

**Zur Numerik  
der Rand- und Eigenwertprobleme  
gewöhnlicher Differenzialgleichungen**

Werner Vogt  
Technische Universität Ilmenau  
Institut für Mathematik  
Postfach 100565  
98684 Ilmenau

Ilmenau, den 6. Februar 2006



## 2 Problemklassen und Standardform

Häufig sind Probleme der Praxis nicht in der Standardform (2) vorgegeben. Die Überführung von Differenzialgleichungen höherer Ordnung und deren Anfangsbedingungen in Systeme 1. Ordnung ist allgemein üblich, um Standardsoftware für Anfangswertprobleme anwenden zu können. Anders verhält es sich bei Vorliegen von *Randbedingungen*, deren Standardform  $g(x(a), x(b)) = 0$  erfahrungsgemäß schwerer verstanden wird. Wir wollen deshalb diese Transformation exemplarisch an einigen wesentlichen Problemklassen demonstrieren.

### 2.1 Differenzialgleichungen höherer Ordnung

Ist das Randwertproblem für eine explizite Gleichung n-ter Ordnung

$$x^{(n)} = f(t, x, \dot{x}, \dots, x^{(n-1)}), \quad (3)$$

mit n allgemeinen Randbedingungen

$$g_i(x(a), \dot{x}(a), \dots, x^{(n-1)}(a), x(b), \dot{x}(b), \dots, x^{(n-1)}(b)) = 0, \quad i = 1(1)n \quad (4)$$

gegeben, so erhalten wir mit den neuen Funktionen

$$x_1(t) = x(t), \quad x_2(t) = \dot{x}(t), \dots, \quad x_n(t) = x^{(n-1)}(t)$$

das spezielle System 1. Ordnung für  $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$

$$\begin{array}{llll} \dot{x}_1 & = & x_2 & , & g_1(x_1(a), \dots, x_n(b)) & = & 0 \\ \dot{x}_2 & = & x_3 & , & g_2(x_1(a), \dots, x_n(b)) & = & 0 \\ \dot{x}_3 & = & x_4 & , & g_3(x_1(a), \dots, x_n(b)) & = & 0 \\ \dots & & & & \dots & & \dots \\ \dot{x}_{n-1} & = & x_n & , & g_{n-1}(x_1(a), \dots, x_n(b)) & = & 0 \\ \dot{x}_n & = & f(t, x_1, x_2, \dots, x_n) & , & g_n(x_1(a), \dots, x_n(b)) & = & 0. \end{array}$$

Dabei kann jede Randbedingung  $g_i$  von sämtlichen  $n$  Lösungswerten  $x_1(a), \dots, x_n(a)$  am linken Rand  $a$  und von den entsprechenden Werten  $x_1(b), \dots, x_n(b)$  am rechten Rand  $b$  abhängen. Ähnlich kann man DGL-Systeme höherer Ordnung auf die Standardform (2) reduzieren.

**Beispiel 1** 1. Das skalare lineare Randwertproblem 2. Ordnung

$$\ddot{x} + x \cdot \cosh t = 0, \quad x(0) = 0, \quad x(1) = 1$$

wird leicht auf die Standardform transformiert:

$$\begin{array}{llll} \dot{x}_1 & = & x_2, & g_1(x_1(0), x_2(0), x_1(1), x_2(1)) := x_1(0) & = & 0 \\ \dot{x}_2 & = & -x_1 \cdot \cosh t, & g_2(x_1(0), x_2(0), x_1(1), x_2(1)) := x_1(1) - 1 & = & 0. \end{array}$$

2. Die Durchbiegungsfunktion  $\eta(\xi)$  einer Eisenbahnschiene der Länge  $2L$  genügt der DGL

$$\begin{aligned} \frac{d^2}{d\xi^2} \left( E \cdot J(\xi) \frac{d^2 \eta}{d\xi^2} \right) + K\eta &= q(\xi), \\ \eta''(-L) &= \eta''(+L) = 0, \\ \eta'''(-L) &= \eta'''(+L) = 0 \end{aligned}$$

auf dem Intervall  $I = [-L, L]$ . Die 4 Randbedingungen bedeuten, dass Biegemoment und Scherungskraft bei  $\xi = \pm L$  verschwinden sollen. Elastizitätsmodul  $E > 0$  der Schiene und der Schienenbettung  $K > 0$  sind gegeben sowie die reellen Funktionen  $J(\xi)$  (Flächenträgheitsmoment) und  $q(\xi)$  (Belastung der Schiene). Hier empfiehlt sich eine problemangepasste Transformation mit den Funktionen

$$x_1(\xi) = \eta(\xi), \quad x_2(\xi) = \eta'(\xi), \quad x_3(\xi) = E \cdot J(\xi)\eta''(\xi), \quad x_4(\xi) = \frac{d}{d\xi} (E \cdot J(\xi)\eta''(\xi)),$$

womit wir durch Differentiation und Einsetzen der Randwerte die Standardform

$$\begin{aligned} x_1' &= x_2 & , & \quad g_1(x_1(-L), \dots, x_4(L)) := x_3(-L) = 0 \\ x_2' &= \frac{1}{EJ(\xi)} \cdot x_3 & , & \quad g_2(x_1(-L), \dots, x_4(L)) := x_3(+L) = 0 \\ x_3' &= x_4 & , & \quad g_3(x_1(-L), \dots, x_4(L)) := x_4(-L) = 0 \\ x_4' &= -Kx_1 + q(\xi) & , & \quad g_4(x_1(-L), \dots, x_4(L)) := x_4(+L) = 0 \end{aligned}$$

für  $x(\xi) = (x_1(\xi), x_2(\xi), x_3(\xi), x_4(\xi))^T$  auf  $I = [-L, L]$  erhalten.

3. Die Bewegung eines Raumschiffes mit Koordinaten  $(x, y, z)$  um die im Koordinatenursprung angenommene Erde (Zwei-Körper-Problem in  $\mathbb{R}^3$ ) wird durch

$$\begin{aligned} \ddot{x}(t) &= -k \cdot x(t)/r^3 \\ \ddot{y}(t) &= -k \cdot y(t)/r^3 & \text{mit } r^2 = x^2 + y^2 + z^2 \\ \ddot{z}(t) &= -k \cdot z(t)/r^3 \end{aligned}$$

mit der normalisierten Gravitationskonstanten  $k$  beschrieben. Sind Startposition bei  $t = 0$  und Endposition des Raumschiffes nach der normalisierten Zeit  $t = 2$  vorgeschrieben, d.h.

$$(x(0), y(0), z(0)) = (1.076, 0, 0), \quad (x(2), y(2), z(2)) = (0, 0.576, 0.997661),$$

so liegt ein Zweipunkt-Randwertproblem vor. Wir setzen für die 3 Ortskoordinaten  $x_1 = x$ ,  $x_2 = y$ ,  $x_3 = z$  und für die Geschwindigkeitskoordinaten  $x_4 = \dot{x}$ ,  $x_5 = \dot{y}$ ,  $x_6 = \dot{z}$  an und erhalten damit ein nichtlineares System 1. Ordnung

$$\begin{aligned} \dot{x}_1 &= x_4 & , & \quad x_1(0) - 1.076 = 0 \\ \dot{x}_2 &= x_5 & , & \quad x_2(0) = 0 \\ \dot{x}_3 &= x_6 & , & \quad x_3(0) = 0 \\ \dot{x}_4 &= -k \cdot x_1/r^3 & , & \quad x_4(2) = 0 & \text{mit } r^2 = x_1^2 + x_2^2 + x_3^2 \\ \dot{x}_5 &= -k \cdot x_2/r^3 & , & \quad x_5(2) - 0.576 = 0 \\ \dot{x}_6 &= -k \cdot x_3/r^3 & , & \quad x_6(2) - 0.997661 = 0 \end{aligned}$$

sowie Randbedingungen in Standardform. □

In den Randbedingungen des Beispiels traten stets nur Lösungen  $x(a)$  am linken Rand oder  $x(b)$  am rechten Rand auf. Allgemein bezeichnet man Bedingungen der Form

$$\begin{aligned} g_i(x(a)) &= 0, & i &= 1(1)p \\ g_i(x(b)) &= 0, & i &= p + 1(1)n \end{aligned} \quad (5)$$

als *separierte Randbedingungen*. Andernfalls spricht man von *nichtseparierten Randbedingungen*. Die explizit vorgegebenen Randwerte in Beispiel 37 bilden offenbar die einfachste, ungekoppelte Form separierter Bedingungen.

## 2.2 Mehrpunktbedingungen und Funktionalnebenbedingungen

Mitunter treten „Randbedingungen“ an inneren Intervallpunkten  $t_i$  auf. Sei eine Segmentierung  $a = t_0 < t_1 < t_2 < \dots < t_m = b$  von  $I = [a, b]$  in  $m$  Intervalle gegeben. Dann lautet das allgemeine *Mehrpunktproblem* mit  $m \geq 1$  Teilintervallen

$$\begin{aligned} \dot{x} &= f(t, x), & f &: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \\ g(x(t_0), x(t_1), \dots, x(t_m)) &= 0, & g &: \mathbb{R}^n \times \mathbb{R}^n \times \dots \times \mathbb{R}^n \rightarrow \mathbb{R}^n. \end{aligned} \quad (6)$$

Auf dem  $i$ -ten Teilintervall  $I_i = [t_{i-1}, t_i]$  führen wir die neue Zeitvariable  $s$  ein und transformieren mittels der Funktionen  $\varphi_i : I_i \rightarrow [0, 1]$  mit

$$s = \varphi_i(t) = \frac{t - t_{i-1}}{t_i - t_{i-1}}, \quad i = 1(1)m \quad (7)$$

alle Teilintervalle auf das Einheitsintervall  $[0, 1]$ . Einsetzen dieser Transformation in  $x(t)$  liefert  $m$  neue Funktionen  $z_i : [0, 1] \rightarrow \mathbb{R}^n$  mit

$$x(t) = z_i(\varphi_i(t)) \quad \text{bzw.} \quad z_i(s) = x((t_i - t_{i-1})s + t_{i-1}), \quad i = 1(1)m.$$

Wir wenden diese Transformation auf das Mehrpunktproblem (6) an und erhalten die  $n \cdot m$  Differenzialgleichungen

$$z'_i(s) = (t_i - t_{i-1}) \cdot f((t_i - t_{i-1})s + t_{i-1}, z_i(s)), \quad i = 1(1)m \quad (8)$$

zusammen mit den  $n$  Randbedingungen

$$g(z_1(0), z_1(1), z_2(1), \dots, z_m(1)) = 0. \quad (9)$$

Die fehlenden  $(n - 1) \cdot m$  Randbedingungen gewinnen wir dadurch, dass wir die Stetigkeit der Lösung  $x(t)$  an den inneren Teilpunkten fordern, d.h.

$$z_{i-1}(1) = z_i(0), \quad i = 2(1)m. \quad (10)$$

Definieren wir den zusammengesetzten Funktionenvektor („Supervektor“)  $z(s) \in \mathbb{R}^{n \cdot m}$  durch  $z(s) = (z_1^T(s), z_2^T(s), \dots, z_m^T(s))^T$ , so erhalten wir ein zu (6) äquivalentes Randwertproblem der Dimension  $n \cdot m$  auf  $[0, 1]$  in der Standardform

$$z'(s) = F(s, z(s)), \quad G(z(0), z(1)) = 0$$

mit den Funktionen  $F$  aus (8) und  $G$  aus (9) und (10). Für größere Werte  $m$  empfiehlt sich allerdings eine direkte Behandlung des Mehrpunktproblems (6) mit angepassten Methoden. Funktionalnebenbedingungen treten meist als *Integralbedingungen* auf, so dass das Randwertproblem nun die Form

$$\begin{aligned} \dot{x} &= f(t, x), & f : \mathbb{R} \times \mathbb{R}^n &\rightarrow \mathbb{R}^n \\ g_i(x(a), x(b)) &= 0, & i &= 1(1)p \\ \int_a^b g_i(t, x(t)) dt &= 0, & i &= p + 1(1)n \end{aligned} \quad (11)$$

mit  $p$  Randbedingungen und  $n - p$  Integralbedingungen hat. Wir führen die zusätzlichen Funktionen  $z_i$  mit

$$z_i(t) = \int_a^t g_{p+i}(\tau, x(\tau)) d\tau, \quad i = 1(1)n - p$$

ein und erhalten durch deren Differentiation  $n - p$  zusätzliche Differentialgleichungen mit Anfangswerten

$$\dot{z}_i(t) = g_{p+i}(t, x(t)), \quad z_i(a) = 0, \quad i = 1(1)n - p.$$

Die Integralnebenbedingungen in (11) bedeuten nunmehr  $z_i(b) = 0$ , womit sich ein erweitertes Randwertproblem

$$\begin{aligned} \dot{x}_i &= f_i(t, x) & , & \quad i = 1(1)n \\ \dot{z}_i &= g_{p+i}(t, x) & , & \quad i = 1(1)n - p \\ g_i(x(a), x(b)) &= 0 & , & \quad i = 1(1)p \\ z_i(a) &= 0 & , & \quad i = 1(1)n - p \\ z_i(b) &= 0 & , & \quad i = 1(1)n - p \end{aligned} \quad (12)$$

mit  $2n - p$  Differentialgleichungen und  $2n - p$  Randbedingungen in der Standardform ergibt.

## 2.3 Periodizitätsprobleme

In zahlreichen Anwendungen der Mechanik und Elektrotechnik treten periodische Lösungen kontinuierlicher dynamischer Systeme auf. Diese Lösungen befinden sich nicht als Ruhelagen im Gleichgewicht, sondern schwingen mit konstanter Schwingungsdauer  $T$ . Der kleinste Wert  $T > 0$  mit  $x(T) = x(0)$  heißt Periode der Bewegung.

**Periodisch erregte Systeme** Die Bestimmung periodischer Lösungen  $x(t)$  ist relativ einfach, wenn das System  $\dot{x} = f(t, x)$  periodisch erregt ist, d.h. falls eine Erregungsperiode  $T_0 > 0$  existiert, so dass

$$f(t + T_0, x) = f(t, x) \quad \forall (t, x) \in \mathbb{R} \times \mathbb{R}^n \quad (13)$$

gilt. Derartige Gleichungen beschreiben angetriebene Schwingungssysteme, deren Systemantwort auf die Erregungsschwingung von Interesse ist. Gesucht ist eine Lösung  $x(t)$  der Periode  $T = T_0$  (harmonische Lösung),  $T = k \cdot T_0$ ,  $k \in \mathbb{N}$ ,  $k > 1$ , ( $k$ -fach subharmonische Lösung) oder  $T = T_0/k$ ,  $k \in \mathbb{N}$ ,  $k > 1$ , (superharmonische Lösung). Charakteristisch ist, dass die

Periode  $T$  der gesuchten Lösung vorgegeben und damit bekannt ist.<sup>1</sup> Dann kann das Problem auf das Periodizitätsintervall  $I = [0, T]$  eingeschränkt werden. Gesucht ist eine  $\mathcal{C}^1$ -Funktion  $x : I \rightarrow \mathbb{R}^n$  mit

$$\dot{x} = f(t, x), \quad x(T) = x(0). \quad (14)$$

Mit den gekoppelten Randbedingungen  $g(x(0), x(T)) = x(T) - x(0) = 0$  liegt nun die Standardform (2) vor.

