Analysis und Numerik linearer differentiell-algebraischer Gleichungen

Peter Kunkel und Volker Mehrmann

16. Dezember 1994

Inhaltsverzeichnis

1 Problemstellung 2

2 Konstante Koeffizienten 6

3 Variable Koeffizienten 13

4 Numerische Verfahren 26

Kapitel 1

Problemstellung

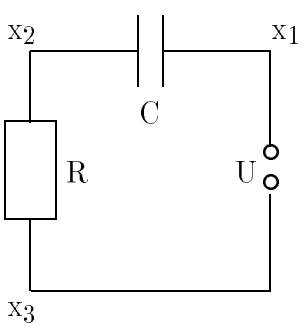
Viele physikalische Vorgänge werden mathematisch durch Differentialgleichungssysteme modelliert. Gibt es im physikalischen System allerdings Einschränkungen an den Zustand durch Erhaltungsgleichungen (etwa in der Art der Kirchhoff'schen Gesetze) oder Zwangsbedingungen (etwa Einschränkung der Bewegung eines Massepunktes auf eine vorgegebene Fläche), so kann das mathematische Modell neben Differentialgleichungen auch algebraische Gleichungen beinhalten. Man nennt solche gemischten Systeme nach diesen beiden Bestandteilen auch differentiell-algebraische Gleichungssyteme. Die allgemeinste Form differentiell-algebraischer Gleichungen ist gegeben durch

$$F(t, x, \dot{x}) = 0, \tag{1.1}$$

wobei $\dot{x}$ die Ableitung von x nach t bezeichnet. Dabei ist die Funktion F gegeben, während eine Funktion x gesucht ist, die in einem vorzugebenden Sinn die Gleichung (1.1) löst. In den Eigenschaften dieser Gleichung werden sich die Eigenschaften sowohl von Differentialgleichungen wie auch von algebraischen Gleichungen wiederfinden lassen. Aber es ist auch das Auftreten neuer Phänomene denkbar.

Beispiel 1 Um ein mathematisches Modell für die Aufladung eines Kondensators über einen Widerstand mittels einer Gleichspannungsquelle zu gewinnen,

Abbildung 1.1 Eine elektronische Schaltung



ordnen wir jedem Leitungsabschnitt ein Potential x_i, $i = 1,2,3$ zu, vergleiche Abbildung 1.1. Die Gleichspannungsquelle hebt das Potential von x_3 auf x_1 um U, d. h. $x_1 - x_3 - U = 0$. Da nach dem ersten Kirchhoff'schen Gesetz in jedem Knoten die Summe der Ströme verschwindet, gilt für den zweiten Knoten unter der Annahme idealer elektronischer Bauteile $C(\dot{x}_1 - \dot{x}_2) + (x_3 - x_2)/R = 0$, wenn R die Größe des Widerstands und C die Kapazität des Kondensators bezeichnet. Schließlich können wir das Nullpotential noch frei wählen, z. B. $x_3 = 0$. Man erhält somit als Modell das differentiell-algebraische Gleichungssystem

$$\begin{aligned} x_1 - x_3 - U &= 0, \\ C(\dot{x}_1 - \dot{x}_2) + (x_3 - x_2)/R &= 0, \\ x_3 &= 0. \end{aligned}$$

Beispiel 2 Ein Massepunkt mit Masse m und kartesischen Koordinaten (x, y) soll sich unter Einfluß der Erdanziehung in einem festen Abstand l um den Ursprung bewegen (physikalisches Pendel, vgl. Abbildung 1.2). Mit der kinetischen Energie $T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2)$ und der potentiellen Energie $U = mgy$, wobei g die Erdbeschleunigung bezeichnet, sowie der Zwangsbedingung $x^2 + y^2 - l^2 = 0$ erhalten wir die sogenannte Lagrange-Funktion

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) - mgy - \lambda(x^2 + y^2 - l^2)$$

mit dem Lagrange-Parameter λ. Die gesuchten Bewegungsgleichungen ergeben sich daraus in der Form

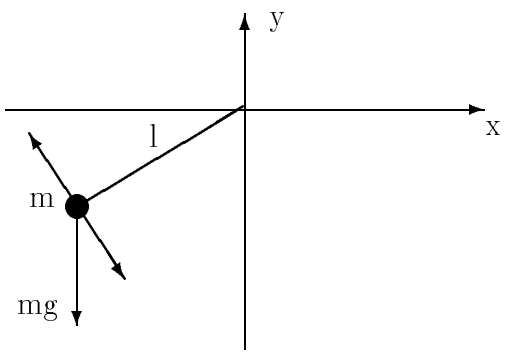
$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0$$

für $q = x, y, \lambda$, d. h.

$$\begin{aligned} m\ddot{x} + 2x\lambda &= 0, \\ m\ddot{y} + 2y\lambda + mg &= 0, \\ x^2 + y^2 - l^2 &= 0. \end{aligned}$$

Natürlich kann man in diesen einfachen Beispielen die erhaltenen differentiell-algebraischen Gleichungen dadurch lösen, daß man die Gleichungen per Hand auflöst oder umparametrisiert, um schließlich auf eine mehr oder weniger leicht lösbare Differentialgleichung zu kommen. Bei komplizierteren Problemen ist eine solche Vorgehensweise nicht mehr möglich. Man benötigt dann numerische Verfahren, die Gleichungen der Form (1.1) direkt lösen. Dabei sollte die numerische Diskretisierung die wesentlichen Eigenschaften des vorgegebenen Problems widerspiegeln. Dazu ist es zuerst nötig, die analytischen Eigenschaften einer differentiell-algebraischen Gleichung (1.1) zu untersuchen. Für den hier vorgebenen Rahmen ist allerdings das allgemeine nichtlineare Problem zu kompliziert und umfangreich. Wir werden uns deshalb hier nur mit linearen Problemen beschäftigen. Man sollte aber dabei nicht vergessen, daß das Beherrschen linearer Probleme wesentlich für das Beherrschen nichtlinearer Probleme ist, löst man doch solche meist durch sukzessives Linearisieren (Newton-Verfahren).

Abbildung 1.2 Ein mechanisches System



Gegenstand der folgenden Betrachtungen sind also lineare differentiell-algebraische Gleichungen mit variablen Koeffizienten in der Form

$$E(t)\dot{x} = A(t)x + f(t), \tag{1.2}$$

wobei t in einem (offenen) Intervall $\mathbb{I} \subseteq \mathbb{R}$ variiert und

$$E, A \in C(\mathbb{I}, \mathbb{C}^{n,n}), \quad f \in C(\mathbb{I}, \mathbb{C}^n),$$

eventuell zusammen mit einer Anfangsbedingung

$$x(t_0) = x_0, \tag{1.3}$$

wobei $t_0 \in \mathbb{I}$ und $x_0 \in \mathbb{C}^n$.

Wir müssen nun noch festlegen, was wir unter einer Lösung dieses Problems verstehen wollen. Berücksichtigt man noch eventuell notwendige Einschränkungen an die Anfangsbedingung (so muß in Beispiel 2 die Anfangsbedingung zumindest beinhalten, daß der Massepunkt den richtigen Abstand zum Ursprung hat) und an die Inhomogenität (beim Speziallfall einer linearen Gleichung $Ax = b$ ist eine Bedingung für die Existenz einer Lösung gegeben durch $\operatorname{rang} A = \operatorname{rang}(A, b)$), gelangt man zur folgenden Definition.

Definition 3 1. Eine Funktion $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ heißt Lösung von (1.2), wenn sie die Gleichung (1.2) punktweise erfüllt.

2. Sie heißt Lösung des Anfangswertproblemes (1.2) mit (1.3), wenn sie zusätzlich der Anfangsbedingung (1.3) genügt.

3. Eine Anfangsbedingung (1.3) heißt konsistent (bezüglich E, A und f), wenn das zugehörige Anfangswertproblem mindestens eine Lösung besitzt.

4. Eine Inhomogenität f heißt konsistent (bezüglich E und A), wenn die zugehörige differentiell-algebraische Gleichung mindestens eine Lösung besitzt.

Wir werden hier stets ein Problem als lösbar bezeichnen, wenn es mindestens eine Lösung besitzt. Das hört sich zwar selbstverständlich an, wird aber in der Literatur oft anders gehandhabt.

Mit den obigen Begriffen ergeben sich eine Reihe von Fragen, mit denen wir uns im folgenden beschäftigen werden.

- Unter welchen Bedingungen hat (1.2) eine Lösung?
- Wie sieht in diesem Fall die Lösungsmenge aus?
- Was sind die konsistenten Anfangsbedingungen bzw. Inhomogenitäten?
- Wann gibt es eine eindeutige Lösung?

Neben diesen theoretischen Fragen werden wir uns auch für geeignete numerische Verfahren interessieren. Dabei wollen wir uns nur um solche Diskretisierungen bemühen, die keine zusätzliche Struktur des Ausgangsproblems verlangen.

Kapitel 2

Konstante Koeffizienten

Wir nehmen zunächst vereinfachend an, daß die Matrixfunktionen E und A in (1.2) zeitlich konstant sind. Statt (1.2) schreibt man dann

$$E\dot{x} = Ax + f(t) \tag{2.1}$$

mit $E, A \in \mathbb{C}^{n,n}$ und nennt (2.1) eine lineare differentiell-algebraische Gleichung mit konstanten Koeffizienten. Die Eigenschaften der Gleichung (2.1) sind schon seit dem letzten Jahrhundert wohlverstanden, im wesentlichen durch Arbeiten von Weierstraß und Kronecker. Das liegt vor allem daran, daß man an (2.1) mit algebraischen Standardmethoden herangehen kann. Diese Vorgehensweise soll im folgenden zumindest in wesentlichen Zügen nachvollzogen werden.

Skaliert man (2.1) mit einer nichtsingulären Matrix $P \in \mathbb{C}^{n,n}$ und die Funktion x gemäß $x = Q\overline{x}$ mit einer nichtsingulären Matrix $Q \in \mathbb{C}^{n,n}$, so erhält man mit

$$\overline{E}\dot{\overline{x}} = \overline{A}\overline{x} + \overline{f}(t), \quad \overline{E} = PEQ, \quad \overline{A} = PAQ, \quad \overline{f} = Pf$$

wiederum eine lineare differentiell-algebraische Gleichung mit konstanten Koeffizienten. Dabei vermittelt $x = Q\overline{x}$ eine umkehrbar eindeutige Abbildung zwischen den zugehörigen Lösungsmengen. Das bedeutet, daß man statt (2.1) auch das transformierte Problem betrachten kann, um die uns interessierenden Fragen zu beantworten. Die folgende Definition ist nun naheliegend.

Definition 4 Zwei Paare (E_i, A_i), $E_i, A_i \in \mathbb{C}^{n,n}$, $i = 1, 2$, von Matrizen heißen (stark) äquivalent, wenn es nichtsinguläre Matrizen $P, Q \in \mathbb{C}^{n,n}$ gibt, sodaß

$$E_2 = PE_1Q, \quad A_2 = PA_1Q. \tag{2.2}$$

Man schreibt dann $(E_1, A_1) \sim (E_2, A_2)$.

Entsprechende Definitionen findet man in der Literatur auch unter den Stichworten Matrixbüschel, lineare Matrixfunktionen oder verallgemeinerte Eigenwertprobleme, oft mit der Schreibweise $\lambda E - A$ statt (E, A). Wie in der Definition schon angedeutet, liegt eine Äquivalenzrelation vor.

Lemma 5 *Die in Definition 4 eingeführte Relation ist eine Äquivalenzrelation.*

Beweis:
Es ist zu zeigen, daß die Relation reflexiv, symmetrisch und transitiv ist.

1. Reflexivität: Es ist $(E, A) \sim (E, A)$ vermöge $P = Q = I$.

2. Symmetrie: Aus $(E_1, A_1) \sim (E_2, A_2)$ folgt $E_2 = PE_1Q$ und $A_2 = PA_1Q$ mit nichtsingulären Matrizen P und Q. Damit gilt aber auch $E_1 = P^{-1}E_2Q^{-1}$, $A_1 = P^{-1}A_2Q^{-1}$ und es folgt $(E_2, A_2) \sim (E_1, A_1)$.

3. Transitivität: Aus $(E_1, A_1) \sim (E_2, A_2)$ und $(E_2, A_2) \sim (E_3, A_3)$ folgt $E_2 = P_1E_1Q_1$, $A_2 = P_1A_1Q_1$ sowie $E_3 = P_2E_2Q_2$, $A_3 = P_2A_2Q_2$ mit nichtsingulären Matrizen P_i und Q_i, $i = 1, 2$. Einsetzen liefert $E_3 = P_2P_1E_1Q_1Q_2$, $A_3 = P_2P_1A_1Q_1Q_2$ und es gilt $(E_1, A_1) \sim (E_3, A_3)$.

□

Damit können wir die übliche Frage nach einer Normalform stellen, d. h. nach einem zu gegebenem (E, A) äquivalenten Matrixpaar von möglichst einfacher Gestalt, an dem man dann hoffentlich Eigenschaften der zugehörigen differentiell-algebraischen Gleichung ablesen kann. Die sich ergebende sogenannte Kronecker-Normalform ist recht schwierig herzuleiten, siehe etwa [5]. Wir werden uns deshalb auf einen Spezialfall beschränken und das allgemeine Resultat hier ohne Beweis angegeben und auch darauf verzichten, Ergebnisse daraus abzuleiten, zumal wir im nächsten Kapitel die wesentlichen Aspekte der Kronecker-Normalform für (2.1) bei der Behandlung variabler Koeffizienten wiederfinden.

Satz 6 *Seien $E, A \in \mathbb{C}^{n,n}$. Dann existieren nichtsinguläre Matrizen $P, Q \in \mathbb{C}^{n,n}$, sodaß*

$$P(\lambda E - A)Q = \operatorname{diag}(\boldsymbol{L}_{\epsilon_1}, \ldots, \boldsymbol{L}_{\epsilon_u}, \boldsymbol{M}_{\eta_1}, \ldots, \boldsymbol{M}_{\eta_u}, \boldsymbol{J}_{\rho_1}, \ldots, \boldsymbol{J}_{\rho_v}, \boldsymbol{N}_{\sigma_1}, \ldots, \boldsymbol{N}_{\sigma_w}), \tag{2.3}$$

wobei gilt:

1. *Es ist $\boldsymbol{L}_{\epsilon_j}$ ein $(\epsilon_j, \epsilon_j + 1)$-Bidiagonalblock, $\epsilon_j \in \mathbb{N}_0$, der Form*

$$\lambda \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

2. *Es ist $\boldsymbol{M}_{\eta_j}$ ein $(\eta_j + 1, \eta_j)$-Bidiagonalblock, $\eta_j \in \mathbb{N}_0$, der Form*

$$\lambda \begin{bmatrix} 1 & & \\ 0 & \ddots & \\ & \ddots & 1 \\ & & 0 \end{bmatrix} - \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{bmatrix}.$$

3. *Es ist $\boldsymbol{J}_{\rho_j}$ ein (ρ_j, ρ_j)-Jordanblock, $\rho_j \in \mathbb{N}$, $\lambda_j \in \mathbb{C}$, der Form*

$$\lambda \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{bmatrix}.$$

4. Es ist $\boldsymbol{N}_{\sigma_j}$ ein (σ_j, σ_j)-nilpotenter Block, $\sigma_j \in \mathbb{N}$, der Form

$$\lambda \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} - \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

Dabei sind Art, Größe und Anzahl der Blöcke charakteristisch für das Matrixpaar (E, A).

Mit Hilfe der Kronecker-Normalform kann man nun direkt das Lösungsverhalten von (2.1) studieren, indem man einzelne Blöcke betrachtet. Dabei muß man die zwei verschiedenen Arten von Bidiagonalblöcken immer paarweise zusammenfassen. Man beachte, daß beide Arten tatsächlich gleich oft vorkommen, da E und A quadratisch sind. Man kann zwar auch nichtquadratische Systeme betrachten, aber das würde den Rahmen dieser Arbeit sprengen.

Ausgangspunkt für unsere Betrachtungen ist der Begriff eines regulären Matrixpaares.

Definition 7 Seien $E, A \in \mathbb{C}^{n,n}$. Dann heißt das Polynom p definiert durch

$$p(\lambda) = \det(\lambda E - A) \tag{2.4}$$

charakteristisches Polynom von (E, A). Ist p das Nullpolynom, so heißt das Paar (E, A) singulär, im anderen Fall regulär.

Lemma 8 *Jedes zu einem regulären Matrixpaar äquivalente Matrixpaar ist selbst regulär.*

Beweis:
Sei $E_2 = PE_1Q$, $A_2 = PA_1Q$ mit nichtsingulärem P und Q. Dann gilt

$$\begin{aligned} p_2(\lambda) &= \det(\lambda E_2 - A_2) = \det(\lambda PE_1Q - PA_1Q) = \\ &= \det P \det(\lambda E_1 - A_1) \det Q = c\, p_1(\lambda) \end{aligned}$$

mit $c \neq 0$. □

Die Regularität eines Matrixpaares hat direkte Auswirkung auf das Lösungsverhalten der zugehörigen differentiell-algebraischen Gleichung. Man kann nämlich sofort zeigen, daß Regularität notwendig ist für die eindeutige Lösbarkeit des Anfangswertproblems. Mit dem für lineare Probleme bekannten Prinzip, daß sich zwei Lösungen des inhomogenen Problems um eine Lösung des homogenen Problems unterscheiden, ist dies äquivalent zu folgender Aussage.

