Dissertation

# Numerische Lösung singulärer Anfangswertprobleme zweiter Ordnung

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Ao. Univ. Prof. Dipl.-Ing. Dr. techn. Ewa Weinmüller
E115
Institut für Angewandte und Numerische Mathematik

eingereicht an der Technischen Universität Wien
Technisch-Naturwissenschaftliche Fakultät

von

Dipl.-Ing. Othmar Koch

Matrikelnummer: 9225898

Schumanngasse 11/1, 1180 Wien.

Wien, am 22. Juli 1999

Diese Arbeit wurde von Othmar Koch mit LaTeX unter Verwendung der $\mathcal{AMS}$–Fonts gesetzt.

# Kurzfassung

Diese Arbeit beschäftigt sich mit der numerischen Lösung von singulären Anfangswertproblemen zweiter Ordnung der Gestalt

$$\begin{aligned} &y''(t) = \frac{A_1(t)}{t}y'(t) + \frac{A_0(t)}{t^2}y(t) + f(t, y(t)), \quad t \in (0, 1], \\ &B_0 y(0) = \beta, \\ &A_0(0)y(0) = 0, \quad y'(0) = 0. \end{aligned} \tag{1}$$

Dabei ist $y$ eine auf $[0, 1]$ stetige $n$-dimensionale Funktion, $A_0(t), A_1(t)$ stetige Matrizen und $f$ eine nichtlineare vektorwertige Funktion, die als Lipschitz-stetig vorausgesetzt wird. $B_0$ ist eine konstante $m \times n$-Matrix und $\beta \in \mathbb{R}^m,\ m \leq n$.

Diese Anfangsbedingungen sind die allgemeinsten, die zu einer stetigen Lösung $y$ führen. Diese Tatsache, sowie weitere analytische Eigenschaften von (1) wie Eindeutigkeit und Glattheit der Lösung, die in einer Arbeit von Koch, Kofler und Weinmüller (sh. [31]) hergeleitet wurden, werden in dieser Arbeit kurz zusammengefasst. Dabei stellt sich heraus, dass das Problem (1) nur unter bestimmten Bedingungen an $A_0(0)$ und $A_1(0)$ sinnvoll gelöst werden kann. Zur numerischen Lösung wird die Gleichung (1) auf ein äquivalentes System erster Ordnung transformiert, und zwar erhält man mit $z(t) = (z_1(t), z_2(t)) = (y(t), ty'(t))$ die Gleichung

$$\begin{aligned} &z'(t) = \frac{M(t)}{t}z(t) + \mathring{f}(t, z(t)), \quad t \in (0, 1], \\ &z(0) = z_0, \end{aligned} \tag{2}$$

mit

$$M(t) := \begin{pmatrix} 0 & I \\ A_0(t) & I + A_1(t) \end{pmatrix}, \quad \mathring{f}(t, z(t)) := \begin{pmatrix} 0 \\ f(t, z_1(t)) \end{pmatrix}.$$

Das Problem (2) wird mit vier verschiedenen Einschrittverfahren gelöst, und zwar dem expliziten und dem impliziten Euler-Verfahren, der Trapezregel und der Mittelpunktsregel. Es wird bewiesen, dass die beiden Euler-Verfahren mit der klassischen Ordnung 1 konvergieren, ebenso zeigt die Trapezregel die klassische Konvergenzordnung 2. In den meisten anwendungsrelevanten Situationen folgt dieses Konvergenzverhalten auch für die Mittelpunktsregel, nur in gewissen Spezialfällen tritt eine Ordnungsreduktion auf ein Niveau $|\ln(h)|^k h^2,\ k \in \mathbb{N}$, ein, diese Tatsache kann auch experimentell untermauert werden. Schließlich wird eine asymptotische Fehlerentwicklung für die mit dem impliziten Euler-Verfahren berechnete Näherungslösung hergeleitet. Eine solche Fehlerentwicklung dient oft als Basis für die Anwendung gewisser Beschleunigungsalgorithmen.

# Inhaltsverzeichnis

# Kapitel 1

# Einleitung

## 1.1 Definition singulärer Probleme

Diese Arbeit beschäftigt sich mit der numerischen Lösung gewisser Klassen *singulärer* Anfangswertprobleme. Ein solches singuläres Problem kann auf zwei Arten formuliert werden, und zwar auf einem endlichen oder einem unendlichen Intervall. Im ersten Fall hat die Gleichung dann die Form

$$x'(t) = \frac{1}{t^\alpha} g(t, x(t)), \quad t \in (0, 1], \tag{1.1}$$

im zweiten Fall

$$x'(t) = t^\beta g(t, x(t)), \quad t \in [1, \infty). \tag{1.2}$$

Dabei ist $g(t, v)$ auf einem geeigneten Gebiet (zumindest) stetig in $t$ und LIPSCHITZ-stetig bezüglich $v$ (gleichmäßig in $t$), weiters gilt $\alpha > 0$ und $\beta \geq 0$. Offensichtlich stehen die beiden Formulierungen über die Transformation $t \mapsto \frac{1}{t}$ in enger Beziehung. Man beachte, dass für $\beta = 0$ auch der reguläre Fall eingeschlossen ist. Abhängig vom Wert von $\alpha$ wird die Art der Singularität in (1.1) klassifiziert. Für $\alpha < 1$ spricht man von einer *schwachen Singularität*, für $\alpha = 1$ von einer *Singularität der ersten Art* und für $\alpha > 1$ von einer *Singularität der zweiten Art*.

Das Hauptinteresse dieser Arbeit liegt an singulären Anfangswertproblemen *zweiter Ordnung* der speziellen Gestalt

$$\begin{aligned} y''(t) &= \frac{A_1}{t} y'(t) + \frac{A_0}{t^2} y(t) + f(t, y(t)), \quad t \in (0, 1], \\ y(0) &= y_0, \quad y'(0) = y_1. \end{aligned} \tag{1.3}$$

Sei $I := (0, 1]$, $G$ ein geeignetes Gebiet im $\mathbb{R}^{n+1}$ und für die Gleichung (1.3) gelte: $y \in C^2(I, \mathbb{R}^n)$ sei eine zweimal stetig differenzierbare $n$-dimensionale Funktion,

$f \in C(G, \mathbb{R}^n)$ stetig, $A_0, A_1 \in C(\overline{I}, \mathbb{R}^{n\times n})$ seien stetige (oder sogar konstante) Matrizen und $y_0, y_1 \in \mathbb{R}^n$ konstante Vektoren, die gewissen Einschränkungen unterliegen können, siehe Kapitel 3[1]. Gesucht werden Lösungen $y$, die auch in 0 stetig fortsetzbar sind. Dass Probleme der Form (1.3) auch in das Schema von Gleichung (1.1) eingeordnet werden können, sieht man folgendermaßen:

Wendet man auf Gleichung (1.3) die Variablentransformation $z(t) := (z_1(t), z_2(t)) := (y(t), ty'(t))$ an, so erhält man eine Gleichung erster Ordnung, nämlich

$$\begin{aligned} z'(t) &= \frac{M}{t} z(t) + \overset{\circ}{f}(t, z(t)), \quad t \in (0, 1], \\ z(0) &= z_0, \end{aligned} \tag{1.4}$$

mit

$$M := \begin{pmatrix} 0 & I \\ A_0 & I + A_1 \end{pmatrix}, \quad \overset{\circ}{f}(t, z(t)) := \begin{pmatrix} 0 \\ f(t, z_1(t)) \end{pmatrix}.$$

Die Frage, wie die Anfangswerte zu wählen sind, um ein sinnvolles Problem zu erhalten, wird in Kapitel 3 geklärt werden. Deshalb soll hier in der Einleitung nicht weiter auf diese Schwierigkeit eingegangen werden.

Damit kann man also bei der Untersuchung von Gleichungen des Typs (1.3) das äquivalente singuläre Problem erster Ordnung mit einer Singularität der ersten Art (1.4) betrachten.

## 1.2 Auftreten singulärer Probleme in den Naturwissenschaften

In diesem Abschnitt soll ausgeführt werden, wo singuläre Probleme in Anwendungen (vor allem in der Physik) auftreten, um ihre wichtige Rolle zu verdeutlichen.

Mathematische Modelle vieler Anwendungen in der Physik und der Chemie nehmen die Form von Systemen zeitabhängiger partieller Differentialgleichungen unter Anfangs- und/oder Randbedingungen an. Das Literaturverzeichnis enthält eine kurze Liste von aktuellen Publikationen, die sich mit Anwendungen der Thomas-Fermi-Gleichungen ([59]–[68]) und der Ginzburg-Landau-Gleichungen ([69]–[78]) beschäftigen, um das rege Interesse an diesen Problemen zu demonstrieren. Thomas-Fermi-Gleichungen treten z. B. in Problemen aus der Quantenmechanik, in statistischen Modellen von Atomen und Molekülen, Modellen für isolierte neutrale Atome, in der Astrophysik oder in Modellen von Materie hoher Dichte auf, Ginzburg-Landau-Gleichungen finden sich z. B. im Kontext ferromagnetischer Systeme, der Supraleitung oder der Laser-Hydrodynamik.

[1] Für die hier verwendeten Schreibweisen siehe Abschnitt 1.4.

Für die Untersuchung stationärer Lösungen können viele dieser Probleme auf singuläre gewöhnliche Differentialgleichungen zurückgeführt werden, besonders, wenn aufgrund dem Problem inhärenter Symmetrien Polar-, Zylinder- oder Kugelkoordinaten verwendet werden.

## 1.3 Ziele und Aufbau dieser Arbeit

Wie die Ausführungen des letzten Abschnitts belegen, kommt der Lösung singulärer Probleme in vielen Gebieten der Naturwissenschaften eine wichtige Bedeutung zu. Aufgrund der Komplexität der Probleme kommt eine solche Lösung normalerweise nur *numerisch* in Betracht. Wie jedoch später in Kapitel 4 ausgeführt wird, kann man zentrale Anliegen der numerischen Mathematik mit den Standardmethoden für *glatte* Probleme nicht immer in den Griff bekommen, da die klassischen Zugänge bei der Untersuchung von numerischen Verfahren versagen[2].

Dennoch wurden und werden numerische Verfahren mit Erfolg auf singuläre Probleme angewendet, ohne dass es dafür immer eine allgemeine theoretische Rechtfertigung gibt. Ziel dieser Arbeit ist es, einige dieser Lücken zu schließen.

In Kapitel 5 wird versucht, eine Konvergenztheorie für einige Einschrittverfahren bei der Anwendung auf singuläre Probleme herzuleiten, Kapitel 6 enthält eine asymptotische Fehlerentwicklung für eines dieser Verfahren. Vor der Untersuchung dieser numerischen Aspekte singulärer Probleme wird jedoch auf einige analytische Aspekte wie etwa Existenz, Eindeutigkeit oder Glattheit der Lösung eingegangen. In Kapitel 2 werden zunächst einige „klassische“ Resultate über die Existenz und Eindeutigkeit der Lösung von Anfangswertproblemen angegeben, die man für hinreichende Glattheit der Datenfunktionen erhalten kann. Die Beweise der entsprechenden Sätze finden sich im Anhang A, da sie nicht zu den eigentlichen Intentionen dieser Arbeit zählen, dem interessierten Leser jedoch nicht vorenthalten werden sollen. Kapitel 3 geht schließlich auf die entsprechenden Fragestellungen nach Existenz, Eindeutigkeit und Glattheitseigenschaften der Lösung singulärer Probleme ein, die in [31] und [33] ausführlich herausgearbeitet wurden. Viele der dort beschriebenen Resultate gehen auch wesentlich in die Überlegungen zur numerischen Lösung singulärer Probleme (Kapitel 5) ein. Um keine zu großen Überschneidungen mit [31] zu provozieren, werden nur zum überblicksmäßigen Verständnis notwendige Beweise und Beweisideen tatsächlich ausgeführt, ansonsten wird auf [31] verwiesen.

[2] Ein Beispiel wäre hier die Konvergenztheorie mittels Stabilitätsüberlegungen, deren Beweis essentiell von Glattheitseigenschaften der betrachteten Gleichung abhängt. Oft kann man feststellen, dass numerische Methoden ihre Eigenschaften ändern, dass also z. B. eine Reduktion der Konvergenzordnung eines Verfahrens für singuläre Probleme auftritt, sh. dazu z. B. [25], [52] und [53].

In Kapitel 4 werden dann einige Grundkonzepte der numerischen Lösung von Anfangswertproblemen vorgestellt. Weiters werden einige klassische Verfahren wie z. B. das *explizite* und das *implizite* EULER-Verfahren eingeführt, sowie Konvergenzbeweise für diese Verfahren bei Anwendung auf glatte Probleme geführt. Das geschieht recht ausführlich, zum einen, weil viele der eingeführten Begriffe und vorgestellten Resultate auch in späteren Kapiteln Verwendung finden, zum anderen um zeigen zu können, wo die Grenzen der präsentierten Methoden liegen, wenn singuläre Probleme gelöst werden sollen.

Ab Kapitel 5 schließlich werden die vom Autor selbst bewiesenen neuen Resultate zu numerischen Aspekten singulärer Anfangswertprobleme ausgeführt.

In Kapitel 5 wird die Konvergenz von Einschrittverfahren (explizites und implizites EULER-Verfahren, Trapez- und Mittelpunktsregel) für das allgemeine nichtlineare Problem (1.3) in seiner Form (1.4) gezeigt. Dabei werden zuerst die entsprechenden Aussagen für *lineare Probleme mit konstanter Koeffizientenmatrix*, dann für *allgemeine lineare Probleme* und schließlich für den *nichtlinearen Fall* gezeigt. Die Beweise für das implizite EULER-Verfahren, die Trapezregel und die Mittelpunktsregel unterscheiden sich dabei nicht wesentlich von denen für das explizite EULER-Verfahren, und werden deshalb etwas weniger ausführlich angegeben. Einige für das Verständnis der Beweise wesentliche Sätze werden ausführlich im Anhang A bewiesen, insbesondere, wenn die Sätze in Kapitel 5 eine Modifikation von Sätzen, auf denen aufgebaut wurde, enthalten (etwa wenn sich die Voraussetzungen bekannter Sätze abschwächen lassen, um eine Verwendung für das untersuchte Problem zu ermöglichen).

Kapitel 6 enthält eine asymptotische Fehlerentwicklung für das implizite EULER-Verfahren. Die Herleitung dieser Entwicklung stützt sich wesentlich auf Resultate und Methoden aus den Kapiteln 3 und 5. Die Existenz einer solchen Fehlerentwicklung ist ein wesentliches Kriterium zum Funktionieren gewisser Beschleunigungsalgorithmen, sh. [12], [13], [14] und [32].

In den Anhängen finden sich einerseits Beweise vieler bereits bekannter Sätze, die für das Verständnis dieser Arbeit sinnvoll sind, aber dennoch den Rahmen des Hauptteils sprengen würden (Anhang A) und andererseits Sätze, deren Aussage in der Arbeit verwendet wird, ohne dass ihr Beweis von weiterem Interesse für die Überlegungen wäre (Anhang B). Dieser Abschnitt ist grob nach den mathematischen Gebieten, denen die jeweiligen Sätze entstammen, gegliedert.

Im Literaturverzeichnis schließlich finden sich sowohl Verweise auf Arbeiten, denen verwendete Sätze entnommen sind, als auch weitere Literatur über singuläre Probleme. Am Ende werden noch aktuelle Arbeiten aus Anwendungsgebieten aufgelistet.

## 1.4 Verwendete Notation und grundlegende Aspekte

In diesem Abschnitt werden die in dieser Arbeit häufig vorkommenden Notationen zusammengestellt, die dann im Weiteren ohne zusätzliche Erläuterungen verwendet werden. Die folgende Liste von Bezeichungen enthält *keine* Erklärungen mathematischer Konstrukte, die als bekannt vorausgesetzt werden und erhebt keinen Anspruch auf Vollständigkeit.

Metrische Räume werden allgemein als Paar $(M, \rho_M)$ angegeben, dabei bezeichnet $\rho_M$ die *Metrik* (*Abstandsfunktion*) und $M$ die Menge, auf der diese definiert ist. Analog werden *normierte Räume* bzw. *Banachräume* mit $(E, \| \ \|_E)$ und *Innenprodukträume* bzw. *Hilberträume* mit $(H, \langle \ \rangle_H)$ bezeichnet. Das Subskript wird dabei oft entfallen, wenn klar ist, zu welchem (Produkt-) Raum die entsprechende Funktion gehört. In einem metrischen Raum bezeichnet $K(x, r)$ die offene und $\overline{K}(x, r)$ die abgeschlossene Kugel um den Punkt $x$ mit Radius $r$. Sei allgemein $M$ eine Menge, dann steht $\overset{\circ}{M}$ für das *Innere*, $\overline{M}$ für den *Abschluss* und $\partial M$ für den *Rand* von $M$.

In reellen bzw. komplexen Vektorräumen wird in der Schreibweise *kein* Unterschied zwischen Skalaren und Vektoren gemacht, wenn dies unmissverständlich ist, auch wird kein Unterschied zwischen Zeilen- und Spaltenvektoren gemacht, wenn aus dem Kontext klar ist, worum es sich handelt, oder diese Tatsache in der jeweiligen Situation irrelevant ist. Dies gilt sowohl für Vektoren $x \in \mathbb{R}^n$ bzw. $\mathbb{C}^n$ als auch für Funktionen $f : G \subseteq \mathbb{R}^m \to \mathbb{R}^n$ bzw. $g : \tilde{G} \subseteq \mathbb{C}^m \to \mathbb{C}^n$.

In $\mathbb{R}$ bzw. $\mathbb{C}$ wird der Betrag $|\ |$ als Norm verwendet, im $\mathbb{R}^n(\mathbb{C}^n)$ die Maximumnorm; diese wird in der Notation auch als $|\ |$ geschrieben. Es gilt also

$$|x| := \max_{1 \le i \le n} |x_i| \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n(\mathbb{C}^n).$$

Wenn nicht anders angegeben, wird für Funktionen $f$ die Supremumsnorm verwendet und mit $\|f\|_\infty$ bezeichnet, wenn der Definitionsbereich klar ist (ansonsten wird im Text näher darauf eingegangen).

Hat die Funktion mehrere (endlich viele!) Koordinaten, so wird zuerst das Supremum jeder Koordinatenfunktion genommen, und dann das Maximum über diese Werte gebildet. In der Schreibweise wird das nicht vom skalaren Fall unterschieden.

Für Matrizen $A \in \mathbb{R}^{m \times n}$ gilt in Analogie zu den obigen Bezeichnungen, dass $|M|$ die Zeilensummennorm der konstanten Matrix $M$ bezeichnet. Ist $M = M(t)$ eine von $t$ abhängige Matrix, so ist

$$\|M\|_\infty = \max_t |M(t)|.$$

$|M(t)|$ bezeichnet dabei natürlich die Zeilensummennorm der Matrix $M$ in $t$.

**Bemerkung:** Die Zeilensummennorm für Matrizen entspricht der von der Maximumnorm im $\mathbb{R}^n$ bzw. $\mathbb{R}^m$ induzierten Operatornorm.

Sind $G$ und $\tilde{G}$ Gebiete, so wird der Raum der stetigen Funktionen $f : G \to \tilde{G}$ mit $C(G, \tilde{G})$ bezeichnet. Ist $f$ $p$-mal stetig differenzierbar, so schreibt man $f \in C^p(G, \tilde{G})$. Allgemein bezeichnet man den Definitionsbereich einer Funktion $f$ mit $\mathcal{D}(f)$ und den Bildbereich mit $\mathcal{R}(f)$. Die identische Abbildung wird stets mit $I$ bezeichnet. Handelt es sich bei $I \in \mathbb{R}^{2n \times 2n}$ um die $2n$-dimensionale Einheitsmatrix, dann steht $I_1$ für die ersten $n$ Zeilen von $I$, und $I_2$ für die letzten $n$ Zeilen.

Häufig werden in dieser Arbeit Vektoren, die auf einem geeigneten *Gitter* $\Delta := (t_{i_0}, \ldots, t_N)$ (cf. Definition 4.1.1) definiert sind, vorkommen. Handelt es sich dabei um *äquidistante Gitter* mit *Schrittweite* $h$ (was fast immer der Fall sein wird), so wird der Index $h$ verwendet, um einen solchen Vektor zu kennzeichnen, also

$$x_h := (x_{i_0}, \ldots, x_N).$$

Die Maximumnorm auf dem Raum dieser Vektoren wird mit $\| \ \|_h$ bezeichnet:

$$\|x_h\|_h := \max_{i_0 \le i \le N} |x_i|.$$

**Bemerkung:** Es kann dabei gelten $x_i \in \mathbb{R}^n(\mathbb{C}^n), i = i_0, \ldots, N$.

Ist $f$ eine Funktion, die auf einer Obermenge von $[t_{i_0}, t_N]$ definiert ist, so wird die Projektion von $f$ auf den Raum der Gitterfunktionen von $\Delta$ mit $R_\Delta$ (bei äquidistanten Gittern auch $R_h$) bezeichnet, d. h.

$$R_\Delta(f) = (f(t_{i_0}), \ldots, f(t_N)).$$

Schließlich steht in dieser Arbeit

$$\delta_{ij} := \begin{cases} 1, & i = j, \\ 0, & \text{sonst} \end{cases}$$

für das KRONECKER-*delta* und $O$ und $o$ sind die LANDAU-Symbole.

# Kapitel 2

# Analytische Resultate für glatte Anfangswertprobleme

In diesem Abschnitt soll die Frage der Existenz und Eindeutigkeit der Lösung $x(t)$ des (expliziten) Anfangswertproblems

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0 \tag{2.1}$$

untersucht werden[1], wobei die „klassische“ Theorie für *glatte* Probleme vorgestellt wird. Anschließend wird aufgezeigt, warum die präsentierten Beweismethoden für singuläre Probleme versagen. Beweisdetails sind im Anhang A zu finden.

**Bemerkung:** Man kann ohne Beschränkung der Allgemeinheit ausschließlich Gleichungen erster Ordnung betrachten, da sich jedes (explizite) Anfangswertproblem $k$-ter Ordnung ($x, f$ wieder wie oben)

$$x^{(k)}(t) = f(t, x(t), x'(t), \ldots, x^{(k-1)}(t)), \quad x^{(j)}(t_0) = x_{0,j},\ j = 0, \ldots, k-1$$

mittels der Transformation $y_j(t) := x^{(j)}(t),\ j = 0, \ldots, k-1$ auf das $nk$-dimensionale Problem erster Ordnung

$$\begin{aligned}
y_0'(t) &= y_1(t), \\
y_1'(t) &= y_2(t), \\
&\vdots \\
y_{k-2}'(t) &= y_{k-1}(t), \\
y_{k-1}'(t) &= f(t, y_0(t), \ldots, y_{k-1}(t)), \\
y(0) &:= (y_0(0), \ldots, y_{k-1}(0)) = (x_{0,0}, \ldots, x_{0,k-1})
\end{aligned}$$

reduzieren lässt.

[1] Dabei gilt $x : [a, b] \to \mathbb{R}^n, f : G \to \mathbb{R}^n, G \subseteq \mathbb{R}^{n+1}, t_0 \in [a, b], x_0 \in \mathbb{R}^n$.

## 2.1 Der Existenz- und Eindeutigkeitssatz von Picard und Lindelöf

Betrachte Gleichung (2.1) auf einem geeigneten Gebiet $G \subseteq \mathbb{R}^{n+1}$ mit $(t_0, x_0) \in G$. Die Funktion $f(t,x)$ sei stetig von $G$ in $\mathbb{R}^n$ und zusätzlich Lipschitz-stetig bezüglich $x$ (gleichmäßig in $t$), d. h. es gelte

$$|f(t,x) - f(t,y)| \leq L|x-y| \quad \forall (t,x),(t,y) \in G \tag{2.2}$$

mit einer geeigneten Konstante $L$ ($L$ bezeichnet man als Lipschitz-*Konstante*).

Dann gilt der folgende Satz:

**Satz 2.1.1 (Existenz- und Eindeutigkeitssatz von Picard-Lindelöf)** *Sei $f(t,x)$ stetig auf dem kompakten Quader*

$$R := \{(t,x) \in \mathbb{R}^{n+1} : |t-t_0| \leq a, |x-x_0| \leq b\}$$

*und dort Lipschitz-stetig bezüglich $x$ mit Konstante $L$. Sei $M$ eine Schranke für $f$ auf $R$, es gelte also*

$$|f(t,x)| \leq M \quad \forall (t,x) \in R.$$

*Dann hat das Anfangswertproblem*

$$x'(t) = f(t,x(t)), \quad x(t_0) = x_0$$

*eine eindeutige Lösung $x(t)$ auf dem Intervall $J := [t_0 - \alpha, t_0 + \alpha]$ mit $\alpha := \min\{a, \frac{b}{M}\}$, die iterativ gewonnen werden kann als $x(t) = \lim_{n\to\infty} \varphi_n(t)$ mit*

$$\varphi_n(t) := x_0 + \int_{t_0}^{t} f(\tau, \varphi_{n-1}(\tau))\, d\tau, \tag{2.3}$$

*wenn der Startwert $\varphi_0(t)$ aus einem geeigneten Bereich gewählt ist. Die Folge $(\varphi_n(t))_{n\in\mathbb{N}}$ konvergiert gleichmäßig gegen $x(t)$. Weiters gelten die Abschätzungen*

$$|x(t) - \varphi_n(t)| \leq \frac{(\alpha L)^n}{n!} \mathrm{e}^{\alpha L} \max_{\tau\in J} |\varphi_1(\tau) - \varphi_0(\tau)|, \quad t \in J, \tag{2.4}$$

$$|x(t) - \varphi_n(t)| \leq \frac{M}{L} \sum_{\nu=n+1}^{\infty} \frac{(L|t-t_0|)^\nu}{\nu!} + \mu \sum_{\nu=n}^{\infty} \frac{(L|t-t_0|)^\nu}{\nu!}, \quad t \in J, \tag{2.5}$$

$$\mu := \max_{\tau\in J} |\varphi_0(\tau) - x_0|.$$

*Beweis:* Siehe Satz A.1.4 in Abschnitt A.1 des Anhangs.

## 2.2 Der Existenzsatz von Peano

Im Anhang A.1 wird gezeigt, dass es mehrere verschiedene Lösungen eines Anfangswertproblems

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0$$

geben kann, wenn die Funktion $f(t, x)$ zwar stetig aber nicht Lipschitz-stetig bezüglich $x$ ist. Die Existenz mindestens einer Lösung ist in diesem Fall jedoch gesichert, wie der folgende Satz zeigt.

**Satz 2.2.1 (Existenzsatz von Peano)** *Sei $f(t, x)$ stetig auf dem kompakten Quader*

$$R := \{(t, x) \in \mathbb{R}^{n+1} : |t - t_0| \leq a, |x - x_0| \leq b\},$$

*und sei $M$ eine Schranke für $f$ auf $R$, gelte also*

$$|f(t, x)| \leq M \quad \forall (t, x) \in R.$$

*Dann hat das Anfangswertproblem*

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0$$

*(mindestens) eine Lösung $x(t)$ auf dem Intervall $J := [t_0 - \alpha, t_0 + \alpha]$ mit $\alpha := \min\{a, \frac{b}{M}\}$.*

*Beweis:* Siehe Satz A.1.5 in Anhang A.1.

**Bemerkung:** Das Intervall $J$, auf dem die Aussage Gültigkeit besitzt, ist dasselbe wie im Existenz- und Eindeutigkeitssatz von Picard-Lindelöf 2.1.1.

## 2.3 Fortsetzung der Lösung

Bei den Resultaten der beiden vorangegangenen Abschnitte 2.1 und 2.2 handelt es sich um *lokale* Aussagen, die nur in einer (eventuell sehr kleinen) Umgebung des betrachteten Anfangswerts Gültigkeit besitzen. Dieser Abschnitt soll das *globale* Verhalten von Lösungen beleuchten.

**Definition 2.3.1** *Sei $I = (a, b)$ ein offenes Intervall, $G \subseteq \mathbb{R}^{n+1}$ ein Gebiet, $f \in C(G, \mathbb{R}^n)$ und $x$ eine Lösung der Differentialgleichung*

$$x'(t) = f(t, x(t)), \quad t \in I. \tag{2.6}$$

*Dann heißt $I$* rechtsmaximales Existenzintervall von *$x$, wenn es keine Fortsetzung von $x$ auf ein Intervall $J = (\tilde{a}, \tilde{b})$ mit folgenden Eigenschaften gibt:*

1. *$I$ ist in $J$ enthalten.*

2. *$I$ und $J$ haben unterschiedliche rechte Randpunkte.*

*Analog definiert man ein* linksmaximales Existenzintervall. *Ist $I$ sowohl links- als auch rechtsmaximales Existenzintervall von $x$, so heißt $I$* maximales Existenzintervall *der Lösung $x$.*

*Man sagt,* die auf dem Intervall $I$ definierte Lösung $x$ kommt dem Rand von $G$ rechts (beziehungsweise links) beliebig nahe*, wenn es zu jedem Kompaktum $K \subset G$ einen Punkt $t_K \in I$ gibt mit $t_K \geq t_0$ (bzw. $t_K \leq t_0$) und $(t_K, x(t_K)) \notin K$, wobei $t_0$ irgendein Punkt aus $I$ mit $(t_0, x(t_0)) \in K$ ist.*

Der Existenzsatz von Peano 2.2.1 besagt, dass das Problem (2.6) für stetiges $f$ *lokal* stets eine Lösung auf einem bestimmten Intervall $J$ besitzt. Es stellt sich nun die Frage, ob eine solche Lösung auf ein größeres Intervall fortgesetzt werden kann. Die Frage kann positiv beantwortet werden, dies ist Gegenstand des folgenden Satzes.

**Satz 2.3.2** *Sei $f \in C(G, \mathbb{R}^n)$ und $x$ eine Lösung von (2.6) auf einem Intervall $J$. Dann gibt es eine Fortsetzung $\tilde{x}$ von $x$ auf ein maximales Existenzintervall $I = (x_-, x_+), -\infty \leq x_- < x_+ \leq \infty$, und $\tilde{x}$ kommt dem Rand von $G$ rechts und links beliebig nahe.*

*Beweis:* Siehe Satz A.1.9 im Anhang.

**Bemerkung:** Da unter den getroffenen Annahmen von Satz 2.3.2 keine *Eindeutigkeit* der Lösung gegeben ist, ist auch das maximale Existenzintervall *$I$ nicht eindeutig.* Stellt man jedoch an $f$ die gleichen Forderungen, wie in Satz 2.1.1, so erhält man analog zu Satz 2.3.2 die Existenz einer *eindeutigen Lösung* und damit auch eines *eindeutigen maximalen Existenzintervalls.*

## 2.4 Singuläre Probleme

Betrachtet man ein *singuläres Problem* der Gestalt

$$x'(t) = \frac{1}{t} M x(t) + t g(t) =: f(t, x(t)), \quad x(0) = x_0,$$

(cf. Kapitel 1), so ist die rechte Seite $f$ im Allgemeinen (außer z. B. für spezielle Gestalt von $M$) nicht *stetig* im Punkt $t = 0$. Außerdem ist wegen

$$|f(t, x) - f(t, y)| \leq \frac{1}{t} |M||x - y|, \quad t \in (0, 1]$$

$f(t, x)$ auch nicht LIPSCHITZ-stetig *gleichmäßig in* $t$ auf $(0, 1]$. Es ist daher nicht zu erwarten, daß für beliebiges $x_0$ eine (eindeutige) Lösung des Anfangswertproblems existiert. Das nächste Kapitel befasst sich eingehend mit der Frage, welche Forderungen an die Matrix $M$ und die Funktion $g$ zu stellen sind und wie die Anfangsbedingung zu wählen ist, damit eine Lösung des Problems existiert, und wie die Glattheitseigenschaften dieser Lösung aussehen.

# Kapitel 3

# Analytische Resultate für singuläre Anfangswertprobleme

Dieses Kapitel befasst sich mit den zentralen analytischen Fragen nach der Existenz, der Eindeutigkeit und den Glattheitseigenschaften der Lösung $y(t)$ des singulären Anfangswertproblems zweiter Ordnung

$$\begin{aligned} y''(t) &= \frac{A_1}{t}y'(t) + \frac{A_0}{t^2}y(t) + f(t, y(t)), \quad t \in (0,1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0, \end{aligned} \tag{3.1}$$

wobei die folgenden Voraussetzungen gelten: Sei $G$ ein geeignetes Gebiet im $\mathbb{R}^{n+1}$, und es gelte für die Gleichung (3.1): $y \in C^2((0,1], \mathbb{R}^n)$ sei eine zweimal stetig differenzierbare $n$-dimensionale Funktion, $f \in C(G, \mathbb{R}^n)$ stetig, $A_0, A_1 \in C([0,1], \mathbb{R}^{n\times n})$ seien stetige (oder sogar konstante) Matrizen, $B_0 \in \mathbb{R}^{m\times n}$ eine konstante Matrix und $\beta \in \mathbb{R}^m$ ein konstanter Vektor. Dabei sei $m \leq 2n$. Gesucht werden Lösungen $y$, die auch in 0 stetig fortsetzbar sind.

Da mit der Variablentransformation $z(t) := (z_1(t), z_2(t)) := (y(t), ty'(t))$ die Gleichung (3.1) in die äquivalente Form

$$\begin{aligned} z'(t) &= \frac{M}{t}z(t) + \overset{\circ}{tf}(t, z(t)), \quad t \in (0,1], \\ z(0) &= z_0, \end{aligned} \tag{3.2}$$

mit

$$M := \begin{pmatrix} 0 & I \\ A_0 & I + A_1 \end{pmatrix}, \quad \overset{\circ}{f}(t, z(t)) := \begin{pmatrix} 0 \\ f(t, z_1(t)) \end{pmatrix}$$

umgeschrieben werden kann[1], werden auch allgemein singuläre Gleichungen er-

[1] Wie $z_0$ aussieht, wird im Weiteren klar.

ster Ordnung der Gestalt

$$\begin{aligned} z'(t) &= \frac{M}{t} z(t) + f(t, z(t)), \quad t \in (0, 1], \\ z(0) &= z_0, \end{aligned} \tag{3.3}$$

untersucht, um durch die dafür erhaltenen Ergebnisse Rückschlüsse auf die Gleichung (3.1) zu ziehen.

Die Untersuchung von (3.3) soll mittels der *formalen* Konstruktion der Lösung dieser Gleichung geschehen (auch wenn sich diese Lösung in der Praxis i. A. *nicht explizit angeben* lässt). Deshalb folgen jetzt allgemeine Resultate zur Theorie linearer Differentialgleichungen (mit variablen Koeffizienten).

## 3.1 Lösungstheorie allgemeiner linearer Differentialgleichungen

Eine allgemeine lineare Differentialgleichung erster Ordnung mit variablen Koeffizienten auf dem Intervall $I$ hat die Form

$$x'(t) = A(t)x(t) + f(t) \tag{3.4}$$

mit *Koeffizientenmatrix* $A : I \to \mathbb{R}^{n\times n}$ und *Inhomogenität* $f : I \to \mathbb{R}^n$. Gilt $f = 0$, dann heißt die Gleichung

$$x'(t) = A(t)x(t) \tag{3.5}$$

*homogen.*

**Bemerkung:** Wie am Anfang von Kapitel 2 bereits ausgeführt, lassen sich alle Differentialgleichungen höherer Ordnung auf Gleichungen erster Ordnung zurückführen, es genügt also stets, solche zu betrachten.

Der folgende Satz zeigt, dass Anfangswertprobleme linearer Differentialgleichungen mit stetigen Datenfunktionen stets eindeutig lösbar sind.

**Satz 3.1.1** *Sei $I$ ein beliebiges Intervall, $A(t) \in C(I, \mathbb{R}^{n\times n})$ eine stetige Matrix und $f(t) \in C(I, \mathbb{R}^n)$ eine stetige (vektorwertige) Funktion auf $I$. Dann ist das Anfangswertproblem*

$$x'(t) = A(t)x(t) + f(t), \quad x(t_0) = x_0 \tag{3.6}$$

*für jedes $t_0 \in I$ und $x_0 \in \mathbb{R}^n$ eindeutig auf ganz $I$ lösbar.*

*Beweis:* Analog wie in Satz A.1.4, siehe [18, S. 501].

Die Lösungen von (3.5) bilden einen linearen Raum, wie im Folgenden gezeigt wird.

**Satz 3.1.2** *Sei $A(t)$ stetig in $I$ und $x_i(t),\ i = 1, \ldots, n$, seien linear unabhängige Lösungen der homogenen Gleichung (3.5). Dann lässt sich jede Lösung $x(t)$ von (3.5) in eindeutiger Weise als (reelle) Linearkombination von $x_1(t), \ldots, x_n(t)$ darstellen.*

*Deshalb wird jede Menge von $n$ linear unabhängigen Lösungen der homogenen Gleichung (3.5) als* Integralbasis *bezeichnet.*

*Beweis:* Siehe [18, S. 504].

Der nächste Satz liefert ein einfaches Kriterium zur Identifikation einer Integralbasis.

**Satz 3.1.3** *Die Koeffizientenmatrix $A(t)$ sei stetig auf $I$. Dann bilden die $n$ Lösungen*

$$x_i(t) = (x_{1i}(t), x_{2i}(t), \ldots, x_{ni}(t)), \quad t \in I, \quad i = 1, \ldots, n$$

*von (3.5) genau dann eine Integralbasis, wenn ihre* Wronski-Determinante

$$W(t) := \begin{vmatrix} x_{11}(t) & x_{12}(t) & \cdots & x_{1n}(t) \\ x_{21}(t) & x_{22}(t) & \cdots & x_{2n}(t) \\ \vdots & & & \vdots \\ x_{n1}(t) & x_{n2}(t) & \cdots & x_{nn}(t) \end{vmatrix}$$

*zumindest für ein $t_0 \in I$ nicht verschwindet. In diesem Fall gilt $W(t) \neq 0\ \forall t \in I$.*

*Beweis:* Siehe [18, S. 505].

Im Folgenden wird eine ganz spezielle Integralbasis eingeführt, die sich später als nützlich erweisen wird.

Sei $t_0 \in I$ ein beliebiger Punkt, und bezeichne

$$e_i := (\delta_{1i}, \ldots, \delta_{ni}), \quad i = 1, \ldots, n$$

den $i$-ten Einheitsvektor im $\mathbb{R}^n$. Dann gibt es nach Satz 3.1.1 genau eine Lösung $x_i(t)$ der homogenen Gleichung (3.5) mit

$$x_i(t_0) = e_i, \quad i = 1, \ldots, n.$$

Sei nun $x(t)$ irgendeine Lösung von (3.5) und gelte

$$x(t_0) = \sum_{i=1}^{n} c_i e_i$$

mit reellen Koeffizienten $c_i$. Dann ist die Funktion

$$y(t) := \sum_{i=1}^{n} c_i x_i(t)$$

offensichtlich eine Lösung von (3.5) mit $y(t_0) = x(t_0)$, und deshalb gilt nach Satz 3.1.1 $y = x$.

Diese Integralbasis erweist sich also als besonders praktisch, da man die Lösung von

$$x'(t) = A(t)x(t), \quad x(t_0) = x_0$$

nun als

$$x(t) = \Phi(t, t_0)x_0$$

mit der WRONSKI-*Matrix*

$$\Phi(t, t_0) := (x_1(t), \ldots, x_n(t))$$

schreiben kann ($x_i$ bezeichnen die soeben eingeführten Basislösungen).

**Bemerkung:** Die oben angegebene Matrix $\Phi$ erfüllt also das *Matrizen-Anfangswertproblem*

$$\Phi'(t, t_0) = A(t)\Phi(t, t_0), \quad \Phi(t_0, t_0) = I. \tag{3.7}$$

Im Weiteren wird untersucht, wie die Lösungen des *inhomogenen* Systems (3.4) aussehen. Für den Lösungsraum gilt:

**Satz 3.1.4** *Man erhält alle Lösungen des inhomogenen Systems (3.4) — und nur diese —, indem man zu irgendeiner festen (*partikulären*) Lösung von (3.4) alle Lösungen des dazugehörigen homogenen Systems (3.5) addiert.*

*Beweis:* Siehe [18, S. 502].

Mittels der *Variation der Konstanten* kann man die allgemeine Lösung der inhomogenen Gleichung (3.4) auch „explizit“ angeben.

**Satz 3.1.5** *Sei $x_i,\ i = 1, \ldots, n$ eine Integralbasis des zur Gleichung (3.4) gehörigen homogenen Problems (3.5) und $\Phi(t, t_0)$ ihre* WRONSKI-*Matrix. Dann erhält man eine Partikulärlösung von (3.4) aus*

$$x_{\mathrm{p}}(t) = \Phi(t, t_0) \int_{t_0}^{t} \Phi^{-1}(\tau, t_0) f(\tau)\, d\tau.$$

*Beweis:* Siehe [18, S. 510].

**Bemerkung:** Man beachte, dass diese partikuläre Lösung sinnvoll definiert ist, da die WRONSKI-Matrix laut 3.1.3 überall regulär ist.

Zusammenfassend gilt also für die Lösung eines Anfangswertproblems einer linearen Differentialgleichung

**Korollar 3.1.6** *Sei $x_i(t),\ i = 1, \ldots, n$ die Integralbasis von (3.5), für die $x_i(t_0) = e_i,\ i = 1, \ldots, n$, gilt, und $\Phi(t, t_0)$ ihre* WRONSKI-*Matrix. Dann ist*

$$x(t) := \Phi(t, t_0)x_0 + \Phi(t, t_0) \int_{t_0}^{t} \Phi^{-1}(\tau, t_0) f(\tau)\, d\tau$$

*die (eindeutige) Lösung des Anfangswertproblems*

$$x'(t) = A(t)x(t) + f(t), \quad x(t_0) = x_0.$$

*Beweis:* Satz 3.1.4 mit Satz 3.1.2 und Satz 3.1.5.

## 3.2 Lösungsdarstellung für lineare singuläre Anfangswertprobleme mit konstanter Koeffizientenmatrix

Gemäß der Resultate von Abschnitt 3.1 muss man also eine Integralbasis (und damit eine WRONSKI-Matrix) eines homogenen linearen Systems kennen, um (formal) die Lösung des inhomogenen Anfangswertproblems anschreiben zu können. Dies soll im Folgenden für die singuläre Anfangswertaufgabe (3.1) über den „Umweg“ von (3.3) geschehen, um im Anschluss ausgehend von dieser Lösungsdarstellung die gewünschten Resultate ableiten zu können. Man beginnt dabei mit dem Problem (3.3) mit *konstanter* Koeffizientenmatrix $M$, um dann die Resultate auf den allgemeinen linearen und den nichtlinearen Fall zu erweitern. Die Überlegungen sind [31] entnommen, und sind zum Teil auch in [33] detailliert ausgeführt.

### 3.2.1 Gleichungen erster Ordnung

In diesem Abschnitt wird das Anfangswertproblem

$$\begin{aligned} z'(t) &= \frac{M}{t} z(t) + f(t), \quad t \in (0,1], && (3.8)\\ z(0) &= z_0, && (3.9) \end{aligned}$$

mit $M \in \mathbb{R}^{n\times n}$ untersucht. Man beginnt dabei mit der Analyse der Lösungsstruktur der Gleichung

$$z'(t) = \frac{M}{t} z(t) + f(t), \quad t \in (0,1]. \tag{3.10}$$

Um eine WRONSKI-Matrix der zu (3.10) gehörenden homogenen Gleichung herzuleiten, wird die Gleichung zunächst so transformiert, dass die Koeffizientenmatrix JORDAN-Normalform (cf. Satz B.2.2) besitzt.

Sei also $J$ eine Matrix in JORDAN-Normalform und $E$ eine Transformationsmatrix, sodass gilt

$$M = EJE^{-1}.$$

Mit der neuen Variable $v(t) := E^{-1}z(t)$ und der Definition $g(t) := E^{-1}f(t)$ erhält man aus (3.10) die Gleichung

$$v'(t) = \frac{1}{t} Jv(t) + g(t), \quad t \in (0,1]. \tag{3.11}$$

Aus der Lösung von (3.11) erhält man offensichtlich sofort die Lösung von (3.10).

Der entscheidende Schritt bei der Konstruktion der Lösung ist die Bestimmung einer WRONSKI-Matrix $\Phi$ der homogenen Gleichung

$$v'(t) = \frac{1}{t} Jv(t), \quad t \in (0,1], \tag{3.12}$$

da sich nach Korollar 3.1.6 für jedes $a \in (0,1]$ die Lösung von (3.11) in der Form

$$v(t) = \Phi(t,a)v(a) + \Phi(t,a) \int_a^t \Phi^{-1}(\tau,a) g(\tau)\, d\tau \tag{3.13}$$

angeben lässt, wenn $\Phi$ die WRONSKI-Matrix ist, die das Anfangswertproblem (3.7) (Bezeichnungen sinngemäß) löst.

Nach Satz B.2.2 genügt es offenbar zunächst, nur Systeme zu betrachten, in denen $J = J_m(\lambda)$ eine JORDAN-Matrix gemäß (B.2) ist, da (3.11) *teilweise entkoppelt* ist.

Beginnt man bei der $m$-ten Gleichung

$$v_m'(t) = \frac{\lambda}{t} v_m(t),$$

und setzt jeweils die Lösung der $i$-ten Gleichung sukzessive in die $(i-1)$-te Gleichung ein für $i = m, \ldots, 2$ und löst diese mittels Variation der Konstanten, so sieht man elementar, dass die Matrix

$$\Psi(t) := t^\lambda \begin{pmatrix} 1 & \ln(t) & \frac{\ln(t)^2}{2!} & \cdots & \frac{\ln(t)^{m-1}}{(m-1)!} \\ 0 & 1 & \ln(t) & \cdots & \frac{\ln(t)^{m-2}}{(m-2)!} \\ 0 & 0 & 1 & & \frac{\ln(t)^{m-3}}{(m-3)!} \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \tag{3.14}$$

eine Wronski-Matrix von (3.12) ist. Für eine detaillierte Herleitung siehe [33]. Die Matrix $\Psi$ aus (3.14) lässt sich auch als eine Matrix-Exponentialfunktion auffassen, das ist Gegenstand des nächsten Lemmas.

**Lemma 3.2.1** *Für die Matrix $\Psi$ aus (3.14) gilt*

$$\Psi(t) = \exp(\ln(t)J) := \sum_{i=0}^{\infty} \frac{1}{i!} (\ln(t)J)^i.$$

*Beweis:* Durch vollständige Induktion, siehe [33].

**Bemerkung:** In Analogie zur reellen Exponentialfunktion wird künftig oft die Abkürzung

$$t^J := \exp(\ln(t)J)$$

verwendet.

Setzt man nun

$$\Phi(t, a) := \Psi(t)\Psi^{-1}(a),$$

dann ist offensichtlich $\Phi$ die gesuchte Lösung von (3.7). Für die folgenden Überlegungen beginnt man einfachheitshalber bei $a = 1$, da gilt $\Phi(t, 1) = \Psi(t)$. Man erhält somit die allgemeine Lösung von (3.11) als

$$v(t) = \Psi(t) \left( \tilde{c} + \int_1^t \Psi^{-1}(\tau) g(\tau)\, d\tau \right).$$

Durch Rücktransformation erhält man für die allgemeine Lösung $z$ von (3.10)

$$z(t) = t^M \left( c + \int_1^t \tau^{-M} f(\tau)\, d\tau \right). \tag{3.15}$$

Für diese gilt offensichtlich $z(1) = c$.

**Bemerkung:** Man sieht sofort, dass sich im Allgemeinen die soeben konstruierte Lösung *nicht* stetig in 0 fortsetzen lässt, das war jedoch eine der Forderungen an die Lösung des Anfangswertproblems (3.8) und (3.9). Es werden also (abhängig vom Spektrum von $M$) nur gewisse Anfangsbedingungen in Betracht kommen, damit die Stetigkeitsforderung erfüllt ist. Dies ist der Kernpunkt der Überlegungen der späteren Abschnitte.

Zunächst soll jedoch noch ausgehend von (3.15) die Lösung des Problems zweiter Ordnung konstruiert werden.

### 3.2.2 Gleichungen zweiter Ordnung

In diesem Abschnitt wird das Anfangswertproblem

$$\begin{aligned} y''(t) &= \frac{A_1}{t}y'(t) + \frac{A_0}{t^2}y(t) + f(t), \quad t \in (0,1], && (3.16)\\ B_0 y(0) &= \beta, && (3.17)\\ A_0 y(0) &= 0, \quad y'(0) = 0 && (3.18) \end{aligned}$$

mit $A_0, A_1 \in \mathbb{R}^{n\times n}$ untersucht. Dazu transformiert man wieder mittels $z(t) := (z_1(t), z_2(t)) := (y(t), ty'(t))$ die Gleichung (3.16) in die äquivalente Form

$$\begin{aligned} z'(t) &= \frac{M}{t}z(t) + \overset{\circ}{f}(t), \quad t \in (0,1], && (3.19)\\ z(0) &= z_0, && (3.20) \end{aligned}$$

mit

$$M := \begin{pmatrix} 0 & I \\ A_0 & I + A_1 \end{pmatrix}, \quad \overset{\circ}{f}(t) := \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.$$

Man sieht sogleich, dass sich die Anfangsbedingung $B_0 y(0) = \beta$ zu $B_0 z_1(0) = \beta$ transformiert. Was das für $z_0$ bedeutet, wird im Weiteren klar.