**Beispiel 2** Subharmonisch reagierende elektrische Netzwerke werden in [13] ausgiebig analysiert. Die Modellgleichung eines derartigen nichtlinearen Schwingkreises lautet

$$\ddot{x} - \varepsilon(1 - x^2 - \dot{x}^2)\dot{x} + x + b(4x^3 - 3x) = B \cdot \cos 3t,$$

wobei  $\varepsilon, b, B > 0$  konstante Parameter sind. Das System 1. Ordnung

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \varepsilon(1 - x_1^2 - x_2^2)x_2 - x_1 - b(4x_1^3 - 3x_1) + B \cdot \cos 3t \end{aligned}$$

mit Erregungsperiode  $T_0 = \frac{2\pi}{3}$  genügt der Periodizitätsbedingung (13). Zu bestimmen ist eine dreifach subharmonische Systemantwort  $x$  der Periode  $T = 2\pi$ , indem das System (14) auf  $[0, 2\pi]$  gelöst wird.  $\square$

**Autonome Systeme** Bestimmen wir periodische Lösungen autonomer Systeme

$$\dot{x} = f(x), \quad f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad (15)$$

so ist im Allgemeinen deren Periodendauer  $T$  unbekannt und muss zusammen mit der Lösung bestimmt werden. Gesucht sind damit eine  $\mathcal{C}^1$ -Funktion  $x(t)$  und ein  $T > 0$ , so dass

$$\dot{x} = f(x), \quad x(T) = x(0) \quad \text{auf} \quad I = [0, T] \quad (16)$$

erfüllt ist. Wegen der Translationsinvarianz bei autonomen Differenzialgleichungen ist die Lösung nicht eindeutig bestimmt. Wir führen deshalb eine  $(n + 1)$ -te Randbedingung ein, die die Lösung eindeutig macht und zugleich den unbekanntem Zahlenwert  $T$  festlegt. Eine derartige „Phasenbedingung“ kann z.B. den Anfangswert  $x(0)$  so festlegen, dass eine Ableitungskomponente  $\dot{x}_i(0) = 0$  wird und damit die zusätzliche Randbedingung

$$g_{n+1}(x(0)) := f_i(x(0)) = 0, \quad i \in \{1, 2, \dots, n\}$$

entsteht. Die geeignete Wahl des Index  $i$  entscheidet oft über den Erfolg des Ansatzes, weshalb in [9, 14] geeignetere, aber aufwändigere Phasenbedingungen (Projektionsbedingung, integrale Phasenbedingung) vorgestellt werden.

Den freien Randpunkt  $T$  eliminieren wir aus den Randbedingungen, indem wir eine Koordinatentransformation  $s : [0, T] \rightarrow [0, 1]$  auf das Einheitsintervall mittels  $s = s(t) = t/T$  durchführen und die Funktionen

$$y(s) = x(t(s)) = x(Ts)$$

---

<sup>1</sup> Existenzaussagen und Berechnungsmethoden sind damit einfacher als im autonomen Fall, weshalb die Transformation in ein  $(n + 1)$ -dimensionales autonomes System vermieden werden sollte.

definieren. Differentiation und Einsetzen von (16) liefert zusammen mit der Phasenbedingung das Randwertproblem

$$\begin{aligned} y'(s) &= T(s) \cdot f(y(s)) \quad , \quad y(1) - y(0) = 0 \\ T'(s) &= 0 \quad , \quad f_i(y(0)) = 0, \end{aligned} \quad (17)$$

wobei die triviale Differentialgleichung  $T'(s) = 0$  für die konstante Funktion  $T(s) = T$  hinzugefügt wurde. Damit ergibt sich auch in diesem Fall mit der Funktion  $y_{n+1}(s) = T(s)$  die Standardform

$$\begin{aligned} y'_1 &= y_{n+1} \cdot f_1(y(s)) \quad , \quad g_1 := y_1(1) - y_1(0) = 0 \\ y'_2 &= y_{n+1} \cdot f_2(y(s)) \quad , \quad g_2 := y_2(1) - y_2(0) = 0 \\ \dots & \quad \dots \quad \quad \dots \quad \dots \quad \cdot \\ y'_n &= y_{n+1} \cdot f_n(y(s)) \quad , \quad g_n := y_n(1) - y_n(0) = 0 \\ y'_{n+1} &= 0 \quad , \quad g_{n+1} := f_i(y_1(0), \dots, y_n(0)) = 0 \end{aligned} \quad (18)$$

auf  $[0, 1]$  mit der Lösung  $T^* = y_{n+1}(0)$  und  $x_i^*(t) = y_i(t/T^*)$ ,  $i = 1(1)n$ . Andere Probleme mit freiem Rand lassen sich in ähnlicher Weise auf das Einheitsintervall  $[0, 1]$  transformieren.

## 2.4 Eigenwertprobleme

Typische Randwertprobleme aus der Ingenieurspraxis enthalten in der Regel zahlreiche Parameter wie elektrische Widerstände, Kapazitäten, Induktivitäten oder aber Materialdichten, Widerstandszahlen, Elastizitätsmodule usw. Mitunter gibt man einen dieser Parameter  $\lambda$  nicht zahlenmäßig vor, sondern möchte einen kritischen Wert  $\lambda^*$  so bestimmen, dass ein bestimmtes Lösungsverhalten eintritt. Derartige Werte nennt man Eigenwerte des Randwertproblems. Das bekannteste Eigenwertproblem ist das Eulersche Knickproblem aus der Statik.

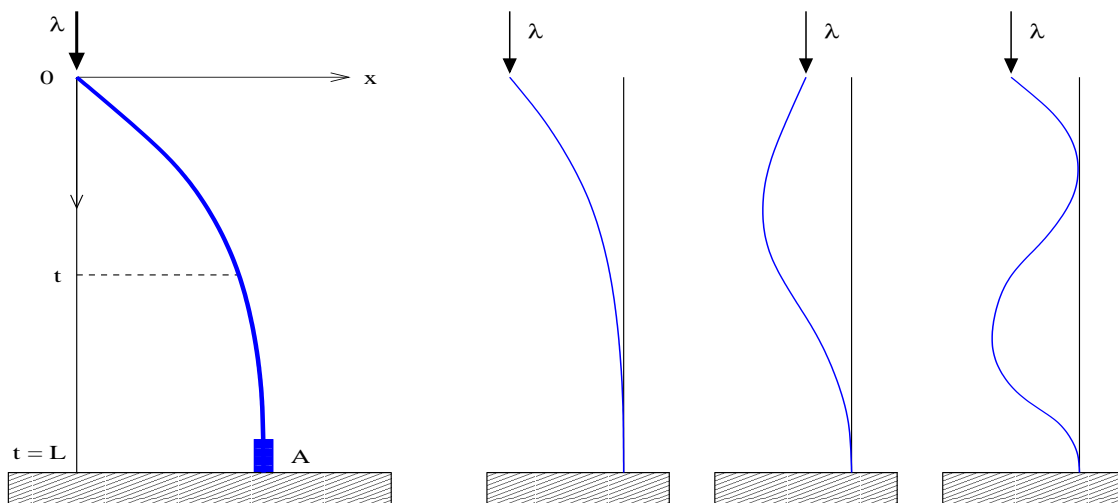


Abbildung 1: Knickstab mit  $\lambda > \lambda^*$  und Ausbiegefunktionen

**Beispiel 3** Im Punkt  $A$  sei ein homogener Stab der Länge  $L$  senkrecht eingespannt, dessen oberes Ende frei ist. Im Schwerpunkt der Endfläche wirke eine Kraft  $\lambda$  senkrecht nach unten. Überschreitet die Kraft einen kritischen Wert  $\lambda^*$ , die so genannte Knicklast, so wird das geradlinige Gleichgewicht instabil und der Stab erhält in der stabilen Gleichgewichtslage eine ausgelenkte Form wie in Abb. 1 dargestellt. Im angegebenen Koordinatensystem wollen wir annehmen, dass die Auslenkung  $x$  klein gegenüber der Stablänge  $L$  ist. Mit dem Elastizitätsmodul  $E$  und dem Flächenträgheitsmoment  $J(t)$ , das ortsabhängig sein kann, erhalten wir unter geeigneten Annahmen das Randwertproblem für die Ausbiegefunktion  $x(t)$

$$-EJ(t)\ddot{x} = \lambda x, \quad x(0) = 0, \quad \dot{x}(L) = 0.$$

Das Randwertproblem ist linear und homogen und besitzt für alle Werte von  $\lambda$  die (nicht interessierende) unausgelenkte Lösung  $x(t) \equiv 0$ . Für welche Werte  $\lambda^*$  existieren jedoch nichttriviale Lösungen  $x^*(t)$  – so genannte Eigenlösungen – wie in Abb. 1?  $\square$

Betrachten wir zwecks Beantwortung dieser Frage passend zum Standardproblem (2) das allgemeine lineare Eigenwertproblem für DGL-Systeme 1. Ordnung

$$\dot{x} - A(t, \lambda)x = 0, \quad B_a(\lambda)x(a) + B_b(\lambda)x(b) = 0 \quad (19)$$

mit der Matrixfunktion  $A$  und den Randmatrizen  $B_a$  und  $B_b$  sowie dem reellen Parameter  $\lambda \in \Lambda = [\lambda_a, \lambda_e]$ . Offenbar sind die Differentialgleichungen und ebenso die Randbedingungen linear-homogen und besitzen deshalb stets die Null-Lösung  $x(t) \equiv 0$ . Die Matrixfunktion  $A : [a, b] \times [\lambda_a, \lambda_e] \rightarrow \mathbb{R}^{n \times n}$  und die Randmatrizen  $B_a, B_b \in \mathbb{R}^{n \times n}$  können dabei in nichtlinearer Weise vom Parameter  $\lambda$  abhängen.

**Definition 4 (Eigenwert, Eigenfunktion)** Ein Wert  $\lambda^* \in \mathbb{R}$  heißt Eigenwert des Randwertproblems (19), falls zu diesem Parameterwert eine nichttriviale Lösung  $x^*(t) \neq 0$  existiert. Jede derartige Lösung heißt Eigenfunktion zu  $\lambda^*$ .

Eine nichttriviale Lösung  $x$  lässt sich erzwingen, wenn wir fordern, dass eine Norm dieser Lösung nicht verschwindet, indem wir

$$\|x\|^2 = (x, x) = \int_a^b x(\tau)^T x(\tau) d\tau = 1$$

setzen. Diese Integralnebenbedingung kann aber durch Einführung einer neuen Funktion  $\theta$  mit

$$\theta(t) = \int_a^t x(\tau)^T x(\tau) d\tau$$

wie in Problem (11) umgewandelt werden. Durch Differentiation und Einsetzen der Randpunkte ergibt sich eine zusätzliche Differentialgleichung, allerdings mit 2 Randbedingungen

$$\dot{\theta}(t) = x(t)^T x(t), \quad \theta(a) = 0, \quad \theta(b) = 1.$$

Eine jetzt fehlende  $(n + 2)$ -te Differenzialgleichung erhalten wir, indem wir die konstante Funktion  $\lambda(t) = \lambda$  einführen. Mittels der DGL  $\dot{\lambda}(t) = 0$  wird dann die Darstellung

$$\begin{aligned} \dot{x}(t) &= A(t, \lambda(t))x(t) & , & & B_a(\lambda(a))x(a) + B_b(\lambda(b))x(b) &= 0 \\ \dot{\lambda}(t) &= 0 & , & & \theta(a) &= 0 \\ \dot{\theta}(t) &= x(t)^T x(t) & , & & \theta(b) - 1 &= 0 \end{aligned} \quad (20)$$

erreicht. Mit den Funktionen  $y(t) = (x(t)^T, \lambda(t), \theta(t))^T$  ist dies aber genau die Standardform für Zweipunkt-Randwertprobleme.

**Beispiel 5** Das bezüglich  $\lambda$  nichtlineare Eigenwertproblem aus [6]

$$\ddot{x} + g(t, \lambda)x = 0, \quad x(0) = 0, \quad \dot{x}(1) = f(\lambda)x(1), \quad t \in [0, 1]$$

mit den Funktionen

$$g(t, \lambda) = \frac{1}{\lambda}(t + 10) - \lambda, \quad f(\lambda) = -\lambda$$

besitzt reelle Eigenwerte  $\lambda_1 > \lambda_2 > \lambda_3 > \dots > 0$ , die sich bei 0 häufen. Mit den Funktionen  $\lambda(t)$  und  $\theta(t)$  überführen wir das Problem in die Form (20) für den Vektor  $(x_1, x_2, x_3, x_4)^T$

$$\begin{aligned} \dot{x}_1 &= x_2 & , & & x_1(0) &= 0 \\ \dot{x}_2 &= -x_1[(t + 10)/x_3 - x_3] & , & & x_2(1) + x_1(1)x_3(1) &= 0 \\ \dot{x}_3 &= 0 & , & & x_4(0) &= 0 \\ \dot{x}_4 &= x_1^2 + x_2^2 & , & & x_4(1) - 1 &= 0 \end{aligned} \quad (21)$$

Die konstante Funktion  $x_3(t) = \lambda$  liefert bei Vorgabe geeigneter Startwerte die gesuchten Eigenwerte und  $x_1(t) = x(t)$  die zugehörigen normierten Eigenfunktionen.  $\square$

Zahlreiche verwandte Aufgabenklassen wie Verzweigungsprobleme, Variationsprobleme und Optimalsteuerprobleme können – allerdings oft mit beträchtlichem Aufwand – ebenfalls auf Randwertprobleme (2) transformiert werden. Wichtig ist dabei der Erhalt der Lösungsmenge: Jede Lösung des Ausgangsproblems muss auch eine Lösung der erhaltenen Aufgabe (2) sein! Unter geeigneten Problemvoraussetzungen lässt sich in den skizzierten Fällen sogar die Lösungsäquivalenz nachweisen, d.h. die Standardform (2) besitzt keine neu hinzukommenden „Geisterlösungen“.

