, daß die linear-implizite $\tilde{L}$-stabile Methode LISM1F1 genauer und stabiler ist als die voll implizite $\tilde{A}_C$-stabile Methode TRAPSP. □

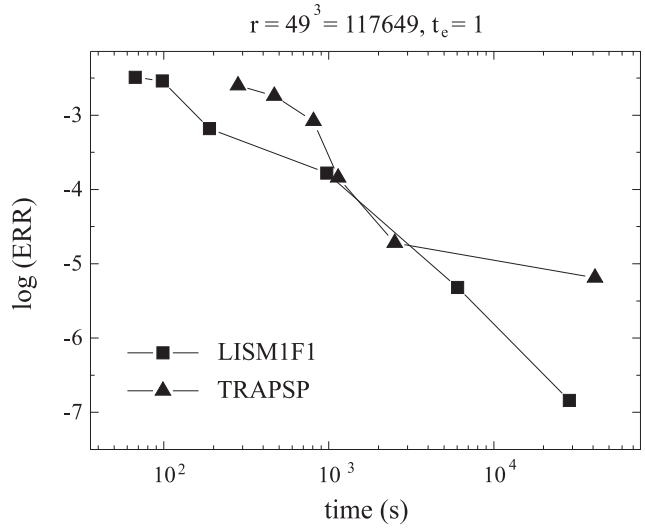

ABBILDUNG 2.6. Beispiel 2.6.3, $r = 49^3 = 117649$, $t_e = 10$

BEISPIEL 2.6.3. Als drittes Beispiel wählen wir die dreidimensionale PDE

$$
\begin{aligned}
u_t = u_{xx} + u_{yy} + u_{zz} + e^t\big(&x(1-x)y(1-y)z(1-z) \\
&+ 2y(1-y)z(1-z) + 2x(1-x)z(1-z) + 2x(1-x)y(1-y)\big)
\end{aligned}
$$

mit $\Omega = [0,1]^3$, $t \in [0,10]$ und homogenen Dirichlet-Randbedingungen. Die Anfangsbedingung sei so gewählt, daß die exakte Lösung

$$u(t,x,y,z) = e^t\,x\,(1-x)\,y\,(1-y)\,z\,(1-z)$$

ist. Bei der Semidiskretisierung tritt erneut kein Ortsfehler auf. Abbildung 2.6 demonstriert, daß beide Methode LISM1F1 und TRAPSP, ähnlich und zufriedenstellend für sehr große Probleme arbeiten. □

Beispiel 2.6.4. Sei das zweidimensionale Problem

$$u_t = u_{xx} + u_{yy} + e^t \left(x\,(1-x)\,y\,(1-y)\,(16+y) \right.$$
$$\left. + 2\,y\,(1-y)\,(16+y) + 6\,x\,(1-x)\,(5+y) \right)$$

für $\Omega = [0,1]^2$, $t \in [0,1]$ gegeben. Die Funktion

$$u(t,x,y) = e^t\,x\,(1-x)\,y\,(1-y)\,(16+y)$$

ist die Lösung des Beispiels, wenn wir $u(0,x,y) = x(1-x)y(1-y)(16+y)$ und homogene Dirichlet-Randbedingungen vorschreiben. Abbildung 2.7 zeigt, daß LISM1F1 weniger genau sind als LTRAP order TRAPSP. Wir nehmen an, daß der Grund dafür die bessere B-Konsistenz der Trapez-Splitting-Methode auf der Problemklasse (2.5.8) ist (siehe Bemerkung 2.5.5). □

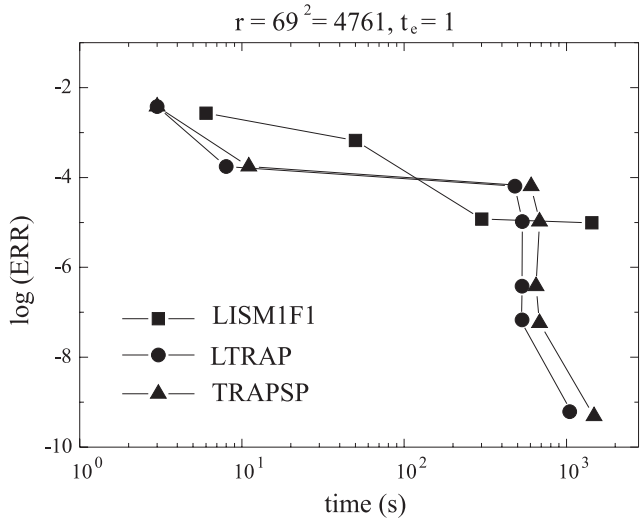

Abbildung 2.7. Beispiel 2.6.4, $r = 69^2 = 4761$, $t_e = 1$

2.7. Zusammenfassung

Zur geeigneten Charakterisierung der Stabiltätseigenschaften von Splitting-Methoden wurden die Begriffe der $\tilde{A}_C$- und $\tilde{A}$-Stabilität (bzw. $\tilde{L}_C$- und $\tilde{L}$-Stabilität) definiert. Wir haben eine Klasse von linear-impliziten Splitting-Methoden von klassischer Konsistenzordnung zwei entwickelt, die $\tilde{A}$- und $\tilde{L}$-stabil sind (unabhängig von der Vertauschbarkeit der Matrizen A_i). Diese Methoden sind effektiv für die numerische Lösung von räumlich mehrdimensionalen parabolischen Differentialgleichungen auf Parallelepipeden, da hierfür nur die Lösung lineare Gleichungssysteme mit tridiagonalen Koeffizientenmatrizen erforderlich ist und der Rechenaufwand erheblich reduziert wird. Obwohl diese, im Gegensatz zu der Trapez-Splitting-Methode TRAPSP (welche nur $\tilde{A}_C$-stabil ist), nur eine B-Konsistenzordnung Null besitzen, zeigen unsere Tests, daß für gewisse Probleme die $\tilde{L}$-Stabilität (und klassische Ordnung zwei) von größerer Bedeutung sein kann als die B-Konsistenz.

KAPITEL 3

Lineare partielle differentiell-algebraische Systeme

3.1. Einführung

Während wir in Kapitel 2 die numerische Lösung von Anfangs-Randwertproblemen einer Klasse von räumlich mehrdimensionalen parabolischen Differentialgleichungen mittels linear-impliziter Splitting-Methoden untersucht haben, wollen wir uns in diesem Kapitel mit einer Klasse partieller Differentialgleichungssystemen beschäftigen, die sich aus unterschiedlichen Typen von Gleichungen zusammensetzen.

Die mathematische Modellierung komplexer Zusammenhänge führt häufig auf Systeme, die zum Beispiel aus einer Kopplung

- parabolischer und elliptischer Differentialgleichungen oder
- elliptischer und gewöhnlicher Differentialgleichungen oder
- parabolischer und gewöhnlicher Differentialgleichungen und algebraischer Gleichungen

bestehen. Es gibt viele Möglichkeiten, solche Gleichungen zu kombinieren. Diese Systeme müssen für eine eindeutige Lösung durch Anfangs- und geeignete Randbedingungen ergänzt werden. Weitere hier einzuordnende Problemklassen sind z.B. Systeme zeitabhängiger partieller Differentialgleichungen und DAEs aus dem Bereich der Simulation von Mehrkörpersystemen, die über Kontaktbedingungen gekoppelt sind ([Arn98c]), oder Systeme hyperbolischen Typs, deren Randbedingungen durch zeitabhängige differentiell-algebraische Systeme bestimmt sind und bei der Modellierung von Netzwerken auftreten ([Gün98]).

Die mathematische Untersuchung des Lösungsverhaltens gekoppelter Systeme von PDEs, ODEs und DAEs ist ein noch junges Forschungsgebiet. Es gibt relativ wenige Arbeiten hierzu. Derartig gekoppelte partielle Differentialgleichungssysteme werden als partielle Differentialgleichungen mit Zwangsbedingungen (engl: *constrained partial differential equations - constrained PDEs*) oder auch als *partielle differentiell-algebraische Systeme* (engl: *partial differential algebraic equations - PDAEs*) bezeichnet. Die Bezeichnung PDAE, die wir im folgenden verwenden wollen, resultiert u.a. daraus, daß ein solches System häufig in eine Folge von differentiell-algebraischen Systemen (engl.: differential algebraic equations - DAEs)

überführt werden kann. PDAEs wurden von verschiedenen Gesichtspunkten aus betrachtet. Erste Untersuchungen linearer PDAEs findet man bei Campbell/Marszalek ([Cam95], [Cam96a], [Cam96b], [Mar97]). Arnold betrachtet im Zusammenhang mit linearen PDAEs den von ihm definierten gleichmäßigen Störungsindex bez. der Semidiskretisierung des Problems ([Arn98a],[Arn98b]). Spezielle PDAEs wurden von Simeon (elastische Mehrkörpersysteme [Sim96],[Arn98c]) und Weickert (Navier-Stokes Gleichungen [Wei96]) betrachtet und dabei mittels der Linienmethode auf DAEs überführt und gelöst.

In diesem Kapitel wird sich darauf beschränkt, nur lineare PDAEs zu betrachten, die genügend oft stetig differenzierbare Lösungen besitzen. Wir wollen anhand der Laplace- und Fouriertransformation einen Weg zur Charakterisierung von PDAEs aufzeigen. In [Mar97] wird darauf hingewiesen, daß PDAEs auch unstetige und Impulslösungen haben können. Es stellt sich dann die Frage, wie partielle Ableitungen in der PDAE zu interpretieren sind. Es werden zunächst zwei Anwendungsbeispiele kurz vorgestellt.

Beispiel 3.1.1. Die Modellierung einer *Populationsdynamik* von n Spezies in Abhängigkeit von m gleichmäßig verteilten Nahrungsquellen führt auf die PDAE ([Leu89],[Wal70])

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

wobei $u = (u_1, \ldots, u_n)^\top$ der Dichte-Vektor der Spezies und $v = (v_1, \ldots, v_m)^\top$ der Dichte-Vektor der Nahrungsquellen seien. Die Größen $D > 0$, f_j, g_i und Anfangs- und geeignete Randbedingungen seien vorgegeben. Die Zahl der Individuen der Spezies ist sowohl abhängig von zeitlichen Änderungen als auch von der räumlichen Verteilung und kann daher durch eine Diffusionsgleichung beschrieben werden. Die Populationen der Nahrungsquellen hingegen ist gleichmäßig verteilt und nur von Änderungen bez. der Zeit (z.B. Jahreszeiten) abhängig; ihre Dichten werden daher durch gewöhnliche Differentialgleichungen beschrieben.

Andere Interpretation: Obiges System kann auch ein Reaktions-Diffusions-System modellieren, wobei die Komponenten u_j Konzentrationen von diffundierenden Substanzen und die v_i Konzentrationen von Substanzen seien, deren Partikel nicht diffundieren können (als ideal durchmischt angenommen sind). Das zu lösende Differentialgleichungssystem dieses Beispiels besteht aus parabolischen und gewöhnlichen Differentialgleichungen. □

Beispiel 3.1.2. *Schmelzen fester Materialien in Elektro-Öfen* (bzw. elektrothermischen Schmelzanlagen, vgl. Abb. 3.1). Die Modellierung des Temperaturverlaufes mittels der Wärmeleitungsgleichung und der Bilanzgleichung für die Stromdichte im Schmelzofen führt für $t > 0$ auf das gekoppelte System bzw. auf die PDAE ([Grö78],[Luc91],[Can73])

$$\rho\, c\, \frac{\partial T}{\partial t} = \operatorname{div}(K\, \nabla T) + \sigma(T)(\nabla \Phi)^{\top}(\nabla \Phi) \qquad \text{in } \Omega \subset \mathbb{R}^k,\ k = 1, 2, 3$$

$$0 = \operatorname{div}(\sigma(T)\, \nabla \Phi) \qquad \text{in } \Omega$$

zur Bestimmung der Temperatur T und des elektrisches Potentials Φ. Hierzu seien Anfangs- und geeignete Randbedingungen vorgegeben, sowie die Dichte ρ, die spezifische Wärme c , die Wärmeleitfähigkeit K und die elektr. Leitfähigkeit $\sigma = \sigma(T)$ bekannt. In diesem Beispiel müssen parabolische und elliptische Differentialgleichun-

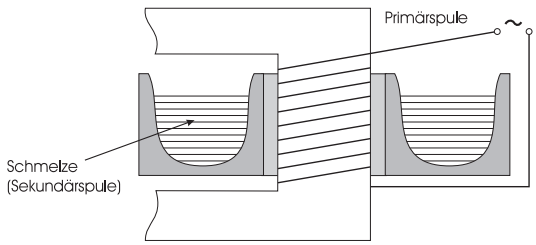

Abbildung 3.1. Schmelzen fester Materialien in Elektro-Öfen

gen simultan gelöst werden. □

Zahlreiche Anwendungsbeispiele gibt es auch in anderen naturwissenschaftlichen Bereichen. So findet man PDAEs auf dem Feld der Navier-Stokes Gleichungen [Cam96b], [Lin97], [Wei96], in der chemischen Verfahrenstechnik [Gri96], [Leu89], [Pip90], [Mar97], bei der Modellierung von Signalübertragungen in elektrischen Anlagen [Cha70] oder in der Magnetohydrodynamik [Sez87], [Cam97].

Diese einführenden Beispiele zeigen bereits, daß man bei gewissen Modellierungen gekoppelte Systeme von Differentialgleichungen unterschiedlichen Typs erhält. Die mathematische Behandlung wurde bisher nur für spezielle Systeme untersucht, so z.B. in [Alt98] und [Leu89] bez. der Existenz und Eindeutigkeit von Lösungen bestimmter Systeme und in [Jod90],[Jod91] bez. einer numerischen Lösung einer sehr speziellen linearen PDAE (siehe Bemerkung 3.5.28). In diesem Kapitel wollen wir lineare PDAEs mit konstanten Koeffizientenmatrizen untersuchen und deren numerische Lösung mittels zweier Diskretisierungsverfahren betrachten. Zur Charakterisierung der betrachteten Problemklasse führen wir in Analogie zu den DAEs

einen einheitlichen differentiellen Zeitindex und einen differentiellen Ortsindex ein und zeigen die Bedeutung dieser Indexe für die numerische Behandlung.

Wir betrachten lineare partielle differentiell-algebraische Gleichungen (lineare PDAEs) der Form

$$A\,u_t(t,x) + B\,u_{xx}(t,x) + C\,u(t,x) = g(t,x), \quad (t,x) \in \mathfrak{I} \times \Omega. \tag{3.1.1}$$

Hierbei seien $\mathfrak{I} = (0, t_e)$, $\Omega = (-l, l)$, $t_e > 0$, $l > 0$ und $A, B, C \in \mathbb{R}^{n\times n}$, $\bar{\mathfrak{I}} = [0, t_e]$ für $t_e < \infty$, $\bar{\mathfrak{I}} = [0, t_e)$ für $t_e = \infty$ und $\bar{\Omega} = [-l, l]$. $u, g : \bar{\mathfrak{I}} \times \bar{\Omega} \to \mathbb{R}^n$. Es kann sein, daß einige Gleichung der PDAE auf einer der Mengen $\bar{\mathfrak{I}} \times \Omega$, $\mathfrak{I} \times \bar{\Omega}$ oder $\bar{\mathfrak{I}} \times \bar{\Omega}$ definiert sind (siehe Beispiel 3.1.4 oder 3.3.3).

Insbesondere sei in unseren Betrachtungen mindestens eine der beiden Matrizen A und B singulär. Die Spezialfälle, daß $A = 0$ oder $B = 0$, wollen wir nicht betrachten, da dann die PDAE (3.1.1) eine parameterabhängige DAE ist. Wir setzen folglich voraus, daß A und B nicht identisch mit der Nullmatrix sind. Bei Anfangs-Randwertproblemen linearer partieller (z.B. parabolischer) Differentialgleichungssysteme (PDEs) der Form (3.1.1) mit regulären Matrizen A, B ist für eine eindeutige Lösbarkeit erforderlich, daß für jede Komponente von u Anfangs- und geeignete Randbedingungen vorzugeben sind. Bei linearen PDAEs der Form (3.1.1) sind hingegen nicht für alle Komponenten sowohl Anfangs- als auch Randbedingungen vorzugeben bzw. müssen gegebene Anfangs- und Randbedingungen gewissen zusätzlichen Bedingungen genügen, sie müssen *konsistent* sein (vgl. Definitionen 3.2.3 und 3.2.7). Bevor wir in Abschnitt 3.3 auf die Anfangs- und Randbedingungen näher eingehen werden, führen wir zwei Mengen $\mathfrak{M}_{AB}$ und $\mathfrak{M}_{RB}$ zur Beschreibung der Komponenten von u ein, für die Anfangs- und Randbedingungen vorgegeben werden können. Sei $u = (u_1, \ldots, u_n)^\top$. Wir schreiben für alle $t \in \bar{\mathfrak{I}}$ Randbedingungen der Form

$$\mathrm{R_B}\ u_j(t,x) := u_j(t, \pm l)\ =\ 0, \tag{3.1.2}$$

für Komponenten u_j von u vor, wenn $j \in \mathfrak{M}_{RB} \subset \{1, \ldots, n\}$. (Der Einfachheit halber wurden hier homogene Dirichlet-Randbedingungen gewählt, genauso sind auch von Neumann-Randbedingungen denkbar.) Weiterhin seien für $x \in \bar{\Omega}$ Anfangsbedingungen der Form

$$u_i(0,x) = u_{0i}(x) \tag{3.1.3}$$

mit $i \in \mathfrak{M}_{AB} \subset \{1, \ldots, n\}$ gegeben. Diese Komponenten $u_{0i}(x)$ von $u_0(x) := u(0,x)$ können beliebig vorgegeben werden. Weiterhin setzen wir voraus, daß die

Verträglichkeitsbedingungen

$$\mathrm{R_B}\, u_{0i}(x) = \mathrm{R_B}\, u_i(0,x) \quad \text{für} \quad i \in \mathfrak{M}_{AB} \cap \mathfrak{M}_{RB} \tag{3.1.4}$$

zwischen den Anfangs- und Randbedingungen gelten. Durch die folgende Definition wird der Lösungsbegriff gegeben:

DEFINITION 3.1.3. *Für eine hinreichend glatte Funktion g ist eine Funktion $u(t,x)$, $t \in \tilde{\mathfrak{I}}$, $x \in \bar{\Omega}$ eine Lösung der PDAE (3.1.1)-(3.1.4), wenn sie hinreichend glatt ist, die PDAE (punktweise) erfüllt und eindeutig durch die Anfangs- und Randbedingungen (3.1.3), (3.1.2) bestimmt ist.*

Das folgende einfache Beispiel illustriert die Besonderheit bei der Vorgabe von Anfangs- und Randbedingungen von Problemen der Form (3.1.1) mit singulären Matrizen A, B gegenüber von solchen mit regulären Matrizen A, B.

BEISPIEL 3.1.4. Es sei die PDAE

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{=A} \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix} + \underbrace{\begin{pmatrix} -1 & 0 & 0 \\ 0 & -b & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{=B} \begin{pmatrix} u_{1xx} \\ u_{2xx} \\ u_{3xx} \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & c_1 & 0 \\ 0 & 0 & c_2 \\ 0 & c_3 & 0 \end{pmatrix}}_{=C} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix}$$

$$u_j(t,-1) = u_j(t,1) \;=\; 0,\; j \in \mathfrak{M}_{RB}$$
$$u_i(0,x) = u_{0i}(x)\, \sin(\pi x),\; i \in \mathfrak{M}_{AB}$$

gegeben für $t \in \mathfrak{I}$, $x \in (-1,1)$ mit $a, b > 0$, $c_i \neq 0$ und hinreichend glatten Funktionen $g_i(t,x)$, $i = 1,2,3$. Der Term $\sin(\pi x)$ in den Anfangsbedingungen garantiert die Verträglichkeit (3.1.4) der Anfangsbedingungen mit den Randbedingungen. Im Gegensatz zu Problemen mit regulären Matrizen A, B, für die $\mathfrak{M}_{RB}, \mathfrak{M}_{AB} = \{1,2,3\}$ gilt, sind diese Mengen echte Teilmengen von $\{1,2,3\}$. Dies zeigt sich ganz einfach darin, daß sich für die zweite und dritte Lösungskomponente direkt ergibt:

$$\begin{aligned} u_2(t,x) &= \frac{1}{c_3} g_3(t,x) \\ u_3(t,x) &= \frac{1}{c_2} \left(g_2(t,x) - \frac{a}{c_3} g_{3t}(t,x) + \frac{b}{c_3} g_{3xx}(t,x) \right) . \end{aligned}$$

D.h., die Anfangs- und Randbedingungen für $u_2(t,x), u_3(t,x)$ sind vollständig durch die Anfangs- und Randwerte der rechten Seite der PDAE und ihrer Ableitungen gegeben. Folglich können wir keinerlei Anfangsbedingungen oder Randbedingungen

für $u_2(t,x), u_3(t,x)$ vorschreiben, und es gilt $\mathfrak{M}_{RB} = \mathfrak{M}_{AB} = \{1\}$. Die Lösungskomponente $u_1(t,x)$ ist die Lösung des parabolischen Anfangs-Randwertproblems

$$\begin{aligned} u_{1t}(t,x) &= u_{1xx}(t,x) + g_1(t,x) - \frac{c_1}{c_3} g_3(t,x), \\ u_1(t,-1) &= u_1(t,1) = 0\ ,\ u_1(0,x) = u_{01}(x)\sin(\pi x), \end{aligned}$$

für das wir $u_{01}(x)$ beliebig vorschreiben können. □

In den folgenden Abschnitten werden wir zunächst den differentiellen Orts- und den einheitlichen differentiellen Zeitindex für lineare PDAEs einführen und anhand von Beispielen einige Besonderheiten linearer PDAEs, für die ein einheitlicher differentieller Zeitindex nicht definiert werden kann, aufführen. In Abschnitt 3.3 werden wir auf die Wahl konsistenter Anfangs- und Randbedingungen eingehen, in Abschnitt 3.4 eine konsistente Darstellung der Lösung des ARWP (3.1.1)-(3.1.4) angeben. In Abschnitt 3.5 werden wir auf die numerische Behandlung linearer PDAEs eingehen, indem wir die Eigenschaften zweier bekannter Differenzenverfahren bei Anwendung auf lineare PDAEs betrachten und numerische Testergebnisse vorstellen.

3.2. Indexe linearer PDAEs

Ähnlich wie im Fall von DAEs (vgl. Abschnitt 1.2) ist es günstig, auch für PDAEs Indexe einzuführen. Im Gegensatz zu (gewöhnlichen) DAEs unterscheiden wir bei den PDAEs zwischen Orts- und Zeitindex. Diese Indexe beschreiben spezielle Eigenschaften des Systems in Bezug sowohl auf die analytische Lösung als auch auf die numerische Behandlung (siehe Abschnitt 3.5). In diesem Abschnitt wird mit Hilfe einer Laplace-Transformation ein differentieller Ortsindex eingeführt und auf der Grundlage einer Fouriertransformation ein einheitlicher differentieller Zeitindex definiert.

Es sei bemerkt, daß Indexe für lineare PDAEs der Form (3.1.1) in jüngster Zeit auch durch andere Autoren definiert wurden (vgl. [Cam96a], [Cam96b], [Mar97]). Die in diesem Kapitel definierten Indexe entsprechen teilweise Indexen der dortigen Definitionen, unterscheiden sich aber von diesen durch einige Details. Aufbauend auf den Indexen, die in den Reports [Luc97a], [Luc97b] eingeführt wurden, findet man Indexkonzepte für semilineare PDAEs in [Luc98] und in [Gün98] für lineare PDAEs, die aus einer Kopplung von DAEs und Gleichungen hyperbolischen Typs bestehen.

Für lineare PDAEs, für die die im folgenden eingeführten Indexe definiert sind, können einerseits Aussagen zur Vorgabe konsistenter Anfangs- und Randbedingungen und andererseits Konvergenzsaussagen zur numerischen Lösung dieser PDAEs für zwei Diskretisierungsverfahren getroffen werden (vgl. Abschnitt 3.5).

Wir nehmen für unsere weiteren Untersuchungen stets an, daß folgende Voraussetzungen erfüllt sind:

VORAUSSETZUNG 3.2.1. Für die betrachteten Probleme gelte:

a) Das Anfangs-Randwertproblem (ARWP) (3.1.1) - (3.1.4) hat eine und genau eine Lösung.
b) Jede Komponente des Lösungsvektors u, der partiellen Ableitung u_t und der Funktion g genüge einer Wachstumsbeschränkung der Form

$$|y(t,x)| \leq M\, e^{\alpha\, t}\ , \quad \alpha \geq 0, t \geq 0$$

(M und α sind unabhängig von x).
c) Das Matrixbüschel $(B, \xi A + C)$, $\xi \in \mathbb{C}$, $\mathrm{Re}(\xi) > \alpha$, ist regulär.
d) Das Matrixbüschel $(A, \rho_k B + C)$ ist regulär für alle k, ρ_k Eigenwert des Operators $\frac{\partial^2}{\partial x^2}$ zu den vorgeschriebenen RBn (3.1.2).
e) Der Vektor $g(t,x)$ der rechten Seite der PDAE (3.1.1) und der Vektor der Anfangswerte $u_0(x) = u(0,x)$ sind hinreichend glatt.

3.2.1. Ortsindex

Sei $t_e = \infty$ und $y : [0,\infty) \to \mathbb{R}$ stetig. Wir setzen voraus, daß y einer Wachstumsbeschränkung

$$|y(t)| \leq M\, e^{\alpha\, t}\ , \quad t \in [0,\infty),\ 0 < M < \infty,\ \alpha > 0.$$

genügt. Bezeichne y_ξ die Laplace-Transformierte von y (siehe z.B. [Smi58],[Bel84]), d.h.

$$y_\xi := \mathfrak{L}\{y(t)\} = \int_0^\infty e^{-t\,\xi}\, y(t)\, d\, t\ , \quad \mathrm{Re}(\xi) \geq \alpha.$$

Das Laplace-Integral y_ξ konvergiert in der rechten Halbebene $\mathrm{Re}(\xi) > \alpha$ (vgl. Abbildung 3.2). Die Inversionsformel des Laplace-Integrals ist gegeben durch

$$y(t) := \mathfrak{L}^{-1}\{y_\xi\} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{\xi t} y_\xi\, d\, \xi, \qquad t > 0,\ c \geq \alpha \text{ beliebig}, \tag{3.2.1}$$

wobei längs einer Geraden durch $\mathrm{Re}(\xi) = c$ integriert wird.

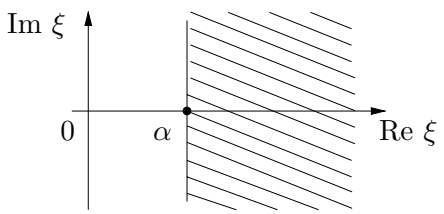

ABBILDUNG 3.2. Konvergenzbereich des Laplace-Integrals

Die Voraussetzung 3.2.1 b) sichert, daß die PDAE (3.1.1) Laplace-transformiert werden kann. Durch Multiplikation mit $e^{-t\xi}$ und anschließende Integration erhält man aus (3.1.1) die Laplace-transformierte PDAE

$$B\, u_\xi''(x) + (\xi A + C)\; u_\xi(x) = g_\xi(x) + A\, u_0(x)\;, \quad \mathrm{Re}(\xi) > \alpha\;, \tag{3.2.2}$$

wobei $u_0(x)$ der Anfangsvektor ist. Wenn die Matrix B singulär ist, dann ist die vom Parameter ξ abhängige Gleichung (3.2.2) eine DAE. Die Kronecker-Normalform der DAE (3.2.2) (vgl. Abschnitt 1.2) bildet die Grundlage zur Definition eines differentiellen Ortsindexes der PDAE und der Bestimmung der Menge $\mathfrak{M}_{RB}$ (vgl. Abschnitt 3.3). Die Voraussetzung 3.2.1 c) sichert, daß es nichtsinguläre Matrizen $S_{L,\xi}, T_{L,\xi} \in \mathbb{C}^{n\times n}$ gibt, so daß

$$S_{L,\xi}\, B\, T_{L,\xi} = \begin{pmatrix} I_{m_1} & 0 \\ 0 & N_{L,\xi} \end{pmatrix}, \quad S_{L,\xi}\; (\xi A + C)\; T_{L,\xi} = \begin{pmatrix} R_{L,\xi} & 0 \\ 0 & I_{m_2} \end{pmatrix}. \tag{3.2.3}$$

$R_{L,\xi} \in \mathbb{C}^{m_1\times m_1}$ ist eine quadratische Matrix und $N_{L,\xi} \in \mathbb{N}^{m_2\times m_2}$ ist eine nilpotente Jordan-Kettenmatrix, wobei $m_1 + m_2 = n$. $I_{m_i} \in \mathbb{N}^{m_i\times m_i}$, $i = 1, 2$, ist die Einheitsmatrix. Der Riesz-Index von $N_{L,\xi}$ sei mit $\nu_{L,\xi}$ bezeichnet. Gleichung (3.2.2) kann mit

$$\begin{pmatrix} v_\xi(x) \\ w_\xi(x) \end{pmatrix} := T_{L,\xi}^{-1} u_\xi(x) \quad \text{und} \quad \begin{pmatrix} r_{\xi,1}(x) \\ r_{\xi,2}(x) \end{pmatrix} := S_{L,\xi}(g_\xi(x) + Au_0(x))$$

in das entkoppelte Differentialgleichungssystem

$$v_\xi''(x) + R_{L,\xi}\, v_\xi(x) = r_{\xi,1}(x) \tag{3.2.4}$$

$$N_{L,\xi}\, w_\xi''(x) + w_\xi(x) = r_{\xi,2}(x) \tag{3.2.5}$$

transformiert werden. Die Differentialgleichung (3.2.4) zeigt, daß wir ein System gewöhnlicher Differentialgleichungen 2. Ordnung der Dimension m_1 zu lösen haben. Es müssen für jede Komponente $v_{\xi,j}$ von v_ξ geeignete Randbedingungen gegeben sein, damit (3.2.4) eine eindeutige Lösung besitzt. D.h., auch für m_1 Komponenten von u_ξ müssen geeignete Randbedingungen vorgegeben werden.