Lemma 9 *Bei singulärem Matrixpaar besitzt das zugehörige homogene Anfangswertproblem eine nichttriviale Lösung.*

Beweis:
Das Matrixpaar (E, A) sei singulär. Dann ist wegen $p \equiv 0$ die Matrix $\lambda E - A$ singulär für jedes $\lambda \in \mathbb{C}$. Seien λ_i, $i = 1, \ldots, n+1$, paarweise verschiedene

komplexe Zahlen. Dann gibt es zu λ_i ein $v_i \in \mathbb{C}^n \setminus \{0\}$ mit $(\lambda_i E - A)v_i = 0$. Die Vektoren v_i, $i = 1, \ldots, n+1$, sind in jedem Fall linear abhängig. Also gibt es komplexe Zahlen α_i, $i = 1, \ldots, n+1$, die nicht alle verschwinden, mit

$$\sum_{i=1}^{n+1} \alpha_i v_i = 0.$$

Für die durch

$$x(t) = \sum_{i=1}^{n+1} \alpha_i v_i e^{\lambda_i (t-t_0)}$$

definierte Funktion x gilt neben $x(t_0) = 0$ auch

$$E\dot{x}(t) = \sum_{i=1}^{n+1} \alpha_i \lambda_i E v_i e^{\lambda_i (t-t_0)} = \sum_{i=1}^{n+1} \alpha_i A v_i e^{\lambda_i (t-t_0)} = Ax(t).$$

Da x sicher nicht die Nullfunktion ist, ist x eine nichttriviale Lösung des homogenen Anfangswertproblems

$$E\dot{x} = Ax, \quad x(t_0) = 0.$$

□

Es bleibt jetzt die Frage, ob umgekehrt Regularität eindeutige Lösbarkeit des Anfangswertproblems impliziert. Dazu kehren wir zum Problem des Auffindens einer Normalform zumindest für reguläre Matrixpaare zurück.

Satz 10 *Seien $E, A \in \mathbb{C}^{n,n}$ und (E, A) regulär. Dann gilt*

$$(E, A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \right), \tag{2.5}$$

wobei J eine Matrix in Jordan'scher Normalform ist und N eine nilpotente Matrix ebenfalls in Jordan'scher Normalform ist. Dabei ist erlaubt, daß der eine oder andere Block in (2.5) nicht vorhanden ist.

Beweis:
Da (E, A) regulär ist, existiert ein $\lambda_0 \in \mathbb{C}$ mit $\det(\lambda_0 E - A) \neq 0$ bzw. $\lambda_0 E - A$ nichtsingulär. Dann gilt

$$\begin{aligned}(E, A) &\sim (E, A - \lambda_0 E + \lambda_0 E) \sim \\ &\sim ((A - \lambda_0 E)^{-1} E, I + \lambda_0 (A - \lambda_0 E)^{-1} E).\end{aligned}$$

Man transformiert nun $(A - \lambda_0 E)^{-1} E$ auf Jordan'sche Normalform. Diese sei gegeben durch $\mathrm{diag}(\overline{J}, \overline{N})$, wobei $\overline{J}$ nichtsingulär ist (also der zu den nichtverschwindenden Eigenwerten gehörige Teil sein soll) und somit $\overline{N}$ eine nilpotente, strikte (obere) Dreiecksmatrix ist. Damit haben wir

$$(E, A) \sim \left(\begin{bmatrix} \overline{J} & 0 \\ 0 & \overline{N} \end{bmatrix}, \begin{bmatrix} I + \lambda_0 \overline{J} & 0 \\ 0 & I + \lambda_0 \overline{N} \end{bmatrix} \right).$$

Wegen der speziellen Gestalt von $\overline{N}$ ist $I + \lambda_0 \overline{N}$ eine nichtsinguläre (obere) Dreiecksmatrix. Somit ergibt sich

$$(E, A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & (I + \lambda_0 \overline{N})^{-1} \overline{N} \end{bmatrix}, \begin{bmatrix} \overline{J}^{-1} + \lambda_0 I & 0 \\ 0 & I \end{bmatrix} \right),$$

wobei $(I + \lambda_0 \overline{N})^{-1} \overline{N}$ wieder strenge (obere) Dreiecksmatrix und demnach nilpotent ist. Transformation der nichttrivialen Einträge auf Jordan'sche Normalform liefert schließlich (2.5) mit der geforderten Blockstruktur. □

Mit dieser auch manchmal nach Weierstraß benannten Normalform kann man nun die Lösungen von (2.1) mit regulärem Matrizenpaar explizit angeben. Dazu nutzt man aus, daß (2.1) in zwei Teilprobleme separiert, wenn (E, A) in Normalform vorliegt. Nennt man jeweils wieder die unbekannte Funktion x und die Inhomogenität f, so stellt der erste Teil mit

$$\dot{x} = Jx + f(t) \tag{2.6}$$

eine gewöhnliche Differentialgleichung dar, während der zweite Teil die Gestalt

$$N\dot{x} = x + f(t) \tag{2.7}$$

besitzt. Da Anfangswertprobleme bei linearen gewöhnlichen Differentialgleichungen (2.6) bekanntermaßen stets eindeutig lösbar sind (siehe etwa [7]), wenden wir uns (2.7) zu.

Lemma 11 *Die differentiell-algebraische Gleichung* (2.7) *mit* $f \in C^\nu(\mathbb{I}, \mathbb{C}^n)$ *besitzt die eindeutige Lösung*

$$x = -\sum_{i=0}^{\nu-1} N^i f^{(i)}, \tag{2.8}$$

wobei ν *durch die Bedingungen* $N^\nu = 0$ *und* $N^{\nu-1} \neq 0$ *gegeben ist.*

Beweis:
Daß eine Lösung von (2.7) die Gestalt (2.8) besitzen muß, kann man dadurch zeigen, daß man die Gleichung (2.7) weiter aufteilt in die einzelnen nilpotenten Jordanblöcke und dann komponentenweise löst. Schneller führt folgendes formale Vorgehen zum Ziel. Sei D der Operator, der einer differenzierbaren Funktion seine Ableitung zuordnet. Dann hat (2.7) die Gestalt $NDx = x + f$ oder $(I - ND)x + f = 0$. Da N nilpotent ist und N und D vertauschbar sind (N ist ein konstanter Faktor), erhält man mit Hilfe der Neumann'schen Reihe

$$x = -(I - ND)^{-1} f = -\sum_{i=0}^{\infty} (ND)^i f = -\sum_{i=0}^{\nu-1} N^i f^{(i)}.$$

Daß (2.8) tatsächlich eine Lösung von (2.7) ist, verifiziert man durch Einsetzen

gemäß

$$N\dot{x} - x - f = -\sum_{i=0}^{\nu-1} N^{i+1} f^{(i+1)} + \sum_{i=0}^{\nu-1} N^i f^{(i)} - f = 0.$$

□

Bei der Lösung (2.8) von (2.7) fallen zwei Dinge auf. Zunächst ist die Lösung ohne Vorgabe eines Anfangswertes eindeutig (oder der vorgegebene Anfangswert muß (2.8) an der Stelle t_0 erfüllen). Außerdem muß man fordern, daß f mindestens ν-mal stetig differenzierbar ist, um zu garantieren, daß x stetig differenzierbar ist. In diesem Sinn ist die Größe ν ein wichtiges Merkmal einer linearen differentiell-algebraischen Gleichung mit konstanten Koeffizienten.

Definition 12 Das Matrixpaar (E, A) sei regulär mit einer Normalform gemäß (2.5). Dann heißt die Größe ν defininert durch $N^\nu = 0$, $N^{\nu-1} \neq 0$ bzw. $\nu = 0$, falls kein nilpotenter Block vorkommt, der Index von (E, A) und man schreibt $\nu = \mathrm{ind}(E, A)$.

Um diese Definition und die Schreibweise $\nu = \mathrm{ind}(E, A)$ zu rechtfertigen, müssen wir noch zeigen, daß ν unabhängig von der speziellen Transformation auf Normalform ist.

Lemma 13 *Seien zum regulären Matrixpaar (E, A) durch*

$$(E, A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & N_i \end{bmatrix}, \begin{bmatrix} J_i & 0 \\ 0 & I \end{bmatrix} \right), \quad i = 1, 2,$$

zwei Normalformen gegeben, wobei die Größe des ersten Blockes d_1 bzw. d_2 sei. Dann gilt $d_1 = d_2$ und $N_1^\nu = 0$, $N_1^{\nu-1} \neq 0$ ist äquivalent zu $N_2^\nu = 0$, $N_2^{\nu-1} \neq 0$.

Beweis:
Zunächst gilt für das charakteristische Polynom der beiden Normalformen

$$p_i(\lambda) = \det \begin{bmatrix} \lambda I - J_i & 0 \\ 0 & \lambda N_i - I \end{bmatrix} = (-1)^{n-d_i} \det(\lambda I - J_i),$$

d. h. p_i ist ein Polynom vom Grad d_i. Da die Normalformen aber äquivalent sind und sich nach dem Beweis von Lemma 8 die zugehörigen charakteristischen Polynome nur um einen nichtverschwindenden Faktor unterscheiden können, folgt $d_1 = d_2$ und die Blockgrößen sind gleich. Aus der Äquivalenz der Normalformen folgt nun

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & N_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N_1 \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

und

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} J_2 & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} J_1 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

mit nichtsingulären Matrizen $(P_{ij})_{i,j=1,2}$ und $(Q_{ij})_{i,j=1,2}$. Durch Ausmultiplizieren ergeben sich die Beziehungen

$$P_{11} = Q_{11}, \quad P_{12}N_2 = Q_{12}, \quad P_{21} = N_1 Q_{21}, \quad P_{22}N_2 = N_1 Q_{22},$$

respektive

$$P_{11}J_2 = J_1Q_{11}, \quad P_{12} = J_1Q_{12}, \quad P_{21}J_2 = Q_{21}, \quad P_{22} = Q_{22}.$$

Damit folgt $P_{21} = N_1P_{21}J_2$ und durch sukzessives Einsetzen von P_{21} wegen der Nilpotenz von N_1 schließlich $P_{21} = 0$. Also müssen $P_{11} = Q_{11}$ und $P_{22} = Q_{22}$ nichtsingulär sein. Insbesondere sind dann J_1 und J_2 wie auch N_1 und N_2 jeweils zueinander ähnlich. Die Behauptung folgt nun aus der Theorie der Jordan'schen Normalform. □

Damit folgt, daß sowohl die Blockgrößen in der Normalform als auch der oben definierte Index charakteristisch für ein Matrixpaar bzw. für eine lineare differentiell-algebraische Gleichung mit konstanten Koeffizienten sind.

Wir können nun die erzielten Ergebnisse im Hinblick auf die anfangs gestellten Fragen wie folgt zusammenfassen.

Satz 14 *Sei (E, A) ein reguläres Matrixpaar und durch nichtsinguläre Matrizen P und Q gemäß*

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad PAQ = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}, \quad Pf = \begin{bmatrix} \overline{f}_1 \\ \overline{f}_2 \end{bmatrix} \tag{2.9}$$

und

$$Q^{-1}x = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \end{bmatrix}, \quad Q^{-1}x_0 = \begin{bmatrix} \overline{x}_{10} \\ \overline{x}_{20} \end{bmatrix} \tag{2.10}$$

eine Transformation von (2.1) *und* (1.3) *auf Normalform gegeben. Weiter sei $\nu = \mathrm{ind}(E, A)$ und $f \in C^\nu(\mathbb{I}, \mathbb{C}^n)$. Dann gilt:*

1. *Die differentiell-algebraische Gleichung* (2.1) *ist lösbar.*

2. *Eine Anfangsbedingung* (1.3) *ist genau dann konsistent, wenn*

$$\overline{x}_{20} = -\sum_{i=0}^{\nu-1} N^i \overline{f}_2^{(i)}(t_0).$$

3. *Jedes Anfangswertproblem mit konsistenter Anfangsbedingung ist eindeutig lösbar.*

Kapitel 3

Variable Koeffizienten

Betrachtet man nun den allgemeinen linearen Fall (1.2), so liegt die Idee nahe, in Anlehnung an Satz 14 Regularität des Matrixpaares $(E(t), A(t))$ für alle $t \in \mathbb{I}$ zu verlangen, um eindeutige Lösbarkeit des Anfangswertproblems bei konsistenter Anfangsbedingung zu erreichen. Leider stellt sich schnell heraus, daß diese beiden Eigenschaften einer linearen differentiell-algebraischen Gleichung voneinander unabhängig sind.

Beispiel 15 Gegeben seien E, A und f durch

$$E(t) = \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad f(t) = 0, \quad \mathbb{I} = (-1, 1).$$

Wegen

$$\det(\lambda E(t) - A(t)) = 1 - \lambda^2 t^2 + \lambda^2 t^2 = 1$$

ist $(E(t), A(t))$ regulär für alle $t \in \mathbb{I}$, aber man rechnet schnell nach, daß x gegeben durch

$$x(t) = c(t) \begin{bmatrix} t \\ 1 \end{bmatrix}$$

für jedes $c \in C^1(\mathbb{I}, \mathbb{C})$ mit $c(t_0) = 0$ das zugehörige homogene Anfangswertproblem löst. Insbesondere gibt es also mehr als eine Lösung.

Beispiel 16 Gegeben seien E, A und f durch

$$E(t) = \begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \quad \mathbb{I} = \mathbb{R}$$

mit $f \in C^2(\mathbb{I}, \mathbb{C}^2)$. Wegen

$$\det(\lambda E(t) - A(t)) = -\lambda t + \lambda t = 0$$

ist $(E(t), A(t))$ singulär für alle $t \in \mathbb{I}$. Mit $x = (x_1, x_2)$ schreibt sich die zugehörige differentiell-algebraischen Gleichung als

$$0 = -x_1(t) + t x_2(t) + f_1(t), \quad \dot{x}_1(t) - t\dot{x}_2(t) = f_2(t).$$

Auflösen der ersten Gleichung liefert $x_1(t) = tx_2(t) + f_1(t)$. Leitet man diese Beziehung ab und setzt in die zweite Gleichung ein, so findet man schließlich als einzige Lösung

$$x_1(t) = tf_2(t) - t\dot{f}_1(t) + f_1(t), \quad x_2(t) = f_2(t) - \dot{f}_1(t).$$

Hier findet man also, daß jedes Anfangsproblem mit konsistenter Anfangsbedingung eindeutig lösbar ist.

Der Grund für dieses Verhalten ist darin zu finden, daß die Äquivalenzrelation (2.2) für lineare differentiell-algebraische Gleichungen mit variablen Koeffizienten nicht adäquat ist. Vielmehr muß man nun umkehrbare zeitabhängige Transformationen zulassen, die (1.2) in eine Gleichung gleicher Form überführen. Sind P und Q Matrixfunktionen, die punktweise nichtsingulär sind, so kann man wie im Fall konstanter Koeffizienten die Gleichung durch Multiplikation von links skalieren. Die durch $x = Q\overline{x}$ gegebene Skalierung der gesuchten Funktion muß zur Transformation der Gleichung abgeleitet werden. Wegen $\dot{x} = Q\dot{\overline{x}} + \dot{Q}\overline{x}$ tritt aber durch die Produktregel ein zusätzlicher Term auf. Man muß also bei (1.2) eine andere Definition von Äquivalenz betrachten.

Definition 17 Zwei Paare (E_i, A_i), $E_i, A_i \in C(\mathbb{I}, \mathbb{C}^{n,n})$, $i = 1, 2$, von Matrixfunktionen heißen (global) äquivalent, wenn es punktweise nichtsinguläre Matrixfunktionen $P \in C(\mathbb{I}, \mathbb{C}^{n,n})$, $Q \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ gibt, sodaß

$$E_2 = PE_1Q, \quad A_2 = PA_1Q - PE_1\dot{Q} \tag{3.1}$$

als Gleichheit von Funktionen. Man schreibt wiederum $(E_1, A_1) \sim (E_2, A_2)$.

Lemma 18 *Die in Definition* 17 *eingeführte Relation ist eine Äquivalenzrelation.*

Beweis:
Wir zeigen die drei bekannten Eigenschaften.

1. Reflexivität: Es ist $(E, A) \sim (E, A)$ vermöge $P = Q = I$.

2. Symmetrie: Aus $(E_1, A_1) \sim (E_2, A_2)$ folgt $E_2 = PE_1Q$ und $A_2 = PA_1Q - PE_1\dot{Q}$ mit punktweise nichtsingulären Matrixfunktionen P und Q. Damit gilt aber auch bei punktweiser Definition der Inversen $E_1 = P^{-1}E_2Q^{-1}$, $A_1 = P^{-1}A_2Q^{-1} + P^{-1}E_2Q^{-1}\dot{Q}Q^{-1}$ und es folgt $(E_2, A_2) \sim (E_1, A_1)$ wegen $d/dt\, Q^{-1} = -Q^{-1}\dot{Q}Q^{-1}$.

3. Transitivität: Aus $(E_1, A_1) \sim (E_2, A_2)$ und $(E_2, A_2) \sim (E_3, A_3)$ folgt $E_2 = P_1E_1Q_1$, $A_2 = P_1A_1Q_1 - P_1E_1\dot{Q}_1$ sowie $E_3 = P_2E_2Q_2$, $A_3 = P_2A_2Q_2 - P_2E_2\dot{Q}_2$ mit punktweise nichtsingulären Matrixfunktionen P_i und Q_i, $i = 1, 2$. Einsetzen liefert $E_3 = P_2P_1E_1Q_1Q_2$, $A_3 = P_2P_1A_1Q_1Q_2 - P_2P_1E_1(\dot{Q}_1Q_2 + Q_1\dot{Q}_2)$ und es gilt $(E_1, A_1) \sim (E_3, A_3)$.

□

Die Regularität des Matrixpaares $(E(t), A(t))$ für festes t ist bezüglich dieser Äquivalenzrelation keine Invariante mehr. Es erhebt sich also die Frage, was nun

unter dieser neuen Äquivalenzrelation die wesentlichen Invarianten sind und was eine mögliche Normalform ist. Dies für beliebige Matrixfunktionen E und A, die ja im allgemeinen nichtlinear von t abhängen, zu untersuchen, gestaltet sich schwierig. So hat man z. B. bei der einfachen skalaren Gleichung $0 = tx + f(t)$ sofort für die Existenz einer Lösung die notwendige Bedingung $f(0) = 0$. Um solche und ähnliche Effekte auszuschließen, werden wir zusätzliche Bedingungen an die Funktionen E und A stellen. Wie diese zu wählen sind, wollen wir uns im folgenden herleiten, indem wir zunächst fragen, was die Äquivalenzrelation (3.1) für einen festen Punkt $t \in \mathbb{I}$ bedeutet. Beachtet man, daß man zu gegebenem $P(t)$, $Q(t)$ und $\dot{Q}(t)$ stets geeignete Funktionen P und Q finden kann, die die vorgegebenen Werte annehmen, so gelangt man zur folgenden lokalen Version der obigen Äquivalenzrelation, siehe auch [9].