Den Versuch einer Problemtransformation sollte man stets unternehmen, weil man damit eine Neuentwicklung spezieller numerischer Verfahren vermeidet. Es ist oft ökonomischer, nach einer Transformation ein stabiles, ausgereiftes Standardverfahren auf ein geringfügig erweitertes Randwertproblem anzuwenden, als selbst ein Verfahren für eine sehr spezielle Problemklasse erfinden und erproben zu müssen.

## 2.5 Problemvoraussetzungen

Um ein numerisches Verfahren sinnvoll auf das Randwertproblem (2) anwenden zu können, muss die Existenz einer  $\mathcal{C}^1$ -Lösung  $x : I \rightarrow \mathbb{R}^n$  garantiert sein. Betrachten wir der Einfachheit

halber das allgemeine *lineare Randwertproblem*

$$\dot{x} = A(t)x + r(t), \quad B_a x(a) + B_b x(b) = c, \quad t \in I = [a, b] \quad (22)$$

mit der Matrixfunktion  $A : I \rightarrow \mathbb{R}^{n \times n}$ , der Vektorfunktion  $r : I \rightarrow \mathbb{R}^n$ , den konstanten Randmatrizen  $B_a, B_b \in \mathbb{R}^{n \times n}$  und dem Randvektor  $c \in \mathbb{R}^n$ . Aus der qualitativen Theorie ist bekannt, dass das zugehörige lineare Anfangswertproblem eine eindeutige Lösung auf ganz  $I$  besitzt, falls nur die Funktionen  $A$  und  $r$  stetig auf  $I$  sind. Auch für nichtlineare Anfangswertprobleme garantiert der Existenz- und Eindeutigkeitssatz von Picard und Lindelöf zumindest in einer Umgebung des Anfangspunktes die eindeutige Lösbarkeit. Bei Randwertproblemen ist der Sachverhalt dagegen ungleich komplizierter, denn hier entscheiden die Randbedingungen ebenfalls über die Lösungsmenge.

**Beispiel 6** Das lineare System  $\dot{x}_1 = x_2$ ,  $\dot{x}_2 = -x_1$  hat die allgemeine Lösung  $x_1(t) = C_1 \sin t + C_2 \cos t$ ,  $x_2(t) = C_1 \cos t - C_2 \sin t$ . Durch Ausrechnen bestätigt man leicht folgende Erkenntnis: Mit den Randbedingungen

- $x_1(0) = 0, \quad x_1(\pi/2) = 1$  existiert die eindeutige Lösung  $x_1(t) = \sin t$ ,
- $x_1(0) = 0, \quad x_1(\pi) = 0$  existieren unendlich viele Lösungen  $x_1(t) = C_1 \sin t$ ,
- $x_1(0) = 0, \quad x_1(\pi) = 1$  existiert keine Lösung. □

Im Falle allgemeiner linearer Systeme (22) oder nichtlinearer Systeme (2) sind keine leicht überprüfbareren Existenz- und Eindeutigkeitssätze bekannt. Andererseits schränken spezielle Zusatzannahmen die Problemklasse oft stark ein. Typische Sätze mit hinreichenden Bedingungen findet man in [7, 4]. Wir wollen deshalb die Problemklasse nicht unnötig einengen und treffen deshalb folgende Standardannahme:

### Voraussetzung 7 (Glattheit und Existenz)

$f$  und  $g$  sind hinreichend glatt<sup>2</sup> auf  $I \times \mathbb{R}^n$  und Problem (2) besitzt eine  $\mathcal{C}^1$ -Lösung  $x(t)$  auf  $I$ .

Führt man eine der behandelten Problemtransformationen durch, so sollte man prüfen, welche Voraussetzungen an das Originalproblem zu stellen sind, um das Erfülltsein der Standardannahme 7 für das Randwertproblem (2) zu garantieren.

## 3 Schießverfahren und Mehrzielmethode

Die Grundidee dieses aus der Ballistik entlehnten Zuganges (deshalb *Schießverfahren*, *ballistisches Verfahren*) besteht in der Zurückführung des Randwertproblems auf eine Folge von Anfangswertproblemen. Diese können mit den leistungsfähigen Anfangswertlösern von MATLAB [11] behandelt werden (vgl. dazu [17]).

<sup>2</sup> Alle partiellen Ableitungen bis zur  $s$ -ten Ableitung existieren und sind stetig, wobei  $s \geq 1$  ist.

### 3.1 Einfaches Schießverfahren

Wir betrachten zum gegebenen Randwertproblem das assoziierte Anfangswertproblem

$$\dot{x} = f(t, x), \quad x(a) = s \quad \text{mit} \quad s \in \mathbb{R}^n, \quad t \in I = [a, b]. \quad (23)$$

Angenommen, diese Aufgabe besitzt zu jedem  $s \in D$  eine auf  $[a, b]$  eindeutige Lösung  $x(t, s)$ . Die geforderten Randbedingungen  $g(s, x(b, s)) = 0$  werden damit im Allgemeinen nicht zu erfüllen sein. Definieren wir deshalb eine Abbildung  $\varphi : D \rightarrow D$  mit

$$\varphi(s) := g(s, x(b, s)), \quad (24)$$

so löst  $x(t, s^*)$  mit  $s^* \in D$  genau dann das Randwertproblem (2), wenn der  $n$ -dimensionale Vektor  $s = s^*$  eine Lösung des Gleichungssystems

$$\varphi(s) = 0, \quad s \in D \subset \mathbb{R}^n \quad (25)$$

ist. Damit wird das Ausgangsproblem formal<sup>3</sup> auf ein Nullstellenproblem in  $\mathbb{R}^n$  reduziert, wofür leistungsfähige Verfahren in [5, Kapitel 21] verfügbar sind. Um den Integrationsaufwand niedrig zu halten, ist ein überlinear konvergentes Verfahren erforderlich. Das Newton-Verfahren lautet in der Standardform

$$s_{k+1} = s_k - [\varphi'(s_k)]^{-1} \varphi(s_k), \quad k = 0, 1, 2, \dots, \quad s_0 \in D. \quad (26)$$

Mit (24) erhalten wir durch Differentiation die Newton-Matrix

$$\varphi'(s) = B_a(s) + B_b(s)X(b, s),$$

worin die partiellen Ableitungen der Randbedingungen (die Randmatrizen) durch

$$B_a(s) := \left. \frac{\partial g(v, w)}{\partial v} \right|_{v=s, w=x(b,s)} \quad \text{und} \quad B_b(s) := \left. \frac{\partial g(v, w)}{\partial w} \right|_{v=s, w=x(b,s)}$$

und die Fundamentallösung durch

$$X(t, s) := \frac{\partial x(t, s)}{\partial s}$$

definiert sind. Die aufwändige Invertierung der Newton-Matrix vermeiden wir durch Umstellung der Verfahrensgleichung (26) nach der Newton-Korrektur  $d_k = s_{k+1} - s_k$ , womit die praktikablere Form des Newton-Verfahrens entsteht:

Iteriere für  $k = 0, 1, 2, \dots$ , beginnend mit  $s_0$ :

- (1) Berechne die Matrix  $\varphi'(s_k) = B_a(s_k) + B_b(s_k)X(b, s_k)$ .
- (2) Löse das lineare Gleichungssystem

$$\varphi'(s_k) d_k = -\varphi(s_k). \quad (27)$$

- (3) Korrigiere  $s_{k+1} = s_k + d_k$ .

<sup>3</sup> Wir setzen vorerst voraus, dass die Anfangswertprobleme (23) exakt gelöst werden können.

Um aus diesem Ansatz einen Algorithmus zu entwickeln, sind folgende Fragen zu klären:

1. Wie kann die Matrixfunktion  $X(t, s)$  ermittelt werden?
2. Unter welchen Bedingungen an das Ausgangsproblem kann die lokale Konvergenz des Newton-Verfahrens garantiert werden?
3. Welche Auswirkungen hat die genäherte Bestimmung von  $x(t, s)$  mit einem Diskretisierungsverfahren auf die Lösung des Randwertproblems?

Beantworten wir zuerst die Frage 1, indem wir eine Lösung  $x(t, s)$  des Anfangswertproblems (23) in die DGL einsetzen

$$\frac{\partial x}{\partial t}(t, s) = f(t, x(t, s)), \quad x(a, s) = s.$$

Differenzieren wir diese Gleichungen nach den Anfangswerten  $s$ , so erhalten wir mit der Glattheitsvoraussetzung ein Anfangswertproblem für die Matrixfunktion  $X(t, s) = \frac{\partial x}{\partial s}(t, s)$

$$\frac{\partial}{\partial t} \left( \frac{\partial x}{\partial s}(t, s) \right) = \frac{\partial f}{\partial x}(t, x(t, s)) \cdot \frac{\partial x}{\partial s}(t, s), \quad \frac{\partial x}{\partial s}(a, s) = I. \quad (28)$$

Darin ist  $I$  die  $n$ -reihige Einheitsmatrix. Folgende Begriffe sind deshalb wesentlich:

**Definition 8 (Variationssystem, Hauptfundamentalmatrix)**

$x(t, s)$  sei die Lösung des Anfangswertproblems (23) mit  $s \in D$ .

(i) Die Matrixfunktion  $A : I \times D \rightarrow \mathbb{R}^{n \times n}$  mit

$$A(t, s) := \frac{\partial f}{\partial x}(t, x(t, s)) = f_x(t, x(t, s))$$

heißt zugehörige Jacobi-Matrix.

(ii) Die Matrix-DGL für  $X : I \times D \rightarrow \mathbb{R}^{n \times n}$  mit

$$\dot{X} = A(t, s) \cdot X, \quad \dot{X} = \frac{\partial}{\partial t} X(t, s) \quad (29)$$

heißt zugehöriges Variationssystem.

(iii) Die bei  $t = a$  normierte Matrixfunktion  $X : I \times D \rightarrow \mathbb{R}^{n \times n}$  mit

$$\dot{X} = A(t, s) \cdot X, \quad X(a, s) = I \quad (I \text{ Einheitsmatrix}) \quad (30)$$

heißt Hauptfundamentalmatrix.

Die Matrixfunktion  $X(t, s)$  kann also durch eine simultan ausführbare Integration des linearen Anfangswertproblems (30) gewonnen werden. Im Vergleich von (26) mit (30) erhalten wir am Intervallende  $b$  die gesuchte Matrix  $X(b, s)$  als Hauptfundamentalmatrix. Damit lässt sich Algorithmus 9 des *einfachen Schießverfahrens* (engl.: *simple shooting*) formulieren, der zu vorgegebener Startnäherung  $s_0 \in \mathbb{R}^n$  und Toleranz  $tol$  eine Lösung des Randwertproblems liefert. Mit dem erhaltenen Startvektor  $s^*$  rekonstruiert man durch Integration des Anfangswertproblems in Schritt 2 die gesuchte Lösung  $x^*(t) = x(t, s^*)$  des Randwertproblems mit  $x^*(a) = s^*$ .

Wir wollen nun Frage 2 beantworten. Aus [5, Kapitel 21] ist bekannt, dass das Gleichungssystem  $\varphi(s) = 0$  mit dem Newton-Verfahren lokal eindeutig lösbar ist, wenn die Lösung  $s^*$  regulär (isoliert) ist. Den entsprechenden Begriff für die Lösung  $x^*(t)$  des Randwertproblems liefert

**Algorithmus 9 (Einfaches Schießverfahren)**Function  $[s^*] = \text{shooting}(f, g, f_x, B_a, B_b, a, b, s_0, \text{tol}, \text{kmax})$ 1. Für  $k = 0, 1, \dots, \text{kmax}$  iteriere:1.1. (Anfangswertprobleme) Löse die  $n + 1$  Systeme für  $t \in [a, b]$ :

$$\begin{aligned} \dot{x} &= f(t, x) & , & \quad x(a, s_k) = s_k \\ \dot{X} &= f_x(t, x(t, s_k)) \cdot X & , & \quad X(a, s_k) = I. \end{aligned}$$

Ergebnis:  $x(b, s_k) \in \mathbb{R}^n$ ,  $X(b, s_k) \in \mathbb{R}^{n \times n}$ 

1.2. (Nullstellenaufgabe) Bilde die Funktionen

$$\begin{aligned} \varphi(s_k) &:= g(s_k, x(b, s_k)) \\ \varphi'(s_k) &:= B_a(s_k) + B_b(s_k)X(b, s_k). \end{aligned}$$

1.3. (Newton-Schritt) Löse das lineare System

$$\varphi'(s_k) \cdot d_k = -\varphi(s_k).$$

1.4. Falls  $\|d_k\| < \text{tol} \cdot (1 + \|\varphi(s_k)\|)$ , so gehe zu Schritt 2.1.5. (Newton-Korrektur) Aktualisiere  $s_{k+1} := s_k + d_k$ .2. Setze  $s^* := s_k$  und löse das Anfangswertproblem

$$\dot{x} = f(t, x), \quad x(a) = s^*.$$

3. Return  $s^*$ 

**Definition 10 (Reguläre Lösung)** Voraussetzung 7 sei erfüllt. Die Lösung  $x^*(t)$  heißt regulär (auch: isoliert), wenn das assoziierte linear-homogene Variationsproblem

$$\dot{z} = f_x(t, x^*(t))z, \quad B_a z(a) + B_b z(b) = 0 \quad (31)$$

zum Randwertproblem nur die Lösung  $z(t) \equiv 0$  besitzt. Die Randmatrizen sind durch

$$B_a := \left. \frac{\partial g(v, w)}{\partial v} \right|_{v=x^*(a), w=x^*(b)} \quad \text{und} \quad B_b := \left. \frac{\partial g(v, w)}{\partial w} \right|_{v=x^*(a), w=x^*(b)}$$

definiert.

Problem (31) entsteht durch Linearisierung der Differenzialgleichung und der Randbedingungen an der Lösung  $x^*(t)$ . Die Eindeutigkeit der Null-Lösung  $z$  bedeutet, dass die Linearisierung nicht singular ist, womit numerische Verfahren wie das Newton-Verfahren anwendbar werden. Es lässt sich nachweisen: Wenn  $x^*(t)$  eine reguläre Lösung ist, so ist sie auch lokal eindeutig. Eine Lösung  $x^*(t)$  heißt lokal eindeutig (auch: geometrisch isoliert), wenn eine Konstante  $r > 0$  existiert, so dass in dem Schlauch um die Lösung  $\mathcal{U}[x^*; r] := \{x \in C(I) \mid \sup_{t \in I} \|x(t) - x^*(t)\| \leq r\}$  keine von  $x^*(t)$  verschiedene Lösung des Randwertproblems existiert.<sup>4</sup> Für derartige Lösungen findet man in [8, S. 9] den

<sup>4</sup> Der Begriff „Isoliertheit“ wird wegen möglicher Fehlinterpretationen nachfolgend nicht benutzt.