Genauso wie in Abschnitt 3.2.1 (cf. (3.15)) leitet man die Lösungsdarstellung

$$z(t) = t^M \left( c + \int_1^t \tau^{-M} \tau \overset{\circ}{f}(\tau)\, d\tau \right) \tag{3.21}$$

mit $z(1) = c \in \mathbb{R}^{2n}$ beliebig, her. Daraus erhält man jetzt direkt

$$y(t) = z_1(t), \quad y'(t) = z_1'(t) = \frac{z_2(t)}{t}.$$

Im Folgenden erweist es sich als praktisch, (3.21) mittels der Variablentransformation

$$\tau \mapsto s = \frac{\tau}{t}$$

umzuschreiben zu

$$\begin{aligned} z(t) &= t^M \left( c - \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau)\, d\tau \right) + t^M \int_0^t \tau\tau^{-M} \overset{\circ}{f}(\tau)\, d\tau \\ &= t^M c^\star + t^2 \int_0^1 s s^{-M} \overset{\circ}{f}(st)\, ds. \end{aligned}$$

## 3.3 Existenz, Eindeutigkeit und Glattheitseigenschaften der Lösung von linearen Gleichungen zweiter Ordnung mit konstanten Koeffizienten

Man sieht an der Gleichung (3.21) unschwer, dass es von den Eigenwerten $\lambda_i = \sigma_i + \mathrm{i}\tau_i,\ i = 1, \ldots, 2n$, (und zwar vom Vorzeichen ihres Realteils $\sigma_i$) der Matrix $M$ abhängt, ob Lösungen von (3.1) stetig (beschränkt) in 0 fortsetzbar sind. Gegebenenfalls kommen nur bestimmte Werte der Konstante $c$ in (3.21) in Frage, um die Stetigkeit von $y$ zu gewährleisten. In diesem Abschnitt sollen die Anfangsbedingungen, die an der Stelle 0 gestellt werden müssen bzw. können, um eine stetige, eindeutige Lösung von (3.1) zu erhalten, angegeben werden[2]. Dazu werden zunächst folgende drei Fälle unterschieden:

1. Alle Eigenwerte von $M$ haben negativen Realteil.
2. Alle Eigenwerte von $M$ sind gleich 0.
3. Alle Eigenwerte von $M$ haben positiven Realteil.

Nicht betrachtet wird der Fall, dass $M$ rein imaginäre Eigenwerte ungleich 0 besitzt, denn diese liefern in $t^M$ Beiträge der Form $\cos(\tau_i \ln(t)) + \mathrm{i}\sin(\tau_i \ln(t))$.

Der allgemeine Fall, dass $M$ Eigenwerte mit verschiedenem Vorzeichen des Realteils besitzt, wird im Anschluss daran analysiert.

### 3.3.1 Systeme mit Eigenwerten mit negativem Realteil

In diesem Fall reicht schon die Forderung nach der Stetigkeit der Lösung in 0 aus, um eine eindeutige Lösung von (3.16) zu erhalten. Die *homogenen* Anfangsbedingungen $y(0) = y'(0) = 0$ sind *notwendig und hinreichend* dafür, dass $y \in C([0,1], \mathbb{R}^n)$ gilt (siehe dazu [31]). Diese Tatsache wird im folgenden Lemma formuliert.

[2]Die Formulierungen der Anfangswertprobleme sind stets so, dass die allgemeinsten Anfangsbedingungen, die notwendig und hinreichend für die Stetigkeit der Lösung sind, vorgeschrieben werden.

**Lemma 3.3.1** *Angenommen, $M$ hat nur Eigenwerte mit negativem Realteil. Für jedes $f \in C^p([0,1],\mathbb{R}^n), p \geq 0$, gibt es eine eindeutige stetige Lösung $z(t)$ von (3.19). Diese erfüllt die Anfangsbedingung $z(0) = 0$ und hat die Gestalt*

$$z(t) = t^2 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau.$$

*Weiters ist $y(t) := z_1(t)$ mit*

$$\begin{aligned} y(t) &= t^2 I_1 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau, \\ y'(t) &= t I_2 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau, \end{aligned}$$

*$y \in C^{p+2}([0,1],\mathbb{R}^n)$, die eindeutige Lösung der Anfangswertaufgabe*

$$\begin{aligned} y''(t) &= \frac{A_1}{t} y'(t) + \frac{A_0}{t^2} y(t) + f(t), \quad t \in (0,1], \\ y(0) &= 0, \quad y'(0) = 0, \end{aligned}$$

*und es gelten die Abschätzungen*

$$\begin{aligned} |y(t)| &\leq \text{const.} t^2 \|f\|_\infty, \\ |y'(t)| &\leq \text{const.} t \|f\|_\infty, \\ |y''(t)| &\leq \text{const.} \|f\|_\infty. \end{aligned}$$

*Beweis:* Siehe [31].

**Bemerkung:** Wegen $z'(t) = (y'(t), y'(t) + ty''(t))$ für $y \in C^2([0,1],\mathbb{R}^n)$ und $y(0) = y'(0) = 0$ gilt auch $z'(0) = 0$.

### 3.3.2 Systeme mit Eigenwert 0

Bezeichne im Weiteren $X_0$ den Eigenraum zum Eigenwert $\lambda = 0$ von $M$ und $R$ die Projektion auf $X_0$ entlang des orthogonalen Komplements von $X_0$[3]. Aus der Definition von $R$ folgt, dass die Abbildung $H := I - R$ die Projektion auf den von den Hauptvektoren zum Eigenwert 0 aufgespannten Unterraum des $\mathbb{R}^{2n}$ ist. Habe $R$ den Rang $r \leq 2n$, dann bezeichne $\tilde{R} \in \mathbb{R}^{2n \times r}$ die Matrix der linear unabhängigen Spalten von $R$.

In diesem Fall ist es notwendig, gewisse Anfangsbedingungen an der Stelle 0 vorzuschreiben, um eine eindeutige Lösung zu erhalten. Die Anfangsbedingungen

[3] $R$ wird auch *Spektralprojektion* genannt.

müssen jedoch bestimmte Bedingungen erfüllen, damit die Lösung auch stetig ist. Die im nächsten Lemma formulierten Bedingungen sind also *notwendig und hinreichend* für die Existenz einer eindeutig bestimmten stetigen Lösung von (3.16) unter der Annahme, dass alle Eigenwerte von $M$ gleich 0 sind (cf. [31]).

**Lemma 3.3.2** *Angenommen, alle Eigenwerte von $M$ sind gleich 0. Für jedes $f \in C^p([0,1],\mathbb{R}^n), p \geq 0$, und $\gamma \in \mathcal{R}(R)$ gibt es eine eindeutige stetige Lösung $z(t)$ von (3.19) mit*

$$z(t) = \gamma + t^2 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau.$$

*Diese Lösung erfüllt $z(0) = \gamma$ und $z'(0) = 0$, und dies ist die allgemeinste Anfangsbedingung, die zu einer stetigen Lösung $z$ führt. Die Funktion $y(t) := z_1(t)$ mit*

$$\begin{aligned} y(t) &= I_1\gamma + t^2 I_1 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau, \\ y'(t) &= tI_2 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau, \end{aligned}$$

*$y \in C^{p+2}([0,1],\mathbb{R}^n)$, ist die (eindeutige) Lösung der Anfangswertaufgabe*

$$\begin{aligned} y''(t) &= \frac{A_1}{t} y'(t) + \frac{A_0}{t^2} y(t) + f(t), \quad t \in (0,1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0, \end{aligned}$$

*falls $m = r$ gilt, die Matrix $B_0 I_1 \tilde{R}$ regulär ist und $B_0 I_1 \gamma = \beta$ gilt. Weiters gelten die folgenden Abschätzungen:*

$$\begin{aligned} |y(t)| &\leq \text{const.} t^2 \|f\|_\infty + |\tilde{R}(B_0 I_1 \tilde{R})^{-1}\beta|, \\ |y'(t)| &\leq \text{const.} t \|f\|_\infty, \\ |y''(t)| &\leq \text{const.} \|f\|_\infty. \end{aligned}$$

*Beweis:* Siehe [31].

### 3.3.3 Systeme mit Eigenwerten mit positivem Realteil

Für Systeme, in denen $M$ nur Eigenwerte mit positivem Realteil besitzt, sind *alle Lösungen der Gestalt (3.21)* stetig in 0. Um eine Lösung eindeutig festzulegen bedarf es also $2n$ Zusatzbedingungen. Wie das folgende Lemma zeigt, ist es nur möglich, diese in einem Punkt $a > 0$ festzulegen, da *alle* Lösungen in 0 verschwinden, dort also keine (anderen) Anfangsbedingungen vorgeschrieben werden können (cf. [31]).

Im Folgenden bezeichne $\sigma_+$ den kleinsten Realteil der Eigenwerte von $M$ und $n_{\max}$ die Dimension der größten JORDAN-Matrix, die in der Darstellung gemäß Satz B.2.2 auftritt.

**Lemma 3.3.3** *Für jedes $f \in C^p([0,1],\mathbb{R}^n), p \geq 0$, und jedes $c \in \mathbb{R}^{2n}$ gibt es eine eindeutige stetige Lösung $z(t)$ von (3.19) mit*

$$z(t) = t^M c + t^M \int_1^t \tau\tau^{-M} \overset{\circ}{f}(\tau)\, d\tau,$$

*$z(1) = c$. Weiters ist $y(t) := z_1(t)$ mit*

$$\begin{aligned} y(t) &= I_1 t^M \left( c + \int_1^t \tau\tau^{-M} \overset{\circ}{f}(\tau)\, d\tau \right), \\ y'(t) &= I_2 t^{M-I} \left( c + \int_1^t \tau\tau^{-M} \overset{\circ}{f}(\tau)\, d\tau \right), \end{aligned}$$

*$y \in C([0,1],\mathbb{R}^n) \cap C^{p+2}((0,1],\mathbb{R}^n)$, eine Lösung von (3.16) und es gelten die Abschätzungen*

$$\begin{aligned} |y(t)| &\leq \begin{cases} \text{const.} t^{\sigma_+}(1+|\ln(t)|^{n_{\max}-1})(|c|+\|f\|_\infty), & \sigma_+ < 2, \\ \text{const.} t^2(1+|\ln(t)|^{n_{\max}})(|c|+\|f\|_\infty), & \sigma_+ = 2, \\ \text{const.} t^2(|c|+\|f\|_\infty), & \sigma_+ > 2, \end{cases} \\ |y'(t)| &\leq \begin{cases} \text{const.} t^{\sigma_+-1}(1+|\ln(t)|^{n_{\max}-1})(|c|+\|f\|_\infty), & \sigma_+ < 2, \\ \text{const.} t(1+|\ln(t)|^{n_{\max}})(|c|+\|f\|_\infty), & \sigma_+ = 2, \\ \text{const.} t(|c|+\|f\|_\infty), & \sigma_+ > 2, \end{cases} \\ |y''(t)| &\leq \begin{cases} \text{const.} t^{\sigma_+-2}(1+|\ln(t)|^{n_{\max}-1})(|c|+\|f\|_\infty), & \sigma_+ < 2, \\ \text{const.} (1+|\ln(t)|^{n_{\max}})(|c|+\|f\|_\infty), & \sigma_+ = 2, \\ \text{const.} (|c|+\|f\|_\infty), & \sigma_+ > 2. \end{cases} \end{aligned}$$

*Beweis:* Siehe [31].

**Bemerkung:** Aus den Abschätzungen in Lemma 3.3.3 erkennt man, dass für jedes $c \in \mathbb{R}^{2n}$ $y(0) = 0$ gilt. Deshalb kann man in 0 keine Anfangswerte vorschreiben, die zu einer eindeutigen stetigen Lösung führen würden. Aus diesem Grund wird in allen weiteren Überlegungen in dieser Arbeit der Fall, dass $M$ Eigenwerte mit positivem Realteil besitzt, ausgeklammert, da er zu keinem sinnvoll gestellten CAUCHYschen Anfangswertproblem der Form (3.16) – (3.18) führt.

### 3.3.4 Allgemeine Systeme

In diesem Abschnitt werden die Resultate der vorangegangenen Abschnitte 3.3.1 – 3.3.3 zusammenfassend auf Systeme mit allgemeinem Spektrum angewendet.

Wie aus Abschnitt 3.3.3 hervorgeht, braucht man nur Systemmatrizen $M$ mit Eigenwerten mit negativem Realteil und mit Eigenwert 0 zu betrachten, da das Anfangswertproblem in 0 nicht sinnvoll gestellt werden kann, wenn $M$ Eigenwerte mit positivem Realteil besitzt.

Der folgende Satz enthält die notwendigen und hinreichenden Bedingungen für die Existenz einer eindeutigen stetigen Lösung.

**Satz 3.3.4** *Angenommen, $M$ habe nur Eigenwerte mit negativem Realteil oder den Eigenwert 0. Für jedes $f \in C^p([0,1],\mathbb{R}^n), p \geq 0$, und $\gamma \in \mathcal{R}(R) \subseteq \mathbb{R}^{2n}$ gibt es eine eindeutige, stetige Lösung $z(t)$ von (3.19) mit*

$$z(t) = \gamma + t^2 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau,$$

*die $z(0) = \gamma$ und $z'(0) = 0$ erfüllt, und diese Anfangsbedingung ist die allgemeinste, die zu einer stetigen Lösung $z$ führt. Ist $m = r$, $B_0 \in \mathbb{R}^{m\times n}, B_0 I_1 \tilde{R}$ regulär, $\beta \in \mathbb{R}^m$ und $\gamma = \tilde{R}(B_0 I_1 \tilde{R})^{-1}\beta$, so ist $y(t) := z_1(t)$ mit*

$$\begin{aligned} y(t) &= I_1\tilde{R}(B_0 I_1 \tilde{R})^{-1}\beta + t^2 I_1 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau, \\ y'(t) &= tI_2 \int_0^1 \tau\tau^{-M} \overset{\circ}{f}(\tau t)\, d\tau, \end{aligned}$$

*$y \in C^{p+2}([0,1],\mathbb{R}^n)$, die Lösung der Anfangswertaufgabe*

$$\begin{aligned} y''(t) &= \frac{A_1}{t} y'(t) + \frac{A_0}{t^2} y(t) + f(t), \quad t \in (0,1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0. \end{aligned}$$

*Es gelten die folgenden Abschätzungen:*

$$\begin{aligned} |y(t)| &\leq \text{const.} t^2 \|f\|_\infty + |\tilde{R}(B_0 I_1 \tilde{R})^{-1}\beta|, \\ |y'(t)| &\leq \text{const.} t \|f\|_\infty, \\ |y''(t)| &\leq \text{const.} \|f\|_\infty. \end{aligned}$$

*Beweis:* Siehe [31].

## 3.4 Lineare singuläre Probleme mit variabler Koeffizientenmatrix

In diesem Kapitel werden Probleme der Gestalt

$$\begin{aligned} y''(t) &= \frac{A_1(t)}{t}y'(t) + \frac{A_0(t)}{t^2}y(t) + f(t), \quad t \in (0,1], \\ B_0 y(0) &= \beta \\ A_0(0)y(0) &= 0, \quad y'(0) = 0 \end{aligned} \tag{3.22}$$

untersucht. Die Anfangsbedingungen sind die allgemeinsten, für die man noch hoffen kann, eine stetige Lösung zu erhalten, da dies auch für den Spezialfall von konstanten Koeffizientenmatrizen der Fall war. Dabei wird angenommen, dass gilt

$$A_i(t) = A_i(0) + t^\gamma C_i(t), \quad \gamma > 0, \quad i \in \{0,1\} \tag{3.23}$$

mit $C_i(t) \in C([0,1], \mathbb{R}^{n\times n})$. Im Fall $\gamma < 1$ stellte sich in [31] heraus, dass die Glattheitseigenschaften der Lösung unbefriedigend sind. Darum wird im Folgenden ausschließlich $\gamma = 1$ betrachtet.

Manchmal werden an $A_0(t)$ stärkere Forderungen gestellt als an $A_1(t)$, es gelte also

$$A_0(t) = A_0(0) + tA_0'(0) + t^2 D_0(t), \quad D_0 \in C([0,1], \mathbb{R}^{n\times n}).$$

Auch in diesem Fall wird das transformierte Problem erster Ordnung

$$\begin{aligned} z'(t) &= \frac{M}{t}z(t) + \overset{\circ}{C}(t)z(t) + t\overset{\circ}{f}(t), \quad t \in (0,1], \\ z(0) &= z_0, \end{aligned} \tag{3.24}$$

mit

$$\begin{aligned} M &:= \begin{pmatrix} 0 & I \\ A_0(0) & I + A_1(0) \end{pmatrix}, \quad \overset{\circ}{f}(t) := \begin{pmatrix} 0 \\ f(t) \end{pmatrix} \\ \overset{\circ}{C}(t) &:= \begin{pmatrix} 0 & 0 \\ C_0(t) & C_1(t) \end{pmatrix} \end{aligned}$$

untersucht. Im Folgenden wird stets angenommen, dass $M$ die Spektraleigenschaften besitzt, die im letzten Abschnitt als sinnvoll herausgearbeitet wurden. Es gilt

**Satz 3.4.1** *Angenommen, $M$ hat nur Eigenwerte mit negativem Realteil oder den Eigenwert 0. Für jedes $\overset{\circ}{f} \in C^p([0,1], \mathbb{R}^{2n}), \overset{\circ}{C} \in C^p([0,1], \mathbb{R}^{2n\times 2n}), p \geq 0$, und $z_0 \in \mathcal{R}(R)$ gibt es eine eindeutige Lösung $z \in C^{p+1}([0,1], \mathbb{R}^{2n})$ von (3.24). Diese Lösung hat die Gestalt*

$$z(t) = \mathcal{K}[z](t) + \psi(t)$$

*mit*

$$\begin{aligned}\mathcal{K}[z](t) &:= t\int_0^1 s^{-M}\overset{\circ}{C}(st)z(st)\,ds,\\ \psi(t) &:= z_0 + t^2\int_0^1 s s^{-M}\overset{\circ}{f}(st)\,ds.\end{aligned}$$

*$y(t) = I_1 z(t)$ ist eine Lösung von (3.22) genau dann, wenn $A_0'(0)y(0) = 0$ und $z_0 = \tilde{R}(B_0 I_1 \tilde{R})^{-1}\beta$ gilt. Diese Lösung erfüllt $y \in C^{p+2}([0,1],\mathbb{R}^n)$ falls $C_0 \in C^{p+1}([0,1],\mathbb{R}^{n\times n})$. Weiters gibt es ein $\delta > 0$, sodass für $t \in [0,\delta]$ die folgenden Abschätzungen gelten:*

$$\begin{aligned}|y(t)| &\le \text{const.}t^2(|\tilde{R}(B_0 I_1\tilde{R})^{-1}\beta| + \|f\|_\infty) + |\tilde{R}(B_0 I_1\tilde{R})^{-1}\beta|,\\ |y'(t)| &\le \text{const.}t(|\tilde{R}(B_0 I_1\tilde{R})^{-1}\beta| + \|f\|_\infty),\\ |y''(t)| &\le \text{const.}(|\tilde{R}(B_0 I_1\tilde{R})^{-1}\beta| + \|f\|_\infty).\end{aligned}$$

*Beweis:* Siehe [31].

**Bemerkung:** Im Allgemeinen ist die Bedingung $y'(0) = 0$ nicht notwendig für die Stetigkeit der Lösung $y$. Auch eine Lösung die

$$y'(0) = -(A_0(0) + A_1(0))^{-1}C_0(0)y(0)$$

erfüllt ist stetig. Man beachte, dass die Matrix $A_0(0) + A_1(0)$ regulär ist. Siehe dazu [31].

## 3.5 Nichtlineare singuläre Probleme

Dieses Kapitel befasst sich mit Problemen der Gestalt

$$\begin{aligned}y''(t) &= \frac{A_1(t)}{t}y'(t) + \frac{A_0(t)}{t^2}y(t) + f(t,y(t)), \quad t\in(0,1],\\ B_0y(0) &= \beta\\ A_0(0)y(0) &= 0, \quad y'(0) = 0.\end{aligned} \tag{3.25}$$

Diese Anfangsbedingungen sind die allgemeinsten, für die man auf eine stetige Lösung hoffen kann, cf. Abschnitt 3.3. $A_i(t)$ habe wieder die Gestalt (3.23).)

Wieder transformiert man auf die äquivalente Gleichung erster Ordnung

$$\begin{aligned}z'(t) &= \frac{M}{t}z(t) + \overset{\circ}{C}(t)z(t) + t\overset{\circ}{f}(t,z(t)), \quad t\in(0,1],\\ z(0) &= z_0,\end{aligned} \tag{3.26}$$

mit

$$\begin{aligned} M &:= \begin{pmatrix} 0 & I \\ A_0(0) & I + A_1(0) \end{pmatrix}, \quad \overset{\circ}{f}(t) := \begin{pmatrix} 0 \\ f(t, I_1 z(t)) \end{pmatrix} \\ \overset{\circ}{C}(t) &:= \begin{pmatrix} 0 & 0 \\ C_0(t) & C_1(t) \end{pmatrix}. \end{aligned}$$

Um die Existenz einer eindeutigen Lösung von (3.26) und (3.25) zeigen zu können muss $f(t, y)$ in diesem Fall auch noch eine LIPSCHITZ-Bedingung bezüglich $y$ erfüllen[4].

Unter den selben Annahmen über $A_0$ und $A_1$ wie im letzten Abschnitt erhält man das folgende Resultat[5].

**Satz 3.5.1** *Angenommen, $M$ hat nur Eigenwerte mit negativem Realteil oder den Eigenwert 0 und es gilt $z_0 \in \mathcal{R}(R)$. Sei $\overset{\circ}{f} \in C^p([0,1] \times \mathbb{R}^{2n}, \mathbb{R}^{2n})$, $\overset{\circ}{C} \in C^p([0,1], \mathbb{R}^{2n \times 2n})$, $p \geq 0$, und darüberhinaus sei $\overset{\circ}{f}(t, z)$ LIPSCHITZ-stetig bezüglich $z$ auf $[0,1] \times \mathbb{R}^{2n}$.[6] Dann gibt es eine eindeutige Lösung $z \in C^{p+1}([0,1], \mathbb{R}^{2n})$ von (3.26). Diese hat die Gestalt*

$$z(t) = \mathcal{K}^*[z](t) + z_0$$

*mit*

$$\mathcal{K}^*[z](t) := t \int_0^1 \tau^{-M} \left( \overset{\circ}{C}(\tau t) z(\tau t) + \tau t f(\tau t, z(\tau t)) \right) d\tau.$$

*$y(t) = I_1 z(t)$ ist die eindeutige Lösung von (3.25) genau dann, wenn $B_0 I_1 \tilde{R}$ regulär ist und $I_1 z_0 = I_1 \tilde{R} (B_0 I_1 \tilde{R})^{-1} \beta$ und $I_2 z_0 = 0$ gilt. Ist $C_0 \in C^{p+1}([0,1], \mathbb{R}^{n \times n})$, dann gilt $y \in C^{p+2}([0,1], \mathbb{R}^n)$.*

*Beweis:* Siehe [31].

**Bemerkung:** Mittels Taylorentwicklung kann man das lineare Gleichungssystem

$$\left( I - A_1(0) - \frac{A_0(0)}{2} \right) y''(0) = \frac{A_0''(0)}{2} y(0) + f(0, y(0))$$

für $y''(0)$ herleiten. Falls die führende Koeffizientenmatrix dieses Systems regulär ist, kann man also $y''(0)$ berechnen, ohne $y(t)$ explizit zu kennen. Die Kenntnis der ersten Ableitungen der Lösung an der Stelle 0 ist für die numerische Lösung singulärer Probleme von entscheidender Bedeutung.

Es wird nochmals betont, dass dieses Kapitel lediglich Resultate enthält, die für die in dieser Arbeit untersuchten Probleme relevant sind, für eine vollständige Darstellung siehe [31].

---

[4] Cf. (2.2).

[5] Der Fall $A_0, A_1 \in \mathbb{R}^{n \times n}$ ist ein trivialer Spezialfall dieser Situation.

[6] Das gilt genau dann, wenn $f(t, y)$ LIPSCHITZ-stetig bezüglich $y$ auf $[0,1] \times \mathbb{R}^n$ ist.

# Kapitel 4

# Numerische Lösung glatter Anfangswertprobleme

In diesem Kapitel werden einige klassische Einschrittverfahren zur näherungsweisen Lösung von Anfangswertaufgaben gewöhnlicher Differentialgleichungen vorgestellt. Weiters wird die Konvergenzordnung dieser Verfahren für *glatte* Probleme untersucht, um im Anschluss daran aufzuzeigen, wo die dabei verwendeten Beweismethoden versagen, wenn sie auf *singuläre* Gleichungen angewendet werden.

## 4.1 Grundbegriffe

Im Folgenden wird also die numerische Approximation der Lösung $x$ des Anfangswertproblems

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0 \tag{4.1}$$

auf einem Intervall $I = [a, b]$ untersucht. Dabei wird ohne Beschränkung der Allgemeinheit $t_0 = a$ gesetzt, also *nach rechts gerechnet*. Da eine *numerische Lösung* des Problems nur Sinn macht, wenn eine Lösung analytisch existiert und eindeutig ist, werden an $f$ zunächst die Glattheitsforderungen, die sich in Kapitel 2 (Satz 2.1.1) ergeben hatten, auf einem geeigneten Gebiet $G \subseteq \mathbb{R}^{n+1}$ gestellt. Es wird sich aber herausstellen, dass für manche Aussagen noch stärkere Bedingungen an $f$ erforderlich sind.

**Definition 4.1.1 (Gitter)** *Bezeichne $x$ die Lösung der Anfangswertaufgabe (4.1). Sie existiere im Intervall $I = [a, b]$ mit $t_0 := a$ und einem $b > a$. Unter einem* Gitter *auf $I$ versteht man eine Punktmenge der Gestalt*

$$I_{\mathbf{h}} := (t_0, t_1, \ldots, t_N) \text{ mit } a = t_0 < t_1 < \cdots < t_N \leq b.$$

$\mathbf{h} = (h_0, \ldots, h_{N-1})$ *bezeichnet dabei den* Schrittweitenvektor*, wenn man setzt* $h_j := t_{j+1} - t_j, j = 0, \ldots, N-1$*. Ein Gitter bezeichnet man als* äquidistant*, wenn gilt* $h_j =: h = \text{const.},\ j = 0, \ldots, N-1$*. In diesem Fall wird der Einfachheit halber als Index einfach nur* $h$ *anstelle des Schrittweitenvektors* $\mathbf{h} = \{h\}^N$ *verwendet. Bei* nicht äquidistanten *Gittern wird immer* $t_N = b$ *vorausgesetzt, bei einem äquidistanten Gitter sei* $N = \left\lfloor \frac{b-a}{h} \right\rfloor$*. Wenn möglich wird man die tatsächlich verwendeten Schrittweiten aber stets so wählen, dass* $\frac{b-a}{h}$ *ganzzahlig ist.*

Bei den folgenden Betrachtungen wird davon ausgegangen, dass zur Realisierung des Verfahrens jeweils ein festes, bekanntes Gitter $I_{\mathbf{h}}$ verwendet wird, wobei die theoretischen Untersuchungen ausschließlich für äquidistante Gitter durchgeführt werden. In der Praxis wird aus Effizienzgründen die Schrittweite während der Rechnung variabel gehalten, um in jedem Schritt zu prüfen, ob die Schrittweite vergrößert oder verkleinert werden soll. In diesem Fall wird $I_{\mathbf{h}}$ erst *während*, und nicht bereits *vor* der Rechnung festgelegt. Auf diese Fragen wird hier nicht eingegangen werden.

**Definition 4.1.2 (Numerisches Verfahren)** *Unter einem* numerischen Verfahren *zur Approximation der Lösung* $x \in C^1(I, \mathbb{R}^n)$ *des Anfangswertproblems (4.1) versteht man ein Verfahren, das*

1. *ein Gitter* $I_h$ *bestimmt und*
2. *eine* Gitterfunktion $u_h : I_h \to \mathbb{R}^n$ *berechnet mit* $(t, u_h(t)) \in G$ *für jedes* $t \in I_h$*.*

*Häufig wird die Abkürzung* $u_j := u_h(t_j),\ t_j \in I_h$ *verwendet werden, wenn die Bedeutung aus dem Kontext klar ist.*

**Definition 4.1.3 (Konvergenz)** *Sei* $x$ *die exakte Lösung von (4.1). Ein numerisches Verfahren, das zu jedem Gitter* $I_h$ *eine Gitterfunktion* $u_h$ *berechnet, heißt* konvergent für das Anfangswertproblem (4.1) auf dem Intervall $[a, b]$*, wenn für den* globalen Fehler $\varepsilon_h$*, definiert durch*

$$\varepsilon_h(t) := x(t) - u_h(t) \quad \forall t \in I_h$$

*die Beziehung*

$$\lim_{h \to 0} \|\varepsilon_h\|_h := \lim_{h \to 0} \max_{t \in I_h} |u_h(t) - x(t)| = 0$$

*gilt. Das Verfahren hat die* Konvergenzordnung $p > 0$*, wenn gilt*

$$\|\varepsilon_h\|_h = O(h^p) \quad \text{für } h \to 0.$$

Im Folgenden wird eine Klasse von numerischen Verfahren eingeführt, in die alle im Weiteren untersuchten Verfahren einzuordnen sind, um sodann allgemein für diese einige Begriffsbildungen vorzunehmen und Standardresultate zu zeigen.

**Definition 4.1.4 (Einschrittverfahren)** *Ein (*explizites*)* Einschrittverfahren *zur Bestimmung einer Näherungslösung $u_h$ von (4.1) auf einem Gitter $I_h$ hat die Form*

$$\begin{aligned} u_0 &:= u_{h;0}, \\ t_j &:= a + jh, \quad j = 0, \ldots, N, \\ u_j &:= u_{j-1} + h\varphi(t_{j-1}, u_{j-1}, h), \quad j = 1, \ldots, N. \end{aligned}$$

*Dabei heißt $\varphi$* Verfahrensfunktion *des zur Differentialgleichung (4.1) bei der Schrittweite $h$ gehörenden Einschrittverfahrens. Die Vorgabe des* Startwerts $u_{h;0}$ *ist Bestandteil des Verfahrens und kann von $h$ abhängen. Im Normalfall wird man $u_{h;0} := x_0$ setzen.*

Die beiden nächsten Konzepte nehmen eine zentrale Stellung in der Untersuchung von Diskretisierungsverfahren zur Lösung von Differentialgleichungen ein, da sie für glatte Probleme zum Beweis der Konvergenz führen.

**Definition 4.1.5 (Konsistenz)** *Sei $x$ die (*exakte*) Lösung von (4.1). Der* lokale (Diskretisierungs-) Fehler $l_h$ *des Einschrittverfahrens mit Verfahrensfunktion $\varphi$ ist dann definiert als*

$$\begin{aligned} l_0 &:= x_0 - u_{h;0} \\ l_j &:= \frac{x(t_j) - x(t_{j-1})}{h} - \varphi(t_{j-1}, x(t_{j-1}), h), \quad j = 1, \ldots, N. \end{aligned}$$

*Ausgehend davon bezeichnet man ein Einschrittverfahren als* konsistent *mit dem gegebenen Anfangswertproblem, wenn gilt*

$$\lim_{h \to 0} \|l_h\|_h = 0.$$

*Die* Konsistenzordnung $p > 0$ *liegt vor, wenn*

$$\|l_h\|_h = O(h^p) \quad \textit{für } h \to 0$$

*gilt.*

Aus der Konsistenz allein folgt jedoch noch *nicht* die Konvergenz, es ist nämlich darüberhinaus noch sicherzustellen, dass die Akkumulation der lokalen Fehler nicht so stark ist, dass die Konvergenz dadurch verhindert wird. Mit dem Konzept der *Stabilität*, wie sie im Folgenden formuliert wird, ist es in vielen Fällen möglich, dies zu gewährleisten, es wird sich jedoch zeigen, dass diese *lokale* Eigenschaft für singuläre Probleme nicht ausreicht, um Konvergenz zu zeigen, und ein *globales* Stabilitätskonzept nötig ist.

**Definition 4.1.6 (Stabilität)** *Betrachte zwei „parallele“ Schritte eines Einschrittverfahrens:*

$$\begin{aligned}(t_{j-1}, u_{j-1}) &\mapsto (t_j, u_j),\\ (t_{j-1}, \tilde{u}_{j-1}) &\mapsto (t_j, \tilde{u}_j).\end{aligned}$$

*Dann heißt das Verfahren* stabil*, wenn es ein $h_0 > 0$ gibt, sodass die Ungleichung*

$$|u_j - \tilde{u}_j| \leq (1 + Sh)|u_{j-1} - \tilde{u}_{j-1}|, \quad j = 1, \ldots, N,$$

gleichmäßig in $h$ *für $h \leq h_0$ gilt (also mit einer von $h$ unabhängigen Konstante $S$).*

Im Folgenden wird gezeigt, dass aus Konsistenz und Stabilität eines Einschrittverfahrens im Sinne der Definitionen 4.1.5 und 4.1.6 die Konvergenz des Verfahrens folgt.

**Satz 4.1.7 (Konvergenz von Einschrittverfahren)** *Sei $u_h$ die Approximationslösung der Gleichung (4.1), die mittels eines Einschrittverfahrens mit Verfahrensfunktion $\varphi(t, u, h)$ mit der Schrittweite $h$ ermittelt wurde. Für die (gestörten) Startwerte gelte $u_0 = x_0 + \varepsilon_0$. Der lokale Fehler werde mit $l_h$ bezeichnet. Ist das Verfahren stabil, dann gilt für den globalen Fehler $\varepsilon_j := u_j - x(t_j)$ mit einem $L > 0$ die Abschätzung*

$$|\varepsilon_j| \leq \mathrm{e}^{Lhj}|\varepsilon_0| + \frac{\mathrm{e}^{Lhj} - 1}{L} M(h), \quad j = 0, \ldots, N, \tag{4.2}$$

*wobei $M(h)$ eine Schranke für den lokalen Diskretisierungsfehler in dem Sinn darstellt, dass gilt*

$$\max_{1 \leq j \leq N} |l_j| \leq M(h).$$

*Beweis:* Betrachte die „parallelen“ Schritte

$$\begin{aligned}u_{j-1} &\mapsto u_j := u_{j-1} + h\varphi(t_{j-1}, u_{j-1}, h),\\ x(t_{j-1}) &\mapsto \tilde{u}_j := x(t_{j-1}) + h\varphi(t_{j-1}, x(t_{j-1}), h).\end{aligned}$$

Aufgrund der Stabilität des Verfahrens gilt für ein $L > 0$

$$|u_j - \tilde{u}_j| \leq (1 + Lh)|u_{j-1} - x(t_{j-1})|.$$

Außerdem folgt aus der Definition von $\tilde{u}_j$

$$x(t_j) - \tilde{u}_j = hl_j,$$

und damit

$$\begin{aligned} |\varepsilon_j| &\leq |u_j - \tilde{u}_j| + |\tilde{u}_j - x(t_j)| \\ &\leq (1 + Lh)|u_{j-1} - x(t_{j-1})| + h|l_j| \\ &\leq (1 + Lh)|\varepsilon_{j-1}| + hM(h), \quad j = 1, \ldots, N, \end{aligned}$$

mit der obigen Definition von $M(h)$. Damit sind die Bedingungen des diskreten GRONWALL-Lemmas B.1.6 erfüllt und es folgt daraus sofort (4.2).

**Bemerkung:** An der Abschätzung aus Satz 4.1.7 sieht man also, dass unter den getroffenen Annahmen ein konsistentes Verfahren auch konvergent ist, und darüberhinaus die Konsistenzordnung gleich der Konvergenzordnung ist. Dies spiegelt ein allgemeines Prinzip von Diskretisierungsverfahren wider, nämlich, dass aus Konsistenz und Stabilität die Konvergenz folgt (wenn die Begriffe der Aufgabe angemessen eingeführt sind).

Das nächste Lemma gibt eine einfache Bedingung für die Stabilität eines Einschrittverfahrens an.

**Lemma 4.1.8** *Ist die Verfahrensfunktion $\varphi(t, u, h)$ eines Einschrittverfahrens gemäß Definition 4.1.4* LIPSCHITZ*-stetig bezüglich $u$ (gleichmäßig in den anderen Variablen), dann ist das Verfahren stabil.*

*Beweis:* Sei die Lipschitzkonstante gleich $L$, dann folgt sofort die Ungleichung aus Definition 4.1.6 mit $S = L$.

In den folgenden Abschnitten werden jene numerischen Verfahren vorgestellt, die in dieser Arbeit untersucht werden sollen.

## 4.2 Das explizite EULER-Verfahren

**Definition 4.2.1 (Explizites EULER-Verfahren)** *Das* explizite EULERverfahren *zur Lösung des Anfangswertproblems (4.1) ist definiert durch*

$$\begin{aligned} u_0 &:= u_{h;0}, \\ u_j &:= u_{j-1} + hf(t_{j-1}, u_{j-1}), \quad j = 1, \ldots, N. \end{aligned}$$

Es handelt sich also um ein explizites Einschrittverfahren mit Verfahrensfunktion $f$. Ist $f$ LIPSCHITZ-stetig (also die Existenz und Eindeutigkeit der Lösung von (4.1) gesichert, cf. Satz 2.1.1), dann ist das Verfahren nach Lemma 4.1.8 stabil. Es ist also nur noch die Konsistenz zu zeigen.

**Lemma 4.2.2** *Sei die Lösung $x$ des Anfangswertproblems (4.1) zweimal stetig differenzierbar, dann besitzt das explizite* EULER*-Verfahren für dieses Problem die Konsistenzordnung 1, wenn auch die Startwerte $u_{h;0}$ von erster Ordnung gegen $x_0$ konvergieren.*

*Beweis:* Aus der Taylorentwickung von $x(t_j)$ um $t_{j-1}$ (siehe dazu Satz B.1.7) erhält man

$$\begin{aligned} l_j &:= \frac{x(t_j) - x(t_{j-1})}{h} - f(t_{j-1}, x(t_{j-1})) \\ &= \frac{x(t_j) - x(t_{j-1})}{h} - x'(t_{j-1}) \\ &= h \int_0^1 (1-\tau) x''(t_{j-1} + \tau h)\, d\tau, \quad j = 1, \ldots, N. \end{aligned}$$

Sei $M_2$ eine Schranke für $x''$ auf dem betrachteten Intervall $[a, b]$, dann folgt für den lokalen Fehler

$$\max_{1 \le j \le N} |l_j| \le \frac{h}{2} M_2$$

nach dem erweiterten Mittelwertsatz der Integralrechnung B.1.10. Dieser Satz ist anwendbar, da die Integration komponentenweise erfolgt. Nimmt man noch die Bedingung an die Startwerte hinzu, so folgt die Konsistenzordnung 1 des Verfahrens gemäß Definition 4.1.5.

**Korollar 4.2.3** *Ist die Lösung $x$ von (4.1) in $C^2([a, b], \mathbb{R}^n)$, ist die Funktion $f(t, x)$* LIPSCHITZ*-stetig bezüglich $x$ und konvergieren die Startwerte $u_{h;0}$ von erster Ordnung gegen den exakten Anfangswert $x_0$, dann ist das explizite* EULER*-Verfahren konvergent von erster Ordnung gegen $x$.*

*Beweis:* Satz 4.1.7 unter Verwendung von Lemma 4.1.8 und Lemma 4.2.2.

**Bemerkung:** Normalerweise wählt man für die Startwerte $u_{h;0} := x_0$, wenn dies möglich ist, und hat die an die Startwerte gestellten Bedingungen der vorausgegangenen Überlegungen trivialerweise erfüllt. Es wird sich jedoch herausstellen, dass dies keinesfalls immer möglich oder sinnvoll ist.

## 4.3 Das implizite EULER-Verfahren

**Definition 4.3.1 (Implizites EULER-Verfahren)** *Das* implizite EULERverfahren *zur Lösung des Anfangswertproblems (4.1) ist definiert durch*

$$\begin{aligned} u_0 &:= u_{h;0}, \\ u_j &:= u_{j-1} + h f(t_j, u_j), \quad j = 1, \ldots, N. \end{aligned}$$

Bei diesem Verfahren handelt es sich auf den ersten Blick *nicht* um ein Einschrittverfahren im Sinne von Definition 4.1.4, da das Verfahren *implizit* ist, d. h. dass man zur Berechnung von $u_j$ jeweils eine Gleichung auflösen muss. Im Folgenden wird aber gezeigt, dass sich das Verfahren unter einer gewissen Einschränkung der zulässigen Schrittweiten sehr wohl als Einschrittverfahren auffassen lässt, und damit die vorangegangenen Überlegungen anwendbar sind.

**Satz 4.3.2** *Sei $f(t,x)$* LIPSCHITZ*-stetig bezüglich $x$ mit* LIPSCHITZ*-Konstante $L$, wobei (der Einfachheit halber) $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ gelte (Einschränkungen des Bereichs haben lediglich eine kompliziertere Notation zur Folge). Dann lässt sich das implizite* EULER*-Verfahren für hinreichend kleine Schrittweiten $h$ als Einschrittverfahren gemäß Definition 4.1.4 auffassen, d. h. durch die implizite Form wird eine Verfahrensfunktion $\varphi(t,u,h)$ definiert, die* LIPSCHITZ*-stetig bezüglich $u$ ist.*

*Beweis:* Ein Schritt des impliziten EULER-Verfahrens im Punkt $t$ mit Schrittweite $h$ hat die Gestalt

$$u(t+h) = u(t) + hf(t+h, u(t+h)).$$

Mit der Abkürzung $u := u(t)$ ist dann zur Bestimmung von $u(t+h)$ ein Fixpunkt $v = v(t,u,h)$ der Abbildung

$$T_u : \mathbb{R}^n \to \mathbb{R}^n, \quad v \mapsto u + hf(t+h, v)$$

gesucht. Aus

$$\begin{aligned} |T_u v - T_u w| &= h|f(t+h,v) - f(t+h,w)| \\ &\leq hL|v-w| \end{aligned}$$

folgt die Kontraktionseigenschaft von $T_u$, falls $hL < 1$ gilt (cf. Definition A.1.2). Damit ist der Fixpunktsatz von BANACH A.1.3 anwendbar, es gibt also genau ein $v = v(t,u,h)$ mit $v = T_u v$. Dann ist

$$\varphi(t,u,h) := f(t+h, v(t,u,h))$$

die gesuchte Verfahrensfunktion für das implizite EULER-Verfahren.

Die LIPSCHITZ-Stetigkeit von $\varphi(t,u,h)$ bezüglich $u$ kann folgendermaßen gezeigt werden: Seien $u, \tilde{u} \in \mathbb{R}^n$ und $v := v(t,u,h)$ und $\tilde{v} := v(t,\tilde{u},h)$ die eindeutigen Fixpunkte von $T_u$ bzw. $T_{\tilde{u}}$. Dann folgt

$$\begin{aligned} |v - \tilde{v}| &= |T_u v - T_{\tilde{u}} \tilde{v}| \\ &\leq |u - \tilde{u}| + h|f(t+h,v) - f(t+h,\tilde{v})| \\ &\leq |u - \tilde{u}| + hL|v - \tilde{v}|, \end{aligned}$$

und daher

$$|v(t,u,h) - v(t,\tilde{u},h)| \leq \frac{1}{1-hL}|u-\tilde{u}|,$$

also die LIPSCHITZ-Stetigkeit von $v(t,u,h)$ bezüglich $u$. Hiermit ergibt sich

$$\begin{aligned} |\varphi(t,u,h) - \varphi(t,\tilde{u},h)| &= |f(t+h, v(t,u,h)) - f(t+h, v(t,\tilde{u},h))| \\ &\leq L|v(t,u,h) - v(t,\tilde{u},h)| \\ &\leq \frac{L}{1-hL}|u-\tilde{u}|, \end{aligned}$$

und damit sind alle Behauptungen gezeigt.

Folglich ist die im Abschnitt 4.1 entwickelte Theorie auch für das implizite EULER-Verfahren anwendbar, der Konvergenzbeweis verläuft ähnlich wie der für das explizite EULER-Verfahren.

**Lemma 4.3.3** *Sei die Lösung $x$ des Anfangswertproblems (4.1) zweimal stetig differenzierbar und $f(t,x)$* LIPSCHITZ*-stetig bezüglich $x$ mit Konstante $L$, dann besitzt das implizite* EULER*-Verfahren für dieses Problem die Konsistenzordnung 1, wenn auch die Startwerte $u_{h;0}$ von erster Ordnung gegen $x_0$ konvergieren.*

*Beweis:* Aus der Taylorentwickung von $x(t_{j-1})$ um $t_j$ (siehe dazu Satz B.1.7) erhält man[1]

$$\begin{aligned} |l_j| &= \left|\frac{x(t_j) - x(t_{j-1})}{h} - f(t_j, v(t_{j-1}, x(t_{j-1}), h))\right| \\ &= \left|\frac{1}{h}\left(x(t_j) - x(t_j) + hx'(t_j) - h^2\int_0^1 (1-\tau)x''(t_j - \tau h)\,d\tau\right)\right. \\ &\quad \left. - f(t_j, v(t_{j-1}, x(t_{j-1}), h))\right| \\ &\leq |x'(t_j) - f(t_j, v(t_{j-1}, x(t_{j-1}), h))| + \frac{h}{2}M_2 \\ &= |f(t_j, x(t_j)) - f(t_j, v(t_{j-1}, x(t_{j-1}), h))| + O(h) \\ &\leq L|x(t_j) - v(t_{j-1}, x(t_{j-1}), h)| + O(h) \\ &\leq L|x(t_j) - x(t_{j-1})| + h|f(t_j, v(t_{j-1}, x(t_{j-1}), h))| + O(h) \\ &= Lh\left|\int_0^1 x'(t_j - \tau h)\,d\tau\right| + O(h) = O(h), \quad j = 1,\ldots,N. \end{aligned}$$

Nimmt man noch die Bedingung an die Startwerte hinzu, so folgt die Konsistenzordnung 1 gemäß Definition 4.1.5.

[1] Zur Definition des lokalen Diskretisierungsfehlers siehe 4.1.5, $v(t,u,h)$ ist aus Satz 4.3.2 übernommen. $M_2$ bezeichnet wieder eine Schranke für $x''$ auf $[a,b]$, für die Abschätzung der auftretenden Integrale verwendet man Satz B.1.10 analog wie in Lemma 4.2.2.

**Korollar 4.3.4** *Ist die Lösung $x$ von (4.1) in $C^2([a,b],\mathbb{R}^n)$, ist die Funktion $f(t,x)$* Lipschitz*-stetig bezüglich $x$ und konvergieren die Startwerte $u_{h;0}$ von erster Ordnung gegen den exakten Anfangswert $x_0$, dann ist das implizite* Euler*-Verfahren konvergent von erster Ordnung gegen $x$.*

*Beweis:* Folgt direkt aus Satz 4.1.7 unter Verwendung von Lemma 4.1.8 und Lemma 4.3.3, wenn man Satz 4.3.2 beachtet.

## 4.4 Die Trapezregel

**Definition 4.4.1 (Trapezregel)** *Die* Trapezregel *zur Lösung des Anfangswertproblems (4.1) ist definiert durch*

$$\begin{aligned} u_0 &:= u_{h;0}, \\ u_j &:= u_{j-1} + \frac{h}{2}(f(t_{j-1},u_{j-1}) + f(t_j,u_j)), \quad j = 1,\ldots,N. \end{aligned}$$

Auch hier handelt es sich um ein implizites Verfahren, also muss noch ein dem Satz 4.3.2 entsprechendes Resultat bewiesen werden, um die Theorie von Abschnitt 4.1 verwenden zu können. Das geschieht im folgenden Satz.

**Satz 4.4.2** *Sei $f(t,x)$* Lipschitz*-stetig bezüglich $x$ mit* Lipschitz*-Konstante $L$, wobei $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ gelte. Dann lässt sich die Trapezregel für hinreichend kleine Schrittweiten $h$ als Einschrittverfahren gemäß Definition 4.1.4 auffassen, d. h. durch die implizite Form wird eine Verfahrensfunktion $\varphi(t,u,h)$ definiert, die* Lipschitz*-stetig bezüglich $u$ ist.*

*Beweis:* Ein Schritt der Trapezregel im Punkt $t$ mit Schrittweite $h$ hat die Gestalt

$$u(t+h) = u(t) + \frac{h}{2}(f(t,u(t)) + f(t+h,u(t+h))).$$

Mit der Abkürzung $u := u(t)$ ist dann zur Bestimmung von $u(t+h)$ ein Fixpunkt $v = v(t,u,h)$ der Abbildung

$$T_u : \mathbb{R}^n \to \mathbb{R}^n, \quad v \mapsto u + \frac{h}{2}(f(t,u) + f(t+h,v))$$

gesucht. Aus

$$\begin{aligned} |T_u v - T_u w| &= \frac{h}{2}|f(t+h,v) - f(t+h,w)| \\ &\leq \frac{hL}{2}|v-w| \end{aligned}$$

folgt die Kontraktionseigenschaft von $T_u$, falls $hL < 2$ gilt (cf. Definition A.1.2). Damit ist der Fixpunktsatz von BANACH A.1.3 anwendbar, es gibt also genau ein $v = v(t, u, h)$ mit $v = T_u v$. Dann ist

$$\varphi(t, u, h) := \frac{1}{2}(f(t, u) + f(t + h, v(t, u, h)))$$

die gesuchte Verfahrensfunktion für die Trapezregel.

Die LIPSCHITZ-Stetigkeit von $\varphi(t, u, h)$ bezüglich $u$ kann folgendermaßen gezeigt werden: Seien $u, \tilde{u} \in \mathbb{R}^n$ und $v := v(t, u, h)$ und $\tilde{v} := v(t, \tilde{u}, h)$ die eindeutigen Fixpunkte von $T_u$ bzw. $T_{\tilde{u}}$. Dann folgt

$$\begin{aligned} |v - \tilde{v}| &= |T_u v - T_{\tilde{u}} \tilde{v}| \\ &\leq |u - \tilde{u}| + \frac{hL}{2}|u - \tilde{u}| + \frac{hL}{2}|v - \tilde{v}| \end{aligned}$$

und daher

$$|v(t, u, h) - v(t, \tilde{u}, h)| \leq \frac{2 + hL}{2 - hL}|u - \tilde{u}|,$$

also die LIPSCHITZ-Stetigkeit von $v(t, u, h)$ bezüglich $u$. Hiermit ergibt sich

$$\begin{aligned} |\varphi(t, u, h) - \varphi(t, \tilde{u}, h)| &\leq \frac{L}{2}|u - \tilde{u}| + \frac{L}{2}|v(t, u, h) - v(t, \tilde{u}, h)| \\ &\leq \frac{2L}{2 - hL}|u - \tilde{u}|, \end{aligned}$$

und damit sind alle Behauptungen gezeigt.