Definition 19 Zwei Paare (E_i, A_i), $E_i, A_i \in \mathbb{C}^{n,n}$, $i = 1, 2$, von Matrizen heißen (lokal) äquivalent, wenn es Matrizen $P, Q, B \in \mathbb{C}^{n,n}$, P, Q nichtsingulär, gibt, sodaß

$$E_2 = PE_1Q, \quad A_2 = PA_1Q - PE_1B. \tag{3.2}$$

Man schreibt ebenfalls $(E_1, A_1) \sim (E_2, A_2)$ und unterscheidet von der zuvor definierten Äquivalenz nach der Art (Matrix oder Matrixfunktion) der Paare.

Lemma 20 *Die in Definition* 19 *eingeführte Relation ist eine Äquivalenzrelation.*

Beweis:
Der Beweis kann analog zum Beweis von Lemma 18 geführt werden. □

Man beachte, daß man für $B = 0$ als Spezialfall die Transformation (2.2) erhält. Damit stehen hier mehr Möglichkeiten zur Verfügung, ein gegebenes Matrixpaar zu vereinfachen. Man darf also eine einfachere Normalform als die Kronecker-Normalform erwarten. Im folgenden benutzen wir zur Vereinfachung die Sprechweise, daß eine Matrix eine Basis eines Vektorraums ist, wenn dies für ihre Spalten zutrifft, und zur Vermeidung von Fallunterscheidungen die Konvention, daß die einzige Basis des Untervektorraumes $\{0\}$ des $\mathbb{C}^n$ die leere Matrix $\emptyset \in \mathbb{C}^{n,0}$ mit $\operatorname{rang} \emptyset = 0$ ist. Außerdem bezeichne der Superskript * den Übergang zur komplex-konjugierten Transponierten.

Satz 21 *Seien $E, A \in \mathbb{C}^{n,n}$ und*

$$\begin{array}{lll} \text{(a)} & T & \text{Basis von } \operatorname{kern} E \\ \text{(b)} & Z & \text{Basis von } \operatorname{cobild} E = \operatorname{kern} E^* \\ \text{(c)} & T' & \text{Basis von } \operatorname{cokern} E = \operatorname{bild} E^* \\ \text{(d)} & V & \text{Basis von } \operatorname{cobild}(Z^*AT). \end{array} \tag{3.3}$$

Dann sind die Größen

$$\begin{array}{lll} \text{(a)} & r = \operatorname{rang} E & \text{(Rang)} \\ \text{(b)} & a = \operatorname{rang}(Z^*AT) & \text{(Algebraischer Teil)} \\ \text{(c)} & s = \operatorname{rang}(V^*Z^*AT') & \text{(Strangeness)} \\ \text{(d)} & d = r - s & \text{(Differentieller Teil)} \\ \text{(e)} & u = n - r - a - s & \text{(Unbestimmter Teil)} \end{array} \tag{3.4}$$

invariant unter (3.2) *und* (E, A) *ist (lokal) äquivalent zu der Normalform*

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 & 0 \\ I_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} . \tag{3.5}$$

Beweis:
Seien (E_i, A_i), $i = 1, 2$, äquivalent. Wegen

$$\operatorname{rang} E_2 = \operatorname{rang}(PE_1Q) = \operatorname{rang} E_1$$

ist r invariant. Für die Größen a und s ist zunächst zu zeigen, daß sie nicht von einer speziellen Wahl der Basen abhängen. Jeder Basiswechsel läßt sich darstellen durch

$$\tilde{T} = TM_T, \quad \tilde{Z} = ZM_Z, \quad \tilde{T}' = T'M_{T'}, \quad \tilde{V} = M_Z^{-1} V M_V$$

mit nichtsingulären Matrizen $M_T, M_Z, M_{T'}, M_V$ und die Wohldefiniertheit von a und s folgt aus

$$\operatorname{rang}(\tilde{Z}^* A \tilde{T}) = \operatorname{rang}(M_Z^* Z^* A T M_T) = \operatorname{rang}(Z^* A T)$$

und

$$\operatorname{rang}(\tilde{V}^* \tilde{Z}^* A \tilde{T}') = \operatorname{rang}(M_V^* V^* M_Z^{*-1} M_Z^* Z^* A T' M_{T'}) = \operatorname{rang}(V^* Z^* A T').$$

Seien nun zu (E_2, A_2) die Basen T_2, Z_2, T_2', V_2 gegeben, d. h.

$$\begin{array}{lll} \operatorname{rang}(E_2 T_2) = 0, & T_2^* T_2 \text{ nichtsingulär}, & \operatorname{rang}(T_2^* T_2) = n - r, \\ \operatorname{rang}(Z_2^* E_2) = 0, & Z_2^* Z_2 \text{ nichtsingulär}, & \operatorname{rang}(Z_2^* Z_2) = n - r, \\ \operatorname{rang}(E_2 T_2') = r, & T_2'^* T_2' \text{ nichtsingulär}, & \operatorname{rang}(T_2'^* T_2') = r, \\ \operatorname{rang}(V_2^* Z_2^* A_2 T_2) = 0, & V_2^* V_2 \text{ nichtsingulär}, & \operatorname{rang}(V_2^* V_2) = \hat{a}_2 \end{array}$$

mit $\hat{a}_2 = \dim \operatorname{cobild}(Z_2^* A_2 T_2)$. Setzt man die Äquivalenzrelation (3.2) ein und definiert

$$T_1 = QT_2, \quad Z_1^* = Z_2^* P, \quad T_1' = QT_2', \quad V_1^* = V_2^*,$$

so erhält man die gleichen Beziehungen für (E_1, A_1) mit den Matrizen T_1, Z_1, T_1' und V_1. Also sind T_1, Z_1, T_1' Basen entsprechend (3.3). Wegen

$$\begin{aligned} \hat{a}_2 &= \dim \operatorname{cobild}(Z_2^* A_2 T_2) = \\ &= \dim \operatorname{cobild}(Z_2^* P A_1 Q T_2 - Z_2^* P E_1 B T_2) = \\ &= \dim \operatorname{cobild}(Z_1^* A_1 T_1) = \hat{a}_1, \end{aligned}$$

wobei $Z_1^* E_1 = 0$ benutzt wurde, trifft dies auch auf V_1 zu. Mit

$$\operatorname{rang}(Z_2^* A_2 T_2) = \operatorname{rang}(Z_2^* P A_1 Q T_2 - Z_2^* P E_1 B T_2) = \operatorname{rang}(Z_1^* A_1 T_1)$$

und

$$\operatorname{rang}(V_2^* Z_2^* A_2 T_2') = \operatorname{rang}(V_2^* Z_2^* P A_1 Q T_2' - V_2^* Z_2^* P E_1 B T_2') = \operatorname{rang}(V_1^* Z_1^* A_1 T_1')$$

folgt die Invarianz von a und s und damit auch die von d und u.

Zur Herleitung der Normalform (3.5) sei zunächst Z' eine Basis von bild E und V' eine Basis von bild(Z^*AT). Die Matrizen $(T', T), (Z', Z), (V', V)$ sind dann nichtsingulär. Außerdem sind $Z'^* E T'$ und entsprechend konstruierte Matrizen ebenfalls nichtsingulär. Wir erhalten damit

$$\begin{aligned}
(E, A) &\sim \left(\begin{bmatrix} Z'^* E T' & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} Z'^* A T' & Z'^* A T \\ Z^* A T' & Z^* A T \end{bmatrix} \right) \sim \\
&\sim \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ Z^* A T' & Z^* A T \end{bmatrix} \right) \sim \\
&\sim \left(\begin{bmatrix} I_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ V'^* Z^* A T' & I_a & 0 \\ V^* Z^* A T' & 0 & 0 \end{bmatrix} \right) \sim \\
&\sim \left(\begin{bmatrix} I_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_a & 0 \\ V^* Z^* A T' & 0 & 0 \end{bmatrix} \right)
\end{aligned}$$

und durch eine entsprechende Transformation von $V^* Z^* A T'$ schließlich (3.5). □

Geht man nun über zu Paaren (E, A) von Matrixfunktionen, so kann man für jedes $t \in \mathbb{I}$ dem Matrixpaar $(E(t), A(t))$ die entsprechenden charakteristischen Werte (3.4) zuordnen. Es ergeben sich so Funktionen $r, a, s : \mathbb{I} \to \mathbb{N}_0$. Eine sinnvolle Annahme scheint jetzt zu sein, daß diese Funktionen konstant sind, d. h. daß die entsprechenden Blockgrößen in (3.5) nicht von t abhängig sind. Dies erlaubt dann die Anwendung der folgenden Eigenschaft einer Matrixfunktion mit konstantem Rang, siehe z. B. [11].

Satz 22 *Sei $E \in C^\ell(\mathbb{I}, \mathbb{C}^{m,n})$, $\ell \in \mathbb{N}_0$, mit $\operatorname{rang} E(t) = r$ für alle $t \in \mathbb{I}$. Dann gibt es punktweise unitäre (insbesondere also nichtsinguläre) Funktionen $U \in C^\ell(\mathbb{I}, \mathbb{C}^{m,m})$ und $V \in C^\ell(\mathbb{I}, \mathbb{C}^{n,n})$ derart, daß*

$$U^* E V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \tag{3.6}$$

mit $\Sigma \in C^\ell(\mathbb{I}, \mathbb{C}^{r,r})$ punktweise nichtsingulär.

Es ergibt sich nun die folgende globale Normalform, die erwartungsgemäß, da man jetzt die Kopplung $B = \dot{Q}$ berücksichtigen muß, etwas komplizierter als die lokale Normalform ausfällt.

Satz 23 *Sei (E, A) ein Paar von hinreichend glatten Matrixfunktionen und gelte*

$$r(t) \equiv r, \quad a(t) \equiv a, \quad s(t) \equiv s. \tag{3.7}$$

Dann ist (E, A) global äquivalent zu der Normalform

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & 0 & 0 & A_{24} & A_{25} \\ 0 & 0 & I_a & 0 & 0 \\ I_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} . \tag{3.8}$$

Dabei sind alle Einträge der Form A_{ij} selbst wieder Matrixfunktionen.

Beweis:
Im folgenden soll das Wort "neu" über dem Äquivalenzzeichen bedeuten, daß die folgende Notation an die neue Blockstruktur angepaßt wurde und deshalb das gleiche Symbol einen anderen Eintrag bezeichnen kann. Unter Verwendung von Satz 22 erhalten wir die folgende Kette von äquivalenten Paaren von Matrixfunktionen.

$$\begin{aligned}
(E,A) &\sim \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right) \overset{\text{neu}}{\sim} \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right) \sim \\
&\sim \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12}V_1 \\ U_1^* A_{21} & U_1^* A_{22} V_1 \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \dot{V}_1 \end{bmatrix}\right) \sim \\
&\overset{\text{neu}}{\sim} \left(\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & I & 0 \\ A_{31} & 0 & 0 \end{bmatrix}\right) \sim \\
&\sim \left(\begin{bmatrix} V_2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11}V_2 & A_{12} & A_{13} \\ A_{21}V_2 & I & 0 \\ U_2^* A_{31} V_2 & 0 & 0 \end{bmatrix} - \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) \sim \\
&\overset{\text{neu}}{\sim} \left(\begin{bmatrix} V_{11} & V_{12} & 0 & 0 & 0 \\ V_{21} & V_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ A_{31} & A_{32} & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \sim \\
&\overset{\text{neu}}{\sim} \left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & A_{13} & A_{14} & A_{15} \\ 0 & A_{22} & A_{23} & A_{24} & A_{25} \\ 0 & A_{32} & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \sim \\
&\sim \left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & A_{13} & A_{14} & A_{15} \\ 0 & A_{22} & A_{23} & A_{24} & A_{25} \\ 0 & A_{32} & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & -A_{32} & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}\right) \sim \\
&\overset{\text{neu}}{\sim} \left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & A_{13} & A_{14} & A_{15} \\ 0 & A_{22} & A_{23} & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \sim
\end{aligned}$$

$$\sim \left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & A_{22} & 0 & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \sim$$

$$\overset{\text{neu}}{\sim} \left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & Q_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & A_{22}Q_2 - \dot{Q}_2 & 0 & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \sim$$

$$\overset{\text{neu}}{\sim} \left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & 0 & 0 & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right).$$

Dabei wurde im letzten Schritt Q_2 als Lösung des Anfangswertproblems

$$\dot{Q}_2 = A_{22}Q_2, \quad Q_2(t_0) = I$$

auf $\mathbb{I}$ gewählt. Wegen der eindeutigen Lösbarkeit ist gesichert, daß Q_2 punktweise nichtsingulär ist. □

Wir wollen nun dieses Ergebnis auf unsere beiden Beispiele vom Anfang dieses Kapitels anwenden.

Beispiel 24 Dem Beispiel 15 entnimmt man sofort $r = \operatorname{rang} E = 1$ für ganz $\mathbb{I}$. Mit der Wahl

$$T = \begin{bmatrix} t \\ 1 \end{bmatrix}, \quad T' = \begin{bmatrix} 1 \\ -t \end{bmatrix}, \quad Z = \begin{bmatrix} 1 \\ -t \end{bmatrix}$$

der Basen ergibt sich $a = \operatorname{rang}(Z^*AT) = 0$ mit $V = [\,1\,]$ und damit $s = \operatorname{rang}(V^*Z^*AT') = 1$.

Beispiel 25 Dem Beispiel 16 entnimmt man ebenfalls sofort $r = \operatorname{rang} E = 1$ für ganz $\mathbb{I}$. Mit der Wahl

$$T = \begin{bmatrix} t \\ 1 \end{bmatrix}, \quad T' = \begin{bmatrix} 1 \\ -t \end{bmatrix}, \quad Z = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

der Basen ergibt sich $a = \operatorname{rang}(Z^*AT) = 0$ mit $V = [\,1\,]$ und damit $s = \operatorname{rang}(V^*Z^*AT') = 1$.

Es ergeben sich also bei beiden Paaren von Matrixfunktionen dieselben charakteristischen Größen (r, a, s). Damit ist klar, daß wesentliche Information noch in den Matrixfunktionen A_{ij} von (3.8) enthalten sein muß. Schreibt man die zu (3.8) gehörige differentiell-algebraische Gleichung aus, so erhält man

$$\begin{aligned}
&\text{(a)} \quad \dot{x}_1 = A_{12}(t)x_2 + A_{14}(t)x_4 + A_{15}(t)x_5 + f_1(t), \\
&\text{(b)} \quad \dot{x}_2 = A_{24}(t)x_4 + A_{25}(t)x_5 + f_2(t), \\
&\text{(c)} \quad 0 = x_3 + f_3(t), \\
&\text{(d)} \quad 0 = x_1 + f_4(t), \\
&\text{(e)} \quad 0 = f_5(t).
\end{aligned} \tag{3.9}$$

Man erkennt sofort in (3.9c) eine algebraische Gleichung für x_3 (algebraischer Teil) und in (3.9e) eine Konsistenzbedingung an die Inhomogenität zusammen mit der freien Wahl von x_5 (unbestimmter Teil). Weiter sieht (3.9b) wie eine Differentialgleichung aus (differentieller Teil). Das eigentliche Problem, das anscheinend für differentiell-algebraische Gleichungen typisch ist, ist die Kopplung der algebraischen Gleichung (3.9d) für x_1 an (3.9a), wo die Ableitung $\dot{x}_1$ vorkommt (Strangeness). Um weitergehende Aussagen machen zu können, bleibt also nichts anderes übrig, als (3.9d) abzuleiten und $\dot{x}_1$ in (3.9a) zu eliminieren, wodurch dies zu einer algebraischen Gleichung wird. Man beachte, daß man bei der Herleitung der Lösungen der beiden Beispiele ebenfalls ableiten muß. Dieses Ableiten und Eliminieren entspricht dem Übergang vom Paar (3.8) zum Paar

$$\left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & 0 & 0 & A_{24} & A_{25} \\ 0 & 0 & I_a & 0 & 0 \\ I_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} . \tag{3.10}$$

Da die algebraische Gleichung (3.9d) erhalten bleibt, kann man diesen Vorgang auch wieder umkehren, indem man (3.9d) ableitet und zu der neuen ersten Gleichung addiert. Damit ist auch klar, daß dabei die Lösungsmenge der differentiell-algebraischen Gleichung (3.9) nicht verändert wird.

Es liegt nun nahe, für das so erhaltene Paar (3.10) wieder die zugehörigen charakteristischen Werte (r, a, s) zu bestimmen, anzunehmen, daß diese über das Intervall konstant sind, und auf globale Normalform zu transformieren. Dabei muß aber zunächst geklärt werden, ob diese neuen Werte überhaupt charakteristisch für das urspüngliche System sind. Es kann ja sein, daß man mit zwei verschiedenen Transformationen auf globale Normalform durch den Übergang auf das neue Paar verschiedene charakteristische Werte erhält.