**Algorithmus 11 (Schießverfahren mit Differenzen)**Function  $[s^*] = \text{difference\_shooting}(f, g, a, b, s_0, \text{tol}, kmax)$ 1. Für  $k = 0, 1, \dots, kmax$  iteriere:1.1. Wähle  $\delta_j = \sqrt{\varepsilon_M}(1 + |s_{kj}|)$ ,  $j = 1(1)n$ .1.2. (Anfangswertprobleme) Löse mit  $\sigma_0 := s_k$  und  $\sigma_j := s_k + \delta_j \cdot e_j$ ,  $j = 1(1)n$  die  $n + 1$  Systeme für  $t \in [a, b]$ :

$$\dot{x} = f(t, x), \quad x(a, \sigma_j) = \sigma_j, \quad j = 0(1)n.$$

Ergebnis:  $x(b, \sigma_j) \in \mathbb{R}^n$ ,  $j = 0(1)n$ 

1.3. (Nullstellenaufgabe) Bilde die Funktionen

$$\varphi(\sigma_j) := g(\sigma_j, x(b, \sigma_j)), \quad j = 0(1)n \quad \text{und damit}$$

$$\varphi(s_k) := \varphi(\sigma_0) \quad \text{ sowie die Differenzenquotienten}$$

$$\varphi_j := \frac{1}{\delta_j} [\varphi(\sigma_j) - \varphi(\sigma_0)], \quad j = 1(1)n,$$

$$\Phi_k := (\varphi_1, \varphi_2, \dots, \varphi_n).$$

1.4. (Newton-Schritt) Löse das lineare Gleichungssystem

$$\Phi_k d_k = -\varphi(s_k).$$

1.5. Falls  $\|d_k\| < \text{tol} \cdot (1 + \|\varphi(s_k)\|)$ , so gehe zu Schritt 2.1.6. (Newton-Korrektur) Aktualisiere  $s_{k+1} := s_k + d_k$ .2. Setze  $s^* := s_k$  und löse das Anfangswertproblem

$$\dot{x} = f(t, x), \quad x(a) = s^*.$$

3. Return  $s^*$ 

**Satz 12** Die reguläre Lösung  $x^*(t)$  erfülle die Voraussetzungen 7 mit  $f \in \mathcal{C}^2(I \times D)$  und der Lipschitz-Konstanten  $L := \sup_{I \times D} |f_x(t, x)|$ . Dann existiert eine Umgebung

$$\mathcal{K}[s^*; \varrho] = \{s \in D \mid \|s - s^*\| \leq \varrho\}, \quad \text{mit } \varrho := re^{-L(b-a)}$$

des Anfangswertes  $s^* = x^*(a)$  mit folgenden Eigenschaften:

(i) Das Anfangswertproblem (23) besitzt zu jedem  $s \in \mathcal{K}[s^*; \varrho]$  eine eindeutige Lösung  $x(t, s) \in \mathcal{C}^1(I)$ .

(ii) Das normierte Fundamentalsystem  $X(t, s)$  existiert als Lösung von (30) und genügt der Identität

$$X(t, s) = \frac{\partial x}{\partial s}(t, s), \quad (t, s) \in I \times \mathcal{K}[s^*; \varrho].$$

(iii)  $x^*(t) = x(t, s^*)$  ist genau dann eine reguläre Lösung des Randwertproblems, wenn  $s^* = x^*(a)$  eine reguläre Lösung des Gleichungssystems  $\varphi(s) = 0$  ist.

Dieser Satz garantiert mit den Aussagen (i) und (ii) die *Durchführbarkeit des Schießverfahrens* 9, wenn nur die Startnäherung  $s_0$  hinreichend nahe der Lösung  $s^*$  gewählt wird. Die *lokale quadratische Konvergenz* der Startwerte  $s_k \rightarrow s^*$  wird wegen Aussage (iii) auf Grundlage des Satzes 21.22 aus [5] gesichert.

Kritisch zu vermerken ist, dass die Bestimmung der Hauptfundamentalmatrix  $X(b, s)$  mittels des Variationssystems aufwändig ist und die Jacobi-Matrizen  $f_x(t, x)$  sowie  $B_a(s)$  und  $B_b(s)$  in der Praxis oft nicht verfügbar sind. Deshalb sind geeignete Differenzen-Approximationen der Newton-Matrix  $\varphi'(s)$  sinnvoll, da sie das Variationssystem vollkommen vermeiden. In Algorithmus 11 wird  $\varphi'(s)$  spaltenweise aufgebaut. Der Startwert  $s_k$  des  $k$ -ten Newton-Schrittes wird dazu mittels eines Schrittweitenvektors  $\delta = (\delta_1, \delta_2, \dots, \delta_n)^\top$ ,  $\delta_j > 0$ , gestört. Sei  $e_j$  der  $j$ -te Einheitsvektor in  $\mathbb{R}^n$ . Damit werden die Vektoren

$$\begin{aligned}\sigma_0 &:= s_k, \\ \sigma_j &:= s_k + \delta_j \cdot e_j, \quad j = 1(1)n\end{aligned}$$

definiert und als Startwerte genutzt.  $\varepsilon_M$  ist die Maschinengenauigkeit. Zu Startnäherung  $s_0 \in \mathbb{R}^n$  und Toleranz  $tol$  liefert das linear konvergente Verfahren eine Lösung  $x^*$  des Randwertproblems.

**Beispiel 13** 1. Wir wenden das Schießverfahren mit Differenzen 11 auf die Gleichung 2. Ordnung

$$\ddot{x}(t) = \frac{3}{2}x(t)^2, \quad x(0) = 4, \quad x(1) = 1 \quad (32)$$

aus [16, S. 185] an. Als Löser wird das DOPRI-Verfahren `ode45` der MATLAB-Bibliothek genutzt, das bei einer Genauigkeit  $tol = 10^{-6}$  zuverlässige Resultate lieferte. Mit den Startlösungen  $s_0 = (4, -1)^T$  und  $s_0 = (4, -10)^T$  ergeben sich die Iterierten der Abbildungen 2 (a) und (b), wogegen mit dem Startvektor  $s_0 = (4, -20)^T$  eine Konvergenz gegen eine zweite Lösung in Abb. 2 (c) erfolgte. Beide Lösungen sind zusammen in Abb. 2 (d) dargestellt.

$tol = 10^{-6}$				$tol = 10^{-12}$		
$\lambda_{start}$	$\lambda_i$	$k$	$N$	$\lambda_i$	$k$	$N$
1.60	1.634 939e-0	4	53	1.634 939 335 260e-0	6	49
0.40	4.472 961e-1	5	93	4.472 961 222 282e-1	6	93
0.16	1.689 512e-1	5	145	1.689 512 535 492e-1	6	145
0.08	8.668 067e-2	5	197	8.668 066 710 515e-2	6	205
0.04	2.517 402e-2	5	361	2.517 401 993 301e-2	6	373

Tabelle 1: Ausgewählte Eigenwerte des Problems (40)

2. Das bezüglich  $\lambda$  nichtlineare Eigenwertproblem aus Beispiel 5

$$\ddot{x} + g(t, \lambda)x = 0, \quad x(0) = 0, \quad \dot{x}(1) = f(\lambda)x(1), \quad t \in [0, 1] \quad (33)$$

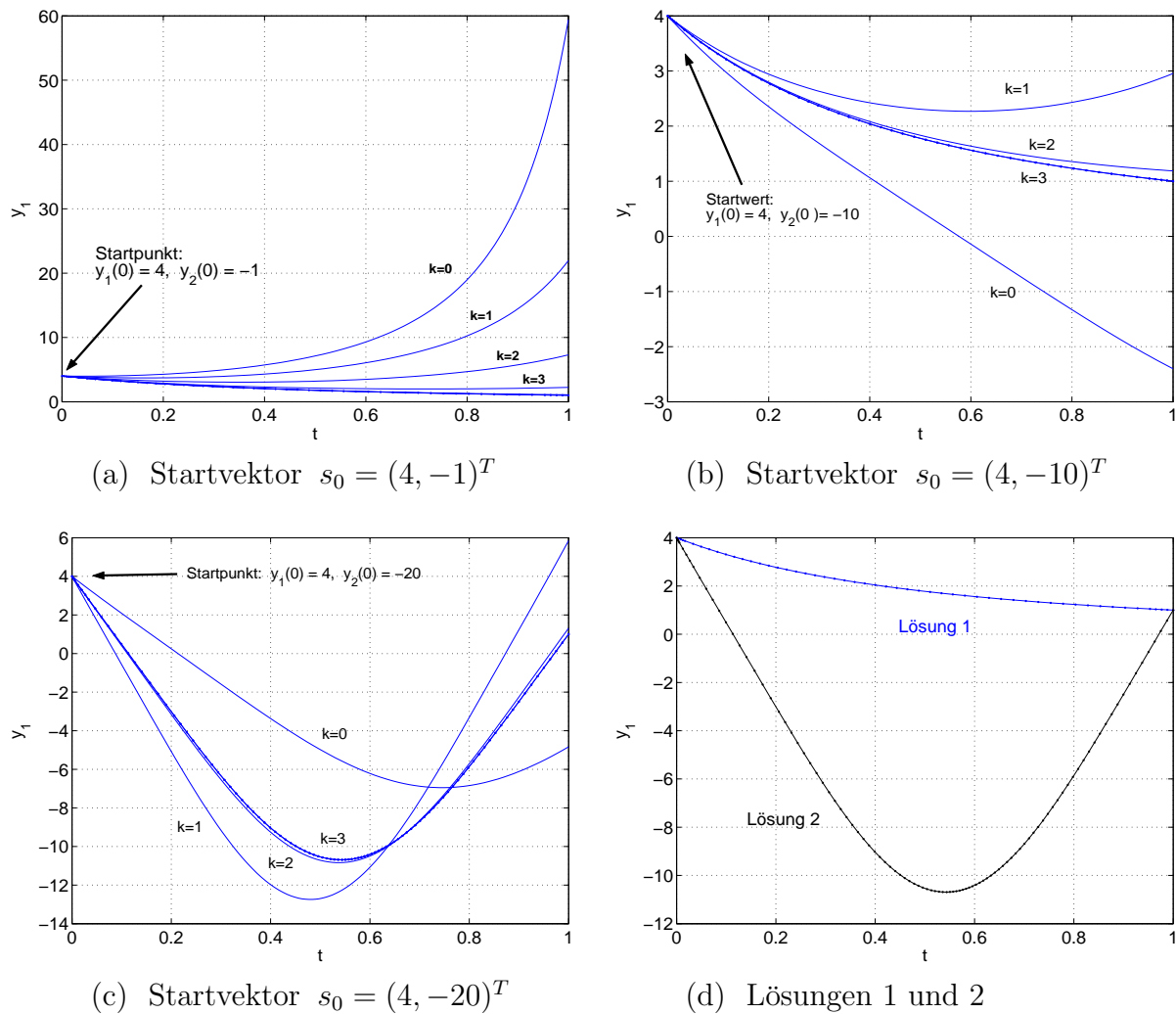


Abbildung 2: Randwertproblem 2. Ordnung (32)

mit den Funktionen

$$g(t, \lambda) = \frac{1}{\lambda}(t + 10) - \lambda, \quad f(\lambda) = -\lambda$$

wurde mit den Hilfsfunktionen  $\lambda(t)$  und  $\theta(t)$  in die Standardform

$$\begin{aligned} \dot{x}_1 &= x_2 & , & & x_1(0) &= 0 \\ \dot{x}_2 &= -x_1[(t+10)/x_3 - x_3] & , & & x_2(1) + x_1(1)x_3(1) &= 0 \\ \dot{x}_3 &= 0 & , & & x_4(0) &= 0 \\ \dot{x}_4 &= x_1^2 + x_2^2 & , & & x_4(1) - 1 &= 0 \end{aligned}$$

für den Vektor  $(x_1, x_2, x_3, x_4)^T$  überführt. Mit dem Löser `ode45` und den Genauigkeiten  $tol = 10^{-6}$  bzw.  $tol = 10^{-12}$  ergaben sich bei geeigneten Startwerten  $s_0 = (0, 1, \lambda_{start}, 1)^T$  die Eigenwertnäherungen  $\lambda = x_3(t)$  der Tabelle 1. Die Spalte  $k$  gibt die Zahl der ausgeführten Newton-Schritte an, während  $N$  die Zahl der Integrationsschritte von `ode45` bei der Integration der erhaltenen Lösung  $x^*(t)$  bezeichnet.  $\square$

Um die Frage 3 zu beantworten, nehmen wir nun an, dass alle  $n + 1$  in Algorithmus 9 auftretenden Anfangswertprobleme mit ein- und demselben numerischen Integrationsverfahren (V) auf demselben Gitter gelöst werden. Unter welchen Voraussetzungen kann dann auch die Konvergenz der Näherungslösungen für das Randwertproblem garantiert werden? Ein grundlegendes Theorem von H. B. Keller [8] basiert auf der Regularitätsforderung 10 und liefert zusammen mit weiteren Annahmen den folgenden

**Satz 14** Die hinreichend glatte Lösung  $x^*(t)$  nach Voraussetzung 7 sei regulär. Das Diskretisierungsverfahren (V) für die Anfangswertprobleme sei nullstabil und konsistent mit Ordnung  $\mathcal{O}(h^p)$ ,  $p \in \mathbb{N}$ . Dann existiert ein  $h_0 > 0$ , so dass für alle Schrittweiten  $h \in (0, h_0]$  gilt:

- (i) Die Gleichung  $\varphi(s) = 0$  besitzt zu jedem  $h$  in einer Kugelumgebung  $\mathcal{K}[s^*; \varrho]$  eine eindeutige Lösung  $s^*(h)$  mit  $\|s^*(h) - s^*\| \leq C_0 h^p$ ,  $C_0 > 0$ , const.
- (ii) Das Verfahren (V), angewendet auf das Anfangswertproblem

$$\dot{x} = f(t, x), \quad x(a) = s^*(h), \quad t \in [a, b],$$

ist ausführbar und liefert die diskrete Lösung  $(u_j(h))$ ,  $j = 0(1)N$ . Diese konvergiert gegen  $x^*(t)$  mit Ordnung  $p$ , d.h.  $\|u_j(h) - x^*(t_j)\| \leq Ch^p$ ,  $j = 0(1)N$ , mit der von  $h$  unabhängigen Konstanten  $C > 0$ .