Gleichung (3.2.5) ist äquivalent zu einer algebraischen Gleichung, wie die folgenden Umformungen zeigen:

$$\begin{aligned} w_\xi(x) &= r_{\xi,2}(x) - N_{L,\xi} w''_\xi(x) \\ &= r_{\xi,2}(x) - N_{L,\xi} r''_{\xi,2}(x) + N^2_{L,\xi} w^{(4)}_\xi(x) \\ &\vdots \\ w_\xi(x) &= r_{\xi,2}(x) - N_{L,\xi} r''_{\xi,2}(x) + \ldots \\ &\qquad + (-1)^{\nu_{L,\xi}-1} N^{\nu_{L,\xi}-1}_{L,\xi} r^{(2\nu_{L,\xi}-2)}_{\xi,2}(x) \\ &\qquad + (-1)^{\nu_{L,\xi}} \underbrace{N^{\nu_{L,\xi}}_{L,\xi}}_{=0} w^{(2\nu_{L,\xi})}_\xi(x). \end{aligned} \tag{3.2.6}$$

Gleichung (3.2.6) legt uns eine Definition für den Ortsindex analog zur Definition des Differentiationsindex für DAEs nah: Differenziert man (3.2.6) einmal nach x, so erhält man

$$w'_\xi(x) = r'_{\xi,2}(x) - N_{L,\xi} r'''_{\xi,2}(x) + \ldots (-1)^{\nu_{L,\xi}-1} N^{\nu_{L,\xi}-1}_{L,\xi} r^{(2\nu_{L,\xi}-1)}_{\xi,2}(x) \ . \tag{3.2.7}$$

D.h., um eine explizite Differentialgleichung für $w_\xi(x)$ zu erhalten, braucht man $2\nu_{L,\xi} - 1$ Differentiationen nach der Variablen x.

Aus Gleichung (3.2.6) erhalten wir für jede Komponente des Vektors $w_\xi \in \mathbb{C}^{m_2}$ einen Ausdruck, der ausschließlich vom gegebenen Vektor $r_{\xi,2}$ und seiner Ableitungen nach x bis zur Ordnung $2\nu_{L,\xi} - 2$ abhängt. Das bedeutet, daß für alle Komponenten von w_ξ keine Randbedingungen vorgeschrieben werden können, da die Werte von w_ξ auf dem Rand durch die Randwerte von $r_{\xi,2}$ und dessen Ableitungen bestimmt sind.

Für welche m_1 Komponenten $u_{\xi j}$ von u_ξ (mit j aus einer Menge $\mathfrak{M}^\xi_{RB} \subset \{1, \ldots, n\}$) wir Randbedingungen vorschreiben können, kann man, ausgehend von w_ξ, anhand der Transformationsmatrix $T_{L,\xi}$ bestimmen. Dies werden wir in Abschnitt 3.3 erläutern. Im folgenden gehen wir davon aus, daß es ein $\alpha^* \geq \alpha$ gibt, so daß für alle $\xi \in \mathbb{C} : \mathrm{Re}(\xi) \geq \alpha^*$ einerseits $\mathfrak{M}^\xi_{RB} \equiv \mathfrak{M}_{RB}$ und andererseits auch $\nu_{L,\xi} \equiv \nu_L$ für die Nilpotenz von $N_{L,\xi}$ gilt.

Die Äquivalenz der transformierten PDAE (3.2.2) mit dem entkoppelten System (3.2.4), (3.2.5) und die Differentialgleichung (3.2.7) für $w_\xi(x)$ legen nahe, den differentiellen Ortsindex $\nu_{d,x}$ unter den Voraussetzungen 3.2.1 b), c) wie folgt zu definieren:

Definition 3.2.2. *Sei $\alpha^* > \alpha \in \mathbb{R}^+$ eine Zahl, so daß bez. der Matrizen A, B, C der PDAE (3.1.1) für alle $\xi \in \mathbb{C} : \mathrm{Re}(\xi) > \alpha^*$*

1. *das Matrixbüschel $(B, \xi A + C)$ regulär ist,*

2. $\mathfrak{M}_{RB}^{\xi} \equiv \mathfrak{M}_{RB}$ *gilt und*
3. *die Nilpotenz der Matrizen* $N_{L,\xi}$ *unabhängig von* ξ *ist, d.h.* $\nu_L \equiv \nu_{L,\xi} \geq 1$.

Dann hat die PDAE (3.1.1) den differentiellen Ortsindex

$$\nu_{d,x} := 2\nu_L - 1.$$

Für B regulär, d.h. wenn $\nu_L = 0$*, definieren wir* $\nu_{d,x} = 0$.

Der differentielle Ortsindex $\nu_{d,x}$ zeigt, welche Differenzierbarkeitseigenschaften bez. der Variablen x die Funktion $r_{\xi,2}(x)$ (und somit auch $g(t,x)$, $u_0(x)$) mindestens haben muß, um die Laplace-transformierte PDAE (3.2.2) in ein explizites Differentialgleichungssystem (3.2.4), (3.2.7) zu überführen. Besitzt eine PDAE einen differentiellen Ortsindex, so kann man die Menge $\mathfrak{M}_{RB}$ bestimmen und somit angeben, für welche Komponenten $u_j(t,x)$ mit $j \in \mathfrak{M}_{RB}$ Randbedingungen (3.1.2) vorgegeben werden können.

In der folgenden Definition wird der Begriff *konsistenter Randwerte*, die nicht beliebig vorgegeben werden können, eingeführt. D.h., wird ein Randwert für eine Komponente u_j mit $j \notin \mathfrak{M}_{RB}$ vorgegeben, so gibt es nur dann eine Lösung der PDAE, die diesem Randwert genügt, wenn dieser gewisse (Konsistenz-)Bedingungen erfüllt, also konsistent ist.

Definition 3.2.3. *Sei* $\mathfrak{M}_{RB}^{\xi} = \mathfrak{M}_{RB}$, $\forall \xi \in \mathbb{C} : \mathrm{Re}(\xi) \geq \alpha^*$. *Dann heißt ein Randwert* $u_j(t, \pm l)$, $j \notin \mathfrak{M}_{RB}$, *einer Komponente von u* konsistent, *wenn seine Laplace-Transformierte*

$$u_{\xi j}(\pm l) = \left(T_{L,\xi} \begin{pmatrix} v_\xi(\pm l) \\ w_\xi(\pm l) \end{pmatrix} \right)_j$$

erfüllt.

Bemerkung 3.2.4. Ist die Matrix $R_{L,\xi}$ in (3.2.4) negativ definit, dann ist das Randwertproblem (3.2.2) mit homogenen Dirichlet-Randbedingungen für $u_{\xi j}$, $j \in \mathfrak{M}_{RB}^{\xi}$, unter der Voraussetzung 3.2.1 c) eindeutig lösbar. □

Bemerkung 3.2.5. Da n in vielen Anwendungen nur klein ist (z.B. $n = 2, 3$ oder 4), kann die Nilpotenz $\nu_{L,\xi}$ der Matrix $N_{L,\xi}$ leicht bestimmt werden. Im Fall, daß $N_{L,\xi}$ nur aus einem Block besteht, gilt $\nu_{L,\xi} = m_2$. □

3.2.2. Zeitindex

Neben dem bereits definierten Ortsindex führen wir nun einen Zeitindex für PDAEs ein. Wir multiplizieren hierzu die PDAE (3.1.1) mit einer geeigneten Funktion $\phi_k(x)$ und integrieren die Gleichung nach x auf dem Intervall $[-l, l]$. Die Funktionen $\phi_k(x)$, $k = 1, 2, ...$, sind orthogonale Eigenfunktionen des Operators

$\frac{\partial^2}{\partial x^2}$ zugehörig zum Eigenwert ρ_k. $\phi_k(x)$ genügen den selben Randbedingungen wie $u_j(t,x)$, $j \in \mathfrak{M}_{RB}$, d.h. die homogenen Bedingungen (3.1.2). Mittels einer endlichen Fouriertransformation bez. der Eigenfunktionen ϕ_k einer Vektorfunktion $\chi(t,x)$

$$\hat{\chi}_k(t) = \frac{1}{l}\int_{-l}^{l} \chi(t,x)\phi_k(x)\,dx,\ \ k = 1,2,... \tag{3.2.8}$$

(vgl. z.B. [Smi58],[Bra65],[Naa72]) erhalten wir

$$A\hat{u}_k'(t) + (\rho_k\, B + C)\,\hat{u}_k(t) = \hat{g}_k(t) + B\varphi_k(t)\ \ =:\ \bar{g}_k(t) \tag{3.2.9}$$

mit $\varphi_k(t) = (\varphi_{k1}(t), .., \varphi_{kn}(t))^\top$ und

$$\begin{aligned} \varphi_{kj}(t) &= 0 && \text{für } j \in \mathfrak{M}_{RB}, \\ \varphi_{kj}(t) &= \frac{1}{l}\Big[\phi_k'(x)u_j(t,x)\Big]_{x=-l}^{x=l} && \text{für } j \notin \mathfrak{M}_{RB}, \end{aligned}$$

wobei φ aus der partiellen Integration des Terms

$$\int_{-l}^{l} u_{xx}(t,x)\phi_k(x)\,dx$$

resultiert.

Bemerkung 3.2.6. Die $\hat{u}_k$ sind die endlichen Fouriertransformierten bez. der Funktionen $\phi_k = \sin\left(\frac{k\pi}{2l}(x+l)\right)$, $k = 1(1)M$. Eine konsistente Darstellung der Lösung $u(t,x)$ wird in Abschnitt 3.4 diskutiert. □

Wenn die Matrix A singulär ist, dann ist Gleichung (3.2.9) eine DAE, die vom Parameter ρ_k abhängt. Diese kann mit Voraussetzungen 3.2.1 d) und e) eindeutig gelöst werden, wenn Anfangsbedingungen nur für Komponenten $\hat{u}_{ki}$ von $\hat{u}_k$ mit i aus einer Menge $\mathfrak{M}_{AB}^k \subset \{1,\dots,n\}$ beliebig vorgegeben werden. Voraussetzung 3.2.1 d) sichert ebenfalls, daß eine Kronecker-Transformation mit regulären Matrizen $S_{F,k}, T_{F,k}$ durchgeführt werden kann, so daß

$$S_{F,k}\,A\,T_{F,k} = \begin{pmatrix} I_{n_1} & 0 \\ 0 & N_{F,k} \end{pmatrix}, \quad S_{F,k}\,(\rho_k B + C)\,T_{F,k} = \begin{pmatrix} R_{F,k} & 0 \\ 0 & I_{n_2} \end{pmatrix} \tag{3.2.10}$$

mit $R_{F,k} \in \mathbb{R}^{n_1\times n_1}$. $N_{F,k} \in \mathbb{N}^{n_2\times n_2}$ ist wiederum eine nilpotente Jordan-Kettenmatrix mit Riesz-Index $\nu_{F,k}$ ($n_1 + n_2 = n$), $I_{n_j} \in \mathbb{N}^{n_j\times n_j}$ ($j = 1,2$) die Einheitsmatrix. Aus dieser Beziehung folgt, daß Gleichung (3.2.9) äquivalent zum entkoppelten Differentialgleichungssystem

$$y_k'(t) + R_{F,k}\,y_k(t) = s_{k,1}(t) \tag{3.2.11}$$

$$N_{F,k}\,z_k'(t) + z_k(t) = s_{k,2}(t) \tag{3.2.12}$$

ist, wobei $(y_k^\top(t), z_k^\top(t))^\top := T_{F,k}^{-1}\hat{u}_k(t)$, $(s_{k,1}^\top(t), s_{k,2}^\top(t))^\top := S_{F,k}\bar{g}_k(t)$. Gleichung (3.2.11) ist ein gewöhnliches Differentialgleichungssystem erster Ordnung, das für jeden beliebigen Anfangswert $y_k(0)$ und jede stetige Funktion $s_{k,1}(t)$ eine eindeutige Lösung hat, die durch

$$y_k(t) = e^{-R_{F,k}t}y_k(0) + \int_0^t e^{-R_{F,k}(t-\tau)}\, s_{k,1}(\tau)\, d\,\tau \tag{3.2.13}$$

gegeben ist. Die Lösung $z_k(t)$ ergibt sich aus (3.2.12) zu

$$z_k(t) = \sum_{i=0}^{\nu_{F,k}-1} (-N_{F,k})^i\, s_{k,2}^{(i)}(t)\ . \tag{3.2.14}$$

Diese Gleichung zeigt, daß wir keine Anfangsbedingungen $z_k(0)$ vorschreiben können, weil durch diese Gleichung die Anfangswerte von $z_k(t)$ durch die Anfangswerte von $s_{k,2}(t)$ und ihrer Ableitungen gegeben sind. Da die Lösungskomponente $z_k(t)$ gewisse Glattheitseigenschaften besitzen soll, hat $s_{k,2}(t)$ für Indexe $\nu_{F,k} \geq 2$ scharfe Forderungen an die Differenzierbarkeit zu erfüllen.

Mit Hilfe der Matrix $T_{F,k}$ kann man die Menge $\mathfrak{M}_{AB}^k$ bestimmen (vgl. Abschnitt 3.3), für die die ABn der Komponenten $\hat{u}_{ki}$ mit $i \in \mathfrak{M}_{AB}^k$ beliebig vorgegeben werden können. Gibt man für Komponenten $u_i(t,x)$, $i \notin \mathfrak{M}_{AB}$, von $u(t,x)$ ABn beliebig vor und transformiert diese mittels der endlichen Fouriertransformation (3.2.8), so fordern wir, daß diese ABn die Gleichung (3.2.14) für alle $k = 1, 2, ..$ nicht verletzen. D.h., wir fordern $\mathfrak{M}_{AB}^k \equiv \mathfrak{M}_{AB}\ \forall k$. Es gibt Beispiele von PDAEs, für die dies nicht erfüllbar ist (siehe Abschnitt 3.2.3)

Das Analogon zu Definition 3.2.3 ist:

Definition 3.2.7. *Sei $\mathfrak{M}_{AB}^k = \mathfrak{M}_{AB}$, $k = 1, 2, ...$ Dann heißt ein Anfangswert $u_j(0,x)$, $j \notin \mathfrak{M}_{AB}$, $x \in [-l,l]$ einer Komponente von u* konsistent*, wenn seine endliche Fouriertransformierte*

$$\hat{u}_{kj}(0) = \left(T_{F,k}\begin{pmatrix} y_k(0) \\ z_k(0)\end{pmatrix}\right)_j$$

erfüllt.

Auf der Grundlage der vorherigen Betrachtungen wird der Zeitindex einer linearen PDAE definiert.

Definition 3.2.8. *Für die PDAE (3.1.1) sei für alle $k = 1, 2, ..$*

1. *das Matrixbüschel $(A, \rho_k\, B + C)$ regulär,*
2. *$\mathfrak{M}_{AB}^k \equiv \mathfrak{M}_{AB}$ und*
3. *die Nilpotenz der Matrizen $N_{F,k}$ unabhängig von k, d.h. $\nu_{F,k} \equiv \nu_F$.*

Dann hat die PDAE (3.1.1) den einheitlichen differentiellen Zeitindex *$\nu_{d,t} := \nu_F$.*

Gleichung (3.2.14) zeigt, daß der einheitliche Zeitindex $\nu_{d,t}$ Informationen darüber gibt, welche Differenzierbarkeitseigenschaften bez. der Variablen t die Funktion $s_{k,2}(t)$ (und somit auch $g(t,x)$) mindestens haben muß. Weiterhin zeigt Gleichung (3.2.11), daß n_1 Anfangsbedingungen für die Lösung der DAE (3.2.9) und somit bei einheitlichem differentiellen Zeitindex (da $\mathfrak{M}_{AB}^k = \mathfrak{M}_{AB}$) für die Lösung der PDAE (3.1.1) beliebig vorgegeben werden können.

Bemerkung 3.2.9. Für $\nu_{d,t} = 0$ und $\nu_{d,x} = 0$ ist (3.1.1) eine PDE. □

Beispiel 3.2.10. Für die PDAE in Beispiel 3.1.4 kann man leicht zeigen, daß $\nu_L = 2$, d.h. $\nu_{d,x} = 3$, $\nu_{d,t} = 2$. □

Beispiel 3.2.11. Sei in der PDAE (3.1.1) $n = 2$

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} b_1 & b_2 \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} c_1 & c_2 \\ c_3 & 0 \end{pmatrix}, \ c_2, c_3 \neq 0,$$

$b_1, b_2, c_1, c_2, c_3 \in \mathbb{R}$. Für $b_2 = 0$ erhalten wir die Indexe $\nu_L = 2$ und $\nu_{d,x} = 3$. Hingegen für $b_2 \neq 0$ folgt $\nu_L = 1$ und $\nu_{d,x} = 1$. Für alle $b_2 \in \mathbb{R}$ ist $\nu_{d,t} = 2$. □

3.2.3. PDAEs mit nichteinheitlichem Zeitindex

Wir betrachten weiter die Fouriertransformierte PDAE (3.2.9). Wenn die Matrizen A, B, C eine spezielle Struktur haben, kann es sein, daß die Voraussetzungen 2. und/oder 3. in der Definition für den einheitlichen Zeitindex nicht erfüllt sind. So kann es vorkommen, daß das Matrixbüschel $(A, \rho_k\, B + C)$ nicht den gleichen Index für alle k hat. Abhängig von n ist dies nur für endlich viele k möglich. Wir sprechen dann vom einem *Indexsprung* (siehe Beispiel 3.2.12). Ferner kann auftreten, daß zwar der Riesz-Index des Büschels $(A, \rho_k\, B + C)$ für alle k gleich ist, aber die Bedingung $\mathfrak{M}_{AB}^k = \mathfrak{M}_{AB}$ nicht für alle $k = 1, 2, ..$ erfüllt ist (siehe Beispiel 3.2.13). Probleme wie diese unterscheiden sich qualitativ von solchen mit einem einheitlichen Zeitindex.

Die folgenden Beispiele illustrieren den Fall des Indexsprungs bzw. des Wechsels von $N_{F,k}$.

Beispiel 3.2.12. Seien die Matrizen $A, B, C \in \mathbb{R}^{2\times 2}$ gegeben durch

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \ B \ = \ \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}, \ C \ = \ \begin{pmatrix} c_1 & c_2 \\ c_3 & c_4 \end{pmatrix}, \ \det(B) \neq 0.$$

Da B regulär ist, folgt $\mathfrak{M}_{RB} = \{1, 2\}$ und in Gleichung (3.2.9) $\varphi_k(t) = 0$. Das Matrixbüschel $(A, \rho_k B + C)$ ist regulär, wenn

$$\begin{aligned} &\text{entweder } b_4\,\rho_k = -c_4 \text{ und } (b_2\,\rho_k + c_2)(b_3\,\rho_k + c_3) \neq 0 \\ &\text{oder } b_4\,\rho_k \neq -c_4. \end{aligned}$$

Sei κ so gewählt, daß die Matrix $\kappa A + (\rho_k B + C)$ regulär ist. Dann berechnet sich der Riesz-Index des Matrixbüschels durch (vgl. Definition 1.2.4)

$$\operatorname{ind}(A, \rho_k B + C) = \operatorname{ind}\Big([\kappa A + (\rho_k B + C)]^{-1} A\Big) = \begin{cases} 1 & \text{für } b_4\,\rho_k \ \neq -c_4 \\ 2 & \text{für } b_4\,\rho_k \ = -c_4\ . \end{cases}$$

Die PDAE hat

- den einheitlichen Zeitindex 1, wenn $b_4\,\rho_k \neq -c_4$ für alle $k = 1, 2, ..,$
- den einheitlichen Zeitindex 2, wenn $b_4 = c_4 = 0$,
- einen Indexsprung, wenn $\rho_k = -\frac{c_4}{b_4}$ für ein k.

Sei der Einfachheit halber die Matrix $\rho_k B + C$ regulär für alle $k > 0$, d.h. $D_k := \det(\rho_k B + C) \neq 0$. Unter Verwendung der Jordanschen Normalform von $(\rho_k B + C)^{-1} A$ können wir die Lösung der DAE (3.2.9) wie folgt angeben (siehe z.B. [Str95]).

1. Für den Fall, daß $b_4\,\rho_k \neq -c_4$ (das Matrixbüschel $(A, \rho_k B + C)$ hat dann den Riesz-Index 1) hat die Lösung $\hat{u}_k = (\hat{u}_{k1}, \hat{u}_{k2})^\top$ von (3.2.9) die Darstellung

$$\begin{aligned} \hat{u}_k(t) = \hat{u}_{k1}(0) \begin{pmatrix} 1 \\ -\frac{b_3\,\rho_k + c_3}{b_4\,\rho_k + c_4} \end{pmatrix} e^{-\frac{D_k}{b_4\,\rho_k + c_4} t} + \begin{pmatrix} 0 \\ \frac{1}{b_4\,\rho_k + c_4} \end{pmatrix} \hat{g}_{k2}(t) \\ + \begin{pmatrix} 1 \\ -\frac{b_3\,\rho_k + c_3}{b_4\,\rho_k + c_4} \end{pmatrix} \int_0^t \Big(\hat{g}_{k1}(\tau) - \tfrac{b_2\,\rho_k + c_2}{b_4\,\rho_k + c_4} \hat{g}_{k2}(\tau) \Big) e^{\frac{D_k}{b_4\,\rho_k + c_4}(\tau - t)}\, d\,\tau\ . \end{aligned}$$

   mit $\hat{g}_k = (\hat{g}_{k1}, \hat{g}_{k2})^\top$. Für $t = 0$ muß dann

$$\hat{u}_k(0) = \begin{pmatrix} \hat{u}_{k1}(0) \\ \hat{u}_{k2}(0) \end{pmatrix} \ = \ \hat{u}_{k1}(0) \begin{pmatrix} 1 \\ -\frac{b_3\,\rho_k + c_3}{b_4\,\rho_k + c_4} \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{b_4\,\rho_k + c_4} \end{pmatrix} \hat{g}_{k2}(0)$$

   erfüllt sein. Das liefert uns die Konsistenzbedingung zwischen den Anfangswerten und der rechten Seite

$$\hat{g}_{k2}(0) = (b_3\,\rho_k + c_3)\hat{u}_{k1}(0) + (b_4\,\rho_k + c_4)\hat{u}_{k2}(0)\ .$$

   Wenn wir $\hat{u}_{k1}(0)$ vorschreiben, dann ist $\hat{u}_{k2}(0)$ eindeutig festgelegt durch $\hat{u}_{k1}(0)$ und $\hat{g}_{k2}(0)$ und kann somit nicht beliebig vorgeschrieben werden.

2. Die Lösung von (3.2.9) für $b_4\,\rho_k = -c_4$ (also Riesz-Index 2) ist gegeben durch

$$\hat{u}_k(t) = \begin{pmatrix} \dfrac{\hat{g}_{k2}(t)}{b_3\,\rho_k + c_3} \\ \dfrac{(b_3\,\rho_k + c_3)\hat{g}_{k1}(t) - (b_1\,\rho_k + c_1)\hat{g}_{k2}(t) - \hat{g}'_{k2}(t)}{(b_2\,\rho_k + c_2)(b_3\,\rho_k + c_3)} \end{pmatrix},$$

und die Konsistenzbedingungen lauten

$$\hat{u}_{k1}(0) = \frac{\hat{g}_{k2}(0)}{b_3\,\rho_k + c_3},$$
$$\hat{u}_{k2}(0) = \frac{(b_3\,\rho_k + c_3)\hat{g}_{k1}(0) - (b_1\,\rho_k + c_1)\hat{g}_{k2}(0) - \hat{g}'_{k2}(0)}{(b_2\,\rho_k + c_2)(b_3\,\rho_k + c_3)},$$

was bedeutet, daß wir im allgemeinen weder $\hat{u}_{k1}(0)$ noch $\hat{u}_{k2}(0)$ vorschreiben können.

Da B regulär ist, werden für alle Komponenten homogene Dirichlet-RBn vorgeschrieben. Die $\hat{u}_k$, $\hat{g}_k$ können als k-te Fourierkoeffizienten von $u(t,x)$, $g(t,x)$ der Fourierreihendarstellung bez. der Basisfunktionen $\phi_k(x)$ interpretiert werden. Dann ist

$$u(t,x) = \sum_{k=1}^{\infty} \hat{u}_k(t)\phi_k(x), \quad g(t,x) = \sum_{k=1}^{\infty} \hat{g}_k(t)\phi_k(x).$$

In diesem Beispiel haben bei einheitlichem differentiellen Zeitindex 1 der PDAE (3.1.1) alle DAEs (3.2.9) den Differentiationsindex 1. Das bedeutet, daß wir dann für alle Anfangswertprobleme (3.2.9) $\hat{u}_{k1}(0)$ beliebig vorschreiben können. Weiterhin kann ein konsistentes $\hat{u}_{k2}(0)$ eindeutig aus $\hat{u}_{k1}(0)$ und der gegebenen rechten Seite berechnet werden. Für $u_0(x)$ ergibt sich folglich, daß nur für die erste Komponente ein Anfangswert $u_1(0,x)$ vorgeschrieben werden kann und sich ein konsistentes $u_2(0,x)$ aus $u_1(0,x)$ und der rechten Seite bestimmen läßt (d.h. $\mathfrak{M}_{AB} = \mathfrak{M}^k_{AB} = \{1\}$). Analog können wir keinerlei Anfangsbedingung für den Fall des einheitlichen differentiellen Zeitindex 2 vorschreiben (d.h. $\mathfrak{M}_{AB} = \mathfrak{M}^k_{AB} = \emptyset$), da die Anfangswerte vollständig durch die rechte Seite bestimmt sind.