Um aus Satz 4.1.7 die Konvergenz der Trapezregel folgern zu können, muss also nur noch die Konsistenz gezeigt werden.

**Lemma 4.4.3** *Sei die Lösung $x$ des Anfangswertproblems (4.1) dreimal stetig differenzierbar und $f(t, y)$* LIPSCHITZ*-stetig bezüglich $y$ mit Konstante $L$, dann besitzt die Trapezregel für dieses Problem die Konsistenzordnung 2, wenn auch die Startwerte $u_{h;0}$ von zweiter Ordnung gegen $x_0$ konvergieren.*

*Beweis:* Mittels Taylorentwicklung[2] folgt

$$|l_j| = \left| \frac{x(t_j) - x(t_{j-1})}{h} - \frac{1}{2}\left(f(t_{j-1}, x(t_{j-1})) + f(t_j, v(t_{j-1}, x(t_{j-1}), h))\right) \right|$$

[2]Satz B.1.7; in der Abschätzung finden des weiteren die Definitionen aus 4.1.5 sowie 4.4.2 und Satz B.1.10 Verwendung. Für $v(t_{j-1}, x(t_{j-1}), h)$ wird manchmal kurz $v$ geschrieben. $M_3$ bezeichnet eine Schranke für $x^{(3)}$ auf $[a, b]$. $Df(t, x)$ bezeichnet die FRÉCHET-Ableitung von $f$ im Punkt $(t, x)$.

$$
\begin{aligned}
&= \left|\frac{x(t_j)-x(t_{j-1})}{h} - x'(t_{j-1}) + \frac{1}{2}(x'(t_{j-1}) - f(t_j,v))\right| \\
&= \left|\frac{1}{h}\left(x(t_{j-1}) + hx'(t_{j-1}) + \frac{h^2}{2}x''(t_{j-1}) + \frac{h^3}{2}\int_0^1 (1-\tau)^2 x^{(3)}(t_{j-1}+\tau h)\,d\tau\right.\right. \\
&\quad \left.\left. - x(t_{j-1})\right) - x'(t_{j-1}) + \frac{1}{2}(x'(t_{j-1}) - x'(t_j) + x'(t_j) - f(t_j,v))\right| \\
&\leq \left|\frac{h}{2}x''(t_{j-1}) + \frac{1}{2}\left(x'(t_{j-1}) - x'(t_{j-1}) - hx''(t_{j-1})\right.\right. \\
&\quad \left.\left. - h^2\int_0^1 (1-\tau)x''(t_{j-1}+\tau h)\,d\tau + f(t_j,x(t_j)) - f(t_j,v)\right)\right| + \frac{h^2}{6}M_3 \\
&\leq \frac{L}{2}|x(t_j) - v| + O(h^2) \\
&= \frac{L}{2}\left|x(t_j) - x(t_{j-1}) - \frac{h}{2}f(t_{j-1},x(t_{j-1}))\right. \\
&\quad \left. - \frac{h}{2}f\left(t_j, x(t_{j-1}) + \frac{h}{2}f(t_{j-1},x(t_{j-1})) + \frac{h}{2}f(t_j,v)\right)\right| + O(h^2) \\
&= \frac{L}{2}\left|hx'(t_{j-1}) - \frac{h}{2}x'(t_{j-1}) - \frac{h}{2}f(t_{j-1},x(t_{j-1}))\right. \\
&\quad - \frac{h}{2}\int_0^1 Df\left(t_{j-1}+\tau h, x(t_{j-1}) + \frac{\tau h}{2}(f(t_{j-1},x(t_{j-1}))\right. \\
&\quad \left.\left. + f(t_j,v))\right) d\tau \left(h, \frac{h}{2}(f(t_{j-1},x(t_{j-1})) + f(t_j,v))\right)\right| + O(h^2) \\
&= O(h^2), \quad j = 1,\ldots,N.
\end{aligned}
$$

Nimmt man noch die Bedingung an die Startwerte hinzu, so folgt die Konsistenzordnung 2 gemäß Definition 4.1.5.

**Korollar 4.4.4** *Ist die Lösung $x$ von (4.1) in $C^3([a,b],\mathbb{R}^n)$, ist die Funktion $f(t,x)$* Lipschitz*-stetig bezüglich $x$ und konvergieren die Startwerte $u_{h;0}$ von zweiter Ordnung gegen den exakten Anfangswert $x_0$, dann ist die Trapezregel konvergent von zweiter Ordnung gegen $x$.*

*Beweis:* Folgt direkt aus Satz 4.1.7 unter Verwendung von Lemma 4.1.8 und Lemma 4.4.3, wenn man Satz 4.4.2 beachtet.

## 4.5 Die Mittelpunktsregel

**Definition 4.5.1 (Mittelpunktsregel)** *Die* Mittelpunktsregel *zur Lösung des Anfangswertproblems (4.1) ist definiert durch*

$$u_0 := u_{h;0},$$

$$u_j := u_{j-1} + hf\left(t_{j-1} + \frac{h}{2}, \frac{u_{j-1}+u_j}{2}\right), \quad j = 1, \ldots, N.$$

*Der Punkt $t_j + \frac{h}{2}$ wird im Folgenden häufig kurz mit $t_{j+1/2}$ bezeichnet, $j = 0, \ldots, N-1$. $t_{j-1/2}$ ist sinngemäß definiert.*

Da es sich auch hier um ein implizites Verfahren handelt, muss ein 4.3.2 bzw. 4.4.2 entsprechendes Resultat bewiesen werden.

**Satz 4.5.2** *Sei $f(t,x)$* Lipschitz*-stetig bezüglich $x$ mit* Lipschitz*-Konstante $L$, wobei $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ gelte. Dann lässt sich die Mittelpunktsregel für hinreichend kleine Schrittweiten $h$ als Einschrittverfahren gemäß Definition 4.1.4 auffassen, d. h. durch die implizite Form wird eine Verfahrensfunktion $\varphi(t,u,h)$ definiert, die* Lipschitz*-stetig bezüglich $u$ ist.*

*Beweis:* Ein Schritt der Mittelpunktsregel im Punkt $t$ mit Schrittweite $h$ hat die Gestalt

$$u(t+h) = u(t) + hf\left(t + \frac{h}{2}, \frac{u(t)+u(t+h)}{2}\right).$$

Mit der Abkürzung $u := u(t)$ ist dann zur Bestimmung von $u(t+h)$ ein Fixpunkt $v = v(t,u,h)$ der Abbildung

$$T_u : \mathbb{R}^n \to \mathbb{R}^n, \quad v \mapsto u + hf\left(t + \frac{h}{2}, \frac{u+v}{2}\right)$$

gesucht. Aus

$$\begin{aligned} |T_u v - T_u w| &= h\left|f\left(t + \frac{h}{2}, \frac{u+v}{2}\right) - f\left(t + \frac{h}{2}, \frac{u+w}{2}\right)\right| \\ &\leq \frac{hL}{2}|v-w| \end{aligned}$$

folgt die Kontraktionseigenschaft von $T_u$, falls $hL < 2$ gilt (cf. Definition A.1.2). Damit ist der Fixpunktsatz von Banach A.1.3 anwendbar, es gibt also genau ein $v = v(t,u,h)$ mit $v = T_u v$. Dann ist

$$\varphi(t,u,h) := f\left(t + \frac{h}{2}, \frac{u+v(t,u,h)}{2}\right)$$

die gesuchte Verfahrensfunktion für die Mittelpunktsregel.

Die Lipschitz-Stetigkeit von $\varphi(t,u,h)$ bezüglich $u$ kann folgendermaßen gezeigt werden: Seien $u, \tilde{u} \in \mathbb{R}^n$ und $v := v(t,u,h)$ und $\tilde{v} := v(t,\tilde{u},h)$ die eindeutigen Fixpunkte von $T_u$ bzw. $T_{\tilde{u}}$. Dann folgt

$$\begin{aligned} |v - \tilde{v}| &= |T_u v - T_{\tilde{u}}\tilde{v}| \\ &\leq |u - \tilde{u}| + \frac{hL}{2}|u - \tilde{u}| + \frac{hL}{2}|v - \tilde{v}| \end{aligned}$$

und daher

$$|v(t,u,h) - v(t,\tilde{u},h)| \leq \frac{2+hL}{2-hL}|u-\tilde{u}|,$$

also die LIPSCHITZ-Stetigkeit von $v(t,u,h)$ bezüglich $u$. Hiermit ergibt sich

$$\begin{aligned} |\varphi(t,u,h) - \varphi(t,\tilde{u},h)| &\leq \frac{L}{2}|u-\tilde{u}| + \frac{L}{2}|v(t,u,h) - v(t,\tilde{u},h)| \\ &\leq \frac{2L}{2-hL}|u-\tilde{u}|, \end{aligned}$$

und damit sind alle Behauptungen gezeigt.

Im folgenden Lemma wird nun noch die Konsistenz der Mittelpunktsregel gezeigt.

**Lemma 4.5.3** *Sei die Lösung $x$ des Anfangswertproblems (4.1) dreimal stetig differenzierbar und $f(t,y)$* LIPSCHITZ*-stetig bezüglich $y$ mit Konstante $L$, dann besitzt die Mittelpunktsregel für dieses Problem die Konsistenzordnung 2, wenn auch die Startwerte $u_{h;0}$ von zweiter Ordnung gegen $x_0$ konvergieren.*

*Beweis:* Mittels Taylorentwicklung folgt[3]

$$\begin{aligned} |l_j| &= \left| \frac{1}{h}\left( x(t_{j-1/2}) + \frac{h}{2}x'(t_{j-1/2}) + \frac{h^2}{8}x''(t_{j-1/2}) \right.\right. \\ &\quad + \frac{h^3}{16}\int_0^1 (1-\tau)^2 x^{(3)}\left(t_{j-1/2} + \tau\frac{h}{2}\right)\,d\tau\Big) \\ &\quad - \frac{1}{h}\left( x(t_{j-1/2}) - \frac{h}{2}x'(t_{j-1/2}) + \frac{h^2}{8}x''(t_{j-1/2}) \right. \\ &\quad - \frac{h^3}{16}\int_0^1 (1-\tau)^2 x^{(3)}\left(t_{j-1/2} - \tau\frac{h}{2}\right)\,d\tau\Big) \\ &\quad \left. - f\left(t_{j-1/2}, \frac{v(t_{j-1}, x(t_{j-1}), h) + x(t_{j-1})}{2}\right)\right| \\ &\leq \left| x'(t_{j-1/2}) - f\left(t_{j-1/2}, \frac{x(t_{j-1})+v}{2}\right)\right| + \frac{h^2}{24}M_3 \\ &= \left| f(t_{j-1/2}, x(t_{j-1/2})) - f\left(t_{j-1/2}, \frac{x(t_{j-1})+v}{2}\right)\right| + O(h^2) \\ &\leq L\left| x(t_{j-1/2}) - \frac{x(t_{j-1})+v}{2}\right| + O(h^2) \\ &= L\left| x(t_{j-1}) + \frac{h}{2}x'(t_{j-1}) + \frac{h^2}{4}\int_0^1 (1-\tau)x''\left(t_{j-1} + \tau\frac{h}{2}\right)\,d\tau\right. \end{aligned}$$

[3]Die verwendeten Definitionen und Sätze sind die gleichen wie in 4.4.3.

$$
\begin{aligned}
& - x(t_{j-1}) - \frac{h}{2} f\left(t_{j-1/2}, \frac{x(t_{j-1}) + v}{2}\right) \Bigg| + O(h^2) \\
= \; & \frac{Lh}{2} \Bigg| f(t_{j-1}, x(t_{j-1})) - f(t_{j-1}, x(t_{j-1})) \\
& - \int_0^1 Df\left(t_{j-1} + \tau\frac{h}{2}, x(t_{j-1}) + \tau\frac{v - x(t_{j-1})}{2}\right) d\tau \left(\frac{h}{2}, \frac{v - x(t_{j-1})}{2}\right) \Bigg| \\
& + O(h^2) \\
= \; & O(h^2), \quad j = 1, \ldots, N.
\end{aligned}
$$

Nimmt man noch die Bedingung an die Anfangswerte dazu, so folgt die Konsistenz gemäß Definition 4.1.5.

Zusammenfassend erhält man wieder das Konvergenzresultat.

**Korollar 4.5.4** *Ist die Lösung $x$ von (4.1) in $C^3([a,b], \mathbb{R}^n)$, ist die Funktion $f(t,x)$* Lipschitz*-stetig bezüglich $x$ und konvergieren die Startwerte $u_{h;0}$ von zweiter Ordnung gegen den exakten Anfangswert $x_0$, dann ist die Mittelpunktsregel konvergent von zweiter Ordnung gegen $x$.*

*Beweis:* Folgt direkt aus Satz 4.1.7 unter Verwendung von Lemma 4.1.8 und Lemma 4.5.3, wenn man Satz 4.5.2 beachtet.

## 4.6 Singuläre Probleme

In diesem Abschnitt soll gezeigt werden, warum die in den vorigen Abschnitten für glatte Probleme verwendeten Beweismethoden versagen, wenn sie für Konvergenzbeweise für singuläre Probleme herangezogen werden. Exemplarisch wird dabei das explizite Euler-Verfahren angewendet auf die *skalare* singuläre Gleichung erster Ordnung

$$x'(t) = \frac{\lambda}{t} x(t) + tg(t), \quad x(0) = 0 \tag{4.3}$$

mit $\lambda < 0$ betrachtet. Die Wahl von $\lambda < 0$, einer Inhomogenität der Gestalt $tg(t)$ und des Anfangswerts $x(0) = 0$ ist aus den Betrachtungen in Kapitel 3 klar. Offensichtlich lässt sich das explizite Euler-Verfahren 4.2.1 wegen der Singularität in $t = 0$ nicht ohne Modifikation in $t_0 = 0$ starten, man denke sich also das Verfahren in einem Punkt $t_0 > 0$ mit einem Startwert $u_h(t_0) = u_{h;0}$ gestartet. Dieser Startwert ist natürlich im Allgemeinen nicht gleich dem exakten Wert der Lösung $x(t_0)$. Diese *Störung des Startwerts* werde mit $\varepsilon_0 := u_{h;0} - x(t_0)$

bezeichnet. Für die Verfahrensfunktion $\varphi$ des expliziten EULER-Verfahrens bei Anwendung auf Gleichung (4.3) gilt

$$\varphi(t,u,h) := \frac{\lambda}{t}u + tg(t).$$

Für $t_0 > 0$ ist $\varphi(t,u,h)$ LIPSCHITZ-stetig bezüglich $u$ auf jedem Intervall $[t_0,\delta]$ mit LIPSCHITZ-Konstante $L := \frac{|\lambda|}{t_0}$. Dabei wird $\delta > t_0$ fest gewählt[4]. Mit dieser Konstante $L$ kann man das Resultat von Satz 4.1.7 unter Beachtung von Lemma 4.1.8 formal nachvollziehen. Nimmt man unter Beachtung der Ergebnisse aus Kapitel 3, speziell Lemma 3.3.1, an, dass für die exakte Lösung von (4.3) $x \in C^2([0,1],\mathbb{R})$ gilt, also Lemma 4.2.2 für alle $t_0 \geq 0$ angewendet werden kann, so erhält man die Abschätzung

$$\begin{aligned} |u_j - x(t_j)| &\leq \exp\left(|\lambda|\frac{hj}{t_0}\right)|\varepsilon_0| + \frac{t_0}{|\lambda|}\left(\exp\left(|\lambda|\frac{hj}{t_0}\right) - 1\right)\frac{M_2 h}{2} \\ &\leq \exp\left(|\lambda|\frac{1}{t_0}\right)|\varepsilon_0| + \frac{t_0}{|\lambda|}\left(\exp\left(|\lambda|\frac{1}{t_0}\right) - 1\right)\frac{M_2 h}{2}, \end{aligned} \tag{4.4}$$

wobei mit $M_2$ wieder eine Schranke für $x''$ bezeichnet wird. Natürlich ist man aber im Endeffekt an einer numerischen Lösung von Gleichung (4.3) auf dem Intervall $[0,1]$ interessiert, also ist es notwendig, sukzessive $t_0 \to 0$ gehen zu lassen. Es erscheint sinnvoll, $t_0 = t_0(h) := t_{i_0} = i_0 h$ zu setzen, da dann $\lim_{h\to 0} t_0 = 0$ gilt, und gleichzeitig die Rechnung für alle $h$ formal gleich beim Gitterpunkt $t_{i_0}$ begonnen werden kann. Lässt man jetzt $h \to 0$ gehen, so führt das zu einem exponentiellen Aufklingen der rechten Seite in Abschätzung (4.4), die durch das polynomiale Abklingen der anderen Faktoren nicht kompensiert werden kann. Insgesamt ist die Abschätzung (4.4) also nicht brauchbar, um die Konvergenz des expliziten EULER-Verfahrens für Gleichung (4.3) zu zeigen. Das bedeutet aber natürlich *nicht*, dass diese Konvergenz nicht trotzdem stattfinden kann. Man beachte etwa, dass in der Berechnung der LIPSCHITZ-Konstanten dem Umstand keine Rechnung getragen wurde, dass $\lambda < 0$ gilt. Zieht man jedoch geeignete Beweismethoden heran, die diese Tatsache ausnützen, indem sie auf die einfache lokale Betrachtensweise der Stabilität, wie sie in diesem Kapitel zugrundegelegt wurde, verzichten, so erhält man unter geeigneten Voraussetzungen an die Daten singulärer Probleme wieder befriedigende Konvergenzresultate. Dies ist Gegenstand des nächsten Kapitels.

[4] Im Weiteren sei o. B. d. A. $\delta = 1$.

# Kapitel 5

# Numerische Lösung singulärer Probleme

In diesem Abschnitt werden neue Resultate des Autors zur Konvergenz von Einschrittverfahren für singuläre Probleme diskutiert. Beweise einiger verwendeter Sätze, die nicht unmittelbar mit der Materie in Zusammenhang stehen, sind im Anhang A angegeben. Sätze, bei denen kein Hinweis auf den Ursprung zu finden ist, wurden vom Autor (zumindest in der jeweiligen Formulierung) bewiesen.

Dieses Kapitel befasst sich mit der numerischen Lösung des singulären Anfangswertproblems zweiter Ordnung

$$\begin{aligned} y''(t) &= \frac{A_1}{t}y'(t) + \frac{A_0}{t^2}y(t) + f(t, y(t)), \quad t \in (0, 1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0. \end{aligned} \tag{5.1}$$

Die Wahl der Anfangsbedingungen ist aus Kapitel 3 klar, cf. speziell Satz 3.5.1. Da Einschrittverfahren angewendet werden sollen, muss man auf Gleichung (5.1) die Variablentransformation $z(t) := (z_1(t), z_2(t)) := (y(t), ty'(t))$ anwenden, um die Gleichung erster Ordnung

$$\begin{aligned} z'(t) &= \frac{M}{t}z(t) + \overset{\circ}{tf}(t, z(t)), \quad t \in (0, 1], \\ z(0) &= z_0, \end{aligned} \tag{5.2}$$

mit

$$M := \begin{pmatrix} 0 & I \\ A_0 & I + A_1 \end{pmatrix}, \quad \overset{\circ}{f}(t, z(t)) := \begin{pmatrix} 0 \\ f(t, z_1(t)) \end{pmatrix}$$

zu erhalten. Wie aus der Bemerkung auf Seite 28 zu entnehmen ist, ist für die Lösung $z$ von (5.2) auch der Wert $z'(0) = (y'(0), y'(0))$ bekannt.

Zuerst wird ausführlich das explizite EULER-Verfahren untersucht, im Anschluss daran das implizite EULER-Verfahren in kürzerer Form, da sich in den Beweismethoden viele Ähnlichkeiten zum expliziten EULER-Verfahren finden.

Auch die Beweise für die Trapez- und die Mittelpunktsregel weisen viele Parallelen in der Vorgangsweise auf und sind deshalb weniger ausführlich.

## 5.1 Das explizite EULER-Verfahren

Bei der Untersuchung der Konvergenz des EULER-Verfahrens 4.2.1 wird zunächst der lineare Fall mit konstanter Koeffizientenmatrix $M$ studiert. Ausgehend von diesen Resultaten wird dann der *allgemeine lineare Fall* ($M = M(t)$) und der *nichtlineare Fall* ($f = f(t, y)$) betrachtet.

### 5.1.1 Konstante Koeffizientenmatrix $M$

Gemäß der Resultate aus Kapitel 3, siehe Satz 3.3.4, wird die Gleichung

$$\begin{aligned} z'(t) &= \frac{M}{t} z(t) + \overset{\circ}{tf}(t), \quad t \in (0, 1], \\ z(0) &= z_0, \end{aligned} \tag{5.3}$$

mit $z_0 \in \mathcal{R}(R)$, wobei $R$ die Spektralprojektion auf den Eigenraum von $M$ zum Eigenwert 0 bezeichnet[1], untersucht. Es wird angenommen, dass $M$ keine Eigenwerte mit positivem Realteil oder rein imaginäre Eigenwerte ungleich 0 besitzt, cf. Kapitel 3.

Um die Untersuchungen zu vereinfachen, transformiert man die Gleichung wieder so, dass die Koeffizientenmatrix JORDAN-Normalform besitzt (siehe dazu Abschnitt 3.2.1, S. 18 bzw. Satz B.2.2), also

$$\begin{aligned} v'(t) &= \frac{1}{t} J v(t) + t g(t) \\ v(0) &= E^{-1} z_0 =: v_0 \end{aligned} \tag{5.4}$$

mit $J = E^{-1}ME$ in JORDAN-Normalform und $v(t) := E^{-1}z(t), g(t) := E^{-1}\overset{\circ}{f}(t)$.

Man sieht sofort, dass sich (wegen der Polstelle bei $t = 0$) das explizite EULER-Verfahren *nicht ohne weiteres in* $t_0 = 0$ *starten lässt*. Deshalb geht man folgendermaßen vor:

Sei ein äquidistantes Gitter

$$\Delta_h := (t_{i_0}, t_{i_0+1}, \ldots, t_N)$$

[1] Siehe dazu Satz 3.3.4.

mit $t_j = hj,\ j = i_0, \ldots, N, h = \frac{1}{N}$ und einem $i_0 > 0$ definiert. Im Folgenden wird das EULER-Verfahren stets auf einer Folge von Gittern $\Delta_h$ betrachtet, wobei $i_0$ festgehalten wird. Es gilt aber natürlich $\lim_{h\to 0} t_{i_0} = 0$ und $\lim_{h\to 0} N = \infty$.

Es folgen jetzt einige Definitionen, die zentraler Gegenstand der anschließenden Überlegungen sind.

Der Operator $F$ sei definiert auf $C^1((0,1], \mathbb{C}^{2n})$ als

$$F(x) := \begin{pmatrix} x'(t) - \frac{1}{t}Jx(t) - tg(t),\ t \in (0,1] \\ x(0) - v_0 \end{pmatrix}, \tag{5.5}$$

wobei $v_0$ der Startwert in (5.4) ist. Die (exakte) Lösung von (5.4) ist also äquivalent zur Lösung von

$$F(v) = 0.$$

Sei $x_h = (x_{i_0}, \ldots, x_N) \in \mathbb{C}^{2n(N-i_0+1)}$, dann sei der Operator $F_h$ erklärt durch

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1}-x_j}{h} - \frac{1}{t_j}Jx_j - t_j g(t_j),\ j = i_0, \ldots, N-1 \\ x_{i_0} - v_{h;i_0} \end{pmatrix}, \tag{5.6}$$

wobei $v_{h;i_0}$ der Startwert für das EULER-Verfahren ist. Dann ist die Berechnung der Approximationslösung $v_h$ zur Schrittweite $h$ mittels des EULER-Verfahrens äquivalent zur Lösung der Gleichung

$$F_h(v_h) = 0.$$

**Definition 5.1.1 (Konsistenz)** *Es sei $v$ eine Lösung der Gleichung $F(v) = 0$ mit $F : U \to E_2, U \subseteq E_1$, wobei $(E_1, \| \|_1)$ und $(E_2, \| \|_2)$ Banachräume seien. Betrachte eine Familie von Näherungsproblemen $F_h(v_h) = 0$ auf den Banachräumen $(E_1^h, \| \|_1^h), (E_2^h, \| \|_2^h)$, wobei $F_h : E_1^h \to E_2^h$ gelte. Um eine Beziehung zwischen den Problemen $F(v) = 0$ und $F_h(v_h) = 0$ zu erhalten, gebe es lineare Abbildungen $R_1^h$ und $R_2^h$, sodass*

$$R_i^h : E_i \to E_i^h, \quad \lim_{h\to 0} \|R_i^h(x)\|_i^h = \|x\|_i \quad \forall x \in E_i, \quad i \in \{1,2\}$$

*gilt. Die Familie $F_h$ heißt* konsistent mit dem Problem $F(v) = 0$, *wenn für die („exakte“) Lösung $v$ gilt*

$$\lim_{h\to 0} \|F_h(R_1^h(v))\|_2^h = 0.$$

*Die Familie hat* Konsistenzordnung $p > 0$, *wenn*

$$\|F_h(R_1^h(v))\|_2^h = O(h^p) \quad \textit{für } h \to 0$$

*gilt.*

**Definition 5.1.2 (Stabilität)** *Seien* $(E_1, \| \ \|_1)$ *und* $(E_2, \| \ \|_2)$ *Banachräume,* $U \subseteq E_1$ *und* $F : U \to E_2$ *ein Operator.* $F$ *heißt* stabil*, wenn es eine Konstante* $K > 0$ *gibt mit*

$$\|x - y\|_1 \leq K\|F(x) - F(y)\|_2 \quad \forall x, y \in U.$$

*Für eine Familie* $F_h, h \in (0, h_0], h_0 > 0$*, von Operatoren, cf. Definition 5.1.1, spricht man von Stabilität, wenn die Ungleichung*

$$\|x_h - y_h\|_1^h \leq K\|F_h(x_h) - F_h(y_h)\|_2^h$$

gleichmäßig*, d. h. mit einer von* $h$ *unabhängigen Konstante* $K$*, gilt.*

**Bemerkung:** Stabilität eines Operators $F$ kommt also der LIPSCHITZ-Stetigkeit des inversen Operators $F^{-1}$ auf $\mathcal{R}(R)$ gleich.

**Korollar 5.1.3** *Ist die Familie* $F_h$ *stabil und konsistent im Sinne der Definitionen 5.1.2 und 5.1.1, dann sind die Näherungslösungen* $v_h$ *von* $F_h(v_h) = 0$ *konvergent gegen die Lösung* $v$ *von* $F(v) = 0$*, wegen*

$$\lim_{h\to 0} \|v_h - R_1^h(v)\|_1^h \leq \lim_{h\to 0} K\|F_h(R_1^h(v))\|_2^h = 0.$$

Im Folgenden wird also versucht, Konsistenz und Stabilität der Operatoren $F_h$ aus (5.6) zu zeigen, um daraus auf die Konvergenz der Näherungslösungen zu schließen. Dies wird jedoch in manchen Fällen nur mit einer Modifikation der obigen Konzepte möglich sein[2], das vorangehende Korollar 5.1.3 weist jedoch die richtige Richtung für die Konvergenzbeweise.

Da Gleichung (5.4) *teilweise entkoppelt* ist, scheint es sinnvoll, das explizite EULER-Verfahren *komponentenweise* zu betrachten. Dabei genügt es vorerst, nur JORDAN-Matrizen $J = J_m(\lambda)$ zu betrachten, da nur innerhalb eines solchen Blocks Wechselwirkungen zwischen den Lösungskoordinaten auftreten können. Es sind also skalare Gleichungen der beiden Bauarten

$$v_i'(t) = \frac{\lambda}{t} v_i(t) + t g_i(t), \tag{5.7}$$

$$v_i'(t) = \frac{\lambda}{t} v_i(t) + \frac{1}{t} v_{i+1}(t) + t g_i(t) \tag{5.8}$$

mit $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}, \Re(\lambda) = \sigma \leq 0$ zu analysieren.

Es ergeben sich, wie auch für die analytischen Überlegungen aus Kapitel 3, Unterschiede für die Fälle, dass der Realteil von $\lambda$ negativ ist oder verschwindet. Die beiden Fälle werden deshalb im Folgenden getrennt untersucht.

[2]Für JORDAN-Matrizen $J_m(0)$ der Größe $m > 1$ zum Eigenwert 0 ist $F_h$ *nicht stabil.*

**Eigenwerte mit negativem Realteil**

Die Lösung der Gleichung (5.7) mittels des expliziten EULER-Verfahrens ist also äquivalent zur Lösung der Gleichung

$$F_{h;i}(v_{h;i}) := \left( \begin{array}{l} \frac{v_{j+1;i}-v_{j;i}}{h} - \frac{\lambda}{t_j} v_{j;i} - t_j g_i(t_j),\ j = i_0, \ldots, N-1 \\ v_{i_0;i} - v_{h;i_0;i} \end{array} \right) = 0. \qquad (5.9)$$

Dabei setzt man $v_{h;i_0;i} := v_i(0) + t_{i_0} v_i'(0)$. Der Anfangswert $v_i(0)$ ist aus der Spezifikation des Problems (5.4) bekannt. $v_i'(0)$ läßt sich mit Hilfe der Resultate aus Kapitel 3 berechnen, cf. die Bemerkung auf S. 28.

**Bemerkung:** Diese Wahl des Anfangswerts entspricht einem EULER-Schritt von 0 weg unter Verwendung der Kenntnis von $v_i'(0)$. Die technischen Gründe, die diese gesonderte Behandlung des ersten Schritts notwendig machen, werden später klar werden. Es wird sich nämlich zeigen, dass das Ergebnis dieses ersten Schritts von zweiter Ordnung gegen den exakten Wert der analytischen Lösung konvergiert.

Um die Notation zu vereinfachen, soll für den Rest dieses Abschnitts der Index $i$ wegfallen und $g_j := g(t_j)$ gesetzt werden.

Sei $v$ die exakte Lösung von (5.7) und $v_h$ die Lösung von (5.9). Die lineare Abbildung $R_h$ ist definiert als $R_h(v) := (v(t_{i_0}), \ldots, v(t_N))$. Das Ziel ist es, den Ausdruck $\|v_h - R_h(v)\|_h$ abzuschätzen. Als ersten Schritt zeigt man die Stabilität von $F_h$. Es gilt

$$\begin{aligned} F_h(R_h(v)) &= \left( \begin{array}{l} \frac{v(t_{j+1})-v(t_j)}{h} - \frac{\lambda}{t_j} v(t_j) - t_j g_j,\ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0} v'(0) \end{array} \right) \\ &= \left( \begin{array}{l} \frac{v(t_{j+1})-v(t_j)}{h} - v'(t_j),\ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0} v'(0) \end{array} \right) \\ &=: \left( \begin{array}{l} l_{j+1},\ j = i_0, \ldots, N-1 \\ l_{i_0} \end{array} \right). \end{aligned}$$

Der Ausdruck $R_h(v) - v_h =: \varepsilon_h$ ist Lösung der Gleichung

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{\lambda}{t_j} \varepsilon_j - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0. \qquad (5.10)$$

Man versucht jetzt, diese Lösung $\varepsilon_h$ in $l_h$ abzuschätzen.

Dies geschieht in etwas allgemeinerer Form in Lemma 5.1.6. Für den Beweis werden vorher noch zwei weitere Lemmata gebraucht, deren Beweise sinngemäß [52] entstammen.

**Lemma 5.1.4** *Sei $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ mit $\sigma = \Re(\lambda) > 0$ fest gewählt. Definiere für $j \geq k \geq 1$*

$$z_{kj}(\lambda) := \begin{cases} 1, & k = j, \\ \prod_{l=k}^{j-1} \left(1 - \frac{\lambda}{l}\right), & 1 \leq k < j, \quad j = 2, 3, \ldots. \end{cases}$$

*Dann gibt es ein $\eta > 0$ und ein $C \geq 1$, sodass*

$$|z_{kj}(\lambda)| \leq C \left(\frac{k}{j}\right)^{\eta}, \quad 1 \leq k \leq j, \quad j = 1, 2, \ldots. \tag{5.11}$$

*Beweis:* Siehe Lemma A.2.1.

**Lemma 5.1.5** *Sei $h > 0, t_j := jh, k > j \geq i_0 > 0$ und $\gamma \in \mathbb{R}$, dann gilt*

$$\sum_{l=j}^{k-1} h t_l^{\gamma-1} \leq \begin{cases} c_1 |t_k^{\gamma} - t_j^{\gamma}|, & \gamma \neq 0, \\ c_2 \ln\left(\frac{t_k}{t_j}\right), & \gamma = 0. \end{cases} \tag{5.12}$$

*Beweis:* Siehe Lemma A.2.2.

**Lemma 5.1.6** *Sei $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1} - \varepsilon_j}{h} - \frac{\lambda}{t_j}\varepsilon_j - t_j^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.13}$$

*gilt*

$$\begin{aligned} \varepsilon_i &= \prod_{l=i_0}^{i-1} \left(1 + \frac{h\lambda}{t_l}\right) l_{i_0} + \sum_{l=i_0}^{i-2} \prod_{k=l+1}^{i-1} \left(1 + \frac{h\lambda}{t_k}\right) h t_l^{\gamma-1} l_{l+1} + h t_{i-1}^{\gamma-1} l_i \\ &=: z_{i_0,i}(-\lambda) l_{i_0} + \sum_{l=i_0}^{i-2} z_{l+1,i}(-\lambda) h t_l^{\gamma-1} l_{l+1} + h t_{i-1}^{\gamma-1} l_i, \\ &\quad i = i_0 + 1, \ldots, N. \end{aligned} \tag{5.14}$$

*$z_{lk}(-\lambda)$ ist wie in Lemma 5.1.4 definiert. Produkte bzw. Summen, bei denen der obere Index kleiner als der untere ist, sind als leer aufzufassen, d. h. solche Summen sind gleich 0, solche Produkte gleich 1.*

*Weiters gelten die Abschätzungen*

$$|\varepsilon_i| \leq \text{const.}(|l_{i_0}| + t_i^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l|) \tag{5.15}$$

$$\leq \text{const.} \|l_h\|_h, \quad i = i_0, \ldots, N. \tag{5.16}$$

*Beweis:* Wird mittels vollständiger Induktion geführt. Für $i = i_0 + 1$ stimmt (5.14) offensichtlich. Setze also (5.14) für $\varepsilon_{i-1}$ voraus und betrachte $\varepsilon_i$:

$$\begin{aligned}
\varepsilon_i &= \left(1 + \frac{h\lambda}{t_{i-1}}\right)\varepsilon_{i-1} + ht_{i-1}^{\gamma-1} l_i \\
&= \left(1 + \frac{h\lambda}{t_{i-1}}\right)\left(\prod_{l=i_0}^{i-2}\left(1 + \frac{h\lambda}{t_l}\right) l_{i_0} + \sum_{l=i_0}^{i-3}\prod_{k=l+1}^{i-2}\left(1 + \frac{h\lambda}{t_k}\right) ht_l^{\gamma-1} l_{l+1} + ht_{i-2}^{\gamma-1} l_{i-1}\right) \\
&\quad + ht_{i-1}^{\gamma-1} l_i \\
&= \prod_{l=i_0}^{i-1}\left(1 + \frac{h\lambda}{t_l}\right) l_{i_0} + \sum_{l=i_0}^{i-2}\prod_{k=l+1}^{i-1}\left(1 + \frac{h\lambda}{t_k}\right) ht_l^{\gamma-1} l_{l+1} + ht_{i-1}^{\gamma-1} l_i.
\end{aligned}$$

Mit Lemma 5.1.4 und Lemma 5.1.5 folgt jetzt für ein $\eta > 0$ die Abschätzung

$$\begin{aligned}
|\varepsilon_i| &\le d_1\left(\frac{t_{i_0}}{t_i}\right)^{\eta} |l_{i_0}| + d_2 \sum_{l=i_0}^{i-2}\left(\frac{t_{l+1}}{t_i}\right)^{\eta} ht_l^{\gamma-1}|l_{l+1}| + ht_{i-1}^{\gamma-1}|l_i| \\
&\le d_1|l_{i_0}| + d_3 t_i^{-\eta} \sum_{l=i_0}^{i-2} t_l^{\eta} ht_l^{\gamma-1}|l_{l+1}| + ht_{i-1}^{\gamma-1}|l_i| \\
&\le d_1|l_{i_0}| + d_4 t_i^{-\eta} \sum_{l=i_0}^{i-1} ht_l^{\eta+\gamma-1} \max_{i_0+1\le k\le N} |l_k| \\
&\le d_1|l_{i_0}| + d_5 t_i^{-\eta}|t_i^{\eta+\gamma} - t_{i_0}^{\eta+\gamma}| \max_{i_0+1\le k\le N} |l_k| \\
&\le d_1|l_{i_0}| + d_5 t_i^{\gamma} \max_{i_0+1\le k\le N} |l_k| \\
&\le d_6\|l_h\|_h, \quad i = i_0, \ldots, N.
\end{aligned}$$

Aus Lemma 5.1.6 folgt für die Lösung $\varepsilon_h$ von (5.10) also die Abschätzung

$$|\varepsilon_i| \le \text{const.}\|l_h\|_h, \quad i = i_0, \ldots, N. \tag{5.17}$$

Mit Hilfe dieser Abschätzung wird letztendlich die Konvergenz der Näherungslösungen $v_h$ mit $F_h(v_h) = 0$ gegen die Lösung $v$ von $F(v) = 0$ gezeigt, doch zuvor folgt noch die Untersuchung von Gleichung (5.8).

In diesem Fall definiert man

$$F_h(v_h, w_h) := \begin{pmatrix} \frac{v_{j+1}-v_j}{h} - \frac{\lambda}{t_j} v_j - \frac{1}{t_j} w_j - t_j g_j, \ j = i_0, \ldots, N-1 \\ v_{i_0} - v_0 - t_{i_0} v'(0) \end{pmatrix}.$$

Die Gleichung $F_h(v_h, w_h) = 0$ ist äquivalent zur Lösung von (5.8) mittels des expliziten EULER-Verfahrens zum Startwert $v_{i_0} = v(0) + t_{i_0} v'(0)$.

**Bemerkung:** Der Index $i$ wurde wieder weggelassen, um die Ausführungen übersichtlicher zu gestalten, man beachte aber, dass es sich um einen Operator handelt, der auf Gittervektoren $x_h \in \mathbb{C}^{N-i_0+1}$ operiert, und der *nicht* gleich dem Operator $F_h$ aus (5.6) ist, sondern eine von dessen Komponenten darstellt!

Im Folgenden wird vorausgesetzt, dass $w_h$ bekannt ist und (sinngemäß) die Abschätzung (5.15) erfüllt. Diese Annahme ist gerechtfertigt, da ja das Gesamtproblem beginnend für die „entkoppelten Komponenten“ auf die anderen Lösungskomponenten hochgerechnet wird. Setzt sich unter diesen Annahmen das Verhalten von $w_h$ auf $v_h$ fort, kann man also wieder eine Abschätzung der Form (5.15) erhalten, so ist das Verhalten *aller* Komponenten der Lösung bekannt.

Sei also $v$ die exakte Lösung von (5.8), $w$ die Lösung der („zu $v$ gehörigen“) Gleichung der Gestalt (5.7) oder (5.8), $v_h$ die Lösung von $F_h(v_h, w_h) = 0$ und $w_h$ die Lösung der zu $v_h$ korrespondierenden Gleichung (die die Diskretisierung der Gleichung für $w$ ist) . Dann kann man $\varepsilon_h := R_h(v) - v_h$ (analog wie in (5.10)) darstellen als Lösung von

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{\lambda}{t_j}\varepsilon_j - \frac{1}{t_j}\delta_j - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.18}$$

mit

$$\begin{aligned} F_h(R_h(v), R_h(w)) &= \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - \frac{\lambda}{t_j}v(t_j) - \frac{1}{t_j}w(t_j) - t_j g_j,\ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0}v'(0) \end{pmatrix} \\ &= \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - v'(t_j),\ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0}v'(0) \end{pmatrix} \\ &=: \begin{pmatrix} l_{j+1},\ j = i_0, \ldots, N-1 \\ l_{i_0} \end{pmatrix} \end{aligned}$$

und $\delta_h = R_h(w) - w_h$.

Die Abschätzung von $\varepsilon_h$ in $l_h$ erfolgt wie in Lemma 5.1.6 in etwas allgemeinerer Form.

**Lemma 5.1.7** *Sei $\gamma > 0$. Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \leq \text{const.}(|\tilde{l}_{i_0}| + t_i^{\gamma} \max_{i_0+1 \leq k \leq N} |\tilde{l}_k|), \quad i = i_0, \ldots, N, \tag{5.19}$$

*wobei $\tilde{l}_h \in \mathbb{C}^{N-i_0+1}$ ein Gittervektor sei. $\varepsilon_h$ sei die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{\lambda}{t_j}\varepsilon_j - \frac{1}{t_j}\delta_j - t_j^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0. \tag{5.20}$$

*Dann gilt*

$$\begin{aligned}\varepsilon_i &= \prod_{l=i_0}^{i-1}\left(1+\frac{h\lambda}{t_l}\right) l_{i_0} + \sum_{l=i_0}^{i-2}\prod_{k=l+1}^{i-1}\left(1+\frac{h\lambda}{t_k}\right) h t_l^{\gamma-1} l_{l+1}^* + h t_{i-1}^{\gamma-1} l_i^* \\ &=: z_{i_0,i}(-\lambda) l_{i_0} + \sum_{l=i_0}^{i-2} z_{l+1,i}(-\lambda) h t_l^{\gamma-1} l_{l+1}^* + h t_{i-1}^{\gamma-1} l_i^*, \\ &\quad i = i_0+1,\ldots,N \end{aligned} \tag{5.21}$$

*mit* $l_i^* := t_{i-1}^{-\gamma}\delta_{i-1} + l_i,\ i = i_0+1,\ldots,N$. *Weiters gelten die Abschätzungen*

$$\begin{aligned}|\varepsilon_i| &\le \text{const.}(\max\{|l_{i_0}|, |\tilde{l}_{i_0}|\} + t_i^{\gamma} \max_{i_0+1\le k\le N} \max\{|l_k|, |\tilde{l}_k|\}) &(5.22)\\ &\le \text{const.}\max\{\|l_h\|_h, \|\tilde{l}_h\|_h\}, \quad i = i_0,\ldots,N. &(5.23)\end{aligned}$$

*Beweis:* Die Darstellung (5.21) folgt ganz analog wie in 5.1.6 mit vollständiger Induktion.

Zur Abschätzung von $\varepsilon_h$ gelangt man mittels der Abschätzung von $l_h^*$ gemäß

$$\begin{aligned}|l_{i+1}^*| &\le d_1 t_i^{-\gamma}(|\tilde{l}_{i_0}| + t_i^{\gamma} \max_{i_0+1\le k\le N} |\tilde{l}_k|) + \max_{i_0+1\le k\le N} |l_k| \\ &\le d_2(t_i^{-\gamma}|\tilde{l}_{i_0}| + \max_{i_0+1\le k\le N} \max\{|\tilde{l}_k|, |l_k|\}), \quad i = i_0,\ldots,N-1,\end{aligned}$$

und dann genauso wie in 5.1.6 mit einem $\eta > 0$

$$\begin{aligned}|\varepsilon_i| &\le d_3|l_{i_0}| + d_4 t_i^{-\eta}\sum_{l=i_0}^{i-2} h t_l^{\eta-1}|\tilde{l}_{i_0}| + d_2 h t_{i-1}^{-1}|\tilde{l}_{i_0}| \\ &\quad + d_5 t_i^{-\eta}\sum_{l=i_0}^{i-2} h t_l^{\eta+\gamma-1} \max_{i_0+1\le k\le N} \max\{|\tilde{l}_k|, |l_k|\} \\ &\quad + d_2 h t_{i-1}^{\gamma-1} \max_{i_0+1\le k\le N} \max\{|\tilde{l}_k|, |l_k|\} \\ &\le d_3|l_{i_0}| + d_6 t_i^{-\eta}\sum_{l=i_0}^{i-1} h t_l^{\eta-1}|\tilde{l}_{i_0}| \\ &\quad + d_7 t_i^{-\eta}\sum_{l=i_0}^{i-1} h t_l^{\eta+\gamma-1} \max_{i_0+1\le k\le N} \max\{|\tilde{l}_k|, |l_k|\} \\ &\le d_3|l_{i_0}| + d_8 t_i^{-\eta}|t_i^{\eta} - t_{i_0}^{\eta}||\tilde{l}_{i_0}| + d_9 t_i^{-\eta}|t_i^{\eta+\gamma} - t_{i_0}^{\eta+\gamma}| \max_{i_0+1\le k\le N} \max\{|\tilde{l}_k|, |l_k|\} \\ &\le d_3|l_{i_0}| + d_8|\tilde{l}_{i_0}| + d_9 t_i^{\gamma} \max_{i_0+1\le k\le N} \max\{|\tilde{l}_k|, |l_k|\}.\end{aligned}$$

Daraus folgen sofort (5.22) und (5.23).

Für $\gamma = 1$ erhält man die gewünschte Abschätzung von $\varepsilon_h = R_h(v) - v_h$ in $l_h$ und $\tilde{l}_h$.

Die vorangegangenen komponentenweisen Betrachtungen sollen jetzt wieder zu einem Resultat für den Operator $F_h$ aus (5.6) zusammengeführt werden. Im Weiteren gelte also für alle betrachteten Vektoren wieder $x_h \in \mathbb{C}^{2n(N-i_0+1)}$. Vorerst werden nur Matrizen $J$ mit Eigenwerten mit negativem Realteil betrachtet.

Sei $v$ die exakte Lösung von $F(v) = 0$ und $v_h$ die Lösung von $F_h(v_h) = 0$. Der globale Diskretisierungsfehler $\varepsilon_h := R_h(v) - v_h$ erfüllt dann die Gleichung

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - J\varepsilon_j - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.24}$$

mit

$$l_h := F_h(R_h(v)) = \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - v'(t_j),\ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0} v'(0) \end{pmatrix}. \tag{5.25}$$

Die Lemmata 5.1.6 und 5.1.7 zeigen, dass

$$|\varepsilon_i| \leq \text{const.}(|l_{i_0}| + t_i \max_{i_0+1 \leq k \leq N} |l_k|), \quad i = i_0, \ldots, N, \tag{5.26}$$

$$\|\varepsilon_h\|_h \leq \text{const.}\|l_h\|_h \tag{5.27}$$

gilt. Gleichung (5.27) bedeutet, dass der Operator $F_h$ stabil im Sinne von Definition 5.1.2 ist. Es bleibt für den Beweis der Konvergenz der Näherungslösungen $v_h$ also nur noch die Konsistenz von $F_h$ gemäß Definition 5.1.1 zu zeigen. Das geschieht im folgenden Lemma.

**Lemma 5.1.8** *Sei $l_h$ definiert wie in (5.25). Dann gilt für $v \in C^1([0,1], \mathbb{C}^{2n})$*

$$\begin{aligned} \lim_{h \to 0} \|l_h\|_h &= 0, \\ |l_{i_0}| &= o(h). \end{aligned}$$

*Ist $v \in C^2([0,1], \mathbb{C}^{2n})$, dann gilt sogar*

$$\begin{aligned} \|l_h\|_h &= O(h), \\ |l_{i_0}| &= O(h^2). \end{aligned}$$

*Beweis:* Ist $v$ stetig differenzierbar auf $[0,1]$, dann gilt nach dem Satz von Taylor B.1.7 (Entwicklung von $v(t_{j+1})$ um $t_j$)

$$\begin{aligned} \left| \frac{v(t_{j+1}) - v(t_j)}{h} - v'(t_j) \right| &= \left| \int_0^1 v'(t_j + \tau h)\, d\tau - v'(t_j) \right| \\ &= \left| \int_0^1 (v'(t_j + \tau h) - v'(t_j))\, d\tau \right| \\ &\leq \max_{\theta \in [t_j, t_{j+1}]} |v'(\theta) - v'(t_j)|, \quad j = i_0, \ldots, N-1. \end{aligned}$$

Da $v'$ auf dem kompakten Intervall $[0,1]$ gleichmäßig stetig ist (siehe Satz B.1.5), gilt also wegen $\lim_{h\to 0}|t_{j+1}-t_j|=0$

$$\lim_{h\to 0}\max_{i_0+1\le k\le N}|l_k|=0.$$

Für die „Anfangswertkomponente“ $l_{i_0}$ von $l_h$ gilt wieder mit Satz B.1.7

$$\begin{aligned}|v(t_{i_0})-v(0)-t_{i_0}v'(0)| &= \left|t_{i_0}\int_0^1[v'(\tau t_{i_0})-v'(0)]\,d\tau\right|\\ &\le hi_0\max_{\theta\in[0,t_{i_0}]}|v'(\theta)-v'(0)|\\ &= o(h)\end{aligned}$$

wegen der gleichmäßigen Stetigkeit von $v'$. Insgesamt folgt also

$$\lim_{h\to 0}\|l_h\|_h=0.$$

Sei jetzt $v\in C^2([0,1],\mathbb{C}^{2n})$. Dann gilt nach dem Satz von Taylor B.1.7 (analog wie in Lemma 4.2.2)

$$\begin{aligned}\left|\frac{v(t_{j+1})-v(t_j)}{h}-v'(t_j)\right| &= h\left|\int_0^1(1-\tau)v''(t_j+\tau h)\,d\tau\right|\\ &\le \frac{h}{2}\max_{\theta\in[t_j,t_{j+1}]}|v''(\theta)|\\ &= O(h),\quad j=i_0,\dots,N-1.\end{aligned}$$

Die Abschätzung des Integrals erfolgte mit Hilfe des erweiterten Mittelwertsatzes der Integralrechnung B.1.10. Dieser ist in der Form B.1.10 anwendbar, da die Integration komponentenweise erfolgt.