Satz 26 *Seien die Paare (E, A) und $(\tilde{E}, \tilde{A})$ von Matrixfunktionen äquivalent und von der Form* (3.8)*. Dann sind die jeweils durch Übergang von* (3.8) *nach* (3.10) *erhaltenen modifizierten Paare $(E_{\mathrm{mod}}, A_{\mathrm{mod}})$ und $(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}})$ ebenfalls äquivalent.*

Beweis:
Nach Voraussetzung gibt es glatte, punktweise nichtsinguläre Matrixfunktionen P und Q derart, daß

$$P\tilde{E} = EQ, \quad P\tilde{A} = AQ - E\dot{Q}.$$

Aus der ersten Beziehung leitet man ab, daß

$$\begin{bmatrix} P_{11} & P_{12} & 0 & 0 & 0 \\ P_{21} & P_{22} & 0 & 0 & 0 \\ P_{31} & P_{32} & 0 & 0 & 0 \\ P_{41} & P_{42} & 0 & 0 & 0 \\ P_{51} & P_{52} & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} & Q_{15} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} & Q_{25} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

wenn wir P und Q entsprechend zu (3.8) partitionieren. Damit erhalten wir für die letzten drei Blockzeilen der zweiten Beziehung

$$\begin{bmatrix} P_{34} & 0 & P_{33} & 0 & 0 \\ P_{44} & 0 & P_{43} & 0 & 0 \\ P_{54} & 0 & P_{53} & 0 & 0 \end{bmatrix} = \begin{bmatrix} Q_{31} & Q_{32} & Q_{33} & Q_{34} & Q_{35} \\ Q_{11} & Q_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Die Funktionen P und Q besitzten also die Gestalt

$$P = \begin{bmatrix} Q_{11} & 0 & P_{13} & P_{14} & P_{15} \\ Q_{21} & Q_{22} & P_{23} & P_{24} & P_{25} \\ 0 & 0 & Q_{33} & Q_{31} & P_{35} \\ 0 & 0 & 0 & Q_{11} & P_{45} \\ 0 & 0 & 0 & 0 & P_{55} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & 0 & 0 & 0 & 0 \\ Q_{21} & Q_{22} & 0 & 0 & 0 \\ Q_{31} & 0 & Q_{33} & 0 & 0 \\ Q_{41} & Q_{42} & Q_{43} & Q_{44} & Q_{45} \\ Q_{51} & Q_{52} & Q_{53} & Q_{54} & Q_{55} \end{bmatrix}$$

und die Matrixfunktionen

$$Q_{11},\ Q_{22},\ Q_{33},\ P_{55},\ \begin{bmatrix} Q_{44} & Q_{45} \\ Q_{54} & Q_{55} \end{bmatrix}$$

müssen punktweise nichtsingulär sein. Aus den ersten zwei Blockzeilen der zweiten Beziehung erhält man nun

$$\begin{bmatrix} Q_{11} & 0 \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} \tilde{A}_{12} & \tilde{A}_{14} & \tilde{A}_{15} \\ 0 & \tilde{A}_{24} & \tilde{A}_{25} \end{bmatrix} =$$
$$= \begin{bmatrix} A_{12} & A_{14} & A_{15} \\ 0 & A_{24} & A_{25} \end{bmatrix} \begin{bmatrix} Q_{22} & 0 & 0 \\ Q_{42} & Q_{44} & Q_{45} \\ Q_{52} & Q_{54} & Q_{55} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ \dot{Q}_{22} & 0 & 0 \end{bmatrix}.$$

Daraus folgt, daß $(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}})$ global äquivalent ist zu

$$\left(\begin{bmatrix} Q_{11} & 0 & & & \\ Q_{21} & Q_{22} & & & \\ & & I & & \\ & & & I & \\ & & & & I \end{bmatrix} \begin{bmatrix} 0 & & & & \\ & I & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} Q_{11} & 0 & & & \\ Q_{21} & Q_{22} & & & \\ & & I & & \\ & & & I & \\ & & & & I \end{bmatrix} \begin{bmatrix} 0 & \tilde{A}_{12} & 0 & \tilde{A}_{14} & \tilde{A}_{15} \\ 0 & 0 & 0 & \tilde{A}_{24} & \tilde{A}_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \sim$$
$$\sim \left(\begin{bmatrix} 0 & & & & \\ & Q_{22} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & 0 & 0 & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \right.$$
$$\left. \cdot \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & Q_{22} & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & Q_{42} & 0 & Q_{44} & Q_{45} \\ 0 & Q_{52} & 0 & Q_{54} & Q_{55} \end{bmatrix} - \begin{bmatrix} 0 & & & & \\ & \dot{Q}_{22} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \right) \sim$$

$$\sim \left(\begin{bmatrix} 0 & & & & \\ & Q_{22} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & Q_{22}^{-1} & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & * & 0 & * & * \\ 0 & * & 0 & * & * \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & 0 & 0 & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} - \right.$$

$$- \begin{bmatrix} 0 & & & & \\ & \dot{Q}_{22} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & Q_{22}^{-1} & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & * & 0 & * & * \\ 0 & * & 0 & * & * \end{bmatrix} -$$

$$\left. - \begin{bmatrix} 0 & & & & \\ & Q_{22} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{d}{dt}Q_{22}^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & * & * \\ 0 & * & 0 & * & * \end{bmatrix} \right) \sim$$

$$\sim \left(\begin{bmatrix} 0 & & & & \\ & I & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} & A_{15} \\ 0 & X & 0 & A_{24} & A_{25} \\ 0 & 0 & I & 0 & 0 \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right)$$

mit

$$X = -(\dot{Q}_{22}Q_{22}^{-1} + Q_{22}\frac{d}{dt}Q_{22}^{-1}) = -\frac{d}{dt}(Q_{22}Q_{22}^{-1}) = -\dot{I} = 0.$$

□

Die Aussage dieses Satzes erlaubt nun ein sukzessives Vorgehen, indem man ausgehend von $(E_0, A_0) = (E, A)$ eine Folge (E_i, A_i), $i \in \mathbb{N}_0$, von Paaren von Matrixfunktionen dadurch konstruiert, daß man (E_i, A_i) auf die Normalform (3.8) transformiert und dann zu (3.10) übergeht, um (E_{i+1}, A_{i+1}) zu erhalten. Man beachte, daß in jedem Schritt eine (3.7) entsprechende Voraussetzung gemacht werden muß. Satz 26 besagt dann, daß die so gewonnene Folge (r_i, a_i, s_i) von Invarianten charakteristisch für das Ausgangspaar ist. Die Beziehung $r_{i+1} = r_i - s_i$, die man direkt aus dem Vergleich der linken Matrizen in (3.8) und (3.10) entnimmt, garantiert nun, daß nach endlich vielen Schritten die Strangeness verschwindet. Ist dies erreicht, so werden alle Folgen von da an stationär, da durch den Übergang von (3.8) auf (3.10) nichts mehr geändert wird. Auch der Folgenindex, ab dem dies eintritt, ist charakteristisch für das vorgegebene Paar.

Definition 27 Sei (E, A) ein Paar von hinreichend glatten Matrixfunktionen. Die Folge (r_i, a_i, s_i) sei wohldefiniert, insbesondere gelte also (3.7) entsprechend für jedes Folgenglied (E_i, A_i) der oben konstruierten Folge. Dann heißt

$$\mu = \min\{i \in \mathbb{N}_0 \mid s_i = 0\} \tag{3.11}$$

Strangeness-Index von (E, A) bzw. von (1.2).

Die vorangegangene Diskussion können wir nun wie folgt zusammenfassen.

Satz 28 *Sei für (E, A) der Strangeness-Index μ wohldefiniert (d. h. seien die Voraussetzungen in Definition 27 erfüllt) und sei $f \in C^{\mu}(\mathbb{I}, \mathbb{C}^n)$. Dann ist (1.2) äquivalent (im dem Sinn, daß es eine umkehrbar eindeutige Abbildung zwischen den Lösungsräumen gibt) zu einer differentiell-algebraischen Gleichung der Form*

$$\begin{array}{ll} \text{(a)} & \dot{x}_1 = A_{13}(t)x_3 + f_1(t), \\ \text{(b)} & 0 = x_2 + f_2(t), \\ \text{(c)} & 0 = f_3(t), \end{array} \tag{3.12}$$

wobei sich die Inhomogenität aus $f, \dot{f}, \ldots, f^{(\mu)}$ bestimmt.

Beweis:
Nach dem Vorangegangen weiß man, daß das Paar (E_μ, A_μ) strangeness-frei ist, daß also in (3.8) zwei der fünf Blöcke nicht auftreten. Außerdem sind alle Umformungen umkehrbar und ändern die Struktur des Lösungsraumes nicht. □

An den übriggebliebenen Gleichungen (3.12) lassen sich jetzt leicht die eingangs gestellten Fragen beantworten.

Korollar 29 *Sei für (E, A) der Strangeness-Index μ wohldefiniert und sei $f \in C^{\mu+1}(\mathbb{I}, \mathbb{C}^n)$. Dann gilt:*

1. *Das Problem (1.2) ist genau dann lösbar, wenn die u_μ funktionalen Konsistenzbedingungen*
$$f_3 \equiv 0 \tag{3.13}$$
erfüllt sind.

2. *Eine Anfangsbedingung (1.3) ist genau dann konsistent, wenn zusätzlich die a_μ Bedingungen*
$$x_2(t_0) = -f_2(t_0) \tag{3.14}$$
von (1.3) impliziert werden.

3. *Das zugehörige Anfangswertproblem ist genau dann eindeutig lösbar, wenn wiederum zusätzlich*
$$u_\mu = 0 \tag{3.15}$$
gilt.

Man beachte, daß die stärkere Glattheitsforderung im vorstehenden Korollar dazu dient zu garantieren, daß die Lösung stetig differenzierbar ist. Man stellt aber anhand von (3.12) fest, daß es eigentlich ausreichen würde zu fordern, daß die Inhomogenität dort stetig ist. Es wäre dann nur x_1 stetig differenzierbar, während für x_2 und x_3 Stetigkeit genügt. Diese Problematik des "richtigen" Lösungsraumes wollen wir allerdings hier nicht vertiefen.

Vergleicht man das obige Resultat mit Satz 14 (der Strangeness-Index ist für Paare von konstanten Matrixfunktionen trivialerweise wohldefiniert), so kann man offensichtlich die Regularität eines Matrixpaares durch die Bedingung (3.15) verallgemeinern, während μ die Rolle von $\nu - 1$ übernimmt.

Statt den Übergang von (3.8) auf (3.10) durch Differentiation tatsächlich auszuführen, kann man auch versuchen, nur mit Äquivalenztransformationen

des ursprünglichen Paares (E, A) auszukommen. Man hätte dann ein zu (E, A) global äquivalentes Paar von Matrixfunktionen zur Verfügung, an dem man den Strangeness-Index wie auch die Folge (r_i, a_i, s_i) der charakteristischen Größen von (E, A) ablesen könnte.

Satz 30 *Sei für (E, A) der Strangeness-Index μ wohldefiniert und (r_i, a_i, s_i), $i = 0, \ldots, \mu$, die Folge der zugehörigen charakteristischen Größen. Weiter sei (in Verbindung mit (3.4d,e))*

$$w_0 = u_0, \qquad w_i = u_i - u_{i-1}, \quad i = 1, \ldots, \mu, \tag{3.16}$$

und

$$c_0 = a_0 + s_0, \qquad c_i = s_{i-1} - w_i, \quad i = 1, \ldots, \mu. \tag{3.17}$$

Dann ist (E, A) global äquivalent zu einem Paar der Form

$$\left(\left[\begin{array}{c|ccc|ccc} I & 0 & \cdots & 0 & 0 & * & * \\ \hline 0 & 0 & \cdots & 0 & 0 & F_\mu & * \\ \vdots & \vdots & & \vdots & & \ddots & \ddots & \\ \vdots & \vdots & & \vdots & & & \ddots & F_1 \\ 0 & 0 & \cdots & 0 & & & & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & G_\mu & * \\ \vdots & \vdots & & \vdots & & \ddots & \ddots & \\ \vdots & \vdots & & \vdots & & & \ddots & G_1 \\ 0 & 0 & \cdots & 0 & & & & 0 \end{array}\right], \left[\begin{array}{c|ccc|ccc} * & * & \cdots & * & 0 & \cdots & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ \hline 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & I & & & \\ \vdots & \vdots & & \vdots & & \ddots & & \\ \vdots & \vdots & & \vdots & & & \ddots & \\ 0 & 0 & \cdots & 0 & & & & I \end{array}\right] \right) \begin{array}{l} d_\mu \\ w_\mu \\ \vdots \\ \vdots \\ w_0 \\ c_\mu \\ \vdots \\ \vdots \\ c_0 \end{array} . \tag{3.18}$$

Dabei gilt

$$\operatorname{rang} \begin{bmatrix} F_i \\ G_i \end{bmatrix} = c_i + w_i = s_{i-1} \le c_{i-1}. \tag{3.19}$$

Die Matrixfunktionen F_i und G_i haben also zusammen punktweise vollen Zeilenrang.

Beweis:
Aus Platzgründen kann der Beweis hier nicht geführt werden. Der interessierte Leser sei auf [9] verwiesen. Der dortige Beweis beruht im wesentlichen darauf, die Transformation, die man nach Übergang auf (3.10) anwendet, soweit wie möglich auf (3.8) zu übertragen. □

Zum Abschluß dieses Kapitels wollen wir die erhaltenen Resultate auf unsere beiden Beispiele anwenden. Dabei verwenden wir die Abkürzung "dif" über dem Äquivalenzzeichen, um den Übergang von (3.8) auf (3.10) zu markieren.

Beispiel 31 Für das Problem aus Beispiel 15 gilt

$$\begin{aligned} (E, A) &= \left(\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \sim \\ &\sim \left(\begin{bmatrix} 0 & 1 \\ 1 & -t \end{bmatrix} \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & -t \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \sim \end{aligned}$$

$$
\begin{aligned}
&\sim \left(\begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & t \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right) \sim \\
&\sim \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right) \sim \\
&\overset{\text{dif}}{\sim} \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right).
\end{aligned}
$$

Damit lauten die charakteristischen Größen

$$
\begin{aligned}
r_0 = 1, \quad a_0 = 0, \quad s_0 = 1, \quad d_0 = 0, \quad u_0 = 0, \\
r_1 = 0, \quad a_1 = 1, \quad s_1 = 0, \quad d_1 = 0, \quad u_1 = 1,
\end{aligned}
$$

außerdem $\mu = 1$. Die zugehörige differentiell-algebraische Gleichung besteht also aus einer algebraischen Gleichung bei einer unbestimmten Komponente. Insbesondere ist also die Lösung des homogenen Anfangswertproblems nicht eindeutig in Übereinstimmung mit den Resultaten von Beispiel 15.

Beispiel 32 Für das Problem aus Beispiel 16 gilt

$$
\begin{aligned}
(E, A) = &\left(\begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix}, \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \right) \sim \\
&\sim \left(\begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right) \sim \\
&\sim \left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \sim \\
&\overset{\text{dif}}{\sim} \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right).
\end{aligned}
$$

Damit lauten die charakteristischen Größen

$$
\begin{aligned}
r_0 = 1, \quad a_0 = 0, \quad s_0 = 1, \quad d_0 = 0, \quad u_0 = 0, \\
r_1 = 0, \quad a_1 = 2, \quad s_1 = 0, \quad d_1 = 0, \quad u_1 = 0,
\end{aligned}
$$

außerdem $\mu = 1$. Die zugehörige differentiell-algebraische Gleichung besteht also aus zwei algebraischen Gleichungen. Insbesondere ist damit die Lösung ohne Vorgabe von Anfangsbedingungen eindeutig in Übereinstimmung mit den Resultaten von Beispiel 16.

Kapitel 4

Numerische Verfahren

Nachdem wir uns in den beiden vorangegangenen Kapiteln mit den analytischen Eigenschaften linearer differentiell-algebraischer Gleichungen auseinandergesetzt haben, wollen wir in diesem Kapitel Verfahren zu deren numerischer Lösung entwickeln. Aus Platzgründen werden wir uns hier nur mit Mehrschrittverfahren beschäftigen. Ausgangspunkt sollen die für die Lösung gewöhnlicher Differentialgleichungen bekannten Resultate sein. Im folgenden bezeichne h eine fest gewählte Schrittweite, $t_l = t_0 + lh$ die zugehörigen Gitterpunkte und x_l Approximationen an $x(t_l)$.

Definition 33 Ein lineares k-Schritt-Verfahren, $k \in \mathbb{N}$, zur Lösung einer gewöhnlichen Differentialgleichung $\dot{x} = f(t,x)$ ist gegeben durch die Verfahrensvorschrift

$$\sum_{l=0}^{k} \alpha_l x_l = h \sum_{l=0}^{k} \beta_l f(t_l, x_l) \tag{4.1}$$

zur Berechnung von x_k mit Koeffizienten $\alpha_l, \beta_l \in \mathbb{R}$, $l = 0, \ldots, k$, wobei $\alpha_k \neq 0$ und $\alpha_0^2 + \beta_0^2 \neq 0$. Die durch

$$\varrho(\lambda) = \sum_{l=0}^{k} \alpha_l \lambda^l, \quad \sigma(\lambda) = \sum_{l=0}^{k} \beta_l \lambda^l \tag{4.2}$$

festgelegten Polynome ϱ und σ heißen charakteristische Polynome des Mehrschrittverfahrens.

Ist $\beta_k = 0$, so kann die Vorschrift (4.1) direkt nach x_k aufgelöst werden (explizites Verfahren). Für $\beta_k \neq 0$ muß im allgemeinen ein nichtlineares Gleichungssystem, bei linearen Problemen zumindest ein lineares Gleichungssystem, gelöst werden (implizites Verfahren). Man beachte, daß man (4.1) durch Multiplikation mit einem Skalar verschieden normieren kann, z. B. durch die Forderung $\alpha_k = 1$ oder $\beta_k = 1$ für implizite Verfahren. Die Eigenschaften von Mehrschrittverfahren bei der Lösung von gewöhnlichen Differentialgleichungen lassen sich wie folgt zusammenfassen, siehe z. B. [7].