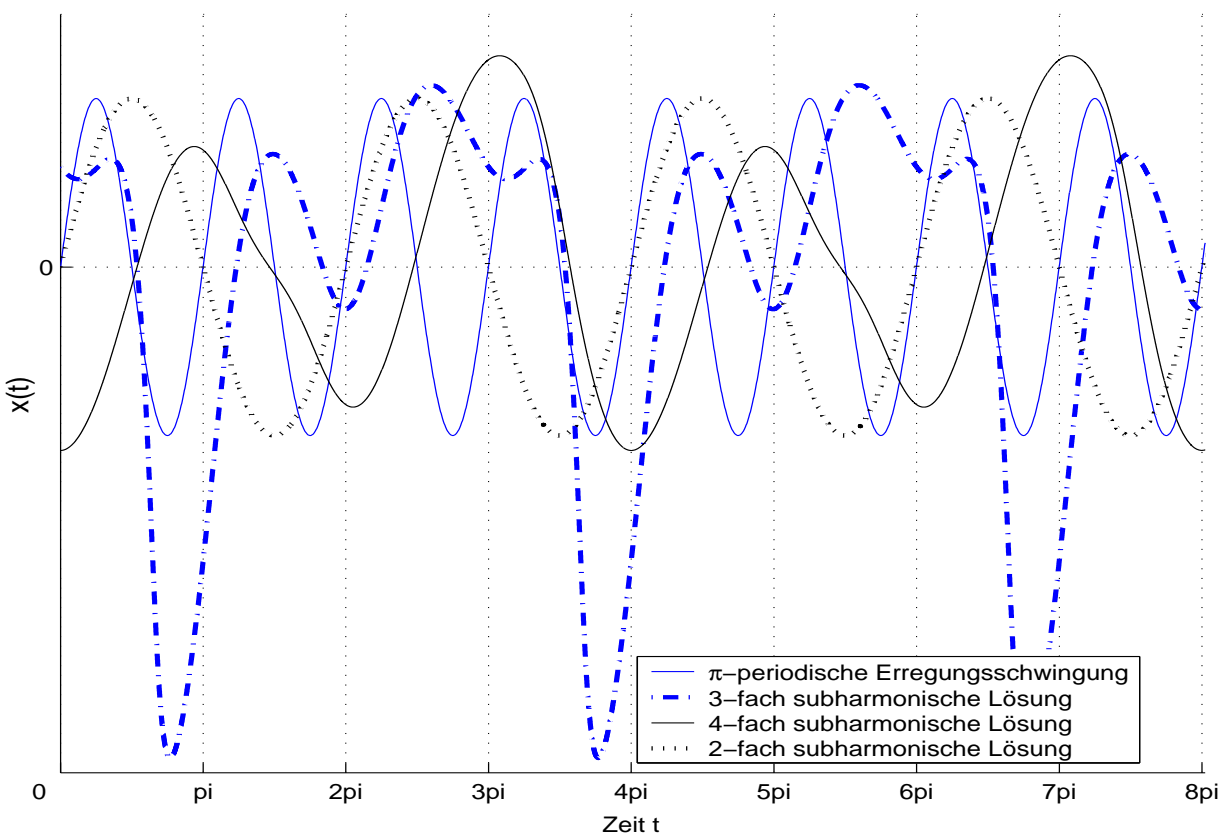


Abbildung 3: Zweifach subharmonisch reagierendes System – Erregungsschwingung und subharmonische Ausgangsschwingungen

**Beispiel 15** Das folgende zweifach subharmonisch reagierende System beschreibt ein elektrisches Netzwerk, das bei harmonischer Erregung als Systemantwort eine exakt subharmonische Schwingung erzeugt. Die Modellgleichung von E. Philippow [13] lautet

$$\ddot{x} - \varepsilon(1 - x^2 - \dot{x}^2)\dot{x} + 2bx\dot{x} + x = \hat{B} \sin 2t$$

mit den reellen Parametern  $b = \hat{B} = 1$  sowie dem freien Parameter  $\varepsilon \in \mathbb{R}$ . Transformation in ein System erster Ordnung liefert das nichtautonome Schwingungsproblem

$$\begin{aligned} \dot{x}_1 &= x_2, & x_1(0) - x_1(T) &= 0 \\ \dot{x}_2 &= \varepsilon(1 - x_1^2 - x_2^2)x_2 - 2bx_1x_2 - x_1 + \hat{B} \sin 2t, & x_2(0) - x_2(T) &= 0 \end{aligned} \quad (34)$$

mit Erregungsperiode  $T_0 = \pi$  und vorgegebener Lösungsperiode  $T = 2\pi$ . Untersucht man das Lösungsverhalten für den festen Parameterwert  $\varepsilon_0 = 0.4$ , so findet man mit Algorithmus 11 leicht einen periodischen Orbit der gewünschten Periodendauer  $2\pi$ . Darüber hinaus existieren aber auch dreifach bzw. vierfach subharmonische Lösungen. Abb. 3 zeigt alle diese Schwingungen.

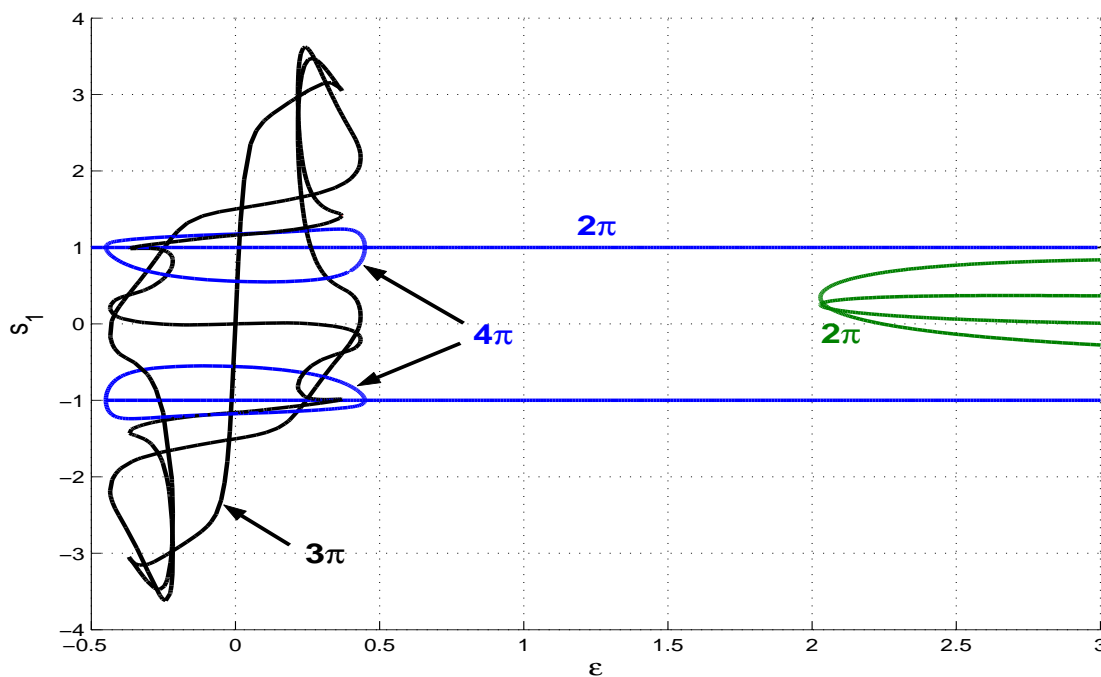


Abbildung 4:  $(\varepsilon, s_1)$ -Diagramm der periodischen Lösungen des Systems (34)

Wenden wir uns den  $3\pi$ -periodischen Lösungen zu, deren Anfangswerte  $s_1 = x_1(0)$  in Abhängigkeit vom Parameter  $\varepsilon$  in Abb. 4 durch eine geschlossene schwarze Kurve dargestellt ist. Der Versuch einer Lösungsfortsetzung nach  $\varepsilon$ , ausgehend von  $\varepsilon = 0.36$  und dem Startpunkt  $s_0 = x(0) = (0.38697629, 0.09266292)^T$ , scheitert mit dem Einfach-Schießverfahren schon nach wenigen Fortsetzungsschritten. Grund hierfür ist die „starke“ Instabilität der periodischen Lösungen, die zum völligen Versagen jedes numerischen Integrationsverfahrens bereits vor Erreichen des Intervallendes  $T = 3\pi$  führt.  $\square$

Das Stabilitätsverhalten der  $T$ -periodischen Lösung  $x(t, s^*)$  des Beispiels kann in Algorithmus 9 leicht durch die abschließende Berechnung der Hauptfundamentalmatrix  $X(t; s^*)$

in Schritt 2 analysiert werden. Mit den Eigenwerten  $m_1, m_2, \dots, m_n$  der *Monodromiematrix*  $M = X(T; s^*)$ , den so genannten charakteristischen Multiplikatoren gilt (vgl. [12])

**Satz 16 (Stabilitätskriterium bei periodisch erregten Systemen)**  $x(t, s^*)$  sei eine Lösung der Periode  $T > 0$  mit charakteristischen Multiplikatoren  $m_i, i = 1(1)n$ .

- (i) Ist  $|m_i| < 1$  für alle  $i \in \{1, \dots, n\}$ , so ist  $x(t, s^*)$  asymptotisch Ljapunov-stabil.
- (ii) Ist  $|m_i| > 1$  für ein  $i \in \{1, \dots, n\}$ , so ist  $x(t, s^*)$  Ljapunov-instabil.

### 3.2 Mehrfach-Schießverfahren

Die Durchführbarkeit des einfachen Schießverfahrens ist nach Satz 12 oft nur in unmittelbarer Lösungsnähe gesichert, denn die Lösungen der Anfangswertprobleme müssen auf dem ganzen Intervall  $I = [a, b]$  existieren. Der Konvergenzbereich des Schießverfahrens kann deshalb bei großer Lipschitz-Konstante  $L$  und langem Intervall  $I$  wegen  $\varrho = re^{-L(b-a)}$  aus Satz 12 extrem klein werden. Im *Mehrfach-Schießverfahren* (*Mehrzielmethode*, engl.: *multiple shooting, parallel shooting*) versuchen wir, diesem Problem mit einer Segmentierung des Lösungsintervalls entgegenzuwirken. Auf  $I = [a, b]$  legen wir ein Gitter von so genannten Schießpunkten

$$a = t_0 < t_1 < t_2 < \dots < t_m = b$$

mit  $m \geq 1$  Teilintervallen  $I_k = [t_k, t_{k+1}]$  fest.  $x(t) = x(t; t_k, s_k)$  bezeichne die auf  $I_k$  definierte Lösung des Anfangswertproblems

$$\dot{x} = f(t, x), \quad x(t_k) = s_k, \quad t_k \leq t \leq t_{k+1}, \quad k = 0(1)m - 1. \quad (35)$$

Abb. 5 zeigt diese Lösungen zu vorgegebenen Anfangswerten  $s_k$ . Mit der Abkürzung

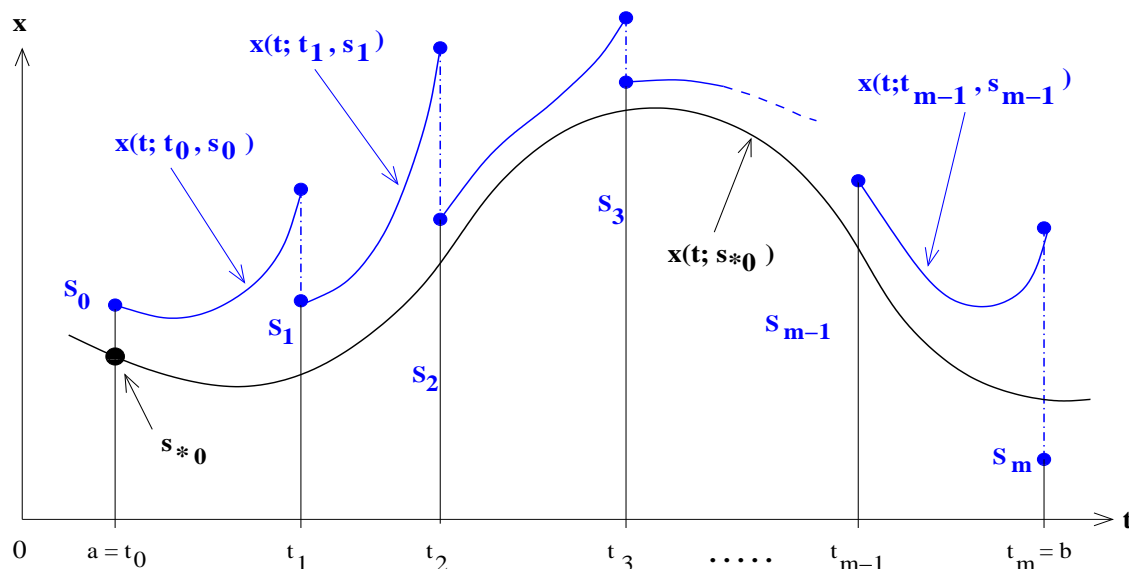


Abbildung 5: Stückweise Lösung mit der Mehrzielmethode

$s_m := x(b)$  ergeben sich  $m$  Stetigkeitsbedingungen (Matching-Bedingungen) an die unbekannten „Schießwerte“  $s_0, s_1, \dots, s_m$ , ergänzt um die Randbedingungen zu

$$\begin{aligned} x(t_{k+1}; t_k, s_k) &= s_{k+1}, & k = 0(1)m - 1 \\ g(s_0, s_m) &= 0. \end{aligned} \quad (36)$$

Das daraus entstehende große Gleichungssystem für die  $n(m+1)$  unbekanntes Zahlenwerte  $s = (s_0, s_1, \dots, s_m)^T \in \mathbb{R}^{n \times (m+1)}$  erhält dann folgende Struktur:

$$F(s) := \begin{pmatrix} F_0(s_0, s_1) \\ F_1(s_1, s_2) \\ \vdots \\ F_{m-1}(s_{m-1}, s_m) \\ F_m(s_0, s_m) \end{pmatrix} = \begin{pmatrix} x(t_1; t_0, s_0) - s_1 \\ x(t_2; t_1, s_1) - s_2 \\ \vdots \\ x(t_m; t_{m-1}, s_{m-1}) - s_m \\ g(s_0, s_m) \end{pmatrix} = 0. \quad (37)$$

Das Einfach-Schießverfahren entspricht genau dem Sonderfall  $m = 1$ . Hier kann  $s_1$  direkt in die letzte Gleichung eingesetzt werden, womit sich (24) für den unbekanntes Wert  $s = s_0$  ergibt. Die theoretischen Grundlagen des Mehrfach-Schießverfahrens findet der interessierte Leser in [8]; wir fassen sie zusammen in

**Satz 17** *Die (hinreichend glatte) Lösung  $x^*(t)$  nach Voraussetzung 7 sei regulär. Mit  $s^* = (s_0^*, s_1^*, \dots, s_m^*)^T$  gilt:*

- (i)  $s^*$  ist Lösung von  $F(s) = 0$  genau dann, wenn  $x^*(t) = x(t; t_k, s_k^*)$ ,  $k = 0(1)m - 1$ , Lösung des Randwertproblems (2) ist.
- (ii) Eine Lösung  $s^*$  von (37) ist regulär genau dann, wenn die Lösung  $x^*(t)$  des Randwertproblems (2) regulär ist.
- (iii) (37) kann mit dem Newton-Verfahren gelöst werden, das  $Q$ -quadratisch konvergiert, falls die Startnäherung  $s^{(0)} = (s_0^{(0)}, s_1^{(0)}, \dots, s_m^{(0)})^T$  hinreichend nahe bei  $s^*$  liegt.
- (iv) Das numerische Integrationsverfahren für die  $m$  Anfangswertprobleme (35) habe die Konvergenzordnung  $p \in \mathbb{N}$ . Dann konvergieren die mit Schrittweite  $h$  berechneten Näherungen  $s^*(h)$  gegen  $s^*$ , d.h.

$$s_k^*(h) - s_k^* = \mathcal{O}(h^p), \quad k = 0(1)m.$$

Die diskrete Lösung  $(u_j(h))$ ,  $j = 0(1)N$ , des Randwertproblems besitzt ebenfalls die Konvergenzordnung  $p$ .