Für $l_{i_0}$ erhält man

$$\begin{aligned}|v(t_{i_0})-v(0)-t_{i_0}v'(0)| &= \left|t_{i_0}^2\int_0^1(1-\tau)v''(\tau t_{i_0})\,d\tau\right|\\ &\le \frac{t_{i_0}^2}{2}\max_{\theta\in[0,t_{i_0}]}|v''(\theta)|\\ &= O(h^2),\end{aligned}$$

und damit ist das Lemma vollständig bewiesen.

**Bemerkung:** Ist $g\in C^p([0,1],\mathbb{C}^{2n})$, dann folgt $v\in C^{p+1}([0,1],\mathbb{C}^{2n}), p\in\{0,1\}$, wegen $v(t)=E^{-1}(y(t),ty'(t))$, und damit können die Voraussetzungen von Lemma 5.1.8 auf Glattheitsforderungen für die Datenfunktion $g$ (und damit $f$) zurückgeführt werden. Die Bedingungen aus 3.3.4 gelten auch, wenn $M$ den (mehrfachen) Eigenwert 0 besitzt. Für variable Koeffizientenmatrix und das nichtlineare Problem kann man die entsprechenden Bedingungen für die Differentiationsklasse der Lösung den Sätzen 3.4.1 und 3.5.1 entnehmen.

Im Folgenden werden die soeben bewiesenen Resultate nur mehr zu einem Konvergenzbeweis für das Problem (5.1) zusammengefasst.

**Korollar 5.1.9** *Betrachte Gleichung (5.4), wobei $J$ nur Eigenwerte mit negativem Realteil besitze. Diese Gleichung werde mit dem expliziten* EULER-*Verfahren gelöst. Die so gewonnenen Näherungslösungen $v_h$ konvergieren für $v \in C^1([0,1],\mathbb{C}^{2n})$ gegen die exakte Lösung $v$ von (5.4). Gilt $v \in C^2([0,1],\mathbb{C}^{2n})$, dann ist die Konvergenzordnung gleich 1.*

*Beweis:* Die Eigenschaften des Operators $F_h$ aus (5.6), die in 5.1.6 und 5.1.7 komponentenweise bewiesen wurden und insgesamt zu (5.26) und (5.27) führten, sowie Lemma 5.1.8, liefern sofort die Aussage.

Definiert man $z_h := Ev_h = (Ev_{i_0}, \ldots, Ev_N)$, so sind offensichtlich diese $z_h$ (von gleicher Ordnung wie die $v_h$) konvergent gegen die Lösung $z$ von (5.2) wegen

$$\|z_h - R_h(z)\|_h = \|Ev_h - ER_h(v)\|_h \leq |E|\|v_h - R_h(v)\|_h.$$

Insbesondere gilt dies für die ersten $n$ Komponenten von $z(t) = (y(t), ty'(t))$, weshalb sich die Konvergenzresultate auch auf $y$ übertragen. Das ist Gegenstand des folgenden, diesen Abschnitt abschließenden Satz.

**Satz 5.1.10** *Sei $y$ die Lösung von (5.1). Sei weiters $y_h := I_1 Ev_h$, wobei $v_h$ die mit dem expliziten* EULER-*Verfahren berechneten Näherungslösungen der Gleichung (5.4) seien. Hat die Matrix $M$ aus (5.2) nur Eigenwerte mit negativem Realteil, dann gilt für $y \in C^2([0,1],\mathbb{R}^n)$*

$$\lim_{h\to 0} \|y_h - R_h(y)\|_h = 0.$$

*Ist $y \in C^3([0,1],\mathbb{R}^n)$, so gilt*

$$\|y_h - R_h(y)\|_h = O(h), \quad h \to 0.$$

*Beweis:* Folgt sofort aus Korollar 5.1.9.

**Bemerkungen:**

1. Man beachte, dass *alle* obigen Resultate vorerst *nur für Koeffizientenmatrizen $M$ bewiesen sind, die ausschließlich Eigenwerte mit negativem Realteil besitzen.* Für Matrizen $M$ mit Eigenwert 0 wird das korrespondierende Ergebnis im nächsten Abschnitt hergeleitet.

2. Wendet man das explizite EULER-Verfahren direkt (in vektorieller Form!) auf Gleichung (5.3) an, ohne auf die Gestalt (5.4) zu transformieren, so liefert das dasselbe Ergebnis wie die oben beschriebene Vorgangsweise, wegen

   $$z_{j+1} := Ev_{j+1} = Ev_j + hEJv_j + ht_jEg_j = z_j + hMz_j + ht_j\overset{\circ}{f}(t_j).$$

   Aus diesem Grund sind die Lösungen $z_h$ und $y_h$ auch *reell*, obwohl $v_h$ im Allgemeinen *komplex* ist.

**Eigenwert 0**

In diesem Abschnitt werden lediglich die Resultate bewiesen, die 5.1.6 und 5.1.7 entsprechen. Der restliche Konvergenzbeweis wird fast identisch den Überlegungen aus Abschnitt 5.1.1 verlaufen, und deshalb nicht ausführlich beschrieben.

Betrachtet werde die Lösung der (skalaren) Gleichungen

$$v'(t) = tg(t), \tag{5.28}$$
$$v'(t) = \frac{1}{t}w(t) + tg(t). \tag{5.29}$$

Zur Lösung von (5.28) definiere einen Operator

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1}-x_j}{h} - t_j g_j,\ j = i_0, \dots, N-1 \\ x_{i_0} - v_{h;i_0} \end{pmatrix}. \tag{5.30}$$

Die Lösung von (5.28) mit dem expliziten EULER-Verfahren ist also äquivalent zur Lösung von $F_h(v_h) = 0$. Sei wieder $v$ die exakte Lösung von (5.28), $v_h$ die Lösung von $F_h(v_h) = 0$ und $\varepsilon_h := R_h(v) - v_h$. Dann erfüllt $\varepsilon_h$ die Gleichung

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - l_{j+1},\ j = i_0, \dots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.31}$$

mit

$$l_h := F_h(R_h(v)) = \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - v'(t_j),\ j = i_0, \dots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0}v'(0) \end{pmatrix}.$$

Für die Stabilitätsabschätzung beweist man wieder ein etwas allgemeineres Resultat.

**Lemma 5.1.11** *Sei $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - t_j^{\gamma-1} l_{j+1},\ j = i_0, \dots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.32}$$

*gilt*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} l_{l+1}, \quad i = i_0, \dots, N.$$

*Weiters gilt die Abschätzung*

$$|\varepsilon_i| \le \text{const.}(|l_{i_0}| + t_i^{\gamma} \max_{i_0+1 \le k \le N} |l_k|).$$

*Beweis:* Die Lösungsdarstellung ist unmittelbar ersichtlich und die Abschätzung folgt aus Lemma 5.1.5.

Die Abschätzung für die Näherungslösung von (5.29) gestaltet sich etwas komplizierter. In diesem Fall treten nämlich logarithmische Terme auf. Mit welcher Potenz diese Terme eingehen, hängt davon ab, die wievielte Zeile einer JORDAN-Matrix $J = J_m(0)$ gerade betrachtet wird. Der Induktionsbeweis im nächsten Lemma zeigt, in welcher Weise sich diese Logarithmen von Komponente zu Komponente fortpflanzen.

**Lemma 5.1.12** *Sei $\gamma > 0$. Betrachte die Gleichung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{t_j}\delta_j - t_j^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0.$$

*Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \le \sum_{l=0}^{k} b_l |\ln(h)|^l + c t_i^{\gamma}, \quad i = i_0, \ldots, N$$

*mit $c, b_l > 0,\ l = 0, \ldots k$. Dann folgt die Lösungsdarstellung*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} l_{l+1}^*, \quad i = i_0, \ldots, N \tag{5.33}$$

*mit $l_l^* := t_{l-1}^{-\gamma}\delta_{l-1} + l_l,\ l = i_0 + 1, \ldots, N$, sowie die Abschätzung*

$$\begin{aligned} |\varepsilon_i| \ &\le\ |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_i^{\gamma} \max\{c, \max_{i_0+1 \le m \le N} |l_m|\}, \\ & \quad i = i_0, \ldots, N. \end{aligned} \tag{5.34}$$

*Beweis:* Die Lösungsdarstellung (5.33) ist wieder offensichtlich. Die Abschätzung (5.34) folgt aus

$$\begin{aligned} |\varepsilon_i| \ &\le\ |l_{i_0}| + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} |l_{l+1}^*| \\ &\le\ |l_{i_0}| + \sum_{l=i_0}^{i-1} h t_l^{-1} \sum_{m=0}^{k} b_m |\ln(h)|^m \\ &\quad + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} c + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} |l_{l+1}| \end{aligned}$$

und weiter mit Lemma 5.1.5 analog wie in Lemma 5.1.6 und unter Verwendung der Abschätzung

$$\left|\ln\left(\frac{t_k}{t_j}\right)\right| \leq |\ln(t_j)| \leq |\ln(h)+\ln(i_0)| \leq 2|\ln(h)|, \quad 0 < t_{i_0} \leq t_j < t_k \leq 1.$$

Sei jetzt $F_h$ definiert als

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1}-x_j}{h} - \frac{1}{t_j}Jx_j - t_j g_j,\ j = i_0, \ldots, N-1 \\ x_{i_0} - v_{h;i_0} \end{pmatrix}, \tag{5.35}$$

wobei $x_h \in \mathbb{C}^{m(N-i_0+1)}$ und $J = J_m(0)$ gilt. Wieder ist die Lösung von $F_h(v_h) = 0$ äquivalent zur Lösung der Gleichung (5.4) mittels des expliziten EULER-Verfahrens. Ist $v$ die exakte Lösung und $v_h$ die Näherungslösung, dann erfüllt der globale Fehler $\varepsilon_h := R_h(v) - v_h$ die Gleichung

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{t_j}J\varepsilon_j - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.36}$$

mit $l_h := F_h(R_h(v))$.

Mit Hilfe von Lemma 5.1.12 kann man nun die Lösung von

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{t_j}J\varepsilon_j - t_j^{\gamma-1}l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.37}$$

abschätzen.

**Lemma 5.1.13** *$\varepsilon_{j;l}$ bezeichne die $l$-te Komponente von $\varepsilon_j$ für $l = 1, \ldots, m$ (und für die anderen Vektoren sinngemäß). Dann gilt für die Lösung $\varepsilon_h$ von (5.37)*

$$\begin{aligned} |\varepsilon_{i;j}| &\leq \text{const.} \left( \sum_{l=j}^{m} \max_{l \leq k \leq m} |l_{i_0;k}| |\ln(h)|^{l-j} + t_i^{\gamma} \max_{j \leq k \leq m} \max_{i_0+1 \leq l \leq N} |l_{l;k}| \right) \\ &\leq \text{const.} \left( \sum_{l=j}^{m} |l_{i_0}| |\ln(h)|^{l-j} + t_i^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l| \right), \\ &\quad i = i_0, \ldots, N, \quad j = 1, \ldots, m. \end{aligned}$$

*Beweis:* Wird mittels vollständiger Induktion beginnend bei der $m$-ten Komponente geführt. Der Induktionsanfang kann Lemma 5.1.11 entnommen werden. Der Induktionsschritt von $j+1$ auf $j$ entspricht Lemma 5.1.12 für $k = m-j-1$.

Lemma 5.1.13 für $\gamma = 1$ zeigt also, dass der Operator $F_h$ aus (5.35) *nicht stabil* im Sinne von Definition 5.1.2 ist. Die Näherungslösungen $v_h$ konvergieren aber trotzdem gegen die exakte Lösung, da die logarithmischen Terme nur in Produkten mit den Anfangswerten $l_{i_0}$ auftreten, und diese nach Lemma 5.1.8 schneller als die anderen Komponenten von $l_h$ konvergieren. Damit folgt also auch im Fall, dass $J = J_m(0)$ gilt, das folgende Korollar.

**Korollar 5.1.14** *Betrachte Gleichung (5.4), wobei $J = J_{2n}(0)$ gelte. Diese Gleichung werde mit dem expliziten* EULER-*Verfahren gelöst. Die so gewonnenen Näherungslösungen $v_h$ konvergieren für $v \in C^1([0,1], \mathbb{C}^{2n})$ gegen die exakte Lösung $v$ von (5.4). Gilt $v \in C^2([0,1], \mathbb{C}^{2n})$, dann ist die Konvergenzordnung gleich 1.*

*Beweis:* Wegen $h|\ln(h)|^k = o(1)$ für $h \to 0$ $\forall k \in \mathbb{N} \cup \{0\}$, gilt für $v \in C^1([0,1], \mathbb{C}^{2n})$ nach Lemma 5.1.8 und nach Lemma 5.1.13

$$\begin{aligned} \|R_h(v) - v_h\|_h &\leq \text{const.}(|\ln(h)|^{2n}|l_{i_0}| + \max_{i_0+1 \leq k \leq N} |l_k|) \\ &\leq |\ln(h)|^{2n} o(h) + o(1) = o(1) \to 0, \quad h \to 0. \end{aligned}$$

Ist $v$ in $C^2([0,1], \mathbb{C}^{2n})$, dann gilt

$$\|R_h(v) - v_h\|_h \leq |\ln(h)|^{2n} O(h^2) + O(h) = O(h), \quad h \to 0.$$

Mit denselben Überlegungen wie in Abschnitt 5.1.1 erhält man das entsprechende Resultat für $y$ wie in Satz 5.1.10 jetzt auch für $J = J_m(0)$ und damit insgesamt für Systeme, die nur Eigenwerte mit negativem Realteil oder den Eigenwert 0 besitzen, das folgende Konvergenzresultat.

**Satz 5.1.15** *Sei $y$ die Lösung von (5.1). Sei weiters $y_h := I_1 E v_h$, wobei $v_h$ die mit dem expliziten* EULER-*Verfahren berechneten Näherungslösungen der Gleichung (5.4) seien. Hat die Matrix $M$ aus (5.2) nur Eigenwerte mit negativem Realteil oder den Eigenwert 0, dann gilt für $y \in C^2([0,1], \mathbb{R}^n)$*

$$\lim_{h \to 0} \|y_h - R_h(y)\|_h = 0.$$

*Gilt $y \in C^3([0,1], \mathbb{R}^n)$, dann folgt*

$$\|y_h - R_h(y)\|_h = O(h), \quad h \to 0.$$

*Beweis:* Analog wie Satz 5.1.10 mit Korollar 5.1.14.

### 5.1.2 Variable Koeffizientenmatrix

In diesem Abschnitt wird die Gleichung

$$\begin{aligned} y''(t) &= \frac{A_1(t)}{t} y'(t) + \frac{A_0(t)}{t^2} y(t) + f(t), \quad t \in (0,1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0 \end{aligned} \tag{5.38}$$

mit von $t$ abhängigen Matrizen $A_0(t), A_1(t)$ betrachtet. Zur Vereinfachung sei angenommen, dass

$$A_i(t) = A_i + t^\gamma C_i(t), \quad i \in \{0,1\} \tag{5.39}$$

mit $\gamma > 0, A_i \in \mathbb{R}^{n\times n}, C_i \in C([0,1],\mathbb{R}^{n\times n}),\ i = 0,1$ gilt[3].

Wieder wendet man die Transformation $z(t) := (y(t), ty'(t))$ an und erhält

$$\begin{aligned} z'(t) &= \frac{M(t)}{t} z(t) + \overset{\circ}{f}(t), \quad t \in (0,1], \\ z(0) &= z_0, \end{aligned} \tag{5.40}$$

mit

$$M(t) := \begin{pmatrix} 0 & I \\ A_0(t) & I + A_1(t) \end{pmatrix} =: M + t^\gamma \overset{\circ}{C}(t), \quad \overset{\circ}{f}(t) := \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.$$

Bringt man $M$ wieder auf JORDAN-Normalform (Satz B.2.2), so erhält man

$$\begin{aligned} v'(t) &= \frac{1}{t} J v(t) + t g(t) + t^{\gamma-1} C(t) v(t), \quad t \in (0,1], \\ v(0) &= E^{-1} z_0 =: v_0 \end{aligned} \tag{5.41}$$

mit $J = E^{-1}ME$ in JORDAN-Normalform, $v(t) := E^{-1}z(t)$, $g(t) := E^{-1}\overset{\circ}{f}(t)$ und $C(t) := E^{-1}\overset{\circ}{C}(t)E$.

Für die Berechnung der Näherungslösung von (5.41) mittels des expliziten EULER-Verfahrens ist also die Gleichung[4]

$$\begin{aligned} F_h(v_h) &:= \begin{pmatrix} \frac{v_{j+1}-v_j}{h} - \frac{1}{t_j} J v_j - t_j g_j - t_j^{\gamma-1} C(t_j) v_j,\ j = i_0, \dots, N-1 \\ v_{i_0} - v_0 - t_{i_0} v'(0) \end{pmatrix} \\ &= 0 \end{aligned} \tag{5.42}$$

zu lösen.

Um die Existenz einer eindeutigen Lösung von (5.42) zu beweisen und zu Abschätzungen für diese zu gelangen, wendet man die Fixpunktiteration gemäß Satz A.1.3 an.

Die Menge, auf der die Iteration durchgeführt wird, sei der (affine) Raum aller $x_h = (x_{i_0}, \dots, x_N) \in \mathbb{C}^{2n(N-i_0+1)}$ mit $x_{i_0} = v_0$. Dieser Raum werde kurz mit $X$ bezeichnet. Für $x_h, y_h \in X$ gelte die folgende Gleichung:

$$\begin{pmatrix} \frac{x_{j+1}-x_j}{h} - \frac{1}{t_j} J x_j - t_j g_j - t_j^{\gamma-1} C(t_j) y_j,\ j = i_0, \dots, N-1 \\ x_{i_0} - v(0) - t_{i_0} v'(0) \end{pmatrix} = 0. \tag{5.43}$$

[3] Ist $A_i \in C^1([0,1],\mathbb{R}^{n\times n})$, so folgt die Darstellung (5.39) direkt aus Satz B.1.7 mit $\gamma = 1$. Ist nämlich $A_i \in C^1([0,1],\mathbb{R}^{n\times n})$, dann ist $C_i(t) := \frac{A_i(t)-A_i(0)}{t}$ stetig auf $[0,1]$, siehe dazu auch [9, S. 189 sq.].

[4] $v'(0)$ ist wieder aus Kapitel 3 bekannt, siehe die Bemerkung auf Seite 28.

Man definiert jetzt eine Abbildung $G : X \to X$ so, dass $x_h = G(y_h)$ die Lösung von Gleichung (5.43) ist. Offensichtlich ist die Lösung von Gleichung (5.42) äquivalent zum Fixpunktproblem $v_h = G(v_h)$.

Um die eindeutige Existenz eines Fixpunkts von $G$ zu zeigen, wendet man den Fixpunktsatz von Banach A.1.3 an. Dazu ist zu zeigen, dass $G$ kontrahierend ist.

Seien $x_h, y_h \in X$. Dann ist der Ausdruck $v_h := G(x_h) - G(y_h)$ die Lösung der Gleichung

$$\begin{pmatrix} \frac{v_{j+1}-v_j}{h} - \frac{1}{t_j} J v_j - t_j^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ v_{i_0} \end{pmatrix} = 0 \tag{5.44}$$

mit

$$l_{j+1} := C(t_j)(x_j - y_j), \quad j = i_0, \ldots, N-1.$$

Für Gleichungen der Gestalt (5.44) wurden bereits in Abschnitt 5.1.1 Abschätzungen berechnet. Aus Lemma 5.1.6 und Lemma 5.1.7 sowie aus Lemma 5.1.11 und Lemma 5.1.12 erhält man direkt (da $l_{i_0} = 0$ gilt)

$$\begin{aligned} |v_i| &\leq c_1 t_i^{\gamma} \max_{i_0+1 \leq k \leq N} |l_k| \\ &\leq c_1 t_i^{\gamma} \|l_h\|_h \\ &\leq c_2 t_i^{\gamma} \|x_h - y_h\|_h, \quad i = i_0, \ldots, N. \end{aligned}$$

Wegen $\gamma > 0$ gibt es ein $\delta > 0$, sodass

$$c_2 t^{\gamma} < 1 \quad \forall t \leq \delta$$

gilt. Wenn man also mit der obigen Vorgangsweise nur auf dem Intervall $(0, \delta]$ rechnet anstatt auf $(0, 1]$ und deshalb auch nur mehr Gitter auf $(0, \delta]$ betrachtet, was natürlich nichts an den obigen Überlegungen ändert, und setzt $L := c_2 \delta^{\gamma} < 1$, so gilt also

$$\|G(x_h) - G(y_h)\|_h \leq L \|x_h - y_h\| \quad \forall x_h, y_h \in X,$$

und die Abbildung $G$ ist kontrahierend gemäß Definition A.1.2. Aus dem Fixpunktsatz von Banach folgt also, dass $G$ einen eindeutig bestimmten Fixpunkt $v_h \in X$ besitzt. Darüberhinaus ist $v_h$ die eindeutige Lösung von (5.42).

Im Folgenden soll aufbauend auf den vorangegangenen Resultaten wieder der globale Fehler $\varepsilon_h := R_h(v) - v_h$ abgeschätzt werden. $\varepsilon_h$ erfüllt die Gleichung

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{t_j} J \varepsilon_j - t_j^{\gamma-1} C(t_j) \varepsilon_j - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0, \tag{5.45}$$

wobei wieder

$$l_h := F_h(R_h(v)) = \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - v'(t_j),\ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v(0) - t_{i_0} v'(0) \end{pmatrix}$$

gilt.

Ganz analog wie für Gleichung (5.42) (die Abbildung $G$ ist dabei sinngemäß definiert) beweist man die Existenz einer eindeutigen Lösung $\varepsilon_h$ von (5.45) mit $G(\varepsilon_h) = \varepsilon_h$ und

$$\|G(x_h) - G(y_h)\|_h \leq L\|x_h - y_h\|_h \quad \forall x_h, y_h \in X$$

mit $L < 1$.

Um $\varepsilon_h$ unter Verwendung dieser Ungleichung abzuschätzen, geht man folgendermaßen vor:

Für $\tilde{0} := (l_{i_0}, 0, \ldots, 0) \in \mathbb{C}^{2n(N-i_0+1)}$ gilt

$$\begin{aligned} \|\varepsilon_h\|_h - \|G(\tilde{0})\|_h &= \|G(\varepsilon_h)\|_h - \|G(\tilde{0})\|_h \leq \|G(\varepsilon_h) - G(\tilde{0})\|_h \\ &\leq L \max_{i_0+1\leq k\leq N} |\varepsilon_k| \leq L\|\varepsilon_h\|_h, \end{aligned} \tag{5.46}$$

mit $L < 1$. Daraus folgt nun sofort, dass gilt

$$\|\varepsilon_h\|_h \leq \frac{1}{1-L}\|G(\tilde{0})\|_h.$$

$x_h = G(\tilde{0})$ ist die Lösung der Gleichung

$$\left( \begin{array}{l} \frac{x_{j+1}-x_j}{h} - \frac{1}{t_j} J x_j - t_j^{\gamma-1} C(t_j)\tilde{0}_j - l_{j+1},\ j = i_0, \ldots, N-1 \\ x_{i_0} - l_{i_0} \end{array} \right) = 0.$$

Diese Gleichung ist linear und inhomogen. Es gilt das Superpositionsprinzip, und man kann $x_h$ aufspalten in zwei Summanden $x_h^{(1)}$ und $x_h^{(2)}$, die jeweils die Gleichung für die Inhomogenität $l_h$ und für

$$\left( \begin{array}{l} t_j^{\gamma-1} C(t_j)\tilde{0}_j,\ j = i_0, \ldots, N-1 \\ 0 \end{array} \right)$$

lösen. Für beide Lösungskomponenten sind aus den vorangehenden Überlegungen schon Abschätzungen bekannt, und zwar

$$\begin{aligned} \|x_h^{(1)}\|_h &\leq \text{const.}(|\ln(h)|^{2n}|l_{i_0}| + \max_{i_0+1\leq k\leq N} |l_k|), \\ \|x_h^{(2)}\|_h &\leq \text{const.}|l_{i_0}|. \end{aligned}$$

Insgesamt folgt also

$$\|x_h\| \leq \text{const.}(|\ln(h)|^{2n}|l_{i_0}| + \max_{i_0+1\leq k\leq N} |l_k|).$$

Für die Konsistenz des Operators $F_h$ kann der Beweis von Lemma 5.1.8 unter Verwendung von Satz 3.4.1 direkt übernommen werden. Dabei ist Satz 3.4.1 zu beachten.

Damit kann man also wieder die Existenz von eindeutigen Näherungslösungen $v_h$ und deren Konvergenz gegen $v$ zeigen. Man beachte aber, dass die Stabilitätsresultate dieses Abschnitts *nur auf einem geeigneten Intervall* $(0, \delta]$ *Gültigkeit besitzen.* Auf jedem Intervall $[\delta, 1], \delta > 0$, ist jedoch die Verfahrensfunktion des expliziten EULER-Verfahrens hinreichend glatt, um die Resultate aus Kapitel 4 verwenden zu können[5]. Es ist also auch auf dem Rest des Intervalls $(0, 1]$ die Konvergenz des expliziten EULER-Verfahrens gesichert. Damit folgt der nächste Satz.

**Satz 5.1.16** *Sei $y$ die Lösung von (5.38). Sei weiters $y_h := I_1 E v_h$, wobei $v_h$ die mit dem expliziten* EULER-*Verfahren berechneten Näherungslösungen der Gleichung (5.41) seien. Hat die Matrix $M \in \mathbb{R}^{n\times n}$ aus (5.40) nur Eigenwerte mit negativem Realteil oder den Eigenwert 0, dann gilt für $y \in C^2([0,1], \mathbb{R}^n)$*

$$\lim_{h\to 0} \|y_h - R_h(y)\|_h = 0.$$

*Gilt sogar $y \in C^3([0,1], \mathbb{R}^n)$, dann folgt*

$$\|y_h - R_h(y)\|_h = O(h), \quad h \to 0.$$

*Beweis:* Analog wie Satz 5.1.15.

**Bemerkung:** Da die Näherungslösung $v_h$ eindeutig ist, braucht sie natürlich nicht über den Umweg einer Fixpunktiteration ermittelt zu werden, sondern kann direkt aus (5.42) bzw. durch das EULER-Verfahren direkt auf (5.40) angewendet berechnet werden.

### 5.1.3 Das nichtlineare Problem

In diesem Abschnitt wird die nichtlineare Gleichung

$$\begin{aligned} y''(t) &= \frac{A_1}{t} y'(t) + \frac{A_0}{t^2} y(t) + f(t, y(t)), \quad t \in (0,1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0 \end{aligned} \tag{5.47}$$

untersucht, wobei $A_0, A_1 \in \mathbb{R}^{n\times n}$ konstant seien. Wieder transformiert man auf die Gleichung erster Ordnung

$$\begin{aligned} z'(t) &= \frac{M}{t} z(t) + \overset{\circ}{tf}(t, z(t)), \quad t \in (0,1], \\ z(0) &= z_0, \end{aligned} \tag{5.48}$$

[5]Man vergleiche als Beispiel dafür die Betrachtungen aus Abschnitt 4.6.

mit

$$M := \begin{pmatrix} 0 & I \\ A_0 & I + A_1 \end{pmatrix}, \quad \overset{\circ}{f}(t, z(t)) := \begin{pmatrix} 0 \\ f(t, z_1(t)) \end{pmatrix}.$$

In diesem Fall ist es für die weiteren Überlegungen nicht notwendig, auf JORDAN-Normalform zu transformieren. Somit bleiben alle auftretenden Größen *reell*, was im nichtlinearen Fall von entscheidender Bedeutung ist. Sonst müsste man nämlich eine Fortsetzung von $\overset{\circ}{f}$ auf komplexe Argumente betrachten.

Die Konvergenz des expliziten EULER-Verfahrens für (5.48) wird mittels der Linearisierung des diskreten Lösungsoperators gezeigt. Dazu wird der folgende Satz, dessen Beweis aus [29, S. 469] stammt, verwendet. Dies ist für Eigenwerte mit negativem Realteil ohne Modifikation möglich, ebenso für den Eigenwert 0, wenn die Dimension der entsprechenden JORDAN-Matrizen gleich 1 ist. Im Falle, dass die JORDAN-Normalform von $M$ JORDAN-Matrizen $J_m(0)$ mit $m > 1$ enthält, ist es nötig, den Beweis von Satz 5.1.17 geringfügig abzuwandeln. Obwohl eine der Voraussetzungen verletzt ist, gelingt der Beweis der Konvergenz aufgrund der speziellen Eigenschaften des konkreten Problems dennoch. Es wird dabei die Notation der Definitionen 5.1.2 und 5.1.1 verwendet. Darüberhinaus bezeichne $DF_h(v_h)$ die FRÉCHET-Ableitung (oder *Linearisierung*) eines Operators $F_h$ im Punkt $v_h$.

**Satz 5.1.17** *Sei $v$ Lösung der Gleichung $F(v) = 0$. Die Familie $F_h$ sei konsistent mit der Aufgabe $F(v) = 0$, also*

$$\|F_h(R_1^h(v))\|_2^h \leq M(h), \quad \lim_{h \to 0} M(h) = 0.$$

*Die Familie $F_h$ besitze* FRÉCHET*-Ableitungen $DF_h(x_h)$ in einer Kugel $\overline{K}(R_1^h(v), \rho_0) \subseteq (E_1^h, \| \ \|_1^h)$ und für ein $h_0 > 0$ gelte für alle $h \in (0, h_0]$:*

1. *$DF_h(R_1^h(v))$ haben (gleichmäßig) beschränkte Inverse in den Mittelpunkten der jeweiligen Kugeln, d. h. es gibt ein $K_0 > 0$ mit*

   $$\|DF_h^{-1}(R_1^h(v))\|_{2,1}^h \leq K_0.$$

   *$\| \ \|_{1,2}^h$ bezeichnet dabei die von $\| \ \|_1^h$ und $\| \ \|_2^h$ induzierte Operatornorm, d. h. für einen linearen Operator $T : E_1^h \to E_2^h$ gilt*

   $$\|T\|_{1,2}^h = \sup_{y \in E_1^h \setminus \{0\}} \frac{\|T(y)\|_2^h}{\|y\|_1^h}.$$

2. *$DF_h$ ist (gleichmäßig)* HÖLDER*-stetig auf $\overline{K}(R_1^h(v), \rho_0)$, d. h. es gilt*

   $$\begin{aligned} \|DF_h(x_h) - DF_h(y_h)\|_{1,2}^h \ &\leq \ K_L(\|x_h - y_h\|_1^h)^\alpha, \\ &\quad 0 < \alpha \leq 1, \quad \forall x_h, y_h \in \overline{K}(R_1^h(v), \rho_0). \end{aligned}$$

*Dann gibt es für hinreichend kleines $h_0$ und $\rho_0$ für jedes $h \in (0, h_0]$ eine eindeutige Lösung $v_h$ von $F_h(v_h) = 0$ in $\overline{K}(R_1^h(v), \rho_0)$. Für diese gilt*

$$\|R_1^h(v) - v_h\|_1^h \leq \text{const.} M(h) \to 0, \quad h \to 0.$$

*Beweis:* Der folgende Beweis stammt aus [29, S. 469]. Definiere eine Abbildung $G_h : E_1^h \to E_1^h$ durch

$$G_h(x_h) := x_h - DF_h^{-1}(R_1^h(v))F_h(x_h).$$

Die Lösung des Fixpunktproblems $v_h = G_h(v_h)$ ist offensichtlich äquivalent zur Lösung von $F_h(v_h) = 0$. Um die Existenz eines solchen (eindeutigen) Fixpunktes zeigen zu können, soll der Fixpunktsatz von Banach A.1.3 angewendet werden. Es ist also zu zeigen, dass die Abbildung $G_h$ kontrahierend im Sinne von Definition A.1.2 ist.

Nach dem Satz von Taylor B.1.7 gilt für $x_h, y_h \in \overline{K}(R_1^h(v), \rho_0)$

$$G_h(x_h) - G_h(y_h) = DF_h^{-1}(R_1^h(v))(DF_h(R_1^h(v)) - \widehat{DF}_h(x_h, y_h))(x_h - y_h)$$

mit

$$\widehat{DF}_h(x_h, y_h) := \int_0^1 DF_h(\tau x_h + (1-\tau)y_h)\, d\tau.$$

Aus der Hölder-Stetigkeit von $DF_h$ folgt die Abschätzung

$$\begin{aligned}
\|DF_h(R_1^h(v)) - \widehat{DF}_h(x_h, y_h)\|_{1,2}^h &\leq \int_0^1 \|DF_h(\tau R_1^h(v) + (1-\tau)R_1^h(v)) \\
&\quad - DF_h(\tau x_h + (1-\tau)y_h)\|_{1,2}^h \, d\tau \\
&\leq \int_0^1 K_L(\|\tau(R_1^h(v) - x_h) \\
&\quad + (1-\tau)(R_1^h(v) - y_h)\|_1^h)^\alpha \, d\tau \\
&\leq K_L \rho_0^\alpha \quad \forall x_h, y_h \in \overline{K}(R_1^h(v), \rho_0),
\end{aligned}$$

und daraus

$$\|G_h(x_h) - G_h(y_h)\|_1^h \leq K_0 K_L \rho_0^\alpha \|x_h - y_h\|_1^h =: L\|x_h - y_h\|_1^h.$$

Wähle jetzt $\rho_0$ so klein, dass $L < 1$ gilt, dann ist also $G$ eine Kontraktion auf $\overline{K}(R_1^h(v), \rho_0)$.

Jetzt bleibt noch zu zeigen, dass $G_h$ eine Selbstabbildung von $\overline{K}(R_1^h(v), \rho_0)$ ist. Das kann über die Wahl der Schrittweite $h$ erreicht werden, wie die nächste Abschätzung zeigt. Sei $x_h \in \overline{K}(R_1^h(v), \rho_0)$, dann gilt

$$\begin{aligned}
\|R_1^h(v) - G_h(x_h)\|_1^h &\leq \|R_1^h(v) - G_h(R_1^h(v))\|_1^h + \|G_h(R_1^h(v)) - G_h(x_h)\|_1^h \\
&\leq K_0 \|F_h(R_1^h(v))\|_2^h + L\|x_h - R_1^h(v)\|_1^h \\
&\leq K_0 M(h) + L\rho_0.
\end{aligned}$$

Ist also $h_0$ so klein, dass

$$K_0 M(h) \leq (1-L)\rho_0 \quad \forall h \leq h_0$$

gilt, dann ist $G_h$ für $h \leq h_0$ eine *kontrahierende Selbstabbildung auf* $\overline{K}(R_1^h(v), \rho_0)$.

Damit folgt aus dem Fixpunktsatz von BANACH die Existenz eines eindeutigen Fixpunktes $v_h = G_h(v_h)$, der die eindeutige Lösung von $F_h(v_h) = 0$ ist. Die Abschätzung, die die Konvergenz dieser Näherungslösungen zeigt, erhält man mittels

$$\begin{aligned} \|R_1^h(v) - v_h\|_1^h &= \|R_1^h(v) - G_h(v_h)\|_1^h \leq K_0 M(h) + L\|R_1^h(v) - v_h\|_1^h \quad \Rightarrow \\ \|R_1^h(v) - v_h\|_1^h &\leq \frac{K_0}{1-L} M(h). \end{aligned}$$

Um den obigen Satz auf das Problem der Lösung von (5.48) mit dem EULER-Verfahren anwenden zu können, definiert man folgende Operatoren:

$$F(x) := \begin{pmatrix} x'(t) - \frac{1}{t}Mx(t) - t\overset{\circ}{f}(t, x(t)),\ t \in (0,1] \\ x(0) - z_0 \end{pmatrix}, \tag{5.49}$$

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1}-x_j}{h} - \frac{1}{t_j}Mx_j - t_j\overset{\circ}{f}(t_j, x_j),\ j = i_0, \ldots, N-1 \\ x_{i_0} - z(0) - t_{i_0} z'(0) \end{pmatrix}. \tag{5.50}$$

Die Lösung von (5.48) ist äquivalent zu $F(z) = 0$, die Berechnung der Approximationslösungen mittels des expliziten EULER-Verfahrens zu $F_h(z_h) = 0$. Für die FRÉCHET-Ableitung von $F_h$ gilt offensichtlich

$$DF_h(x_h)y_h = \begin{pmatrix} \frac{y_{j+1}-y_j}{h} - \frac{1}{t_j}My_j - t_j D_2\overset{\circ}{f}(t_j, x_j)y_j,\ j = i_0, \ldots, N-1 \\ y_{i_0} \end{pmatrix},$$

wobei $D_2\overset{\circ}{f}(t,x)$ für die FRÉCHET-Ableitung der Funktion $\overset{\circ}{f}$ nach dem zweiten Argument im Punkt $(t,x)$ steht.

Nun versucht man, für die soeben eingeführten Operatoren die Voraussetzungen von Satz 5.1.17 zu zeigen. Dies geschieht zuerst für den Fall, dass die Matrix $M$ nur Eigenwerte mit negativem Realteil besitzt, dann für JORDAN-Matrizen $J = J_m(0)$ und abschließend für den allgemeinen Fall, da sich bei den Stabilitätsresultaten ja schon im Abschnitt 5.1.2 Unterschiede für die beiden ersten Fälle ergeben hatten.

Die Konsistenzresultate folgen in jedem Fall genauso wie in Lemma 5.1.8 unter Verwendung von Satz 3.5.1.

Die HÖLDER-Stetigkeit von $DF_h$ folgt, falls die lineare Abbildung $D_2\overset{\circ}{f}(t,x)$ HÖLDER-stetig bezüglich $x$ auf $\overline{K}(v,\rho_0) \subset (C([0,1],\mathbb{R}^{2n}),\| \ \|_\infty)$ ist, mittels der folgenden Abschätzung:

$$\begin{aligned}
\|(DF_h(x_h) - DF_h(y_h))z_h\|_h &\leq \max_{i_0\leq k\leq N-1} |(t_k D_2\overset{\circ}{f}(t_k,x_k) - t_k D_2\overset{\circ}{f}(t_k,y_k))z_k| \\
&\leq \max_{i_0\leq k\leq N-1} |t_k D_2\overset{\circ}{f}(t_k,x_k) - t_k D_2\overset{\circ}{f}(t_k,y_k)||z_k| \\
&\leq \max_{i_0\leq k\leq N-1} |D_2\overset{\circ}{f}(t_k,x_k) - D_2\overset{\circ}{f}(t_k,y_k)|\|z_h\|_h \\
&\leq \max_{i_0\leq k\leq N-1} K_L|x_k - y_k|^\alpha \|z_h\|_h \\
&\leq K_L\|x_h - y_h\|_h^\alpha \|z_h\|_h.
\end{aligned}$$

**Bemerkung:** Mit der HÖLDER-Stetigkeit von $D_2\overset{\circ}{f}$ ist also folgendes gemeint: $D_2\overset{\circ}{f}(t,x)$ ist für alle $(t,x)$ eine lineare Abbildung von $\mathbb{R}^{2n}$ in $\mathbb{R}^{2n}$. Bezeichnet $|D_2\overset{\circ}{f}(t,x)|$ die Zeilensummennorm dieser Abbildung, dann gelte

$$|D_2\overset{\circ}{f}(t,x) - D_2\overset{\circ}{f}(t,y)| \leq K_L|x-y|^\alpha \quad \forall (t,x),(t,y) \in [0,1]\times\overline{K}(v,\rho_0).$$

Nun ist noch die Beschränktheit der Inversen der linearen Abbildung $DF_h(R_h(z))$ zu zeigen.

Sei also $y_h = DF_h^{-1}(R_h(z))l_h$. Dann ist $y_h$ die Lösung der Gleichung

$$\begin{pmatrix} \frac{y_{j+1}-y_j}{h} - \frac{1}{t_j}My_j - t_jD_2\overset{\circ}{f}(t_j,z(t_j))y_j - l_{j+1},\ j = i_0,\ldots,N-1 \\ y_{i_0} - l_{i_0} \end{pmatrix} = 0.$$

Gleichungen dieser Bauart wurden bereits in Abschnitt 5.1.2 untersucht (cf. Gleichung (5.45)). Falls $D_2\overset{\circ}{f} \in C([0,1],\mathbb{R}^{2n\times 2n})$ ist, kann man die entsprechenden Ergebnisse für Eigenwerte mit negativem Realteil anwenden und erhält

$$\|y_h\|_h \leq c\|l_h\|_h.$$

Damit gilt also für die induzierte Operatornorm $\| \ \|$

$$\|DF_h^{-1}(R_h(z))\| \leq c.$$

Im Fall, dass alle Eigenwerte von $M$ negativen Realteil haben, sind also die Voraussetzungen von Satz 5.1.17 erfüllt, und es folgt die Konvergenz der (eindeutigen) Näherungslösungen $z_h$ gegen die exakte Lösung $z$ von (5.48). Für $y$ folgt damit der nächste Satz:

**Satz 5.1.18** *Sei $y$ die Lösung von (5.47). Sei weiters $y_h := I_1 z_h$, wobei $z_h$ die mit dem expliziten* EULER-*Verfahren berechneten Näherungslösungen der Gleichung (5.48) seien. Sei $\overset{\circ}{f}$ definiert wie in (5.48) und $D_2\overset{\circ}{f}(t,x)$* HÖLDER*-stetig bezüglich $x$. Hat die Matrix $M$ aus (5.48) nur Eigenwerte mit negativem Realteil, dann gilt für $y \in C^2([0,1],\mathbb{R}^n)$*

$$\lim_{h\to 0} \|y_h - R_h(y)\|_h = 0.$$

*Gilt $y \in C^3([0,1],\mathbb{R}^n)$, dann folgt*

$$\|y_h - R_h(y)\|_h = O(h), \quad h \to 0.$$

*Beweis:* Folgt direkt aus den obigen Überlegungen für $z_h$.

**Bemerkung:** Die obigen Resultate sind nur *lokal* in dem Sinne, dass sie nur in einer Umgebung der exakten Lösung und für kleine Schrittweiten gelten.

Im Fall, dass in der JORDAN-Normalform von $M$ JORDAN-Matrizen $J_m(0)$ mit $m > 1$ auftreten, kann Satz 5.1.17 nicht angewendet werden, da in der Abschätzung der Inversen des Operators $DF_h(R_h(z))$ logarithmische Terme auftreten. Man kann jedoch den Beweis (siehe dazu 5.1.17) so modifizieren, dass man trotzdem das Konvergenzresultat erhält. Die logarithmischen Terme treten nämlich nur bei den „Anfangswerten“ auf, und diese heben sich im konkreten Fall entweder weg, oder es kann die Tatsache ausgenützt werden, dass die Anfangswertkomponente des Fehlers $F_h(R_h(z))$ eine höhere Konvergenzordnung besitzt als die restlichen Komponenten. Dies geschieht im nächsten Satz.

**Satz 5.1.19** *Sei $z$ die exakte Lösung von (5.48) und sei $M = J_m(0), m \geq 1$. $D_2\overset{\circ}{f}(t,x)$ sei* HÖLDER*-stetig bezüglich der zweiten Variable. Dann gilt für ein hinreichend kleines $\rho_0$ und $h_0$, dass die Gleichung $F_h(z_h) = 0$ für $h \in (0,h_0]$ eine eindeutige Lösung in $\overline{K}(R_h(z),\rho_0)$ besitzt. Für die Lösung gilt*

$$\lim_{h\to 0} \|z_h - R_h(z)\|_h = 0,$$

*falls $z \in C^1([0,1],\mathbb{R}^{2n})$ gilt. Gilt sogar $z \in C^2([0,1],\mathbb{R}^{2n})$, dann gilt*

$$\|z_h - R_h(z)\|_h = O(h), \quad h \to 0.$$

*Beweis:* Es wird hier besonders auf die Unterschiede zu Satz 5.1.17 hingewiesen, für weitere Details siehe den Beweis 5.1.17.

Definiere wie in 5.1.17 die Abbildung

$$G_h(x_h) := x_h - DF_h^{-1}(R_h(z))F_h(x_h).$$

Es soll wieder gezeigt werden, dass $G_h$ eine Kontraktion auf einer geeigneten Kugel um die exakte Lösung ist. Es gilt

$$v_h := G_h(x_h) - G_h(y_h) = DF_h^{-1}(R_h(z))(DF_h(R_h(z)) - \widehat{DF}_h(x_h, y_h))(x_h - y_h)$$

mit

$$\widehat{DF}_h(x_h, y_h) := \int_0^1 DF_h(\tau x_h + (1-\tau)y_h)\, d\tau.$$

Dieser Ausdruck ist die Lösung $v_h$ der Gleichung

$$\begin{aligned} DF_h(R_h(z))v_h &= (DF_h(R_h(z)) - \widehat{DF}_h(x_h, y_h))(x_h - y_h) \\ &=: \begin{pmatrix} g_{j+1},\ j = i_0, \ldots, N-1 \\ g_{i_0} \end{pmatrix}. \end{aligned}$$

Dabei gilt $g_{i_0} = 0$, wie man sofort sieht. Aus den Beweisen aus Abschnitt 5.1.2 folgt

$$\begin{aligned} \|v_h\|_h &\leq K_0(|\ln(h)|^{m-1}|g_{i_0}| + \max_{i_0+1\leq k\leq N} |g_k|) \\ &= K_0 \max_{i_0+1\leq k\leq N} |g_k| \\ &\leq K_0 K_L \rho_0^\alpha \|x_h - y_h\|_h, \end{aligned}$$

analog wie in 5.1.17. Also ist $G_h$ wieder kontrahierend auf einer Kugel $\overline{K}(R_h(z), \rho_0)$ für geeignetes $\rho_0$. Über die Wahl einer hinreichend kleinen Schrittweite zeigt man wieder, dass $G_h$ auch eine Selbstabbildung auf dieser Kugel ist, und zwar

$$\begin{aligned} \|R_h(z) - G_h(x_h)\|_h &\leq \|DF_h^{-1}(R_h(z))F_h(R_h(z))\|_h + L\|R_h(z) - x_h\|_h \\ &\leq c_1|\ln(h)|^{m-1}|l_{i_0}| + c_2 \max_{i_0+1\leq k\leq N} |l_k| + L\rho_0, \end{aligned}$$

mit $l_h := F_h(R_h(z))$. Aus den Konsistenzresultaten von Lemma 5.1.8 folgt also, dass man für ein hinreichend kleines $h$ eine Selbstabbildung von $\overline{K}(R_h(z), \rho_0)$ erhält. Aus Satz A.1.3 folgt also die Existenz einer eindeutigen Lösung von $F_h(z_h) = 0$. Die Konvergenzabschätzung verläuft aufbauend auf der letzten Abschätzung genauso wie in 5.1.17.

Damit folgt schließlich das Konvergenzresultat für allgemeines Spektrum von $M$.

**Satz 5.1.20** *Sei $y$ die Lösung von (5.47). Sei weiters $y_h := I_1 z_h$, wobei $z_h$ die mit dem expliziten* EULER*-Verfahren berechneten Näherungslösungen der Gleichung (5.48) seien. Sei $\overset{\circ}{f}$ definiert wie in (5.48) und $D_2\overset{\circ}{f}(t,x)$* HÖLDER*-stetig bezüglich $x$. Hat die Matrix $M$ aus (5.48) nur Eigenwerte mit negativem Realteil oder den Eigenwert 0, dann gilt für $y \in C^2([0,1], \mathbb{R}^n)$*

$$\lim_{h\to 0} \|y_h - R_h(y)\|_h = 0.$$

*Gilt $y \in C^3([0,1], \mathbb{R}^n)$, dann ist*

$$\|y_h - R_h(y)\|_h = O(h), \quad h \to 0.$$

*Beweis:* Folgt aus den vorangegangenen Überlegungen.

**Bemerkung:** Manchmal ist es auch sinnvoll, Probleme der Gestalt

$$\begin{aligned} y''(t) &= \frac{A_1(t)}{t} y'(t) + \frac{A_0(t)}{t^2} y(t) + f(t, y(t)), \quad t \in (0,1], \\ B_0 y(0) &= \beta, \\ A_0 y(0) &= 0, \quad y'(0) = 0 \end{aligned} \tag{5.51}$$

zu betrachten. Erfüllt $A_i(t), i = 0, 1$ die gleichen Bedingungen wie in Abschnitt 5.1.2 und $f(t, y)$ dieselben Bedingungen wie in Abschnitt 5.1.3, dann folgt das Konvergenzresultat von Satz 5.1.20 ganz analog zum Beweis dieses Satzes.

## 5.2 Das implizite EULER-Verfahren

Die Überlegungen in diesem Abschnitt verlaufen über weite Strecken gleich wie im Abschnitt 5.1, wenn man die untersuchten Operatoren sinngemäß definiert, deshalb werden hier nur die Resultate, die Lemma 5.1.6, 5.1.7, 5.1.11, 5.1.12, sowie Lemma 5.1.8 entsprechen, ohne weitere Zwischenbemerkungen bewiesen. Alle anderen Resultate lassen sich analog wie in 5.1 daraus ableiten.

Betrachte also den Fall, dass die Koeffizientenmatrix $M$ konstant und das Problem linear ist.