Im Fall eines Indexsprungs kann $\mathfrak{M}_{AB}$ auf der Grundlage der Fouriertransformation nicht ermittelt werden. Dann ist $\rho_j = -\frac{c_4}{b_4}$ für ein $j > 0$, und der j-te Fourierkoeffizient von $u_1(0,x)$ ist durch den Anfangswert der rechten Seite festgelegt. Alle anderen Fourierkoeffizienten $\hat{u}_{k1}(0)$, $k \neq j$, sind freie Parameter. Es ist $\mathfrak{M}^j_{AB} = \emptyset$ und $\mathfrak{M}^k_{AB} = \{1\}$ für $k \neq j$, so daß die zweite Bedingung in Definition 3.2.8 verletzt ist und kein einheitlicher differentieller Zeitindex festgelegt werden kann. □

BEISPIEL 3.2.13. Sei $n = 4$ und die Matrizen A, B, C gegeben durch

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; B = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & c \end{pmatrix}.$$

Das Matrixbüschel $(A, \rho_k\, B + C)$ hat den Riesz-Index 2 für alle $k \in \mathbb{N}^+$. Sei $c = \rho_{\bar{k}}$ für ein $\bar{k} \in \mathbb{N}^+$. Die Kroneckertransformationen (3.2.10) des Büschels $(A, \rho_{\bar{k}}\, B + C)$ und des Büschels $(A, \rho_k\, B + C)$ liefern

$$S_{F,\bar{k}}\, A \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-c & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{T_{F,\bar{k}}} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{N_{F,\bar{k}}},$$

$$S_{F,k}\, A \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1-\rho_k & 0 \\ \rho_k - c & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}}_{T_{F,k}} = \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \underbrace{0 \quad 0 \quad 0}_{N_{F,k}} & & \end{array}\right).$$

D.h., die Matrix $N_{F,\bar{k}} \in \mathbb{R}^{4\times 4}$ ist verschieden von den Matrizen $N_{F,k} \in \mathbb{R}^{3\times 3}$, $k \neq \bar{k}$. Somit ist $\mathfrak{M}_{AB}^{\bar{k}} = \emptyset$, aber $\mathfrak{M}_{AB}^{k} \neq \emptyset$ $(k \neq \bar{k})$ hat ein Element $(n_1 = 1)$. Aus (3.2.14) folgt, daß $u_{\bar{k}}(0)$ vollständig durch die rechte Seite bestimmt ist. Aber aus (3.2.11) und der Beziehung $u_k(t) = T_{F,k}(y_k^\top, z_k^\top)^\top$ ergibt sich, daß für alle $k \neq \bar{k}$ z.B. die dritte Komponente von $u_k(0)$ beliebig gewählt werden kann, d.h. $\mathfrak{M}_{AB}^{k} = \{3\}$. Es kann für eine PDAE (3.1.1) mit diesen Matrizen kein einheitlicher Zeitindex festgelegt werden. □

In den vorangegangenen Beispielen hatten die Mengen $\mathfrak{M}_{AB}^{\bar{k}}$ und $\mathfrak{M}_{AB}^{k}$ $(k \neq \bar{k})$ unterschiedlich viele Elemente und insbesondere waren die Matrizen $N_{F,\bar{k}}$ und $N_{F,k}$ unterschiedlich. Im folgenden Beispiel sind zwar die Anzahl der Elemente der Mengen $\mathfrak{M}_{AB}^{\bar{k}}$ und die Matrizen $N_{F,k}$ für alle k gleich, aber es ist nicht möglich, eine Menge $\mathfrak{M}_{AB}$ zu bestimmen.

Beispiel 3.2.14. [Wen98] Seien in (3.1.1)

$$A = \frac{1}{a-b}\begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} -b & a \\ a & b \end{pmatrix}, \quad a \neq b.$$

Dieses Beispiel besitzt den differentiellen Ortsindex $\nu_{d,x} = 0$, es kann aber kein einheitlicher differentieller Zeitindex definiert werden:

Die Transformationsmatrizen $S_{F,k}, T_{F,k}$ sind bis auf einen Parameter $\beta \neq 0$ durch

$$S_{F,k} = \begin{pmatrix} -\frac{2\rho_k+a+b}{a-b} & 1 \\ \frac{1}{\beta(a-b)} & 0 \end{pmatrix}, \qquad T_{F,k} = \begin{pmatrix} a-\rho_k & \beta \\ b-\rho_k & \beta \end{pmatrix}$$

bestimmt. Gilt $\rho_k \neq a, b$ für alle $k \in \mathbb{N}^+$, so kann man $\mathfrak{M}_{AB} = \mathfrak{M}_{AB}^k = \{1\}$ oder $\mathfrak{M}_{AB} = \mathfrak{M}_{AB}^k = \{2\}$ wählen. Gilt hingegen $\rho_{k_a} = a$ und $\rho_{k_b} = b$ für $k_a, k_b \in \mathbb{N}^+$, so kann man $\mathfrak{M}_{AB}$ nicht angeben. In diesem Fall ist

$$T_{F,k_a} = \begin{pmatrix} 0 & \beta \\ b-a & \beta \end{pmatrix} \qquad \text{für} \qquad \rho_{k_a} = a$$

$$\text{und} \qquad T_{F,k_b} = \begin{pmatrix} a-b & \beta \\ 0 & \beta \end{pmatrix} \qquad \text{für} \qquad \rho_{k_b} = b.$$

D.h., einerseits ist der Anfangswert für die erste Komponente von u_{k_a} durch die rechte Seite bestimmt, da $u_{k_a 1} = \beta z_{k_a}$, und $u_{k_a 2}(0)$ kann beliebig gewählt werden. Andererseits ist $u_{k_b 2}(0)$ durch die rechte Seite bestimmt und $u_{k_b 1}(0)$ kann beliebig vorgeschrieben werden. Es ist stets $\mathfrak{M}_{AB}^{k_a} \neq \mathfrak{M}_{AB}^{k_b}$. Setzt man in diesem Beispiel $a, b > 0$ voraus, so gilt stets $\rho_k \neq a, b$, da $\rho_k < 0$, und man kann $\mathfrak{M}_{AB} = \{1\}$ oder $\mathfrak{M}_{AB} = \{2\}$ wählen.

Bemerkt sei, daß man bei diesem Beispiel die PDAE durch eine Variablentransformation in eine PDAE mit einheitlichem Zeitindex überführen kann: Sei $\bar{u}_1 := \frac{1}{a-b}(u_1 - u_2)$ und $\bar{u}_2 := \frac{1}{a-b}(u_1 + u_2)$. Dann lautet die PDAE für $\bar{u} = (\bar{u}_1, \bar{u}_2)^\top$

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \bar{u}_t + \begin{pmatrix} a-b & 0 \\ 0 & a-b \end{pmatrix} \bar{u}_{xx} + \frac{a-b}{2} \begin{pmatrix} -(a+b) & a-b \\ (a-b) & a+b \end{pmatrix} \bar{u} = g.$$

Aus

$$S_{F,k} = \begin{pmatrix} -\frac{2\rho_k+a+b}{(a-b)^2} & \frac{1}{a-b} \\ \frac{2}{(a-b)^2} & 0 \end{pmatrix} \text{ und } T_{F,k} = \begin{pmatrix} a-b & 0 \\ -\rho_k+a+b & 1 \end{pmatrix} \quad \forall\ k = 1, 2, ...$$

folgt $\nu_{d,t} \equiv \nu_{F,k} = 1$, und man kann $\mathfrak{M}_{AB} \equiv \mathfrak{M}_{AB}^k = \{1\}$ wählen. □

Bemerkung 3.2.15. Die Indexdefinitionen in [Cam96a], [Cam96b] basieren ebenfalls auf der Fouriertransformierten DAE (3.2.9), unterscheiden aber nicht zwischen Problemen mit und ohne Indexsprung bzw. verschiedenen $\mathfrak{M}_{AB}^k$. □

3.3. Konsistenz von Anfangs- und Randbedingungen

In diesem Abschnitt gehen wir auf die Vorgabe konsistenter Anfangs- und Randbedingungen ein. Hierbei werden die Bezeichnungen aus Abschnitt 3.2.1 und 3.2.2 verwendet und nur PDAEs betrachtet, für die ein differentieller Ortsindex und ein einheitlich differentieller Zeitindex definiert werden kann. Wir betrachten zunächst die Bestimmung von $\mathfrak{M}_{AB}^k$. Es wurde bereits festgestellt, daß nur Anfangsbedingungen für $y_k(t)$, aber nicht für $z_k(t)$ vorgeschrieben werden können. Da $u_k = T_{F,k}(y_k^\top, z_k^\top)^\top$ und $T_{F,k}$ regulär ist, folgt, daß n_1 Elemente von $u_k(0)$ beliebig vorgegeben werden können und daß man für $n_2 = m - n_1$ Komponenten von $u_k(t)$ keinerlei Anfangsbedingungen vorgeben kann. Die Matrix $T_{F,k}$ zerlegen wir in $T_{F,k} = (T_{k,1}\, T_{k,2})$ mit $T_{k,i} \in \mathbb{R}^{n\times n_i}$, $i = 1, 2$, dann ist $u_k = T_{k,1}y_k + T_{k,2}z_k$. Seien mit $\theta_{ki} = (\theta_{ki1}, \dots, \theta_{kin_1}) \in \mathbb{R}^{n_1}$, $i = 1(1)n$, die Zeilen von $T_{k,1}$ bezeichnet. Aus diesen θ_{ki} lassen sich stets n_1 linear unabhängige Zeilen $i_1, \dots, i_{n_1}$ auswählen. Sind für alle k diese Zeilen $\theta_{ki_1}, \dots, \theta_{ki_{n_1}}$ linear unabhängig, so kann man $\mathfrak{M}_{AB} = \{i_1, \dots, i_{n_1}\}$ setzen. Diese Auswahl linear unabhängiger Zeilen von $T_{k,1}$ und somit die Menge $\mathfrak{M}_{AB}^k$ nicht immer eindeutig bestimmt ist (vgl. Beispiele 3.3.1, 3.3.2). Analog kann man die Menge $\mathfrak{M}_{RB}$ bestimmen, die m_1 Elemente hat, die nicht immer eindeutig bestimmt sind. Die folgenden Beispiele illustrieren die Bestimmung der Mengen $\mathfrak{M}_{AB}$ und $\mathfrak{M}_{RB}$.

Beispiel 3.3.1. Seien

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix},\ C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

Mit den Bezeichnungen aus Abschnitt 3.2 erhält man für diese Matrizen

$$S_{F,k} = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{4}\rho_k \\ 1 & -1 & \rho_k \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}, \qquad T_{F,k} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{pmatrix},$$

$$S_{L,\xi} = \begin{pmatrix} 0 & -1 & \xi+1 \\ \frac{1}{\xi+1} & -\frac{1}{\xi+1} & 1 \\ 0 & 0 & -\frac{1}{\xi+1} \end{pmatrix}, \qquad T_{L,\xi} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & -(\xi+1) \\ 1 & 0 & 0 \end{pmatrix},$$

$$N_{F,k} = 0,\ R_{F,k} = \begin{pmatrix} 1-\frac{\rho_k}{2} & 0 \\ 0 & 1 \end{pmatrix}, \qquad N_{L,\xi} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ R_{L,\xi} = -2(\xi+1)$$

und somit $\nu_{d,t} = 1, \nu_{d,x} = 3$, $n_1 = 2, n_2 = 1, m_1 = 1, m_2 = 2$. D.h., es können $n_1 = 2$ Anfangsbedingungen beliebig vorgegeben werden. Es ist

$$T_{k,1} = \begin{pmatrix} \theta_{k1} \\ \theta_{k2} \\ \theta_{k3} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Zwei linear unabhängige Zeilen von $T_{k,1}$ sind für alle $k = 1, 2, ...$ die Zeilen θ_{k1} und θ_{k2} sowie auch θ_{k1} und θ_{k3}. Somit kann man $\mathfrak{M}^k_{AB} = \{1, 2\}$ oder $\mathfrak{M}^k_{AB} = \{1, 3\}$ für alle k wählen. $\mathfrak{M}^k_{AB} = \{2, 3\}$ ist nicht möglich, da $\theta_{k2} = \theta_{k3}$ und linear abhängig sind. Somit kann $\mathfrak{M}_{AB} = \{1, 2\}$ oder $\mathfrak{M}_{AB} = \{1, 3\}$ gesetzt werden. Da $m_1 = 1$, hat $\mathfrak{M}_{RB}$ nur ein Element. Da alle Elemente der ersten Spalte von $T_{L,\xi}$ ungleich Null sind, kann man sich eine beliebige Komponente von u für $\mathfrak{M}_{RB}$ auswählen, z.B. $\mathfrak{M}_{RB} = \{3\}$. □

Beispiel 3.3.2. Seien

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ -1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & c \\ 1 & 0 \end{pmatrix}, \quad c > 0.$$

Mit den Matrizen

$$S_{F,k} = \begin{pmatrix} \frac{1}{c} & \frac{2-c-\rho_k}{c(\rho_k-1)} \\ 0 & \frac{1}{\rho_k-1} \end{pmatrix}, \quad T_{F,k} = \begin{pmatrix} 0 & -1 \\ c & 1 \end{pmatrix},$$

$$S_{L,\xi} = \begin{pmatrix} 0 & \frac{1}{\xi+c} \\ \frac{1}{\xi+c} & -\frac{1}{\xi+c} \end{pmatrix}, \quad T_{L,\xi} = \begin{pmatrix} -(\xi+c) & 0 \\ \xi+1 & 1 \end{pmatrix}$$

und den Bezeichnungen aus Abschnitt 3.2 erhält man ($n_1 = n_2 = 1$, $m_1 = m_2 = 1$)

$$\begin{pmatrix} \hat{u}_{k1} \\ \hat{u}_{k2} \end{pmatrix} = T_{F,k} \begin{pmatrix} y_k \\ z_k \end{pmatrix} = \begin{pmatrix} -z_k \\ cy_k + z_k \end{pmatrix}, \tag{3.3.1}$$

$$\begin{pmatrix} u_{\xi 1} \\ u_{\xi 2} \end{pmatrix} = T_{L,\xi} \begin{pmatrix} v_\xi \\ w_\xi \end{pmatrix} = \begin{pmatrix} -(\xi + c)v_\xi \\ (\xi + 1)v_\xi + w_\xi \end{pmatrix}. \tag{3.3.2}$$

Gleichung (3.3.1) legt für alle $k = 1, 2, ..$ die Komponente $\hat{u}_{k1}(t)$ und somit $u_1(t, x)$ als diejenige fest, für die wir keine Anfangsbedingung vorschreiben können, d.h. $\mathfrak{M}_{AB} \equiv \mathfrak{M}^k_{AB} = \{2\}$ (nur $\theta_{k2} \neq 0$). Andererseits können die Randbedingungen für $v_\xi(x)$ beliebig vorgegeben werden, und wir folgern aus (3.3.2), daß $\mathfrak{M}_{RB} \equiv \mathfrak{M}^\xi_{RB} = \{1\}$. Somit können wir für $u_{\xi 2}(x)$ (und folglich auch für $u_2(t, x)$) keine Randbedingungen vorschreiben, da sich die Randwerte für $u_{\xi 2}$ aus der zweiten Komponente in Gleichung (3.3.2) berechnen. Andererseits ist es auch denkbar, daß wir die RBn für $u_{\xi 2}$ beliebig vorschreiben, d.h. $\mathfrak{M}_{RB} \equiv \mathfrak{M}^\xi_{RB} = \{2\}$. Dann ist $u_{\xi 1} = -\frac{\xi+c}{\xi+1}(u_{\xi 2} - w_\xi)$. D.h., die Menge $\mathfrak{M}_{RB}$ ist nicht eindeutig bestimmt. Allerdings sollte man in diesem Beispiel $\mathfrak{M}_{RB} = \{1\}$ wählen, da wir in Hinblick auf die numerische Lösung nur für diese Komponente RBn zur Diskretisierung der Ortsableitung benötigen (vgl. Abschnitt 3.5). Wir schließen weiterhin, daß dieses Beispiel einen einheitliche differentiellen Zeitindex $\nu_{d,t} = 1$ und $\nu_{d,x} = 1$ besitzt. □

Im folgenden Beispiel wollen wir die Berechnung konsistenter Randbedingungen demonstrieren.

Beispiel 3.3.3. Seien in (3.1.1)

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & -1 \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix}, \qquad b < 0. \tag{3.3.3}$$

Da A regulär ist, folgt $\mathfrak{M}_{AB} = \{1, 2\}$. D.h., für alle Komponenten von u können beliebige Anfangsbedingungen vorgeschrieben werden. Zur Bestimmung von $\mathfrak{M}_{RB}$ und konsistenten Randbedingungen führen wir eine Laplace-Transformation der PDAE durch (vgl. Abschnitt 3.2.1) und erhalten eine Differentialgleichung der Form (3.2.2). Die Kronecker-transformierte Differentialgleichung lautet

$$\begin{pmatrix} v''_\xi \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{\xi^2}{b-\xi} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_\xi \\ w_\xi \end{pmatrix} = S_{L,\xi} \tilde{g}_\xi, \quad \text{wobei} \quad \begin{pmatrix} u_{\xi,1} \\ u_{\xi,2} \end{pmatrix} = \begin{pmatrix} -\frac{\xi^2}{\xi-b} v_\xi - w_\xi \\ \frac{b\xi}{\xi-b} v_\xi + w_\xi \end{pmatrix}.$$

Hieraus ergibt sich, daß $\nu_{d,x} = 1$ und daß die Menge $\mathfrak{M}_{RB}$ nicht eindeutig bestimmt ist. Wir können entweder $\mathfrak{M}_{RB} = \{1\}$ oder $\mathfrak{M}_{RB} = \{2\}$ wählen. Wir nehmen den

ersten Fall. Nach (3.1.1) ergeben sich die Randwerte für u_2 aus der Lösung des AWPs

$$\frac{d}{dt}u_2(t,\pm l) = g_2(t,\pm l) - bu_1(t,\pm l) \tag{3.3.4}$$

mit der gegebenen Anfangsbedingung $u_2(0,\pm l) = u_{0,2}(\pm l)$. Die Lösung dieser ODE ergibt die konsistenten Randbedingungen für u_2. □

Im folgenden Beispiel sind Anfangswerte von Komponenten u_j mit $j \notin \mathfrak{M}_{AB}$ als Parameter in der Laplace-transformierten PDAE vorhanden. Diese Parameter treten nach einer Rücktransformation in der Lösung der PDAE nicht mehr auf.

Beispiel 3.3.4. Wir betrachten erneut Beispiel 3.1.4. Für dieses erhält man mit den Bezeichnungen aus Abschnitt 3.2

$$S_{F,k} = \begin{pmatrix} 1 & 0 & -\frac{c_1}{c_3} \\ 0 & \frac{1}{c_2} & \frac{\rho_k b - a}{c_2 c_3} \\ 0 & 0 & \frac{a}{c_2 c_3} \end{pmatrix}, \qquad T_{F,k} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \frac{c_2}{a} \\ 0 & 1 & 1 \end{pmatrix},$$

$$S_{L,\xi} = \begin{pmatrix} 1 & 0 & -\frac{c_1}{c_3} \\ 0 & \frac{1}{c_2} & \frac{b - \xi a}{c_2 c_3} \\ 0 & 0 & -\frac{b}{c_2 c_3} \end{pmatrix}, \qquad T_{L,\xi} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & -\frac{c_2}{b} \\ 0 & 1 & 1 \end{pmatrix},$$

$$N_{F,k} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ R_{F,k} = -\rho_k, \qquad N_{L,\xi} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ R_{L,\xi} = -\xi$$

und somit $\nu_{d,t} = 2$ ($n_1 = 1, n_2 = 2$), $\nu_{d,x} = 3$ ($m_1 = 1, m_2 = 2$). Insbesondere sieht man anhand der Matrizen $T_{F,k}, T_{L,\xi}$, daß $\mathfrak{M}_{RB} = \mathfrak{M}_{AB} = \{1\}$ gilt. Wir betrachten daher nur noch die verbleibenden Komponenten u_j, $j \in \{2,3\}$, von u. Es gilt mit Gleichung (3.2.6)

$$\begin{aligned} \begin{pmatrix} u_{\xi 2} \\ u_{\xi 3} \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{c_2}{b} \\ 1 & 1 \end{pmatrix} w_\xi = \begin{pmatrix} 0 & -\frac{c_2}{b} \\ 1 & 1 \end{pmatrix} \left(r_{\xi,2} - N_{L,\xi} r'_{\xi,2} \right) \\ &= \begin{pmatrix} \frac{1}{c_3} g_{\xi 3} \\ \frac{1}{c_2}(g_{\xi 2} + a u_{02}) - \frac{\xi a}{c_2 c_3} g_{\xi 3} + \frac{b}{c_2 c_3} g''_{\xi 3} \end{pmatrix}. \end{aligned}$$

Die inverse Laplace-Transformation liefert:

$$u_2(t,x) = \mathfrak{L}^{-1}\left\{u_{\xi 2}(x)\right\} = \mathfrak{L}^{-1}\left\{\frac{1}{c_3} g_{\xi 3}(x)\right\} = \frac{1}{c_3} g_3(t,x)$$

$$u_3(t,x) = \mathfrak{L}^{-1}\{u_{\xi 3}(x)\} =$$
$$\ldots = \frac{1}{c_2}\left(g_2(t,x) + \frac{b}{c_3}g_{3,xx}(t,x) + a\mathfrak{L}^{-1}\left\{u_{02}(x) - \frac{\xi}{c_3}g_{\xi 3}(x)\right\}\right),$$

woraus $u_2(0,x) = \frac{1}{c_3}g_3(0,x)$ folgt. Es ist

$$\mathfrak{L}^{-1}\left\{u_{02}(x) - \frac{\xi}{c_3}g_{\xi 3}(x)\right\} = \frac{1}{c_3}\mathfrak{L}^{-1}\{g_3(0,x) - \xi g_{\xi 3}(x)\} = -\frac{1}{c_3}g_{t,3}(t,x)$$

und somit $u_3(t,x) = \frac{1}{c_2}g_2(t,x) - \frac{a}{c_2c_3}g_{3,t}(t,x) + \frac{b}{c_2c_3}g_{3,xx}(t,x)$. Dieses Beispiel zeigt, daß als Parameter vorhandene Anfangswerte von Komponenten u_j mit $j \notin \mathfrak{M}_{AB}$ in den rechten Seiten der transformierten PDAEs (3.2.2) in die Rücktransformation einfließen und in der Lösung u nicht auftreten. Andererseits verbleiben z.B. die Anfangswerte für u_i mit $i \in \mathfrak{M}_{AB}$ als frei wählbare Parameter in der Lösungsdarstellung. □

Es zeigt sich, daß die Mengen $\mathfrak{M}_{AB}$ und $\mathfrak{M}_{RB}$ nicht immer eindeutig bestimmt sind. Es sei bemerkt, daß man auch einen anderen Weg bez. der Konsistenz von Anfangs- und Randbedingungen verfolgen kann. Gibt man unabhängig von den Mengen $\mathfrak{M}_{AB}$ und $\mathfrak{M}_{RB}$ für alle Komponenten von u Anfangs- und Randbedingungen vor, so überprüfe man, ob diese konsistent sind (vgl. Definitionen 3.2.3 und 3.2.7).

3.4. Eine konsistente Darstellung der Lösung

In diesem Abschnitt soll eine konsistente Darstellung der Lösung des linearen ARWP (3.1.1)-(3.1.4) gegeben werden unter der Voraussetzung, daß dieses einen einheitlichen differentiellen Zeitindex $\nu_{d,t}$ und einen differentiellen Ortsindex $\nu_{d,x}$ besitzt.

Die Anfangs- bzw. Randwerte der Komponenten $u_j(t,x)$ mit $j \notin \mathfrak{M}_{AB}$ bzw. $j \notin \mathfrak{M}_{RB}$ seien wie bereits beschrieben bestimmt (vgl. z.B. Beispiel 3.3.3).
Sei $\theta : [0,\infty) \times [-l,l] \to \mathbb{R}^n$ eine glatte Vektorfunktion mit stetigen partiellen Ableitungen θ_t, θ_{xx}, für die

$$\theta(t,\pm l) = u(t,\pm l) \tag{3.4.1}$$

gilt, d.h. insbesondere

$$\theta_j = 0 \quad \text{für} \quad j \in \mathfrak{M}_{RB}. \tag{3.4.2}$$

Sei $U(t,x)$ definiert durch

$$U(t,x) := u(t,x) - \theta(t,x). \tag{3.4.3}$$

Setzen wir $u = U + \theta$ in Gleichung (3.1.1) ein, so ergibt sich die PDAE

$$A\,U_t(t,x) + B\,U_{xx}(t,x) + C\,U(t,x) = G(t,x), \tag{3.4.4}$$

wobei $G := g - A\,\theta_t - B\,\theta_{xx} - C\,\theta$. Die Anfangsbedingung ist gegeben durch

$$U_j(0,x) = u_{0j}(x) - \theta_j(0,x), \qquad j \in \mathfrak{M}_{AB}. \tag{3.4.5}$$

Offensichtlich genügt U homogenen Dirichlet-Randbedingungen in alle Komponenten, d.h.

$$U(t,-l) = U(t,l) = 0. \tag{3.4.6}$$

Da nach Voraussetzung sowohl die Lösung der PDAE $u(t,x)$ als auch $\theta(t,x)$ stetig sind, ist $U(t,x)$ stetig. Somit ist U in eine Fourierreihe bez. der ϕ_k

$$U(t,x) = \sum_{k=1}^{\infty} \hat{U}_k(t)\phi_k(x) \;=\; \sum_{k=1}^{\infty} \hat{U}_k(t) \sin\left(\tfrac{k\pi}{2l}(x+l)\right) \tag{3.4.7}$$

mit $\hat{U}_k$ aus (3.2.8) entwickelbar. Sei nun vorausgesetzt, daß G in eine Fourierreihe bez. der ϕ_k entwickelbar ist mit den entsprechenden Fourierkoeffizienten $\hat{G}_k$. Multiplikation von (3.4.4) mit $\frac{1}{l}\phi_k(x)$ und anschließende Integration bez. x von $x = -l$ bis $x = l$ liefert

$$A\,\hat{U}_k'(t) + (\rho_k\,B + C)\,\hat{U}_k(t) = \hat{G}_k(t), \quad k = 1,2,\dots, \tag{3.4.8}$$

$$\text{mit} \qquad \rho_k = -\left(\tfrac{k\pi}{2l}\right)^2 \quad \text{und} \quad U_{kj0} = U_{kj}(0), \quad j \in \mathfrak{M}_{AB}.$$

Nach Abschnitt 3.2 existiert unter den Voraussetzungen 3.2.1 d) und 3.2.1 e) eine eindeutige Lösung $\hat{U}_k(t)$ der DAE (3.4.8) für $k = 1,2,\dots$. Nach (3.4.3) gilt

$$u(t,x) = \sum_{k=0}^{\infty} \hat{U}_k(t)\phi_k(x) + \theta(t,x). \tag{3.4.9}$$

Aufgrund der Voraussetzungen an u und θ folgt die Stetigkeit von U_t und U_{xx} und die gleichmäßige Konvergenz der Reihen

$$U_t(t,x) = \sum_{k=1}^{\infty} \hat{U}_k'(t)\phi_k(x) \quad \text{bzw.} \quad U_{xx}(t,x) = \sum_{k=1}^{\infty} \rho_k \hat{U}_k(t)\phi_k(x). \tag{3.4.10}$$

Einsetzen von (3.4.9) in die linke Seite der PDAE (3.1.1) und gliedweise Differentiation liefert

$$\begin{aligned}
&A\,u_t(t,x) + B\,u_{xx}(t,x) + C\,u(t,x) \\
&= \sum_{k=0}^{\infty} \Big(A\,\hat{U}_k'(t) + (\rho_k\,B + C)\,\hat{U}_k(t)\Big)\,\phi_k(x) + A\,\theta_t(t,x) + B\,\theta_{xx}(t,x) + C\,\theta(t,x) \\
&= \sum_{k=0}^{\infty} \hat{G}_k(t)\phi_k(x) + A\,\theta_t(t,x) + B\,\theta_{xx}(t,x) + C\,\theta(t,x)
\end{aligned}$$

$$= \sum_{k=0}^{\infty} \hat{G}_k(t)\phi_k(x) - G(t,x) + g = g.$$

D.h., die Darstellung der Lösung u in der Form (3.4.9) erfüllt die PDAE (3.1.1). Wegen (3.4.6) und (3.4.2) erfüllt u die RBn (3.1.2) und nach (3.4.5) die ABn (3.1.3). Die ABn und die RBn für die Komponenten $u_j(t,x)$ mit $j \notin \mathfrak{M}_{AB}$ bzw. $j \notin \mathfrak{M}_{RB}$ sind wegen (3.4.1) und (3.4.3) nach Konstruktion konsistent.