Um das System  $F(s) = 0$  zu lösen, wenden wir wie beim Einfach-Schießverfahren das Newton-Verfahren

$$s^{(\nu+1)} = s^{(\nu)} - F'(s^{(\nu)})^{-1}F(s^{(\nu)}), \quad \nu = 0, 1, 2, \dots \quad (38)$$

an.<sup>5</sup> Die Jacobi-Matrix hat wegen (37) die zyklisch bidiagonale Blockstruktur

$$F'(s) = \begin{pmatrix} G_0 & -I & 0 & \cdots & 0 \\ 0 & G_1 & -I & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & G_{m-1} & -I \\ B_a & 0 & \cdots & 0 & B_b \end{pmatrix} \quad \text{mit} \quad \begin{aligned} G_k &:= \frac{\partial x}{\partial s_k}(t_{k+1}; t_k, s_k) \\ B_a &:= \left. \frac{\partial g(v, w)}{\partial v} \right|_{v=s_0, w=s_m} \\ B_b &:= \left. \frac{\partial g(v, w)}{\partial w} \right|_{v=s_0, w=s_m}, \end{aligned}$$

<sup>5</sup> Hier bezeichnet  $s^{(\nu)}$  die  $\nu$ -te Newton-Näherung, wogegen  $s_k$  die  $k$ -te Komponente von  $s$  ist.

wobei  $I \in \mathbb{R}^{n \times n}$  die Einheitsmatrix ist. Die  $n \times n$ -Matrizen  $G_k$  können wir wie beim Einfach-Schießen durch Integration des Variationssystems auf dem Intervall  $I_k = [t_k, t_{k+1}]$

$$\dot{X} = A(t; t_k, s_k)X, \quad X(t; t_k, s_k) = I, \quad k = 0(1)m - 1 \quad (39)$$

bestimmen, wobei die zugehörigen Jacobi-Matrizen  $A$  und die Matrizen  $G_k$  durch

$$A(t; t_k, s_k) := f_x(t, x(t; t_k, s_k)) \quad \text{und} \quad G_k = X(t_{k+1}; t_k, s_k) = \frac{\partial x}{\partial s_k}(t_{k+1}; t_k, s_k)$$

definiert sind. Mit diesen Erweiterungen können wir nun den Grundalgorithmus 9 zum *Mehrfach-Schießverfahren* weiterentwickeln. Bei gegebener Startnäherung  $s_0 \in \mathbb{R}^n$  und To-

### Algorithmus 18 (Mehrfach-Schießverfahren)

Function  $[s^*] = \text{multiple\_shooting}(f, g, f_x, B_a, B_b, a, b, s_0, tol, \nu max)$

1. Wähle Schießpunkte  $t_0 = a < t_1 < \dots < t_m = b$  und Startnäherungen  $s = (s_0, s_1, \dots, s_m)^T$ .
2. Für  $\nu = 0(1)\nu max$  iteriere:
  - 2.1. (Anfangswertprobleme) Für  $k = 0(1)m - 1$  wiederhole:
    - Löse auf  $I_k = [t_k, t_{k+1}]$  die  $n + 1$  Systeme (35) und (39).  
Ergebnisse:  $x(t_{k+1}; t_k, s_k) \in \mathbb{R}^n$ ,  $X(t_{k+1}; t_k, s_k) \in \mathbb{R}^{n \times n}$
  - 2.2. (Nullstellenaufgabe)
    - Bilde den Vektor  $F(s)$  nach (37) und die Matrix  $F'(s)$ .
  - 2.3. (Newton-Schritt) Löse das lineare System
 
$$F'(s) \cdot d = -F(s).$$
  - 2.4. Falls  $\|d\| < tol \cdot (1 + \|F'(s)\|)$ , so gehe zu Schritt 3.
  - 2.5. (Newton-Korrektur) Aktualisiere  $s := s + d$ .
3. Setze  $s^* := s$  und bestimme die Lösung durch
 
$$\dot{x} = f(t, x), \quad x(t_k) = s_k, \quad t_k \leq t \leq t_{k+1}, \quad k = 0(1)m - 1.$$
4. Return  $s^*$

leranz  $tol$  müssen die Schießpunkte  $t_k$  und zugehörigen Anfangswerte  $s_k$  so gewählt werden, dass die Anfangswertprobleme des Schrittes 2.1 in den Intervallen  $I_k = [t_k, t_{k+1}]$  eindeutig lösbar sind. Satz 12 garantiert dies für alle  $s_k \in \mathcal{S}[s_k^*; \varrho_k]$  mit den Radien

$$\varrho_k := r e^{-L_k(t_{k+1}-t_k)} \quad \text{und Lipschitz-Konstanten} \quad L_k := \sup_{I_k \times D} |f_x(t, x)|.$$

Mit einer hinreichend feinen Zerlegung von  $I$  kann man damit stets die Durchführbarkeit des Mehrfach-Schießverfahrens garantieren. Problematisch bleibt die Bestimmung einer

Startlösung  $s_0$  und der Schießpunkte  $t_k$ . Mittels Parameterfortsetzung Da der arithmetische Aufwand mit zunehmender Gitterfeinheit rasch ansteigt, sollten möglichst wenige Schießpunkte verwendet werden. Algorithmusschritt 2.1 zeigt, dass der numerische Integrationsaufwand gegenüber dem Einfach-Schießverfahren nicht zunimmt. Der Mehraufwand resultiert aus den entstehenden größeren linearen Gleichungssystemen. Diese sollten mittels LU-Zerlegung mit Pivotisierung gelöst werden, wobei die schwache (sparse) Besetzungsstruktur bei der Speicherung zu berücksichtigen ist. (vgl. [5, Kapitel 21]) lassen sich in komplizierten Anwendungen die Schießpunkte schrittweise anpassen.

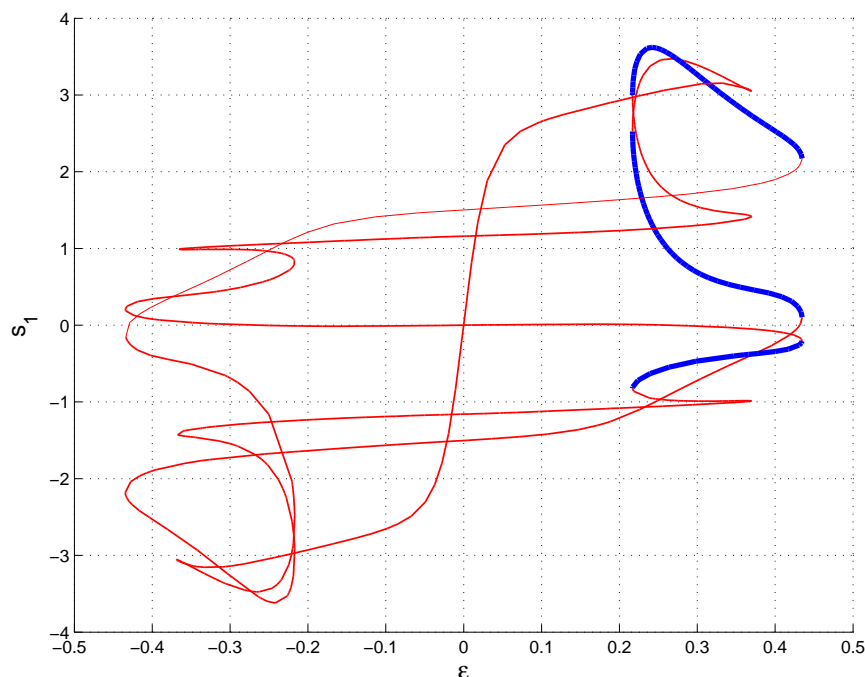


Abbildung 6:  $(\varepsilon, s_1)$ -Diagramm mit Stabilitätsverhalten der periodischen Lösungen

**Beispiel 19** Wir wenden uns der Berechnung der  $3\pi$ -periodischen Lösung des subharmonisch reagierenden Systems aus Beispiel 15 zu. Abb. 6 zeigt die 1. Komponente  $s_1$  der Lösung in Abhängigkeit vom Parameter  $\varepsilon$ , den  $3\pi$ -periodischen Lösungszweig. Asymptotisch stabile Abschnitte gemäß Satz 16 sind darin dick gezeichnet. Beginnend mit  $\varepsilon = 0.36$  und dem Startpunkt  $s_0 = x(0) = (0.38697629, 0.09266292)^T$  versagt das Einfach-Schießverfahren schon nach wenigen Fortsetzungsschritten für  $\varepsilon$  bei einem charakteristischen Multiplikator<sup>6</sup>  $m_i$  der Größenordnung  $10^4$ . Abb. 7 stellt die maximalen charakteristischen Multiplikatoren der Lösungskurve in Abhängigkeit von deren Bogenlänge logarithmisch skaliert dar. Setzen wir das Mehrfach-Schießverfahren mit  $m$  äquidistant gewählten Schießpunkten  $t_k = kT/m$  ein, so berechnen wir die Monodromiematrix nun mit den Fundamentallösungen auf den Teilintervallen durch

$$M = X(t_m; t_{m-1}, s_{m-1}^*) \cdot X(t_{m-1}; t_{m-2}, s_{m-2}^*) \cdot \dots \cdot X(t_1; t_0, s_0^*).$$

In Abb. 7 sind die Abbruchpunkte der Fortsetzung mit der Anzahl  $m$  gekennzeichnet. Erst mit 12 Schießpunkten wird die Fortsetzung in diesem stark instabilen Bereich mit

<sup>6</sup> Vgl. Satz 16

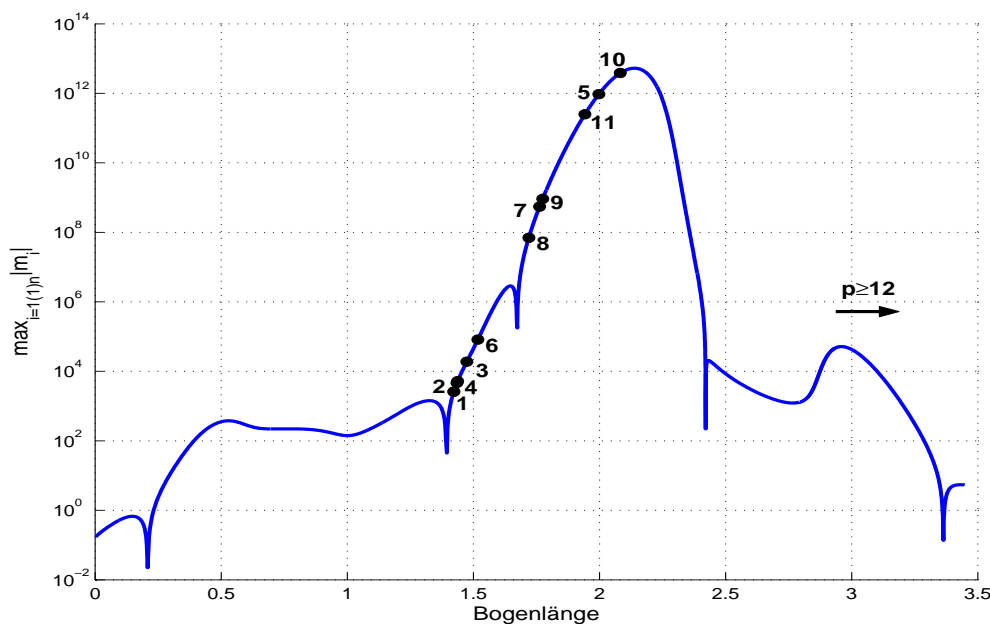


Abbildung 7: Betragsgrößerer Multiplikator in Abhängigkeit von der Bogenlänge

$\max_{i=1(1)n} |m_i| > 10^{12}$  gemeistert. Auffällig ist, dass man mit nur fünf Schießpunkten die Instabilität beinahe überwindet, aber erst die Verwendung der doppelten Anzahl besser abzuschneiden vermag. Eine automatische Schießpunktanpassung der Lösungsfortsetzung im instabilen Bereich ist deshalb ratsam.  $\square$

## 4 Finite Differenzen und Kollokationsverfahren

Während Schießverfahren das Randwertproblem als ein dynamisches System betrachten und wie in Beispiel 15 auf eine Folge von Anfangswertproblemen zurückführen, gehen die nun behandelten Verfahren eher vom statischen Charakter des Zweipunkt-Randwertproblems aus, wie er bei Biegeproblemen in Beispiel 37(2) oder Eigenwertproblemen in Beispiel 3 auftritt. Die Näherungslösung wird gleichzeitig an vielen Lösungspunkten bestimmt, weshalb dieser Zugang mitunter als *globale Methode* charakterisiert wird. Wir wollen die unabhängige „Zeitvariable“ weiterhin mit  $t$  bezeichnen und das Randwertproblem wie bisher in Vektorschreibweise

$$\dot{x} = f(t, x), \quad g(x(a), x(b)) = 0, \quad f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad g : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n \quad (40)$$

notieren. Die Standardvoraussetzungen 7 zu Glattheit und Existenz sollen erfüllt sein. Im Unterschied zur Mehrzielmethode nehmen wir nun eine Segmentierung des Grundintervalles  $I = [a, b]$  in viele Teilintervalle vor, indem wir endliche Gitter

$$I_N = \{ t_j \mid t_j = t_{j-1} + h_j, \quad h_j > 0, \quad j = 1(1)N, \quad t_0 = a, \quad t_N = b \} \quad (41)$$

mit den  $N + 1$  Gitterpunkten  $t_j$  und den positiven Schrittweiten  $h_j$  einführen. Zu jedem Gitterpunkt  $t_j$  bestimmen wir eine diskrete Lösung  $u_j$ , die die exakte Lösung  $x(t_j)$  approximiert. *Finite-Differenzenverfahren* (engl.: *finite difference method*, *FDM*) nutzen dafür einfache

Diskretisierungen mit niedriger Konvergenzordnung und dementsprechend feiner Segmentierung, wogegen *Kollokationsverfahren* genauere Approximationen auf gröberen Gittern  $I_N$  vornehmen.