Untersucht wird die (schon wie im vorigen Abschnitt teilweise entkoppelte) Gleichung

$$\begin{aligned} v'(t) &= \frac{1}{t} J v(t) + t g(t), \quad t \in (0,1], \\ v(0) &= v_0. \end{aligned} \tag{5.52}$$

Betrachte den Operator

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1} - x_j}{h} - \frac{1}{t_{j+1}} J x_{j+1} - t_{j+1} g_{j+1}, \ j = i_0, \ldots, N-1 \\ x_{i_0} - v_{h;i_0} \end{pmatrix}. \tag{5.53}$$

Dann ist die Lösung von (5.52) mit dem impliziten EULER-Verfahren (siehe Definition 4.3.1) äquivalent zur Lösung von $F_h(v_h) = 0$. Dabei sei $v_{h;i_0}$ als Ergebnis eines Schrittes des impliziten Eulerverfahrens von 0 weg gewählt. Dieser Anfangswert wird deshalb gesondert behandelt, um die formale Übereinstimmung

mit Abschnitt 5.1 zu erhalten. Um die Konvergenz der Näherungslösungen $v_h$ zu zeigen, geht man genau so vor wie in Abschnitt 5.1. Es ist offensichtlich, dass der einzige Unterschied in der Lösungsdarstellung und Abschätzung der skalaren Gleichungen besteht. Weiters gibt es in den Überlegungen zur Konsistenz eine formale Abweichung, deshalb wird dieser Aspekt eingehender diskutiert.

Es folgt nun ein Hilfslemma und die für den Konvergenzbeweis benötigten Resultate.

**Lemma 5.2.1** *Sei $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ mit $\sigma = \Re(\lambda) > 0 festgeähltt$. Definiere für $j \geq k \geq 1$*

$$\tilde{z}_{kj}(\lambda) := \begin{cases} 1, & k = j, \\ \prod_{l=k}^{j-1} \left(1 + \frac{\lambda}{l}\right)^{-1}, & 1 \leq k < j, \quad j = 2, 3, \ldots. \end{cases}$$

*Dann gibt es ein $\eta > 0$ und ein $C \geq 1$, sodass*

$$|\tilde{z}_{kj}(\lambda)| \leq C \left(\frac{k}{j}\right)^{\eta}, \quad 1 \leq k \leq j, \quad j = 1, 2, \ldots. \tag{5.54}$$

*Beweis:* Siehe Lemma A.2.3.

**Lemma 5.2.2** *Sei $\lambda \in \mathbb{C}$ mit $\Re(\lambda) < 0$ und $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1} - \varepsilon_j}{h} - \frac{\lambda}{t_{j+1}} \varepsilon_{j+1} - t_{j+1}^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.55}$$

*gilt*

$$\begin{aligned} \varepsilon_i &= \prod_{l=i_0+1}^{i} \left(1 - \frac{h\lambda}{t_l}\right)^{-1} l_{i_0} + \sum_{l=i_0+1}^{i} \prod_{k=l}^{i} \left(1 - \frac{h\lambda}{t_k}\right)^{-1} h t_l^{\gamma-1} l_l \\ &=: \tilde{z}_{i_0+1,i+1}(-\lambda) l_{i_0} + \sum_{l=i_0+1}^{i} \tilde{z}_{l,i+1}(-\lambda) h t_l^{\gamma-1} l_l, \quad i = i_0 + 1, \ldots, N. \end{aligned} \tag{5.56}$$

*$\tilde{z}_{lk}(-\lambda)$ ist wie in Lemma 5.2.1 definiert. Produkte bzw. Summen, bei denen der obere Index kleiner als der untere ist, sind als leer aufzufassen, d. h. solche Summen sind gleich 0, solche Produkte gleich 1.*

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_i| &\leq \text{const.}(|l_{i_0}| + t_{i+1}^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l|) && (5.57) \\ &\leq \text{const.} \|l_h\|_h, \quad i = i_0, \ldots, N. && (5.58) \end{aligned}$$

*Beweis:* Die Lösungsdarstellung folgt sofort mittels vollständiger Induktion: Für $i = i_0 + 1$ ist (5.56) offensichtlich. Gelte (5.56) für $\varepsilon_{i-1}$ und betrachte $\varepsilon_i$:

$$\begin{aligned}
\varepsilon_i &= \left(1 - \frac{h\lambda}{t_i}\right)^{-1} (\varepsilon_{i-1} + ht_i^{\gamma-1} l_i) \\
&= \left(1 - \frac{h\lambda}{t_i}\right)^{-1} \left( \prod_{l=i_0+1}^{i-1} \left(1 - \frac{h\lambda}{t_l}\right)^{-1} l_{i_0} + \sum_{l=i_0+1}^{i-1} \prod_{k=l}^{i-1} \left(1 - \frac{h\lambda}{t_k}\right)^{-1} ht_l^{\gamma-1} l_l + ht_i^{\gamma-1} l_i \right) \\
&= \prod_{l=i_0+1}^{i} \left(1 - \frac{h\lambda}{t_l}\right)^{-1} l_{i_0} + \sum_{l=i_0+1}^{i} \prod_{k=l}^{i} \left(1 - \frac{h\lambda}{t_k}\right)^{-1} ht_l^{\gamma-1} l_l.
\end{aligned}$$

Daraus folgt mit Lemma 5.2.1 und Lemma 5.1.5 sofort die Abschätzung

$$\begin{aligned}
|\varepsilon_i| &\leq c_1 \left(\frac{t_{i_0+1}}{t_{i+1}}\right)^{\eta} |l_{i_0}| + c_2 \sum_{l=i_0+1}^{i} \left(\frac{t_l}{t_{i+1}}\right)^{\eta} ht_l^{\gamma-1} |l_l| \\
&\leq \text{const.}(|l_{i_0}| + t_{i+1}^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l|) \\
&\leq \text{const.} \|l_h\|_h, \quad i = i_0, \ldots, N
\end{aligned}$$

(cf. dazu den Beweis von Lemma 5.1.6).

**Lemma 5.2.3** *Sei $\lambda \in \mathbb{C}$ mit $\Re(\lambda) < 0$ und $\gamma > 0$. Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \leq \text{const.}(|\tilde{l}_{i_0}| + t_{i+1}^{\gamma} \max_{i_0+1 \leq k \leq N} |\tilde{l}_k|), \quad i = i_0, \ldots, N, \tag{5.59}$$

*mit $\tilde{l}_h \in \mathbb{C}^{N-i_0+1}$. $\varepsilon_h$ sei die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{\lambda}{t_{j+1}} \varepsilon_{j+1} - \frac{1}{t_{j+1}} \delta_{j+1} - t_{j+1}^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0. \tag{5.60}$$

*Dann gilt*

$$\begin{aligned}
\varepsilon_i &= \prod_{l=i_0+1}^{i} \left(1 - \frac{h\lambda}{t_l}\right)^{-1} l_{i_0} + \sum_{l=i_0+1}^{i} \prod_{k=l}^{i} \left(1 - \frac{h\lambda}{t_k}\right)^{-1} ht_l^{\gamma-1} l_l^* \\
&=: \tilde{z}_{i_0+1,i+1}(-\lambda) l_{i_0} + \sum_{l=i_0+1}^{i} \tilde{z}_{l,i+1}(-\lambda) ht_l^{\gamma-1} l_l^*, \quad i = i_0 + 1, \ldots, N
\end{aligned} \tag{5.61}$$

*mit $l_i^* := t_i^{-\gamma} \delta_i + l_i$, $i = i_0 + 1, \ldots, N$. Weiters gelten die Abschätzungen*

$$\begin{aligned}
|\varepsilon_i| &\leq \text{const.}(\max\{|l_{i_0}|, |\tilde{l}_{i_0}|\} + t_{i+1}^{\gamma} \max_{i_0+1 \leq k \leq N} \max\{|l_k|, |\tilde{l}_k|\}) &(5.62)\\
&\leq \text{const.} \max\{\|l_h\|_h, \|\tilde{l}_h\|_h\}, \quad i = i_0, \ldots, N. &(5.63)
\end{aligned}$$

*Beweis:* Die Lösungsdarstellung (5.61) folgt direkt mittels Induktion genauso wie in Lemma 5.2.2. Die Abschätzungen folgen ebenfalls wie in 5.2.2, wenn man die Abschätzung von $|\delta_i|$ in gleicher Weise, wie das in Lemma 5.1.7 geschehen war, einsetzt.

**Lemma 5.2.4** *Sei $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - t_{j+1}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0 \tag{5.64}$$

*gilt*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0+1}^{i} h t_l^{\gamma-1} l_l.$$

*Weiters gilt die Abschätzung*

$$|\varepsilon_i| \leq \text{const.}(|l_{i_0}| + t_{i+1}^{\gamma} \max_{i_0+1 \leq k \leq N} |l_k|).$$

*Beweis:* Die Lösungsdarstellung ist unmittelbar ersichtlich und die Abschätzung folgt aus Lemma 5.1.5.

**Lemma 5.2.5** *Sei $\gamma > 0$. Betrachte die Gleichung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{t_{j+1}} \delta_{j+1} - t_{j+1}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0.$$

*Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \leq \sum_{l=0}^{k} b_l |\ln(h)|^l + c t_{i+1}^{\gamma}, \quad i = i_0, \ldots, N$$

*mit $c, b_l > 0,\ l = 0, \ldots, k$. Dann folgt die Lösungsdarstellung*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0+1}^{i} h t_l^{\gamma-1} l_l^*, \quad i = i_0, \ldots, N-1 \tag{5.65}$$

*mit $l_l^* := t_l^{-\gamma} \delta_l + l_l,\ l = i_0 + 1, \ldots, N$, sowie die Abschätzung*

$$\begin{aligned} |\varepsilon_i| \ &\leq \ |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_{i+1}^{\gamma} \max\{c, \max_{i_0+1 \leq m \leq N} |l_m|\}, \\ &\quad i = i_0, \ldots, N. \end{aligned} \tag{5.66}$$

*Beweis:* Die Lösungsdarstellung (5.65) ist wieder offensichtlich. Die Abschätzung (5.66) folgt aus

$$\begin{aligned}|\varepsilon_i| &\leq |l_{i_0}| + \sum_{l=i_0+1}^{i} h t_l^{\gamma-1} |l_l^*| \\ &\leq |l_{i_0}| + \sum_{l=i_0+1}^{i} h t_l^{-1} \sum_{m=0}^{k} b_m |\ln(h)|^m \\ &\quad + c_1 \sum_{l=i_0+1}^{i} h t_l^{\gamma-1} c + \sum_{l=i_0+1}^{i} h t_l^{\gamma-1} |l_l|\end{aligned}$$

unter Beachtung von $t_{l+1} \leq c_2 t_l, l \geq 1$, und weiter mit Lemma 5.1.5 analog wie in Lemma 5.1.6 und unter Verwendung der Abschätzung

$$\left|\ln\left(\frac{t_k}{t_j}\right)\right| \leq |\ln(t_j)| \leq |\ln(h) + \ln(i_0+1)| \leq 2|\ln(h)|, \quad 0 < t_{i_0+1} \leq t_j < t_k \leq 1.$$

Mit dem obigen Lemma folgt direkt die Abschätzung, die Lemma 5.1.13 entspricht, für die Gleichung

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{t_{j+1}} J_m(0)\varepsilon_{j+1} - t_{j+1}^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0. \tag{5.67}$$

**Lemma 5.2.6** *Sei $\varepsilon_h$ die Lösung von (5.67). $\varepsilon_{j;l}$ bezeichne die $l$-te Komponente von $\varepsilon_j$ für $l = 1, \ldots, m$ (und für die anderen Vektoren sinngemäß). Dann gilt*

$$\begin{aligned}|\varepsilon_{i;j}| &\leq \text{const.} \left( \sum_{l=j}^{m} \max_{l \leq k \leq m} |l_{i_0;k}| |\ln(h)|^{l-j} + t_{i+1}^{\gamma} \max_{j \leq k \leq m} \max_{i_0+1 \leq l \leq N} |l_{l;k}| \right) \\ &\leq \text{const.} \left( \sum_{l=j}^{m} |l_{i_0}| |\ln(h)|^{l-j} + t_{i+1}^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l| \right), \\ &\quad i = i_0, \ldots, N, \quad j = 1, \ldots, m.\end{aligned}$$

*Beweis:* Wird mittels vollständiger Induktion beginnend bei der $m$-ten Komponente geführt. Der Induktionsanfang kann Lemma 5.2.4 entnommen werden. Der Induktionsschritt von $j+1$ auf $j$ entspricht Lemma 5.2.5 für $k = m - j - 1$.

Abschließend wird noch die Konsistenz von $F_h$ aus (5.53) bewiesen. Es gilt

$$\begin{aligned}F_h(R_h(v)) &= \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - \frac{1}{t_{j+1}} Jv(t_{j+1}) - t_{j+1} g_{j+1}, \ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v_{h;i_0} \end{pmatrix} \\ &= \begin{pmatrix} \frac{v(t_{j+1})-v(t_j)}{h} - v'(t_{j+1}), \ j = i_0, \ldots, N-1 \\ v(t_{i_0}) - v_{h;i_0} \end{pmatrix} \\ &=: \begin{pmatrix} l_{j+1}, \ j = i_0, \ldots, N-1 \\ l_{i_0} \end{pmatrix}.\end{aligned} \tag{5.68}$$

**Lemma 5.2.7** *Sei $l_h$ definiert wie in (5.68) und $g(t,v)$ sei* LIPSCHITZ*-stetig bezüglich $v$ mit Konstante $L$. Dann gilt für $v \in C^1([0,1],\mathbb{C}^{2n})$*

$$\begin{aligned} \lim_{h\to 0} \|l_h\|_h &= 0, \\ |l_{i_0}| &= o(h). \end{aligned}$$

*Ist $v \in C^2([0,1],\mathbb{C}^{2n})$, dann gilt sogar*

$$\begin{aligned} \|l_h\|_h &= O(h), \\ |l_{i_0}| &= O(h^2). \end{aligned}$$

*Beweis:* Ist $v$ stetig differenzierbar auf $[0,1]$ dann gilt nach dem Satz von TAYLOR B.1.7 (Entwicklung von $v(t_j)$ um $t_{j+1}$)

$$\begin{aligned} \left|\frac{v(t_{j+1})-v(t_j)}{h} - v'(t_{j+1})\right| &= \left|\int_0^1 v'(t_{j+1}-\tau h)\,d\tau - v'(t_{j+1})\right| \\ &= \left|\int_0^1 (v'(t_{j+1}-\tau h) - v'(t_{j+1}))\,d\tau\right| \\ &\le \max_{\theta\in[t_j,t_{j+1}]} |v'(\theta)-v'(t_{j+1})|, \quad j = i_0,\ldots,N-1, \end{aligned}$$

und weiter wie in Lemma 5.1.8.

Sei jetzt $v \in C^2([0,1],\mathbb{C}^{2n})$. Dann ist nach dem Satz von TAYLOR B.1.7

$$\begin{aligned} \left|\frac{v(t_{j+1})-v(t_j)}{h} - v'(t_{j+1})\right| &= h\left|-\int_0^1 (1-\tau)v''(t_{j+1}-\tau h)\,d\tau\right| \\ &\le \frac{h}{2}\max_{\theta\in[t_j,t_{j+1}]} |v''(\theta)| \\ &= O(h), \quad j = i_0,\ldots,N-1. \end{aligned}$$

Der Rest ist wieder analog zu Lemma 5.1.8.

Für die Anfangswerte gilt unter Verwendung der LIPSCHITZ-Bedingung an $g$ und mit der Abkürzung $N := |(I-J)^{-1}|$[6]

$$\begin{aligned} |l_{i_0}| &= |v(t_{i_0}) - v_{i_0}| = |v(t_{i_0}) - (I-J)^{-1}(v_0 + t_{i_0}^2 g(t_{i_0}, v_{i_0}))| \\ &\le N|v_0 + t_{i_0}^2 g(t_{i_0}, v_{i_0}) - t_{i_0}^2 g(t_{i_0}, v(t_{i_0})) + t_{i_0}^2 g(t_{i_0}, v(t_{i_0})) - v(t_{i_0}) + Jv(t_{i_0})| \\ &\le N|v_0 + t_{i_0}v'(t_{i_0}) - Jv(t_{i_0}) - v(t_{i_0}) + Jv(t_{i_0})| + t_{i_0}^2 NL|v_{i_0} - v(t_{i_0})| \\ &\le Nt_{i_0}\int_0^1 |v'(t_{i_0}) - v'(\tau t_{i_0})|\,d\tau + t_{i_0}^2 NL|v(t_{i_0}) - v_{i_0}| \\ &= o(h) + t_{i_0}^2 NL|v(t_{i_0}) - v_{i_0}|, \end{aligned}$$

[6] $I-J$ ist invertierbar, da $J$ nur Eigenwerte mit nichtpositivem Realteil besitzt.

und damit für hinreichend kleines $h$

$$|l_{i_0}| \leq \frac{1}{1 - t_{i_0}^2 NL} o(h) = o(h).$$

Gilt $v \in C^2([0,1], \mathbb{C}^{2n})$ dann folgt ganz analog

$$|l_{i_0}| \;\leq\; N t_{i_0}^2 \left| \int_0^1 (1-\tau) v''((1-\tau) t_{i_0})\, d\tau \right| + t_{i_0}^2 NL |l_{i_0}| = O(h^2).$$

Die Konvergenzresultate und ihre Beweise sowohl für konstante Koeffizientenmatrix $M$ als auch für variable Koeffizienten und das nichtlineare Problem lassen sich jetzt (bei sinngemäßer Definition aller vorkommenden Operatoren) aufbauend auf den vorangegangenen Resultaten ganz genauso herleiten, wie das im Abschnitt 5.1 für das explizite EULER-Verfahren geschehen war.

**Bemerkungen:**

1. Genauso wie in der Anmerkung auf Seite 56 sieht man, dass man das implizite EULER-Verfahren auch direkt auf Gleichung (5.2) anwenden kann, ohne vorher auf JORDAN-Normalform zu transformieren, da die Matrix $I - hM$ aufgrund der Voraussetzungen an die Eigenwerte von $M$ stets regulär ist.

   *Beweis:* $E^{-1}(I - hM)E = I - hJ$ ist eine obere Dreiecksmatrix mit Diagonaleinträgen, deren Realteil $\geq 1$ ist.

2. Beim impliziten EULER-Verfahren spricht auch nichts dagegen, $i_0 = 0$ zu wählen. In den vorangegangenen Berechnungen wurde nur formal von einem Punkt $t_{i_0}$ ausgegangen, um größtmögliche Allgemeinheit der Aussagen zu sichern.

### 5.2.1 Das NEWTON-Verfahren

Im Allgemeinen ist es bei impliziten Verfahren nötig, zur Bestimmung der Näherungslösung $z_{j+1}$ aus $z_j$ und den Problemdaten ein *nichtlineares Gleichungssystem* zu lösen. Da diese Aufgabe üblicherweise nicht exakt erfüllt werden kann, ist man auf Näherungsverfahren angewiesen. Hierbei bietet sich das NEWTON-Verfahren an, das in der Folge mit dem Hinweis auf einige klassische Resultate beschrieben wird. Diese Darstellung ist [34] entnommen. Danach wird gezeigt, dass auch für singuläre Probleme wie sie in diesem Zusammenhang behandelt werden eine zufriedenstellende Konvergenzgeschwindigkeit erreicht werden kann. Dafür verwendet man Ideen aus [28] und [29].

**Definition 5.2.8 (NEWTON-Verfahren)** *Seien $(E_1, \| \ \|_1)$ und $(E_2, \| \ \|_2)$ normierte Räume, $U$ eine offene Teilmenge von $E_1$ und $F : U \to Y$ eine* FRÉCHET-*differenzierbare Funktion. Dann sei eine Folge $(x_k)_{k\in\mathbb{N}}$ definiert durch*

1. *$x_1 \in U$ beliebig,*
2. *$x_{k+1}$ sei eine Nullstelle der Funktion*[7]

$$F_k : U \to Y$$
$$F_k(x) := F(x_k) + DF(x_k)(x - x_k).$$

*Dieser Algorithmus wird* NEWTON-Verfahren *genannt, die Folge $(x_k)_{k\in\mathbb{N}}$ die zugehörige* Iterationsfolge.

Im Folgenden wird der Begriff der Konvergenzordnung für Folgen eingeführt, um sodann zu zeigen, von welcher Ordnung die zum soeben eigeführten NEWTON-Verfahren gehörige Iterationsfolge gegen eine Nullstelle der nichtlinearen Abbildung $F$ konvergiert.

**Definition 5.2.9 (Konvergenz einer Iterationsfolge)** *Sei $(E, \| \ \|)$ ein normierter Raum. Eine Folge $(x_k)_{k\in\mathbb{N}}$ in $E$ heißt* konvergent von (mindestens) $p$-ter Ordnung*, $p \geq 1$, falls gilt:*

1. *$(x_k)_{k\in\mathbb{N}}$ konvergiert gegen ein $x^\star \in E$.*
2. *$\exists C > 0, p \geq 1, i_0 \in \mathbb{N}$, sodass gilt*

$$\|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^p \quad \forall k \geq i_0.$$

*Für $p = 1$ spricht man von* linearer*, für $p = 2$ von* quadratischer *Konvergenz.*

*Gibt es ein Nullfolge $(C_k)_{k\in\mathbb{N}}$ nichtnegativer reeller Zahlen sodass*

$$\|x_{k+1} - x^\star\| \leq C_k\|x_k - x^\star\| \quad \forall i \in \mathbb{N}$$

*gilt, dann spricht man von* superlinearer *Konvergenz.*

**Definition 5.2.10** *Seien $(E_1, \| \ \|_1)$ und $(E_2, \| \ \|_2)$ normierte Räume und $A : E_1 \to E_2$ eine lineare Abbildung, dann heißt $A$* regulär*, wenn es ein $m > 0$ gibt, sodass gilt*

$$\|Ax\|_2 \geq m\|x\|_1 \quad \forall x \in E_1.$$

[7] $DF$ bezeichnet die FRÉCHET-Ableitung von $F$.

**Satz 5.2.11** *Sei $(E_1, \| \|_1)$ ein normierter Raum, $(E_2, \| \|_2)$ ein Banachraum, $U$ eine offene Teilmenge von $E_1$ und $F \in C^1(U, E_2)$. Weiters sei für alle $\overline{x} \in U$ die Gleichung*

$$F(\overline{x}) + DF(\overline{x})(x - \overline{x}) = 0 \tag{5.69}$$

*lösbar. Ist für ein $x^\star \in U$, $F(x^\star) = 0$ und ist $DF(x^\star)$ regulär, dann existiert eine Umgebung $U_0$ von $x^\star$, sodass das* NEWTON*-Verfahren für jeden Startwert aus $U_0$ durchführbar ist und die zugehörige Iterationsfolge bleibt in $U_0$ und konvergiert mindestens superlinear gegen $x^\star$. Ist darüberhinaus $DF$ in einer Umgebung von $x^\star$* LIPSCHITZ*-stetig, dann ist die Konvergenz sogar quadratisch.*

*Beweis:* Siehe [34, S. 65 sq.].

Diese klassische Konvergenzordnung für das NEWTON-Verfahren ist in dem Sinne optimal unter allen Näherungsverfahren zur Berechnung von Nullstellen, als jede superlinear (quadratisch) konvergente Iterationsfolge bis auf Terme der Ordnung $o(\|x_{k+1} - x_k\|_1)$ (bzw. $O(\|x_{k+1} - x_k\|_1^2)$) die „NEWTON-Gleichung“ (5.69) erfüllt. Für eine Präzisierung dieser Aussage siehe [34, S. 68 sqq.].

Dass unter entsprechenden Voraussetzungen diese Konvergenzordnung auch für die Anwendung auf die Bestimmung der Näherungslösungen der oben betrachteten singulären Anfangswertprobleme gilt, wird unter Verwendung der folgenden Sätze bewiesen.

**Satz 5.2.12** *Sei $(E, \| \|)$ ein Banachraum und $x_0 \in E$ und $r > 0$ so, dass gilt:*

1. *Die Abbildung $F : E \to E$ besitzt eine* FRÉCHET*-Ableitung $DF$ auf $U := K(x_0, r)$, die fast überall auf der Verbindungsstrecke zwischen zwei beliebigen Punkten von $U$ beschränkt und stetig ist.*

2. *$DF(x_0)$ ist invertierbar.*

3. *Es gilt*[8]

$$\|DF^{-1}(x_0)(DF(x) - DF(y))\|_D \leq \omega(\|x - y\|) \quad \forall x, y \in U,$$

*wobei $\omega(t)$ stetig und nichtfallend für $t > 0$ sei und $\omega(0) \geq 0$ gelte. Mit dieser Funktion $\omega$ gelte*

$$\omega(r) < \frac{1}{2},$$
$$|DF^{-1}(x_0)F(x_0)| \leq (1 - 2\omega(r))r.$$

[8] $\| \|_D$ bezeichne die von $\| \|$ induzierte Operatornorm.

*Dann hat die Gleichung*

$$F(x) = 0$$

*eine in $U$ eindeutige Lösung $x^\star$ und diese kann mit dem* NEWTON*-Verfahren ausgehend vom Startwert $x_0$ berechnet werden.*

*Gilt sogar*

$$\omega(t) = Kt^\alpha, \quad \alpha \in (0, 1],$$

*dann ist die Konvergenzordnung der Iterationsfolge gleich $1 + \alpha$.*

*Beweis:* Siehe [28].

**Bemerkung:** Ist $DF$ LIPSCHITZ-stetig, dann wird also die optimale Konvergenzordnung 2 erreicht.

Der nächste Satz wird mit Beweis angeführt, da ähnlich wie in Satz 5.1.19 eine Modifikation des Beweises notwendig ist um die gewünschten Resultate für den Fall des Auftretens einer JORDAN-Matrix $J_m(0)$ mit $m > 1$ zeigen zu können.

**Satz 5.2.13** *Sei $v$ Lösung der Gleichung $F(v) = 0$. Die Familie $F_h$ sei konsistent mit der Aufgabe $F(v) = 0$, also*

$$\|F_h(R_1^h(v))\|_2^h \leq M(h), \quad \lim_{h\to 0} M(h) = 0.$$

*Die Familie $F_h$ besitze* FRÉCHET*-Ableitungen $DF_h(x_h)$ in einer Kugel $\overline{K}(R_1^h(v), \rho_0) \subseteq (E_1^h, \| \|_1^h)$ und für ein $h_0 > 0$ gelte für alle $h \in (0, h_0]$:*

1. *$DF_h(R_1^h(v))$ haben (gleichmäßig) beschränkte Inverse in den Mittelpunkten der jeweiligen Kugeln, d. h. es gibt ein $K_0 > 0$ mit*

$$\|DF_h^{-1}(R_1^h(v))\|_{2,1}^h \leq K_0.$$

   *$\| \|_{1,2}^h$ bezeichnet dabei die von $\| \|_1^h$ und $\| \|_2^h$ induzierte Operatornorm.*

2. *$DF_h$ ist (gleichmäßig)* HÖLDER*-stetig auf $\overline{K}(R_1^h(v), \rho_0)$, d. h. es gilt*

$$\begin{aligned} \|DF_h(x_h) - DF_h(y_h)\|_{1,2}^h &\leq K_L(\|x_h - y_h\|_1^h)^\alpha, \\ &\quad 0 < \alpha \leq 1, \quad \forall x_h, y_h \in \overline{K}(R_1^h(v), \rho_0). \end{aligned}$$

*Dann gibt es $h_0, \rho_1, \rho_0 > 0, \rho_1 \leq \rho_0$, sodass für alle $h \in (0, h_0]$ und für $x_h^{(0)} \in \overline{K}(R_1^h(v), \rho_1)$ das* NEWTON*-Verfahren mit Startwert $x_h^{(0)}$ von Ordnung $1+\alpha$ gegen die eindeutige Lösung von $F_h(v_h) = 0$ in $\overline{K}(R_1^h(v), \rho_0)$ konvergiert.*

*Beweis:* Zuerst ist zu zeigen, dass $DF_h(x_h)$ invertierbar ist für alle $x_h \in \overline{K}(R_1^h(v), \rho_0)$, falls $\rho_0$ hinreichend klein ist. Man schreibt

$$DF_h(x_h) = DF_h(R_1^h(v))(I + DF_h^{-1}(R_1^h(v))(DF_h(x_h) - DF_h(R_1^h(v)))).$$

$DF_h(R_1^h(v))$ ist laut Voraussetzung invertierbar und die Inverse durch $K_0$ beschränkt. Um die Invertierbarkeit des zweiten Faktors zu zeigen, verwendet man das VON NEUMANN-Lemma, siehe Lemma B.1.13, unter Beachtung von Lemma B.1.12. Es folgt nämlich sofort

$$\|DF_h^{-1}(R_1^h(v))(DF_h(x_h) - DF_h(R_1^h(v)))\| \leq K_0 K_L \rho_0^\alpha < 1 \tag{5.70}$$

für hinreichend kleines $\rho_0$[9], und weiters

$$\|DF_h^{-1}(x_h)\| \leq \frac{K_0}{1 - K_0 K_L \rho_0^\alpha} \quad \forall x_h \in \overline{K}(R_1^h(v), \rho_0). \tag{5.71}$$

Beginnt man eine NEWTON-Iteration in einem Punkt $x_h^{(0)} \in \overline{K}(R_1^h(v), \rho_1)$, so gilt

$$\begin{aligned} x_h^{(1)} - x_h^{(0)} &= -DF_h^{-1}(x_h^{(0)})F_h(x_h^{(0)}) \\ &= -DF_h^{-1}(x_h^{(0)})F_h(R_1^h(v)) + DF_h^{-1}(x_h^{(0)})(F_h(R_1^h(v)) - F_h(x_h^{(0)})) \\ &= -DF_h^{-1}(x_h^{(0)})F_h(R_1^h(v)) \\ &\quad + DF_h^{-1}(x_h^{(0)})(\widehat{DF}_h(R_1^h(v), x_h^{(0)})(R_1^h(v) - x_h^{(0)})) \end{aligned}$$

mit

$$\widehat{DF}_h(x_h, y_h) := \int_0^1 DF_h(\tau x_h + (1-\tau) y_h)\, d\tau,$$

siehe Satz B.1.7.

Wegen

$$\begin{aligned} \|\widehat{DF}_h(R_1^h(v), x_h^{(0)}) - DF_h(x_h^{(0)})\|_{1,2}^h &\leq \int_0^1 \|DF_h(\tau R_1^h(v) + (1-\tau)x_h^{(0)}) \\ &\quad - DF_h(x_h^{(0)})\|_{1,2}^h \, d\tau \\ &\leq \int_0^1 K_L(\|\tau(R_1^h(v) - x_h^{(0)})\|_1^h)^\alpha \, d\tau \\ &\leq K_L \frac{\rho_1^\alpha}{1+\alpha} \end{aligned}$$

gilt

$$\begin{aligned} \|DF_h^{-1}(x_h^{(0)})\widehat{DF}_h(R_1^h(v), x_h^{(0)})\|_{1,1}^h &= \|I + DF_h^{-1}(x_h^{(0)})(\widehat{DF}_h(R_1^h(v), x_h^{(0)}) \\ &\quad - DF_h(x_h^{(0)}))\|_{1,1}^h \\ &\leq 1 + K_L \frac{\rho_1^\alpha}{1+\alpha} \frac{K_0}{1 - K_0 K_L \rho_0^\alpha} =: c, \end{aligned} \tag{5.72}$$

[9]Da dieses $\rho_0$ dasselbe ist, das sich in Satz 5.1.17 als Radius der Kugel ergeben hatte, in der die eindeutige Lösung der Gleichung liegt, ist die NEWTON-Iteration sinnvoll.

und damit

$$\|x_h^{(1)} - x_h^{(0)}\|_1^h \leq \frac{K_0}{1 - K_0 K_L \rho_0^\alpha} M(h) + c\rho_1. \tag{5.73}$$

Für hinreichend kleines $h_0, \rho_1$ wird diese Schranke beliebig klein. Damit sind die Voraussetzungen von Satz 5.2.12 erfüllt. Denn laut Voraussetzung ist $F$ FRÉCHET-differenzierbar auf $\overline{K}(R_1^h(v), \rho_0)$ mit einer beschränkten Inversen in $R_1^h(v)$. Aus der HÖLDER-Stetigkeit von $DF_h$ folgt, dass $\omega(t) = K_0 K_L t^\alpha$ ist. Man kann jetzt $h_0$ und $\rho_1$ so wählen, dass die Bedingungen an $\omega$ und $x_h^{(0)}$ erfüllt sind.

Mit diesem Satz folgt also sofort die Konvergenz des NEWTON-Verfahrens für die Bestimmung der mit dem impliziten EULER-Verfahren berechneten Näherungslösung $z_{j+1}$ aus $z_j$, wenn in der JORDAN-Normalform von $M$ keine JORDAN-Matrizen $J_m(0)$ mit $m > 1$ auftreten. Für den Beweis, dass die Voraussetzungen von Satz 5.2.13 erfüllt sind, siehe Seite 67 sqq.

Für eine JORDAN-Matrix $J_m(0), m > 1$ sind die Abbildungen $DF_h^{-1}(R_h(z))$ nicht gleichmäßig beschränkt. Deshalb muss der obige Beweis in gleicher Weise modifiziert werden wie in Satz 5.1.19, indem man beachtet, dass für die Vektoren $x_h$, für die $\|DF_h^{-1}(R_h(z))x_h\|_h$ abgeschätzt wird, jeweils die Anfangswertkomponente verschwindet. Damit kann man zunächst zwar (5.70) zeigen, nicht jedoch (5.71). Also muss man die Existenz der Inversen $DF_h^{-1}(x_h)$ auf $\overline{K}(R_h(v), \rho_0)$ voraussetzen. Das bedeutet aber keine Einschränkung, da dazu schon die Stetigkeit von $D_2\overset{\circ}{f}(t,x)$ auf $\overline{K}(R_h(z), \rho_0) \subseteq (C([0,1], \mathbb{R}^{2n}), \| \ \|_\infty)$ ausreicht, und diese ist ohnehin notwendig für die HÖLDER-Stetigkeit von $DF_h$. Die Abschätzung (5.72) lässt sich in gleicher Weise herleiten. Abschätzung (5.73) hat jetzt die Gestalt

$$\|x_h^{(1)} - x_h^{(0)}\|_1^h \leq \frac{K_0}{1 - K_0 K_L \rho_0^\alpha} |\ln(h)|^{m-1} O(h) + c\rho_1.$$

Damit lassen sich die weiteren Argumente analog zum Satz durchführen.

Insgesamt erhält man also, dass die Konvergenzgeschwindigkeit des NEWTON-Verfahrens gleich $1+\alpha$ für eine HÖLDER-stetige, für eine LIPSCHITZ-stetige Funktion $f$ sogar gleich 2 ist.

## 5.3 Die Trapezregel

In diesem Abschnitt wird die Trapezregel[10] auf Konvergenz untersucht.

Die Überlegungen in diesem Abschnitt verlaufen über weite Strecken gleich wie im Abschnitt 5.1, wenn man die untersuchten Operatoren sinngemäß definiert, deshalb ist die Darstellung nicht so ausführlich und es werden im wesentlichen

[10] siehe Definition 4.4.1

nur die Resultate, die Lemma 5.1.6, 5.1.7, 5.1.11, 5.1.12, sowie Lemma 5.1.8 entsprechen, sowie einige Lemmata, die von diesen nur in unwichtigen formalen Details abweichen, aber in der jeweiligen Form benötigt werden, bewiesen.

Sei also der Operator $F_h$ definiert als

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1}-x_j}{h} - \frac{1}{2}\left(\frac{1}{t_j}Jx_j + t_j g_j + \frac{1}{t_{j+1}}Jx_{j+1} + t_{j+1}g_{j+1}\right),\ j = i_0, \ldots, N-1 \\ x_{i_0} - v_{h;i_0} \end{pmatrix}.$$

Dann ist die Berechnung von Näherungslösungen von[11]

$$\begin{aligned} v'(t) &= \frac{1}{t}Jv(t) + tg(t), \quad t \in (0,1], \\ v(0) &= v_0 \end{aligned} \tag{5.74}$$

mittels der Trapezregel äquivalent zur Lösung von

$$F_h(v_h) = 0.$$

**Bemerkung:** Für die Wahl des Anfangswerts $v_{h;i_0}$ geht man folgendermaßen vor: Man wählt $i_0 := 1$ und wählt $v_{h;1}$ als Resultat eines Schritts der Trapezregel ausgehend vom exakten Startwert in 0. Dies ist nur dann möglich, wenn $v'(0)$ bekannt ist. Nach den Resultaten von Kapitel 3, sh. S. 28, ist dies der Fall. Der erste Schritt hat demgemäß die Gestalt

$$v_1 = v_0 + \frac{h}{2}\left(v'(0) + \frac{1}{h}Jv_1 + hg_1\right).$$

Dass diese Vorgangsweise zur „klassischen“ Konvergenzordnung 2 führt, wird später bewiesen, im Folgenden wird formal weiterhin von einem beliebigen (festen) $i_0$ ausgegangen.

Setzt man $F_h(R_h(v)) =: l_h$, dann erfüllt der globale Fehler $\varepsilon_h := R_h(v) - v_h$ die Gleichung

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\left(\frac{1}{t_j}J\varepsilon_j + \frac{1}{t_{j+1}}J\varepsilon_{j+1}\right) - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0.$$

Die Lösung $\varepsilon_h$ dieser Gleichung soll wieder in $l_h$ abgeschätzt werden. Dies geschieht wie schon in den vorangegangenen Abschnitten komponentenweise, wobei zwischen den vollständig entkoppelten und den nur teilweise entkoppelten Gleichungen unterschieden werden muss und außerdem Eigenwerte mit negativem

[11] Die Gleichung ist bereits teilweise entkoppelt, so wie in Abschnitt 5.1.

Realteil und der Eigenwert 0 getrennt behandelt werden. Die Formulierung der Resultate ist etwas allgemeiner als für die Abschätzung von $\varepsilon_h$ benötigt wird, die allgemeineren Aussagen finden jedoch in späteren Abschnitten Verwendung.

Sei also zunächst $\lambda \in \mathbb{C}$ mit $\Re(\lambda) < 0$.

**Lemma 5.3.1** *Sei $\gamma > 0$ und $\Re(\lambda) < 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\left(\frac{\lambda}{t_j}\varepsilon_j + \frac{\lambda}{t_{j+1}}\varepsilon_{j+1}\right) - t_j^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \qquad (5.75)$$

*gilt*

$$\begin{aligned} \varepsilon_j &= \prod_{l=i_0+1}^{j} \left(1 - \frac{h\lambda}{2t_l}\right)^{-1} \prod_{k=i_0}^{j-1} \left(1 + \frac{h\lambda}{2t_k}\right) l_{i_0} \\ &\quad + \sum_{m=i_0}^{j-1} \prod_{p=m}^{j-1} \left(1 - \frac{h\lambda}{2t_{p+1}}\right)^{-1} \prod_{q=m}^{j-2} \left(1 + \frac{h\lambda}{2t_{q+1}}\right) h t_m^{\gamma-1} l_{m+1} \\ &=: \tilde{z}_{i_0+1,j+1}\left(-\frac{\lambda}{2}\right) z_{i_0,j}\left(-\frac{\lambda}{2}\right) l_{i_0} + \sum_{m=i_0}^{j-1} \tilde{z}_{m+1,j+1}\left(-\frac{\lambda}{2}\right) z_{m+1,j}\left(-\frac{\lambda}{2}\right) h t_m^{\gamma-1} l_{m+1}, \\ &\quad j = i_0+1, \ldots, N. \end{aligned} \qquad (5.76)$$

*$z_{lk}$ und $\tilde{z}_{lk}$ sind definiert wie in Lemma 5.1.4 bzw. wie in Lemma 5.2.1. Produkte oder Summen, bei denen der obere Index kleiner ist als der untere sind als leer aufzufassen.*

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_j| &\le \text{const.}(|l_{i_0}| + t_j^{\gamma} \max_{i_0+1 \le k \le N} |l_k|) && (5.77) \\ &\le \text{const.}\|l_h\|_h, \quad j = i_0, \ldots, N. && (5.78) \end{aligned}$$

*Beweis:* Die Lösungsdarstellung folgt mit vollständiger Induktion aus der Darstellung

$$\varepsilon_{j+1} = \frac{1 + \frac{h\lambda}{2t_j}}{1 - \frac{h\lambda}{2t_{j+1}}} \varepsilon_j + \frac{h t_j^{\gamma-1}}{1 - \frac{h\lambda}{2t_{j+1}}} l_{j+1}.$$

Der Induktionsanfang lautet

$$\varepsilon_{i_0+1} = \frac{1 + \frac{h\lambda}{2t_{i_0}}}{1 - \frac{h\lambda}{2t_{i_0+1}}} l_{i_0} + \frac{h t_{i_0}^{\gamma-1}}{1 - \frac{h\lambda}{2t_{i_0+1}}} l_{i_0+1}.$$

Gelte die Darstellung (5.76) für $\varepsilon_j$, dann folgt

$$\begin{aligned}
\varepsilon_{j+1} &= \left(1-\frac{h\lambda}{2t_{j+1}}\right)^{-1}\left(1+\frac{h\lambda}{2t_j}\right)\prod_{l=i_0+1}^{j}\left(1-\frac{h\lambda}{2t_l}\right)^{-1}\prod_{k=i_0}^{j-1}\left(1+\frac{h\lambda}{2t_k}\right)l_{i_0}\\
&\quad+\left(1-\frac{h\lambda}{2t_{j+1}}\right)^{-1}\left(1+\frac{h\lambda}{2t_j}\right)\cdot\\
&\quad\cdot\sum_{m=i_0}^{j-1}\prod_{p=m}^{j-1}\left(1-\frac{h\lambda}{2t_{p+1}}\right)^{-1}\prod_{q=m}^{j-2}\left(1+\frac{h\lambda}{2t_{q+1}}\right)ht_m^{\gamma-1}l_{m+1}\\
&\quad+\left(1-\frac{h\lambda}{2t_{j+1}}\right)^{-1}ht_j^{\gamma-1}l_{j+1}\\
&= \prod_{l=i_0+1}^{j+1}\left(1-\frac{h\lambda}{2t_l}\right)^{-1}\prod_{k=i_0}^{j}\left(1+\frac{h\lambda}{2t_k}\right)l_{i_0}\\
&\quad+\sum_{m=i_0}^{j}\prod_{p=m}^{j}\left(1-\frac{h\lambda}{2t_{p+1}}\right)^{-1}\prod_{q=m}^{j-1}\left(1+\frac{h\lambda}{2t_{q+1}}\right)ht_m^{\gamma-1}l_{m+1}.
\end{aligned}$$

Die Abschätzung erfolgt ähnlich wie in 5.1.6 bzw. 5.2.2 unter Verwendung von 5.1.4, 5.2.1 und 5.1.5.

$$\begin{aligned}
|\varepsilon_j| &\le c_1\left(\frac{t_{i_0+1}}{t_{j+1}}\right)^{\eta}c_2\left(\frac{t_{i_0}}{t_j}\right)^{\eta}|l_{i_0}|\\
&\quad+\sum_{m=i_0}^{j-1}c_1\left(\frac{t_{m+1}}{t_{j+1}}\right)^{\eta}c_2\left(\frac{t_{m+1}}{t_j}\right)^{\eta}ht_m^{\gamma-1}|l_{m+1}|\\
&\le c_3|l_{i_0}|+c_4(t_{j+1}t_j)^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{\gamma-1}t_{m+1}^{2\eta}\max_{i_0+1\le k\le N}|l_k|\\
&\le \text{const.}(|l_{i_0}|+(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{\gamma-1+2\eta}\max_{i_0+1\le k\le N}|l_k|)\\
&\le \text{const.}(|l_{i_0}|+(t_jt_{j+1})^{-\eta}|t_j^{\gamma+2\eta}-t_{i_0}^{\gamma+2\eta}|\max_{i_0+1\le k\le N}|l_k|)\\
&\le \text{const.}(|l_{i_0}|+t_j^{\gamma}\max_{i_0+1\le k\le N}|l_k|)\\
&\le \text{const.}\|l_h\|_h,\quad j=i_0+1,\dots,N.
\end{aligned}$$

Das folgende Lemma unterscheidet sich nur durch eine Indexverschiebung vom vorangegangenen, das Resultat wird aber später in dieser Form gebraucht und soll deshalb angeführt werden.

**Lemma 5.3.2** *Sei $\gamma > 0$ und $\Re(\lambda) < 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\left(\frac{\lambda}{t_j}\varepsilon_j + \frac{\lambda}{t_{j+1}}\varepsilon_{j+1}\right) - t_{j+1}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0 \qquad (5.79)$$

*gilt*

$$\begin{aligned} \varepsilon_j &= \prod_{l=i_0+1}^{j} \left(1 - \frac{h\lambda}{2t_l}\right)^{-1} \prod_{k=i_0}^{j-1} \left(1 + \frac{h\lambda}{2t_k}\right) l_{i_0} \\ &\quad + \sum_{m=i_0}^{j-1} \prod_{p=m}^{j-1} \left(1 - \frac{h\lambda}{2t_{p+1}}\right)^{-1} \prod_{q=m}^{j-2} \left(1 + \frac{h\lambda}{2t_{q+1}}\right) h t_{m+1}^{\gamma-1} l_{m+1} \\ &=: \tilde{z}_{i_0+1,j+1}\left(-\frac{\lambda}{2}\right) z_{i_0,j}\left(-\frac{\lambda}{2}\right) l_{i_0} + \sum_{m=i_0}^{j-1} \tilde{z}_{m+1,j+1}\left(-\frac{\lambda}{2}\right) z_{m+1,j}\left(-\frac{\lambda}{2}\right) h t_{m+1}^{\gamma-1} l_{m+1}, \\ &\quad j = i_0+1, \ldots, N. \end{aligned} \qquad (5.80)$$

*$z_{lk}$ und $\tilde{z}_{lk}$ sind definiert wie in Lemma 5.1.4 bzw. wie in Lemma 5.2.1. Produkte oder Summen, bei denen der obere Index kleiner ist als der untere sind als leer aufzufassen.*

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_j| &\leq \text{const.}(|l_{i_0}| + t_j^{\gamma} \max_{i_0+1 \leq k \leq N} |l_k|) && (5.81) \\ &\leq \text{const.}\|l_h\|_h, \quad j = i_0, \ldots, N. && (5.82) \end{aligned}$$

*Beweis:* Genauso wie in 5.3.1 unter Beachtung von $t_{j+1}^{\gamma} \leq \text{const.} t_j^{\gamma}$.