Beispiel 3.4.1. Wir betrachten erneut Beispiel 3.3.3. Dort haben wir die konsistenten Randbedingungen berechnet, die wir für die Transformation (3.4.3) benötigen. Somit können wir diese Transformation durchführen und U kann in der Form (3.4.7) dargestellt werden. □

Beispiel 3.4.2. Wir führen die Betrachtungen aus Beispiel 3.3.4 fort. Wir können

$$\theta(t,x) := \begin{pmatrix} 0 \\ \frac{1}{c_3} g_3(t,x) \\ \frac{1}{c_2} g_2(t,x) - \frac{a}{c_2 c_3} g_{3,t}(t,x) + \frac{b}{c_2 c_3} g_{3,xx}(t,x) \end{pmatrix}$$

definieren und erhalten $\theta(t,\pm l) = u(t,\pm l)$, d.h., für U in (3.4.3) ist $U(t,\pm l) = 0$. Weiterhin sei

$$G := g - A\,\theta_t - B\,\theta_{xx} - C\,\theta \;=\; \left(g_1 - \frac{c_1}{c_3} g_3,\; 0,\; 0\right)^{\top}.$$

Angenommen $g_1 - \frac{c_1}{c_3} g_3$ und $u_1(0,x) = u_{01}(x)\sin(\pi x)$ besitzen eine Fourierentwicklung bez. der ϕ_k, so erhalten wir aus der Kronecker-Transformation (vgl. Abschnitt 3.2.2) für die Lösung von (3.4.8)

$$\hat{U}_{1,k}(t) = y_k(t) = \; e^{\rho_k t}\hat{U}_{1,k}(0) + \int_0^t e^{\rho_k (t-\tau)} \left(\hat{g}_{1,k}(\tau) - \frac{c_1}{c_3}\hat{g}_{3,k}(\tau)\right) d\,\tau$$

$$\begin{pmatrix} \hat{U}_{2,k}(t) \\ \hat{U}_{3,k}(t) \end{pmatrix} = \begin{pmatrix} 0 & \frac{c_2}{a} \\ 1 & 1 \end{pmatrix} z_k(t) = \begin{pmatrix} 0 & \frac{c_2}{a} \\ 1 & 1 \end{pmatrix} \left(\underbrace{s_{k,2}}_{=0} - \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \underbrace{s'_{k,2}}_{=0} \right) \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

und schließlich die Darstellung (3.4.9)

$$u(t,x) = \sum_{k=1}^{\infty} \begin{pmatrix} \hat{U}_{1,k}(t) \\ 0 \\ 0 \end{pmatrix} \sin\left(\frac{k\pi}{2l}(x+l)\right) + \theta(t,x).$$

für die Lösung der PDAE, welche identisch ist mit der in Beispiel 3.1.4 gegebenen Lösung. □

3.5. Zwei Diskretisierungsverfahren zur numerischen Behandlung von linearen PDAEs

Zur numerischen Lösung der linearen PDAE (3.1.1) werden zwei auf der (vertikalen) Linienmethode basierende Diskretisierungsverfahren betrachtet, die im Fall parabolischer Differentialgleichungssysteme in das *BTCS Schema* und das *Crank-Nikolson Verfahren* übergehen. In Abhängigkeit von den eingeführten Indexen werden Konsistenz- und Konvergenzaussagen getroffen. Wir setzen stets voraus, daß wir konsistente Anfangs- und Randbedingungen gegeben haben.

3.5.1. Ortsdiskretisierung und Konvergenz

Für die Ortsdiskretisierung werden finite Differenzen verwendet und die partiellen Ortsableitungen u_{xx} mittels des zentralen Differenzenquotienten 2. Ordnung (1.1.6) approximiert (vgl Abschnitt 1.1.1). Hierzu legen wir auf das Ortsintervall $[-l, l]$ das äquidistantes Gitter (1.1.2), d.h.

$$\Omega_h = \left\{ x_k \ : \ x_k = -l + k\,h,\ k = 0(1)M+1,\ h = \frac{2l}{M+1} \right\},$$

wobei $M \in \mathbb{N}^+$ die Anzahl der inneren Ortsgitterpunkte ist. Für $k \in \{1, \ldots, M\}$ erhalten wir nach Einsetzen der Approximation (1.1.8) in (3.1.1) die *semidiskrete Gleichung*

$$A\,w_k'(t) + \frac{1}{h^2} B\ (w_{k+1}(t) - 2w_k(t) + w_{k-1}(t)) + C\,w_k(t) = g_k(t), \tag{3.5.1}$$

wobei $g_k(t) := g(t, x_k)$ und $w_k(t)$ eine Approximation an $u(t, x_k)$ ist. Unter Verwendung des Kronecker-Produktes $\otimes$ (siehe Seite 110) und den Bezeichnungen

$$\begin{aligned}
w(t) &:= (w_1^\top(t), \ldots, w_M^\top(t))^\top \in \mathbb{R}^{nM}, \\
r(t) &:= \left(\frac{1}{h^2} I_M \otimes B \right) (u^\top(t, -l), 0 \ldots, 0, u^\top(t, l))^\top \in \mathbb{R}^{nM}, \\
G(t) &:= (g_1^\top(t), \ldots, g_M^\top(t))^\top \in \mathbb{R}^{nM}, \\
\tilde{G}(t) &:= G(t) - r(t)
\end{aligned}$$

erhält man das *semidiskrete Problem* der PDAE (3.1.1)-(3.1.3) in Matrixschreibweise

$$\begin{aligned}
(I_M \otimes A)\, w'(t) + \left(\frac{1}{h^2} P \otimes B + I_M \otimes C \right) w(t) &= \tilde{G}(t) \\
w(0) &= w_0.
\end{aligned} \tag{3.5.2}$$

Hierbei ist I_M die $(M \times M)$-Einheitsmatrix, die Matrix P durch

$$P = \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & & \ddots & \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{M\times M} \tag{3.5.3}$$

definiert und $w_0 := (\tilde{u_0}^\top(x_1), \ldots, \tilde{u_0}^\top(x_M))^\top \in \mathbb{R}^{nM}$der mit (3.5.2) konsistente Anfangswert. Die Differenz $\tilde{u}_0 - u_0$ verschwinde für $h \to 0$ komponentenweise.

Bemerkung 3.5.1. Die Gestalt des semidiskreten Problems (3.5.2) beruht auf der von uns gewählten Anordnung der Komponenten des Vektors $w(t)$. Hierbei haben wir eine Anordnung bez. der Komponenten vorgenommen [Str92]. Wählt man eine Anordnung der Komponenten nach den Gitterpunkten, d.h., wenn $\bar{w}_i(t) \approx (u_i(t, x_1), \ldots, u_i(t, x_M))^\top$, $i = 1(1)n$, und $\bar{w} := (\bar{w}_1^\top, \ldots, \bar{w}_n^\top)^\top$ ($\bar{G}$ entsprechend), so lautet das semidiskrete System

$$(A \otimes I_M)\, \bar{w}'(t) + \left(B \otimes \frac{1}{h^2} P + C \otimes I_M \right) \bar{w}(t) = \bar{G}(t), \quad \bar{w}(0) = \bar{w}_0, \tag{3.5.4}$$

das dem System (3.5.2) äquivalent ist. □

Der lokale Ortsdiskretisierungsfehler (vgl. Definition 1.1.2) ist gegeben durch:

$$\alpha_h(t) = (I_M \otimes A) u_h'(t) + \left(\frac{1}{h^2} P \otimes B + I_M \otimes C \right) u_h(t) - \tilde{G}(t) \tag{3.5.5}$$

Sei $u_{xx,h}$ die Restriktion von u_{xx} auf das Ortsgitter (vgl. Abschnitt 1.1.1). Aus (1.1.6) erhält man

$$\left(\frac{1}{h^2} P \otimes B \right) u_h(t) + r(t) = (I_M \otimes B)\, u_{xx,h}(t) + h^2 (I_M \otimes B)\, \beta_h(t), \tag{3.5.6}$$

wobei $\beta_h(t) := (\beta_1(t), \ldots, \beta_{nM}(t))^\top$ mit

$\beta_{i+(k-1)n}(t) := \frac{1}{24}\left(\frac{\partial^4}{\partial x^4} u_i(t, \zeta_{ik}) + \frac{\partial^4}{\partial x^4} u_i(t, \bar{\zeta}_{ik}) \right), \quad \zeta_{ik} \in (x_{k-1}, x_k), \bar{\zeta}_{ik} \in (x_k, x_{k+1})$

für $i = 1(1)n,\ k = 1(1)M$. Mit (3.5.2) folgt somit

$$\alpha_h(t) = h^2 \left(I_M \otimes B \right) \beta_h(t). \tag{3.5.7}$$

Unter der Annahme[1] $\max\limits_{-l \le x \le l} \left\| \frac{\partial^4}{\partial x^4} u(t, x) \right\|_{2,n} \le C$ $(C > 0)$ für $t \in \mathfrak{I}^*$, gilt

$$\|\alpha_h(t)\| = \mathcal{O}(h^2), \qquad t \in \mathfrak{I}^*, \tag{3.5.8}$$

d.h., die Ortsdiskretisierung ist auf $\mathfrak{I}^*$ von zweiter Ordnung konsistent.

[1] $\|.\|_{2,n}$ Euklidische Vektornorm im $\mathbb{R}^n$, siehe Seite 110

Aus (3.5.2) und (3.5.5) ergibt sich für den globalen Ortsdiskretisierungsfehler $\eta_h(t)$ die DAE

$$\begin{aligned}(I_M \otimes A)\eta_h'(t) + \left(\frac{1}{h^2}P \otimes B + I_M \otimes C\right)\eta_h(t) &= \alpha_h(t)\\ \eta_h(0) &= u_h(0) - w(0).\end{aligned} \tag{3.5.9}$$

Dies zeigt, daß der globale Ortsdiskretisierungsfehler vom lokalen Ortsdiskretisierungsfehler abhängt.

Für die weiteren Untersuchungen werden folgende Bezeichnungen eingeführt:

$$\begin{aligned}&\tilde{A} = I_M \otimes A, && \tilde{B} = \frac{1}{h^2}P \otimes B + I_M \otimes C,\\ &\hat{A} = (c\tilde{A} + \tilde{B})^{-1}\tilde{A}, && \hat{B} = (c\tilde{A} + \tilde{B})^{-1}\tilde{B}, && \hat{\alpha}_h(t) = (c\tilde{A} + \tilde{B})^{-1}\alpha_h(t),\end{aligned}$$

wobei c so gewählt sei, daß $(c\tilde{A} + \tilde{B})$ invertierbar ist. Weiterhin sei $\hat{A}^D$ bzw. $\hat{B}^D$ die Drazin-Inverse von $\hat{A}$ bzw. $\hat{B}$ ([Gri86],[Ans87],[Sim93]). Dann ist die eindeutige Lösung von (3.5.9) gegeben durch

$$\begin{aligned}\eta_h(t) = e^{-\hat{A}^D\hat{B}\,t}\hat{A}\hat{A}^D(u_h(0) - w(0)) + \hat{A}^D\int_0^t e^{-\hat{A}^D\hat{B}(t-s)}\,\hat{\alpha}_h(s)\,d\,s\\ + (I - \hat{A}\hat{A}^D)\sum_{l=0}^{\nu_{d,t}-1}(-1)^l(\hat{A}\hat{B}^D)^l\hat{B}^D\hat{\alpha}_h^{(l)}(t)\end{aligned} \tag{3.5.10}$$

wobei

$$\begin{aligned}\eta_h(0) = \hat{A}\hat{A}^D(u_h(0) - w(0))\\ + (I - \hat{A}\hat{A}^D)\sum_{l=0}^{\nu_{d,t}-1}(-1)^l(\hat{A}\hat{B}^D)^l\hat{B}^D\hat{\alpha}_h^{(l)}(0),\end{aligned} \tag{3.5.11}$$

gilt, d.h., die Anfangswerte sind konsistent (siehe [Cam76]). Wenn A regulär ist, so ist $\hat{A}^D = \hat{A}^{-1}$ und die Beziehung (3.5.11) ist trivial.

Für eine Matrix sei $\|.\|$ die durch die diskrete L_2-Norm induzierte Matrixnorm und $\mu[.]$ die zugehörige logarithmische Matrixnorm. Mit $\mu_h = \mu[-\hat{A}^D\hat{B}]$ erhalten

wir aus (3.5.10) die Abschätzung

$$\begin{aligned}\|\eta_h(t)\| &\le e^{\mu_h t}\|\hat{A}\hat{A}^D\|\|u_h(0)-w(0)\| + \|\hat{A}^D\|\int_0^t e^{\mu_h(t-s)}\|\hat{\alpha}_h(s)\|\,d\,s\\ &\quad + \|I-\hat{A}\hat{A}^D\|\|\hat{B}^D\|\sum_{l=0}^{\nu_{d,t}-1}\|(\hat{A}\hat{B}^D)\|^l\|\hat{\alpha}_h^{(l)}(t)\|\\ &\le e^{\mu_h t}\|\hat{A}\hat{A}^D\|\|u_h(0)-w(0)\| + \frac{e^{\mu_h t}-1}{\mu_h}\|\hat{A}^D\|\max_{0\le\tau\le t}\|\hat{\alpha}_h(\tau)\|\\ &\quad + \|I-\hat{A}\hat{A}^D\|\|\hat{B}^D\|\max_{l=0}^{\nu_{d,t}-1}\|\hat{\alpha}_h^{(l)}(t)\|\sum_{l=0}^{\nu_{d,t}-1}\|(\hat{A}\hat{B}^D)\|^l,\quad 0\le t\le t_e.\end{aligned}$$

Somit haben wir folgendes Konvergenzergebnis.

Satz 3.5.2. *Die Ortsdiskretisierung sei von zweiter Ordnung konsistent. Sei weiterhin* $\|u_h(0)-w(0)\| = \mathcal{O}(h^2)$ *für* $h\to 0$ *und* $|\mu_h|$ *beschränkt für* $h\to 0$. *Dann ist die Ortsdiskretisierung von zweiter Ordnung konvergent, d.h.*

$$\|\eta_h(t)\| = \mathcal{O}(h^2) \quad \textit{für} \quad h\to 0,\quad t\in\mathfrak{I}^*.$$

Bemerkung 3.5.3. Wenn $(c\tilde{A}+\tilde{B})^{-1}$ für ein c existiert, dann ist die Lösung (3.5.10) unabhängig von c (vgl. [Cam76]). □

3.5.2. Index der MOL-DAE

Zwischen dem einheitlichen differentiellen Zeitindex der PDAE (3.1.1) und dem Differentiationsindex der MOL-DAE (3.5.2) gilt folgender Zusammenhang.

Satz 3.5.4. *Die PDAE (3.1.1) habe einen einheitlichen differentiellen Zeitindex* $\nu_{d,t}$. *Dann gilt für den Differentiationsindex* di *der linearen DAE (3.5.2)* $\mathrm{di} = \nu_{d,t}$ *für hinreichend kleine* h.

Die Behauptung dieses Satzes läßt sich aus den folgenden Betrachtungen herleiten. Seien die Eigenvektoren der Matrix $\frac{1}{h^2}P$ mit Φ_k, $k=1(1)M$, bezeichnet. Die Eigenwerte von $\frac{1}{h^2}P$ sind $\lambda_k = -\frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(M+1)}\right)$ (vgl. z.B. [Tho95]), d.h.

$$\frac{1}{h^2}P\,\Phi_k = \lambda_k\Phi_k,\qquad k=1(1)M. \tag{3.5.12}$$

Die Eigenvektoren

$$\Phi_k := \kappa_k\left(\sin\left(\frac{k\pi}{M+1}\right),\ldots,\sin\left(\frac{kM\pi}{M+1}\right)\right)^\top\in\mathbb{R}^M,$$

$\kappa_k\in\mathbb{R},\ \kappa_k\neq 0,\ k=1(1)M$, bilden eine orthogonale Basis des $\mathbb{R}^M$. Sei $\kappa_k=\sqrt{l^{-1}}$, $k=1(1)M$. Dann sind die Φ_k normiert bezüglich der im folgenden verwendeten

diskreten L_2-Norm $\|.\|$ (vgl. (1.1.4) mit $d = 1$), d.h. $\|\Phi_k\| = \sqrt{h\Phi_k^\top \Phi_k} = 1$ für $k = 1(1)M$.

Sei $\Phi := \sqrt{h}\,(\Phi_1 \dots \Phi_M) \in \mathbb{R}^{M\times M}$. Aufgrund der speziellen Struktur von P und der Orthonormalität der Φ_k gilt $\Phi^{-1} = \Phi^\top = \Phi$ und $\Phi\Phi_k = \sqrt{h^{-1}}e_k = (\sqrt{h^{-1}}\delta_{ki})_{i=1(1)M} \in \mathbb{R}^M$, $k = 1(1)M$, wobei δ_{ki} das Kronecker-Symbol ist. Sei $\Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_M)$. Mit (3.5.12) gilt dann $\Phi\frac{1}{h^2}P\Phi = \Lambda$. Nach Multiplikation von (3.5.2) von links mit $\Phi \otimes I_n$ und Anwendung von Eigenschaften des Kronecker-Produktes erhält man

$$\begin{aligned}&(\Phi \otimes A\,)w'(t) + \left(\tfrac{1}{h^2}\Phi P \otimes B + \Phi \otimes C\right) w(t)\\ &= (I_M \otimes A)(\Phi \otimes I_n)\, w'(t) + \left(\tfrac{1}{h^2}\Phi P\Phi \otimes B + I_M \otimes C\right)(\Phi \otimes I_n)\, w(t)\end{aligned}$$

und somit

$$(I_M \otimes A)\,\omega'(t) + (\Lambda \otimes B + I_M \otimes C)\,\omega(t) = (\Phi \otimes I_n)\,\tilde{G}(t), \tag{3.5.13}$$

wobei $\omega = (\omega_1^\top, \dots, \omega_M^\top)^\top := (\Phi \otimes I_n)\, w$ und $\omega_k \in \mathbb{R}^n$ $(k = 1(1)M)$. Dann ist $\omega_k = (\sqrt{h}\,\Phi_k^\top \otimes I_n)\, w$. Somit erhält man aus (3.5.2) ein System von DAEs, das in M entkoppelte DAEs der Form

$$\begin{aligned}A\,\omega_k'(t) + (\lambda_k\, B + C)\,\omega_k(t) &= (\sqrt{h}\,\Phi_k^\top \otimes I_n)\,\tilde{G}(t), \quad k = 1(1)M,\\ \omega_k(0) &= (\sqrt{h}\,\Phi_k^\top \otimes I_n)\, w_0\end{aligned} \tag{3.5.14}$$

zerfällt.

LEMMA 3.5.5. *Der Differentiationsindex der linearen DAE (3.5.2) ist gleich dem Maximum der Differentiationsindexe der linearen DAEs (3.5.14).*

BEWEIS. Da $\Phi \otimes I_n$ regulär ist, folgt mit Hilfe von Lemma 1.2.5

$$\begin{aligned}\nu &= \mathrm{ind}\left(I_M \otimes A, \tfrac{1}{h^2}P \otimes B + I_M \otimes C\right)\\ &= \mathrm{ind}\left((\Phi \otimes I_n)(I_M \otimes A)(\Phi \otimes I_n), (\Phi \otimes I_n)(\tfrac{1}{h^2}P \otimes B + I_M \otimes C)(\Phi \otimes I_n)\right)\\ &= \mathrm{ind}\left(I_M \otimes A, \Phi\tfrac{1}{h^2}P\Phi \otimes B + I_M \otimes C\right) = \mathrm{ind}\left(I_M \otimes A, \Lambda \otimes B + I_M \otimes C\right).\end{aligned}$$

Somit ist ν gleich dem Index des Systems (3.5.13), das die Aneinanderreihung der DAEs (3.5.14) ist. Man überlegt sich leicht, daß daher der Index dieses Systems gleich dem Maximum der Indexe der DAEs (3.5.14) ist, woraus die Behauptung folgt. □

Die Eigenwerte λ_k $(k \in \{1, \dots, M\})$ von P konvergieren für $h \to 0$ gegen die $\rho_k = -\left(\frac{k\pi}{2l}\right)^2$. Wenn die zugrundeliegende PDAE einen einheitlichen differentiellen Zeitindex $\nu_{d,t} = \nu_F$ hat, ist für alle Büschel $(A, \rho_k\, B + C)$ der Riesz-Index der gleiche, d.h. $\nu_{F,k} = \nu_F$. Dann ist für hinreichend kleine h auch $\mathrm{ind}(A, \lambda_k\, B + C) = \nu_{F,k} = \nu_F$.

Somit ist das Maximum $\max_{k=1(1)M}\{\mathrm{ind}(A, \lambda_k B+C)\} = \nu_F = \nu_{d,t}$, was die Aussage von Satz 3.5.4 beinhaltet.

3.5.3. Zeit-Diskretisierungen und Konvergenz der Gesamtdiskretisierung

In der Zeitintegration wird das semidiskrete Problem (3.5.2) mit einem Einschrittverfahren gelöst, dem *impliziten Euler-Verfahren* bzw. der *Trapez-Regel*. Das implizite Euler-Verfahren zur Lösung einer linearen ODE/DAE mit konstanten Koeffizienten

$$X\,w'(t) + Y\,w(t) = g(t), \quad w(0) = w_0, \quad t > 0,\ X, Y \in \mathbb{R}^{r\times r},\ w, g \in R^r, \quad (3.5.15)$$

ist gegeben durch die Verfahrensvorschrift

$$\begin{aligned} \left(\tfrac{1}{\tau_m}X + Y\right)\, v_{m+1} &= \tfrac{1}{\tau_m}X\, v_m + g(t_{m+1}) \\ v_0 &= w_0. \end{aligned} \quad (3.5.16)$$

Die Trapezregel zur Lösung von (3.5.15) hat die Vorschrift

$$\begin{aligned} \left(\tfrac{1}{\tau_m}X + \tfrac{1}{2}Y\right)\, v_{m+1} &= \left(\tfrac{1}{\tau_m}X - \tfrac{1}{2}Y\right)\, v_m + \tfrac{1}{2}\left(g(t_m) + g(t_{m+1})\right) \\ v_0 &= w_0. \end{aligned} \quad (3.5.17)$$

3.5.3.1. BTCS Schema. Verwendet man für die Zeitdiskretisierung des semidiskreten Problems (3.5.2) das implizite Euler-Verfahren (3.5.16), so wird in Analogie zur numerischen Lösung parabolischer ARWP die Gesamtdiskretisierung als *backward-in-time-centered in space (BTCS)* Schema bezeichnet. Sei in (3.5.15) $X = I_M \otimes A$ und $Y = \frac{1}{h^2}P \otimes B + I_M \otimes C$. Dann lautet das BTCS Schema zur Lösung von (3.1.1) mit den Bezeichnungen aus Abschnitt 3.5.1

$$\begin{aligned} \Gamma(\tau, h^2)\, v_{m+1} &= \left(\frac{1}{\tau}I_M \otimes A\right)\, v_m + \tilde{G}(t_{m+1}) \\ v_0 &= w_0, \end{aligned} \quad (3.5.18)$$

wobei
$$\Gamma(\tau, h^2) := \frac{1}{\tau}I_M \otimes A + \frac{1}{h^2}P \otimes B + I_M \otimes C. \quad (3.5.19)$$

v_{m+1} ist eindeutig aus v_m bestimmbar, wenn $\Gamma(\tau, h^2)$ regulär ist. Das folgende Lemma zeigt, daß die eindeutige Lösung von (3.5.18) direkt mit der Regularität der $n\times n$ Matrizen

$$\Gamma_k(\tau, h^2) = \tfrac{1}{\tau}A + \lambda_k\, B + C, \quad k = 1(1)M, \quad (3.5.20)$$

verbunden ist.

LEMMA 3.5.6. *Sind die Matrizen $\Gamma_k(\tau, h^2)$ $(k = 1(1)M)$ regulär für $0 < \tau \leq \tau_0$, $0 < h \leq h_0\,(\tau, h\,fest)$, dann ist Γ regulär, und es existiert eine eindeutige Lösung v_{m+1} von (3.5.18).*

BEWEIS. Mit den Matrizen Φ und Λ aus Abschnitt 3.5.2 und den Eigenschaften $\frac{1}{h^2}P = \Phi\Lambda\Phi$, $\Phi^{-1} = \Phi$ erhält man

$$\begin{aligned}\Gamma(\tau, h^2) &= (\Phi \otimes I_n)\left(\tfrac{1}{\tau} I_M \otimes A + \Lambda \otimes B + I_M \otimes C\right)(\Phi \otimes I_n) \\ &= (\Phi \otimes I_n)\begin{pmatrix} \Gamma_1(\tau,h^2) & & \\ & \ddots & \\ & & \Gamma_M(\tau,h^2) \end{pmatrix}(\Phi \otimes I_n),\end{aligned}$$

woraus mit der Regularität der Matrix Φ die Behauptung folgt. □

LEMMA 3.5.7. *Sei $\Gamma(\tau, h^2)$ regulär und $\nu_{d,t}$ bzw. $\nu_{d,x}$ der einheitliche differentielle Zeitindex bzw. der differentielle Ortsindex der PDAE (3.1.1). Dann gelten für die reguläre Matrix $\Gamma^{-1}(\tau, h^2)$ die folgenden Beziehungen $(i, j = 1(1)nM)$*

(i) *für $\nu_{d,t} \geq 0$ und festes $h > 0$*

$$\lim_{\tau \to 0} \tau^{\nu_{d,t}} \left(\Gamma^{-1}(\tau, h^2)\right)_{i,j} = 0, \tag{3.5.21}$$

(ii) *für $\nu_{d,t} \geq 1$ und festes $h > 0$*

$$\lim_{\tau \to 0} \tau^{\nu_{d,t}} \left(\Gamma^{-1}(\tau, h^2)\frac{1}{\tau}(I_M \otimes A)\right)_{i,j} = 0, \tag{3.5.22}$$

(iii) *für $\nu_{d,x} \geq 0$ und festes $\tau > 0$*

$$\lim_{h \to 0} h^{\nu_{d,x}} \left(\Gamma^{-1}(\tau, h^2)\right)_{i,j} = 0, \tag{3.5.23}$$

(iv) *für $\nu_{d,x} \geq 1$ und festes $\tau > 0$*

$$\lim_{h \to 0} h^{\nu_{d,x}} \left(\Gamma^{-1}(\tau, h^2)\frac{1}{h^2}(I_M \otimes B)\right)_{i,j} = 0. \tag{3.5.24}$$

BEWEIS. Wir betrachten das Matrixbüschel $(I_M \otimes A, \frac{1}{h^2}P \otimes B + I_M \otimes C)$, welches für hinreichend kleine h regulär ist. Wir führen eine Weierstraß-Kronecker-Transformation mit regulären Matrizen $S, T \in \mathbb{R}^{nM \times nM}$ durch. Dann kann Γ^{-1} geschrieben werden als

$$\Gamma^{-1}(\tau, h^2) = S \begin{pmatrix} \tau(I_{M_1} + \tau R)^{-1} & 0 \\ 0 & (I_{M_2} + \frac{1}{\tau}N)^{-1} \end{pmatrix} T,$$

wobei $R \in \mathbb{R}^{M_1 \times M_1}$, $N \in \mathbb{R}^{M_2 \times M_2}$, $M_1 + M_2 = nM$, und $N^{\nu_{d,t}} = 0$ nach Satz 3.5.4. Die Behauptungen (i) und (ii) folgen nun leicht aus der Transformation von $I_M \otimes A$ und aus der Identität $(I + \frac{1}{\tau}N)^{-1} = \sum_{l=0}^{\nu_{d,t}-1}(-\frac{1}{\tau}N)^l$. Analog zeigt man mit der

Weierstraß-Kronecker-Transformation des Matrixbüschels $\left(P \otimes B, I_M \otimes \left(\frac{1}{\tau}A + C\right)\right)$ die Behauptungen (iii) und (iv). □

Der einheitliche differentielle Zeitindex und der differentielle Ortsindex (vgl. Definitionen 3.2.2 und 3.2.8) haben für das BTCS Schema die folgende äquivalente Bedeutung:

Satz 3.5.8. *Der einheitliche differentielle Zeitindex $\nu_{d,t}$ bzw. der differentielle Ortsindex $\nu_{d,x}$ sind die kleinsten natürlichen Zahlen, so daß (3.5.21) bzw. (3.5.23) gilt.*

Definition 3.5.9. *Die Matrizen $\tau^{\nu_{d,t}}\Gamma^{-1}(\tau, h^2)$ bzw. $h^{\nu_{d,x}}\Gamma^{-1}(\tau, h^2)$, die den Beziehungen (3.5.21) bzw. (3.5.23) genügen, heißen* τ-proper *bzw.* h-proper.

Bemerkung 3.5.10. Von Campbell/Marszalek wurden in [Cam96a],[Mar97] die Proper-Eigenschaften der Matrix $R(s, z) = (sA + z^2B + C)^{-1}$ für $|s| \to \infty$ oder $z^2 \to \infty$ eingeführt. □

Anhand der Gleichungen (3.5.22) und (3.5.24) sehen wir, daß die Matrizen $\tau^{\nu_{d,t}}\Gamma^{-1}(\tau, h^2)\frac{1}{\tau}(I_M \otimes A)$ bzw. $h^{\nu_{d,x}}\Gamma^{-1}(\tau, h^2)\frac{1}{h^2}(I_M \otimes B)$ für $\nu_{d,t}, \nu_{d,x} \geq 1$ die gleiche Proper-Eigenschaften wie die Matrizen $\tau^{\nu_{d,t}}\Gamma^{-1}(\tau, h^2)$ bzw. $h^{\nu_{d,x}}\Gamma^{-1}(\tau, h^2)$ haben.