Satz 34 *Das lineare k-Schritt-Verfahren* (4.1) *sei konsistent der Ordnung p,* $p \in \mathbb{N}$, *d. h. es gelte für jedes hinreichend glatte* $x \in C(\mathbb{I}, \mathbb{C}^n)$, *daß*

$$\sum_{l=0}^{k} (\alpha_l x(t_l) - h\beta_l \dot{x}(t_l)) = \mathcal{O}(h^{p+1}) \quad \text{für } h \to 0 \tag{4.3}$$

bzw. äquivalent

$$\sum_{l=0}^{k} \alpha_l l^q = q \sum_{l=0}^{k} \beta_l l^{q-1}, \quad q = 0, \ldots, p, \tag{4.4}$$

und stabil, d. h. die Wurzeln des Polynoms ϱ *liegen in der Einheitskreisscheibe und die Wurzeln auf dem Rand sind einfach. Dann ist das Mehrschrittverfahren konvergent der Ordnung p, d. h. für jede Lösung x von* $\dot{x} = f(t, x)$ *mit hinreichend glattem* $f \in C(\mathbb{I} \times \mathbb{C}^n, \mathbb{C}^n)$ *und für Startwerte* $x_0, \ldots, x_{k-1}$ *mit*

$$x_l - x(t_l) = \mathcal{O}(h^p) \quad \text{für } h \to 0 \tag{4.5}$$

gilt für festes $T = t_0 + Kh \in \mathbb{I}$, $K \in \mathbb{N}_0$, *daß*

$$x(T; h) - x(T) = \mathcal{O}(h^p) \quad \text{für } h \to 0, \tag{4.6}$$

wobei $x(T; h)$ *die mit Schrittweite h an der Stelle T erzielte numerische Approximation bezeichnet.*

Bemerkung 35 Die Konsistenzbedingungen (4.4) für $q = 0, 1$ zusammen mit der Stabilitätsforderung implizieren

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1) \neq 0. \tag{4.7}$$

Es ist also $\lambda = 1$ einfache Nullstelle des Polynoms ϱ. Stabilität eines Mehrschrittverfahrens impliziert außerdem die Existenz einer Vektornorm, sodaß in der induzierten Matrixnorm (bei der Normierung $\alpha_k = 1$)

$$\|M\| \leq 1, \quad M = \begin{bmatrix} -\alpha_{k-1} & \cdots & -\alpha_1 & -\alpha_0 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}. \tag{4.8}$$

gilt.

Es ist nun die Frage, wie man Mehrschrittverfahren verallgemeinern kann, um damit differentiell-algebraische Gleichungen zu lösen. Etwas einfacher gestaltet sich die Beantwortung dieser Frage bei einem den Mehrschrittverfahren verwandten Verfahrenstyp, den sogenannten One-leg-Verfahren.

Definition 36 Ein One-leg-Verfahren mit k Schritten, $k \in \mathbb{N}$, zur Lösung von $\dot{x} = f(t, x)$ ist gegeben durch die Verfahrensvorschrift

$$\frac{1}{h} \sum_{l=0}^{k} \alpha_l x_l = f\Big(\sum_{l=0}^{k} \beta_l t_l, \sum_{l=0}^{k} \beta_l x_l\Big) \tag{4.9}$$

zur Berechnung von x_k mit Koeffizienten $\alpha_l, \beta_l \in \mathbb{R}$, $l = 0, \ldots, k$, wobei $\alpha_k \neq 0$, $\alpha_0^2 + \beta_0^2 \neq 0$ und $\sigma(1) = 1$ mit σ gemäß (4.2).

Hier kann man die Vorschrift (4.9) so interpretieren, daß man in $\dot{x} = f(t,x)$ die Ersetzungen

$$t \leftarrow \sum_{l=0}^{k} \beta_l t_l, \quad x \leftarrow \sum_{l=0}^{k} \beta_l x_l, \quad \dot{x} \leftarrow \frac{1}{h} \sum_{l=0}^{k} \alpha_l x_l \tag{4.10}$$

vornimmt. Dies ist natürlich sofort umsetzbar auf Gleichungen der Form (1.1) gemäß

$$F(\sum_{l=0}^{k} \beta_l t_l, \sum_{l=0}^{k} \beta_l x_l, \frac{1}{h} \sum_{l=0}^{k} \alpha_l x_l) = 0. \tag{4.11}$$

Wendet man diese so erhaltenen Verfahren auf die rein algebraische lineare Gleichung

$$0 = A(t)x + f(t)$$

mit punktweise nichtsingulärem A an, so ergibt sich

$$0 = A(\bar{t}) \sum_{l=0}^{k} \beta_l x_l + f(\bar{t}), \quad \bar{t} = \sum_{l=0}^{k} \beta_l t_l$$

oder

$$\beta_k x_k = -A(\bar{t})^{-1} f(\bar{t}) - \sum_{l=0}^{k-1} \beta_l x_l.$$

Um nach x_k auflösen zu können, muß $\beta_k \neq 0$ sein. Weiter erkennt man, daß im allgemeinen ein lineares Gleichungssystem nicht exakt gelöst wird, sondern daß vielmehr sogar Fehler von vorangegangen Schritten übertragen werden. Ein möglicher Ausweg ist die Forderung

$$\beta_0 = \cdots = \beta_{k-1} = 0, \quad \beta_k = 1, \tag{4.12}$$

passend zu $\sigma(1) = 1$. In diesem Fall gilt $\bar{t} = t_k$ und man erhält mit $x_k = -A(t_k)^{-1} f(t_k)$ als numerische Approximation die tatsächliche Lösung. Darüberhinaus stimmen unter der Bedingung (4.12) die Verfahrensvorschriften (4.1) und (4.9) überein und wir können die für Mehrschrittverfahren bekannten Resultate verwenden. Fordert man zusätzlich zu (4.12) noch möglichst hohe Ordnung in (4.4), so erhält man die sogenannten BDF-Verfahren. Diese lassen sich auch so konstruieren, daß man als Approximation an $\dot{x}(t_k)$ die Ableitung des in $(t_l, x(t_l))$, $l = 0, \ldots, k$, interpolierenden Polynoms an der Stelle t_k verwendet, woher auch der Name "Backward Differentiation Formulae" stammt. Für die BDF-Verfahren gilt $p = k$ in (4.4). Sie sind allerdings nur stabil für $k \leq 6$. Die Koeffizienten α_l sind in Tabelle 4.1 angegeben.

Es ist nun zu untersuchen, inwiefern BDF-Verfahren, gegeben durch die Verfahrensvorschrift

$$F(t_k, x_k, \frac{1}{h} \sum_{l=0}^{k} \alpha_l x_l) = 0, \tag{4.13}$$

für die numerische Lösung differentiell-algebraischer Gleichungen geeignet sind. Wie auch schon bei der analytischen Untersuchung von (1.1) werden wir uns auf lineare Probleme beschränken und mit dem Fall konstanter Koeffizienten beginnen. Für Konvergenzaussagen ist es wesentlich, daß eine eindeutige Lösung existiert. Die Einschränkung auf reguläre Matrixpaare ist daher naheliegend.

Tabelle 4.1 Koeffizienten der BDF-Verfahren

α_{k-l}	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
$k=1$	1	-1					
$k=2$	$\frac{3}{2}$	-2	$\frac{1}{2}$				
$k=3$	$\frac{11}{6}$	-3	$\frac{3}{2}$	$-\frac{1}{3}$			
$k=4$	$\frac{25}{12}$	-4	3	$-\frac{4}{3}$	$\frac{1}{4}$		
$k=5$	$\frac{137}{60}$	-5	5	$-\frac{10}{3}$	$\frac{5}{4}$	$-\frac{1}{5}$	
$k=6$	$\frac{147}{60}$	-6	$\frac{15}{2}$	$-\frac{20}{3}$	$\frac{15}{4}$	$-\frac{6}{5}$	$\frac{1}{6}$

Satz 37 *Seien $E, A \in \mathbb{C}^n$ und (E,A) regulär. Dann sind die BDF-Verfahren* (4.13) *angewandt auf* (2.1) *für $k \leq 6$ konvergent der Ordnung $p = k$ im Sinne von Satz* 34.

Beweis:
Wendet man (4.13) auf (2.1) an, so erhält man

$$E\frac{1}{h}\sum_{l=0}^{k}\alpha_l x_l = Ax_k + f(t_k).$$

Wegen der Regularität von (E,A) kann man gemäß (2.9) auf die Weierstraß-Normalform transformieren. Dabei zerfällt die obige Gleichung in zwei ungekoppelte Teile. Nennt man jeweils wieder die numerischen Approximationen x_l und die Inhomogenität f, so ist der erste Teil nichts anderes als das BDF-Verfahren angewandt auf die gewöhnliche Differentialgleichung (2.6). Hierfür gilt die Behauptung nach Satz 34. Zu untersuchen bleibt der zweite Teil, der dann die Form

$$N\frac{1}{h}\sum_{l=0}^{k}\alpha_l x_l = x_k + f(t_k).$$

besitzt. Um daraus die x_l zu gewinnen, kann man die Gleichung zunächst mit $N^{\nu-1}$ multiplizieren, um $N^{\nu-1}x_k = -N^{\nu-1}f(x_k)$ zu erhalten. Unabhängig von den Startwerten erfüllen nach k Schritten alle Iterierten, die in die weitere Rechnung eingehen, die entsprechende Beziehung. Als nächstes multipliziert man die Gleichung mit $N^{\nu-2}$ und nutzt die eben erhaltenen Beziehungen aus. Nach ν-maligem Anwenden entsprechender Argumente erhält man schließlich x_k in Termen von $f(t_l)$, zumindest wenn das Mehrschrittverfahren genügend weit fortgeschritten ist. Wesentlich ist nur, daß die Zahl der dazu nötigen Schritte von h unabhängig ist. Schließlich muß das Ergebnis in eine Taylorreihe entwickelt werden und mit (2.8) verglichen werden. Die geschilderte Vorgehensweise in dieser Form durchzuführen gestaltet sich jedoch extrem technisch. Es soll deshalb hier ein formaler Zugang genommen werden entsprechend zum Beweis von Lemma 11. Dazu definieren wir einen diskreten Ableitungsoperator D_h durch

$$D_h x_k = \frac{1}{h}\sum_{l=0}^{k}\alpha_l x_l.$$

Die obige Gleichung wird dadurch zu

$$ND_h x_k = x_k + f(t_k).$$

Da D_h linear ist und mit N vertauscht, erhält man

$$x_k = -(I - ND_h)^{-1} f(t_k) = -\sum_{i=0}^{\nu-1} N^i D_h^i f(t_k).$$

Wegen (4.3) zusammen mit (4.12) und $p = k$ für die BDF-Verfahren gilt

$$D_h f(t_k) = \dot{f}(t_k) + \sum_{q \geq k} c_q \frac{h^q}{q!} f^{(q)}(t_k) = \dot{f}(t_k) + \mathcal{O}(h^k).$$

Anwenden von D_h liefert

$$\begin{aligned} D_h^2 f(t_k) &= D_h \dot{f}(t_k) + \sum_{q \geq k} c_q \frac{h^q}{q!} D_h f^{(q)}(t_k) = \\ &= \ddot{f}(t_k) + \sum_{q \geq k} \overline{c}_q \frac{h^q}{q!} f^{(q+1)}(t_k) = \\ &= \ddot{f}(t_k) + \mathcal{O}(h^k). \end{aligned}$$

Durch iteratives Vorgehen erhält man daraus

$$D_h^i f(t_k) - f^{(i)}(t_k) = \mathcal{O}(h^k) \quad \text{für } h \to 0, \quad i = 0, \ldots, \nu - 1$$

und schließlich

$$x_k - x(t_k) = -\sum_{i=0}^{\nu-1} N^i (D_h^i f(t_k) - f^{(i)}(t_k)) = \mathcal{O}(h^k) \quad \text{für } h \to 0,$$

womit auch für den nilpotenten Teil Konvergenz der BDF-Verfahren folgt. □

Damit ist gezeigt, daß die BDF-Verfahren für $k \leq 6$ auch zum Lösen von differentiell-algebraischen Gleichungen mit konstanten Koeffizienten geeignet sind, zumindest wenn man die Schrittweite konstant hält. Für gewöhnliche Differentialgleichungen gibt es entsprechende Resultate, wenn man Änderungen der Schrittweite in einem bestimmten beschränkten Maß zuläßt. Im vorliegenden Fall können sich Schrittweitenänderungen allerdings katastrophal auswirken.

Bemerkung 38 Ist $h_{\max}$ die größte auftretende Schrittweite und beschränkt man die Schrittweitenänderungen wie oben erwähnt, so kann man zeigen (vgl. etwa [1]), daß im Fall von Satz 37 Konvergenz gemäß

$$x(T; h) - x(T) = \mathcal{O}(h_{\max}^p) \quad \text{für } h_{\max} \to 0 \tag{4.14}$$

mit $p = \min(k, k - \nu + 2)$ vorliegt, falls $k \geq \nu - 1$ ist. Es tritt also für $\nu \geq 3$ zumindest eine Ordnungsreduktion auf.

Noch schlimmer sieht die Situation im Fall von linearen differentiell-algebraischen Gleichungen mit variablen Koeffizienten aus, selbst bei der Verwendung konstanter Schrittweite. Diskretisiert man (1.2) mit einem BDF-Verfahren, so erhält man

$$E(t_k)\frac{1}{h}\sum_{l=0}^{k}\alpha_l x_l = A(t_k)x_k + f(t_k) \tag{4.15}$$

bzw. durch Umstellen nach der neuen Iterierten x_k

$$(\alpha_k E(t_k) - hA(t_k))x_k = hf(t_k) - E(t_k)\sum_{l=0}^{k-1}\alpha_l x_l. \tag{4.16}$$

Es gibt also genau dann eine eindeutige numerische Lösung x_k, wenn die Matrix $\alpha_k E(t_k) - hA(t_k)$ nichtsingulär ist. Dies ist für bestimmte Schrittweiten h nur dann möglich, wenn das Matrixpaar $(E(t_k), A(t_k))$ regulär ist. Um die BDF-Verfahren überhaupt sinnvoll durchführen zu können, müßte man also voraussetzen, daß $(E(t), A(t))$ für alle $t \in \mathbb{I}$ regulär ist. Wir haben allerdings zu Anfang von Kapitel 3 bereits gesehen, daß diese Eigenschaft im allgemeinen völlig unabhängig von den Eigenschaften des zu lösenden Problems ist. So liefern die BDF-Verfahren in Beispiel 15 unabhängig von einer vielleicht zusätzlichen Inhomogenität eine eindeutige Lösung, obwohl das gegebene Problem je nach Inhomogenität entweder keine oder unendlich viele, aber nie eine eindeutige Lösung besitzt. Bei Beispiel 16 hingegen sind die BDF-Verfahren überhaupt nicht durchführbar, obwohl das Problem eine eindeutige Lösung besitzt. Man könnte sich jetzt natürlich fragen, für welche Teilprobleme BDF-Verfahren trotzdem geeignet sind. Wir wollen uns jedoch dieser Frage nur insoweit stellen, wie sie uns später bei der Entwicklung allgemeinerer Verfahren weiterhilft. So ist in beiden Beispielen $s_0 \neq 0$. Eine Idee wäre nun, daß für Gleichungen mit $s_0 = 0$ und damit $\mu = 0$ keine Probleme auftreten.

Lemma 39 *Sei für das Paar (E, A) von Matrixfunktionen der Strangeness-Index μ wohldefiniert mit $\mu = 0$. Dann ist $(E(t), A(t))$ regulär für alle $t \in \mathbb{I}$ genau dann, wenn $u_0 = 0$ gilt.*

Beweis:
Wegen $\mu = 0$ ist $s_0 = 0$. Sei zunächst $u_0 = 0$. Mit $U = (Z', Z)$ und $V = (T', T)$ nach Satz 22 gilt punktweise für starke Äquivalenz

$$(E, A) \sim \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} Z'^* A T' & Z'^* A T \\ Z^* A T' & Z^* A T \end{bmatrix} \right).$$

Dabei ist wegen $r_0 + a_0 = n$ der Eintrag $Z^* A T$ nichtsingulär. Damit gilt weiter

$$(E, A) \sim \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} * & 0 \\ 0 & Z^* A T \end{bmatrix} \right) \sim \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \right).$$

Das Matrixpaar $(E(t), A(t))$ ist also für alle $t \in \mathbb{I}$ regulär mit $\mathrm{ind}(E(t), A(t)) = 0$, falls $a_0 = 0$, und $\mathrm{ind}(E(t), A(t)) = 1$, falls $a_0 \neq 0$. Sei nun andererseits $u_0 \neq 0$. In diesem Fall gibt es zu jedem $t \in \mathbb{I}$ einen Vektor $\overline{v} \in \mathbb{C}^n$, $\overline{v} \neq 0$, mit

$\overline{v}^*\overline{E}(t) = \overline{v}^*\overline{A}(t) = 0$, wobei $(\overline{E}, \overline{A})$ eine globale Normalform von (E, A) gemäß (3.8) sei. Rücktransformation liefert sofort die Existenz eines nichtverschwindenden Vektors v mit $v^*E(t) = v^*A(t) = 0$ und $\lambda E(t) - A(t)$ kann für kein λ nichtsingulär sein. □

Tatsächlich kann man nun für diesen Spezialfall von linearen differentiell-algebraischen Gleichungen Konvergenz der BDF-Verfahren bei Verwendung konstanter Schrittweite zeigen. Dazu benötigen wir zunächst folgendes Hilfsresultat (vgl. [1]).

Lemma 40 *Sei $\tilde{M}$ eine quadratische Matrix der Gestalt*

$$\tilde{M} = \begin{bmatrix} M_0 + \mathcal{O}(h) & \mathcal{O}(h) \\ 0 & N \end{bmatrix} \tag{4.17}$$

mit $\|M_0\| \leq 1$ und sei N nilpotent mit $\nu = \mathrm{ind}(N, I)$. Dann gilt

$$\tilde{M}^i = \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(h) \\ 0 & 0 \end{bmatrix} \tag{4.18}$$

für $i \geq \nu$ und $ih \leq T - t_0$, $T \in \mathbb{I}$ fest.