**Lemma 5.3.3** *Sei $\gamma > 0$ und $\Re(\lambda) < 0$. Für $\delta_h$ gelte die Abschätzung*

$$|\delta_j| \leq \text{const.}(|\tilde{l}_{i_0}| + t_j^{\gamma} \max_{i_0+1 \leq k \leq N} |\tilde{l}_k|), \quad j = i_0, \ldots, N,$$

*wobei $\tilde{l}_h \in \mathbb{C}^{N-i_0+1}$ gilt. $\varepsilon_h$ sei die Lösung der linearen Differenzengleichung erster Ordnung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\left(\frac{\lambda}{t_j}\varepsilon_j + \frac{1}{t_j}\delta_j + \frac{\lambda}{t_{j+1}}\varepsilon_{j+1} + \frac{1}{t_{j+1}}\delta_{j+1}\right) - t_j^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0. \qquad (5.83)$$

*Dann gilt*

$$\varepsilon_j = \prod_{l=i_0+1}^{j} \left(1 - \frac{h\lambda}{2t_l}\right)^{-1} \prod_{k=i_0}^{j-1} \left(1 + \frac{h\lambda}{2t_k}\right) l_{i_0}$$

$$
\begin{aligned}
&+\sum_{m=i_0}^{j-1}\prod_{p=m}^{j-1}\left(1-\frac{h\lambda}{2t_{p+1}}\right)^{-1}\prod_{q=m}^{j-2}\left(1+\frac{h\lambda}{2t_{q+1}}\right)ht_m^{\gamma-1}l^*_{m+1}\\
=:\ &\tilde{z}_{i_0+1,j+1}\left(-\frac{\lambda}{2}\right)z_{i_0,j}\left(-\frac{\lambda}{2}\right)l_{i_0} \qquad (5.84)\\
&+\sum_{m=i_0}^{j-1}\tilde{z}_{m+1,j+1}\left(-\frac{\lambda}{2}\right)z_{m+1,j}\left(-\frac{\lambda}{2}\right)ht_m^{\gamma-1}l^*_{m+1},\\
&j=i_0+1,\dots,N, \qquad (5.85)
\end{aligned}
$$

*mit*

$$
l_j^* := \frac{1}{2}\left(t_{j-1}^{-\gamma}\delta_{j-1}+\frac{t_{j-1}^{1-\gamma}}{t_j}\delta_j\right)+l_j,\quad j=i_0+1,\dots,N.
$$

*$z_{lk}$ und $\tilde{z}_{lk}$ sind definiert wie in Lemma 5.1.4 bzw. wie in Lemma 5.2.1. Produkte oder Summen, bei denen der obere Index kleiner ist als der untere sind als leer aufzufassen.*

*Weiters gelten die Abschätzungen*

$$
\begin{aligned}
|\varepsilon_j| &\le \text{const.}(\max\{|l_{i_0}|,|\tilde{l}_{i_0}|\}+t_j^{\gamma}\max_{i_0+1\le k\le N}\max\{|l_k|,|\tilde{l}_k|\}) \qquad (5.86)\\
&\le \text{const.}\max\{\|l_h\|_h,\|\tilde{l}_h\|_h\},\quad j=i_0,\dots,N. \qquad (5.87)
\end{aligned}
$$

*Beweis:* Die Lösungsdarstellung folgt analog wie in Lemma 5.3.1. Die Abschätzung erhält man mittels

$$
\begin{aligned}
|\varepsilon_j| &\le \text{const.}\left(|l_{i_0}|+(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{\gamma-1}t_{m+1}^{2\eta}|l^*_{m+1}|\right)\\
&\le \text{const.}\left(|l_{i_0}|+(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta+\gamma-1}|l^*_{m+1}|\right)\\
&\le \text{const.}\Bigg(|l_{i_0}|+\frac{1}{2}(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta+\gamma-1}t_m^{-\gamma}|\delta_m|\\
&\quad+\frac{1}{2}(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta+\gamma-1}\frac{t_m^{1-\gamma}}{t_{m+1}}|\delta_{m+1}|\\
&\quad+(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta+\gamma-1}|l_{m+1}|\Bigg)\\
&\le \text{const.}\Bigg(|l_{i_0}|+\frac{1}{2}(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta-1}|\tilde{l}_{i_0}|
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{2}(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta+\gamma-1}\max_{i_0+1\le k\le N}|\tilde{l}_k|\\
&+\frac{1}{2}(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}h\frac{t_m^{2\eta}}{t_{m+1}}|\tilde{l}_{i_0}|+\frac{1}{2}(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}h\frac{t_m^{2\eta}}{t_{m+1}}t_{m+1}^{\gamma}\max_{i_0+1\le k\le N}|\tilde{l}_k|\\
&+(t_jt_{j+1})^{-\eta}\sum_{m=i_0}^{j-1}ht_m^{2\eta+\gamma-1}\max_{i_0+1\le k\le N}|l_k|\Bigg)\\
\le\ &\text{const.}\Bigg(|l_{i_0}|+\frac{1}{2}(t_jt_{j+1})^{-\eta}t_j^{2\eta}|\tilde{l}_{i_0}|+\frac{1}{2}(t_jt_{j+1})^{-\eta}t_j^{2\eta+\gamma}\max_{i_0+1\le k\le N}|\tilde{l}_k|\\
&+\frac{1}{2}(t_jt_{j+1})^{-\eta}t_{j+1}^{2\eta}|\tilde{l}_{i_0}|+\frac{1}{2}(t_jt_{j+1})^{-\eta}t_{j+1}^{2\eta+\gamma}\max_{i_0+1\le k\le N}|\tilde{l}_k|\\
&+(t_jt_{j+1})^{-\eta}t_j^{2\eta+\gamma}\max_{i_0+1\le k\le N}|l_k|\\
\le\ &\text{const.}\left(\max\{|l_{i_0}|,|\tilde{l}_{i_0}|\}+t_j^{\gamma}\max_{i_0+1\le k\le N}\max\{|l_k|,\tilde{l}_k|\}\right)\\
\le\ &\text{const.}\max\{\|l_h\|_h,\|\tilde{l}_h\|_h\}.
\end{aligned}
$$

**Lemma 5.3.4** *Sei $\gamma>0$ und $\Re(\lambda)<0$. Für $\delta_h$ gelte die Abschätzung*

$$|\delta_j|\le\text{const.}(|\tilde{l}_{i_0}|+t_j^{\gamma}\max_{i_0+1\le k\le N}|\tilde{l}_k|),\quad j=i_0,\dots,N,$$

*wobei $\tilde{l}_h\in\mathbb{C}^{N-i_0+1}$ gilt. $\varepsilon_h$ sei die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix}\frac{\varepsilon_{j+1}-\varepsilon_j}{h}-\frac{1}{2}\left(\frac{\lambda}{t_j}\varepsilon_j+\frac{1}{t_j}\delta_j+\frac{\lambda}{t_{j+1}}\varepsilon_{j+1}+\frac{1}{t_{j+1}}\delta_{j+1}\right)-t_{j+1}^{\gamma-1}l_{j+1},\ j=i_0,\dots,N-1\\ \varepsilon_{i_0}-l_{i_0}\end{pmatrix}=0. \tag{5.88}$$

*Dann gilt*

$$
\begin{aligned}
\varepsilon_j\ =\ &\prod_{l=i_0+1}^{j}\left(1-\frac{h\lambda}{2t_l}\right)^{-1}\prod_{k=i_0}^{j-1}\left(1+\frac{h\lambda}{2t_k}\right)l_{i_0}\\
&+\sum_{m=i_0}^{j-1}\prod_{p=m}^{j-1}\left(1-\frac{h\lambda}{2t_{p+1}}\right)^{-1}\prod_{q=m}^{j-2}\left(1+\frac{h\lambda}{2t_{q+1}}\right)ht_m^{\gamma-1}l_{m+1}^*\\
=:\ &\tilde{z}_{i_0+1,j+1}\left(-\frac{\lambda}{2}\right)z_{i_0,j}\left(-\frac{\lambda}{2}\right)l_{i_0}+\sum_{m=i_0}^{j-1}\tilde{z}_{m+1,j+1}\left(-\frac{\lambda}{2}\right)z_{m+1,j}\left(-\frac{\lambda}{2}\right)ht_m^{\gamma-1}l_{m+1}^*,\\
&j=i_0+1,\dots,N,
\end{aligned}
\tag{5.89}
$$

*mit*

$$l_j^*:=\frac{1}{2}\left(t_{j-1}^{-\gamma}\delta_{j-1}+\frac{t_{j-1}^{1-\gamma}}{t_j}\delta_j\right)+\left(\frac{t_j}{t_{j-1}}\right)^{\gamma-1}l_j,\quad j=i_0+1,\dots,N.$$

*$z_{lk}$ und $\tilde{z}_{lk}$ sind definiert wie in Lemma 5.1.4 bzw. wie in Lemma 5.2.1. Produkte oder Summen, bei denen der obere Index kleiner ist als der untere sind als leer aufzufassen.*

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_j| &\leq \text{const.}(\max\{|l_{i_0}|, |\tilde{l}_{i_0}|\} + t_j^\gamma \max_{i_0+1\leq k\leq N} \max\{|l_k|, |\tilde{l}_k|\}) && (5.90)\\ &\leq \text{const.}\max\{\|l_h\|_h, \|\tilde{l}_h\|_h\}, \quad j = i_0, \ldots, N. && (5.91) \end{aligned}$$

*Beweis:* Genauso wie Lemma 5.3.3. Bei der Abschätzung ist zu beachten, dass

$$\left(\frac{t_{j+1}}{t_j}\right)^{\gamma-1} \leq \text{const.} \quad \forall \gamma > 0$$

wegen

$$1 \leq \frac{t_{j+1}}{t_j} = \left(1 + \frac{1}{j}\right) \leq 2$$

gilt.

Die nächsten Lemmata behandeln den Fall, dass 0 als Eigenwert der Koeffizientenmatrix auftritt, beginnend mit einer entkoppelten Komponente. In diesem Fall treten genau die selben Gleichungen auf wie für das explizite bzw. das implizite EULER-Verfahren.

**Lemma 5.3.5** *Sei $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - t_j^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \qquad (5.92)$$

*gilt*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} l_{l+1}, \quad i = i_0, \ldots, N.$$

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_i| &\leq \text{const.}(|l_{i_0}| + t_i^\gamma \max_{i_0+1\leq k\leq N} |l_k|) \\ &\leq \text{const.}\|l_h\|_h, \quad i = i_0, \ldots, N. \end{aligned}$$

*Beweis:* Die Lösungsdarstellung ist unmittelbar ersichtlich und die Abschätzung folgt aus Lemma 5.1.5.

**Lemma 5.3.6** *Sei $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - t_{j+1}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0 \tag{5.93}$$

*gilt*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_{l+1}^{\gamma-1} l_{l+1}, \quad i = i_0, \ldots, N.$$

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_i| &\leq \text{const.}(|l_{i_0}| + t_{i+1}^{\gamma} \max_{i_0+1 \leq k \leq N} |l_k|) \\ &\leq \text{const.} \|l_h\|_h. \end{aligned}$$

*Beweis:* Genauso wie Lemma 5.3.5.

Die nächsten vier Lemmata ermöglichen es, analog wie in Lemma 5.1.13 induktiv die Auswirkungen des Auftretens logarithmischer Terme in den Abschätzungen herzuleiten.

**Lemma 5.3.7** *Sei $\gamma > 0$. Betrachte die Gleichung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\left(\frac{1}{t_j}\delta_j + \frac{1}{t_{j+1}}\delta_{j+1}\right) - t_j^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0.$$

*Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \leq \sum_{l=0}^{k} b_l |\ln(h)|^l + c t_i^{\gamma}, \quad i = i_0, \ldots, N$$

*mit $c, b_l > 0,\ l = 0, \ldots k$. Dann folgt die Lösungsdarstellung*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} l_{l+1}^*, \quad i = i_0, \ldots, N \tag{5.94}$$

*mit $l_l^* := \frac{1}{2} t_{l-1}^{-\gamma} \delta_{l-1} + \frac{1}{2} \frac{t_{l-1}^{1-\gamma}}{t_l} \delta_l + l_l,\ l = i_0 + 1, \ldots, N$, sowie die Abschätzung*

$$\begin{aligned} |\varepsilon_i| &\leq |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_i^{\gamma} \max\{c, \max_{i_0+1 \leq m \leq N} |l_m|\}, \\ & \quad i = i_0, \ldots, N. \end{aligned} \tag{5.95}$$

*Beweis:* Die Lösungsdarstellung (5.94) ist wieder offensichtlich. Die Abschätzung folgt aus

$$\begin{aligned}
|\varepsilon_i| &\leq |l_{i_0}| + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} |l_{l+1}^*| \\
&\leq |l_{i_0}| + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_j^{-1} \sum_{l=0}^{k} b_l |\ln(h)|^l \\
&\quad + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_j^{-1} c t_j^{\gamma} + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_{j+1}^{-1} \sum_{l=0}^{k} b_l |\ln(h)|^l \\
&\quad + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_{j+1}^{-1} c t_{j+1}^{\gamma} + \sum_{j=i_0}^{i-1} h t_j^{\gamma-1} |l_{j+1}| \\
&\leq |l_{i_0}| + \text{const.} \Bigg( \frac{1}{2} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \frac{c}{2} t_i^{\gamma} \\
&\quad + \frac{1}{2} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \frac{c}{2} t_{i+1}^{\gamma} + t_i^{\gamma} \max_{i_0+1 \leq j \leq N} |l_j| \Bigg) \\
&\leq |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_i^{\gamma} \max\{c, \max_{i_0+1 \leq m \leq N} |l_m|\}, \\
&\quad i = i_0 + 1, \ldots, N.
\end{aligned}$$

**Lemma 5.3.8** *Sei $J = J_m(0)$. $\varepsilon_{j;l}$ bezeichne die $l$-te Komponente von $\varepsilon_j$ für $l = 1, \ldots, m$ (und für die anderen Vektoren sinngemäß). Dann gilt für die Lösung $\varepsilon_h$ von*

$$\begin{pmatrix} \frac{\varepsilon_{j+1} - \varepsilon_j}{h} - \frac{1}{2} \frac{1}{t_j} J \varepsilon_j - \frac{1}{2} \frac{1}{t_{j+1}} J \varepsilon_{j+1} - t_j^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0$$

$$\begin{aligned}
|\varepsilon_{i;j}| &\leq \text{const.} \left( \sum_{l=j}^{m} \max_{l \leq k \leq m} |l_{i_0;k}| |\ln(h)|^{l-j} + t_i^{\gamma} \max_{j \leq k \leq m} \max_{i_0+1 \leq l \leq N} |l_{l;k}| \right) \\
&\leq \text{const.} \left( \sum_{l=j}^{m} |l_{i_0}| |\ln(h)|^{l-j} + t_i^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l| \right), \\
&\quad i = i_0, \ldots, N, \quad j = 1, \ldots, m.
\end{aligned}$$

*Beweis:* Wird mittels vollständiger Induktion beginnend bei der $m$-ten Komponente geführt. Der Induktionsanfang kann Lemma 5.3.5 entnommen werden. Der Induktionsschritt von $j+1$ auf $j$ entspricht Lemma 5.3.7 für $k = m - j - 1$.

**Lemma 5.3.9** *Sei $\gamma > 0$. Betrachte die Gleichung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_j}\delta_j - \frac{1}{2}\frac{1}{t_{j+1}}\delta_{j+1} - t_{j+1}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0.$$

*Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \leq \sum_{l=0}^{k} b_l |\ln(h)|^l + c t_i^{\gamma}, \quad i = i_0, \ldots, N$$

*mit $c, b_l > 0,\ l = 0, \ldots k$. Dann folgt die Lösungsdarstellung*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_l^{\gamma-1} l_{l+1}^*, \quad i = i_0, \ldots, N \tag{5.96}$$

*mit $l_l^* := \frac{1}{2} t_{l-1}^{-\gamma}\delta_{l-1} + \frac{1}{2}\frac{t_{l-1}^{1-\gamma}}{t_l}\delta_l + \left(\frac{t_l}{t_{l-1}}\right)^{\gamma-1} l_l,\ l = i_0+1, \ldots, N$, sowie die Abschätzung*

$$\begin{aligned} |\varepsilon_i| \ &\leq \ |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_i^{\gamma} \max\{c, \max_{i_0+1 \leq m \leq N} |l_m|\}, \\ &\quad i = i_0, \ldots, N. \end{aligned} \tag{5.97}$$

*Beweis:* Analog wie Lemma 5.3.7 mit den gleichen Modifikationen wie in Lemma 5.3.4.

**Lemma 5.3.10** *Sei $J = J_m(0)$. $\varepsilon_{j;l}$ bezeichne die $l$-te Komponente von $\varepsilon_j$ für $l = 1, \ldots, m$ (und für die anderen Vektoren sinngemäß). Dann gilt für die Lösung $\varepsilon_h$ von*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - J\varepsilon_j - t_{j+1}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0$$

$$\begin{aligned} |\varepsilon_{i;j}| \ &\leq \ \text{const.} \left( \sum_{l=j}^{m} \max_{l \leq k \leq m} |l_{i_0;k}| |\ln(h)|^{l-j} + t_{i+1}^{\gamma} \max_{j \leq k \leq m} \max_{i_0+1 \leq l \leq N} |l_{l;k}| \right) \\ &\leq \ \text{const.} \left( \sum_{l=j}^{m} |l_{i_0}| |\ln(h)|^{l-j} + t_{i+1}^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l| \right), \\ &\quad i = i_0, \ldots, N, \quad j = 1, \ldots, m. \end{aligned}$$

*Beweis:* Wird mittels vollständiger Induktion beginnend bei der $m$-ten Komponente geführt. Der Induktionsanfang kann Lemma 5.3.6 entnommen werden. Der Induktionsschritt von $j+1$ auf $j$ entspricht Lemma 5.3.9 für $k = m - j - 1$.

Zusammen mit den Konsistenzresultaten für den Operator $F_h$, die am Ende dieses Abschnitts bewiesen werden, folgt mit den vorangegangenen Lemmata die Konvergenzordnung zwei für die Trapezregel ganz analog wie die Konvergenzordnung eins für das explizite EULER-Verfahren, cf. Abschnitt 5.1.

Für den Fall mit variablen Koeffizientenmatrizen der gleichen Bauart wie in Abschnitt 5.1.2 ist es möglich aufbauend auf den vorangegangenen Resultaten ganz genauso vorzugehen wie in 5.1.2. Für die Fixpunktgleichung

$$\begin{aligned}\frac{x_{i+1}-x_i}{h} &= \frac{1}{2}\left(\frac{1}{t_i}Jx_i+\frac{1}{t_{i+1}}Jx_{i+1}+t_ig_i+t_{i+1}g_{i+1}\right)\\ &\quad+\frac{1}{2}t_i^{\gamma-1}C(t_i)y_i+\frac{1}{2}t_{i+1}^{\gamma-1}C(t_{i+1})y_{i+1},\quad i=i_0,\ldots,N-1,\end{aligned}$$

ist die Lösung nach dem Superpositionsprinzip in die Anteile zu den Inhomogenitäten $\frac{1}{2}t_i^{\gamma-1}C(t_i)y_i$ und $\frac{1}{2}t_{i+1}^{\gamma-1}C(t_{i+1})y_{i+1}$ aufzuspalten und diese beiden Anteile getrennt abzuschätzen[12].

Auch der nichtlineare Fall folgt analog wie in Abschnitt 5.1.3 aus Satz 5.1.17 und Satz 5.1.19.

Zum Abschluss dieses Abschnitts wird jetzt noch die Konsistenz des Operators $F_h$ bewiesen. Dazu wird die bereits auf Seite 83 angedeutete Vorgangsweise eingeschlagen. Man setzt $i_0 := 1$ und wählt als Startwert in $t_1$ den mittels der Trapezregel aus der Kenntnis von $z(0)$ und $z'(0) = 0$ berechneten Wert.

**Lemma 5.3.11** *Sei die Lösung $z$ von*

$$\begin{aligned}z'(t) &= \frac{1}{t}Mz(t)+\overset{\circ}{tf}(t,z(t)),\quad t\in(0,1],\\ z(0) &= z_0\end{aligned}\tag{5.98}$$

*dreimal stetig differenzierbar. Außerdem sei $\overset{\circ}{f}(t,z)$* LIPSCHITZ*-stetig bezüglich $z$ mit Konstante $L$. Der Operator $F_h$ sei folgendermaßen definiert*[13]*:*

$$F_h(x_h) := \begin{pmatrix}\frac{x_{j+1}-x_j}{h}-\frac{1}{2}\left(\frac{1}{t_j}Mx_j+\overset{\circ}{t_jf}_j+\frac{1}{t_{j+1}}Mx_{j+1}+\overset{\circ}{t_{j+1}f}_{j+1}\right),\ j=1,\ldots,N-1\\ x_1-v_{h;1}\end{pmatrix}.$$

*Für $l_h := F_h(R_h(z))$ gilt dann*

$$\begin{aligned}|l_1| &= O(h^3),\quad h\to 0,\\ |l_j| &= O(h^2),\quad h\to 0,\quad j=2,\ldots,N.\end{aligned}$$

[12]Dafür wurden auch die Lemmata mit den geänderten Indizes benötigt.

[13]Man setze wie gewohnt $\overset{\circ}{f}_j := \overset{\circ}{f}(t_j,x_j)$

*Beweis:* Mittels Taylorentwicklung folgt

$$\begin{aligned}
|l_i| &= \left|\frac{z(t_{i+1})-z(t_i)}{h} - \frac{1}{2}(z'(t_i)+z'(t_{i+1}))\right| \\
&= \frac{1}{2h}|z(t_{i+1}) - z(t_i) - hz'(t_i) + z(t_{i+1}) - z(t_i) - hz'(t_{i+1})| \\
&= \frac{1}{2h}\Bigg|z(t_i) + hz'(t_i) + \frac{h^2}{2}z''(t_i) + \frac{h^3}{2}\int_0^1 (1-\tau)^2 z^{(3)}(t_i+\tau h)\,d\tau \\
&\quad - z(t_i) - hz'(t_i) + z(t_{i+1}) - z(t_{i+1}) + hz'(t_{i+1}) - \frac{h^2}{2}z''(t_{i+1}) \\
&\quad + \frac{h^3}{2}\int_0^1 (1-\tau)^2 z^{(3)}(t_{i+1}-\tau h)\,d\tau - hz'(t_{i+1})\Bigg| \\
&= \frac{h}{4}|z''(t_i) - z''(t_{i+1})| + O(h^2) \\
&= \frac{h}{4}\left|z''(t_i) - z''(t_i) - h\int_0^1 z^{(3)}(t_i+\tau h)\,d\tau\right| + O(h^2) \\
&= O(h^2), \quad i = 2,\dots,N.
\end{aligned}$$

Im Folgenden wird $|l_1| = |z_1 - z(t_1)| = O(h^3)$ gezeigt. Um die Notation zu verkürzen, setze $L(z_1, z(t_1)) := \frac{h^2}{2}(\overset{\circ}{f}(t_1, z_1) - \overset{\circ}{f}(t_1, z(t_1)))$ und $N := \left|\left(I - \frac{1}{2}M\right)^{-1}\right|$. Es gilt[14]

$$z_1 = \left(I - \frac{1}{2}M\right)^{-1}\left(z_0 + \frac{h}{2}z'(0) + \frac{h^2}{2}\overset{\circ}{f}(t_1, z_1)\right).$$

Dann folgt unter Verwendung von $z_0 \in \ker(M)$

$$\begin{aligned}
|z_1 - z(t_1)| &\le N\left|z_0 + \frac{h}{2}z'(0) + \frac{h^2}{2}\overset{\circ}{f}(t_1, z_1) - \left(I - \frac{1}{2}M\right)z(t_1)\right| \\
&\le N\Bigg|z_0 + \frac{h}{2}z'(0) + L(z_1, z(t_1)) + \frac{h^2}{2}\overset{\circ}{f}(t_1, z(t_1)) - \left(I - \frac{1}{2}M\right)z_0 \\
&\quad - h\left(I - \frac{1}{2}M\right)z'(0) - \frac{h^2}{2}\left(I - \frac{1}{2}M\right)z''(0) \\
&\quad - \frac{h^3}{2}\left(I - \frac{1}{2}M\right)\int_0^1 (1-\tau)^2 z^{(3)}(\tau h)\,d\tau\Bigg| \\
&\le N\Bigg|\frac{h}{2}z'(0) + L(z_1, z(t_1)) + \frac{h}{2}\left(z'(t_1) - \frac{1}{h}Mz(t_1)\right) \\
&\quad - h\left(I - \frac{1}{2}M\right)z'(0) - \frac{h^2}{2}\left(I - \frac{1}{2}M\right)z''(0)\Bigg| + O(h^3)
\end{aligned}$$

[14]Die Matrix $\left(I - \frac{1}{2}M\right)$ ist invertierbar, da $M$ nur Eigenwerte mit nichtpositivem Realteil besitzt.

$$
\begin{aligned}
&\leq N\left|\frac{h}{2}z'(0) + L(z_1, z(t_1)) + \frac{h}{2}\left(z'(0) + hz''(0) - \frac{1}{h}Mz_0 - Mz'(0)\right.\right.\\
&\quad \left. - \frac{h}{2}Mz''(0)\right) - h\left(I - \frac{1}{2}M\right)z'(0)\\
&\quad \left. - \frac{h^2}{2}\left(I - \frac{1}{2}M\right)z''(0)\right| + O(h^3)\\
&= N|L(z_1, z(t_1))| + O(h^3)\\
&\leq NL\frac{h^2}{2}|z_1 - z(t_1)| + O(h^3).
\end{aligned}
$$

Für hinreichend kleines $h$ gilt also

$$|l_1| = O(h^3), \quad h \to 0.$$

**Bemerkung:** Der lineare Fall ist natürlich ein Spezialfall des obigen, da in diesem Fall $g(t, z)$ trivialerweise LIPSCHITZ-stetig ist.

Mit den oben bewiesenen Resultaten lässt sich nun ganz analog zu Abschnitt 5.1 die Konvergenz von zweiter Ordnung der Trapezregel zeigen. Auch die Berechnung der Näherungslösungen im nichtlinearen Fall mit dem NEWTON-Verfahren ist möglich und die Konvergenzgeschwindigkeit ist die gleiche wie die im Abschnitt 5.2.1 für das implizite EULER-Verfahren gezeigte.

## 5.4 Die Mittelpunktsregel

In diesem Abschnitt wird die Mittelpunktsregel gemäß Definition 4.5.1 auf Konvergenz untersucht.

In diesem Abschnitt wird die Notation $t_{j+1/2} := t_j + \frac{h}{2}$ verwendet, Ausdrücke wie $t_{j-1/2}, t_{j+3/2}$ o. ä. sind sinngemäß definiert.

Die Beweise sind praktisch identisch zu Abschnitt 5.3, wenn man anstelle der Lemmata 5.1.4, 5.1.5 und 5.2.1 die folgenden Resultate verwendet.

**Lemma 5.4.1** *Sei $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ mit $\sigma = \Re(\lambda) > 0$ fest gewählt. Definiere für $j \geq k \geq 0$*

$$z^*_{kj}(\lambda) := \begin{cases} 1, & k = j,\\ \prod_{l=k}^{j-1}\left(1 - \frac{\lambda}{l+\frac{1}{2}}\right), & 0 \leq k < j, \quad j = 1, 2, \ldots. \end{cases}$$

*Dann gibt es ein $\eta > 0$ und ein $C \geq 1$, sodass*

$$|z^*_{kj}(\lambda)| \leq C\left(\frac{k+\frac{1}{2}}{j+\frac{1}{2}}\right)^{\eta}, \quad 0 \leq k \leq j, \quad j = 0, 1, \ldots. \tag{5.99}$$

*Beweis:* Ähnlich wie im Beweis von Lemma A.2.1 zeigt man für $j > k$

$$\begin{aligned} |z^*_{kj}(\lambda)| &= \left| \prod_{l=k}^{j-1} \frac{l - \lambda + \frac{1}{2}}{l + \frac{1}{2}} \right| \\ &= \left| \frac{\Gamma(j + \frac{1}{2} - \lambda)}{\Gamma(j + \frac{1}{2})} \frac{\Gamma(k + \frac{1}{2})}{\Gamma(k + \frac{1}{2} - \lambda)} \right| \\ &\leq C \left( \frac{k + \frac{1}{2}}{j + \frac{1}{2}} \right)^{\eta}. \end{aligned}$$

**Lemma 5.4.2** *Sei* $h > 0, t_{j+1/2} := \left(j + \frac{1}{2}\right) h, k > j \geq i_0 \geq 0$ *und* $\gamma \in \mathbb{R}$*, dann gilt*

$$\sum_{l=j}^{k-1} h t_{l+1/2}^{\gamma-1} \leq \begin{cases} c_1 |t_{k+1/2}^{\gamma} - t_{j+1/2}^{\gamma}|, & \gamma \neq 0, \\ c_2 \ln \left( \frac{t_{k+1/2}}{t_{j+1/2}} \right), & \gamma = 0. \end{cases} \tag{5.100}$$

*Beweis:* Analog zum Beweis von Lemma A.2.2.

**Lemma 5.4.3** *Sei* $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ *mit* $\sigma = \Re(\lambda) > 0$ *fest gewählt. Definiere für* $j \geq k \geq 0$

$$z^{\dagger}_{kj}(\lambda) := \begin{cases} 1, & k = j, \\ \prod_{l=k}^{j-1} \left(1 + \frac{\lambda}{l + \frac{1}{2}}\right)^{-1}, & 0 \leq k < j, \quad j = 1, 2, \ldots. \end{cases}$$

*Dann gibt es ein* $\eta > 0$ *und ein* $C \geq 1$*, sodass*

$$|z^{\dagger}_{kj}(\lambda)| \leq C \left( \frac{k + \frac{1}{2}}{j + \frac{1}{2}} \right)^{\eta}, \quad 0 \leq k \leq j, \quad j = 0, 1, \ldots. \tag{5.101}$$

*Beweis:* Analog zu Lemma A.2.3.

Für den linearen Fall mit konstanter Koeffizientenmatrix wird also der Operator $F_h$ mit

$$F_h(x_h) := \begin{pmatrix} \frac{x_{j+1} - x_j}{h} - \frac{1}{2} \frac{1}{t_{j+1/2}} J (x_j + x_{j+1}) - t_{j+1/2} g_{j+1/2}, \ j = i_0, \ldots, N-1 \\ x_{i_0} - v_{h;i_0} \end{pmatrix}$$

betrachtet. Dann ist die Berechnung von Näherungslösungen von[15]

$$\begin{aligned} v'(t) &= \frac{1}{t} J v(t) + t g(t), \quad t \in (0, 1], \\ v(0) &= v_0 \end{aligned} \tag{5.102}$$

mittels der Mittelpunktsregel äquivalent zur Lösung von

$$F_h(v_h) = 0.$$

[15]Die Gleichung ist bereits teilweise entkoppelt, so wie in Abschnitt 5.1.

**Bemerkung:** Aus den nun folgenden Abschätzungen sieht man, dass es in diesem Fall auch möglich ist, $i_0 := 0$ zu wählen. Damit kann man den exakten Startwert $v_0 = v(0)$ verwenden.

Setzt man $F_h(R_h(v)) =: l_h$, dann erfüllt der globale Fehler $\varepsilon_h := R_h(v) - v_h$ die Gleichung

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}} J\left(\varepsilon_j + \varepsilon_{j+1}\right) - l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0.$$

Die Lösung $\varepsilon_h$ dieser Gleichung soll wieder in $l_h$ abgeschätzt werden, das geschieht fast identisch wie in Abschnitt 5.3 unter Verwendung der Lemmata 5.4.1, 5.4.2 und 5.4.3.

Sei also zunächst $\lambda \in \mathbb{C}$ mit $\Re(\lambda) < 0$.

**Lemma 5.4.4** *Sei $\gamma > 0$ und $\Re(\lambda) < 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{\lambda}{t_{j+1/2}}(\varepsilon_j + \varepsilon_{j+1}) - t_{j+1/2}^{\gamma-1} l_{j+1},\ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0 \qquad (5.103)$$

*gilt*

$$\begin{aligned} \varepsilon_j &= \prod_{l=i_0}^{j-1} \left(1 - \frac{h\lambda}{2t_{l+1/2}}\right)^{-1} \prod_{k=i_0}^{j-1} \left(1 + \frac{h\lambda}{2t_{k+1/2}}\right) l_{i_0} \\ &\quad + \sum_{m=i_0}^{j-1} \prod_{p=m}^{j-1} \left(1 - \frac{h\lambda}{2t_{p+1/2}}\right)^{-1} \prod_{q=m+1}^{j-1} \left(1 + \frac{h\lambda}{2t_{q+1/2}}\right) h t_{m+1/2}^{\gamma-1} l_{m+1} \\ &=: z_{i_0,j}^{\dagger}\left(-\frac{\lambda}{2}\right) z_{i_0,j}^{*}\left(-\frac{\lambda}{2}\right) l_{i_0} + \sum_{m=i_0}^{j-1} z_{m,j}^{\dagger}\left(-\frac{\lambda}{2}\right) z_{m+1,j}^{*}\left(-\frac{\lambda}{2}\right) h t_{m+1/2}^{\gamma-1} l_{m+1}, \\ &\quad j = i_0 + 1, \ldots, N. \end{aligned} \qquad (5.104)$$

*$z_{lk}^{*}$ und $z_{lk}^{\dagger}$ sind definiert wie in Lemma 5.4.1 bzw. wie in Lemma 5.4.3. Produkte oder Summen, bei denen der obere Index kleiner ist als der untere sind als leer aufzufassen.*

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_j| &\leq \text{const.}(|l_{i_0}| + t_j^{\gamma} \max_{i_0+1 \leq k \leq N} |l_k|) && (5.105) \\ &\leq \text{const.} \|l_h\|_h, \quad j = i_0, \ldots, N. && (5.106) \end{aligned}$$

*Beweis:* Die Lösungsdarstellung folgt mit vollständiger Induktion analog wie in Lemma 5.3.1 unter Verwendung der Beziehung

$$\varepsilon_{j+1} = \frac{1+\frac{h\lambda}{2t_{j+1/2}}}{1-\frac{h\lambda}{2t_{j+1/2}}}\varepsilon_j + \frac{ht_{j+1/2}^{\gamma-1}}{1-\frac{h\lambda}{2t_{j+1/2}}} l_{j+1}.$$

Die Abschätzung erfolgt ähnlich wie in 5.1.6 bzw. 5.2.2 unter Verwendung von 5.4.1, 5.4.3 und 5.4.2.

$$\begin{aligned}
|\varepsilon_j| &\leq c_1\left(\frac{t_{i_0+1/2}}{t_{j+1/2}}\right)^{\eta} c_2\left(\frac{t_{i_0+1/2}}{t_{j+1/2}}\right)^{\eta}|l_{i_0}| \\
&\quad + \sum_{m=i_0}^{j-1} c_1\left(\frac{t_{m+1/2}}{t_{j+1/2}}\right)^{\eta} c_2\left(\frac{t_{m+3/2}}{t_{j+1/2}}\right)^{\eta} ht_{m+1/2}^{\gamma-1}|l_{m+1}| \\
&\leq c_3|l_{i_0}| + c_4 t_{j+1/2}^{-2\eta}\sum_{m=i_0}^{j-1} ht_{m+1/2}^{\eta+\gamma-1}t_{m+3/2}^{\eta}\max_{i_0+1\leq k\leq N}|l_k| \\
&\leq \text{const.}(|l_{i_0}| + t_{j+1/2}^{-2\eta}\sum_{m=i_0}^{j-1} ht_{m+1/2}^{\gamma-1+2\eta}\max_{i_0+1\leq k\leq N}|l_k|) \\
&\leq \text{const.}(|l_{i_0}| + t_{j+1/2}^{-2\eta}|t_{j+1/2}^{\gamma+2\eta} - t_{i_0+1/2}^{\gamma+2\eta}|\max_{i_0+1\leq k\leq N}|l_k|) \\
&\leq \text{const.}(|l_{i_0}| + t_{j+1/2}^{\gamma}\max_{i_0+1\leq k\leq N}|l_k|) \\
&\leq \text{const.}\|l_h\|_h, \quad j = i_0+1,\ldots,N.
\end{aligned}$$

**Lemma 5.4.5** *Sei $\gamma > 0$ und $\Re(\lambda) < 0$. Für $\delta_h$ gelte die Abschätzung*

$$|\delta_j| \leq \text{const.}(|\tilde{l}_{i_0}| + t_{j+1/2}^{\gamma}\max_{i_0+1\leq k\leq N}|\tilde{l}_k|), \quad j = i_0,\ldots,N,$$

*wobei $\tilde{l}_h \in \mathbb{C}^{N-i_0+1}$ gilt. $\varepsilon_h$ sei die Lösung der linearen Differenzengleichung erster Ordnung*

$$\left(\begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}}(\lambda(\varepsilon_j+\varepsilon_{j+1}) + \delta_j + \delta_{j+1}) - t_{j+1/2}^{\gamma-1}l_{j+1},\ j = i_0,\ldots,N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array}\right) = 0. \tag{5.107}$$

*Dann gilt*

$$\begin{aligned}
\varepsilon_j &= \prod_{l=i_0}^{j-1}\left(1-\frac{h\lambda}{2t_{l+1/2}}\right)^{-1}\prod_{k=i_0}^{j-1}\left(1+\frac{h\lambda}{2t_{k+1/2}}\right)l_{i_0} \\
&\quad + \sum_{m=i_0}^{j-1}\prod_{p=m}^{j-1}\left(1-\frac{h\lambda}{2t_{p+1/2}}\right)^{-1}\prod_{q=m+1}^{j-1}\left(1+\frac{h\lambda}{2t_{q+1/2}}\right)ht_{m+1/2}^{\gamma-1}l_{m+1}^{*}
\end{aligned}$$

$$=: \; z^{\dagger}_{i_0,j}\left(-\frac{\lambda}{2}\right) z^{*}_{i_0,j}\left(-\frac{\lambda}{2}\right) l_{i_0} + \sum_{m=i_0}^{j-1} z^{\dagger}_{m,j}\left(-\frac{\lambda}{2}\right) z^{*}_{m+1,j}\left(-\frac{\lambda}{2}\right) h t^{\gamma-1}_{m+1/2} l^{*}_{m+1},$$
$$j = i_0+1,\ldots,N. \tag{5.108}$$

*mit*

$$l^{*}_{j} := \frac{1}{2} t^{-\gamma}_{j-1/2}(\delta_{j-1}+\delta_j) + l_j, \quad j = i_0+1,\ldots,N.$$

*$z^{*}_{lk}$ und $z^{\dagger}_{lk}$ sind definiert wie in Lemma 5.4.1 bzw. wie in Lemma 5.4.3. Produkte oder Summen, bei denen der obere Index kleiner ist als der untere sind als leer aufzufassen.*

*Weiters gelten die Abschätzungen*

$$\begin{aligned} |\varepsilon_j| &\le \text{const.}(\max\{|l_{i_0}|,|\tilde{l}_{i_0}|\} + t^{\gamma}_{j+1/2} \max_{i_0+1\le k\le N} \max\{|l_k|,|\tilde{l}_k|\}) && (5.109)\\ &\le \text{const.}\max\{\|l_h\|_h, \|\tilde{l}_h\|_h\}, \quad j = i_0,\ldots,N. && (5.110)\end{aligned}$$

*Beweis:* Die Lösungsdarstellung folgt analog wie in Lemma 5.4.4. Die Abschätzung erhält man mittels

$$\begin{aligned} |\varepsilon_j| &\le \text{const.}\left(|l_{i_0}| + t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{\eta+\gamma-1}_{m+1/2} t^{\eta}_{m+3/2}|l^{*}_{m+1}|\right)\\
&\le \text{const.}\left(|l_{i_0}| + t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta+\gamma-1}_{m+1/2}|l^{*}_{m+1}|\right)\\
&\le \text{const.}\Bigg(|l_{i_0}| + \frac{1}{2}t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta+\gamma-1}_{m+1/2} t^{-\gamma}_{m+1/2}|\delta_m|\\
&\quad + \frac{1}{2}t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta+\gamma-1}_{m+1/2} t^{-\gamma}_{m+1/2}|\delta_{m+1}|\\
&\quad + t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta+\gamma-1}_{m+1/2}|l_{m+1}|\Bigg)\\
&\le \text{const.}\Bigg(|l_{i_0}| + \frac{1}{2}t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta-1}_{m+1/2}|\tilde{l}_{i_0}|\\
&\quad + \frac{1}{2}t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta+\gamma-1}_{m+1/2}\max_{i_0+1\le k\le N}|\tilde{l}_k|\\
&\quad + \frac{1}{2}t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h\frac{t^{2\eta}_{m+1/2}}{t_{m+3/2}}|\tilde{l}_{i_0}| + \frac{1}{2}t^{-2\eta}_{j+1/2}\sum_{m=i_0}^{j-1} h t^{2\eta-1}_{m+1/2} t^{\gamma}_{m+3/2}\max_{i_0+1\le k\le N}|\tilde{l}_k|\end{aligned}$$

$$
\begin{aligned}
&+ t_{j+1/2}^{-2\eta} \sum_{m=i_0}^{j-1} h t_{m+1/2}^{2\eta+\gamma-1} \max_{i_0+1\le k\le N} |l_k| \Big) \\
\le\ & \text{const.} \Big( |l_{i_0}| + \frac{1}{2} t_{j+1/2}^{-2\eta} t_{j+1/2}^{2\eta} |\tilde{l}_{i_0}| + \frac{1}{2} t_{j+1/2}^{-2\eta} t_{j+1/2}^{2\eta+\gamma} \max_{i_0+1\le k\le N} |\tilde{l}_k| \\
&+ \frac{1}{2} t_{j+1/2}^{-2\eta} t_{j+1/2}^{2\eta} |\tilde{l}_{i_0}| + \frac{1}{2} t_{j+1/2}^{-2\eta} t_{j+1/2}^{2\eta+\gamma} \max_{i_0+1\le k\le N} |\tilde{l}_k| \\
&+ t_{j+1/2}^{-2\eta} t_{j+1/2}^{2\eta+\gamma} \max_{i_0+1\le k\le N} |l_k| \\
\le\ & \text{const.} \left( \max\{|l_{i_0}|, |\tilde{l}_{i_0}|\} + t_{j+1/2}^{\gamma} \max_{i_0+1\le k\le N} \max\{|l_k|, \tilde{l}_k|\} \right) \\
\le\ & \text{const.} \max\{\|l_h\|_h, \|\tilde{l}_h\|_h\}.
\end{aligned}
$$

Die nächsten Lemmata behandeln den Fall, dass 0 als Eigenwert der Koeffizientenmatrix auftritt, beginnend mit einer entkoppelten Komponente.

**Lemma 5.4.6** *Sei $\gamma > 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$
\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - t_{j+1/2}^{\gamma-1} l_{j+1},\ j = i_0, \dots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0 \tag{5.111}
$$

*gilt*

$$
\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_{l+1/2}^{\gamma-1} l_{l+1}, \quad i = i_0, \dots, N.
$$

*Weiters gelten die Abschätzungen*

$$
\begin{aligned}
|\varepsilon_i| &\le \text{const.}(|l_{i_0}| + t_{i+1/2}^{\gamma} \max_{i_0+1\le k\le N} |l_k|) \\
&\le \text{const.} \|l_h\|_h, \quad i = i_0, \dots, N.
\end{aligned}
$$

*Beweis:* Die Lösungsdarstellung ist unmittelbar ersichtlich und die Abschätzung folgt aus Lemma 5.4.2.

Die nächsten zwei Lemmata ermöglichen es, analog wie in Lemma 5.1.13 induktiv die Auswirkungen des Auftretens logarithmischer Terme in den Abschätzungen herzuleiten.

**Lemma 5.4.7** *Sei $\gamma > 0$. Betrachte die Gleichung*

$$
\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}}(\delta_j + \delta_{j+1}) - t_{j+1/2}^{\gamma-1} l_{j+1},\ j = i_0, \dots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{pmatrix} = 0.
$$

*Für $\delta_h$ gelte die Abschätzung*

$$|\delta_i| \le \sum_{l=0}^{k} b_l |\ln(h)|^l + c t_{i+1/2}^{\gamma}, \quad i = i_0, \ldots, N$$

*mit $c, b_l > 0,\ l = 0, \ldots k$. Dann folgt die Lösungsdarstellung*

$$\varepsilon_i = l_{i_0} + \sum_{l=i_0}^{i-1} h t_{l+1/2}^{\gamma-1} l_{l+1}^*, \quad i = i_0, \ldots, N \tag{5.112}$$

*mit $l_l^* := \frac{1}{2} t_{l-1/2}^{-\gamma}(\delta_{l-1} + \delta_l) + l_l,\ l = i_0 + 1, \ldots, N$, sowie die Abschätzung*

$$\begin{aligned} |\varepsilon_i| \ &\le\ |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_{i+1/2}^{\gamma} \max\{c, \max_{i_0+1 \le m \le N} |l_m|\}, \\ &\quad i = i_0, \ldots, N. \end{aligned} \tag{5.113}$$

*Beweis:* Die Lösungsdarstellung (5.112) ist wieder offensichtlich. Die Abschätzung folgt aus

$$\begin{aligned} |\varepsilon_i| \ &\le\ |l_{i_0}| + \sum_{l=i_0}^{i-1} h t_{l+1/2}^{\gamma-1} |l_{l+1}^*| \\ &\le\ |l_{i_0}| + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_{j+1/2}^{-1} \sum_{l=0}^{k} b_l |\ln(h)|^l \\ &\quad + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_{j+1/2}^{-1} c t_{j+1/2}^{\gamma} + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_{j+1/2}^{-1} \sum_{l=0}^{k} b_l |\ln(h)|^l \\ &\quad + \frac{1}{2} \sum_{j=i_0}^{i-1} h t_{j+1/2}^{-1} c t_{j+3/2}^{\gamma} + \sum_{j=i_0}^{i-1} h t_{j+1/2}^{\gamma-1} |l_{j+1}| \\ &\le\ |l_{i_0}| + \text{const.} \Bigg( \frac{1}{2} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \frac{c}{2} t_{i+1/2}^{\gamma} \\ &\quad + \frac{1}{2} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \frac{c}{2} t_{i+1/2}^{\gamma} + t_{i+1/2}^{\gamma} \max_{i_0+1 \le j \le N} |l_j| \Bigg) \\ &\le\ |l_{i_0}| + \text{const.} \sum_{l=0}^{k} b_l |\ln(h)|^{l+1} + \text{const.} t_{i+1/2}^{\gamma} \max\{c, \max_{i_0+1 \le m \le N} |l_m|\}, \\ &\quad i = i_0 + 1, \ldots, N. \end{aligned}$$

**Lemma 5.4.8** *Sei $J = J_m(0)$. $\varepsilon_{j;l}$ bezeichne die $l$-te Komponente von $\varepsilon_j$ für $l = 1, \ldots, m$ (und für die anderen Vektoren sinngemäß). Dann gilt für die Lösung $\varepsilon_h$ von*

$$\left( \begin{array}{l} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}} J(\varepsilon_j + \varepsilon_{j+1}) - t_{j+1/2}^{\gamma-1} l_{j+1}, \ j = i_0, \ldots, N-1 \\ \varepsilon_{i_0} - l_{i_0} \end{array} \right) = 0$$

$$\begin{aligned} |\varepsilon_{i;j}| &\leq \text{const.} \left( \sum_{l=j}^{m} \max_{l \leq k \leq m} |l_{i_0;k}| |\ln(h)|^{l-j} + t_{i+1/2}^{\gamma} \max_{j \leq k \leq m} \max_{i_0+1 \leq l \leq N} |l_{l;k}| \right) \\ &\leq \text{const.} \left( \sum_{l=j}^{m} |l_{i_0}| |\ln(h)|^{l-j} + t_{i+1/2}^{\gamma} \max_{i_0+1 \leq l \leq N} |l_l| \right), \\ &\quad i = i_0, \ldots, N, \quad j = 1, \ldots, m. \end{aligned}$$

*Beweis:* Wird mittels vollständiger Induktion beginnend bei der $m$-ten Komponente geführt. Der Induktionsanfang kann Lemma 5.4.6 entnommen werden. Der Induktionsschritt von $j+1$ auf $j$ entspricht Lemma 5.4.7 für $k = m - j - 1$.

Bevor der allgemeine lineare und der nichtlineare Fall behandelt werden, wird die Konsistenz des Operators $F_h$ untersucht. Dabei stellt sich heraus, dass die Konsistenzordnung gemäß Definition 5.1.1 nur 1 beträgt. Eine einfache Modifikation der bisherigen Vorgangsweise führt jedoch in den meisten Fällen trotzdem zur Konvergenzordnung 2. Nur im Fall, dass in der JORDAN-Normalform von $M$ eine JORDAN-Matrix $J_m(0)$ mit $m > 1$ auftritt, tritt die Ordnungsreduktion tatsächlich ein. Außerdem setzt man $i_0 := 0$[16] und untersucht nur $|l_j|, \ j = 1, \ldots, N$, da der exakte Startwert bekannt ist und damit $l_0 = 0$ gilt.

**Lemma 5.4.9** *Angenommen, $M$ hat nur Eigenwerte mit negativem Realteil oder den Eigenwert 0. Sei die Lösung $z$ von*

$$\begin{aligned} z'(t) &= \frac{1}{t} M z(t) + t \overset{\circ}{f}(t, z(t)), \quad t \in (0, 1], \\ z(0) &= z_0 \end{aligned} \tag{5.114}$$

*dreimal stetig differenzierbar. Außerdem sei $\overset{\circ}{f}(t, x)$* FRÉCHET*-differenzierbar. Der Operator $F_h$ sei folgendermaßen definiert:*

$$F_h(x_h) := \left( \begin{array}{l} \frac{x_{j+1}-x_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}} M(x_j + x_{j+1}) - t_{j+1/2} \overset{\circ}{f}_{j+1/2}, \ j = 0, \ldots, N-1 \\ x_0 - z_0 \end{array} \right).$$

[16] Die vorangegangenen Überlegungen zeigen, dass dies in diesem Fall möglich ist.

*Für $l_h := F_h(R_h(z))$ gilt dann*

$$|l_{j+1}| \;=\; \frac{1}{t_{j+1/2}} O(h^2) = O(h), \quad j = 0, \ldots, N-1.$$

*Beweis:* Mittels Taylorentwicklung[17] erhält man

$$\begin{aligned}
|l_{i+1}| &= \left| \frac{z(t_{i+1}) - z(t_i)}{h} - \frac{1}{t_{i+1/2}} M \frac{z(t_{i+1}) + z(t_i)}{2} \right. \\
&\quad \left. - t_{i+1/2} \mathring{f}\left( t_{i+1/2}, \frac{z(t_{i+1}) + z(t_i)}{2} \right) \right| \\
&= \left| \frac{1}{h} \left( z(t_{i+1/2}) + \frac{h}{2} z'(t_{i+1/2}) + \frac{h^2}{8} z''(t_{i+1/2}) \right. \right. \\
&\quad + \frac{h^3}{16} \int_0^1 (1-\tau)^2 z^{(3)} \left( t_{i+1/2} + \tau \frac{h}{2} \right) d\tau \\
&\quad - z(t_{i+1/2}) + \frac{h}{2} z'(t_{i+1/2}) \\
&\quad \left. - \frac{h^2}{8} z''(t_{i+1/2}) + \frac{h^3}{16} \int_0^1 (1-\tau)^2 z^{(3)} \left( t_{i+1/2} - \tau \frac{h}{2} \right) d\tau \right) \\
&\quad - \frac{1}{2} \frac{1}{t_{i+1/2}} M \left( z(t_{i+1/2}) + \frac{h}{2} z'(t_{i+1/2}) + \frac{h^2}{4} \int_0^1 (1-\tau) z'' \left( t_{i+1/2} + \tau \frac{h}{2} \right) d\tau \right. \\
&\quad \left. + z(t_{i+1/2}) - \frac{h}{2} z'(t_{i+1/2}) + \frac{h^2}{4} \int_0^1 (1-\tau) z'' \left( t_{i+1/2} - \tau \frac{h}{2} \right) d\tau \right) \\
&\quad \left. - t_{i+1/2} \mathring{f}(t_{i+1/2}, z(t_{i+1/2})) - t_{i+1/2} \int_0^1 D\mathring{f}(t_{i+1/2}, z(t_{i+1/2}) + \tau \tilde{h})\, d\tau (0, \tilde{h}) \right| \\
&\leq \left| z'(t_{i+1/2}) - \frac{1}{t_{i+1/2}} M z(t_{i+1/2}) - t_{i+1/2} \mathring{f}(t_{i+1/2}, z(t_{i+1/2})) \right| \\
&\quad + \frac{h^2}{24} M_3 + \frac{1}{t_{i+1/2}} |M| \frac{h^2}{4} M_2 + O(h^2) \\
&= \frac{1}{t_{i+1/2}} O(h^2) = O(h), \quad i = 1, \ldots, N
\end{aligned}$$

mit

$$\tilde{h} := \frac{h^2}{4} \int_0^1 (1-\tau) \left( z'' \left( t_{i+1/2} + \tau \frac{h}{2} \right) + z'' \left( t_{i+1/2} - \tau \frac{h}{2} \right) \right) d\tau = O(h^2).$$

[17] Siehe Satz B.1.7. Im Weiteren bezeichnet $Dg(t,x)$ die FRÉCHET-Ableitung von $g$ im Punkt $(t,x)$ und $M_k$ eine Schranke für die $k$-te Ableitung von $z$ auf $[0,1]$. Die Abschätzung der Integrale erfolgt mit Hilfe von Satz B.1.10.