Bemerkung 3.5.11. Die Charakterisierung der Indexe von PDAEs durch die Proper-Eigenschaft der Matrix $\Gamma^{-1}(\tau, h^2)$ des BTCS Schemas ist von rein algebraischer Natur. Im Gegensatz dazu nutzen ihre Definitionen 3.2.2 und 3.2.8 die differentiellen Eigenschaften der algebraischen Lösungskomponente, die man aus einer Weierstraß-Kronecker-Transformation erhält. □

Es wird nun der lokale Gesamtdiskretisierungsfehler $le_h(t_{m+1})$ betrachtet (vgl. Definition 1.1.19). Für das BTCS Schema ist

$$\hat{v}_{m+1} = \Gamma^{-1}(\tau, h^2)\left(\left(\tfrac{1}{\tau}I_M \otimes A\right)u_h(t_m) + \tilde{G}(t_{m+1})\right).$$

Es sei bemerkt, daß dieser Fehler im Gegensatz zum lokalen Zeitdiskretisierungsfehler bezüglich der Lösung der PDAE $u_h(t)$ und nicht bez. der Lösung $w(t)$ der DAE definiert wurde (vgl. Abschnitt 1.1.3).

Definition 3.5.12. *Ein Diskretisierungsverfahren zur Lösung von (3.1.1) ist* konsistent *mit der PDAE (3.1.1), wenn der lokale Gesamtdiskretisierungsfehler die Beziehung*

$$\frac{1}{\tau}\|le_h(t_{m+1})\| \to 0 \quad \textit{für} \quad \tau,\, h \to 0, \quad m = 0, 1, \ldots \tag{3.5.25}$$

erfüllt.

DEFINITION 3.5.13. *Das BTCS Schema (3.5.18) ist* konsistent von der Ordnung (p,q), $p,q \geq 1$, *wenn*

$$\frac{1}{\tau}\|le_h(t_{m+1})\| = \mathcal{O}(h^p) + \mathcal{O}(\tau^q) \quad \textit{für} \quad \tau,\, h \to 0. \tag{3.5.26}$$

BEMERKUNG 3.5.14. Gelten die Beziehungen (3.5.25) bzw. (3.5.26) unter einer Orts-Zeit-Bedingung der Form

$$c_0 \leq \frac{\tau}{h^2} \quad \text{oder} \quad \frac{\tau}{h^2} \leq c_1 \quad \text{oder} \quad c_0 \leq \frac{\tau}{h^2} \leq c_1, \qquad c_0, c_1 \in \mathbb{R}^+, \tag{3.5.27}$$

so nennen wir das BTCS Schema *bedingt konsistent* bzw. *bedingt konsistent der Ordnung* (p,q). Wenn $c_0 = 0$ und $c_1 = \infty$, so haben wir keine Beschränkung an τ und h. □

Mit Gleichung (3.5.5) erhalten wir

$$\begin{aligned} le_h(t_{m+1}) = \Gamma^{-1}(\tau,h^2)\Big\{ (\tfrac{1}{\tau} I_M \otimes A)\Big[u_h(t_{m+1}) - u_h(t_m)\Big] \\ - (I_M \otimes A)u_h'(t_{m+1}) + \alpha_h(t_{m+1})\Big\}, \end{aligned}$$

woraus mit einer Taylor-Entwicklung von $u_h(t_{m+1})$ und $u_h'(t_{m+1})$ in t_m folgt

$$\begin{aligned} le_h(t_{m+1}) = \tau\,\Gamma^{-1}(\tau,h^2)(I_M \otimes A)\Big[\tfrac{1}{2}u_h''(t_m + \zeta\tau) - u_h''(t_m + \overline{\zeta}\tau)\Big] \\ + h^2\,\Gamma^{-1}(\tau,h^2)(I_M \otimes B)\,\beta_h(t_{m+1}) \end{aligned} \tag{3.5.28}$$

mit $\zeta, \overline{\zeta} \in (0,1)$ (u.U. komponentenweise verschieden). Daher gilt für den lokalen Gesamtdiskretisierungsfehler für $t \in \mathfrak{I}^* = [0,t^*)$, $t^* > 0$,die Abschätzung

$$\|le_h(t_{m+1})\| \leq C_0 \left(\tau\|\Gamma^{-1}(\tau,h^2)(I_M \otimes A)\| + h^2\|\Gamma^{-1}(\tau,h^2)(I_M \otimes B)\|\right), \tag{3.5.29}$$

wobei C_0 eine von τ und h unabhängige, positive Konstante ist. Die Terme der rechten Seite dieser Ungleichung können leicht mit Lemma 3.5.7 abgeschätzt werden, da dieses die folgenden Beziehungen für $\tau, h \to 0$ impliziert:

$$\begin{aligned} \tau^{\nu_{d,t}-1}\|\Gamma^{-1}(\tau,h^2)\| &\leq \kappa_1 \quad \text{für } \nu_{d,t} \geq 0, \\ h^{\nu_{d,x}-1}\|\Gamma^{-1}(\tau,h^2)\| &\leq \kappa_2 \quad \text{für } \nu_{d,x} \geq 0, \\ \tau^{\nu_{d,t}-2}\|\Gamma^{-1}(\tau,h^2)(I_M \otimes A)\| &\leq \kappa_3 \quad \text{für } \nu_{d,t} \geq 1, \\ h^{\nu_{d,x}-3}\|\Gamma^{-1}(\tau,h^2)(I_M \otimes B)\| &\leq \kappa_4 \quad \text{für } \nu_{d,x} \geq 1, \end{aligned} \tag{3.5.30}$$

wobei möglicherweise τ und h nicht unabhängig von einander gegen Null gehen können. D.h., eventuell muß eine der Bedingungen (3.5.27) erfüllt sein.

Die Abschätzung (3.5.29) und die asymptotischen Beziehungen (3.5.30) können wir nun verwenden, um Aussagen zur Konsistenz im Sinne von Definition 3.5.13 zu treffen:

SATZ 3.5.15. *Seien für eine gegebene PDAE (3.1.1) die asymptotischen Beziehungen (3.5.30) erfüllt. Sei $\nu_{d,t}$ der einheitliche differentielle Zeitindex und $\nu_{d,x}$ der differentielle Ortsindex der PDAE. Dann ist das BTCS Schema für $t \in \mathfrak{I}^*$ (bedingt) konsistent der Ordnung (2,1) für*

- $\nu_{d,t} = 0$ *und* $\nu_{d,x} \geq 0$,
- $\nu_{d,t} \geq 0$ *und* $\nu_{d,x} = 0$,
- $\nu_{d,t} = 1$ *und* $\nu_{d,x} = 1$

(eventuell unter einer Bedingung (3.5.27)).

BEWEIS. Im folgenden seien $K_i \geq 0$ Konstanten unabhängig von τ und h und die Argumente von $\Gamma^{-1}(\tau, h^2)$ zur Vereinfachung weggelassen.

Sei zunächst $\nu_{d,t} = 0, \nu_{d,x} \geq 0$. Mit der ersten Ungleichung in (3.5.30) können wir die rechte Seite von (3.5.29) weiter durch

$$K\left(\tau^2\|\tau^{-1}\Gamma^{-1}\| + \tau h^2\|\tau^{-1}\Gamma^{-1}\|\right) = \mathcal{O}(\tau^2) + \mathcal{O}(\tau h^2)$$

abschätzen, woraus die Behauptung für diese Indexe folgt.
Analog erhalten wir im Fall $\nu_{d,t} \geq 0$, $\nu_{d,x} = 0$ nach der zweiten Ungleichung in (3.5.30) die obere Schranke

$$K\left(\tau h^2\|h^{-2}\Gamma^{-1}\| + h^4\|h^{-2}\Gamma^{-1}\|\right) = \mathcal{O}(\tau h^2) + \mathcal{O}(h^4).$$

Mit der Bedingung $0 < \tilde{c}_0 \leq \frac{\tau}{h^2}$ reduziert sich die rechte Seite zu $\mathcal{O}(\tau^2) + \mathcal{O}(\tau h^2)$.
Die dritte Behauptung kann unter Verwendung der letzten beiden Ungleichungen in (3.5.30) gezeigt werden. Wir erhalten für die rechte Seite von (3.5.29) die Abschätzung

$$K\left(\tau^{3-\nu_{d,t}}\|\tau^{\nu_{d,t}-2}\Gamma^{-1}(I_M \otimes A)\| + h^{5-\nu_{d,x}}\|h^{\nu_{d,x}-3}\Gamma^{-1}(I_M \otimes B)\|\right).$$

Setzt man $\nu_{d,t} = \nu_{d,x} = 1$ in diese Schranke ein, so erhält man einen Ausdruck der Form $\mathcal{O}(\tau^2) + \mathcal{O}(h^4)$, der äquivalent zu $\mathcal{O}(\tau^2) + \mathcal{O}(\tau h^2)$ ist, da $0 < \tilde{c}_0 \leq \frac{\tau}{h^2}$, womit die dritte Behauptung des Satzes bewiesen ist. □

Es wird nun der Gesamtdiskretisierungsfehler $\varepsilon_h(t_{m+1})$ betrachtet (vgl. Definition 1.1.15).

DEFINITION 3.5.16. *Ein Diskretisierungsverfahren zur Lösung von (3.1.1) heißt* konvergent der Ordnung (p, q)*, wenn die asymptotische Beziehung*

$$\|\varepsilon_h(t_{m+1})\| = \mathcal{O}(h^p) + \mathcal{O}(\tau^q) \quad \textit{für} \quad \tau, h \to 0,\ m = 0, 1, ... \tag{3.5.31}$$

erfüllt ist. Gilt diese Beziehung nur unter einer Zeit-Orts-Bedingung (3.5.27), so heißt es bedingt konvergent der Ordnung (p, q).

Mit Gleichung (1.1.20) für den lokalen Gesamtdiskretisierungsfehler erhalten für wir $\varepsilon_h(t_{m+1})$ die Rekursion

$$\varepsilon_h(t_{m+1}) = \Gamma^{-1}(\tau, h^2)\left(\tfrac{1}{\tau} I_M \otimes A\right)\varepsilon_h(t_m) + le_h(t_{m+1}), \tag{3.5.32}$$

und durch Induktion erhält man

$$\begin{aligned}\varepsilon_h(t_{m+1}) = &\left(\Gamma^{-1}(\tau, h^2)\frac{1}{\tau} I_M \otimes A\right)^{m+1} \varepsilon_h(0) \\ &+ \sum_{i=0}^{m} \left(\Gamma^{-1}(\tau, h^2)\frac{1}{\tau} I_M \otimes A\right)^{i} le_h(t_{m+1-i}).\end{aligned} \tag{3.5.33}$$

Es folgt

$$\begin{aligned}\|\varepsilon_h(t_{m+1})\| \le &\left\|\left(\Gamma^{-1}(\tau, h^2)\frac{1}{\tau} I_M \otimes A\right)^{m+1}\right\| \|\varepsilon_h(0)\| \\ &+ \max_{i=0}^{m} \|le_h(t_{i+1})\| \sum_{i=0}^{m} \left\|\left(\Gamma^{-1}(\tau, h^2)\frac{1}{\tau} I_M \otimes A\right)^{i}\right\|.\end{aligned} \tag{3.5.34}$$

Voraussetzung 3.5.17. Für $0 < m\,\tau \le t^*$ gelte

$$\begin{aligned}\sup_{i\in\mathbb{N}} \Big\{ \left\| \left(\Gamma^{-1}(\tau, h^2)\tfrac{1}{\tau} I_M \otimes A\right)^{i} \right\| : \tau,\ h \ \text{ eventuell durch} \\ \text{eine Bedingung (3.5.27) eingeschränkt} \Big\} < \infty.\end{aligned} \tag{3.5.35}$$

Angenommen $\|\varepsilon_h(0)\| = \mathcal{O}(h^2)$, so erhalten wir für $t \in \mathfrak{I}^*$ die Abschätzung

$$\|\varepsilon_h(t_{m+1})\| \le \overline{K}\,(m+1) \max_{i=0}^{m} \|le_h(t_{i+1})\| \;\le\; K \max_{i=0}^{m} \frac{\|le_h(t_{i+1})\|}{\tau}, \tag{3.5.36}$$

wobei $\overline{K}$ und K positive Konstanten unabhängig von τ und h sind für $\tau,\ h \ \to 0$ unter der Bedingung $\tau\, m = t \in \mathfrak{I}^*$ (t fest). Mit Theorem 3.5.15 erhalten wir direkt

Satz 3.5.18. *Seien die Voraussetzung von Satz 3.5.15 und die Voraussetzung 3.5.17 erfüllt. Dann ist das BTCS Schema für $t \in \mathfrak{I}^*$ bez. der Norm $\|.\|$ (bedingt) konvergent der Ordnung (2,1) für*

- $\nu_{d,t} = 0$ *und* $\nu_{d,x} \ge 0$,
- $\nu_{d,t} \ge 0$ *und* $\nu_{d,x} = 0$,
- $\nu_{d,t} = \nu_{d,x} = 1$

(eventuell unter einer Bedingung (3.5.27)).

Neben dem Gesamtdiskretisierungsfehler $\varepsilon_h(t_{m+1})$ wollen wir eine diskrete Fourierkomponente dieses Fehlers in Abhängigkeit von den $n \times n$-Matrizen $\frac{1}{\tau}\Gamma_k^{-1}A$ und $\frac{1}{h^2}\Gamma_k^{-1}B$ betrachten. Es bezeichne $\|.\|_{2,n}$ die Euklidische Vektornorm im $\mathbb{R}^n$ bzw. für Matrizen die der Euklidischen Norm zugeordnete Spektralnorm (siehe Seite 110).

LEMMA 3.5.19. *Sei für festes $\tau, h > 0$*

(i) Γ_k^{-1} *regulär für* $k = 1(1)M$,

(ii) $u(t,x)$ *(und damit* $u_h(t)$ *und* $\beta_h(t)$*) hinreichend glatt für* $t \in \mathfrak{I}^*$, $x \in [-l, l]$,

(iii) *die Linearkombination von* $\varepsilon_h(t_j)$ *bez. der* Φ_k *gegeben durch*

$$\varepsilon_h(t_j) = \sum_{k=1}^{M} \Phi_k \otimes e_{h,k}^j, \quad j = 0, 1, ..., \text{ mit } e_{h,k}^j \in \mathbb{R}^n . \tag{3.5.37}$$

Dann kann $e_{h,k}^{m+1}$ *für* $t \in \mathfrak{I}^*$ *durch*

$$\begin{aligned} \|e_{h,k}^{m+1}\|_{2,n} \le & \left\| \left(\Gamma_k^{-1} \frac{A}{\tau} \right)^{m+1} \right\|_{2,n} \|e_{h,k}^0\|_{2,n} \\ & + K \sum_{i=0}^{m} \left\| \left(\Gamma_k^{-1} \frac{A}{\tau} \right)^{i} \right\|_{2,n} \left[h^2 \|\Gamma_k^{-1} B\|_{2,n} + \tau \left\|\Gamma_k^{-1} A\right\|_{2,n} \right] \end{aligned} \tag{3.5.38}$$

abgeschätzt werden, wobei die Konstante $K > 0$ *unabhängig von* τ *und* h *ist.*

BEWEIS. Wir schreiben Gleichung (3.5.32) in der Form

$$\Gamma(\tau, h^2)\, \varepsilon_h(t_{m+1}) = \tfrac{1}{\tau}(I_M \otimes A)\, \varepsilon_h(t_m) + \Gamma(\tau, h^2)\, l e_h(t_{m+1})$$

und multiplizieren sie von links mit $h(\Phi_k^\top \otimes I_n)$, $k \in \{1, \dots, M\}$. Wir verwenden für $\Gamma(\tau, h^2)\, l e_h(t_{m+1})$ Gleichung (3.5.28) und erhalten mit der Darstellung (3.5.37)

$$\begin{aligned} \Gamma_k\, e_{h,k}^{m+1} = & \frac{A}{\tau}\, e_{h,k}^m + h^2 (h\Phi_k^\top \otimes B) \beta_h(t_{m+1}) \\ & + \tau (h\Phi_k^\top \otimes A) \left[\tfrac{1}{2} u_h''(t_m + \zeta\tau) - u_h''(t_m + \overline{\zeta}\tau) \right] . \end{aligned}$$

Kombiniert man diesen Ausdruck mit den Voraussetzungen (i) und (ii), so erhält man die Behauptung durch Bildung von Abschätzung bez. der Norm. □

BEMERKUNG 3.5.20. Wie zu erwarten war, hat die Abschätzung (3.5.38) formal die gleichen τ- und h-Abhängigkeiten wie die Abschätzung (3.5.34). □

Wir werden nun unter Verwendung der Fehlerkomponente $e_{h,k}^{m+1}$ eine notwendige Bedingung für $\|\varepsilon_h(t_{m+1})\| \to 0$ für $\tau, h \to 0$ angeben.

LEMMA 3.5.21. *Seien die Voraussetzungen von Lemma 3.5.19 für* $0 < \tau \le \tau_0$, $0 < h \le h_0$ *erfüllt, und* $\|\varepsilon_h(t_{m+1})\|$ *strebe gegen Null für* $\tau, h \to 0$. *Dann ist die Beziehung*

$$\|e_{h,k}^{m+1}\|_{2,n} \to 0 \quad \textit{für} \quad \tau, h \to 0$$

erfüllt für $t \in \mathfrak{I}^*$. *Weiterhin ist die Konvergenzordnung (bez.* τ *und* h*) von* $\|e_{h,k}^{m+1}\|_{2,n}$ *die gleiche wie die Konvergenzordnung von* $\|\varepsilon_h(t_{m+1})\|$.

BEWEIS. Wir multiplizieren Gleichung (3.5.37) von links mit $h(\Phi_k^\top \otimes I_n)$, nutzen die Orthonormalität der Eigenvektoren Φ_k und erhalten

$$e_{h,k}^{m+1} = h(\Phi_k^\top \otimes I_n)\varepsilon_h(t_{m+1}).$$

Aus dieser Formel folgt die Behauptung des Lemmas. □

Im folgenden Beispiel nehmen wir an, daß die Fehlerkomponente $\|e_{h,k}^0\|$ zum Zeitpunkt $t = 0$ in Ungleichung (3.5.38) $\mathcal{O}(h^2)$ ist.

BEISPIEL 3.5.22. Sei $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} b_1 & b_2 \\ 0 & 0 \end{pmatrix}$ mit $b_1, b_2 < 0$, $C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Es ist leicht zu zeigen, daß für den einheitlichen differentiellen Zeitindex $\nu_{d,t} = 1$ und den differentiellen Ortsindex $\nu_{d,x} = 1$ gilt. In der Spektralnorm ist $\|\Gamma_k^{-1}\frac{A}{\tau}\|_{2,n} = \mathcal{O}(1)$ für $\tau, h \to 0$, wobei $\|\Gamma_k^{-1}B\|_{2,n} = \mathcal{O}(\tau)$ sowohl für $\frac{\tau}{h^2} \geq c_0$ als auch für $\frac{\tau}{h^2} \leq c_1$ ist. Daher wird keine Bedingung der Form (3.5.27) für dieses Beispiel benötigt. Man erhält

$$\left\|\Gamma_k^{-1}\frac{A}{\tau}\right\|_{2,n} = \frac{1}{|1 + \tau(\lambda_k b_1 + 1)|}.$$

Da $\|\Gamma_k^{-1}\frac{A}{\tau}\|_{2,n} \leq 1$ für τ, $h \to 0$, ist

$$\sum_{i=0}^{m} \left\| \left(\Gamma_k^{-1}\frac{A}{\tau}\right)^i \right\|_{2,n} \leq m + 1,$$

und mit $m\tau = t \in \mathfrak{I}^*$ (t fest) erhalten wir die Beziehung

$$\sum_{i=0}^{m} \left\| \left(\Gamma_k^{-1}\frac{A}{\tau}\right)^i \right\|_{2,n} = \mathcal{O}\left(\frac{1}{\tau}\right)$$

für $\tau, h \to 0$. Setzen wir dies in Ungleichung (3.5.38) ein und nehmen $\|e_{h,k}^0\|_{2,n} = \mathcal{O}(h^2)$ an, so erhalten wir für das BTCS Schema, angewandt auf dieses Beispiel, Konvergenz der Fehlerkomponenten $e_{h,k}^{m+1}$ der Ordnung

$$\|e_{h,k}^{m+1}\|_{2,n} = \mathcal{O}(\tau) + \mathcal{O}(h^2), \qquad \tau, h \to 0.$$

□

3.5.3.2. Crank-Nicolson Schema. Das *Crank-Nicolson Schema* entspricht der Gesamtdiskretisierung, die man aus der Linienmethode erhält, wenn man für die Diskretisierung der Ortsableitungen die Approximation (1.1.8) und für die Zeitintegration die Trapezregel (3.5.17) verwendet. Mit

$$\Gamma(\tau, h^2) := \frac{1}{\tau} I_M \otimes A + H_h, \qquad H_h := \frac{1}{2}\left(\frac{1}{h^2} P \otimes B + I_M \otimes C\right) \tag{3.5.39}$$

ist das Crank-Nicolson Schema zur Lösung von (3.1.1) unter Verwendung der Bezeichnungen aus Abschnitt 3.5.1 gegeben durch

$$\begin{aligned} \Gamma(\tau,h^2)\, v_{m+1} &= \left(\frac{1}{\tau} I_M \otimes A - H_h\right) v_m + \frac{1}{2}\left(\tilde{G}(t_m) + \tilde{G}(t_{m+1})\right) \\ v_0 &= w_0. \end{aligned} \tag{3.5.40}$$

Wir betrachten erneut den lokalen Gesamtdiskretisierungsfehler $le_h(t_{m+1})$, wobei nun

$$\hat{v}_{m+1} = \Gamma^{-1}(\tau,h^2)\left\{\left[\frac{1}{\tau} I_M \otimes A - H_h\right] u_h(t_m) + \frac{1}{2}\left[\tilde{G}(t_{m+1}) + \tilde{G}(t_m)\right]\right\}. \tag{3.5.41}$$

Analog zum vorherigen Abschnitt erhalten wir für $le_h(t_{m+1})$ die folgende Darstellung

$$\begin{aligned} le_h(t_{m+1}) &= \Gamma^{-1}(\tau,h^2)\Big\{ \left(\tfrac{1}{\tau} I_M \otimes A\right)\left[u_h(t_{m+1}) - u_h(t_m) - \tfrac{\tau}{2}\left(u_h'(t_{m+1}) + u_h'(t_m)\right)\right] \\ &\qquad + \tfrac{1}{2}\left(\alpha_h(t_m) + \alpha_h(t_{m+1})\right)\Big\} \\ &= \tau^2 \Gamma^{-1}(\tau,h^2)(I_M \otimes A)\left[\tfrac{1}{6} u_h'''(t_m + \zeta\tau) - \tfrac{1}{4} u_h'''(t_m + \overline{\zeta}\tau)\right] \\ &\qquad + h^2\, \Gamma^{-1}(\tau,h^2)(I_M \otimes B)\tfrac{1}{2}\left[\beta_h(t_m) + \beta_h(t_{m+1})\right] \end{aligned} \tag{3.5.42}$$

mit $\zeta, \overline{\zeta} \in (0,1)$ (u.U. komponentenweise verschieden). Dies führt zu der Abschätzung

$$\|le_h(t_{m+1})\| \leq C_1\left(\tau^2 \left\|\Gamma^{-1}(\tau,h^2)(I_M \otimes A)\right\| + h^2 \|\Gamma^{-1}(\tau,h^2)(I_M \otimes B)\|\right),$$

wobei $C_1 > 0$ und unabhängig von τ und h ist. Dies ist das Analogon zu Ungleichung (3.5.29) für das BTCS Schema.

Berücksichtigt man, daß die Propereigenschaft auch für die Matrix Γ^{-1} des Crank-Nicolson Schemas gilt (siehe Lemma 3.5.7), so erhält man analog zu Satz (3.5.15) aus obiger Abschätzung die folgenden Konsistenzeigenschaften:

Satz 3.5.23. *Seien für eine gegebene PDAE (3.1.1) die asymptotischen Beziehungen (3.5.30) für* Γ *aus (3.5.39) erfüllt. Dann ist das Crank-Nicolson Schema für* $t \in \mathfrak{I}^*$ *(bedingt) konsistent der Ordnung (2,2) für*

- $\nu_{d,t} = 0$ *und* $\nu_{d,x} \geq 0$,
- $\nu_{d,t} \geq 0$ *und* $\nu_{d,x} = 0$,
- $\nu_{d,t} = \nu_{d,x} = 1$

(möglicherweise unter einer der Bedingungen (3.5.27)). Es ist bedingt konsistent der Ordnung (2,1) für

- $\nu_{d,t} = 2$ *und* $\nu_{d,x} = 1$ *unter einer Bedingung* $0 < \hat{c}_0 \leq \frac{\tau}{h^2}$.

Um eine Konvergenzaussage für das Crank-Nicolson Schema (analog zu Satz 3.5.18 für das BTCS Schema) zu erhalten, betrachten wir den Gesamtdiskretisierungsfehler $\varepsilon_h(t_{m+1}) = u_h(t_{m+1}) - v_{m+1}$ unter der Voraussetzung, daß die Matrix $\Gamma(\tau, h^2)$ regulär ist. Da

$$v_{m+1} = \Gamma^{-1}\Big(\left[\frac{1}{\tau} I_M \otimes A - H_h\right] v_m + \frac{1}{2}\left(\tilde{G}(t_{m+1}) + \tilde{G}(t_m)\right)\Big)$$

(dies folgt aus Gleichung (3.5.40)) und $u_h(t_{m+1}) = le_h(t_{m+1}) + \hat{v}_{m+1}$ (siehe Gleichung (1.1.20)), können wir

$$\varepsilon_h(t_{m+1}) = \Gamma^{-1}\left(\frac{1}{\tau} I_M \otimes A - H_h\right)\varepsilon_h(t_m) + le_h(t_{m+1}) \tag{3.5.43}$$

schreiben, was analog zu Gleichung (3.5.32) ist. Daher erhalten wir das Äquivalent von Ungleichung (3.5.34) für das Crank-Nicolson Schema

$$\begin{aligned}\|\varepsilon_h(t_{m+1})\| \le &\left\| \left(\Gamma^{-1}(\tau, h^2)\left[\frac{1}{\tau} I_M \otimes A - H_h\right]\right)^{m+1}\right\| \cdot \|\varepsilon_h(0)\| \\ &+ \max_{j=0}^{m} \|le_h(t_{j+1})\| \sum_{i=0}^{m} \left\| \left(\Gamma^{-1}(\tau, h^2)\left[\frac{1}{\tau} I_M \otimes A - H_h\right]\right)^{i}\right\|\end{aligned} \tag{3.5.44}$$

Voraussetzung 3.5.24. Für $0 < m\,\tau \le t^*$ gelte

$$\begin{aligned}\sup_{i\in\mathbb{N}} \Big\{ \| \left(\Gamma^{-1}(\tau, h^2)\left[\tfrac{1}{\tau} I_M \otimes A - H_h\right]\right)^{i} \| : \tau,\ h \ \text{eventuell durch} \\ \text{eine Bedingung (3.5.27) eingeschränkt} \Big\} < \infty.\end{aligned} \tag{3.5.45}$$

Satz 3.5.25. *Seien die Voraussetzung von Satz 3.5.23 und die Voraussetzung 3.5.24 erfüllt. Dann ist das Crank-Nicolson Schema für $t \in \mathfrak{I}^*$ (bedingt) konvergent der Ordnung (2,2) für*

- $\nu_{d,t} = 0$ *und* $\nu_{d,x} \ge 0$,
- $\nu_{d,t} \ge 0$ *und* $\nu_{d,x} = 0$,
- $\nu_{d,t} = \nu_{d,x} = 1$

(eventuell unter einer Bedingung (3.5.27)). Es ist bedingt konvergent der Ordnung (2,1) für

- $\nu_{d,t} = 2$ *und* $\nu_{d,x} = 1$ *unter einer Bedingung* $0 < \hat{c}_0 \le \frac{\tau}{h^2}$.