Beweis:
Wegen der oberen Blockdreiecksgestalt von $\tilde{M}$ gilt für $i \geq \nu$

$$\tilde{M}^i = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix}^i =$$

$$= \begin{bmatrix} M_{11}^i & \sum_{j=0}^{i-1} M_{11}^{i-j-1} M_{12} M_{22}^j \\ 0 & M_{22}^i \end{bmatrix} = \begin{bmatrix} M_{11}^i & \sum_{j=0}^{\nu-1} M_{11}^{i-j-1} M_{12} M_{22}^j \\ 0 & 0 \end{bmatrix}.$$

Da $M_{12} = \mathcal{O}(h)$ in jedem Summanden des $(1,2)$-Eintrages vorkommt und der Rest beschränkt ist, folgt die Behauptung für den $(1,2)$-Eintrag von $\tilde{M}^i$. Für den $(1,1)$-Eintrag gilt

$$\begin{aligned} \|M_{11}^i\| = \|(M_0 + \mathcal{O}(h))^i\| \leq \|M_0 + \mathcal{O}(h)\|^i \leq \\ \leq (1 + Ch)^i \leq (e^{Ch})^i \leq e^{C(T-t_0)} \end{aligned}$$

und damit $M_{11}^i = \mathcal{O}(1)$. □

Wir können nun Konvergenz der BDF-Verfahren mit $k \leq 6$ für strangeness-freie, für konsistente Anfangsbedingungen eindeutig lösbare lineare differentiell-algebraische Gleichungen zeigen.

Satz 41 *Sei für das Paar (E, A) von hinreichend glatten Matrixfunktionen der Strangeness-Index μ wohldefiniert mit $\mu = 0$ und gelte neben $s_0 = 0$ auch $u_0 = 0$. Dann sind die BDF-Verfahren (4.13) angewandt auf (1.2) für $k \leq 6$ konvergent der Ordnung $p = k$ im Sinne von Satz 34.*

Beweis:
Sei x die Lösung von (1.2) mit (1.3) und bezeichne τ_i den zugehörigen lokalen Fehler der BDF-Verfahren entsprechend (4.3), d. h. sei

$$\tau_i = \sum_{l=0}^{k} \alpha_{k-l} x(t_{i-l}) - h\dot{x}(t_i) = \mathcal{O}(h^{k+1}) \quad \text{für } h \to 0.$$

Dann gilt mit $e_{i-l} = x_{i-l} - x(t_{i-l})$

$$
\begin{aligned}
0 &= E(t_i)\sum_{l=0}^{k}\alpha_{k-l}x_{i-l} - hA(t_i)x_i - hf(t_i) = \\
&= E(t_i)\sum_{l=0}^{k}\alpha_{k-l}(x(t_{i-l}) + e_{i-l}) - hA(t_i)(x(t_i) + e_i) - hf(t_i) = \\
&= E(t_i)\sum_{l=0}^{k}\alpha_{k-l}e_{i-l} + E(t_i)(\tau_i + h\dot{x}(t_i)) - hA(t_i)(x(t_i) + e_i) - hf(t_i) = \\
&= h(E(t_i)\dot{x}(t_i) - A(t_i)x(t_i) - f(t_i)) + E(t_i)\sum_{l=0}^{k}\alpha_{k-l}e_{i-l} - hA(t_i)e_i + E(t_i)\tau_i,
\end{aligned}
$$

d. h. also

$$E(t_i)\sum_{l=0}^{k}\alpha_{k-l}e_{i-l} - hA(t_i)e_i + E(t_i)\tau_i = 0.$$

Nach Lemma 39 und dem dort geführten Beweis gibt es glatte, punktweise nichtsinguläre Matrixfunktionen P und Q mit

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad PAQ = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}.$$

Definiert man $\tilde{e}_{i-l} = Q(t_{i-l})^{-1}e_{i-l}$ und $\tilde{\tau}_{i-l} = Q(t_{i-l})^{-1}\tau_{i-l}$, so folgt

$$
\begin{aligned}
P(t_i)E(t_i)Q(t_i)\sum_{l=0}^{k}\alpha_{k-l}Q(t_i)^{-1}Q(t_{i-l})Q(t_{i-l})^{-1}e_{i-l} - \\
-hP(t_i)A(t_i)Q(t_i)Q(t_i)^{-1}e_i + P(t_i)E(t_i)Q(t_i)Q(t_i)^{-1}\tau_i = 0
\end{aligned}
$$

bzw.

$$
\begin{aligned}
P(t_i)E(t_i)Q(t_i)(\alpha_k\tilde{e}_i + \sum_{l=1}^{k}\alpha_{k-l}Q(t_i)^{-1}Q(t_{i-l})\tilde{e}_{i-l}) - \\
-hP(t_i)A(t_i)Q(t_i)\tilde{e}_i + P(t_i)E(t_i)Q(t_i)\tilde{\tau}_i = 0
\end{aligned}
$$

und schließlich mit $\gamma_l = -\alpha_{k-l}/\alpha_k$

$$\tilde{e}_i = Q(t_i)^{-1}(E(t_i) - \frac{h}{\alpha_k}A(t_i))^{-1}E(t_i)Q(t_i)(\sum_{l=1}^{k}\gamma_l Q(t_i)^{-1}Q(t_{i-l})\tilde{e}_{i-l} - \frac{1}{\alpha_k}\tilde{\tau}_i).$$

Setzt man $S_i = Q(t_i)^{-1}(E(t_i) - \frac{h}{\alpha_k}A(t_i))^{-1}E(t_i)Q(t_i)$, so ist

$$
\begin{aligned}
S_i &= (P(t_i)E(t_i)Q(t_i) - \frac{h}{\alpha_k}P(t_i)A(t_i)Q(t_i))^{-1}P(t_i)E(t_i)Q(t_i) = \\
&= \begin{bmatrix} (I - \frac{h}{\alpha_k}J(t_i))^{-1} & 0 \\ 0 & 0 \end{bmatrix}.
\end{aligned}
$$

Unterteilt man nun $\tilde{e}_i = (\tilde{e}_i^{(1)}, \tilde{e}_i^{(2)})$ entsprechend dieser Blockstruktur und geht über auf ein formales Einschrittverfahren gemäß

$$\boldsymbol{e}_i = (\tilde{e}_i^{(1)}, \ldots, \tilde{e}_{i-k+1}^{(1)}, \tilde{e}_i^{(2)}, \ldots, \tilde{e}_{i-k+1}^{(2)}),$$

so gilt

$$\begin{bmatrix} \tilde{e}_i^{(1)} \\ \tilde{e}_{i-1}^{(1)} \\ \vdots \\ \tilde{e}_{i-k+1}^{(1)} \\ \hline \tilde{e}_i^{(2)} \\ \tilde{e}_{i-1}^{(2)} \\ \vdots \\ \tilde{e}_{i-k+1}^{(2)} \end{bmatrix} = \left[\begin{array}{cccc|cccc} M_1^{(i)} & \cdots & \cdots & M_k^{(i)} & L_1^{(i)} & \cdots & \cdots & L_k^{(i)} \\ I & 0 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & I & 0 & & & & \\ \hline & & & & 0 & & & \\ & & & & I & 0 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & I & 0 \end{array}\right] \begin{bmatrix} \tilde{e}_{i-1}^{(1)} \\ \tilde{e}_{i-2}^{(1)} \\ \vdots \\ \tilde{e}_{i-k}^{(1)} \\ \hline \tilde{e}_{i-1}^{(2)} \\ \tilde{e}_{i-2}^{(2)} \\ \vdots \\ \tilde{e}_{i-k}^{(2)} \end{bmatrix} + \begin{bmatrix} b_1^{(i)} \\ 0 \\ \vdots \\ 0 \\ \hline 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

mit

$$\begin{aligned} [\, M_l^{(i)} \quad L_l^{(i)} \,] &= [\, \gamma_l (I - \frac{h}{\alpha_0} J(t_i))^{-1} \quad 0 \,]\, Q(t_i)^{-1} Q(t_{i-l}) = \\ &= [\, \gamma_l I + \mathcal{O}(h) \quad 0 \,] \begin{bmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & I + \mathcal{O}(h) \end{bmatrix} = \\ &= [\, \gamma_l I + \mathcal{O}(h) \quad \mathcal{O}(h) \,] \end{aligned}$$

und

$$b_1^{(i)} = -\frac{1}{\alpha_k}(I - \frac{h}{\alpha_k} J(t_i))^{-1} \tilde{\tau}_i = \mathcal{O}(h^{k+1}).$$

Damit lautet die obige Rekursion

$$\begin{aligned} \boldsymbol{e}_i &= M^{(i)} \boldsymbol{e}_{i-1} + \boldsymbol{b}^{(i)} = \\ &= M^{(i)} (M^{(i-1)} \boldsymbol{e}_{i-2} + \boldsymbol{b}^{(i-1)}) + \boldsymbol{b}^{(i)} = \\ &= M^{(i)} \cdots M^{(1)} \boldsymbol{e}_0 + \sum_{j=0}^{i} M^{(i)} \cdots M^{(j+1)} \boldsymbol{b}^{(j)} \end{aligned}$$

mit $\boldsymbol{e}_0 = \mathcal{O}(h^k)$, außerdem

$$M^{(i)} = \begin{bmatrix} M_0 + \mathcal{O}(h) & \mathcal{O}(h) \\ 0 & N \end{bmatrix},$$

wobei M_0 die zum Verfahren gehörige Stabilitätsmatrix aus (4.8) ist, mit dem einzigen Unterschied, daß dort jeder Eintrag mit dem Faktor I zu versehen ist. Entsprechend zu Bemerkung 35 garantiert Stabilität die Existenz einer Vektornorm, sodaß in der zugehörigen Matrixnorm $\|M_0\| \leq 1$ gilt. Mit Lemma 40 folgt nun

$$\begin{aligned} \boldsymbol{e}_i &= \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(h) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{O}(h^k) \\ \mathcal{O}(h^k) \end{bmatrix} + \sum_{j=0}^{i} \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(h) \\ 0 & N^{i-j} \end{bmatrix} \begin{bmatrix} \mathcal{O}(h^{k+1}) \\ 0 \end{bmatrix} = \\ &= \begin{bmatrix} \mathcal{O}(h^k) \\ 0 \end{bmatrix} + \begin{bmatrix} \mathcal{O}(h^{-1}) & \mathcal{O}(1) \\ 0 & \mathcal{O}(1) \end{bmatrix} \begin{bmatrix} \mathcal{O}(h^{k+1}) \\ 0 \end{bmatrix} = \mathcal{O}(h^k), \end{aligned}$$

gleichbedeutend mit der Behauptung. □

Unter Voraussetzungen wie in Bemerkung 38 läßt sich für variable Schrittweite Konvergenz mit gleicher Ordnung zeigen. Wie soll man aber lineare differentiell-algebraische Gleichungen numerisch angehen, wenn die Voraussetzungen von Satz 41 nicht erfüllt sind? Wünschenswert wäre im Hinblick auf die theoretischen Ergebnisse aus Kapitel 3 ein numerisches Verfahren zumindest für den Fall, daß der Strangeness-Index des zugehörigen Paares von Matrixfunktionen wohldefiniert ist und keine unbestimmten Lösungskomponenten vorkommen. Zwar zeigen die Ergebnisse aus Kapitel 3, daß man theoretisch zu einer (in dem dort angegebenen Sinn) äquivalenten linearen differentiell-algebraischen Gleichung übergehen kann, die die Voraussetzungen von Satz 41 erfüllt. Leider ist diese Vorgehensweise numerisch nicht durchführbar, da sie aufeinander aufbauende Transformationen mittels Matrixfunktionen und deren Ableitung verwendet. Dabei wird die spezielle Form aus (3.12) zur Durchführung der BDF-Verfahren gar nicht benötigt. Es ist nur wichtig, daß die entsprechenden Voraussetzungen erfüllt sind.

Die Frage, mit der wir uns nun auseinandersetzen wollen, ist, ob es eine numerisch zugängliche, die Voraussetzungen von Satz 41 erfüllende lineare differentiell-algebraische Gleichung gibt, deren Lösungsmenge umkehrbar eindeutig auf die Lösungsmenge des Ausgangsproblems abbildbar ist oder, besser noch, deren Lösungsmenge die gleiche ist wie die des Ausgangsproblems. Ausgangspunkt unserer Überlegungen ist die Konstruktion eines numerischen Verfahrens, das die Bestimmung der charaktistischen Größen (r_i, a_i, s_i) erlaubt. Da während der iterativen Prozedur aus Kapitel 3 Ableitungen gebildet werden, ist klar, daß Ableitungsinformation zumindest für die Funktionen E und A in diese numerische Prozedur einfließen muß. Nach einer Idee von Campbell (siehe [3]) leiten wir dazu die Gleichung (1.2) sukzessive ab, bringen alle Ableitungen von x auf die linke Seite und fassen die erhaltenen Gleichungen zu neuen großen differentiell-algebraischen Gleichungssystemen zusammen. Man erhält damit für hinreichend glatte Funktionen E und A für $\ell = 1, \ldots, \mu$ Gleichungen der Form

$$M_\ell(t)\dot{z}_\ell = N_\ell(t)z_\ell + g_\ell(t), \tag{4.19}$$

wobei die Matrixfunktionen M_ℓ, N_ℓ und die Vektorfunktionen z_ℓ, g_ℓ blockweise gegeben sind durch

$$\begin{aligned}
(M_\ell)_{i,j} &= \tbinom{i}{j} E^{(i-j)} - \tbinom{i}{j+1} A^{(i-j-1)},\ i,j = 0,\ldots,\ell,\\
(N_\ell)_{i,j} &= \begin{cases} A^{(i)} & \text{für } i = 0,\ldots,\ell,\ j = 0,\\ 0 & \text{sonst,} \end{cases}\\
(z_\ell)_i &= x^{(i)},\ i = 0,\ldots,\ell,\\
(g_\ell)_i &= f^{(i)},\ i = 0,\ldots,\ell.
\end{aligned} \tag{4.20}$$

Es soll nun gezeigt werden, daß für festes $t \in \mathbb{I}$ die lokalen charakteristischen Größen $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$ des Matrixpaares (gemäß (3.4)) invariant unter globalen Äquivalenztransformationen des Paares (E, A) von Matrixfunktionen sind. Dazu benötigen wir zunächst einige Identitäten für Binomialkoeffizienten. Dabei verwenden wir die Konvention, daß $\binom{i}{j} = 0$ für $i < 0$, $j < 0$ oder $j > i$.

Lemma 42 *Für alle $i,j,k,l \in \mathbb{Z}$ gilt*

$$\begin{aligned}
&\text{(a)} \quad \tbinom{i}{j}\tbinom{i-j}{k}\tbinom{i-j-k}{l} + \tbinom{i}{j+1}\tbinom{i-j-1}{k}\tbinom{i-j-k-1}{l} = \tbinom{i}{k}\tbinom{i-k}{l}\tbinom{i-k-l+1}{j+1},\\
&\text{(b)} \quad \tbinom{i}{j+1}\tbinom{i-j-1}{k}\tbinom{i-j-k-1}{l} = \tbinom{i}{k}\tbinom{i-k}{l}\tbinom{i-k-l}{j+1},\\
&\text{(c)} \quad \tbinom{i}{k-1}\tbinom{i-k+1}{l} + \tbinom{i}{k}\tbinom{i-k}{l-1} + \tbinom{i}{k}\tbinom{i-k}{l} = \tbinom{i+1}{k}\tbinom{i+1-k}{l}.
\end{aligned} \tag{4.21}$$

Beweis:
Die Behauptungen folgen direkt durch Einsetzen der Definition der Binomialkoeffizienten. □

Außerdem benötigen wir eine Beziehung für die Ableitungen eines Produkts von drei Matrixfunktionen.

Lemma 43 *Sei $D = ABC$ das Produkt dreier hinreichend glatter Matrixfunktionen passender Dimensionen. Dann gilt*

$$D^{(i)} = \sum_{j=0}^{i} \sum_{k=0}^{i-j} \tbinom{i}{j}\tbinom{i-j}{k} A^{(j)} B^{(k)} C^{(i-j-k)}. \tag{4.22}$$

Beweis:
Die Behauptung folgt durch Induktion unter Verwendung von (4.21c). □

Als zentrales Ergebnis für das weitere Vorgehen erhalten wir den folgenden Zusammenhang zwischen globaler und lokaler Äquivalenz.

Satz 44 *Seien die Paare (E,A) und $(\tilde{E},\tilde{A})$ von Matrixfunktionen hinreichend glatt und global äquivalent gemäß*

$$\tilde{E} = PEQ, \quad \tilde{A} = PAQ - PE\dot{Q} \tag{4.23}$$

und seien (M_ℓ, N_ℓ) und $(\tilde{M}_\ell, \tilde{N}_\ell)$ die zugehörigen erweiterten Paare nach (4.20). Sei außerdem der Strangeness-Index μ von (E,A) wohldefiniert. Dann sind die Matrixpaare $(M_\ell(t), N_\ell(t))$ und $(\tilde{M}_\ell(t), \tilde{N}_\ell(t))$ lokal äquivalent für jedes $t \in \mathbb{I}$ und jedes $\ell = 0, \ldots, \mu$.