Da also die Konsistenzordnung laut Definition 5.1.1 nur 1 beträgt, muss von der Vorgangsweise der letzten Abschnitte abgegangen werden. Es zeigt sich, dass man mit geringfügig modifizierten Abschätzungen zumindest im Fall, dass alle Eigenwerte der Koeffizientenmatrix $M$ negativen Realteil haben, trotzdem die Konvergenzordnung 2 erhält. Das ist Gegenstand der folgenden zwei Lemmata. Im Fall, dass der Eigenwert 0 auftritt, sind weitere Modifikationen nötig.

**Lemma 5.4.10** *Sei $\Re(\lambda) < 0$. Für die Lösung der linearen Differenzengleichung erster Ordnung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{\lambda}{t_{j+1/2}}(\varepsilon_j + \varepsilon_{j+1}) - l_{j+1},\ j = 0, \ldots, N-1 \\ \varepsilon_0 \end{pmatrix} = 0$$

*mit $l_h$ aus Lemma 5.4.9 gilt*

$$\|\varepsilon_h\|_h = O(h^2), \quad h \to 0.$$

*Beweis:* Aus der Lösungsdarstellung aus Lemma 5.4.4 erhält man ähnlich wie dort

$$\begin{aligned} |\varepsilon_j| &\leq \sum_{m=0}^{j-1} c_1 \left(\frac{t_{m+1/2}}{t_{j+1/2}}\right)^{\eta} c_2 \left(\frac{t_{m+3/2}}{t_{j+1/2}}\right)^{\eta} h|l_{m+1}| \\ &\leq c_3 t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{\eta} t_{m+3/2}^{\eta} t_{m+1/2}^{-1} O(h^2) \\ &\leq \text{const.} \left( t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{2\eta-1} O(h^2) \right) \\ &\leq \text{const.} t_{j+1/2}^{-2\eta} |t_{j+1/2}^{2\eta} - t_{1/2}^{2\eta}| O(h^2) \\ &= O(h^2), \quad j = 1, \ldots, N. \end{aligned}$$

**Lemma 5.4.11** *Sei $\Re(\lambda) < 0$ und gelte $\|\delta_h\|_h = O(h^2)$. $\varepsilon_h$ sei die Lösung der linearen Differenzengleichung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}}(\lambda(\varepsilon_j + \varepsilon_{j+1}) + \delta_j + \delta_{j+1}) - l_{j+1},\ j = 0, \ldots, N-1 \\ \varepsilon_0 \end{pmatrix} = 0$$

*mit $l_h$ aus Lemma 5.4.9. Dann gilt*

$$\|\varepsilon_h\|_h = O(h^2), \quad h \to 0.$$

*Beweis:* Sei

$$l_j^* := \frac{1}{2} t_{j-1/2}^{-1} (\delta_{j-1} + \delta_j) + l_j, \quad j = 1, \ldots, N.$$

Unter Verwendung der Lösungsdarstellung aus Lemma 5.4.5 erhält man ähnlich wie dort

$$\begin{aligned}
|\varepsilon_j| &\leq \text{const.} \left( t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{\eta} t_{m+3/2}^{\eta} |l_{m+1}^*| \right) \\
&\leq \text{const.} \Bigg( \frac{1}{2} t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{2\eta} t_{m+1/2}^{-1} |\delta_m| + \frac{1}{2} t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{2\eta} t_{m+1/2}^{-1} |\delta_{m+1}| \\
&\quad + t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{2\eta} |l_{m+1}| \\
&\leq O(h^2) + t_{j+1/2}^{-2\eta} \sum_{m=0}^{j-1} h t_{m+1/2}^{2\eta-1} O(h^2) \\
&= O(h^2), \quad j = 1, \ldots, N.
\end{aligned}$$

Sei jetzt 0 ein Eigenwert von $M$. Die Modifikation, die in den Lemmata 5.4.10 und 5.4.11 zum Beweis der Konvergenzordnung 2 führte, führt in diesem Fall nicht zum Ziel. Für Komponenten, die vollständig entkoppelt sind, ist jedoch die Konsistenzordnung im Gegensatz zum allgemeinen Fall gleich 2, wie in Lemma 5.4.12 gezeigt wird, und deshalb kann einfach Lemma 5.4.6 zum Beweis der Konvergenz benützt werden.

**Lemma 5.4.12** *Sei $v \in C^3([0,1], \mathbb{R})$ die Lösung von*

$$\begin{aligned}
v'(t) &= t g(t), \quad t \in (0,1], \\
v(0) &= v_0.
\end{aligned}$$

*Dann gilt*

$$\left| \frac{v(t_{j+1}) - v(t_j)}{h} - t_{j+1/2} g_{j+1/2} \right| = O(h^2), \quad j = 0, \ldots, N-1.$$

*Beweis:* Folgt mittels Taylorentwicklung um $t_{j+1/2}$ genauso wie in Lemma 5.4.9, wenn man $t_{j+1/2} g_{j+1/2} = v'(t_{j+1/2})$ beachtet.

Mit dem bisher bewiesenen kann man also ganz analog zu Abschnitt 5.3 zeigen, dass die Mittelpunktsregel für lineare Probleme mit konstanter Koeffizientenmatrix $M$, in deren JORDAN-Normalform kein JORDAN-Block $J_m(0)$ mit $m > 1$ auftritt, von zweiter Ordnung konvergiert. Ganz analog zu Abschnitt 5.1.2 folgt

diese Aussage auch für lineare Probleme mit variabler Koeffizientenmatrix, falls der konstante Anteil von $M(t)$ die entsprechenden Bedingungen erfüllt. Für das nichtlineare Problem lässt sich Satz 5.1.17 anwenden, wenn man mit Lemma 5.4.10 und Lemma 5.4.11 zeigt, dass

$$\|DF_h^{-1}(R_1^h(v))F_h(R_1^h(v))\|_h = O(h^2)$$

und diese Tatsache in der letzten Abschätzung verwendet.

Tritt in der JORDAN-Normalform ein JORDAN-Block $J_m(0)$ mit $m > 1$ auf, so versagen die obigen Modifikationen, und man kann für solche Probleme mit den bisher verwendeten Methoden nur die Konvergenzordnung 1 zeigen.

Genauer gesagt liefert dieselbe Modifikation wie Lemma 5.4.10 und 5.4.11 in Lemma 5.4.7 nur die Konvergenzordnung $O(|\ln(h)|^{m-1}h^2)$, wie das folgende Lemma zeigt:

**Lemma 5.4.13** *Sei $\varepsilon_h$ die Lösung der Differenzengleichung*

$$\begin{pmatrix} \frac{\varepsilon_{j+1}-\varepsilon_j}{h} - \frac{1}{2}\frac{1}{t_{j+1/2}}(\delta_j + \delta_{j+1}) - l_{j+1},\ j = 0, \ldots, N-1 \\ \varepsilon_0 \end{pmatrix} = 0,$$

*wobei*

$$\|\delta_h\|_h = O(|\ln(h)|^k h^2), \quad k \geq 0$$

*und*

$$l_j = \frac{1}{t_{j-1/2}} O(h^2), \quad j = 1, \ldots, N$$

*gelte. Dann folgt*

$$\varepsilon_j = \sum_{l=0}^{j-1} h l_{l+1}^*, \quad j = 1, \ldots, N$$

*mit $l_l^* = \frac{1}{2}\frac{1}{t_{l-1/2}}(\delta_{l-1} + \delta_l) + l_l,\ l = 1, \ldots, N$. Es gilt die Abschätzung*

$$\|\varepsilon_h\|_h = O(|\ln(h)|^{k+1}h^2), \quad h \to 0.$$

*Beweis:* Folgt wegen

$$|\varepsilon_j| \leq \sum_{l=0}^{j-1} h t_{l+1/2}^{-1} O(|\ln(h)|^k h^2) + \sum_{l=0}^{j-1} h t_{l+1/2}^{-1} O(h^2)$$

direkt aus Lemma 5.4.2.

Die soeben geschilderte Ordnungsreduktion kann auch experimentell beobachtet werden, wie das folgende nichtlineare Beispiel, das aus [49, p. 32] stammt, zeigt.

$$y''(t) = -\frac{1}{t}y'(t) - \nu\Big(\mathrm{e}^{y(t)} - \frac{B^2 - \nu(8-2\nu)B + 1}{Bt^2+1}\Big),\ \nu = \pm 1, \quad t \in (0,1],$$
$$y(0) = 2\ln(B+1), \quad y'(0) = 0.$$

Die übliche Transformation $z(t) = (z_1(t), z_2(t)) = (y(t), ty'(t))$ führt auf die äquivalente Gleichung erster Ordnung

$$z'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot z(t) - \nu t \begin{pmatrix} 0 \\ \mathrm{e}^{z_1(t)} - \frac{B^2-\nu(8-2\nu)B+1}{Bt^2+1} \end{pmatrix}, \quad t \in (0,1],$$
$$z(0) = \begin{pmatrix} 2\ln(B+1) \\ 0 \end{pmatrix}, \quad z'(0) = 0.$$

Die Gleichung zweiter Ordnung hat die analytische Lösung

$$y(t) = 2\ln\left(\frac{B+1}{Bt^2+1}\right),$$

wobei die Beziehung

$$B = \frac{(8-2\nu) \pm \sqrt{(8-2\nu)^2 - 4\nu^2}}{2\nu}$$

gelten muss.

Die experimentelle Bestimmung der Konvergenzordnung soll anhand dieses Beispiels mit den Paramaterwerten $\nu = 1, B = 3 + \sqrt{8}$ erfolgen. Diese Parameterwerte wurden als Beispiel 1 in [21, p. 215] verwendet.

Um zu überprüfen, ob tatsächlich die Konvergenzordnung $|\ln(h)|h^2$ vorliegt geht man folgendermaßen vor: Man löst das Problem numerisch (hier mittels der Mittelpunktsregel) mit zwei verschiedenen Schrittweiten $h_1$ und $h_2$. Da die exakte Lösung bekannt ist, lässt sich der globale Fehler $\|\varepsilon_{h_1}\|_{h_1}$ bzw. $\|\varepsilon_{h_2}\|_{h_2}$ berechnen. In diesem Fall wurde der Fehler für den ganzen Vektor $z(t)$ (und nicht etwa nur für $y(t)$) bestimmt. Es gilt, die Annahme zu überprüfen, dass mit einer von $h$ unabhängigen Konstante $c$ gilt

$$\|\varepsilon_{h_1}\|_{h_1} \approx c h_1^p |\ln(h_1)|,$$
$$\|\varepsilon_{h_2}\|_{h_2} \approx c h_2^p |\ln(h_2)|,$$

wobei für klein werdende Schrittweiten $p \approx 2$ gelten sollte, und $c$ sich auf einen moderaten Wert einpendeln sollte. Aus diesen Beziehungen erhält man nun

$$p \approx \frac{\ln\left(\frac{|\ln(h_2)|\|\varepsilon_{h_1}\|_{h_1}}{|\ln(h_1)|\|\varepsilon_{h_2}\|_{h_2}}\right)}{\left|\ln\left(\frac{h_1}{h_2}\right)\right|},$$
$$c \approx \frac{\|\varepsilon_{h_1}\|_{h_1}}{h_1^p |\ln(h_1)|}.$$

In der folgenden Tabelle wird für immer kleinere Schrittweite $h$ aus dem Fehler $\delta$ zweier aufeinanderfolgender Näherungen in der oben beschriebenen Weise $p$ und $c$ ausgerechnet. Man sieht, dass das gewünsche Verhalten, dass sich $p \approx 2$ und $c$ für $h \to 0$ kaum mehr ändern, tatsächlich eintritt, und hat damit eine experimentelle Bestätigung dafür, dass sich die Konvergenzordnung der Mittelpunktsregel tatsächlich auf $|\ln(h)|h^2$ reduziert.

| $h$ | $\delta$ | $p$ | $c$ |
|---|---|---|---|
| $1/5$ | $2.534 \cdot 10^{-01}$ | 1.900778 | $3.355 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-1}$ | $9.709 \cdot 10^{-01}$ | 2.011403 | $4.329 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-2}$ | $3.133 \cdot 10^{-02}$ | 1.977776 | $3.914 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-3}$ | $9.795 \cdot 10^{-02}$ | 1.985776 | $4.031 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-4}$ | $2.937 \cdot 10^{-03}$ | 2.011007 | $4.502 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-5}$ | $8.441 \cdot 10^{-03}$ | 2.011153 | $4.506 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-6}$ | $2.380 \cdot 10^{-04}$ | 2.003675 | $4.315 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-7}$ | $6.648 \cdot 10^{-04}$ | 2.002210 | $4.275 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-8}$ | $1.837 \cdot 10^{-05}$ | 2.002972 | $4.298 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-9}$ | $5.028 \cdot 10^{-05}$ | 2.002564 | $4.284 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-10}$ | $1.365 \cdot 10^{-06}$ | 2.001823 | $4.257 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-11}$ | $3.686 \cdot 10^{-06}$ | 2.001548 | $4.246 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-12}$ | $9.898 \cdot 10^{-07}$ | 2.001537 | $4.246 \cdot 10^{+01}$ |
| $1/5 \cdot 2^{-13}$ | $2.644 \cdot 10^{-08}$ | | |

Die Konvergenz des NEWTON-Verfahrens zur Berechnung der Näherungslösungen für den nichtlinearen Fall ist aber nichtsdestotrotz von der Ordnung, die in Abschnitt 5.2.1 für das implizite EULER-Verfahren gezeigt wurde.

# Kapitel 6

# Asymptotische Fehlerentwicklungen

In diesem Kapitel wird eine asymptotische Fehlerentwicklung für das implizite Eulerverfahren hergeleitet.

Sei $z(t)$ die exakte Lösung der Gleichung

$$\begin{aligned} z'(t) &= \frac{M}{t} z(t) + f(t, z(t)), \quad t \in (0, 1], \\ z(0) &= \zeta_0 \end{aligned} \tag{6.1}$$

und $z_h$ die mit einem numerischen Verfahren berechnete Näherung für $z$. Es wird folgender Ansatz für den globalen Fehler gemacht:

$$z_i - z(t_i) = \sum_{j=1}^{k} h^j e_j(t_i) + r_i, \quad i = 0, \ldots, N. \tag{6.2}$$

In diesem Abschnitt wird gezeigt, dass es unter entsprechenden Voraussetzungen an die Daten des Problems (6.1) glatte, von $h$ unabhängige Funktionen $e_j,\ j = 1, \ldots, k$ gibt, die gewissen *Variationsgleichungen* genügen, und dass für das *Restglied* gilt

$$\|r_h\|_h = O(h^{k+1}).$$

Eine andere Sichtweise von (6.2) ergibt sich, wenn man annimmt, dass die Werte $z_i$ durch eine glatte Interpolierende $\tilde{z}(t)$ verbunden werden. Dann erhält man den Ansatz

$$\tilde{z}(t) = z(t) + \sum_{j=1}^{k} h^j e_j(t) + r(t), \quad t \in [0, 1]. \tag{6.3}$$

Durch Auswertung von (6.3) in den Punkten $t_i$ und $t_{i+1}$ erhält man durch Einsetzen in die definierende Gleichung des numerischen Verfahrens dann Variationsgleichungen (die lineare Differentialgleichungen von der Form von (6.1) sind) für

die $e_j$ und eine Differenzengleichung für das Restglied $r_h$. Aus der Analyse dieser Gleichungen folgen die Existenz und die Glattheitseigenschaften der Entwicklung (6.2). Allgemeine Resultate für reguläre Probleme sind z. B. in [50, S. 21 sqq.] zu finden. Eine asymptotische Fehlerentwicklung für singuläre Probleme wird im Folgenden für das implizite EULERverfahren hergeleitet.

## 6.1 Das implizite EULERverfahren

Betrachte das implizite EULERverfahren (siehe Definition 4.3.1) zur Lösung des Problems (6.1). Im Gegensatz zu den Betrachtungen aus Kapitel 5 nimmt man an, dass mit dem exakten Startwert in $t_{i_0} = 0$ begonnen wird. Für die Näherungslösung $z_h \approx R_h(z)$ soll eine Entwicklung gemäß (6.2) hergeleitet werden. Wie man zu den Variationsgleichungen und zur Abschätzung des Restglieds gelangt, wird exemplarisch für $k = 1$ vorgeführt, woraus die allgemeine Vorgangsweise klar wird. Anschließend werden als Beispiel die Variationsgleichungen für $k = 8$ angegeben und analysiert. Aus dieser Darstellung wird klar, wie die Entwicklung für beliebiges $k$ berechnet werden kann, wenn die Funktion $f(t, z)$ die entsprechenden Glattheitsbedingungen erfüllt.

Sei $f(t, z)$ zweimal stetig differenzierbar auf $[0, 1] \times \mathbb{R}^{2n}$. Dann gilt für die Lösung von (6.1), $z \in C^3([0, 1], \mathbb{R}^{2n})$, siehe Satz 3.5.1. Um eine asymptotische Fehlerentwicklung bis Ordnung 1 für die Näherungslösung $z_h$, die mit dem impliziten EULERverfahren berechnet wurde, zu erhalten, wird der Ansatz

$$z_i = z(t_i) + he_1(t_i) + r_i$$

gemacht. Um die entsprechenden Taylorentwicklungen machen zu können, nimmt man an, dass $e_1 \in C^2([0, 1], \mathbb{R}^{2n})$ gilt. Aus der Analyse der resultierenden Variationsgleichung für $e_1$ wird sich dann zeigen, dass sich unter der gemachten Voraussetzung an $f$ tatsächlich ein zweimal stetig differenzierbares $e_1$ ergibt.

Einsetzen des Ansatzes in die definierende Gleichung des impliziten EULERverfahrens,

$$\frac{z_i - z_{i-1}}{h} = \frac{1}{t_i} M z_i + f(t_i, z_i), \quad i = 1, \ldots, N,$$

und Taylorentwicklung um $t_i$ ergibt für $i = 1, \ldots, N$

$$\begin{aligned} &\frac{1}{h}\left(z(t_i) + he_1(t_i) + r_i - z(t_{i-1}) - he_1(t_{i-1}) - r_{i-1}\right) \\ &\qquad = \frac{1}{t_i} M(z(t_i) + he_1(t_i) + r_i) + f(t_i, z(t_i) + he_1(t_i) + r_i) \\ &\qquad\qquad \Updownarrow \end{aligned}$$

$$
\begin{aligned}
&\frac{1}{h}\Bigg( z(t_i) + he_1(t_i) + r_i - z(t_i) + hz'(t_i) - \frac{h^2}{2}z''(t_i) \\
&+\frac{h^3}{2}\int_0^1 z^{(3)}(t_i - \tau h)(1-\tau)^2\, d\tau \\
&-he_1(t_i) + h^2 e_1'(t_i) - h^3\int_0^1 e_1''(t_i - \tau h)(1-\tau)\, d\tau - r_{i-1}\Bigg) \\
&= \frac{1}{t_i}Mz(t_i) + \frac{h}{t_i}Me_1(t_i) + \frac{1}{t_i}Mr_i \\
&+f(t_i, z(t_i) + he_1(t_i)) + \int_0^1 D_2 f(t_i, z(t_i) + he_1(t_i) + \tau r_i)\, d\tau \cdot r_i \\
&\qquad\qquad \Updownarrow \\
&z'(t_i) - \frac{h}{2}z''(t_i) + he_1'(t_i) + \frac{r_i - r_{i-1}}{h} + O(h^2) \\
&= \frac{1}{t_i}Mz(t_i) + \frac{h}{t_i}Me_1(t_i) + \frac{1}{t_i}Mr_i \\
&+\int_0^1 D_2 f(t_i, z(t_i) + he_1(t_i) + \tau r_i)\, d\tau \cdot r_i \\
&+f(t_i, z(t_i)) + hD_2 f(t_i, z(t_i))e_1(t_i) \\
&+h^2\int_0^1 D_2^2 f(t_i, z(t_i) + \tau h e_1(t_i))(1-\tau)\, d\tau \cdot e_1^2(t_i).
\end{aligned}
$$

Verlangt man, dass die obige Gleichung nicht nur in den Punkten $t_i,\ i = 1, \ldots, N$ gilt, sondern für alle $t \in (0, 1]$, so erhält man durch Vergleich der Koeffizienten von $h$ die Variationsgleichung für $e_1(t)$,

$$
\begin{aligned}
e_1'(t) - \frac{1}{t}Me_1(t) &= D_2 f(t, z(t))e_1(t) + \frac{1}{2}z''(t), \quad t \in (0, 1], \\
e_1(0) &= 0.
\end{aligned}
$$

Die Anfangsbedingung $e_1(0) = 0$ ergibt sich aus $z_0 = z(0)$. Diese Variationsgleichung ist genau vom Typ (6.1), und damit kann man aus Satz 3.5.1 folgern, dass tatsächlich eine Lösung $e_1 \in C^2([0, 1], \mathbb{R}^{2n})$ existiert.

Für das Restglied erhält man unter Beachtung von (6.1) die Differenzengleichung

$$
\begin{aligned}
\frac{r_i - r_{i-1}}{h} - \frac{1}{t_i}Mr_i &= g(t_i, r_i) + l_i, \quad i = 1, \ldots, N, \\
r_0 &= 0,
\end{aligned}
$$

wobei

$$
g(t_i, r_i) := \int_0^1 D_2 f(t_i, z(t_i) + he_1(t_i) + \tau r_i)\, d\tau \cdot r_i
$$

und

$$
l_i = O(h^2), \quad i = 1, \ldots, N,
$$

gilt. Es gilt $g(t,0) = 0$ für alle $t$. Darüberhinaus nimmt man an, dass $D_2 g(t,r)$ existiert und auf $[0,1] \times \mathbb{R}^{2n}$ beschränkt ist. Nun definiert man ähnlich wie in Abschnitt 5.1.2 einen Operator $x_h = G_h(y_h)$ als Lösung der Gleichung

$$\begin{aligned} \frac{x_i - x_{i-1}}{h} - \frac{1}{t_i} M x_i &= g(t_i, y_i) + l_i, \quad i = 1, \ldots, N, \\ x_0 &= 0, \end{aligned}$$

und zeigt wie dort, dass $G_h$ auf einem Intervall $[0, \delta]$ eine Kontraktion mit Konstante $L$ ist, wobei man die Gleichung auf den linearen Fall zurückführt, indem man schreibt

$$g(t_i, x_i) - g(t_i, y_i) = \int_0^1 D_2 g(t_i, y_i + \tau(x_i - y_i))\, d\tau \cdot (x_i - y_i).$$

Damit gibt es einen eindeutigen Fixpunkt $r_h$ von $G_h$, und für diesen gilt

$$\|r_h\|_h \leq \frac{1}{1-L} \|G_h(0)\|_h \leq \text{const.} \|l_h\|_h = O(h^2), \quad h \to 0.$$

**Bemerkung:** Man beachte, dass man unter entsprechenden Voraussetzungen an $f(t,z)$ auch die Konvergenzbeweise in Abschnitt 5.1.3 durch die oben beschriebene Vorgangsweise ersetzen könnte.

Um asymptotische Fehlerentwicklungen von höherer Ordnung zu erhalten, geht man ganz genauso vor, das Ergebnis für Ordnung 8 ist im folgenden Satz formuliert.

**Satz 6.1.1** *Betrachte das Problem (6.1). Für $f(t,z)$ gelte $f \in C^9([0,1] \times \mathbb{R}^{2n}, \mathbb{R}^{2n})$. Dann existiert eine asymptotische Fehlerentwicklung der Gestalt*

$$z_i - z(t_i) = \sum_{j=1}^{8} h^j e_j(t_i) + r_i, \quad i = 0, \ldots, N. \tag{6.4}$$

*Dabei gilt $e_j \in C^{10-j}([0,1], \mathbb{R}^n)$, $j = 1, \ldots, 8$, und die Funktionen $e_j$ erfüllen die folgenden Variationsgleichungen:*

$$\begin{aligned} e_1'(t) - \frac{1}{t} M e_1(t) &= D_2 f(t, z(t)) e_1(t) + \frac{1}{2} z''(t), \quad t \in (0,1], \\ e_1(0) &= 0, \\ e_2'(t) - \frac{1}{t} M e_2(t) &= D_2 f(t, z(t)) e_2(t) + \frac{1}{2} D_2^2 f(t, z(t)) e_1^2(t) \\ &\quad + \frac{1}{2} e_1''(t) - \frac{1}{6} z^{(3)}(t), \quad t \in (0,1], \\ e_2(0) &= 0, \end{aligned}$$

$$
\begin{aligned}
e_3'(t) - \frac{1}{t}Me_3(t) &= D_2f(t,z(t))e_3(t) + D_2^2f(t,z(t))e_1(t)e_2(t) \\
&\quad + \frac{1}{6}D_2^3f(t,z(t))e_1^3(t) + \frac{1}{2}e_2''(t) \\
&\quad - \frac{1}{6}e_1^{(3)}(t) + \frac{1}{24}z^{(4)}(t), \quad t \in (0,1], \\
e_3(0) &= 0, \\
e_4'(t) - \frac{1}{t}Me_4(t) &= D_2f(t,z(t))e_4(t) + D_2^2f(t,z(t))\left(e_1(t)e_3(t) + \frac{1}{2}e_2^2(t)\right) \\
&\quad + \frac{1}{2}D_2^3f(t,z(t))e_1^2(t)e_2(t) + \frac{1}{24}D_2^4f(t,z(t))e_1^4(t) \\
&\quad + \frac{1}{2}e_3''(t) - \frac{1}{6}e_2^{(3)}(t) + \frac{1}{24}e_1^{(4)}(t) - \frac{1}{120}z^{(5)}(t), \quad t \in (0,1], \\
e_4(0) &= 0, \\
e_5'(t) - \frac{1}{t}Me_5(t) &= D_2f(t,z(t))e_5(t) + D_2^2f(t,z(t))\left(e_1(t)e_4(t) + e_2(t)e_3(t)\right) \\
&\quad + \frac{1}{2}D_2^3f(t,z(t))\left(e_1^2(t)e_3(t) + e_1(t)e_2^2(t)\right) \\
&\quad + \frac{1}{6}D_2^4f(t,z(t))e_1^3(t)e_2(t) + \frac{1}{120}D_2^5f(t,z(t))e_1^5(t) \\
&\quad + \frac{1}{2}e_4''(t) - \frac{1}{6}e_3^{(3)}(t) + \frac{1}{24}e_2^{(4)}(t) \\
&\quad - \frac{1}{120}e_1^{(5)}(t) + \frac{1}{720}z^{(6)}(t), \quad t \in (0,1], \\
e_5(0) &= 0, \\
e_6'(t) - \frac{1}{t}Me_6(t) &= D_2f(t,z(t))e_6(t) \\
&\quad + \frac{1}{2}D_2^2f(t,z(t))\left(2e_1(t)e_5(t) + 2e_2(t)e_4(t) + e_3^2(t)\right) \\
&\quad + \frac{1}{6}D_2^3f(t,z(t))\left(6e_1(t)e_2(t)e_3(t) + 3e_1^2(t)e_4(t) + e_2^3(t)\right) \\
&\quad + \frac{1}{12}D_2^4f(t,z(t))\left(4e_1^2(t)e_2^2(t) + e_1^3(t)e_3(t)\right) \\
&\quad + \frac{1}{24}D_2^5f(t,z(t))e_1^4(t)e_2(t) + \frac{1}{720}D_2^6f(t,z(t))e_1^6(t) \\
&\quad + \frac{1}{2}e_5''(t) - \frac{1}{6}e_4^{(3)}(t) + \frac{1}{24}e_3^{(4)}(t) - \frac{1}{120}e_2^{(5)}(t) \\
&\quad + \frac{1}{720}e_1^{(6)}(t) - \frac{1}{5040}z^{(7)}(t), \quad t \in (0,1], \\
e_6(0) &= 0, \\
e_7'(t) - \frac{1}{t}Me_7(t) &= D_2f(t,z(t))e_7(t) \\
&\quad + D_2^2f(t,z(t))\left(e_1(t)e_6(t) + e_2(t)e_5(t) + e_3(t)e_4(t)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{2}D_2^3 f(t,z(t))(2e_1(t)e_2(t)e_4(t)+e_1^2(t)e_5(t)\\
&+e_1(t)e_3^2(t)+e_2^2(t)e_3(t))\\
&+\frac{1}{6}D_2^4 f(t,z(t))\left(e_1^3(t)e_4(t)+3e_1^2(t)e_2(t)e_3(t)+e_1(t)e_2^3(t)\right)\\
&+\frac{1}{24}D_2^5 f(t,z(t))\left(2e_1^3(t)e_2^2(t)+e_1^4(t)e_3(t)\right)\\
&+\frac{1}{120}D_2^6 f(t,z(t))e_1^5(t)e_2(t)\\
&+\frac{1}{5040}D_2^7 f(t,z(t))e_1^7(t)+\frac{1}{2}e_6''(t)\\
&-\frac{1}{6}e_5^{(3)}(t)+\frac{1}{24}e_4^{(4)}(t)-\frac{1}{120}e_3^{(5)}(t)\\
&+\frac{1}{720}e_2^{(6)}(t)-\frac{1}{5040}e_1^{(7)}(t)+\frac{1}{40320}z^{(8)}(t),\quad t\in(0,1],\\
e_7(0) &= 0,\\
e_8'(t)-\frac{1}{t}Me_8(t) &= D_2 f(t,z(t))e_8(t)\\
&+\frac{1}{2}D_2^2 f(t,z(t))\left(2e_1(t)e_7(t)+2e_2(t)e_6(t)+2e_3(t)e_5(t)+e_4^2(t)\right)\\
&+\frac{1}{2}D_2^3 f(t,z(t))(2e_1(t)e_2(t)e_5(t)+e_1^2(t)e_6(t)\\
&+2e_1(t)e_3(t)e_4(t)+e_2^2(t)e_4(t)+e_2(t)e_3^2(t))\\
&+\frac{1}{24}D_2^4 f(t,z(t))(12e_1^2(t)e_2(t)e_4(t)+12e_1(t)e_2^2(t)e_3(t)\\
&+6e_1^2(t)e_3^2(t)+4e_1^3(t)e_5(t)+e_2^4(t))\\
&+\frac{1}{24}D_2^5 f(t,z(t))\left(4e_1^3(t)e_2(t)e_3(t)+e_1^4(t)e_4(t)+2e_1^2(t)e_2^3(t)\right)\\
&+\frac{1}{240}D_2^6 f(t,z(t))\left(2e_1^5(t)e_3(t)+5e_1^4(t)e_2^2(t)\right)\\
&+\frac{1}{720}D_2^7 f(t,z(t))e_1^6(t)e_2(t)\\
&-\frac{1}{40320}D_2^8 f(t,z(t))e_1^8(t)+\frac{1}{2}e_7''(t)-\frac{1}{6}e_6^{(3)}(t)\\
&+\frac{1}{24}e_5^{(4)}(t)-\frac{1}{120}e_4^{(5)}(t)+\frac{1}{720}e_3^{(6)}(t)\\
&-\frac{1}{5040}e_2^{(7)}(t)+\frac{1}{40320}e_1^{(8)}(t)-\frac{1}{362880}z^{(9)}(t),\quad t\in(0,1],\\
e_8(0) &= 0.
\end{aligned}
$$

*Das Restglied $r_h$ erfüllt die Differenzengleichung*

$$
\begin{aligned}
\frac{r_i-r_{i-1}}{h}-\frac{1}{t_i}Mr_i &= g(t_i,r_i)+l_i,\quad i=1,\ldots,N,\\
r_0 &= 0,
\end{aligned}
$$

*wobei*

$$g(t_i, r_i) := \int_0^1 D_2 f \left( t_i, z(t_i) + \sum_{j=1}^{8} h^j e_j(t_i) + \tau r_i \right) d\tau \cdot r_i$$

*gesetzt wurde. Existiert* $D_2 g(t, r)$ *und ist es auf* $[0, 1] \times \mathbb{R}^n$ *beschränkt, dann gilt*

$$\|r_h\|_h = O(h^9).$$

**Bemerkung:** Wie schon betont wurde, lässt sich der obige Satz beliebig verallgemeinern, für $f \in C^k([0, 1] \times \mathbb{R}^{2n}, \mathbb{R}^{2n})$ erhält man eine asymptotische Fehlerentwicklung der Ordnung $k - 1$ mit $e_j \in C^{k+1-j}([0, 1], \mathbb{R}^{2n})$ und $\|r_h\|_h = O(h^k)$.

# Anhang A

# Beweise einiger Sätze

Um die Lesbarkeit dieser Arbeit zu erhöhen, werden in diesem Abschnitt Beweise einiger im Hauptteil dieser Arbeit angeführter Sätze detailliert durchgeführt. Es wird dabei versucht, die einzelnen Teile dieses Anhangs soweit als möglich in sich abgeschlossen zu halten, um die Lektüre zu erleichtern. Dabei werden gewisse Überschneidungen zum Hauptteil in Kauf genommen.

## A.1 Existenz- und Eindeutigkeitssätze für glatte Probleme

Die im Folgenden bewiesenen Sätze werden in Kapitel 2 verwendet.

**Satz A.1.1 (Fixpunktsatz von** Weissinger**)** *Sei $U \neq \emptyset$ eine abgeschlossene Teilmenge eines Banachraums $(E, \| \|)$, $\sum_{n=1}^{\infty} a_n$ eine konvergente Reihe, $a_n > 0$, und $A : U \to U$ eine Selbstabbildung, für die gilt*

$$\|A^n u - A^n v\| \leq a_n \|u - v\| \quad \forall u, v \in U, n \in \mathbb{N}. \tag{A.1}$$

*Dann hat $A$ genau einen Fixpunkt $u$ in $U$, der iterativ gewonnen werden kann als $u = \lim_{n\to\infty} u_n, u_n := A^n u_0$, für einen beliebigen Startwert $u_0 \in U$. Weiters gilt die (*a priori*) Fehlerabschätzung*

$$\|u - u_n\| \leq \left( \sum_{\nu=n}^{\infty} a_\nu \right) \|u_1 - u_0\|. \tag{A.2}$$

*Beweis:* Wegen (A.1) gilt

$$\|u_{n+1} - u_n\| = \|A^n u_1 - A^n u_0\| \leq a_n \|u_1 - u_0\|,$$

und damit nach der Dreiecksungleichung

$$\|u_{n+k} - u_n\| \leq \sum_{\nu=n}^{n+k-1} \|u_{\nu+1} - u_\nu\| \leq \left( \sum_{\nu=n}^{n+k-1} a_\nu \right) \|u_1 - u_0\| \quad \forall k, n \in \mathbb{N}. \quad \text{(A.3)}$$

Aus der Konvergenz der Reihe $\sum_{n=1}^{\infty} a_n$ folgt also, dass die Folge $(u_n)_{n\in\mathbb{N}}$ eine Cauchyfolge ist (der Reihenrest $\sum_{\nu=n}^{\infty} a_\nu$ muss ja für $n \to \infty$ beliebig klein werden), und da $(E, \| \|)$ vollständig ist, gibt es einen Grenzwert $u = \lim_{n\to\infty} u_n$ in $U$. Dieser ist ein Fixpunkt von $A$, da wegen

$$\lim_{n\to\infty} \|u_n - Au\| = \lim_{n\to\infty} \|Au_{n-1} - Au\| \leq \lim_{n\to\infty} a_1 \|u_{n-1} - u\| = 0$$

$(u_n)_{n\in\mathbb{N}}$ nicht nur gegen $u$, sondern auch gegen $Au$ strebt, und deshalb gelten muss $Au = u$. Die Eindeutigkeit des Fixpunkts $u$ sieht man folgendermaßen: Ist $v$ ein weiterer Fixpunkt, dann muss gelten $u = A^j u,\ j = 1, 2, \ldots$ und ebenso $v = A^j v,\ j = 1, 2, \ldots$, und deshalb

$$\|u - v\| = \|A^n u - A^n v\| \leq a_n \|u - v\| \to 0, \quad n \to \infty,$$

also folgt $u = v$. Die Fehlerabschätzung (A.2) erhält man sofort aus (A.3) für $k \to \infty$.

**Bemerkung:** Dieser Beweis stammt aus [57, S. 194 sqq.].

**Definition A.1.2** *Sei $U$ eine Teilmenge eines Banachraums $(E, \| \|)$ und $f : U \to U$ eine Selbstabbildung. Dann heißt $f$* kontrahierend*, wenn es ein $0 \leq L < 1$ gibt mit*

$$\|f(u) - f(v)\| \leq L\|u - v\| \quad \forall u, v \in U.$$

Mit dieser Definition erhält man den folgenden Satz, der einen Spezialfall des Weissingerschen Fixpunktsatzes darstellt.

**Satz A.1.3 (Fixpunktsatz von Banach)** *Sei $U \neq \emptyset$ eine abgeschlossene Teilmenge des Banachraums $(E, \| \|)$ und $A : U \to U$ eine kontrahierende Selbstabbildung gemäß Definition A.1.2. Dann besitzt $A$ genau einen Fixpunkt $u$ in $U$, der iterativ als $u = \lim_{n\to\infty} u_n, u_n := A^n u_0$, mit einem beliebigen Startwert $u_0 \in U$ erhalten werden kann. Weiters gilt die (*a priori*) Fehlerabschätzung*

$$\|u - u_n\| \leq \frac{L^n}{1 - L} \|u_1 - u_0\|.$$

*Beweis:* Analog wie im Weissingerschen Fixpunktsatz mit $a_n = L^n$.

Unter Verwendung des WEISSINGERschen Fixpunktsatzes kann man nun den folgenden Existenz- und Eindeutigkeitssatz beweisen:

**Satz A.1.4 (Existenz- und Eindeutigkeitssatz von PICARD-LINDELÖF)** *Sei $f(t,x)$ stetig auf dem kompakten Quader*

$$R := \{(t,x) \in \mathbb{R}^{n+1} : |t-t_0| \le a, |x-x_0| \le b\}$$

*und dort LIPSCHITZ-stetig bezüglich $x$ mit Konstante $L$. Sei $M$ eine Schranke für $f$ auf $R$, es gelte also*

$$|f(t,x)| \le M \quad \forall (t,x) \in R.$$

*Dann hat das Anfangswertproblem*

$$x'(t) = f(t,x(t)), \quad x(t_0) = x_0 \tag{A.4}$$

*eine eindeutige Lösung $x(t)$ auf dem Intervall $J := [t_0-\alpha, t_0+\alpha]$ mit $\alpha := \min\{a, \frac{b}{M}\}$, die iterativ gewonnen werden kann als $x(t) = \lim_{n\to\infty} \varphi_n(t)$ mit*

$$\varphi_n(t) := x_0 + \int_{t_0}^{t} f(\tau, \varphi_{n-1}(\tau))\, d\tau, \tag{A.5}$$

*wenn der Startwert $\varphi_0(t)$ aus einem geeigneten Bereich gewählt ist. Die Folge $(\varphi_n(t))_{n\in\mathbb{N}}$ konvergiert gleichmäßig gegen $x(t)$. Weiters gelten die Abschätzungen*

$$|x(t)-\varphi_n(t)| \le \frac{(\alpha L)^n}{n!} \mathrm{e}^{\alpha L} \max_{\tau\in J} |\varphi_1(\tau)-\varphi_0(\tau)|, \quad t \in J, \tag{A.6}$$

$$|x(t)-\varphi_n(t)| \le \frac{M}{L} \sum_{\nu=n+1}^{\infty} \frac{(L|t-t_0|)^\nu}{\nu!} + \mu \sum_{\nu=n}^{\infty} \frac{(L|t-t_0|)^\nu}{\nu!}, \quad t \in J, \tag{A.7}$$

$$\mu := \max_{\tau\in J} |\varphi_0(\tau) - x_0|.$$

*Beweis:* Sei $\emptyset \neq U := \overline{K}(x_0, b)$ die abgeschlossene Kugel mit Radius $b$ um den Startwert $x_0$ im Banachraum $(C(J,\mathbb{R}^n), \|\ \|_\infty)$ der stetigen $n$-dimensionalen Funktionen auf dem Intervall $J$ mit der Maximumnorm $\|g\|_\infty := \max_{t\in J} |g(t)|$. Sei eine Abbildung $A$ definiert durch

$$A(x)(t) := x_0 + \int_{t_0}^{t} f(\tau, x(\tau))\, d\tau. \tag{A.8}$$

Aus der obigen Wahl von $\alpha$ folgt sofort

$$\begin{aligned} |A(x)(t) - x_0| &= \left| \int_{t_0}^{t} f(\tau, x(\tau))\, d\tau \right| \\ &\le M|t-t_0| \le M\alpha \le b, \end{aligned}$$

also gilt $A : U \to U$.
Mit vollständiger Induktion wird im Folgenden gezeigt, dass

$$|A^n(u)(t) - A^n(v)(t)| \leq \frac{(L|t-t_0|)^n}{n!}\|u-v\|_\infty \tag{A.9}$$

gilt. Den Induktionsanfang für $n = 1$ erhält man aus

$$\begin{aligned} |A(u)(t) - A(v)(t)| &= \left| \int_{t_0}^{t} (f(\tau, u(\tau)) - f(\tau, v(\tau)))\, d\tau \right| \\ &\leq L \left| \int_{t_0}^{t} |u(\tau) - v(\tau)|\, d\tau \right| \leq L|t-t_0|\|u-v\|_\infty. \end{aligned}$$

Gilt (A.9) für $n-1$, dann folgt

$$\begin{aligned} |A^n(u)(t) - A^n(v)(t)| &\leq \left| \int_{t_0}^{t} |f(\tau, A^{n-1}(u)(\tau)) - f(\tau, A^{n-1}(v)(\tau))|\, d\tau \right| \\ &\leq \left| \int_{t_0}^{t} L|A^{n-1}(u)(\tau) - A^{n-1}(v)(\tau)|\, d\tau \right| \\ &\leq \frac{L^n}{(n-1)!}\|u-v\|_\infty \left| \int_{t_0}^{t} |\tau - t_0|^{n-1}\, d\tau \right| \\ &\leq \frac{(L|t-t_0|)^n}{n!}\|u-v\|_\infty. \end{aligned}$$

Übergang zur Maximumnorm liefert daraus

$$\|A^n(u) - A^n(v)\|_\infty \leq \frac{(\alpha L)^n}{n!}\|u-v\|_\infty =: a_n\|u-v\|_\infty.$$

Damit sind die Voraussetzungen von Satz A.1.1 gezeigt. Daraus folgt die Existenz und Eindeutigkeit der Lösung $x(t)$ von (A.4) sowie die Abschätzung (A.6), da die Lösung von (A.4) äquivalent zur Lösung der VOLTERRAschen Integralgleichung

$$x(t) = x_0 + \int_{t_0}^{t} f(\tau, x(\tau))\, d\tau$$

und damit zur Bestimmung eines Fixpunkts der Abbildung $A$ aus (A.8) ist.
Die Fehlerabschätzung (A.7) zeigt man durch

$$\begin{aligned} x(t) - \varphi_n(t) &= \sum_{\nu=n+1}^{\infty} (\varphi_\nu(t) - \varphi_{\nu-1}(t)), \\ |\varphi_1(t) - \varphi_0(t)| &\leq |\varphi_0(t) - x_0| + \left| \int_{t_0}^{t} f(\tau, \varphi_0(\tau))\, d\tau \right| \\ &\leq \mu + M|t-t_0|, \qquad (A.10) \\ |\varphi_\nu(t) - \varphi_{\nu-1}(t)| &\leq \frac{M}{L}\frac{(L|t-t_0|)^\nu}{\nu!} + \mu\frac{(L|t-t_0|)^{\nu-1}}{(\nu-1)!} \qquad (A.11) \end{aligned}$$

und Anwendung der Dreiecksungleichung. Für den Beweis der Abschätzung genügt es offenbar, $M$ und $\mu$ so zu finden, daß (A.10) gilt. (A.11) berechnet man dann analog zu (A.9) unter Verwendung von (A.10).

**Bemerkungen:**

1. Man kann Satz A.1.4 auch unter Benützung des BANACHschen Fixpunktsatzes A.1.3 beweisen. Benützt man dabei jedoch einfach die Maximumnorm und macht die triviale Abschätzung

$$\|A^n(u) - A^n(v)\|_\infty \leq (\alpha L)^n \|u - v\|_\infty,$$

so erhält man im Allgemeinen eine Einschränkung des Intervalls, für das man die Existenz und Eindeutigkeit zeigen kann. Damit nämlich $A$ kontrahierend ist, muss man zusätzlich noch $\alpha < \frac{1}{L}$ fordern. Man erhält in diesem Fall die (gröbere) Abschätzung

$$\|x - \varphi_n\|_\infty \leq \frac{(\alpha L)^n}{1 - \alpha L} \|\varphi_1 - \varphi_0\|_\infty.$$

Definiert man aber eine andere Norm auf dem Raum der stetigen Funktionen, so erhält man auch bei Verwendung des BANACHschen Fixpunktsatzes das gleiche Gültigkeitsintervall $J := [t_0 - \alpha, t_0 + \alpha], \alpha := \min\{a, \frac{b}{M}\}$ wie bei den Überlegungen in Satz A.1.4:

Sei

$$\begin{aligned} w(t) &:= \mathrm{e}^{L|t-t_0|}, \\ \|f\|_w &:= \sup_{t \in J} \left| \frac{f(t)}{w(t)} \right| \quad \forall f \in C(J, \mathbb{R}^n). \end{aligned}$$

Man erkennt sofort, dass $\| \ \|_w$ eine zu $\| \ \|_\infty$ äquivalente Norm ist, also ist $(C(J, \mathbb{R}^n), \| \ \|_w)$ ein Banachraum. Wähle $\emptyset \neq U := \overline{K}(x_0, b) \in (C(J, \mathbb{R}^n), \| \ \|_w)$, und $A$ wie in (A.8). Dann gilt $A : U \to U$ wegen

$$\|A(x) - x_0\|_w = \sup_{t \in J} \left| \frac{1}{w(t)} \int_{t_0}^t f(\tau, x(\tau))\, d\tau \right| \leq \frac{1}{\inf_{t \in J} |w(t)|} \alpha M \leq b.$$

Die Kontraktionseigenschaft erhält man aus

$$\begin{aligned} \|A(u) - A(v)\|_w &= \sup_{t \in J} \left| \frac{1}{w(t)} \int_{t_0}^t (f(\tau, u(\tau)) - f(\tau, v(\tau)))\, d\tau \right| \\ &\leq \sup_{t \in J} \left| \frac{L}{w(t)} \int_{t_0}^t \left| \frac{u(\tau) - v(\tau)}{w(\tau)} \right| w(\tau)\, d\tau \right| \\ &\leq \sup_{t \in J} \left| \frac{L}{w(t)} \int_{t_0}^t w(\tau)\, d\tau \right| \|u - v\|_w \\ &= \sup_{t \in J} \frac{L}{\mathrm{e}^{L|t-t_0|}} \frac{1}{L} \left( \mathrm{e}^{L|t-t_0|} - 1 \right) \|u - v\|_w \\ &= \left( 1 - \mathrm{e}^{-\alpha L} \right) \|u - v\|_w =: \tilde{L} \|u - v\|_w. \end{aligned}$$

Damit folgen wegen $\tilde{L} < 1$ Existenz und Eindeutigkeit der Lösung von (A.4) auf $U$ aus Satz A.1.3.

2. Ist die rechte Seite $f(t,x)$ in (A.4) *nicht* LIPSCHITZ-stetig bezüglich $x$, so verliert man eventuell die Eindeutigkeit der Lösung, wie das folgende Beispiel beweist:

   **Beispiel:** Gegeben sei das (skalare) Anfangswertproblem

   $$x'(t) = \sqrt{|x(t)|}, \quad x(0) = 0.$$

   Dann sind sowohl die konstante Funktion $x = 0$ als auch die Funktion

   $$x(t) = \begin{cases} -\left(\frac{t}{2}\right)^2, & t \leq 0, \\ \left(\frac{t}{2}\right)^2, & t \geq 0 \end{cases}$$

   Lösungen dieser Gleichung. $\sqrt{|x|}$ ist zwar stetig, jedoch auf keiner Umgebung $U_0$ von $x = 0$ LIPSCHITZ-stetig, wie im Folgenden gezeigt wird.

   Angenommen, es gibt ein $K > 0$ mit

   $$\left|\sqrt{|x|} - \sqrt{|y|}\right| \leq K|x - y| \quad \forall x, y \in U_0 \supseteq \{0\}.$$

   Betrachte die Folgen $(x_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}}$ mit $x_n := \frac{1}{n^2}$ und $y_n := \frac{1}{4n^2}$. Dann müsste gelten

   $$\begin{aligned} \left|\frac{1}{n} - \frac{1}{2n}\right| &\leq K\left|\frac{1}{n^2} - \frac{1}{4n^2}\right| \quad \Leftrightarrow \\ \frac{1}{2n} &\leq K\frac{3}{4n^2} \quad \Leftrightarrow \\ 1 &\leq K\frac{3}{2n} \quad \forall n \in \mathbb{N}. \end{aligned}$$

   Da dies nicht möglich ist, ist $\sqrt{|x|}$ nicht LIPSCHITZ-stetig, und so ist die Existenz zweier verschiedener Lösungen des obigen Anfangswertproblems zu erklären.

Wie das obige Beispiel zeigt, geht eventuell die Eindeutigkeit der Lösung des Anfangswertproblems (A.4) verloren, wenn $f(t,x)$ stetig, nicht jedoch LIPSCHITZ-stetig bezüglich $x$ ist. Es stellt sich nun die Frage, ob man in diesem Fall wenigstens noch etwas über die Existenz (mindestens) einer Lösung aussagen kann. Antwort darauf gibt der folgende Satz.