Wie für das BTCS Schema legen wir die Matrizen

$$\Gamma_k(\tau, h^2) := \tfrac{1}{\tau} A + \tfrac{1}{2}\left(\lambda_k B + C\right), \qquad k = 1(1)M,$$

unseren folgenden Betrachtungen zugrunde. Das Analogon von Lemma 3.5.19 ist gegeben durch

LEMMA 3.5.26. *Sei für festes $\tau, h > 0$*

(i) Γ_k^{-1} *regulär für* $k = 1(1)M$,
(ii) $u(t,x)$ *(und damit* $u_h(t)$ *und* $\beta_h(t)$*) hinreichend glatt für* $t \in \mathfrak{I}^*$, $x \in [-l,l]$,
(iii) *die Linearkombination von* $\varepsilon_h(t_j)$ *bez. der* Φ_k *gegeben durch*

$$\varepsilon_h(t_j) = \sum_{k=1}^{M} \Phi_k \otimes e_{h,k}^j, \quad j = 0,1,..., \text{ mit } e_{h,k}^j \in \mathbb{R}^n .$$

Dann kann $e_{h,k}^{m+1}$ *für* $t \in \mathfrak{I}^*$ *durch*

$$\begin{aligned}\|e_{h,k}^{m+1}\|_{2,n} \le{}& \left\| \left(\Gamma_k^{-1}\left[\tfrac{1}{\tau}A - \tfrac{1}{2}(\lambda_k B + C)\right]\right)^{m+1} \right\|_{2,n} \|e_{h,k}^0\|_{2,n} \\ &+ K \sum_{i=0}^{m} \left\| \left(\Gamma_k^{-1}\left[\tfrac{1}{\tau}A - \tfrac{1}{2}(\lambda_k B + C)\right]\right)^{i} \right\|_{2,n} \left[h^2 \|\Gamma_k^{-1}B\|_{2,n} + \tau^2 \left\|\Gamma_k^{-1}A\right\|_{2,n}\right]\end{aligned}$$

abgeschätzt werden, wobei die Konstante $K > 0$ *unabhängig von* τ *und* h *ist.*

Das Lemma 3.5.21 gilt analog auch für das Crank-Nicolson Schema. So ist $\|e_{h,k}^{m+1}\|_{2,n} \to 0$ $(k = 1(1)M)$ für $\tau, h \to 0$ eine notwendige Bedingung für $\|e_h^{m+1}\| \to 0$ für $\tau, h \to 0$.

BEISPIEL 3.5.27. Wir betrachten erneut Beispiel 3.5.22 und wenden das Crank-Nicolson Schema darauf an. Hierfür ist

$$\Gamma_k^{-1}A = \begin{pmatrix} \frac{\tau}{1+\frac{\tau}{2}(1-\lambda_k)} & 0 \\ 0 & 0 \end{pmatrix}, \qquad \Gamma_k^{-1}B = \begin{pmatrix} -\frac{\tau}{1+\frac{\tau}{2}(1-\lambda_k)} & -\frac{\tau}{1+\frac{\tau}{2}(1-\lambda_k)} \\ 0 & 0 \end{pmatrix},$$

$$\|\Gamma_k^{-1}A\|_{2,n} = \frac{\tau}{1+\frac{\tau}{2}(1-\lambda_k)} \le \tau, \qquad \|\Gamma_k^{-1}B\|_{2,n} = \frac{\sqrt{2}\,\tau}{1+\frac{\tau}{2}(1-\lambda_k)} \le \sqrt{2}\,\tau,$$

$$\left\|\Gamma_k^{-1}\left[\tfrac{1}{\tau}A - \tfrac{1}{2}(\lambda_k B + C)\right]\right\|_{2,n} = \max\left\{1, \left|\frac{1+\frac{\tau}{2}(\lambda_k - 1)}{1-\frac{\tau}{2}(\lambda_k - 1)}\right|\right\} = 1$$

für alle $\tau, h \ge 0$. Hieraus folgt aus Lemma 3.5.26 $\|e_{h,k}^{m+1}\|_{2,n} = \mathcal{O}(\tau^2) + \mathcal{O}(h^2)$. □

BEMERKUNG 3.5.28. In einer Arbeit von L. Jodar [Jod91] ([Jod90]) werden homogene, lineare PDAEs mit $C = 0$, $g \equiv 0$ und homogenen Dirichlet RBn betrachtet. Es wird für diese Probleme durch Differenzenapproximationen basierend auf dem Crank-Nicolson-Schema mittels diskreter Fourierreihen eine numerische Lösung explizit angegeben. □

3.5.4. Schwach gekoppelte lineare PDAEs

Für spezielle Klassen linearer PDAEs sollen nun deren Indexe angegeben und einfach zu überprüfende Konvergenzbedingungen hergeleitet werden, die zum Teil auch Aussagen für höhere Indexe als die in Satz 3.5.18 erfaßten liefern. Anhand der folgenden Fehleranalysen werden Aussagen zur Konvergenz der Gesamtdiskretisierung des BTCS Schemas für diese speziellen Klassen getroffen. Dieses Vorgehen kann auch auf das Crank-Nicolson Schema übertragen werden.

Die nun betrachteten linearen PDAEs seien *schwach gekoppelt*. D.h., die einzelnen Gleichungen der PDAE (3.1.1) sind linear über die Komponenten der Funktion u, aber nicht über die der partiellen Ableitungen u_t und u_{xx}, gekoppelt.

Gekoppelte parabolische und elliptische Differentialgleichungen, die nur über die Funktion $u = (y^\top, w^\top)^\top$, aber nicht über deren partiellen Ableitungen, linear gekoppelt sind, können in der Form der semi-expliziten PDAE

$$\begin{aligned} y_t(t,x) - y_{xx}(t,x) \qquad\qquad + C_{11}\, y(t,x) + C_{12}\, w(t,x) &= g_1(t,x) \\ - w_{xx}(t,x) + C_{21}\, y(t,x) + C_{22}\, w(t,x) &= g_2(t,x). \end{aligned} \tag{3.5.46}$$

mit $(t,x) \in \mathfrak{I} \times \Omega$, $y \in \mathbb{R}^{n_1}$, $w \in \mathbb{R}^{n_2}$, $g_i \in \mathbb{R}^{n_i}$, $C_{ij} \in \mathbb{R}^{n_i \times n_j}$ für $i,j = 1,2$ und $n_1 + n_2 = n$ angegeben werden. Weiterhin sind Anfangs- und Randbedingungen gemäß (3.1.2)-(3.1.4) gegeben.

Satz 3.5.29. *Die lineare PDAE (3.5.46) besitzt den differentiellen Ortsindex $\nu_{d,x} = 0$ und entweder einen einheitlichen differentiellen Zeitindex $\nu_{d,t} = 1$ oder keinen einheitlichen Zeitindex.*

Sind die Eigenwerte von C_{22} für alle $k \in \mathbb{N}$ ungleich ρ_k (z.B. wenn die Eigenwerte von C_{22} nichtnegativ sind), so ist $\nu_{d,t} = 1$, und es gilt $\mathfrak{M}_{AB} = \{1, \dots, n_1\}$, $\mathfrak{M}_{RB} = \{1, \dots, n\}$. Der Zeitindex ist nicht von den Matrizen C_{11}, C_{12}, C_{21} abhängig.

Beweis. Da für die PDAE (3.5.46) $B = I_n$ regulär ist, ist $\nu_{d,x} = 0$. Nach einer Fouriertransformation von (3.5.46) ist die transformierte Gleichung (3.2.9) bereits eine lineare DAE der Form (1.2.6) und Folgerung 1.2.9 kann angewendet werden. Somit ist $\nu_{F,k} = 1$, wenn die Matrix $C_{22} - \rho_k I_{n_2}$ regulär ist. Dies wiederum ist genau dann der Fall, wenn die Eigenwerte von C_{22} von ρ_k verschieden sind. Gilt dies für alle $k \in \mathbb{N}$, d.h. $\nu_{F,k} \equiv \nu_F$, so hat die PDAE (3.5.46) den einheitlichen differentiellen Zeitindex $\nu_{d,t} = \nu_F = 1$, da aus den Transformations-Matrizen

$$S_{F,k} = \begin{pmatrix} I_{n_1} & -C_{12}(C_{22} - \rho_k I_{n_2})^{-1} \\ 0 & (C_{22} - \rho_k I_{n_2})^{-1} \end{pmatrix} \text{ und } T_{F,k} = \begin{pmatrix} I_{n_1} & 0 \\ -(C_{22} - \rho_k I_{n_2})^{-1} C_{21} & I_{n_2} \end{pmatrix}$$

direkt $\mathfrak{M}_{AB} \equiv \mathfrak{M}_{AB}^k = \{1, \dots, n_1\}$ bestimmt werden kann. Die PDAE besitzt andererseits $\nu_{d,t} > 1$, wenn die Matrix $C_{22} - \rho_k I_{n_2}$ singulär für alle k ist, d.h. $\det(C_{22} - \rho_k I_{n_2}) \equiv 0 \ \forall k \in \mathbb{N}$. Da dies der Eigenwertgleichung der Matrix C_{22} entspricht und C_{22} nur n_2 Eigenwerte besitzt, kann $C_{22} - \rho_k I_{n_2}$ nur für maximal n_2 verschiedene k singulär sein, und (3.5.46) kann keinen einheitlichen Zeitindex > 1 besitzen. □

Zur Vorbereitung der Konvergenzanalyse des BTCS Schemas bei Anwendung auf PDAEs (3.5.46) gehen wir zunächst auf die logarithmische Matrixnorm einer blockdiagonalen Matrix im $\mathbb{R}^{nM}$ ein: Für die der diskreten L_2-Norm zugeordneten logarithmische Matrixnorm einer Matrix $D \in \mathbb{R}^{nM}$ gilt (vgl. [Dek84])

$$\begin{aligned}\mu[D] &= \max_{y\neq 0} \frac{\langle Dy, y\rangle}{\langle y, y\rangle} = \max_{y\neq 0} \frac{hy^\top D^\top y}{h\, y^\top y} = \max_{y\neq 0} \frac{y^\top D^\top y}{y^\top y} = \max_{y\neq 0} \frac{\langle Dy, y\rangle_{2,nM}}{\langle y, y\rangle_{2,nM}} \\ &= \mu_{2,nM}[D] = \max\left\{\kappa \ : \ \kappa \text{ Eigenwert von } \tfrac{1}{2}(D^\top + D)\right\}.\end{aligned}$$

Für eine Matrix $D = \operatorname{diag}(D_1, \dots, D_M) \in \mathbb{R}^{nM\times nM}$, $D_k \in \mathbb{R}^{n\times n}$, $k = 1(1)M$, ist

$$\begin{aligned}\tfrac{1}{2}(D^\top + D) &= \operatorname{diag}\left(\tfrac{1}{2}(D_1^\top + D_1), \dots, \tfrac{1}{2}(D_M^\top + D_M)\right) \\ \text{und folglich} \qquad \mu[D] &= \mu_{2,nM}[D] = \max_{k=1(1)M} \{\mu_{2,n}[D_k]\}. \end{aligned} \tag{3.5.47}$$

Unter Benutzung dieser Beziehungen erhalten wir folgende Konvergenzaussage:

Satz 3.5.30. *Das BTCS Schema zur Lösung einer PDAE der Form (3.5.46) mit einheitlichem differentiellen Zeitindex $\nu_{d,t} = 1$ ist konvergent der Ordnung $(2, 1)$, wenn für hinreichend kleine h*

$$\max_{k=1(1)M} \{\lambda_k + \mu_{2,n}[C_{12}(C_{22} - \lambda_k I_{n_2})^{-1} C_{21}]\} + \mu_{2,n}[-C_{11}] \leq 0,$$

wobei die λ_k die Eigenwerte der Matrix $\frac{1}{h^2}P$ sind.

Beweis. Wir betrachten das semidiskrete Problem der PDAE (3.5.46) in der Form (3.5.2) für hinreichend kleine h. Dieses hat den Index 1, da nach Satz 3.5.4 der Riesz-Index des semidiskreten Systems mit dem einheitlichen differentiellen Zeitindex der PDAE $\nu_{d,t} = 1$ übereinstimmt. Nach Folgerung 1.2.9 schließen wir, daß die Matrix $E := -\frac{1}{h^2}P \otimes I_{n_2} + I_M \otimes C_{22}$ für hinreichend kleine h regulär ist. Das BTCS Schema lautet

$$\begin{aligned}\left\{I_{n_1 M} + \tau\left(-\tfrac{1}{h^2}P \otimes I_{n_1} + I_M \otimes C_{11}\right)\right\} Y_{m+1} + \tau\,(I_M \otimes C_{12})\, W_{m+1} &= Y_M + \tau G_1^{m+1} \\ (I_M \otimes C_{21}) Y_{m+1} + \left(-\tfrac{1}{h^2}P \otimes I_{n_2} + I_M \otimes C_{22}\right) W_{m+1} &= G_2^{m+1}\end{aligned}$$

$$\begin{aligned}\text{mit} \quad Y_m \approx Y_h(t_m) &= (y(t_m, x_1)^\top, \dots, y(t_m, x_M)^\top)^\top, \\ W_m \approx W_h(t_m) &= (w(t_m, x_1)^\top, \dots, w(t_m, x_M)^\top)^\top,\end{aligned}$$

$$G_i(t) := \left(g_i(t,x_1)^\top, \dots, g_i(t,x_M)^\top\right)^\top,\ i=1,2.$$

Sei $D := \frac{1}{h^2}P \otimes I_{n_1} - I_M \otimes C_{11} + (I_M \otimes C_{12})E^{-1}(I_M \otimes C_{21}) \in \mathbb{R}^{n_1 M \times n_1 M}$. Unter Verwendung der Matrizen Φ und Λ (siehe Seite 75) erhalten wir

$$\begin{aligned}(\Phi \otimes I_{n_1})D(\Phi \otimes I_{n_1}) &= \Lambda \otimes I_{n_1} - I_M \otimes C_{11} \\ &\quad + (I_M \otimes C_{12})(I_M \otimes C_{22} - \Lambda \otimes I_{n_2})^{-1}(I_M \otimes C_{21}).\end{aligned}$$

$$\begin{aligned}\text{Aus}\quad (I_M \otimes C_{22} - \Lambda \otimes I_{n_2})^{-1} &= \operatorname{diag}((C_{22}-\lambda_1 I_{n_2}), \dots, C_{22}-(\lambda_M I_{n_2}))^{-1} \\ &= \operatorname{diag}\left((C_{22}-\lambda_1 I_{n_2})^{-1}, \dots, (C_{22}-\lambda_M I_{n_2})^{-1}\right) \\ \text{und}\qquad D_k &:= \lambda_k I_{n_1} - C_{11} + C_{12}(C_{22}-\lambda_k I_{n_2})^{-1}C_{21}\end{aligned}$$

folgt $(\Phi \otimes I_{n_1})D(\Phi \otimes I_{n_1}) = \operatorname{diag}(D_1, \dots, D_M)$. Nach (3.5.47) ist

$$\begin{aligned}\mu[(\Phi \otimes I_{n_1})D(\Phi \otimes I_{n_1})] &= \max_{k=1(1)M}\{\mu_{2,n}[D_k]\} \\ &\le \max_{k=1(1)M}\{\lambda_k + \mu_{2,n}[C_{12}(C_{22}-\lambda_k I_{n_2})^{-1}C_{21}]\} + \mu_{2,n}[-C_{11}] \ \le\ 0\end{aligned}$$

$$\begin{aligned}\text{und}\qquad \mu[D] &= \max_{x\neq 0}\frac{x^\top D^\top D x}{x^\top x} = \max_{x=(\Phi\otimes I_{n_1})y\neq 0}\frac{y^\top(\Phi\otimes I_{n_1})^\top D^\top D(\Phi\otimes I_{n_1})y}{y^\top(\Phi\otimes I_{n_1})^\top(\Phi\otimes I_{n_1})y} \\ &= \max_{y\neq 0}\frac{y^\top((\Phi\otimes I_{n_1})D(\Phi\otimes I_{n_1}))^\top((\Phi\otimes I_{n_1})D(\Phi\otimes I_{n_1}))\,y}{y^\top y} \\ &= \mu[(\Phi\otimes I_{n_1})D(\Phi\otimes I_{n_1})] \ \le\ 0.\end{aligned}$$

Aus $\mu[D] \le 0$ folgt $I_{n_1 M} - \tau D$ regulär und $\|(I_{n_1 M} - \tau D)^{-1}\| \le 1$ für $\tau > 0$ (vgl. z.B. Lemma 2.3.4 aus Kapitel 2). Aus der Regularität von E und $I_{n_1 M} - \tau D$ erhält man aus dem BTCS Schema

$$\begin{aligned}W_{m+1} &= E^{-1}\left\{G_2(t_{m+1}) - (I_M \otimes C_{21})Y_{m+1}\right\}, \qquad (3.5.48)\\ Y_{m+1} &= (I_{n_1 M} - \tau D)^{-1}\left\{Y_m + \tau\left(G_1(t_{m+1}) - (I_M \otimes C_{12})E^{-1}G_2(t_{m+1})\right)\right\}.\end{aligned}$$

Aus der PDAE (3.5.46) und (3.5.6) erhält man

$$G_2(t) = (I_M \otimes C_{21})Y_h(t) + E\,W_h(t) + \underbrace{(\tfrac{1}{h^2}P \otimes I_{n_1})W_h(t) - W_{h,xx}(t)}_{=h^2\beta_{h,w}(t)}$$

$$\begin{aligned}&G_1(t) - (I_M \otimes C_{12})E^{-1}G_2(t) \\ &\quad = Y_{h,t}(t) - D\,Y_h(t) + \underbrace{(\tfrac{1}{h^2}P \otimes I_{n_1})Y_h(t) - Y_{h,xx}(t)}_{=h^2\beta_{h,y}(t)} - (I_M \otimes C_{12})E^{-1}h^2\beta_{h,w}(t) \\ &\quad = Y_{h,t}(t) - D\,Y_h(t) + h^2\underbrace{\left\{\beta_{h,y}(t) - (I_M \otimes C_{12})E^{-1}\beta_{h,w}(t)\right\}}_{=:\tilde{\beta}_h(t)}.\end{aligned}$$

Für den lokalen Gesamtdiskretisierungsfehler bez. der Komponente y folgt hieraus

$$\begin{aligned} le_{h,y}(t_{m+1}) &= Y_h(t_{m+1}) - \hat{Y}_{m+1} \\ &= -(I_{n_1 M} - \tau D)^{-1} \left(\tfrac{1}{2}\tau^2 Y_{h,tt}(t_m + \zeta\tau) + \tau h^2 \tilde{\beta}_h(t_{m+1}) \right), \quad \zeta \in (0,1). \end{aligned}$$

Somit ist

$$\|le_{h,y}(t_{m+1})\| \leq C_0 \|(I_{n_1 M} - \tau D)^{-1}\| (\tau^2 + \tau h^2) \leq C_0(\tau^2 + \tau h^2),$$

$C_0 > 0$, unabhängig von τ und h, und

$$\|\varepsilon_{h,y}(t_{m+1})\| \leq \|\varepsilon_{h,y}(0)\| + \max_{i=0}^{m} \|le_{h,y}(t_{i+1})\| \sum_{i=0}^{m} \|(I_{n_1 M} - \tau D)^{-i}\|$$

Wegen $\mathfrak{M}_{AB} = \{1, \ldots, n_1\}$ und $Y_0 = Y_h(0)$ ist $\|\varepsilon_{h,y}(0)\| = 0$ und $\|\varepsilon_{h,y}(t_{m+1})\| = \mathcal{O}(\tau) + \mathcal{O}(h^2)$. Aus (3.5.48) und der Gleichung für $G_2(t)$ folgt

$$\begin{aligned} \varepsilon_{h,w}(t_{m+1}) = W_h(t_{m+1}) - W_{m+1} &= E^{-1} \Big[G_2(t_{m+1}) - h^2 \beta_{h,w}(t_{m+1}) \\ &\qquad - (I_M \otimes C_{21}) Y_h(t_{m+1}) - G_2(t_{m+1}) + (I_M \otimes C_{21}) Y_{m+1} \Big] \\ &= -E^{-1} \Big[h^2 \beta_{h,w}(t_{m+1}) + (I_M \otimes C_{21}) \varepsilon_{h,y}(t_{m+1}) \Big] \end{aligned}$$

die Behauptung für die Konvergenz der Gesamtdiskretisierung. □

Ein analoges Ergebnis erhält man für linear gekoppelte parabolische und gewöhnliche Differentialgleichungen der Form (siehe Beispiel 3.1.1, [Leu89])

$$\begin{aligned} y_t(t,x) \qquad\quad - y_{xx}(t,x) \; + C_{11}\, y(t,x) + \; C_{12}\, w(t,x) \;&=\; g_1(t,x) \\ w_t(t,x) \qquad\qquad\quad + C_{21}\, y(t,x) + \; C_{22}\, w(t,x) \;&=\; g_2(t,x). \end{aligned} \tag{3.5.49}$$

mit $(t,x) \in \mathfrak{I} \times \Omega$, $y \in \mathbb{R}^{m_1}$, $w \in \mathbb{R}^{m_2}$, $g_i \in \mathbb{R}^{m_i}$, $C_{ij} \in \mathbb{R}^{m_i \times m_j}$ für $i,j = 1,2$ und $m_1 + m_2 = n$. D.h., die Kopplung erfolgt nur über die Funktion $u = (y^\top, w^\top)^\top$, aber nicht über deren partiellen Ableitungen. Weiterhin sind Anfangs- und Randbedingungen gemäß (3.1.2)-(3.1.4) gegeben.

Satz 3.5.31. *Die PDAE (3.5.49) besitzt den einheitlichen differentiellen Zeitindex $\nu_{d,t} = 0$ und einen differentiellen Ortsindex $\nu_{d,x} = 1$ und es gilt $\mathfrak{M}_{RB} = \{1, \ldots, m_1\}$, $\mathfrak{M}_{AB} = \{1, \ldots, n\}$.*

Beweis. Der Beweis erfolgt analog zum Beweis von Satz 3.5.29, in dem wir die Laplace-Transformation von (3.5.49) betrachten. Die Matrix $\xi I_{m_1} + C_{22}$ ist für alle ξ mit hinreichend großem Realteil stets regulär. □

SATZ 3.5.32. *Das BTCS Schema zur Lösung einer PDAE der Form (3.5.49) ist konvergent der Ordnung* $(2,1)$*, wenn für hinreichend kleine* h

$$\max_{k=1(1)M}\{\lambda_k\} + \mu_{2,n}\left[-\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22}\end{pmatrix}\right] \leq 0,$$

wobei λ_k *die Eigenwerte der Matrix* $\frac{1}{h^2}P$ *sind.*

BEWEIS. Der Beweis erfolgt ähnlich zum Beweis von Satz 3.5.30. Da $A = I_{m_1}$, gilt für den lokalen Gesamtdiskretisierungsfehler des BTCS Schemas die Gleichung (3.5.28) mit $\Gamma(\tau,h^2)^{-1}(I_M \otimes A) = \tau\,(I_{nM} - \tau\, D)^{-1}$ und

$$D = \tfrac{1}{h^2}P \otimes \begin{pmatrix} I_{m_1} & 0 \\ 0 & 0\end{pmatrix} - I_M \otimes \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22}\end{pmatrix}.$$

Es folgt für Gleichung (3.5.34)

$$\begin{aligned}\|\varepsilon_h(t_{m+1})\| \leq \|(I_{nM} - \tau D)^{-(m+1)}\|\|\varepsilon_h(0)\| \\ + \max_{i=0}^{m}\|le_{h,y}(t_{i+1})\| \sum_{i=0}^{m}\|(I_{nM} - \tau D)^{-i}\|.\end{aligned}$$

Wegen

$$(\Phi \otimes I_n)D(\Phi \otimes I_n) = \operatorname{diag}(D_1,\dots,D_M) \;\text{ mit }\; D_k = \lambda_k \begin{pmatrix} I_{m_1} & 0 \\ 0 & 0\end{pmatrix} - \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22}\end{pmatrix}$$

ist unter Verwendung der Voraussetzung

$$\mu[D] = \max_{k=1(1)M}\{\mu_{n,2}[D_k]\} = \max_{k=1(1)M}\{\lambda_k\} + \mu_{n,2}\left[-\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22}\end{pmatrix}\right] \leq 0,$$

woraus $\|(I_{nM} - \tau D)^{-1}\| \leq 1$ folgt. Da alle AWe beliebig vorgegeben werden können, d.h. $\mathfrak{M}_{AB} = \{1,\dots,n\}$ und $Y_0 = Y_h(0)$, $W_0 = W_h(0)$, ist $\varepsilon_h(0) = 0$ und wegen (3.5.28) folgt die Behauptung. □

Koppelt man parabolische Differentialgleichung und algebraischen Gleichungen linear über die Funtion $u = (y^\top, w^\top)^\top$, so erhält man eine lineare PDAE der Form

$$\begin{aligned} y_t(t,x) \;-\; y_{xx}(t,x) + \; C_{11}\,y(t,x) \;+\; C_{12}\,w(t,x) \;&=\; g_1(t,x) \\ C_{21}\,y(t,x) \;+\; C_{22}\,w(t,x) \;&=\; g_2(t,x). \end{aligned} \tag{3.5.50}$$

mit $(t,x) \in \mathfrak{I} \times \Omega$, $y \in \mathbb{R}^{n_1}$, $w \in \mathbb{R}^{n_2}$, $g_i \in \mathbb{R}^{n_i}$, $C_{ij} \in \mathbb{R}^{n_i \times n_j}$ für $i,j = 1,2$ und $n_1 + n_2 = n$. Weiterhin sind Anfangs- und Randbedingungen gemäß (3.1.2)-(3.1.4) gegeben. Ebenfalls analoge Betrachtungen zum Beweis von Satz 3.5.29 liefern den folgenden Satz:

SATZ 3.5.33. *Die PDAE (3.5.50) besitzt*
(i) für reguläre Matrizen C_{22} *den einheitlichen differentiellen Zeitindex* $\nu_{d,t} = 1$ *und den differentiellen Ortsindex* $\nu_{d,x} = 1$ *und es gilt* $\mathfrak{M}_{RB} = \mathfrak{M}_{AB} = \{1,\dots,n_1\}$,

(ii) für $C_{22} = 0$ und $C_{21}C_{12}$ regulär den einheitlichen differentiellen Zeitindex $\nu_{d,t} = 2$ und den differentiellen Ortsindex $\nu_{d,x} = 3$ und es ist $\operatorname{rang}(C_{21}) = n_2$. *Wenn o.B.d.A. $C_{21} = (0\ I_{n_2})$, so gilt $\mathfrak{M}_{RB} = \mathfrak{M}_{AB} = \{1, \dots, n_1 - n_2\}$.*

BEWEIS. Der Beweis der Behauptung für reguläres C_{22} erfolgt analog zu den Beweisen der Sätze 3.5.30 und 3.5.32, wobei statt der dort betrachteten Matrizen $C_{22} - \rho_k I_{n_2}$ und $C_{22} + \xi I_{n_2}$ hier nur die Matrix C_{22} zugrunde gelegt wird, die insbesondere für alle k und alle ξ regulär ist.

Sei $C_{22} = 0$. Aus den Voraussetzungen 3.2.1 c) und d) folgt $\operatorname{rang}(C_{21}) = n_2$. Somit kann o.B.d.A. angenommen werden, daß die Komponenten von y in (3.5.50) bereits so angeordnet sind, daß z.B. $C_{21} = (0\ I_{n_2})$. Mit den Bezeichnungen $\bar{n} = n_1 - n_2$, $C_{11} = \begin{pmatrix} D_1 & D_2 \\ D_3 & D_4 \end{pmatrix}$, $C_{12} = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}$ ($D_1 \in \mathbb{R}^{\bar{n}\times\bar{n}}, D_2, E_1 \in \mathbb{R}^{\bar{n}\times n_2}, D_3 \in \mathbb{R}^{n_2\times\bar{n}}, D_4, E_2 \in \mathbb{R}^{n_2\times n_2}$), $F = \left(E_1E_2^{-1}D_3 - D_1\right)E_1E_2^{-1} - D_2 + E_1E_2^{-1}D_4 \in \mathbb{R}^{\bar{n}\times n_2}$ kann man die Matrizen $S_{F,k}$ und $T_{F,k}$ angeben:

$$S_{F,k} = \begin{pmatrix} I_{\bar{n}} & -E_1E_2^{-1} & F \\ 0 & I_{n_2} & \rho_k I_{n_2} - D_4 - D_3E_1E_2^{-1} \\ 0 & 0 & I_{n_2} \end{pmatrix},\ T_{F,k} = \begin{pmatrix} I_{\bar{n}} & 0 & E_1E_2^{-1} \\ 0 & 0 & I_{n_2} \\ -E_2^{-1}D_3 & E_2^{-1} & 0 \end{pmatrix}.$$

Man erhält

$$S_{F,k}\,AT_{F,k} = \begin{pmatrix} I_{\bar{n}} & 0 & 0 \\ 0 & 0 & I_{n_2} \\ 0 & 0 & 0 \end{pmatrix},\ S_{F,k}\,(\rho_k B + C)T_{F,k} = \begin{pmatrix} -\rho_k I_m + D_1 - E_1E_2^{-1}D_3 & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & I_{n_2} \end{pmatrix},$$

woraus $\nu_{d,t} = 2$ und $\mathfrak{M}_{AB} = \mathfrak{M}_{AB}^k = \{1, \dots, n_1 - n_2\}$ folgt (analog $\nu_{d,x} = 3$, $\mathfrak{M}_{RB} = \mathfrak{M}_{RB}^\xi = \{1, \dots, n_1 - n_2\}$). □

Analog zu Satz 3.5.30 erhält man

SATZ 3.5.34. *Das BTCS Schema zur Lösung einer PDAE der Form (3.5.50) mit regulärer Matrix C_{22} ($\nu_{d,t} = 1$, $\nu_{d,x} = 1$) ist konvergent der Ordnung $(2, 1)$, wenn für hinreichend kleine h*

$$\max_{k=1(1)M} \{\lambda_k\} + \mu_{2,n}[C_{12}C_{22}^{-1}C_{21}] + \mu_{2,n}[-C_{11}] \le 0$$

gilt, wobei λ_k die Eigenwerte der Matrix $\frac{1}{h^2}P$ sind.