Beweis:
Wir definieren Matrixfunktionen Π, Θ und Ψ durch

$$\begin{aligned}
&(\Pi_\ell)_{i,j} = \tbinom{i}{j} P^{(i-j)}, \quad (\Theta_\ell)_{i,j} = \tbinom{i+1}{j+1} Q^{(i-j)},\\
&(\Psi_\ell)_{i,j} = \begin{cases} Q^{(i+1)} & \text{für } i = 0,\ldots,\ell,\ j = 0,\\ 0 & \text{sonst,} \end{cases}
\end{aligned}$$

und beweisen die Behauptung, indem wir zeigen, daß

$$(\tilde{M}_\ell(t), \tilde{N}_\ell(t)) = \Pi_\ell(t)(M_\ell(t), N_\ell(t)) \begin{bmatrix} \Theta_\ell(t) & -\Psi_\ell(t) \\ 0 & \Theta_\ell(t) \end{bmatrix}$$

gilt. Da alle beteiligten Matrizen untere Blockdreiecksform besitzen, genügt es, dies für $\ell = \mu$ zu zeigen. Man beachte außerdem, daß N_ℓ, $\tilde{N}_\ell$ und Ψ_ℓ nur in

der ersten Blockspalte nichtverschwindende Einträge besitzen. Ausgehend von Lemma 43 erhalten wir zunächst

$$\begin{aligned}\tilde{E}^{(i)} &= \textstyle\sum_{k_1=0}^{i}\sum_{k_2=0}^{i-k_1}\binom{i}{k_1}\binom{i-k_1}{k_2}P^{(k_1)}E^{(k_2)}Q^{(i-k_1-k_2)},\\ \tilde{A}^{(i)} &= \textstyle\sum_{k_1=0}^{i}\sum_{k_2=0}^{i-k_1}[\binom{i}{k_1}\binom{i-k_1}{k_2}P^{(k_1)}A^{(k_2)}Q^{(i-k_1-k_2)} - \\ &\qquad -\binom{i}{k_1}\binom{i-k_1}{k_2}P^{(k_1)}E^{(k_2)}Q^{(i+1-k_1-k_2)}].\end{aligned}$$

Läßt man das Argument t der Einfachheit halber weg, so liefert die Definition der Matrixfunktionen

$$\begin{aligned}&(\Pi_\mu M_\mu \Theta_\mu)_{i,j} = \\ &\quad = \textstyle\sum_{l_1=j}^{i}\sum_{l_2=j}^{l_1}(\Pi_\mu)_{i,l_1}(M_\mu)_{l_1,l_2}(\Theta_\mu)_{l_2,j} = \\ &\quad = \textstyle\sum_{l_1=j}^{i}\sum_{l_2=j}^{l_1}\binom{i}{l_1}P^{(i-l_1)}[\binom{l_1}{l_2}E^{(l_1-l_2)} - \binom{l_1}{l_2+1}A^{(l_1-l_2-1)}]\binom{l_2+1}{j+1}Q^{(l_2-j)}.\end{aligned}$$

Verschieben bzw. Invertieren der Summation ergibt zusammen mit (4.21a,b)

$$\begin{aligned}&(\Pi_\mu M_\mu \Theta_\mu)_{i,j} = \\ &\quad = \binom{i}{j}\textstyle\sum_{k_1=0}^{i-j}\sum_{k_2=0}^{i-j-k_1}\binom{i-j}{k_1}P^{(k_1)}\binom{i-j-k_1}{k_2}E^{(k_2)}Q^{(i-j-k_1-k_2)} - \\ &\qquad -\binom{i}{j+1}\textstyle\sum_{k_1=0}^{i-j-1}\sum_{k_2=0}^{i-j-1-k_1}[\binom{i-j-1}{k_1}P^{(k_1)}\binom{i-j-k_1-1}{k_2}A^{(k_2)}Q^{(i-j-1-k_1-k_2)} - \\ &\qquad\quad -\binom{i-j-1}{k_1}P^{(k_1)}\binom{i-j-1-k_1}{k_2}E^{(k_2)}Q^{(i-j-k_1-k_2)}] = \\ &\quad = \binom{i}{j}\tilde{E}^{(i-j)} - \binom{i}{j+1}\tilde{A}^{(i-j-1)} = (\tilde{M}_\mu)_{i,j}.\end{aligned}$$

Entsprechend folgt, daß

$$\begin{aligned}&(\Pi_\mu N_\mu \Theta_\mu)_{i,0} - (\Pi_\mu M_\mu \Psi_\mu)_{i,0} = \\ &\quad = \textstyle\sum_{l_1=0}^{i}(\Pi_\mu)_{i,l_1}(N_\mu)_{l_1,0}(\Theta_\mu)_{0,0} - \\ &\qquad - \textstyle\sum_{l_1=0}^{i}\sum_{l_2=0}^{l_1}(\Pi_\mu)_{i,l_1}(M_\mu)_{l_1,l_2}(\Psi_\mu)_{l_2,0} = \\ &\quad = \textstyle\sum_{k_1=0}^{i}\binom{i}{k_1}P^{(k_1)}A^{(i-k_1)}Q^{(0)} + \\ &\qquad + \textstyle\sum_{k_1=0}^{i}\sum_{k_2=0}^{i-1-k_1}\binom{i}{k_1}P^{(k_1)}\binom{i-k_1}{k_2}A^{(k_2)}Q^{(i-k_1-k_2)} - \\ &\qquad - \textstyle\sum_{k_1=0}^{i}\sum_{k_2=0}^{i-k_1}\binom{i}{k_1}P^{(k_1)}\binom{i-k_1}{k_2}E^{(k_2)}Q^{(i-k_1-k_2+1)} = \\ &\quad = \tilde{A}^{(i)} = (\tilde{N}_\mu)_{i,0}.\end{aligned}$$

□

Damit ist gezeigt, daß die lokalen charakteristischen Größen $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$ entsprechend (3.4) von $(M_\ell(t), N_\ell(t))$ selbst wieder charakteristisch für das ursprüngliche Paar (E, A) von Matrixfunktionen und somit auch für die zugehörige differentiell-algebraische Gleichung sind. Die lokalen charakteristischen Größen $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$ sind aber numerisch bestimmbar. Im wesentlichen sind nach (3.4) für jedes ℓ drei aufeinanderfolgende numerische Rangbestimmungen durchzuführen, die etwa mittels Singulärwertzerlegung oder QR-Zerlegung mit Spaltentausch bewerkstelligt werden können. Es bleibt damit die Frage, ob man von der Kenntnis von $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$, $\ell = 0, \ldots, \mu$, zurückschließen kann auf die gewünschten charakteristischen Größen (r_i, a_i, s_i), $i = 0, \ldots, \mu$.

Satz 45 *Sei für das Paar (E, A) von hinreichend glatten Matrixfunktionen der Strangeness-Index μ wohldefiniert mit globalen charakteristischen Größen*

(r_i, a_i, s_i), $i = 0, \ldots, \mu$. Seien weiter $(M_\ell(t), N_\ell(t))$, $\ell = 0, \ldots, \mu$, $t \in \mathbb{I}$, die zu (E, A) gehörigen erweiterten Matrixpaare mit lokalen charakteristischen Größen $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$. Dann gilt

$$
\begin{aligned}
\tilde{r}_\ell &= (\ell+1)n - \sum_{i=0}^{\ell} c_i - \sum_{i=0}^{\ell} u_i, & \tilde{r}_\mu &= (\mu+1)n - a_\mu - \sum_{i=0}^{\mu} u_i, \\
\tilde{a}_\ell &= s_{\ell-1} - w_\ell - s_\ell = c_\ell - s_\ell, & \tilde{a}_\mu &= c_\mu, \\
\tilde{s}_\ell &= s_\ell + \sum_{i=0}^{\ell-1} c_i, & \tilde{s}_\mu &= \sum_{i=0}^{\mu-1} c_i = a_\mu - c_\mu, \\
\tilde{d}_\ell &= \tilde{r}_\ell - s_\ell = (\ell+1)n - c_\ell - \sum_{i=0}^{\ell} u_i, & \tilde{d}_\mu &= (\mu+1)n - c_\mu - \sum_{i=0}^{\mu} u_i, \\
\tilde{u}_\ell &= (\ell+1)n - \tilde{r}_\ell - \tilde{a}_\ell - \tilde{s}_\ell = \sum_{i=0}^{\ell} u_i, & \tilde{u}_\mu &= \sum_{i=0}^{\mu} u_i, \\
& \ell = 0, \ldots, \mu.
\end{aligned}
\tag{4.24}
$$

Beweis:
Wegen Satz 44 können wir ohne Einschränkung annehmen, daß das Paar (E, A) in der Normalform (3.18) vorliegt. Für festes $t \in \mathbb{I}$ wenden wir nun lokale Äquivalenztransformationen (3.2) auf $(M_\ell(t), N_\ell(t))$ an. Dabei werden wir im folgenden das Argument t weglassen. Zuerst bilden wir durch Umsortieren das Paar $(\overline{M}_\ell, \overline{N}_\ell)$ gemäß

$$
(\overline{M}_\ell)_{i,j} = (M_\ell)_{\ell-i,\ell-j}, \quad (\overline{N}_\ell)_{i,j} = (N_\ell)_{\ell-i,\ell-j}
$$

und erhalten

$$
\overline{M}_\ell = \begin{bmatrix} T_{0,0} & \cdots & T_{0,\ell} \\ & \ddots & \vdots \\ & & T_{\ell,\ell} \end{bmatrix}, \quad \overline{N}_\ell = \begin{bmatrix} 0 & \cdots & 0 & S_{0,\ell} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & S_{\ell,\ell} \end{bmatrix}.
$$

Dabei besitzt jeder Blockeintrag $T_{i,i}$, $i = 0, \ldots, \ell$ die Form

$$
\left[\begin{array}{c|ccc|ccc|c}
I & 0 & \cdots & 0 & 0 & * & \cdots & * \\ \hline
0 & 0 & \cdots & 0 & 0 & F_\mu & & * \\
\vdots & \vdots & & \vdots & & \ddots & \ddots & \\
\vdots & \vdots & & \vdots & & & \ddots & F_1 \\
0 & 0 & \cdots & 0 & & & & 0 \\ \hline
0 & 0 & \cdots & 0 & 0 & G_\mu & & * \\
\vdots & \vdots & & \vdots & & \ddots & \ddots & \\
\vdots & \vdots & & \vdots & & & \ddots & G_1 \\
0 & 0 & \cdots & 0 & & & & 0
\end{array}\right],
$$

jeder Blockeintrag $T_{i,i+1}$, $i = 0, \ldots, \ell - 1$ die Form

$$\left[\begin{array}{c|ccc|c|ccc} * & * & \cdots & * & 0 & * & \cdots & * \\ \hline 0 & 0 & \cdots & 0 & 0 & * & & * \\ \vdots & \vdots & & \vdots & & \ddots & \ddots & \\ \vdots & \vdots & & \vdots & & & \ddots & * \\ 0 & 0 & \cdots & 0 & & & & 0 \\ \hline 0 & 0 & \cdots & 0 & -I & * & & * \\ \vdots & \vdots & & \vdots & & \ddots & \ddots & \\ \vdots & \vdots & & \vdots & & & \ddots & * \\ 0 & 0 & \cdots & 0 & & & & -I \end{array}\right]$$

und jeder andere Blockeintrag im oberen rechten Teil die Form

$$\left[\begin{array}{c|ccc|c|ccc} * & * & \cdots & * & 0 & * & \cdots & * \\ \hline 0 & 0 & \cdots & 0 & 0 & * & & * \\ \vdots & \vdots & & \vdots & & \ddots & \ddots & \\ \vdots & \vdots & & \vdots & & & \ddots & * \\ 0 & 0 & \cdots & 0 & & & & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & * & & * \\ \vdots & \vdots & & \vdots & & \ddots & \ddots & \\ \vdots & \vdots & & \vdots & & & \ddots & * \\ 0 & 0 & \cdots & 0 & & & & 0 \end{array}\right] .$$

In $\overline{N}_\ell$ besitzt jeder Blockeintrag $S_{i,\ell}$, $i = 0, \ldots, \ell - 1$ die Form

$$\left[\begin{array}{c|ccc|c|cc} * & * & \cdots & * & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & & 0 \end{array}\right]$$

und $S_{\ell,\ell}$ besitzt die Form

$$\left[\begin{array}{c|ccc|c|cc} * & * & \cdots & * & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & & 0 \\ \hline 0 & 0 & \cdots & 0 & I & & \\ \vdots & \vdots & & \vdots & & \ddots & \\ 0 & 0 & \cdots & 0 & & & I \end{array}\right] .$$

Wir können nun die Matrix $\overline{M}_\ell$ folgendermaßen mit Hilfe von Block-Gauß-Elimination über die Identitäten vereinfachen. Die Identität in der oberen linken Ecke jedes Blockes $T_{i,i}$ wird verwendet, um alle anderen Blockeinträge in der zugehörigen Blockzeile zu eliminieren. Dann kann man mit der unteren rechten Identität im $(0,1)$-Block alle anderen Einträge in dieser Blockspalte eliminieren, anschließend mit den unteren zwei Identitäten im $(1,2)$-Block alle anderen Einträge in diesen Blockspalten und induktiv so weiter mit den unteren ℓ Identitäten im $(\ell-1,\ell)$-Block. Sei $(\hat{M}_\ell, \hat{N}_\ell)$ das Paar, das wir auf diese Weise erhalten. Es besitzt die gleiche grobe Blockstruktur wie $(\overline{M}_\ell, \overline{N}_\ell)$. Insbesondere wurde kein Nullblock zerstört. Bezeichnen wir die Blöcke von $\hat{M}_\ell$ und $\hat{N}_\ell$ mit $\hat{T}_{i,j}$ und $\hat{S}_{i,j}$, so können wir wie folgt weiterschließen. Der Diagonalblock $\hat{T}_{i,i}$ besitzt $c_0 + \cdots + c_i + w_0 + \cdots + w_i$ Nullzeilen. Durch die Identitäten in $\hat{T}_{i,i+1}$, falls vorhanden, tragen die $c_0 + \cdots + c_i$ aber nichts zum Rangdefekt von $\hat{M}_\ell$ bei. Da der Rest im wesentlichen Stufenform besitzt, erhalten wir

$$\begin{aligned}\tilde{r}_\ell &= (n-w_0) + (n-w_0-w_1) + (n-w_0-w_1-w_2) + \cdots + \\ &\quad + (n-w_0-\cdots-w_{\ell-1}) + (n-w_0-\cdots-w_\ell-c_0-\cdots-c_\ell) = \\ &= (\ell+1)n - \textstyle\sum_{i=0}^{\ell} u_i - \sum_{i=0}^{\ell} c_i.\end{aligned}$$

Um die anderen Invarianten entsprechend Satz 21 zu bestimmen, benötigen wir $Z_\ell^* \hat{N}_\ell T_\ell$ und $Z_\ell^* \hat{N}_\ell T_\ell'$, wobei die Spalten von Z_ℓ das Kobild von $\hat{M}_\ell$ und die Spalten von T_ℓ, T_ℓ' den Kern bzw. den Kokern von $\hat{M}_\ell$ aufspannen. Da $\hat{N}_\ell$ nur in der letzten Blockspalte nichtverschwindende Einträge besitzt und Z_ℓ einen Nullblock in der oberen linken Ecke (wegen der Identität im entsprechenden Block von $\hat{M}_\ell$) besitzt, ist der einzige relevante Teil in $Z_\ell^* \hat{N}_\ell$, der einen Beitrag zu den gesuchten Rängen liefert, derjenige, der zur letzten Blockzeile von $\hat{N}_\ell$ gehört. Dieser hat die Form

$$\left[\begin{array}{c|ccc|ccc} 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \hline & 0 & \cdots & 0 & 0 & \cdots & 0 \\ & \vdots & & \vdots & \vdots & & \vdots \\ & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \hline & & & & I & & \\ & & & & & \ddots & \\ & & & & & & I \end{array}\right] \begin{array}{l} \\ w_\ell \\ \vdots \\ w_0 \\ c_\ell \\ \vdots \\ c_0 \end{array} .$$

Wir erhalten nun den Kern von $\hat{M}_\ell$ über die vorhandenen Nullspalten und die Kerne der Einträge F_i und G_i. Sei

$$U_i = \begin{bmatrix} F_i \\ G_i \end{bmatrix}$$

mit orthonormalen Basen K_i und K_i' von $\operatorname{kern} U_i$ und $\operatorname{cokern} U_i$. Wegen (3.19) ist U_i eine $(c_i + w_i, c_{i-1})$-Matrix mit vollem Zeilenrang. Damit folgt $\operatorname{rang} K_i = c_{i-1} - c_i - w_i$ und

$$\dim \operatorname{kern} \hat{M}_\ell = \sum_{i=0}^{\ell} \Big(\sum_{j=i+1}^{\mu} (c_{j-1} - c_j - w_j) + c_\mu + \sum_{j=0}^{\mu} w_j \Big) = \sum_{i=0}^{\ell} (c_i + u_i).$$

Wegen der speziellen Form von $Z_\ell \hat{N}_\ell$ ist von $\hat{T}_\ell$ ebenfalls nur der letzte Teil relevant, welcher die Form

$$\left[\begin{array}{c|ccc|ccc} 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \hline & I & & & 0 & \cdots & 0 \\ & & \ddots & & \vdots & & \vdots \\ & & & I & 0 & \cdots & 0 \\ \hline & & & & K_{\mu+1} & & * \\ & & & & & \ddots & \\ & & & & & & K_{\ell+1} \\ & & & & & & 0 \end{array}\right] \begin{array}{l} d_\mu \\ w_\mu \\ \vdots \\ w_0 \\ c_\mu \\ \vdots \\ c_\ell \\ c_0 + \cdots + c_{\ell-1} \end{array}$$

mit $K_{\mu+1} = I$ besitzt. Für $\hat{T}'_\ell$ erhält man entsprechend als relevanten Teil

$$\left[\begin{array}{ccc|c|ccc} 0 & \cdots & 0 & I & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \hline & & & 0 & 0 & \cdots & 0 \\ & & & & K'_\ell & & \\ & & & & & \ddots & \\ & & & & & & K'_1 \end{array}\right] \begin{array}{l} d_\mu \\ w_\mu \\ \vdots \\ w_0 \\ c_\ell + \ldots + c_\mu \\ c_{\ell-1} \\ \vdots \\ c_0 \end{array} .$$

Damit erhalten wir schließlich

$$\tilde{a}_\ell = \operatorname{rang}(Z_\ell^* \hat{N}_\ell \hat{T}_\ell) = \operatorname{rang} K_{\ell+1} = c_\ell - c_{\ell+1} - w_{\ell+1} = s_{\ell-1} - w_\ell - s_\ell$$

bzw.

$$\tilde{s}_\ell = c_0 + \cdots + c_{\ell-1} + \operatorname{rang} K'_{\ell+1} = c_0 + \cdots + c_{\ell-1} + c_{\ell+1} + w_{\ell+1} = s_\ell + \sum_{i=0}^{\ell-1} c_i .$$

□

Die Beziehungen (4.24) erlauben jetzt ein sukzessives Berechnen der gewünschten charakteristischen Größen von (E, A).