## 4.1 Finite-Differenzenverfahren (FDM)

Setzen wir die Lösung  $x(t)$  in die DGL ein und integrieren über das  $j$ -te Teilintervall  $[t_{j-1}, t_j]$ , so lässt sich  $x(t_j)$  durch den Ausdruck

$$x(t_j) - x(t_{j-1}) = \int_{t_{j-1}}^{t_j} f(t, x(t)) dt \quad (42)$$

darstellen. Mit den einfachen numerischen Integrationsregeln aus [5] gewinnen wir daraus

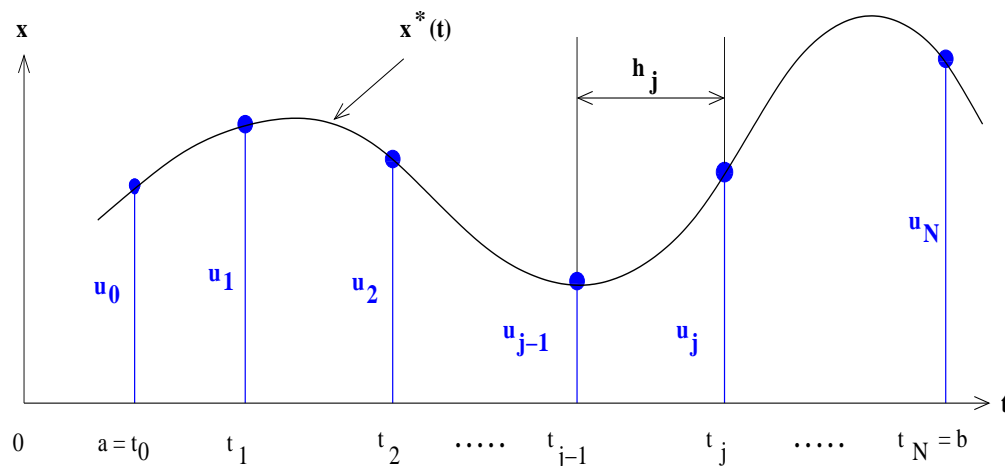


Abbildung 8: Lösung mit dem Finite-Differenzenverfahren

Gitterfunktionen  $u = (u_0, u_1, \dots, u_N)$  wie in Abb. 8, z.B.

$$u_j - u_{j-1} = h_j f(t_{j-1}, u_{j-1}) \quad (\text{Rechteckregel/links})$$

$$u_j - u_{j-1} = h_j f(t_j, u_j) \quad (\text{Rechteckregel/rechts})$$

$$u_j - u_{j-1} = \frac{h_j}{2} [f(t_{j-1}, u_{j-1}) + f(t_j, u_j)] \quad (\text{Trapezregel})$$

Diese Ansätze führen auf die bekannten expliziten und impliziten Euler-Verfahren und auf das implizite Heun-Verfahren (vgl. [17]). Da wegen der ebenfalls zu erfüllenden Randbedingung  $g(u_0, u_N) = 0$  nun explizite Verfahren keinen Vorteil gegenüber impliziten Verfahren bedeuten und das Heun-Verfahren die gegenüber dem impliziten Euler-Verfahren höhere Konsistenzordnung 2 besitzt, wollen wir die *Trapezregel* in der Darstellung

$$\begin{aligned} \frac{1}{h_j}(u_j - u_{j-1}) &= \frac{1}{2} [f(t_{j-1}, u_{j-1}) + f(t_j, u_j)], \quad j = 1(1)N \\ g(u_0, u_N) &= 0 \end{aligned} \quad (43)$$

als Prototyp eines Einschrittverfahrens notieren. Offenbar wird darin die Lösungsableitung  $\dot{x}(t_{j-1} + h_j/2)$  durch den zentralen Differenzenquotienten  $\frac{1}{h_j}(u_j - u_{j-1})$  approximiert und

die beiden Funktionswerte  $f_{j-1}$  und  $f_j$  werden gemittelt. Ein ähnlich aufgebautes System entsteht, wenn wir eine Mittelung der  $t_j$ -Werte und der  $u_j$ -Werte vornehmen, womit sich die *implizite Mittelpunkregel*

$$\begin{aligned} \frac{1}{h_j}(u_j - u_{j-1}) &= f\left(\frac{t_{j-1} + t_j}{2}, \frac{u_{j-1} + u_j}{2}\right), \quad j = 1(1)N \\ g(u_0, u_N) &= 0 \end{aligned} \quad (44)$$

als Alternative zur Trapezregel ergibt.

**Bemerkung 20** Auch für allgemeine lineare  $k$ -Schrittverfahren können wir derartige Verfahrenspaare gewinnen. Ausgehend von einem vorgegebenen linearen Mehrschrittverfahren mit konstanter Schrittweite  $h$

$$\sum_{i=0}^k \alpha_i u_{j-k+i} = h \sum_{i=0}^k \beta_i f(t_{j-k+i}, u_{j-k+i}), \quad j = k(1)N \quad (45)$$

lässt sich unter Beachtung der Normierungsbedingungen  $\beta_0 + \beta_1 + \dots + \beta_k = 1$  das zugehörige *one-leg-Verfahren* (dt.: *Einbeinverfahren*) in der Form

$$\sum_{i=0}^k \alpha_i u_{j-k+i} = h \cdot f\left(\sum_{i=0}^k \beta_i t_{j-k+i}, \sum_{i=0}^k \beta_i u_{j-k+i}\right), \quad j = k(1)N \quad (46)$$

notieren.<sup>7</sup> Im Gegensatz zu den linearen Mehrschrittverfahren wird die Approximation der Lösungsableitung  $\dot{x}(t)$  nicht durch eine gewichtete Summe, sondern durch den Funktionswert an einer einzigen Stelle („one leg“) vorgenommen. Das populärste derartige Formelpaar bildet die Trapezregel (43) mit der impliziten Mittelpunkregel (44) als der zugehörigen one-leg-Implementierung.  $\square$

Im Unterschied zu Einschrittverfahren erfordern Mehrschrittformeln (45) und (46) spezielle Approximationen in Randnähe, weshalb wir darauf hier nicht eingehen werden. Mit der Trapezregel (43) definieren wir für die  $n(N+1)$  unbekanntenen Werte  $u = (u_0, u_1, \dots, u_N)^T \in \mathbb{R}^{n \times (N+1)}$  das Gleichungssystem

$$F(u) := \begin{pmatrix} F_0(u_0, u_N) \\ F_1(u_0, u_1) \\ \vdots \\ F_N(u_{N-1}, u_N) \end{pmatrix} = \begin{pmatrix} g(u_0, u_N) \\ \frac{1}{h_1}(u_1 - u_0) - \frac{1}{2}[f(t_0, u_0) + f(t_1, u_1)] \\ \vdots \\ \frac{1}{h_N}(u_N - u_{N-1}) - \frac{1}{2}[f(t_{N-1}, u_{N-1}) + f(t_N, u_N)] \end{pmatrix} = 0. \quad (47)$$

Wir notieren zur Lösung dieses großen Systems  $F(u) = 0$  das Newton-Verfahren, das wir in der praktikablen Form

$$F'(u^{(\nu)})(u^{(\nu+1)} - u^{(\nu)}) = -F(u^{(\nu)}), \quad \nu = 0, 1, 2, \dots \quad (48)$$

<sup>7</sup>Die Klasse der one-leg-Verfahren wurde von G. Dahlquist für Anfangswertprobleme eingeführt.

angeben.<sup>8</sup> Die Jacobi-Matrix  $F'(u)$  ist von zyklisch bidiagonaler Blockstruktur mit den  $n \times n$ -Blöcken

$$B_a := \left. \frac{\partial g(v, w)}{\partial v} \right|_{v=u_0, w=u_N} \quad \text{und} \quad B_b := \left. \frac{\partial g(v, w)}{\partial w} \right|_{v=u_0, w=u_N}$$

$$S_j := -\frac{1}{h_j} I - \frac{1}{2} f_x(t_{j-1}, u_{j-1}),$$

$$R_j := +\frac{1}{h_j} I - \frac{1}{2} f_x(t_j, u_j), \quad j = 1(1)N$$

und der Einheitsmatrix  $I$ . Mit den Abkürzungen  $d := u^{(\nu+1)} - u^{(\nu)} = (d_0, \dots, d_N)^T$  für die Newton-Korrektur und  $b := -F(u^{(\nu)}) = (b_0, \dots, b_N)^T$  für das negative Residuum ergibt sich so das folgende lineare Gleichungssystem im  $\nu$ -ten Newton-Schritt:

$$\begin{pmatrix} B_a & 0 & 0 & \cdots & B_b \\ S_1 & R_1 & 0 & \cdots & 0 \\ 0 & S_2 & R_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & S_N & R_N \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix}. \quad (49)$$

Die Grundform des Verfahrens ist in Algorithmus 21 dargestellt. Vorzugeben sind insbe-

#### Algorithmus 21 (Finite-Differenzenverfahren)

Function  $[u^*] = \text{FDM}(f, g, f_x, B_a, B_b, a, b, N, tol, \nu max)$

1. Wähle Gitterpunkte  $t_0 = a < t_1 < \dots < t_N = b$  und Startnäherungen  $u = (u_0, u_1, \dots, u_N)^T$ .
2. Für  $\nu = 0(1)\nu max$  iteriere:
  - 2.1. Bestimme Residuum  $b := -F(u)$  und Randmatrizen  $B_a$  und  $B_b$ .
  - 2.2. Berechne die Matrizen  $S_j$  und  $R_j$  für  $j = 0(1)N$ .
  - 2.3. (Newton-Schritt) Löse das lineare System (48), d.h.

$$F'(u) \cdot d = -F(u).$$

- 2.4. Falls  $\|d\| < tol \cdot (1 + \|b\|)$ , so gehe zu Schritt 3.
- 2.5. (Newton-Korrektur) Aktualisiere  $u := u + d$ .
3. Return  $u^* := u$

sondere der Diskretisierungsparameter  $N$  und die Abbruchgenauigkeit  $tol > 0$  des Newton-Verfahrens. Über die Genauigkeit der gewonnenen Gitterfunktion  $u^*$ , d.h. über ihren Diskretisierungsfehler  $e_j := u_j^* - x(t_j)$ , ist allerdings keine Aussage möglich.

<sup>8</sup>  $u^{(\nu)}$  kennzeichnet auch hier die  $\nu$ -te Newton-Näherung, wogegen  $u_j$  die  $j$ -te Komponente von  $u$  ist.

Den zentralen Teil des Algorithmus bildet dessen Schritt 2.3. In vielen Fällen kann das lineare Gleichungssystem (49) direkt durch Blockelimination gelöst werden. Dazu multiplizieren wir die letzte Blockzeile mit einer Transformationsmatrix  $T_1 \in \mathbb{R}^{n \times n}$  mit der Eigenschaft  $T_1 \cdot R_N = -B_b$  und addieren die Zeile zur ersten Blockzeile:

$$\begin{pmatrix} B_a & 0 & 0 & \cdots & T_1 S_N & 0 \\ S_1 & R_1 & 0 & \cdots & 0 & 0 \\ 0 & S_2 & R_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & S_N & R_N \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} T_1 b_N + b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix}.$$

Die sukzessive Elimination der Nicht-Nullelemente oberhalb der Diagonalen erfolgt dann mit Transformationsmatrizen  $T_i$ , die aus den Gleichungssystemen

$$T_i \cdot R_{N-(i-1)} = -T_{i-1} \cdot S_{N-(i-2)}, \quad i = 2(1)N \quad \text{und} \quad T_1 \cdot R_N = -B_b \quad (50)$$

gewonnen werden. Im Ergebnis entsteht das untere Block-Dreieckssystem

$$\begin{pmatrix} T_N S_1 + B_a & 0 & 0 & \cdots & 0 & 0 \\ S_1 & R_1 & 0 & \cdots & 0 & 0 \\ 0 & S_2 & R_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & S_N & R_N \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} T_N b_1 + \dots + T_1 b_N + b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix}.$$

Dessen Lösung kann nun in einem Durchlauf berechnet werden, indem wir die  $N + 1$  Gleichungssysteme sukzessive nach den  $d_i$  auflösen:

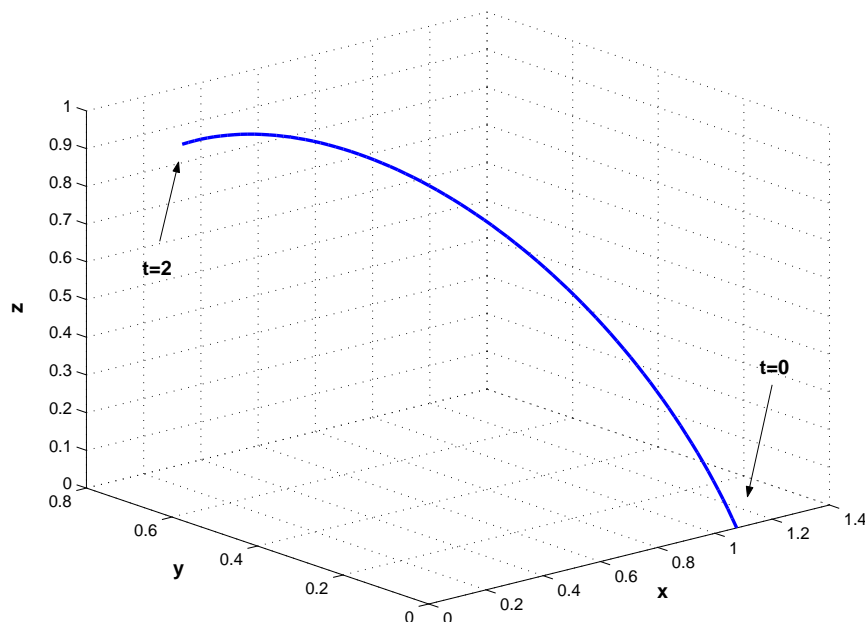
$$\begin{aligned} d_0 &= (T_N S_1 + B_a)^{-1} (T_N b_1 + \dots + T_1 b_N + b_0) \\ d_i &= R_i^{-1} (b_i - S_i d_{i-1}), \quad i = 1(1)N \end{aligned} \quad (51)$$

Die Berechnung inverser Matrizen sollte dabei aus Effizienzgründen zugunsten der Lösung linearer Gleichungssysteme vermieden werden. Ist (49) schlecht konditioniert, so empfiehlt sich eine LU-Zerlegung der Gesamtmatrix mit Pivotisierung und sparser Speicherung oder eine QR-Zerlegung (vgl. [5, Kapitel 20]).