SATZ 3.5.35. *Sei in (3.5.50) $C_{22} = 0$ und $C_{21}C_{12}$ regulär ($\nu_{d,t} = 2$, $\nu_{d,x} = 3$). O.D.d.A. sei $C_{21} = (0\ I_{n_2})$, $C_{11} = \begin{pmatrix} D_1 & D_2 \\ D_3 & D_4 \end{pmatrix}$, $C_{12} = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}$ ($\bar{n} = n_1 - n_2$, $D_1 \in \mathbb{R}^{\bar{n}\times\bar{n}}, D_2, E_1 \in \mathbb{R}^{\bar{n}\times n_2}, D_3 \in \mathbb{R}^{n_2\times\bar{n}}, D_4, E_2 \in \mathbb{R}^{n_2\times n_2}$). Das BTCS Schema zur Lösung einer PDAE der Form (3.5.50) mit diesen Matrizen ist konvergent der Ordnung $(2, 1)$, wenn für hinreichend kleine h*

$$\max_{k=1(1)M} \{\lambda_k\} + \mu_{2,n}[E_1E_2^{-1}D_3 - D_1] \le 0$$

gilt, wobei λ_k die Eigenwerte der Matrix $\frac{1}{h^2}P$ sind.

BEWEIS. Für eine Matrix A sei $\bar{A} = I_M \otimes A$. Auf die Dimensionsangaben für Einheitsmatrizen wird zur Vereinfachung verzichtet. Mit $y = (y_1^\top, y_2^\top)^\top$, $g_1 = (g_{11}^\top, g_{12}^\top)^\top$, $y_i, g_{1i} \in \mathbb{R}^{n_i}$, $i = 1,2$, sowie $Y_{i,h}(t) = (y_i(t,x_k)^\top)_{k=1(1)M}^\top$, $Y_{i,m} \approx Y_{i,h}(t_m)$, $W_h = (w(t,x_k)^\top)_{k=1(1)M}^\top$, $W_m \approx W_h(t_m)$ und entsprechenden $\tilde{G}_{1i}, G_2$ lautet das BTCS Schema

$$\begin{aligned}
\big(\tfrac{1}{\tau}I + (-\tfrac{1}{h^2}P \otimes I + \bar{D}_1)\big)\, Y_{1,m+1} + \bar{D}_2 Y_{2,m+1} + \bar{E}_1 W_{m+1} &= \tfrac{1}{\tau}Y_{1,m} + \tilde{G}_{11}(t_{m+1}) \\
\bar{D}_3 Y_{1,m+1} + \big(\tfrac{1}{\tau}I + (-\tfrac{1}{h^2}P \otimes I + \bar{D}_4)\big)\, Y_{2,m+1} + \bar{E}_2 W_{m+1} &= \tfrac{1}{\tau}Y_{2,m} + \tilde{G}_{12}(t_{m+1}) \\
Y_{2,m+1} &= G_2(t_{m+1}).
\end{aligned}$$

Hieraus folgt mit $R = \frac{1}{h^2}P \otimes I - \bar{D}_1 + \bar{E}_1 \bar{E}_2^{-1} \bar{D}_3$

$$\begin{aligned}
Y_{2,m+1} &= G_2(t_{m+1}) \\
W_{m+1} &= \bar{E}_2^{-1}\Big(- \bar{D}_3 Y_{1,m+1} + \tilde{G}_{12}(t_{m+1}) - \tfrac{1}{\tau}(G_2(t_{m+1}) - G_2(t_m)) \\
&\qquad - (-\tfrac{1}{h^2}P \otimes I + \bar{D}_4) G_2(t_{m+1})\Big) \\
Y_{1,m+1} &= (I - \tau R)^{-1}\Big(Y_{1,m} + \tau\big[\tilde{G}_{11}(t_{m+1}) - \bar{D}_2 G_2(t_{m+1}) \\
&\qquad - \bar{E}_1 \bar{E}_2^{-1}\big\{\tilde{G}_{12}(t_{m+1}) - \tfrac{1}{\tau}(G_2(t_{m+1}) - G_2(t_m))\big\}\big]\Big).
\end{aligned}$$

Wie man leicht sieht, ist $\varepsilon_{h,y_2}(t_{m+1}) = 0$. Verwendet man (3.5.6), die sich aus der PDAE (3.5.50) für $\tilde{G}_{11}, \tilde{G}_{12}, G_2$ ergebenden Ausdrücke und Taylorentwicklungen, so erhält man

$$\begin{aligned}
\varepsilon_{h,w}(t_{m+1}) &= W_h(t_{m+1}) - \bar{E}_2^{-1}\Big(- \bar{D}_3 Y_{1,m+1} + Y'_{2,h}(t_{m+1}) + h^2 \beta_{h,y_2}(t_{m+1}) \\
&\qquad + \bar{D}_3 Y_{1,h}(t_{m+1}) + (-\tfrac{1}{h^2}P \otimes I + \bar{D}_4) Y_{2,h}(t_{m+1}) + \bar{E}_2 W_h(t_{m+1}) \\
&\qquad - Y'_{2,h}(t_{m+1}) + \tfrac{\tau}{2} Y'_{2,h}(t_m + \zeta\tau) - (-\tfrac{1}{h^2}P \otimes I + \bar{D}_4) Y_{2,h}(t_{m+1})\Big) \\
&= -\bar{E}_2^{-1}\Big(\bar{D}_3\, \varepsilon_{h,y_1}(t_{m+1}) + h^2 \beta_{h,y_2}(t_{m+1}) + \tfrac{\tau}{2} Y'_{2,h}(t_m + \zeta\tau)\Big)
\end{aligned}$$

($\zeta \in (0,1)$). Weiterhin ist

$$\begin{aligned}
le_{h,y_1}(t_{m+1}) &= (I - \tau R)^{-1}\Big(\tfrac{\tau^2}{2}\big(\bar{E}_1 \bar{E}_2^{-1} Y''_{2,h}(t_m + \zeta_2\tau) - Y''_{1,h}(t_m + \zeta_1\tau)\big) \\
&\qquad + \tau h^2 \big(\bar{E}_1 \bar{E}_2^{-1} \beta_{h,y_2}(t_{m+1}) - \beta_{h,y_1}(t_{m+1})\big)\Big), \zeta_1, \zeta_2 \in (0,1), \\
\|le_{h,y_1}(t_{m+1})\| &\le C_0 \|(I - \tau R)^{-1}\|(\tau^2 + \tau h^2)
\end{aligned}$$

mit C_0 unabhängig von τ, h (ζ, ζ_1, ζ_2 komponentenweise verschieden). Wegen $\mathfrak{M}_{AB} = \{1, \ldots, n_1 - n_2\}$ ist $Y_{1,0} = Y_h(0)$. Aus der Voraussetzung folgt in Analogie zum Beweis von Satz 3.5.30 $\mu[R] \le 0$ und somit

$$\|\varepsilon_{h,y_1}(t_{m+1})\| = \mathcal{O}(\tau) + \mathcal{O}(h^2) \quad \Longrightarrow \quad \|\varepsilon_{h,w}(t_{m+1})\| = \mathcal{O}(\tau) + \mathcal{O}(h^2)$$

für $\tau, h \to 0$. □

Beispiel 3.5.36. Die PDAE aus Beispiel 3.1.4 ist für $a = b = 1$ von der Form (3.5.50) mit $C_{22} = 0$ und $C_{21}C_{12} = c_3c_1 \neq 0$. Aus Satz 3.5.33 folgt $\nu_{d,t} = 2$ und $\nu_{d,x} = 3$. Sei o.B.d.A. $c_3 = 1$, dann ist $C_{12} = (0\ 1)$, $C_{11} = \left(\begin{smallmatrix} 0 & c_1 \\ 0 & 0 \end{smallmatrix}\right)$, $C_{12} = \left(\begin{smallmatrix} 0 \\ c_2 \end{smallmatrix}\right)$ und $E_1E_2^{-1}D_3 - D_1 = 0$ mit den Bezeichnungen aus Satz 3.5.35. Da die Eigenwerte der Matrix $\frac{1}{h^2}P$ negativ sind (vgl. Seite 74), ist die Voraussetzung von Satz 3.5.35 erfüllt und das BTCS Schema für dieses Beispiel konvergent der Ordnung (2,1). □

Bemerkung 3.5.37. Eine lineare PDAE der Form (3.5.50) mit $C_{22} = 0$ repräsentiert die größere Klasse der PDAEs (3.5.50) mit $C_{22} \neq 0$ und C_{22} singulär, da Probleme (3.5.50) mit singülärem $C_{22} \neq 0$ stets auf solche mit $C_{22} = 0$ überführt werden können (vgl. [Hai96]). □

3.5.5. Numerische Beispiele

Im folgenden werden einige numerische Testrechnungen vorgestellt. Hierfür wird zumeist das BTCS Schema verwendet. Für spezielle Beispiele werden die Konvergenzergebnisse aus Abschnitt 3.5.3 numerisch bestätigt. Wir werden anhand dieses Schemas demonstrieren, daß die Konsistenz der Randbedingungen wichtig für die Genauigkeit einer numerischen Lösung von PDAE (3.1.1) ist. Weiterhin wird das in Abschnitt 3.2.2 diskutierte Problem eines Indexsprunges betrachtet. Unter Verwendung des BTCS Schemas wird gezeigt, daß bei spezieller Wahl der Ortsschrittweite h das semidiskrete Problem (3.5.2) einen Indexsprung haben kann, obwohl die PDAE einen einheitlichen Zeitindex besitzt. Hieraus können sich große Fehler in der zugehörigen numerischen Lösung ergeben.

In den unten betrachteten Beispielen sei $l = 1$ und der inhomogene Term $g(t, x)$ der rechten Seite der PDAE (3.1.1) so gewählt, daß eine von uns vorgegebene Funktion die Lösung der PDAE ist. Somit können Randbedingungen (3.1.2) bzw. Anfangsbedingungen (3.1.3) für u_i mit $i \notin \mathfrak{M}_{RB}$ bzw. $i \notin \mathfrak{M}_{AB}$, die für die numerische Rechnung benötigt werden, exakt gewählt werden, so daß diese als konsistent vorausgesetzt werden können. Somit kann stets $v_0 = U_h(0)$ gewählt werden und es gilt $\varepsilon_h(0) = 0$.

3.5.5.1. Ordnungsbestimmungen. Zunächst sei bemerkt, daß der in Abschnitt 3.5.3 betrachtete Gesamtdiskretisierungsfehler ε_h eine Funktion von τ und h ist. Daher wollen wir hier $\varepsilon_{h,\tau}$ schreiben und die diskrete L_2-Norm des Fehlers betrachten. Nach Definition 1.1.16 gilt für eine konvergente Gesamtdiskretisierung der Ordnung (p, q) für $\tau, h \to 0$ die Relation $\|\varepsilon_{h,\tau}(t_{m+1})\| = \mathcal{O}(h^p) + \mathcal{O}(\tau^q)$. Zur

Bestimmung der numerischen Konvergenzordnung setzt man daher an

$$\|\varepsilon_{h,\tau}(t_{m+1})\| = C_1\, h^p + C_2\, \tau^q, \tag{3.5.51}$$

wobei C_1, C_2 unabhängig von den äquidistante Orts- und Zeitschrittweiten h und τ sind. Aus (3.5.7), (3.5.28) (bzw. (3.5.42)) und (3.5.36) erhält man bei Anwendung des BTCS Schemas (bzw. Crank-Nicolson Verfahrens) $C_1 = 0$, wenn $\frac{\partial^4}{\partial x^4}u(t,x) = 0$, und $C_2 = 0$, wenn $\frac{\partial^2}{\partial t^2}u(t,x) = 0$ (bzw. $\frac{\partial^3}{\partial t^3}u(t,x) = 0$). Dies nutzen wir nun für die numerischen Tests:

Sei $t_{m+1} = 1$ und bezeichne $\varepsilon_{h,\tau} = \varepsilon_{h,\tau}(1)$. Eine spezielle PDAE habe eine Lösung $u_I(t,x)$, die ein Polynom in x mit einem Grad nicht größer als 3 ist. Dann ist $\frac{\partial^4}{\partial x^4}u(t,x) = 0$ und $C_1 = 0$. Somit kann man aus

$$\|\varepsilon_{h,\tau}\| = C_2\, \tau^q, \qquad \left\|\varepsilon_{h,\frac{\tau}{2}}\right\| = C_2\, \left(\frac{\tau}{2}\right)^q$$

numerisch die Konvergenzordnung bez. τ durch

$$q_{num,\tau} = q = \log_2 \|\varepsilon_{h,\tau}\| - \log_2 \left\|\varepsilon_{h,\frac{\tau}{2}}\right\|$$

bestimmen. Analog wählen wir eine PDAE mit einer Lösung $u_{II}(t,x)$, so daß $C_2 = 0$ ist und sich die numerische Konvergenzordnung bez. h durch

$$p_{num,h} = p = \log_2 \|\varepsilon_{h,\tau}\| - \log_2 \left\|\varepsilon_{\frac{h}{2},\tau}\right\|$$

bestimmen läßt.

Beispiel 3.5.38. Wir betrachten Beispiel 3.5.22 mit $b_1 = b_2 = -1$. Sei $g(t,x)$ so gewählt, daß

$$u_I(t,x) = \left(x(x^2-1)\,\cos \pi t,\ (x^2-1)\, e^{-t}\right)^\top \in \mathbb{R}^2$$

die exakte Lösung ist. Da die Komponenten von $u_I(t,x)$ ein Polynom dritten Grades in x ist, wird kein Ortsfehler durch die Semidiskretisierung (1.1.8) von u_{xx} in das semidiskrete Problem (3.5.2) eingeführt, d.h. $\alpha_h(t) = 0$ ($C_1 = 0$). Wenn wir analog $g(t,x)$ so wählen, daß

$$u_{II}(t,x) = \left(x^6(x^2-l^2)\, t,\ x^4(x^2-l^2)\, t\right)^\top \in \mathbb{R}^2,$$

die exakte Lösung ist, so verschwinden die zweiten Zeitableitungen von $u_h(t)$ und es ist $C_2 = 0$. Tabelle 3.1 zeigt, daß für das BTCS Schema die τ-Ordnung gleich 1 und die h-Ordnung gleich 2 ist. Gleiche Konvergenzordnungen wurde für die Errornorm $\|e^m_{h,k}\|$ in Beispiel 3.5.22 hergeleitet.

Die numerisch bestimmte Konvergenzordnungen für das Crank-Nicolson Schema ist Tabelle 3.2 zu entnehmen. Wie in Beispiel 3.5.27 zeigt sich eine τ-Ordnung 2 und eine h-Ordnung 2. □

	$q_{num,\tau}$ mit $u_I(t,x)$						$p_{num,h}$ mit $u_{II}(t,x)$					
$10^{-1}\,\tau^{-1}$	2^2	2^3	2^4	2^5	2^6	2^7	2^2	2^3	2^4	2^5	2^6	2^7
$10^{-1}\,h^{-1}$												
1	0.95	0.97	0.99	0.99	1.00	1.00	1.92	1.92	1.92	1.92	1.92	1.92
2^1	0.95	0.97	0.99	0.99	1.00	1.00	1.98	1.98	1.98	1.98	1.98	1.98
2^2	0.95	0.97	0.99	0.99	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00
2^3	0.95	0.97	0.99	0.99	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00
2^4	0.95	0.97	0.99	0.99	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00
2^5	0.95	0.97	0.99	0.99	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00

TABELLE 3.1. Numerische Konvergenzordnung des BTCS Schemas für Beispiel 3.5.38

	$q_{num,\tau}$ mit $u_I(t,x)$						$p_{num,h}$ mit $u_{II}(t,x)$					
$10^{-1}\,\tau^{-1}$	2^2	2^3	2^4	2^5	2^6	2^7	2^2	2^3	2^4	2^5	2^6	2^7
$10^{-1}\,h^{-1}$												
1	2.01	2.00	2.00	2.00	2.00	2.00	1.92	1.92	1.92	1.92	1.92	1.92
2^1	2.01	2.00	2.00	2.00	2.00	2.00	1.98	1.98	1.98	1.98	1.98	1.98
2^2	2.01	2.00	2.00	2.00	2.00	2.00	1.99	1.99	1.99	1.99	1.99	1.99
2^3	2.01	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
2^4	2.01	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
2^5	2.01	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00

TABELLE 3.2. Numerische Konvergenzordnung des Crank-Nicolson Schemas für Beispiel 3.5.38

BEISPIEL 3.5.39. Sei die PDAE (3.1.1) mit den Matrizen

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

gegeben. Ihre Indexe sind $\nu_{d,t} = 2, \nu_{d,x} = 3$. Obwohl Satz 3.5.18 nicht angewendet werden kann, ist die mit dem BTCS Schema erhaltene numerische Lösung für diese PDAE möglicherweise eine geeignete Approximation an die exakte Lösung:

Die numerisch bestimmte Konvergenzordnungen mit

$$u_I(t,x) = \left(x(x^2-l^2)e^{-t},\ (x^2-l^2)e^{t},\ (x+l)(x^2-l^2)e^{10\,t}\right)^\top \in \mathbb{R}^3$$
$$u_{II}(t,x) = \left(x^6(x^2-l^2)\,t,\ x^4(x^2-l^2)\,t,\ (x^4-l^4)\,t\right)^\top \in \mathbb{R}^3,$$

die Tabelle 3.3 zu entnehmen sind, zeigen Konvergenz der Ordnung $\mathcal{O}(\tau) + \mathcal{O}(h^2)$. □

	$q_{num,\tau}$ mit $u_I(t,x)$						$p_{num,h}$ mit $u_{II}(t,x)$					
$10^{-1}\,\tau^{-1}$	2^2	2^3	2^4	2^5	2^6	2^7	2^2	2^3	2^4	2^5	2^6	2^7
$10^{-1}\,h^{-1}$												
1	0.99	0.99	1.00	1.00	1.00	1.00	1.77	1.77	1.77	1.77	1.77	1.77
2^2	0.99	0.99	1.00	1.00	1.00	1.00	1.95	1.95	1.95	1.95	1.95	1.95
2^4	0.99	0.99	1.00	1.00	1.00	1.00	1.99	1.99	1.99	1.99	1.99	1.99
2^6	1.03	1.01	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00

TABELLE 3.3. Numerische Konvergenzordnung des BTCS Schemas für Beispiel 3.5.39

3.5.5.2. Konsistenz der Randbedingungen. Das folgende Beispiel zeigt, daß es wichtig ist zu wissen für welche Komponenten von u Randbedingungen beliebig vorgeschrieben werden können. Hierfür betrachten wir erneut Beispiel 3.3.3 mit $b = 1$, d.h.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} u_t + \begin{pmatrix} -1 & -1 \\ 0 & 0 \end{pmatrix} u_{xx} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} u = g, \tag{3.5.52}$$

dessen rechte Seite $g = (g_1, g_2)^\top$ so gewählt sei, daß die exakte Lösung dieser PDAE durch

$$u(t,x) = \begin{pmatrix} v(t,x) \\ w(t,x) \end{pmatrix} := \begin{pmatrix} (x^2-1)(x^2+4)x\cos(\pi\,t) \\ (x^2-1)(x^5-2x^2+5)e^{-t} \end{pmatrix}$$

gegeben ist. Da die Matrix A regulär ist, ist das zugehörige MOL-System (3.5.2) ein gewöhnliches Differentialgleichungssystem. Die Menge $\mathfrak{M}_{RB}$ kann $\mathfrak{M}_{RB} = \{1\}$ sein (eine weitere Möglichkeit ist $\mathfrak{M}_{RB} = \{2\}$).

Um den Unterschied zwischen konsistenten und inkonsistenten RBn für die zweite Lösungskomponente zu illustrieren, wenden wir das BTCS Schema an und vergleichen die numerische Lösung $u_h(t,x)$ mit einer anderen numerischen Lösung $\tilde{u}_h(t,x)$ wobei $\tilde{v}_h(t,\pm 1) = v_h(t,\pm 1)$ und $\tilde{w}_h(t,\pm 1) = w_h(t,\pm 1) + e^{2t} - 1$, d.h., $\tilde{w}_h(t,\pm l)$ ist eine inkonsistente RB, da g am Rand verschwindet und die Konsistenzbedingung (3.3.4) $\tilde{w}_{h\,t}(t,\pm 1) \neq g_2(t,\pm 1)$ gilt. Die ABn $u_h(0,x)$ und $\tilde{u}_h(0,x)$ haben den Wert $u(0,x)$, so daß die Verträglichkeitsbedingung (3.1.4) gewährleistet ist.

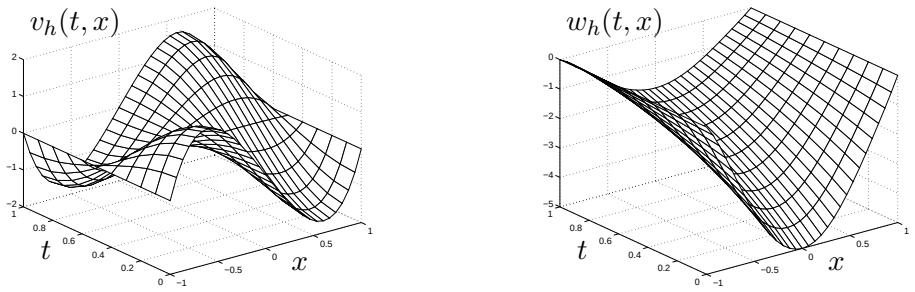

Abbildung 3.3. PDAE (3.5.52), konsistente RBn, $h = \frac{1}{15}$, $\tau = 0.1$

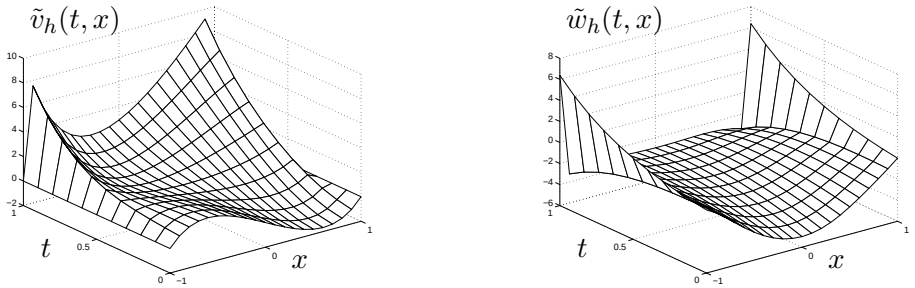

Abbildung 3.4. PDAE (3.5.52), inkonsistente RB, $h = \frac{1}{15}$, $\tau = 0.1$

Vergleichen wir die Abbildungen 3.3 und 3.4, so sehen wir, daß die Lösungen für konsistente und inkonsistente Randbedingungen sehr unterschiedlich sind. Dies war zu erwarten, da wir aufgrund der unterschiedlichen Randbedingungen auch unterschiedliche Probleme gelöst haben. Aber insbesondere die zweite Komponente $\tilde{w}_h(t,x)$ der numerischen Lösung, die mit inkonsistenten RBn berechnet wurde hat in der Nähe der Randpunkte ± 1 teilweise sehr große Gradienten.

3.5.5.3. Das Problem eines Indexsprungs in der MOL-DAE. Angenommen die Matrizen A, B, C sind durch

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \; B = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 1 \\ 1 & c \end{pmatrix} \tag{3.5.53}$$

definiert. Sei weiterhin $\rho_k := -\left(\frac{k\pi}{2l}\right)^2$, $k = 1, 2, \ldots$, der k-te Eigenwert des Operators $\frac{\partial^2}{\partial x^2}$, den wir bereits in Voraussetzung 3.2.1 d) eingeführt haben. Es ist leicht zu sehen, daß das Matrixbüschel $(A, \rho_k B + C)$ den Riesz-Index 1 für $c \neq \rho_k$ und den Riesz-Index 2 für $c = \rho_{\bar{k}}$ für ein $\bar{k}$ hat, d.h., die PDAE hat den einheitlichen differentiellen Zeitindex $\nu_{d,t} = 1$, wenn $c \neq \rho_k \; \forall k \in \mathbb{N}^+$ gilt, und einen Indexsprung, $c = \rho_{\bar{k}}$ für ein $\bar{k} \in \mathbb{N}^+$. Es kann natürlich vorkommen, daß die MOL-DAE (3.5.2) einen Indexsprung für spezielle Matrizen A, B, C und gewisse h besitzt auch wenn die PDAE selbst einen einheitlichen Zeitindex hat. In diesem Beispiel würde dieser Fall eintreten, wenn $c = \lambda_{\bar{k}}$ ($c \neq \rho_k \forall\, k$) für ein gewisses $h > 0$ (dies kann man zeigen durch eine Lineartransformation unter Verwendung des Kronecker Produktes).

Sei $g(t, x)$ so gewählt, daß die exakte Lösung der PDAE durch

$$u(t, x) = \begin{pmatrix} v(t, x) \\ w(t, x) \end{pmatrix} := \begin{pmatrix} x^5 \left(x^4 - 1\right) \cos(\pi t) \\ x^2 \left(x^2 - 1\right) e^{-t} \end{pmatrix}.$$

gegeben ist. Sei $c := \lambda_M(h) = -\frac{4}{h^2} \sin^2\left(\frac{\pi h}{4}\right)$ für $h = h_0 = 0.05$ ($M = 39$). (Der numerische Wert von $\lambda_M(h_0)$ ist -1597.5.) Die Abbildungen 3.5-3.7 zeigen den absoluten Fehler err_h^w der w-Komponente von der numerischen Lösung zur exakten Lösung in jedem Ortsgitterpunkt und den Zeitintegrationschritten. Wir sehen, daß

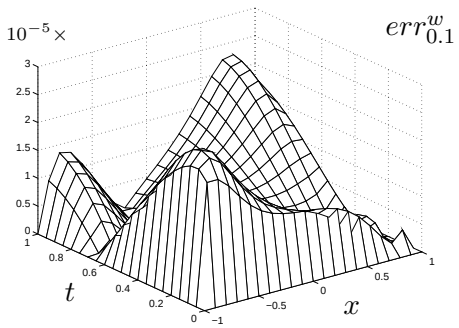

ABBILDUNG 3.5. Bsp. (3.5.53): $|w_{num}(t, x) - w(t, x)|$, $h = 0.1$, $\tau = 0.05$

der Fehler für $h = h_0$ (Abb. 3.6) größer ist als für $h \neq h_0$ (Abb. 3.5,3.7). Solche Effekte können auch in komplizierteren Beispielen auftreten. Es sei bemerkt, daß da $\lambda_k(h) = \rho_k + \mathcal{O}(h^2)$ für $h \to 0$, für hinreichend kleine h kein Indexsprung in der

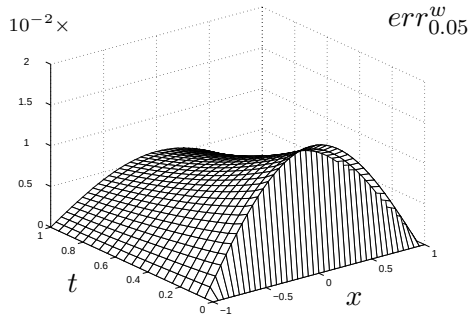

ABBILDUNG 3.6. Bsp. (3.5.53): $|w_{num}(t,x) - w(t,x)|$, $h = 0.05$, $\tau = 0.05$

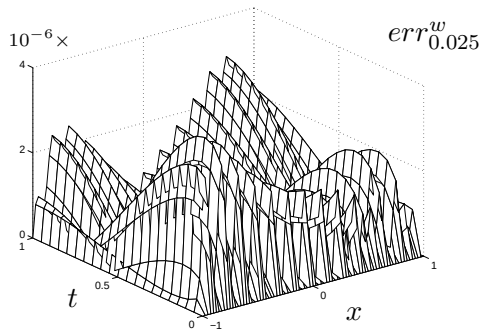

ABBILDUNG 3.7. Bsp. (3.5.53): $|w_{num}(t,x) - w(t,x)|$, $h = 0.025$, $\tau = 0.05$

MOL-DAE auftritt. Diese Beispiel demonstriert uns schließlich, daß wir vorsichtig bei der numerischen Behandlung von PDAEs sein müssen.