Korollar 46 *Unter den Voraussetzungen von Satz* 45 *gelten die Rekursionen*

$$\begin{array}{ll} r_0 = \tilde{r}_0, & r_i = r_{i-1} - s_{i-1} \\ a_0 = \tilde{a}_0, & s_i = \tilde{s}_i - v_{i-1} \\ s_0 = \tilde{s}_0, & w_i = s_{i-1} - s_i - \tilde{a}_i \\ d_0 = \tilde{d}_0, & u_i = w_i - u_{i-1} \\ u_0 = \tilde{u}_0, & a_i = n - r_i - s_i - u_i \\ w_0 = u_0, & d_i = r_i - s_i \\ c_0 = a_0 + s_0, & c_i = s_{i-1} - w_i \\ v_0 = s_0, & v_i = v_{i-1} + c_i \end{array} \tag{4.25}$$

Beweis:
Die Werte r_i ergeben sich direkt aus der Konstruktion der Folge. Aus der dritten Beziehung von (4.24) erhalten wir s_i, um anschließend aus der zweiten Beziehung w_i zu bestimmen. Damit hat man drei unabhängige charakteristische Größen für den i-ten Schritt bestimmt und alle anderen charakteristischen Größen ergeben sich aus den entsprechenden Definitionen. □

Damit haben wir zumindest ein numerisches Verfahren entwickelt, um die zu gegebenem Paar (E, A) gehörigen charakteristischen Größen (r_i, a_i, s_i) samt Strangeness-Index μ zu bestimmen. Man beachte, daß dieses Verfahren nur Information von E und A und deren Ableitungen ausgewertet an einer festen Stelle $t \in \mathbb{I}$ verwendet. Dieses Verfahren erlaubt also umgekehrt auch die Definition dieser charakteristischen Größen, im Gegensatz zu dem Vorgehen in Kapitel 3 sogar mit dem Vorteil, daß man diese punktweise festlegen kann. In diesem Sinn können wir also z. B. angeben, welchen Strangeness-Index eine lineare differentiell-algebraische Gleichung an einer bestimmten Stelle besitzt.

Wir kommen nun zur Frage, wie wir die gewonnenen Resultate verwenden können, um lineare differentiell-algebraische Gleichungen tatsächlich auch numerisch lösen zu können. Aus der bisherigen Diskussion ist klar, daß die wesentliche Information im Paar (M_μ, N_μ) enthalten sein muß. Im folgenden lassen wir hier den Index weg, schreiben also kurz (M, N), ebenso $(\tilde{M}, \tilde{N})$ für das zur Normalform $(\tilde{E}, \tilde{A})$ gehörige große Paar, und nehmen stillschweigend an, daß die Voraussetzungen der Sätze 44 und 45 erfüllt sind. Insbesondere sind dann die Größen $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$ nicht von $t \in \mathbb{I}$ abhängig. Mit Blick auf Satz 28 ist unser Ziel, aus (M, N) punktweise ein Paar $(\hat{E}, \hat{A})$ von Matrixfunktionen zu gewinnen, für das der Strangeness-Index verschwindet und die lokalen charakteristischen Größen gegeben sind durch $\hat{r} = d_\mu$, $\hat{a} = a_\mu$ und $\hat{s} = 0$. Dabei ist wesentlich, daß dieses Verfahren numerisch durchführbar sein soll. Um einen möglichst stabilen Algorithmus zu erhalten, wird darauf geachtet, daß alle Transformationen möglichst kleine Norm besitzen. Das kann man dadurch erreichen, daß man versucht, die Spalten der auftretenden Matrizen oder Matrixfunktionen orthonormal zu wählen.

Wir betrachten zunächst die zur Normalform gehörigen Matrixfunktionen. Nach Satz 45 gilt gerade $a_\mu = \tilde{a}_\mu + \tilde{s}_\mu$. Wegen der Konstanz der Ränge garantiert Satz 22 die Existenz einer glatten Matrixfunktion $\tilde{Z}_2$ der Größe $((\mu+1)n, a_\mu)$ mit orthonormalen Spalten derart, daß gilt

$$\tilde{Z}_2^* \tilde{M} = 0, \quad \operatorname{rang}(\tilde{Z}_2^* \tilde{N} \begin{bmatrix} I_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}) = \operatorname{rang}(\begin{bmatrix} 0 \\ 0 \\ I_{a_\mu} \end{bmatrix}^* \begin{bmatrix} * & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{a_\mu} \end{bmatrix}) = a_\mu.$$

Setzt man nun

$$\tilde{T}_2 = \begin{bmatrix} I_{d_\mu} & 0 \\ 0 & I_{u_\mu} \\ 0 & 0 \end{bmatrix},$$

so gilt

$$\tilde{Z}_2^* \tilde{N} \begin{bmatrix} I_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tilde{T}_2 = 0$$

und

$$\operatorname{rang}(\tilde{E}\tilde{T}_2) = \operatorname{rang}(\begin{bmatrix} I_{d_\mu} & 0 & * \\ 0 & 0 & F \\ 0 & 0 & G \end{bmatrix} \begin{bmatrix} I_{d_\mu} & 0 \\ 0 & I_{u_\mu} \\ 0 & 0 \end{bmatrix}) = d_\mu.$$

Damit haben wir zunächst einmal für die zu einer Normalform gehörigen Matrixfunktionen Teile herausprojiziert, die die richtigen Ränge aufweisen. Diese Resultate müssen jetzt auf die zum ursprünglichen Paar gehörigen Matrixfunktionen übertragen werden. Ausgehend von

$$\tilde{M} = \Pi M \Theta, \quad \tilde{N} = \Pi N \Theta - \Pi M \Psi$$

erhalten wir

$$\tilde{Z}_2^* \tilde{M} = \tilde{Z}_2^* \Pi M \Theta = 0$$

und wegen der speziellen Struktur von N

$$\begin{aligned} a_\mu &= \operatorname{rang}(\tilde{Z}_2^* \tilde{N} [\, I_n \quad 0 \quad \dots \quad 0 \,]^*) = \\ &= \operatorname{rang}(\tilde{Z}_2^* \Pi N \Theta [\, I_n \quad 0 \quad \dots \quad 0 \,]^*) = \\ &= \operatorname{rang}(\tilde{Z}_2^* \Pi N [\, Q^{(0)} \quad * \quad \dots \quad * \,]^*) = \\ &= \operatorname{rang}(\tilde{Z}_2^* \Pi N [\, Q^{(0)} \quad 0 \quad \dots \quad 0 \,]^*) = \\ &= \operatorname{rang}(\tilde{Z}_2^* \Pi N [\, I_n \quad 0 \quad \dots \quad 0 \,]^* Q^{(0)}). \end{aligned}$$

Damit ist gezeigt, daß es eine glatte Matrixfunktion Z_2 der Größe $((\mu+1)n, a_\mu)$ gibt, die ohne Einschränkung orthonormale Spalten besitzt (durch Orthonormierung von $\Pi^* \tilde{Z}_2$) und die Beziehungen

$$Z_2^* M = 0, \quad \operatorname{rang}(Z_2^* N \begin{bmatrix} I_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}) = a_\mu$$

erfüllt. Entsprechend folgt aus

$$\begin{aligned} 0 &= \tilde{Z}_2^* \tilde{N} [\, I_n \quad 0 \quad \dots \quad 0 \,]^* \tilde{T}_2 = \\ &= \tilde{Z}_2^* \Pi N \Theta [\, I_n \quad 0 \quad \dots \quad 0 \,]^* \tilde{T}_2 = \\ &= \tilde{Z}_2^* \Pi N [\, I_n \quad 0 \quad \dots \quad 0 \,]^* Q^{(0)} \tilde{T}_2 \end{aligned}$$

und

$$d_\mu = \operatorname{rang}(\tilde{E}\tilde{T}_2) = \operatorname{rang}(P^{(0)} E Q^{(0)} \tilde{T}_2) = \operatorname{rang}(E Q^{(0)} \tilde{T}_2)$$

die Existenz einer glatten Matrixfunktion T_2 der Größe $(n, d_\mu + u_\mu)$, wiederum ohne Einschränkung mit orthonormalen Spalten, mit

$$Z_2^* N \begin{bmatrix} I_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} T_2 = 0, \quad \operatorname{rang}(ET_2) = d_\mu.$$

Damit existiert schließlich eine glatte Matrixfunktion Z_1 der Größe (n, d_μ) mit orthonormalen Spalten und

$$\operatorname{rang}(Z_1^* E T_2) = d_\mu.$$

Setzt man nun

$$(\hat{E}, \hat{A}) = \left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} \right)$$

mit

$$\hat{E}_1 = Z_1^* E, \quad \hat{A}_1 = Z_1^* A, \quad \hat{A}_2 = Z_2^* N [\, I_n \quad 0 \quad \dots \quad 0 \,]^*,$$

so erhalten wir aus dem Paar (M, N) ein Paar $(\hat{E}, \hat{A})$, das die gleiche Größe wie das ursprüngliche Paar (E, A) besitzt. Es bleibt zu überprüfen, ob es auch die gewünschten charakteristischen Größen besitzt.

Satz 47 *Die lokalen charakteristischen Größen von $(\hat{E}, \hat{A})$ sind unabhängig von $t \in \mathbb{I}$ gegeben durch*

$$(\hat{r}, \hat{a}, \hat{s}) = (d_\mu, a_\mu, 0). \tag{4.26}$$

Beweis:
Wir setzen mit der obigen Notation $T_2 = [T_1', T_3]$, wobei T_1' in den Kokern von $\hat{E}_1$ multipliziert. Multipliziert außerdem T_2' in den Kokern von $\hat{A}_2$, so sind $\hat{E}_1 T_1'$ und $\hat{A}_2 T_2'$ sowie $[T_1', T_2', T_3]$ punktweise nichtsingulär und wir erhalten punktweise die folgenden lokalen Äquivalenzen.

$$\begin{aligned}
(\hat{E}, \hat{A}) &= \left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} \right) \sim \\
&\overset{\mathrm{Q}}{\sim} \left(\begin{bmatrix} \hat{E}_1 T_1' & \hat{E}_1 T_2' & \hat{E}_1 T_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 T_1' & \hat{A}_1 T_2' & \hat{A}_1 T_3 \\ 0 & \hat{A}_2 T_2' & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \sim \\
&\overset{\mathrm{Q}}{\sim} \left(\begin{bmatrix} \hat{E}_1 T_1' & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} * & * & * \\ 0 & \hat{A}_2 T_2' & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \sim \\
&\overset{\mathrm{P}}{\sim} \left(\begin{bmatrix} I_{d_\mu} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} * & * & * \\ 0 & I_{a_\mu} & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \sim \\
&\overset{\mathrm{B}}{\sim} \left(\begin{bmatrix} I_{d_\mu} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_{a_\mu} & 0 \\ 0 & 0 & 0 \end{bmatrix} \right)
\end{aligned}$$

Vom letzten Paar können wir nun direkt ablesen, daß $\hat{r} = d_\mu$, $\hat{a} = a_\mu$ und $\hat{s} = 0$. □

Insbesondere verschwindet also für $(\hat{E}, \hat{A})$ der Strangeness-Index. Setzt man nun noch

$$\hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix} = \begin{bmatrix} Z_1^* f \\ Z_2^* g \\ 0 \end{bmatrix}, \tag{4.27}$$

so hat man aus der differentiell-algebraischen Gleichung (4.19) für $\ell = \mu$ die differentiell-algebraischen Gleichung

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t) \tag{4.28}$$

gewonnen. Die Besonderheit der Herleitung ist, daß die gesuchte Funktion x dabei nicht mittransformiert wurde. Ist insbesondere $u_\mu = n - d_\mu - a_\mu = 0$, so hat (4.28) die gleiche Lösungsmenge wie (1.2). Das gleiche gilt, wenn die Anfangsbedingung (1.3) hinzugenommen wird. Ist hingegen $u_\mu \neq 0$, so gilt dies wegen der Setzung $\hat{f}_3 = 0$ nur, wenn die ursprüngliche Gleichung lösbar war. Im Fall der eindeutigen Lösbarkeit des Anfangswertproblems garantiert dann Satz 41 die Anwendbarkeit der BDF-Verfahren. Außerdem kann man an Hand von (4.28) sofort entscheiden, ob eine Anfangsbedingung konsistent ist. Es muß nämlich gelten, daß

$$0 = \hat{A}_2(t_0)x_0 + \hat{f}_2(t_0). \tag{4.29}$$

Es muß noch bemerkt werden, daß die Funktionen $\hat{E}$, $\hat{A}$ und $\hat{f}$ in der oben angegebenen Art wegen der Glattheitsforderung numerisch nicht zugänglich sind, da man zwar Orthonormalbasen der auftretenden Räume numerisch bestimmen kann, dies aber im allgemeinen nicht zu glatten Funktionen Z_1 und Z_2 führt. Stattdessen führt die numerische Bestimmung der Orthonormalbasen zu Funktionen $Z_1 U_1$ und $Z_2 U_2$ mit punktweise unitären Funktionen U_1 und U_2, die im allgemeinen nicht glatt sind. Damit erreicht man numerisch (4.28) nur bis auf eine eventuell nichtglatte, punktweise unitäre Skalierung von links. Die BDF-Verfahren sind aber glücklicherweise invariant gegen solche Transformationen der Gleichung. Man darf sich also auf den Standpunkt stellen, daß man glatte Funktionen zur Verfügung hat. Damit haben wir also letztendlich ein numerisches Verfahren konstruiert, das für eindeutig lösbare Anfangswertprobleme bei linearen differentiell-algebraischen Gleichungen mit wohl-definiertem Strangeness-Index ohne Einschränkung an dessen Größe anwendbar ist.

Beispiel 48 Für das Problem aus Beispiel 15 gilt mit $\mu = 1$

$$M(t) = \left[\begin{array}{cc|cc} -t & t^2 & 0 & 0 \\ -1 & t & 0 & 0 \\ \hline 0 & 2t & -t & t^2 \\ 0 & 2 & -1 & t \end{array}\right], \quad N(t) = \left[\begin{array}{cc|cc} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Es ist $\operatorname{rang} M(t) = 2$ unabhängig von $t \in \mathbb{I}$ und

$$Z_2^*(t) = \left[\begin{array}{cc|cc} 1 & -t & 0 & 0 \\ \hline 0 & 0 & 1 & -t \end{array}\right], \quad Z_2^*(t)N(t) = \left[\begin{array}{cc|cc} -1 & t & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right].$$

Damit erhält man

$$\hat{E}(t) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{A}(t) = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}$$

und man kann die in Beispiel 15 angegebene allgemeine Lösung des zugehörigen homogenen Anfangswertproblems direkt ablesen.

Beispiel 49 Für das Problem aus Beispiel 16 gilt mit $\mu = 1$

$$M(t) = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ 1 & -t & 0 & 0 \\ \hline 1 & -t & 0 & 0 \\ 0 & -1 & 1 & -t \end{array}\right], \quad N(t) = \left[\begin{array}{cc|cc} -1 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Es ist wiederum $\operatorname{rang} M(t) = 2$ unabhängig von $t \in \mathbb{I}$ und

$$Z_2^*(t) = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & -1 & 0 \end{array}\right], \quad Z_2^*(t)N(t) = \left[\begin{array}{cc|cc} -1 & t & 0 & 0 \\ \hline 0 & -1 & 0 & 0 \end{array}\right].$$

Damit erhält man

$$\hat{E}(t) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{A}(t) = \begin{bmatrix} -1 & t \\ 0 & -1 \end{bmatrix}, \quad \hat{f}(t) = \begin{bmatrix} f_1(t) \\ f_2(t) - \dot{f}_1(t) \end{bmatrix}.$$

Auch hier kann man die in Beispiel 16 angegebene Lösung direkt ablesen. Wegen $u_\mu = 0$ kann man für dieses Beispiel bei konsistenter Anfangsbedingung die BDF-Verfahren zur numerischen Berechnung der eindeutigen Lösung verwenden. Da die zu $(\hat{E}, \hat{A})$ gehörige lineare differentiell-algebraische Gleichung wegen $d_\mu = 0$ nur aus algebraischen Gleichungen besteht, wird diese durch die BDF-Verfahren bis auf Rundungsfehler exakt gelöst. Es sei hier noch einmal darauf hingewiesen, daß direktes Anwenden der BDF-Verfahren auf (1.2) bei diesem Beispiel nicht möglich ist.

Im Anschluß folgt noch ein Literaturverzeichnis für ergänzende und weiterführende Studien.

Literaturverzeichnis

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*. Elsevier, North Holland, New York, 1989.

[2] S. L. Campbell. *Singular Systems of Differential Equations*. Pitman, San Francisco, 1980.

[3] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18:1101–1115, 1987.

[4] F. R. Gantmacher. *The Theory of Matrices*, volume I. Chelsea Publishing Company, New York, 1959.

[5] F. R. Gantmacher. *The Theory of Matrices*, volume II. Chelsea Publishing Company, New York, 1959.

[6] E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment*. Teubner Verlag, Leipzig, 1986.

[7] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I.* Springer-Verlag, Berlin, 1987.

[8] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II.* Springer-Verlag, Berlin, 1991.

[9] P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. Erscheint in J. Comput. Appl. Math., 1995.

[10] P. Kunkel and V. Mehrmann. A new class of discretization methods for the solution of linear differential-algebraic equations. Erscheint in SIAM J. Numer. Anal., 1995.

[11] W. C. Rheinboldt. On the computation of multi-dimensional solution manifolds of parameterized equations. *Numer. Math.*, 53:165–181, 1988.