**Satz A.1.5 (Existenzsatz von PEANO)** *Sei $f(t,x)$ stetig auf dem kompakten Quader*

$$R := \{(t,x) \in \mathbb{R}^{n+1} : |t-t_0| \le a, |x-x_0| \le b\},$$

*und sei $M$ eine Schranke für $f$ auf $R$, gelte also*

$$|f(t,x)| \le M \quad \forall (t,x) \in R.$$

*Dann hat das Anfangswertproblem*

$$x'(t) = f(t,x(t)), \quad x(t_0) = x_0$$

*(mindestens) eine Lösung $x(t)$ auf dem Intervall $J := [t_0-\alpha, t_0+\alpha]$ mit $\alpha := \min\{a, \frac{b}{M}\}$.*

*Beweis:* Es wird wieder wie in Satz A.1.4 gezeigt, dass die Abbildung

$$A(x)(t) := x_0 + \int_{t_0}^{t} f(\tau, x(\tau))\, d\tau \tag{A.12}$$

unter den gegebenen Voraussetzungen (in diesem Fall mindestens) einen Fixpunkt besitzt. Dazu wird der zweite Fixpunktsatz von SCHAUDER B.1.2 verwendet. Definiere $U := \overline{K}(x_0, b) \subseteq (C(J, \mathbb{R}^n), \| \ \|_\infty)$, dann ist $U$ laut Konstruktion abgeschlossen, beschränkt und konvex. $A$ ist eine Selbstabbildung in $U$, wie man aus der Definition von $\alpha$ sofort sieht. Dass $A$ stetig ist folgt aus der Kompaktheit des Quaders $R$. Wegen Satz B.1.5 hat man nämlich die gleichmäßige Stetigkeit der rechten Seite $f$ der Differentialgleichung, und daher gilt

$$\forall \varepsilon > 0 \ \exists \delta > 0 : |u-v| < \delta \Rightarrow |f(t,u) - f(t,v)| < \frac{\varepsilon}{\alpha} \quad \forall t \in J.$$

Es gilt also: $|u-v| < \delta \Rightarrow$

$$\begin{aligned} |A(u)(t) - A(v)(t)| &= \left| \int_{t_0}^{t} (f(\tau, u(\tau)) - f(\tau, v(\tau)))\, d\tau \right| \\ &< |t-t_0| \frac{\varepsilon}{\alpha} \le \varepsilon \quad \forall t \in J. \end{aligned}$$

Damit ist also nur noch zu zeigen, dass $A(U)$ relativ kompakt (cf. Definition B.1.1) ist, um Satz B.1.2 anwenden zu können. Die Menge $C := \overline{A(U)} \subseteq \overline{U} = U$ ist trivialerweise beschränkt und abgeschlossen, damit folgt also aus dem Satz von ARZELÀ-ASCOLI B.1.4, dass $C$ kompakt ist, wenn man noch die gleichgradige Stetigkeit (cf. Definition B.1.3) von $C$ nachweisen kann. Diese zeigt man in zwei Schritten:

- $A(U)$ ist gleichgradig stetig, denn es gilt für alle $A(x) =: \tilde{x} \in A(U), x \in U$, dass
$$|\tilde{x}(t_2) - \tilde{x}(t_1)| = \left| \int_{t_1}^{t_2} f(\tau, x(\tau))\, d\tau \right| \le M |t_2 - t_1|.$$

- Sei $\tilde{y} \in \overline{A(U)}$. Dann gilt

$$\forall \varepsilon > 0 \; \exists \tilde{x} \in A(U) \text{ mit } \|\tilde{x} - \tilde{y}\|_\infty \leq \frac{\varepsilon}{3}.$$

  Weiters folgt aus der gleichgradigen Stetigkeit von $A(U)$

$$\forall \varepsilon > 0 \; \exists \delta > 0 : |t_2 - t_1| < \delta \Rightarrow |\tilde{x}(t_2) - \tilde{x}(t_1)| < \frac{\varepsilon}{3} \quad \forall \tilde{x} \in A(U).$$

  Insgesamt erhält man also, dass aus $|t_2 - t_1| < \delta$ folgt

$$|\tilde{y}(t_2) - \tilde{y}(t_1)| \leq |\tilde{y}(t_2) - \tilde{x}(t_2)| + |\tilde{x}(t_2) - \tilde{x}(t_1)| + |\tilde{x}(t_1) - \tilde{y}(t_1)| \leq \varepsilon,$$

  also die gleichgradige Stetigkeit von $\overline{A(U)}$, und damit ist der Satz bewiesen.

**Bemerkung:** Das Intervall $J$, auf dem die Überlegungen Gültigkeit haben, ist dasselbe wie in Satz A.1.4.

Das folgende Korollar wird benötigt, um Aussagen über die Fortsetzung von Lösungen beweisen zu können.

**Korollar A.1.6** *Sei $V$ eine Teilmenge des $\mathbb{R}^{n+1}$ mit nichtleerem Inneren $\overset{\circ}{V}$ und $f \in C(V, \mathbb{R}^n)$. Sei $\|f\|_\infty \leq M$ auf $V$ und sei $K$ eine kompakte Teilmenge von $\overset{\circ}{V}$. Dann gibt es ein $\delta = \delta(\overset{\circ}{V}, K, M) > 0$, so dass für jedes $(t_0, x_0) \in K$ der Definitionsbereich einer Lösung des Anfangswertproblems*

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0$$

*das Intervall $(t_0 - \delta, t_0 + \delta)$ umfasst.*

*Beweis:* Es kann ohne Einschränkung der Allgemeinheit $M > 0$ angenommen werden. Sei

$$a := \operatorname{dist}(K, \partial V) := \inf_{x \in K} \inf_{y \in \partial V} |x - y|$$

der Abstand von $K$ zum Rand von $V$. Wegen $K \subset \overset{\circ}{V}$ ist $a > 0$. Für jedes $(t_0, x_0) \in K$ und für

$$\delta := \frac{1}{2} \min\{a, \frac{a}{M}\}$$

ist dann der Quader

$$R_\delta(t_0, x_0) := \{(t, x) \in \mathbb{R}^{n+1} : |t - t_0| \leq \delta, |x - x_0| \leq \delta M\}$$

in $\overset{\circ}{V}$ enthalten, und man erhält mit diesem $\delta$ die Existenzaussage aus Satz A.1.5.

Alle bisherigen Existenzaussagen für Lösungen von Differentialgleichungen waren *lokal*, d. h. es konnte nur ein (möglicherweise sehr kleines) Intervall angegeben werden, auf dem die Aussagen Gültigkeit haben. Im Weiteren wird die Frage untersucht, inwieweit man eine solche lokale Lösung fortsetzen kann. Es wird also die Frage nach dem *globalen* Verhalten von Lösungen der Differentialgleichung

$$x'(t) = f(t, x(t)) \tag{A.13}$$

untersucht, wobei $f$ wieder auf einem geeigneten Gebiet $G \subseteq \mathbb{R}^{n+1}$ definiert ist.

Das nächste Lemma gibt eine einfache Bedingung dafür an, dass sich eine Lösung einer Differentialgleichung von einem offenen oder halboffenen Intervall auf den Abschluss des Intervalls fortsetzen lässt.

**Lemma A.1.7** *Sei $f \in C(G, \mathbb{R}^n)$ und $x$ eine Lösung der Differentialgleichung (A.13) auf dem halboffenen Intervall $I := [a, b)$. Gibt es ein Kompaktum $K \subset G$, das den Graphen von $x$, also die Punktmenge $\{(t, x(t)) \in \mathbb{R}^{n+1} : t \in I\}$, enthält, so existiert $\lim_{t\to b} x(t)$ und die Funktion $\tilde{x} : [a, b] \to \mathbb{R}^n$, definiert durch*

$$\tilde{x}(t) := \begin{cases} x(t), & \text{für } t \in I, \\ \lim_{\tau\to b} x(\tau), & \text{für } t = b, \end{cases}$$

*ist stetig differenzierbar und eine Fortsetzung von $x$ auf das abgeschlossene Intervall $\overline{I} = [a, b]$.*

*Beweis:* Die Lösung $x$ genügt auf $I$ der VOLTERRAschen Integralgleichung

$$x(t) = x(a) + \int_a^t f(\tau, x(\tau))\, d\tau, \quad t \in I. \tag{A.14}$$

Da $f$ als stetige Funktion auf dem Kompaktum $K$ (durch eine Konstante $M$) beschränkt ist, gilt für jede Wahl von $t_1, t_2 \in I$ die Ungleichung

$$|x(t_2) - x(t_1)| \le \left| \int_{t_1}^{t_2} f(\tau, x(\tau))\, d\tau \right| \le M|t_2 - t_1|.$$

Also bildet $(x(t_i))_{i\in\mathbb{N}}$ für jede gegen $b$ konvergente Folge $(t_i)_{i\in\mathbb{N}}$ mit $t_i < b, i = 1, 2, \ldots,$ eine Cauchyfolge. Daher existiert $\tilde{x}(b) = \lim_{\tau\to b} x(\tau)$ und es gilt $(b, \tilde{x}(b)) \in K \subset G$. Wegen der gerade nachgewiesenen Stetigkeit von $x$ in $b$ gilt (A.14) auch für $t = b$. Somit genügt $\tilde{x}$ auf dem abgeschlossenen Intervall $[a, b]$ der VOLTERRAschen Gleichung (A.14), ist also insbesondere eine Lösung der Differentialgleichung (A.13) und folglich eine Fortsetzung von $x$ auf $[a, b]$.

**Bemerkung:** Natürlich lässt sich der Beweis für halboffene Intervalle der Form $(a, b]$ und für offene Intervalle $(a, b)$ ganz analog führen.

**Definition A.1.8** *Sei $I = (a, b)$ ein offenes Intervall, $G \subseteq \mathbb{R}^{n+1}$ ein Gebiet, $f \in C(G, \mathbb{R}^n)$ und $x$ eine Lösung der Differentialgleichung (A.13). Dann heißt $I$* rechtsmaximales Existenzintervall von $x$*, wenn es keine Fortsetzung von $x$ auf ein Intervall $J = (\tilde{a}, \tilde{b})$ mit folgenden Eigenschaften gibt:*

1. *$I$ ist in $J$ enthalten.*
2. *$I$ und $J$ haben unterschiedliche rechte Randpunkte.*

*Analog definiert man ein* linksmaximales Existenzintervall*. Ist $I$ sowohl links- als auch rechtsmaximales Existenzintervall von $x$, so heißt $I$* maximales Existenzintervall *der Lösung $x$.*

*Man sagt,* die auf dem Intervall $I$ definierte Lösung $x$ kommt dem Rand von $G$ rechts (beziehungsweise links) beliebig nahe*, wenn es zu jedem Kompaktum $K \subset G$ einen Punkt $t_K \in I$ gibt mit $t_K \geq t_0$ (bzw. $t_K \leq t_0$) und $(t_K, x(t_K)) \notin K$, wobei $t_0$ irgendein Punkt aus $I$ mit $(t_0, x(t_0)) \in K$ ist.*

Mit dieser Definition folgt der nachstehende Satz A.1.9, der die Frage der Fortsetzbarkeit von Lösungen auf maximale Existenzintervalle klärt.

**Satz A.1.9** *Sei $f \in C(G, \mathbb{R}^n)$ und $x$ eine Lösung von (A.13) auf einem Intervall $J$. Dann gibt es eine Fortsetzung $\tilde{x}$ von $x$ auf ein maximales Existenzintervall $I = (x_-, x_+), -\infty \leq x_- < x_+ \leq \infty$, und $\tilde{x}$ kommt dem Rand von $G$ rechts und links beliebig nahe.*

*Beweis:* Offensichtlich genügt es, eine Lösung $x$ von (A.13) auf einem Intervall $I = [a, b)$ zu betrachten und nachzuweisen, dass in diesem Fall eine Fortsetzung auf ein rechtsmaximales Intervall $[a, x_+)$ existiert. Wenn $I$ nicht schon rechtsmaximales Intervall ist, dann hat die Punktmenge $\{(t, x(t)) \in \mathbb{R}^{n+1} : t \in I\}$ einen positiven Abstand von $\partial G$, und man kann nach Lemma A.1.7 das Definitionsintervall durch seinen rechten Randpunkt $b$ abschließen. Seien nun $G_j \subset G, j = 1, 2, \ldots$, offene Mengen mit folgenden Eigenschaften:

1. Der Abschluss $\overline{G_j}$ von $G_j$ ist kompakt in $G$ und es gilt $\overline{G_j} \subset G_{j+1}$ für $j = 1, 2, \ldots$.
2. $\bigcup_{j=1}^{\infty} G_j = G$.

**Bemerkung:** Man kann z. B.

$$G_j := \left\{ (t, x) \in \mathbb{R}^{n+1} : (t, x) \in G, |x| < j, |y| < j, \operatorname{dist}(\{(t, x)\}, \partial G) > \frac{1}{j} \right\}$$

nehmen.

Sei $i$ so groß, dass der Punkt $(b, x(b))$ in $\overline{G_i}$ enthalten ist. Setzt man $K := \overline{G_i}, D := G_{i+1}$, so ist $f$ wegen der Kompaktheit von $\overline{D} \subset G$ auf $D$ beschränkt, und nach Korollar A.1.6 existiert ein $\delta > 0$ und eine Lösung $\hat{x}$ der Anfangswertaufgabe

$$\hat{x}'(t) = f(t, \hat{x}(t)), \quad \hat{x}(b) = x(b)$$

im Intervall $[b, b + \delta]$. Dann ist

$$x^*(t) := \begin{cases} x(t) & \text{für } t \in [a, b], \\ \hat{x}(t) & \text{für } t \in [b, b + \delta] \end{cases}$$

eine Fortsetzung von $x$ auf $[a, b + \delta]$. Da $\overline{G_i}$ beschränkt ist, muss dieser Fortsetzungsprozess *mit dem gleichen* $\delta$ nach endlich vielen (etwa $k$) Wiederholungen zu einem Punkt $b_i := b + k\delta$ und einer Fortsetzung $x_i$ von $x$ auf $[a, b_i]$ mit

$$(b_i - \delta, x_i(b_i - \delta)) \in \overline{G_i},$$

aber

$$(b_i, x_i(b_i)) \in G_{i+1} \backslash \overline{G_i}$$

führen.

Setzt man diesen Prozess fort, so erhält man eine monoton wachsende Folge $(b_i, b_{i+1}, \ldots)$ und zugehörige Lösungen $x_i, x_{i+1}, \ldots$, für die gilt:

$$\begin{gathered} x_{j+1} \text{ ist Fortsetzung von } x_j \text{ — und damit von } x \text{ — auf } [a, b_{j+1}] \\ (b_j, x(b_j)) \in G_{j+1} \backslash \overline{G_j} \text{ für } j = i, i+1, \ldots. \end{gathered} \tag{A.15}$$

Aufgrund dieser Eigenschaft kann eine Fortsetzung $\tilde{x}$ von $x$ auf $[a, x_+)$ mit $x_+ := \lim_{j\to\infty} b_j \in (b, \infty]$ durch

$$\tilde{x}(t) = x_j(t), \text{ falls } t \in [a, b_j]$$

erklärt werden. Offenbar ist $[a, x_+)$ rechtsmaximales Existenzintervall von $\tilde{x}$, da wegen der Eigenschaften der Mengen $G_i$ und der Eigenschaft (A.15) kein Häufungspunkt der Folge $((b_{i+j}, \tilde{x}(b_{i+j})))_{j\in\mathbb{N}}$ im Inneren von $G$ liegt. Dies beweist, dass $\tilde{x}$ dem Rand von $G$ rechts beliebig nahe kommt.

**Bemerkung:** Da unter den getroffenen Annahmen von Satz A.1.9 keine *Eindeutigkeit* der Lösung gegeben ist, ist auch das maximale Existenzintervall $I$ *nicht eindeutig.*

Fasst man nun die *lokale* Aussage des Satzes von Peano A.1.5 und des Fortsetzungssatzes A.1.9 zusammen, so erhält man das folgende Korollar über die *globale* Existenz von Lösungen:

**Korollar A.1.10** *Ist $f \in C(G, \mathbb{R}^n), G \subseteq \mathbb{R}^{n+1}$, und $(t_0, x_0) \in G$, dann besitzt die Anfangswertaufgabe*

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0$$

*eine Lösung, die auf einem maximalen Existenzintervall definiert ist und dem Rand von $G$ rechts und links beliebig nahe kommt.*

**Bemerkung:** Ist $f(t, x)$ nicht nur stetig auf $G$, sondern auch LIPSCHITZ-stetig bezüglich $x$, dann lässt sich der Beweis von Korollar A.1.6 und Satz A.1.9 direkt übernehmen, um in analoger Weise die *eindeutige* Existenz des maximalen Existenzintervalls der *eindeutigen Lösung* von (A.13) zu zeigen.

## A.2 Numerische Lösung singulärer Probleme

**Lemma A.2.1** *Sei $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ mit $\sigma = \Re(\lambda) > 0$ fest gewählt. Definiere für $j \geq k \geq 1$*

$$z_{kj}(\lambda) := \begin{cases} 1, & k = j, \\ \prod_{l=k}^{j-1} \left(1 - \frac{\lambda}{l}\right), & 1 \leq k < j, \quad j = 2, 3, \ldots. \end{cases}$$

*Dann gibt es ein $\eta > 0$ und ein $C \geq 1$, sodass*

$$|z_{kj}(\lambda)| \leq C \left(\frac{k}{j}\right)^{\eta}, \quad 1 \leq k \leq j, \quad j = 1, 2, \ldots. \tag{A.16}$$

*Beweis:* Der folgende Beweis stammt aus [52, S. 443]. Für die *Gammafunktion*

$$\Gamma(z) := \lim_{n \to \infty} \frac{n! n^z}{z(z+1)\cdots(z+n)}$$

(„GAUSSsche Produktdarstellung“) gilt

$$\Gamma(z+1) = z\Gamma(z) \quad \forall z \in \mathbb{C} \backslash \{0, -1, -2, \ldots\}.$$

Für einen Beweis siehe [46, S. 34].

Daraus folgt induktiv sofort

$$\Gamma(z+n) = \Gamma(z) \prod_{k=0}^{n-1} (z+k) \quad \forall n \in \mathbb{N}.$$

Darüberhinaus besagt Satz B.1.9, dass für $a, b \in \mathbb{C}$

$$\frac{\Gamma(s+a)}{\Gamma(s+b)} = s^{a-b} \left(1 + O\left(\frac{1}{s}\right)\right), \quad s \to \infty, \quad \Re(s) > 0$$

gilt.

Damit gilt für hinreichend große $k$ und $j$ mit $j > k$

$$\begin{aligned} z_{kj}(\lambda) &= \prod_{l=k}^{j-1} \frac{l-\lambda}{l} = \frac{\frac{\Gamma(j-\lambda)}{\Gamma(k-\lambda)}}{\frac{\Gamma(j)}{\Gamma(k)}} \\ &= \frac{\Gamma(j-\lambda)}{\Gamma(j)} \frac{\Gamma(k)}{\Gamma(k-\lambda)} = j^{-\lambda}\left(1+O\left(\frac{1}{j}\right)\right) k^{\lambda}\left(1+O\left(\frac{1}{k}\right)\right) \\ &= \left(\frac{k}{j}\right)^{\lambda}\left(1+O\left(\frac{1}{j}\right)\right) = \left(\frac{k}{j}\right)^{\sigma} \mathrm{e}^{\mathrm{i}(\ln(k)-\ln(j))\kappa}\left(1+O\left(\frac{1}{j}\right)\right). \end{aligned}$$

Wie groß $k$ und $j$ gewählt werden müssen, damit die asymptotische Aussage verwendet werden kann, hängt lediglich von $\lambda$ ab, es ist also in jedem Fall nur eine feste, endliche Anzahl von Indizes noch nicht berücksichtigt, auf diese kommt es jedoch nicht an und somit folgt sofort die Abschätzung (A.16).

**Lemma A.2.2** *Sei $h > 0, t_j := jh, k > j \geq i_0 > 0$ und $\gamma \in \mathbb{R}$, dann gilt*

$$\sum_{l=j}^{k-1} h t_l^{\gamma-1} \leq \begin{cases} c_1 |t_k^{\gamma} - t_j^{\gamma}|, & \gamma \neq 0, \\ c_2 \ln\left(\frac{t_k}{t_j}\right), & \gamma = 0. \end{cases} \tag{A.17}$$

*Beweis:* Dieser Beweis stammt sinngemäß aus [52, S. 443]. Es wird die Fallunterscheidung $\gamma \leq 1$ oder $\gamma \geq 1$ gemacht.

1. $\underline{\gamma \geq 1}$: In diesem Fall ist $t_l^{\gamma-1}$ monoton wachsend und es gilt deshalb

$$\begin{aligned} h t_l^{\gamma-1} &\leq \int_{t_l}^{t_{l+1}} \tau^{\gamma-1}\, d\tau = \frac{1}{\gamma}\left(\tau^{\gamma}\Big|_{t_l}^{t_{l+1}}\right) \\ &= \frac{1}{\gamma}(t_{l+1}^{\gamma} - t_l^{\gamma}) \quad \forall l \geq i_0. \end{aligned}$$

   Insgesamt folgt also

$$\sum_{l=j}^{k-1} h t_l^{\gamma-1} \leq \frac{1}{\gamma}|t_k^{\gamma} - t_j^{\gamma}| =: d_1 |t_k^{\gamma} - t_j^{\gamma}|,$$

   da es sich hier um eine Teleskopsumme handelt.

2. $\underline{\gamma \leq 1}$: In diesem Fall gilt

$$h t_l^{\gamma-1} = h\left(t_{l+1}\left(1 - \frac{h}{t_{l+1}}\right)\right)^{\gamma-1}$$

$$
\begin{aligned}
&\leq h\left(1-\frac{1}{i_0+1}\right)^{\gamma-1} t_{l+1}^{\gamma-1} =: d_2 h t_{l+1}^{\gamma-1} \\
&\leq d_2 \int_{t_l}^{t_{l+1}} \tau^{\gamma-1}\, d\tau \\
&= \begin{cases} \frac{d_2}{\gamma}(t_{l+1}^{\gamma}-t_l^{\gamma}), & \gamma \neq 0, \\ d_2 \ln\left(\frac{t_{l+1}}{t_l}\right), & \gamma = 0, \end{cases} \qquad \forall l \geq i_0,
\end{aligned}
$$

da für $\gamma \leq 1$ $t_l^{\gamma-1}$ monoton fallend ist. Zusammenfassend erhält man also (A.17) mit $c_1 = \max\left\{\frac{1}{|\gamma|}, \frac{d_2}{|\gamma|}\right\}$ und $c_2 = d_2$.

**Lemma A.2.3** *Sei $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ mit $\sigma = \Re(\lambda) > 0$ fest gewählt. Definiere für $j \geq k \geq 1$*

$$
\tilde{z}_{kj}(\lambda) := \begin{cases} 1, & k = j, \\ \prod_{l=k}^{j-1}\left(1+\frac{\lambda}{l}\right)^{-1}, & 1 \leq k < j, \quad j = 2,3,\ldots. \end{cases}
$$

*Dann gibt es ein $\eta > 0$ und ein $C \geq 1$, sodass*

$$
|\tilde{z}_{kj}(\lambda)| \leq C\left(\frac{k}{j}\right)^{\eta}, \quad 1 \leq k \leq j, \quad j = 1,2,\ldots. \tag{A.18}
$$

*Beweis:* Ähnlich wie im Beweis von Lemma A.2.1 zeigt man für $j > k$

$$
\begin{aligned}
|\tilde{z}_{kj}| &= \left|\prod_{l=k}^{j-1}\left(\frac{l+\lambda}{l}\right)^{-1}\right| = \left|\prod_{l=k}^{j-1}\frac{l}{l+\lambda}\right| \\
&= \left|\frac{\Gamma(j)}{\Gamma(j+\lambda)}\frac{\Gamma(k+\lambda)}{\Gamma(k)}\right| \\
&= \left|j^{-\lambda}\left(1+O\left(\frac{1}{j}\right)\right)k^{\lambda}\left(1+O\left(\frac{1}{k}\right)\right)\right| \\
&\leq C\left(\frac{k}{j}\right)^{\eta}.
\end{aligned}
$$

# Anhang B

# Verwendete Sätze

In diesem Abschnitt werden Sätze formuliert, die in dieser Arbeit Verwendung finden, deren Beweis aber weder in den Rahmen noch in die Ziele dieser Arbeit passen würde.

## B.1 Sätze aus der Analysis

**Definition B.1.1** *Eine Teilmenge $A$ eines metrischen Raumes $M$ heißt* relativ kompakt*, wenn $\overline{A}$ kompakt ist. Dies ist gleichbedeutend dazu, dass jede Folge in $A$ eine in $M$ konvergente Teilfolge besitzt (cf. dazu [8, S. 188 sq.]).*

**Satz B.1.2 (Zweiter Fixpunktsatz von** Schauder**)** *Sei $U$ eine konvexe, abgeschlossene Teilmenge des normierten Raums $(E, \| \|)$, $f : U \to U$ eine stetige Selbstabbildung und $f(U)$ relativ kompakt. Dann besitzt $f$ (mindestens) einen Fixpunkt.*

*Beweis:* Siehe [17, S. 608]

**Definition B.1.3** *Sei $A$ eine Menge von $n$-dimensionalen Funktionen über der kompakten Menge $\Omega \subset \mathbb{R}^m$. $A$ heißt* gleichgradig stetig*, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, sodass gilt*

$$|t_2 - t_1| < \delta \Rightarrow |f(t_2) - f(t_1)| < \varepsilon \quad \forall t_2, t_1 \in \Omega, \quad \forall f \in A.$$

**Satz B.1.4 (Satz von** Arzelà-Ascoli**)** *Sei $\Omega \subset \mathbb{R}^m$ eine kompakte Menge und $A \subseteq C(\Omega, \mathbb{R}^n)$ eine gleichgradig stetige und bezüglich der Maximumnorm $\| \|_\infty$ abgeschlossene und beschränkte Menge von Funktionen. Dann ist $A$ kompakt in $(C(\Omega, \mathbb{R}^n), \| \|_\infty)$.*

*Beweis:* Siehe [1, S. 87] unter Beachtung der Definitionen von [1, S. 84].

**Satz B.1.5** *Seien $(M, \rho_1)$ und $(N, \rho_2)$ metrische Räume und $U \subseteq M$ eine kompakte Menge. Dann ist jede stetige Funktion $f : U \to N$* gleichmäßig stetig.

*Beweis:* Siehe [16, S. 226]. Der Beweis ist dort zwar nur für reelle Funktionen ausgeführt, lässt sich jedoch bis auf Änderung der Notation für den hier formulierten allgemeinen Fall übernehmen.

**Lemma B.1.6 (Diskretes GRONWALL-Lemma)** *Sei $(x_i)_{i\in\mathbb{N}\cup\{0\}}$ eine Folge nichtnegativer reeller Zahlen, für die gilt*

$$\begin{aligned} x_0 &\leq \delta_0, \\ x_j &\leq (1+\omega)x_{j-1} + \delta, \quad j = 1, 2, \ldots, \end{aligned}$$

*mit $\omega, \delta_0, \delta > 0$. Dann gilt für alle $j \in \mathbb{N} \cup \{0\}$ die Abschätzung*

$$x_j \leq \mathrm{e}^{j\omega}\delta_0 + \frac{\mathrm{e}^{j\omega} - 1}{\omega}\delta, \quad j = 0, 1, \ldots.$$

*Beweis:* Für $\omega \geq 0$ gilt

$$1 + \omega \leq \mathrm{e}^{w}.$$

Mit Hilfe dieser Abschätzung und der folgenden offensichtlichen Rekursion erhält man

$$\begin{aligned} x_j &\leq (1+\omega)^j\delta_0 + \delta\sum_{i=0}^{j-1}(1+\omega)^i \\ &= (1+\omega)^j\delta_0 + \frac{(1+\omega)^j - 1}{\omega}\delta \\ &\leq \mathrm{e}^{j\omega}\delta_0 + \frac{\mathrm{e}^{j\omega} - 1}{\omega}\delta. \end{aligned}$$

**Satz B.1.7 (TAYLORscher Lehrsatz)** *Seien $(E, \| \ \|_E)$ und $(F, \| \ \|_F)$ zwei Banachräume, $A$ eine offene Teilmenge von $E$ und $f \in C^p(A, F)$. Ist dann die Verbindungsstrecke von $x$ und $x + y$ in $A$, so gilt*

$$f(x+y) = \sum_{j=0}^{p-1}\frac{1}{j!}f^{(j)}(x)(y^j) + \left(\int_0^1 \frac{(1-\tau)^{p-1}}{(p-1)!}f^{(p)}(x+\tau y)\,d\tau\right)(y^p),$$

*wobei $y^k$ für $(y, \ldots, y) \in E^k$ steht. Speziell gilt, dass für jedes $\varepsilon > 0$ ein $r > 0$ existiert, sodass für $\|y\|_E \leq r$ gilt*

$$\|f(x+y) - \sum_{j=0}^{p}\frac{1}{j!}f^{(j)}(x)(y^j)\|_F \leq \varepsilon\|y\|_E^p.$$

*Beweis:* Siehe [9, S. 186].

**Korollar B.1.8** *Sei $I := [a, b]$ ein kompaktes Intervall und $f \in C^{p-1}(I, \mathbb{R})$ und die $p$-te Ableitung existiere wenigstens im Inneren $\overset{\circ}{I}$ von $I$. Dann gibt es ein $\theta \in \overset{\circ}{I}$, sodass gilt*

$$f(b) = \sum_{j=0}^{p-1} \frac{1}{j!} f^{(j)}(a)(b-a)^j + \frac{1}{p!} f^{(p)}(\theta)(b-a)^p.$$

*Beweis:* Siehe [16, S. 352 sqq.].

**Satz B.1.9** *Für die* Gammafunktion

$$\Gamma(z) := \lim_{n\to\infty} \frac{n! n^z}{z(z+1)\cdots(z+n)}, \quad z \in \mathbb{C}$$

*(„Gauss*sche *Produktdarstellung“) gilt*

$$\frac{\Gamma(s+a)}{\Gamma(s+b)} = s^{a-b}\left(1 + O\left(\frac{1}{s}\right)\right), \quad s \to \infty, \quad \Re(s) > 0.$$

*Beweis:* Siehe [36, S. 33].

**Satz B.1.10 (Erweiterter Mittelwertsatz der Integralrechnung)** *Die Funktionen $f, g$ seien auf dem Intervall $[a, b]$* Riemann*-integrierbar und es sei $g \leq 0$ oder $g \geq 0$. Dann gibt es ein $\mu \in \mathbb{R}$ mit $\inf_{\theta\in[a,b]} f(\theta) \leq \mu \leq \sup_{\theta\in[a,b]} f(\theta)$, sodass gilt*

$$\int_a^b f(\tau)g(\tau)\,d\tau = \mu \int_a^b g(\tau)\,d\tau.$$

*Ist $f$ stetig, dann gibt es ein $\xi \in [a, b]$ mit $\mu = f(\xi)$.*

*Beweis:* Siehe [16, S. 477].

**Definition B.1.11 (Banachalgebra)** *Sei $(E, \| \|)$ ein Banachraum über $\mathbb{C}$, auf dem eine Multiplikation erklärt ist für die gilt*

1. *$E$ ist ein Ring mit Einselement $I$.*
2. *$\alpha(xy) = (\alpha x)y = x(\alpha y) \quad \forall \alpha \in \mathbb{C}, x, y \in E.$*
3. *$\|xy\| \leq \|x\|\|y\| \quad \forall x, y \in E.$*

*Dann heißt $(E, \| \|)$ eine* Banachalgebra.

**Lemma B.1.12** *Sei $(E, \| \|)$ ein Banachraum, dann ist die Menge aller stetigen linearen Abbildungen von $E$ in $E$ eine Banachalgebra.*

*Beweis:* Siehe [19, S. 115].

**Lemma B.1.13 (**von Neumann**-Lemma)** *Sei $(E, \| \|)$ eine Banachalgebra und $x \in E$ mit $\|x\| < 1$. Dann ist $I - x$ invertierbar und es gilt*

$$(I-x)^{-1} = \sum_{i=0}^{\infty} x^i,$$
$$\|(I-x)^{-1}\| \leq \frac{1}{1-\|x\|}.$$

*Beweis:* Siehe [34, S. 27 sq.]. Der dort angegebene Beweis für die Menge der linearen Abbildungen $L : \mathbb{R}^n \to \mathbb{R}^n$ lässt sich direkt für eine allgemeine Banachalgebra verwenden.

## B.2 Sätze aus der linearen Algebra

**Satz B.2.1** *Zwei Matrizen $A, B \in \mathbb{C}^{n\times n}$ heißen* ähnlich*, wenn es eine invertierbare Matrix $P$ gibt mit*

$$B = P^{-1}AP.$$

*Das* charakteristische Polynom $\chi_A(x)$ *einer Matrix $A \in \mathbb{C}^{n\times n}$ ist erklärt als*

$$\chi_A(x) := \det(A - xI),$$

*wobei $I$ die (n-dimensionale) Einheitsmatrix bezeichnet.*

*Die* Spur *einer quadratischen Matrix ist definiert als die Summe ihrer Diagonalelemente.*

*Es gilt: Für zwei ähnliche Matrizen stimmen die Koeffizienten des charakteristischen Polynoms überein, insbesondere haben ähnliche Matrizen dieselbe Determinante und dieselbe Spur.*

*Beweis:* Siehe [30, S. 73].

**Satz B.2.2 (**Jordan**-Normalform)** *Sei $A \in \mathbb{R}^{n\times n}$ und das charakteristische Polynom (cf. B.2.1) zerfalle in Linearfaktoren (eventuell über $\mathbb{C}$), also*

$$\chi_A(x) = \prod_{k=1}^{r} (x - \mu_k)^{m_k}$$

*mit paarweise verschiedenen $\mu_k$. Dann ist $A$ ähnlich zu einer Matrix der Form*

$$\begin{pmatrix} A(\mu_1) & 0 & \cdots & 0 \\ 0 & A(\mu_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A(\mu_r) \end{pmatrix}. \tag{B.1}$$

*Dabei ist $A(\mu_k)$ eine Matrix, welche aus* JORDAN*-Matrizen $J_m(\mu_k)$ längs der Diagonale aufgebaut ist.* JORDAN*-Matrizen $J_m$ haben die Gestalt*

$$J_m(\lambda) := \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ \vdots & \vdots & 0 & \ddots & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \in \mathbb{C}^{m\times m}, \tag{B.2}$$

*wobei $\lambda$ ein Eigenwert von $A$ ist.*

*Die Matrix (B.1) heißt* JORDAN-Normalform *von $A$. Zwei Matrizen mit in Linearfaktoren zerfallenden charakteristischen Polynomen sind genau dann ähnlich, wenn sie dieselbe* JORDAN*-Normalform besitzen.*

*Beweis:* Siehe [30, S.84].

# Literaturverzeichnis


[1] Alt, H. W. (1992): Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung (2. Aufl.). Berlin, Heidelberg: Springer.

[2] Anderson, N., Arthurs, A. M. (1981): Variational Solutions of the Thomas-Fermi Equations. Quart. Appl. Math. 39, pp. 127–129.

[3] Brabston, D. C. (1974): Numerical Solution of Singular Endpoint Boundary Value Problems. Ph. D. thesis, part II, Applied Mathematics, California Institute of Technology, Pasadena.

[4] Brabston, D. C., Keller, H. B. (1977): Numerical Methods for Singular Two-Point Boundary Value Problems. SIAM J. Numer. Anal. 14, pp. 779–791.

[5] Chan, C. Y., Hon, Y. C. (1987): A Constructive Solution for a Generalized Thomas-Fermi Theory of Ionized Atoms. Quart. Appl. Math. 45, pp. 591–599.

[6] Chan, C. Y, Du, S. W. (1986): A Constructive Method for the Thomas-Fermi Equation. Quart. Appl. Math. 44, pp. 303–307.

[7] Chen, X. (1996): Lorenz Equations: Existence and Nonexistence of Homoclinic Orbits. SIAM J. Math. Anal. 27, pp. 1057–1069.

[8] Cryer, C. W. (1982): Numerical Functional Analysis. New York: Oxford University Press.

[9] Dieudonné, J. (1960): Foundations of Modern Analysis. New York: Academic Press.

[10] Fichtenholz, G. M. (1990): Differential- und Integralrechnung, Band II. VEB Deutscher Verlag der Wissenschaften, Berlin.

[11] Flagg, R. C., Luning, C. D., Perry, W. L. (1980): Implementation of New Iterative Techniques for Solutions of Thomas-Fermi and Emden-Fowler Equations. J. Comp. Phys. 38, pp. 396–405.

[12] Frank, R. (1976): The Method of Iterated Defect Correction and its Application to Two-Point Boundary Value Problems, Part I. Numer. Math. 25, pp. 409–419.

[13] Frank, R. (1977): The Method of Iterated Defect Correction and its Application to Two-Point Boundary Value Problems, Part II. Numer. Math. 27, pp. 407–420.

[14] Frommlet, F., Weinmüller, E. (1995): Asymptotic Error Expansions for Singular BVP's. Eingereicht bei $M^3AS$ (Richard Weiss Memorial Issue). Vorabdruck: Technical Report No. 118E(1995), Institut für Angewandte und Numerische Mathematik, Technische Universität Wien.

[15] Grossmann, C. (1992): Enclosures of the Solution of the Thomas-Fermi Equation by Monotone Discretization. J. Comp. Phys. 98, pp. 26–38.

[16] Heuser, H. (1991): Lehrbuch der Analysis, Teil 1 (9. Aufl.). Stuttgart: Teubner.

[17] Heuser, H. (1991): Lehrbuch der Analysis, Teil 2 (7. Aufl.). Stuttgart: Teubner.

[18] Heuser, H. (1991): Gewöhnliche Differentialgleichungen: Einführung in Lehre und Gebrauch (2. Aufl.). Stuttgart: Teubner.

[19] Heuser, H. (1992): Funktionalanalysis (3. Aufl.). Stuttgart: Teubner.

[20] de Hoog, F. R., Weiss, R. (1976): Difference Methods for Boundary Value Problems with a Singularity of the First Kind. SIAM J. Numer. Anal. 13, pp. 775–813.

[21] de Hoog, F. R., Weiss, R. (1978): Collocation Methods for Singular Boundary Value Problems. SIAM J. Numer. Anal. 15, pp. 198–217.

[22] de Hoog, F. R., Weiss, R. (1979): The numerical Solution of Boundary Value Problems with an Essential Singularity. SIAM J. Numer. Anal. 16, pp. 637–669.

[23] de Hoog, F. R., Weiss, R. (1980): An Approximation Theory for Boundary Value Problems on Infinite Intervals. Computing 24, pp. 227–239.

[24] de Hoog, F. R.,Weiss, R. (1980): On the Boundary Value Problem for Systems of Ordinary Differential Equations with a Singularity of the Second Kind. SIAM J. Math. Anal. 11, pp. 41–60.

[25] de Hoog, F. R., Weiss, R. (1985): The Application of Runge-Kutta Schemes to Singular IVPs. Math. Comp. 44, pp. 93–103.

[26] Jamet, P. (1970): On the Convergence of Finite Difference Approximations to One-Dimensional Singular Boundary Value Problems. Numer. Math. 14, pp. 335–378.

[27] Keller, H. B.,Wolfe, A. W. (1965): On the Nonunique Equilibrium States and Buckling Mechanism of Spherical Shells. SIAM J. 13, pp. 674–705.

[28] Keller, H. B. (1970): Newton's Method under Mild Differentiability Conditions. J. Comput. System Sci. 4, pp. 15-28.

[29] Keller, H. B. (1975): Approximation Methods for Nonlinear Problems with Application to Two-Point Boundary Value Problems. Math. Comp. 29, pp. 464–474.

[30] Klingenberg, W. (1992): Lineare Algebra und Geometrie (3. Aufl. ). Berlin–Heidelberg: Springer Verlag.

[31] Koch, O., Kofler, P., Weinmüller, E. (1998): On the Initial Value Problems for Systems of Ordinary Differential Equations with a Singularity of the First Kind. Eingereicht bei SIAM J. Numer. Anal.

[32] Koch, O., Kofler, P., Weinmüller, E. (1998): Analysis and Numerical Treatment of Singular Initial Value Problems. Eingereicht bei Applied Numerical Mathematics.

[33] Kofler, P. (1998): Dissertation.

[34] Kosmol, P. (1989): Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben. Stuttgart: Teubner.

[35] Liu, L., Moore, G., Russell, R. D. (1997): Computation and Continuation of Homoclinic and Heteroclinic Orbits with Arclength Parameterization. SIAM J. Sci. Comp. 18, pp. 69–94.

[36] Luke, Y. L. (1969): The Special Functions and their Approximations. New York: Academic Press.

[37] Luning, C. D. (1978): An Iterative Technique for Obtaining Solutions of a Thomas-Fermi Equation. SIAM J. Math. Anal. 9, pp. 515–522.

[38] Luning, C. D., Perry, W. L. (1977): An Iterative Technique for Solutions of the Thomas-Fermi Equation Utilizing a Nonlinear Eigenvalue Problem. Quart. Appl. Math. 35, pp. 257–268.

[39] Meissner, H., Tholfsen, P. (1968): Cylindrically Symmetric Solutions of the Ginzburg-Landau Equation. Phys. Rev. 169, pp. 413-416.

[40] Moore, G. (1995): Computation and Parametrization of Periodic and Connecting Orbits. IMA J. Num. Anal. 15, pp. 245–263.

[41] Moore, G. (1995): Geometric Methods for Computing Invariant Manifolds. Appl. Num. Math. 17, pp. 319–331.

[42] Natterer, F. (1973): A Generalized Spline Method for Singular Boundary Value Problems in Ordinary Differential Equations. Linear Algebra and Appl. 7, pp. 189–216.

[43] Natterer, F. (1975): Das Differenzenverfahren für singuläre Rand-Eigenwertaufgaben gewöhnlicher Differentialgleichungen. Numer. Math. 23, pp. 387–409.

[44] Reddien, G. W. (1973): Projection Methods for Singular Two-Point Boundary Value Problems. Numer. Math. 21, pp. 193–203.

[45] Reddien, G. W., Schumaker, L. L. (1976): On a Collocation Method for Singular Two-Point Boundary Value Problems. Numer. Math. 25, pp. 427–432.

[46] Remmert, R. (1995): Funktionentheorie II (2. Aufl.). Berlin, Heidelberg: Springer.

[47] Rentrop, P. (1977): Eine Taylorreihenmethode zur numerischen Lösung von Zwei-Punkt Randwertproblemen mit Anwendung auf singuläre Probleme der nichtlinearen Schalentheorie. TUM-MATH-7733, Technische Universität München.

[48] Rentrop, P. (1976): Numerical Solution of the Singular Ginzburg-Landau Equations by Multiple Shooting. Computing 16, pp. 61–67.

[49] Russel, R. D., Shampine, L. F. (1975): Numerical Methods for Singular Boundary Value Problems. SIAM J. Numer. Anal. 12, pp. 13–36.

[50] Stetter, H. J. (1973): Analysis of Discretization Methods for Ordinary Differential Equations. Berlin: Springer Verlag.

[51] Weinmüller, E. (1984): On the Boundary Value Problems for Systems of Ordinary Second Order Differential Equations with a Singularity of the First Kind. SIAM J. Math. Anal. 15, pp. 287–307.

[52] Weinmüller, E. (1984): A Difference Method for a Singular Boundary Value Problem of Second Order. Math. Comp. 42, pp. 441–464.

[53] Weinmüller, E. (1986): On the Numerical Solution of Singular Boundary Value Problems of Second Order by a Difference Method. Math. Comp. 46, pp. 93–117.

[54] Weinmüller, E. (1986): Collocation for Singular Boundary Value Problems of Second Order. SIAM J. Numer. Anal. 23, pp. 1062–1095.

[55] Weinmüller, E. (1989): Stability of Singular Boundary Value Problems and Their Discretization by Finite Differences. SIAM J. Numer. Anal. 26, pp. 180–213.

[56] Weinmüller, E., Winkler, R. (1988): Pathfollowing Algorithm for Singular Boundary Value Problems. ZAMM 68, pp. 527–537.

[57] Weissinger, J. (1952): Zur Theorie und Anwendung des Iterationsverfahrens. Math. Nachr. 8, S. 193-212.

[58] Werner, H., Arndt, H. (1986): Gewöhnliche Differentialgleichungen: eine Einführung in Theorie und Praxis. Berlin-Heidelberg: Springer.

## Aktuelle Publikationen aus einigen Bereichen der Naturwissenschaften:

### Thomas-Fermi-Gleichungen

[59] Cedillo, A. (1993): A Perturbative Approach of the Thomas-Fermi Equation in Terms of the Density. J. Math. Phys. 34, pp. 2713–2717.

[60] Fernandez, F. M., Ogilvie, J. F. (1990): Approximate Solutions to the Thomas-Fermi Equation. Phys. Rev. A 42, pp. 149–154.

[61] Leung, Y. C., Pei, S. Y. (1989): High-Density Expansions of the Relativistic Thomas-Fermi Equation. Phys. Rev. A 40, pp. 2731–2737.

[62] Mooney, J. W. (1993): Solution of a Thomas-Fermi Problem Using Linear Approximants. Comp. Phys. Comm. 76, pp. 51–57.

[63] Morales, D. A. (1994): An N-Dimensional Interpretation of the Delta-Expansion for the Thomas-Fermi Equation. J. Math. Phys. 35, pp. 3916–3921.

[64] Parker, G. W. (1990): Solution of the Thomas-Fermi Equation for Molecules by an Efficient Relaxation Method. J. Math. Phys. 31, pp. 2535–2537.

[65] Pfalzner, S., Rose, S. J. (1993): Thomas-Fermi Model for Dense Plasmas with Strong Electric Fields. Comp. Phys. Comm. 75, pp. 98–104.

[66] Siringo, F., Piccitto, G., Pucci, R. (1993): Thomas-Fermi Model for the $C_{60}$ Molecule. Phys. Rev. E 48, pp. 263–272.

[67] Tu, K. (1991): Analytic Solution to the Thomas-Fermi and Thomas-Fermi-Dirac-Weizsacker Equations. J. Math. Phys. 32, pp. 2250–2253.

[68] Zimmerer, P., Zimmermann, M., Grun, N., Scheid, W. (1991): Trajectory Method for the Time-Dependent Schrödinger and Thomas-Fermi Equations. Comp. Phys. Comm. 63, pp. 21–27.

### Ginzburg-Landau-Gleichungen

[69] Bauman, P., Phillips, D., Tang, Q. (1996): On the Onset of Superconductivity for the Ginzburg-Landau Equations. Z. Ang. Math. Mech. 76, pp. 277–279.

[70] Carr, T. W., Erneux, T. (1994): Understanding the Bifurcation to Traveling Waves in a Class-B Laser Using a Degenerate Ginzburg-Landau Equation. Phys. Rev. A 50, pp. 4219–4227.

[71] Descalzi, O., Graham, R. (1994): Nonequilibrium Potential for the Ginzburg-Landau Equation in the Phase-Turbulent Regime. Z. Phys. B 93, pp. 509–513.

[72] Fleckinger-Pelle, J., Kaper, H. G. (1996): Gauges for the Ginzburg-Landau equations of superconductivity. Z. Ang. Math. Mech. 76, pp. 345–348.

[73] Jimbo, S., Morita, Y. (1996): Ginzburg-Landau Equations and Stable Solutions in a Rotational Domain. SIAM J. Math. Anal. 27, pp. 1360–1385.

[74] Jinqiao, D., Ly, H. V., Titi, E. S. (1996): The Effect of Nonlocal Interactions on the Dynamics of the Ginzburg-Landau Equation. Z. Ang. Math. Phys. 47, pp. 432–455.

[75] Kapitula, T. (1996): Existence and Stability of Singular Heteroclinic Orbits for the Ginzburg-Landau Equation. Nonlinearity 9, pp. 669–685.

[76] Kapitula, T., Maier-Paape, S. (1996): Spatial Dynamics of Time Periodic Solutions for the Ginzburg-Landau Equation. Z. Ang. Math. Phys. 47, pp. 265–305.

[77] Rezlescu, N., Agop, M., Buzea, C., Buzea, C. G. (1996): Perturbative Solutions of the Ginzburg-Landau Equation and the Superconducting Parameters. Phys. Rev. B 53, pp. 2229–2232.

[78] Tang, Q., Wang, S. (1995): Time Dependent Ginzburg-Landau Equations of Superconductivity. Phys. Rev. D 88, pp. 139–166.

# Index

# Lebenslauf

## Persönliche Daten

| | |
|---|---|
| Name: | Othmar Gerhard Koch. |
| geboren am: | 26.10.1973. |
| Eltern: | Edith Koch, Ing. Otmar Koch. |
| Familienstand: | ledig. |
| Anschrift: | Schumanngasse 11/1, 1180 Wien. |
| Interessen: | Schach, Tischtennis, Darts. |

## Schulbildung

| | |
|---|---|
| 09/80 – 06/84 | Volksschule in 2243 Matzen (NÖ). |
| 09/84 – 06/92 | BG 2230 Gänserndorf, neusprachlicher Zug,<br>Gewinn der niederösterreichischen Englischolympiade. |

## Studium

| | |
|---|---|
| 10/92 – 09/96 | Diplomstudium Technische Mathematik,<br>Mathematik in den Naturwissenschaften,<br>an der Technischen Universität Wien.<br>Diplomarbeit:<br>„Multi–Spezies–Systeme in der Populationsdynamik“<br>am Institut für Algebra und Diskrete Mathematik (E118).<br>Hauptfächer: Biomathematik, Numerik, Funktionalanalysis. |
| 03/97 – 12/98 | Doktoratsstudium der Technischen Wissenschaften.<br>Hauptfach: Singuläre Differentialgleichungen.<br>Zweitfach: Partielle Differentialgleichungen. |
| 07/95 – 06/99 | Studienrichtungsvertreter für Technische Mathematik,<br>2 Jahre stellvertretender Vorsitzender. |
| 07/97 – 06/99 | Vorsitzender der Fakultätsvertretung an der<br>Technisch–Naturwissenschaftlichen Fakultät. |
| 03/96 – ??? | Tutor für die Übungen „Höhere Analysis für TPH“<br>und „Analysis 1 für TPH“. |

Wien, am 22. Juli 1999