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-

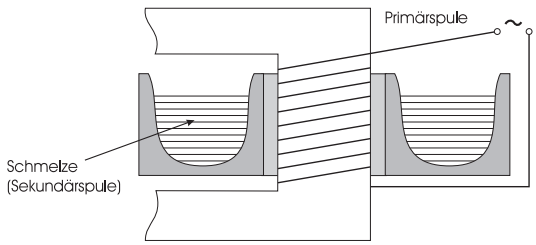

Abbildung 3.1. Schmelzen fester Materialien in Elektro-Öfen

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.

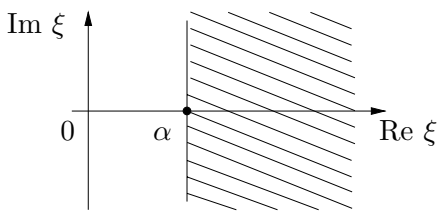

ABBILDUNG 3.2. Konvergenzbereich des Laplace-Integrals

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.

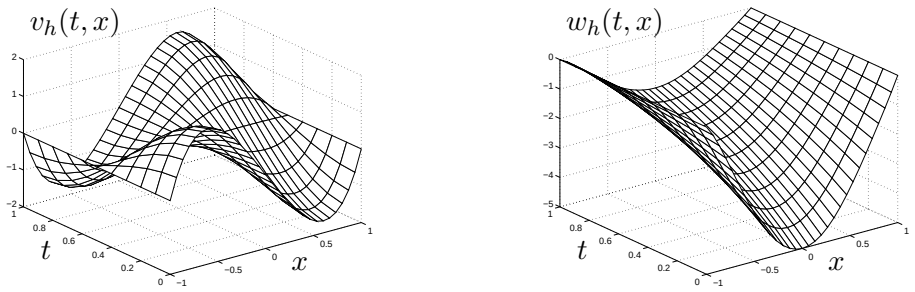

Abbildung 3.3. PDAE (3.5.52), konsistente RBn, $h = \frac{1}{15}$, $\tau = 0.1$

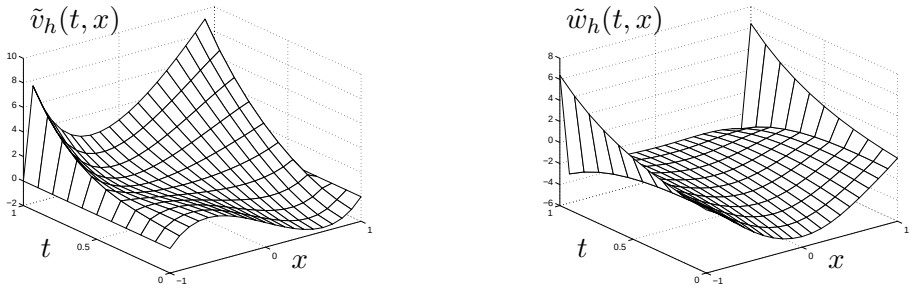

Abbildung 3.4. PDAE (3.5.52), inkonsistente RB, $h = \frac{1}{15}$, $\tau = 0.1$

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

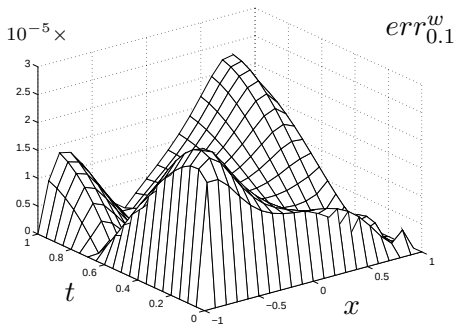

ABBILDUNG 3.5. Bsp. (3.5.53): $|w_{num}(t, x) - w(t, x)|$, $h = 0.1$, $\tau = 0.05$

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

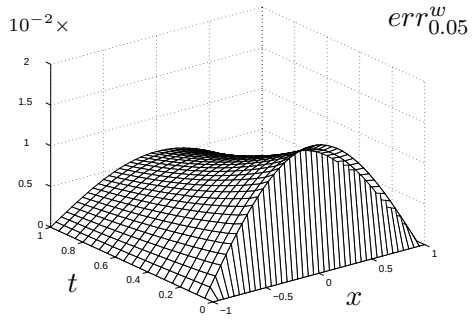

ABBILDUNG 3.6. Bsp. (3.5.53): $|w_{num}(t,x) - w(t,x)|$, $h = 0.05$, $\tau = 0.05$

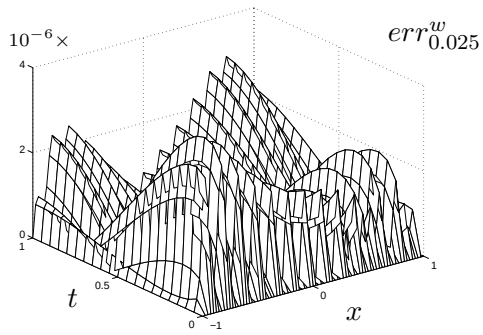

ABBILDUNG 3.7. Bsp. (3.5.53): $|w_{num}(t,x) - w(t,x)|$, $h = 0.025$, $\tau = 0.05$

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.