Wir bemerken weiterhin, daß der Fehler in Abbildung 3.6 auch von einem anderen Standpunkt aus erklärt werden kann: Hierfür betrachten wir die M DAEs (3.5.14) für den zeitabhängigen Vektor $\omega_k \in \mathbb{R}^n$. In dem hier betrachteten Beispiel kann mit $\omega_k = (\omega_{k1}, \omega_{k2})^\top$ die DAE für ein k geschrieben werden als

$$\begin{aligned} \omega'_{k1} + [1 - \lambda_k(h)\,]\omega_{k1} + \omega_{k2} &= g_{k1} \\ \omega_{k1} + [c - \lambda_k(h)\,]\omega_{k2} &= g_{k2}, \end{aligned}$$

was für $\omega_{k1}(t)$ die einfache ODE

$$w'_{k1} + \sigma_k(h)\omega_{k1} = f_{k1} - \frac{f_{k2}}{c - \lambda_k(h)}$$

ergibt, wobei zur Abkürzung $\sigma_k(h) = 1 - \lambda_k(h) + 1/[\lambda_k(h) - c]$ sei. Für unsere Beispielrechnung wählen wir $k = 1$. Die Stabiltät von ω_1 und das hier betrachtete numerische Verfahren hängen vom Vorzeichen von $\sigma_1(h)$ ab, für das $\sigma_1(h) > 0$ für $h > h_0$ und $\sigma_1(h) < 0$ für $\underline{h}_0 \le h < h_0$ gilt, wobei $\underline{h}_0$ eine Schrittweite nahe h_0 ist.

Weiterhin gilt

$$\lim_{h \to h_0} |\sigma_1(h)| = \infty.$$

Aus diesen Beziehungen erhalten wir, daß unsere numerische Methode instabil für $\underline{h_0} \leq h < h_0$ ist. Dies wird wiedergegeben in Abbildung 3.6 für den Fall $h = 0.05$. Ein ähnliches Phänomen wird auch in [Arn98a], [Söd92] betrachtet.

3.6. Zusammenfassung

In diesem Kapitel wurden lineare partielle differentiell-algebraische Systeme vorgestellt und auf deren Besonderheiten eingegangen. Es wurde ein Paar differentieller Indexe $(\nu_{d,x}, \nu_{d,t})$ zur Charakterisierung linearer PDAEs mit konstanten Koeffizientenmatrizen der Form (3.1.1) definiert. Darüberhinaus wurde die Problematik der Vorgabe konsistenter Anfangs- bzw. Randbedingungen erläutert.

Im Anschluß wurde die numerische Behandlung linearer PDAEs anhand zweier Diskretisierungsverfahren, dem BTCS Schema und dem Crank-Nicolson Verfahren, betrachtet. Die Untersuchung der Konvergenz dieser beiden Verfahren wurde für den Fall geführt, daß die äquidistanten Orts- und Zeitschrittweiten gegen Null streben. Im Gegensatz zum Konvergenzverhalten dieser beiden Verfahren für reguläre Systeme (3.1.1) (d.h., wenn beide Matrizen A und B regulär sind), war das Ergebnis, daß eine starke Abhängigkeit der Konvergenz von den beiden eingeführten Indexen besteht. Dieses Ergebnis wurde unter starken Voraussetzungen gefunden. Für spezielle lineare PDAEs kann die Struktur der PDAE bei der Konvergenzanalyse berücksichtigt werden und eine Konvergenzaussage unter weniger starken Voraussetzungen getroffen werden. Numerische Testrechnungen illustrieren die theoretischen Ergebnisse.

Aus den geführten Betrachtungen schließen wir, daß die allgemeinen Betrachtungen von linearen PDAEs für spezielle PDAEs vereinfacht werden können. Somit ist es für PDAEs, die aus der Modellierung praktischer Anwendungen resultieren, günstiger, direkt die Semidiskretisierungen dieser Systeme zu untersuchen, wie dies auch in [Sim96],[Arn98c], [Kun94],[Kun96], [Wei96] getan wurde.

Die theoretischen Untersuchungen und numerischen Experimente dieses Kapitels bilden die Grundlage für weitere Untersuchungen, insbesondere zur Entwicklung effektiver numerischer Verfahren zur Lösung von PDAEs.

Literaturverzeichnis

[Alt98] H. Altenbach, P. Deuring, K. Naumenko. A system of ordinary and partial differential equations describing creep behaviour of thin-walled shells. Technical Report 98-10, Martin-Luther-Universität Halle, Fachbereich Mathematik und Informatik, 1998.

[Ans87] K.M. Anstreicher, U.G. Rotblum. Using Gauss-Jordan elimination to compute the index, generalized nullspace and Drazin inverse. *Linear Algebra Appl.*, 85:221–239, 1987.

[Arn98a] M. Arnold. A note on the uniform perturbation index. Rostocker Math. Kolloquium, Heft 52, 1998.

[Arn98b] M. Arnold. Zur Theorie und zur numerischen Lösung von Anfangswertproblemen für differentiell-algebraische Systeme vom höheren Index (Habilitationsschrift). Forschungsberichte des VDI Verlag Düsseldorf, Reihe 20, 1998.

[Arn98c] M. Arnold, B. Simeon. The simulation of pantograph and catenary: a PDAE approach. TU Darmstadt, Fachbereich Mathematik, Preprint Nr. 1990, 1998.

[Auz90] W. Auzinger. On error structures and extrapolation for stiff systems, with application in the method of lines. *Computing*, 44(4):331–356, 1990.

[Bel84] R.E. Bellman. *The Laplace transform*. Robert S. Roth. - Singapore: World Scientific, 1984.

[Bra65] R. Bracewell. *The Fourier transform and its applications*. New York [u.a.]: McGraw-Hill, 1965.

[Bre89] K.E. Brenan, S.L. Campbell, L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. North-Holland Publ. Co., Amsterdam, 1989.

[Cam76] 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 J. Appl. Math.*, 31(3):411–425, November 1976.

[Cam95] S.L. Campbell, C.W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

[Cam96a] S.L. Campbell, W. Marszalek. The index of an infinite dimensional implicit system. *Math. Mod. of Syst.*, 1(1):1–25, 1996.

[Cam96b] S.L. Campbell, W. Marszalek. ODE/DAE integrators and MOL problems. *ZAMM*, 76:S1, 251–254, 1996.

[Cam97] S.L. Campbell, W. Marszalek. DAEs arising from traveling wave solutions of PDEs. *J. Comput. Appl. Math.*, 82(1-2):41–58, 1997.

[Can73] J.R. Cannon, R.E. Klein. On the observability and stability of the temperature distribution in a composite heat conductor. *SIAM J. Appl. Math.*, 24:596–602, 1973.

[Cha70] Fung-Yual Chang. Transient analysis of lossless transmission lines in a nonhomogeneous dielectric medium. *IEEE Trans. Microwave Theory Tech.*, MIT-18:616–626, 1970.

[Cro79] M. Crouzeix. Sur la B-stabilité des méthodes de Runge-Kutta. *Numer. Math.*, 32:75–82, 1979.

[Dah59] G. Dahlquist. Stability and error bounds in the numerical integration of ordinary differential equations. *Trans. of Royal Inst. of Techn., No. 130, Stockholm*, 1959.

[Dah85] G. Dahlquist. 33 years of numerical instability. I. *BIT*, 25:188–204, 1985.

[Dek84] K. Dekker, J.G. Verwer. *Stability of Runge-Kutta methods for stiff nonlinear differential equations*. North Holland, 1984.

[Deu94] P. Deuflhard, F. Bornemann. *Numerische Mathematik II: Integration gewöhnlicher Differentialgleichungen*. de Gruyter, Berlin, 1994.

[Dor93] J.L.M. van Dorsselaer, J.F.B.M. Kraaijevanger, M.N Spijker. Linear stability analysis in the numerical solution of initial value problems. In A. Iserles, editor, *Acta Numerica*. Cambridge University Press, 1993.

[Dou56] J. Douglas, H. H. Rachford. On the numerical solution of heat conduction problems in two and tree space variables. *Trans. Amer. Math. Soc.*, 82:421–453, 1956.

[Dou62] J. Douglas. Alternating direction method for three space variables. *Numer.Math.*, 4:41–63, 1962.

[EL98a] C. Eichler-Liebenow, N.H. Cong, R. Weiner, K. Strehmel. Linearly implicit splitting methods for higher space-dimensional parabolic differential equations. *Appl. Numer. Math.*, 28:259–274, 1998.

[EL98b] C. Eichler-Liebenow, P. J. van der Houwen, B. P. Sommeijer. Analysis of approximate factorization in iteration methods. *Appl. Numer. Math.*, 28:245–258, 1998.

[ES98] E. Eich-Soellner, C. Führer. *Numerical methods in multibody dynamics*. B.G. Teubner Stuttgart, 1998.

[Fra81] R. Frank, J. Schneid, C.W. Ueberhuber. The concept of B-convergence. *SIAM J. Numer. Anal.*, 18:753–780, 1981.

[Gan86] F.R. Gantmacher. *Matrizentheorie*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1986.

[Gea88] C.W. Gear. Differential-algebraic index transformations. *SIAM J. Sci. Stat. Comp.*, 9:39–47, 1988.

[Gea90] C.W. Gear. Differential-algebraic equations, indices, and integral algebraic equations. *SIAM J. Numer. Anal.*, 27:1527–1534, 1990.

[Got77] D. Gottlieb, S.A. Orszag. Numerical analysis of spectral methods: theory and applications. *CBMS-NSF Regional Conference Series in Applied Mathematics*, No. 26, 1977.

[Gri86] E. Griepentrog, R. März. *Differential-algebraic equations and their numerical treatment*. Teubner-Texte zur Mathematik, Band 88, Leipzig, 1986.

[Gri96] P. Grindrod. *The theory and applications of reaction–diffusion equations*. Clarendon Press, Oxford, 1996.

[Grö78] K. Gröger, M. Schleiff. Bestimmung der Temperaturverteilung in einem stromdurchflossenen Körper. *ZAMM*, 12:547–553, 1978.

[Gro92] Ch. Großmann, H.-G. Roos. *Numerik partieller Differentialgleichungen*. B.G. Teubner, Stuttgart, 1992.

[Gün98] Y. Günther, M. Wagner. Index Concepts for Linear Mixed Systems of Differential-Algebraic and Hyperbolic-Type Equations. Technical Report Preprint Nr. 2012, Technische Universität Darmstadt, Fachbereich Mathematik, 1998.

[Hai82] E. Hairer, G. Bader, Ch. Lubich. On the stability of semi-implicit methods for ordinary differential equations. *BIT*, 22:211–232, 1982.

[Hai89] E. Hairer, Ch. Lubich, M. Roche. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, volume 1409. Springer-Verlag, 1989. Lecture Notes in Mathematics.

[Hai93] E. Hairer, S.P. Nørsett, G. Wanner. *Solving ordinary differential equations I.* Springer-Verlag, 1993.

[Hai96] E. Hairer, G. Wanner. *Solving Ordinary Differential Equations II.* Springer-Verlag, 1996.

[Hou79] P. J. van der Houwen, J. G. Verwer. One-Step Splitting Methods for Semi-Discrete Parabolic Equations. *Computing*, 22:291–309, 1979.

[Hou97] P.J. van der Houwen, B.P. Sommeijer, J. Kok. The iterative solution of fully implicit discretizations of three-dimensional transport models. *Appl. Numer. Math.*, 25(2-3):243–256, 1997.

[Hun89] W. H. Hundsdorfer, J. G. Verwer. Stability and Convergence of the Peaceman-Rachford ADI Method for Initial-Boundary Value Problems. *Mathematics of Computations*, 53(187):81–101, 1989.

[Hun92] W. H. Hundsdorfer. Unconditional convergence of some Crank-Nicolson methods for initial-boundary value problems. *Math. Comput.*, 58(197):35–53, 1992.

[Hun98a] W. Hundsdorfer. A note on stability of the Douglas splitting method. *Math. Comput.*, 67(221):183–190, 1998.

[Hun98b] Hundsdorfer, Willem. Trapezoidal and midpoint splittings for initial-boundary value problems. *Math. Comput.*, 67(223):1047–1062, 1998.

[Jod90] Jodar, Lucas. Computing accurate solutions for coupled systems of second order partial differential equations. *Int. J. Comput. Math.*, 37(3/4):201–212, 1990.

[Jod91] L. Jodar, M. Legua Fernandez. An implicit difference method for the numerical solution of coupled systems of partial differential equations. *Appl. Math. Comput.*, 46(2):127–134, 1991.

[Jor94] J.C. Jorge, F. Lisbona. Contractivity results for alternating direction schemes in Hilbert spaces. *Appl. Numer. Math.*, 15:65–75, 1994.

[Kro90] L. Kronecker. Algebaische Reduktion der Schaaren bilinearer Formen. *Akad. der Wiss. Berlin*, Werke vol. III:141–155, 27. Nov. 1890.

[Kun94] P. Kunkel, V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *J. Comp. Appl. Math.*, 56:225–259, 1994.

[Kun96] P. Kunkel, V. Mehrmann. Local and global invariants of linear differential-algebraic equations and their relation. *Electron. Trans. Numer. Anal.*, 4:138–157, 1996.

[Leu89] A.W. Leung. *Systems of nonlinear partial differential equations.* Kluwer Acad. Publ., Dortrecht, Boston, London, 1989.

[Lie94] C. Liebenow. Linear-implizte Splitting-Methoden für zweidimensionale parabolische Differentialgleichungen. Diplomarbeit, Martin-Luther-Universität Halle-Wittenberg, Fachbereich Mathematik und Informatik, 1994.

[Lin97] Lin, Ping. A sequential regularization method for time-dependent incompressible Navier-Stokes equations. *SIAM J. Numer. Anal.*, 34(3):1051–1071, 1997.

[Luc91] W. Lucht, Th. Radke. A model and a method for the determination of electric fields. In *Numerical treatment of differential equations (Halle,1989)*, S. 351–355. Teubner, Stuttgart, 1991.

[Luc97a] W. Lucht, K. Strehmel, C. Eichler-Liebenow. Linear partial differential algebraic equations. Part I: Indexes, consistent boundary/initial conditions. Technical Report 97-17, Martin-Luther-Universität Halle, Fachbereich Mathematik und Informatik, 1997.

[Luc97b] W. Lucht, K. Strehmel, C. Eichler-Liebenow. Linear partial differential algebraic equations. Part II: Numerical solution. Technical Report 97-18, Martin-Luther-Universität Halle, Fachbereich Mathematik und Informatik, 1997.

[Luc98] W. Lucht, K. Strehmel. Discretization based indices for semilinear partial differential algebraic equations. *Appl. Numer. Math.*, 28:371–386, 1998.

[Luc99] W. Lucht, K. Strehmel, C. Eichler-Liebenow. Indexes and special discretization methods for linear partial differential algebraic equations. *BIT*, 39(3):484–512, 1999.

[Mar90] G. I. Marchuk. Splitting and alternating direction methods. In P. G. Ciarlet, J. L. Lions, editors, *Handbook of numerical analysis. Vol. I.*, S. 197–462. North-Holland Publishing Co., Amsterdam, 1990.

[Mar97] W. Marszalek. *Analysis of partial differential algebraic equations.* Dissertationsschrift, North Carolina State University, Raleigh, 1997.

[Mär98] R. März. EXTRA-ordinary differential equations: Attempts to an analysis of differential-algebraic systems. In A. et. al. Balog, editor, *European congress of mathematics (ECM), Budapest, Hungary, July 22-26, 1996*, volume I of *Prog. Math. 168*, S. 313–334. Basel: Birkhaeuser, 1998.

[Mat86] E. J. W. ter Maten. Splitting methods for fourth order parabolic partial differential equations. *Computing*, 37(4):335–350, 1986.

[Mei90] P. Meinhold, E. Wagner. *Partielle Differentialgleichungen.* Leipzig: BSB B.G. Teubner Verlagsgesellschaft, 1990.

[Mit78] A.R. Mitchell, R. Wait. *The finite element methode in partial differential equations.* John Wiley & Sons, Chichester, 1978.

[Mit80] A.R. Mitchell, D.F. Griffiths. *The finite difference methode in partial differential equations.* John Wiley & Sons, Chichester, 1980.

[Naa72] J. Naas, H. L. Schmid. *Mathematisches Wörterbuch.* Akademie-Verlag GMBH Berlin, BSB B. G. Teubner Verlagsgesellschaft Leipzig, 1972.

[Neu51] J. von Neumann. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. *Math. Nachr.*, 4:258–281, 1951.

[Pea55] D. W. Peaceman, H. H. Jr. Rachford. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3(1):March, 1955.

[Pet82] L.R. Petzold. Differential-algebraic equations are not ODE's. *SIAM J. Sci. Stat. Comput.*, 3:367–384, 1982.

[Pip90] K.G. Pipilis. *Higher order moving finite elements method for systems described by partial differential-algebraic equations.* Dissertationsschrift, Dept. of Chemical Engineering, Imperial College of Science, Technology and Medicine, 1990.

[Ren96] P. Rentrop, K. Strehmel, R. Weiner. Ein Überblick über Einschrittverfahren zur numerischen Integration in der technischen Simulation. *GAMM-Mitteilungen*, 1:9 – 43, 1996.

[Ric67] R. D. Richtmyer, K. W. Morton. *Difference methods for initial–value problems.* Interscience Publ., New York, 1967.

[Sam84] A.A. Samarskij. *Theorie der Differenzenverfahren.* Akad. Verlagsgesellschaft Geest & Portig K.-G., Leipzig, 1984.

[Sez87] M. Sezgin. Magnetohydrodynamic flow in a rectangular channel. *Internat. J. Numer. Methods Fluids*, 7:697–718, 1987.

[Sha94] L. F. Shampine. ODE Solvers and the Method of Lines. *Numer. Methods Partial Differ. Equations*, 10:739–755, 1994.

[She89] Qin Sheng. Solving linear partial differential equations by exponential splitting. *IMA Journal of Numerical Analysis*, 9:199–212, 1989.

[She93] Qin Sheng. Global errror estimates for exponential splitting. *IMA Journal of Numerical Analysis*, 14:27–56, 1993.

[Sim93] B. Simeon, C. Führer, P. Rentrop. The Drazin inverse in multibody system dynamics. *Numer. Math.*, 64(4):521–539, 1993.

[Sim96] B. Simeon. Modelling a flexible slider crank mechanism by a mixed system of DAEs and PDEs. *Math. Mod. of Syst.*, 2(1):1–18, 1996.

[Smi58] W.L. Smirnov. *Lehrgang der höheren Mathematik. Teil IV*. Verlag der Wissenschaften, Berlin, 1958.

[Söd92] G. Söderlind. Remarks on the stability of high–index DAEs with respect to parametric perturbations. *Computing*, 49:303–314, 1992.

[SS86] 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. *Numer. Math.*, 50:405–418, 1986.

[Str92] K. Strehmel, R. Weiner. *Linear-implizite Runge-Kutta-Methoden und ihre Anwendung*. B.G. Teubner Stuttgart-Leipzig, 1992.

[Str95] K. Strehmel, R. Weiner. *Numerik gewöhnlicher Differentialgleichungen*. B.G. Teubner Stuttgart-Leipzig, 1995.

[Tho95] J.W. Thomas. *Numerical partial differential equations: Finite difference methods*. Springer, New York, 1995.

[Ver84] J.G. Verwer, J.M. Sanz-Serna. Convergence of method of lines approximations to partial differential equations. *Computing*, 33:297–313, 1984.

[Wal70] W. Walter. *Differential and Integral Inequalities*. Springer-Verlag New York, 1970.

[War79] R.F. Warming, R.M. Beam. An extension of A-stability to alternating direction methods. *BIT*, 19:395–417, 1979.

[Wei68] K. Weierstraß. Zur Theorie der bilinearen und quadratischen Formen. *Akad. der Wiss. Berlin*, Werke Vol. II:19–44, 18. Mai 1868.

[Wei96] J. Weickert. Navier-Stokes equations as a differential–algebraic system. Preprint SFB 393/96-08, Technische Universität Chemnitz–Zwickau, 1996.

[Wen98] J. Wensch. interne Diskussion. Martin-Luther-Universität Halle, Fachbereich Mathematik und Informatik, 1998.

[Yan71] N. N. Yanenko. *The method of fractional steps*. Springer Berlin-Heidelberg-New York, 1971.

Verwendete Abkürzungen und Bezeichnungen

Abkürzungen:

PDE	partielle Differentialgleichung (partial differential equation)
PDAE	partielle differentiell-algebraische Gleichung (engl.: partial differential algebraic equation)
ODE	gewöhnliche Differentialgleichung (ordinary differential equation)
DAE	(gewöhnliche) differentiell-algebraische Gleichung ((ordinary) differential algebraic equation)
AB	Anfangsbedingung
RB	Randbedingung
ARWP	Anfangs-Randwertproblem
RWP	Randwertproblem
AWP	Anfangswertproblem

Bezeichnungen:

u	Lösung der PDE/PDAE
t	Zeitvariable
x	Ortsvariable
g	rechte Seite der PDE/PDAE
d	Zahl der Raumdimensionen
n	Anzahl der partiellen Differentialgleichungen
u_h	Restriktion der exakten Lösung auf das Ortsgitter

w	exakte Lösung der semidiskreten Problems
v_m	numerische Lösung des semidiskreten Problems zum Zeitpunkt t_m
m	Laufindex der Zeitintegration
M	Zahl der Ortsgitterpunkte einer Dimension bei äquidistantem Gitter
r	Anzahl der gewöhnlichen Differentialgleichungen des semidiskreten Problems ($r = n\,M^d$)
$\mathfrak{I}$	Zeitintervall ($\mathfrak{I} = (0, t_e)$)
t_e	Endpunkt des Zeitintervalls
τ	Zeitschrittweite
$\mathfrak{I}_\tau$	Punktgitter auf Zeitintervall $[0, t_e]$
Ω	Gebiet bez. der Ortsvariablen (z.B. $\Omega = (a, b)^d$)
l	Intervallgrenze des Ortsintervalls bei PDAEs, d.h. $\Omega = (-l, l)$
h	Ortsschrittweite (bei äquidistanten Ortsgitter: $h = \frac{b-a}{M+1}$)
Ω_h	Ortsgitter
$\alpha_h(t_m)$	lokaler Ortsdiskretisierungsfehler zum Zeitpunkt t_m
$\eta_h(t_m)$	globaler Ortsdiskretisierungsfehler zum Zeitpunkt t_m
$le_\tau(t_m)$	lokaler Zeitdiskretisierungsfehler zum Zeitpunkt t_m
$e_\tau(t_m)$	globaler Zeitdiskretisierungsfehler zum Zeitpunkt t_m
$le_h(t_m)$	lokaler Gesamtdiskretisierungsfehler zum Zeitpunkt t_m
$\varepsilon_h(t_m)$	globaler Gesamtdiskretisierungsfehler zum Zeitpunkt t_m
$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathbb{N}^+$	$= \mathbb{N} \setminus \{0\}$ Menge der positiven natürlichen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{C}$	Menge der komplexen Zahlen
$\mathbb{C}^-$	Menge der komplexen Zahlen mit negativen Realteil

$\mathcal{O}(.)$	Landau-Symbol: $f(t) = \mathcal{O}(g(t))$, falls $f(t)/g(t)$ beim zugrunde liegenden Grenzprozeß $t \to \alpha$ beschränkt bleibt
$I_k \in \mathbb{N}^{k\times k}$	Einheitsmatrix
$\mathfrak{M}_{RB}, \mathfrak{M}_{RB}^{\xi}$	Indexmenge für Komponenten, für die RBn vorgegeben werden können
$\mathfrak{M}_{AB}, \mathfrak{M}_{AB}^{k}$	Indexmenge für Komponenten, für die ABn vorgegeben werden können
$\nu_{d,t}$	einheitlicher differentieller Zeitindex der PDAE
$\nu_{d,x}$	differentieller Ortsindex der PDAE
$S_{L,\xi}, T_{L,\xi}$	reguläre Matrizen der Kronecker-Transformation im Laplace-Raum
$S_{F,k}, T_{F,k}$	reguläre Matrizen der Kronecker-Transformation im Fourier-Raum
$\|.\|_{2,n}$	Euklidische Norm (2-Norm) im $\mathbb{R}^n$: $\|y\|_{2,n} = \sqrt{y^\top y} = \sqrt{\sum_{i=1}^n y_i^2}$, $y \in \mathbb{R}^n$ für Matrizen $M \in \mathbb{R}^{n\times n}$ sei $\|M\|_{2,n}$ die Spektralnorm, die der Euklidischen Norm zugeordnet ist
$\|.\|_{L_2}$	L_2-Norm: $\|f\|_{L_2} = \sqrt{\int_a^b f^2(y)\,dy}$ für $f : [a,b] \to \mathbb{R}$ mit $\int_a^b f^2(y)\,dy < \infty$ (d.h., $f \in L_2[a,b]$)
$\|.\|$	diskrete L_2-Norm: $\|y\| = \sqrt{h\sum_{k=1}^M \|y_k\|_{2,n}^2}$, $y = (y_1^\top, \ldots, y_M^\top)^\top \in \mathbb{R}^{nM}$, $y_k \in \mathbb{R}^n$ $(k = 1(1)M)$, für Matrizen die zugeordnete Matrixnorm
$\mu[.]$	die durch die gewählte Vektornorm induzierte logarithmische Matrixnorm
$\otimes$	Kronecker-Produkt: für $A \in \mathbb{R}^{m\times n}$, B beliebige Matrix ist $A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \ldots & a_{1n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \ldots & a_{mn}B \end{pmatrix}$
δ_{ij}	Kronecker-Symbol: $\delta_{ij} = \begin{cases} 1 & \text{für} \quad i = j \\ 0 & \text{für} \quad i \neq j \end{cases}$

Lebenslauf

Persönliche Daten:	Claudia Eichler-Liebenow
	Nickel-Hoffmann-Str. 20 06110 Halle (Saale)
	geboren am: 14.05.1971 in Mühlhausen/Thür.
	Familienstand: verheiratet
1977 - 1987	Besuch der Polytechnischen Oberschule in Mühlhausen/Thür.
1987 - 1989	Besuch der Spezialklassen für Mathematik und Physik der Martin-Luther-Universität Halle-Wittenberg, Abitur
1989 - 1994	Mathematikstudium und Diplom an der Martin-Luther-Universität Halle-Wittenberg
Okt. 1994 - März 1997	Stipendium nach Graduiertenförderung des Landes Sachsen-Anhalt zum Promotionsstudium am Institut für Numerische Mathematik des Fachbereich Mathematik und Informatik der Martin-Luther-Universität Halle-Wittenberg
Apr. 1997 - März 1999	wissenschaftliche Mitarbeiterin am Fachbereich Mathematik und Informatik der Martin-Luther-Universität Halle-Wittenberg im Rahmen des Drittmittelprojektes „Math-Net - Informationsdienste für die Mathematik im Internet“

Erklärung

Hiermit erkläre ich, Claudia Eichler-Liebenow, an Eides statt, daß ich die vorliegende Arbeit selbständig, ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die aus anderen Werken wörtlich oder inhaltlich entnommenen Daten, Fakten und Konzepte sind unter Angabe der ensprechenden Quelle als solche gekennzeichnet.

Diese Arbeit wurde bisher weder im In- noch Ausland in gleicher oder ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt.

Halle/Saale, im Februar 1999

Claudia Eichler-Liebenow