**Beispiel 22** Das Zwei-Körper-Problem in  $\mathbb{R}^3$  aus Beispiel 37 (3) führt auf das System

$$\begin{aligned} \dot{x}_1 &= x_4 & , & & x_1(0) - 1.076 &= 0 \\ \dot{x}_2 &= x_5 & , & & x_2(0) &= 0 \\ \dot{x}_3 &= x_6 & , & & x_3(0) &= 0 \\ \dot{x}_4 &= -k \cdot x_1/r^3 & , & & x_1(2) &= 0 \\ \dot{x}_5 &= -k \cdot x_2/r^3 & , & & x_2(2) - 0.576 &= 0 \\ \dot{x}_6 &= -k \cdot x_3/r^3 & , & & x_3(2) - 0.997661 &= 0 \end{aligned} \quad \text{mit} \quad r^2 = x_1^2 + x_2^2 + x_3^2$$

mit Randbedingungen in Standardform. Auf einem äquidistanten Gitter mit  $N = 80$  wird durch eine numerische Integration, beginnend mit dem Anfangswert  $u_0 = (1, 1, 1, 0, 0, 0)^T$ , eine Startlösung  $u = (u_0, u_1, \dots, u_N)^T$  generiert. Darauf wird Algorithmus 21 in einer durch ein gedämpftes Newton-Verfahren mit Differenzenapproximation der Jacobi-Matrix verbesserten Version angewandt. Bei vorgegebener Abbruchgenauigkeit  $tol = 10^{-8}$  benötigt er 5 Iterationsschritte und liefert die Lösung  $(x, y, z)$  in Abb. 9. Die in den 6 Randbedingungen vorgegebenen Lösungswerte wurden auf 8 Dezimalstellen genau approximiert.  $\square$

Abbildung 9: Zwei-Körper-Problem in  $\mathbb{R}^3$  mit Finite-Differenzenverfahren

## 4.2 Konvergenzanalyse von FDM

Während die Lösung von Anfangswertproblemen, beginnend mit dem gegebenen Anfangswert  $y(a) = y_0$ , schrittweise mittels eines Ein- oder Mehrschrittverfahrens erfolgen kann, ist diese Vorgehensweise bei Randwertproblemen der allgemeinen Form

$$\dot{y} = f(t, y), \quad g(y(a), y(b)) = 0 \quad \text{mit } f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad g : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n \quad (52)$$

nicht möglich. Deshalb sind die Begriffe *Konsistenz*, *Stabilität* und *Konvergenz* nun so zu definieren, dass sie auch für diese und weitere Problemklassen zutreffen. Der von H. B. Keller [8] und H. J. Stetter [15] entwickelte und in [1] genutzte Zugang soll hier vorgestellt werden.

Zuerst notieren wir das Randwertproblem als Operatorgleichung  $F(y) = 0$  mit einem geeigneten Operator  $F$ , der neben den Differentialgleichungen auch die Randbedingungen enthält. Mit einer beliebigen, fest gewählten Norm  $\|x\|_n$  in  $\mathbb{R}^n$  definieren wir dazu die Banach-Räume  $B = C^1(I; \mathbb{R}^n)$  als Urbildraum und  $B^0 = \mathbb{R}^n \times C(I; \mathbb{R}^n)$  als Bildraum, versehen mit den Normen

$$\|y\|_B := \max_{t \in I} \|y(t)\|_n + \max_{t \in I} \|\dot{y}(t)\|_n \quad \text{und} \quad (53)$$

$$\|z\|_{B^0} := \|z_0\|_n + \max_{t \in I} \|z_1(t)\|_n,$$

wobei  $z = (z_0, z_1)^T \in B^0$  mit  $z_0 \in \mathbb{R}^n$  und  $z_1 \in C(I; \mathbb{R}^n)$  ist. Dann kann Problem (52) als Operatorgleichung

$$F(y) = 0 \quad \text{mit} \quad f : B \rightarrow B^0 \quad (54)$$

geschrieben werden, wenn der Operator  $F : B \rightarrow B^0$  durch

$$F(y(t)) := \begin{pmatrix} g(y(a), y(b)) \\ \dot{y}(t) - f(t, y(t)) \end{pmatrix} \quad (55)$$

definiert wird. Das Tripel  $\mathcal{P} = (B, B^0, F)$  bildet das *Ausgangsproblem*, für das ein Element  $y^* \in B$  gesucht wird, das der Operatorgleichung  $F(y^*) = 0$  genügt. Jedes derartige  $y^*$  heißt *klassische Lösung*.

**Bemerkung 23** Eine geeignete Wahl der beiden Funktionenräume  $B$  und  $B^0$  ist wesentlich, wenn bestimmte Lösungseigenschaften erwünscht sind. So bedeutet die Abschätzung  $\|y\|_B \leq C$  mit einer Konstanten  $C$ , dass außer den Funktionswerten auch die Ableitungswerte der Lösung durch  $C$  beschränkt sind, wogegen dies bei  $\|y\|_{B^0} \leq C$  nur für die Funktionswerte gilt.

Auf  $I = [a, b]$  konstruieren wir zu jeder natürlichen Zahl  $N \in \mathbb{N}$  ein Gitter

$$I_N := \{t_j \in I \mid t_0 = a, t_N = b, t_j = t_{j-1} + h_j, j = 1(1)N\} \quad (56)$$

mit den Schritten  $h_j > 0$ ,  $j = 1(1)N$  und dem Maximalschritt

$$h := \max_{j=1(1)N} h_j. \quad (57)$$

Wir wollen des Weiteren eine lineare Konvergenz der Gitterfolge  $\{I_N\}$ ,  $N \in \mathbb{N}$  fordern und geben dazu

**Definition 24** Die Gitterfolge  $\{I_N\}$ ,  $N \in \mathbb{N}$ , konvergiert linear, wenn zwei Konstanten  $c_1 > 0$ ,  $c_2 > 0$  existieren, für die mit den Schritten  $h_j$  jedes Gitters  $I_N$  die Bedingung

$$c_1/N \leq h_j \leq c_2/N, \quad j = 1(1)N \quad (58)$$

für alle  $N \in \mathbb{N}$  erfüllt ist.

Wegen (58) zieht der Grenzübergang  $N \rightarrow \infty$  die Konvergenz aller Schrittweiten  $h_j \rightarrow 0$  nach sich. Zudem sind nun die Terme  $\mathcal{O}(h_j^p)$ ,  $\mathcal{O}(h^p)$  und  $\mathcal{O}(N^{-p})$  asymptotisch äquivalent, weshalb Fehlerordnungen stets durch  $\mathcal{O}(h^p)$  angegeben werden, wie dies bei Finite-Differenzenverfahren üblich ist. Existiert eine von  $N$  unabhängige Konstante  $K > 0$  mit

$$\max_{j=1(1)N} h_j / \min_{j=1(1)N} h_j \leq K \quad \text{für alle Gitter mit } N \in \mathbb{N}, \quad (59)$$

so ist für derartige *quasi-äquidistante Gitter* auch (58) erfüllt.

Um das Ausgangsproblem numerisch auf einem Computer darstellen und lösen zu können, muss man es „diskretisieren“. Dazu betrachten wir zwei Folgen endlich dimensionaler Banach-Räume,

- die Urbildräume  $\{B_N\}$ ,  $N \in \mathbb{N}$ , mit  $\dim B_N < \infty$ ,
- die Bildräume  $\{B_N^0\}$ ,  $N \in \mathbb{N}$ , mit  $\dim B_N^0 < \infty$

und definieren eine Folge von Operatoren  $\{F_N\}$ ,  $N \in \mathbb{N}$ , mit  $F_N : B_N \rightarrow B_N^0$ . Die Folge  $\{\mathcal{P}_N\}$  der Tripel  $\mathcal{P}_N = (B_N, B_N^0, F_N)$  wird als *diskretes Problem* bezeichnet. Jede Elementfolge  $\{u_N^*\}$ ,  $N \in \mathbb{N}$ , mit  $u_N^* \in B_N$  ist eine *Lösung des diskreten Problems*, falls

$$\boxed{F_N(u_N^*) = 0, \quad F_N : B_N \rightarrow B_N^0, \quad N \in \mathbb{N}} \quad (60)$$

gilt. Zur Realisierung dieser diskreten Aufgabe für unser Randwertproblem führen wir diskrete Analoga  $B_N$  und  $B_N^0$  ein, die Räume von Gitterfunktionen darstellen. Seien  $B_N = \mathcal{C}_N^1(I_N; \mathbb{R}^n)$  und  $B_N^0 = \mathbb{R}^n \times C_N(I_N; \mathbb{R}^n)$  derartige Räume von Gitterfunktionen  $u = (u_0, u_1, \dots, u_N)^T$  bzw.  $v = (v_0, v_1, \dots, v_N)^T$  mit  $u_j, v_j \in \mathbb{R}^n$ ,  $j = 0(1)N$ . Die zugehörigen Normen

$$\begin{aligned} \|u\|_{B_N} &:= \max_{j=0(1)N} \|u_j\|_n + \max_{j=1(1)N} \|\partial u_j\|_n, \\ \|v\|_{B_N^0} &:= \|v_0\|_n + \max_{j=1(1)N} \|v_j\|_n, \end{aligned} \quad (61)$$

sind diskrete Analoga zu den Normen (53), wobei  $\partial u_j = (u_j - u_{j-1})/h_j$  bezeichnet.

Um die Beziehung zwischen dem Ausgangsproblem  $\mathcal{P}$  und dem diskreten Problem  $\{\mathcal{P}_N\}$ ,  $N \in \mathbb{N}$ , herzustellen, werden wir lineare Abbildungen zwischen den entsprechenden Räumen, so genannte Restriktionsoperatoren, einführen.

**Definition 25 (Restriktion)** Die Folge  $\{p_N\}$ ,  $N \in \mathbb{N}$ , linearer beschränkter Operatoren mit  $p_N : B \rightarrow B_N$  sei normkonsistent, d.h. für alle  $y \in B$  gilt

$$\lim_{N \rightarrow \infty} \|p_N y\|_{B_N} = \|y\|_B, \quad N \in \mathbb{N}.$$

Dann heißt  $p_N$  Restriktionsoperator (Restriktion, Einschränkung) von  $B$  auf  $B_N$ . Analoges gelte für die Restriktion  $p_N^0 : B^0 \rightarrow B_N^0$  des Bildraumes.

Wegen der Normkonsistenz ist die Folge der Quotienten  $\|p_N y\|_{B_N} / \|y\|_B$  für jedes Element  $y \in B$ ,  $x \neq 0$ , beschränkt. Nach dem Satz von Banach und Steinhaus<sup>9</sup> sind dann die Normen der Operatoren  $p_N$  und  $p_N^0$  sogar gleichmäßig beschränkt. Es existieren also Konstanten  $P \geq 1$  und  $P^0 \geq 1$ , so dass

$$\|p_N\| \leq P, \quad \|p_N^0\| \leq P^0, \quad N \in \mathbb{N} \quad (62)$$

mit den induzierten Operatornormen gilt. Für die Restriktionsoperatoren  $p_N : B \rightarrow B_N$  und  $p_N^0 : B^0 \rightarrow B_N^0$  benutzen wir im Hinblick auf die Definition von  $F_N$

$$\begin{aligned} \{p_N y\}_j &:= y(t_j), \quad j = 0(1)N \quad \text{und} \\ \{p_N^0 z\}_j &:= \begin{cases} z_0 & , \quad j = 0 \\ z_1(t_j - h_j/2) & , \quad j = 1(1)N. \end{cases} \end{aligned} \quad (63)$$

<sup>9</sup> Vgl. [10], S. 511, Satz 4

Offensichtlich sind  $p_N$  und  $p_N^0$  normkonsistent und erfüllen (62) mit  $P = P^0 = 1$ . Eine andere Wahl von  $p_N^0$  wäre

$$\{p_N^0 z\}_j := \begin{cases} z_0 & , j = 0 \\ \frac{1}{2}[z_1(t_j) + z_1(t_{j-1})] & , j = 1(1)N, \end{cases} \quad (64)$$

die offenbar denselben Beziehungen mit  $P = P^0 = 1$  genügt.

Zur Definition der Differenzenoperatoren  $F_N$  empfehlen sich für (55) die eingeführten Ein-schrittformeln, um spezielle Approximationen in Randnähe zu vermeiden und den Aufwand zur Lösung der entstehenden finiten Gleichungssysteme niedrig zu halten. Die *Trapezregel* (43) lautet nun in unserer Operatorschreibweise

$$\{F_N u\}_j := \begin{cases} g(u_0, u_N) & , j = 0 \\ \frac{1}{h_j}(u_j - u_{j-1}) - \frac{1}{2}(f(t_j, u_j) + f(t_{j-1}, u_{j-1})) & , j = 1(1)N, \end{cases} \quad (65)$$

wogegen die *implizite Mittelpunkregel* (44) auf die Operatoren

$$\{F_N u\}_j := \begin{cases} g(u_0, u_N) & , j = 0 \\ \frac{1}{h_j}(u_j - u_{j-1}) - f\left(\frac{t_j + t_{j-1}}{2}, \frac{u_j + u_{j-1}}{2}\right) & , j = 1(1)N \end{cases} \quad (66)$$

führt. Faßt man die 4 eingeführten Banachräume  $B, B^0, B_N, B_N^0$  und die 4 Operatoren  $F, F_N, p_N, p_N^0$  zusammen, so liefern sie ein *Diskretisierungsverfahren*. Den Zusammenhang zwischen Ausgangsproblem  $\mathcal{P}$  und diskretem Problem  $\{\mathcal{P}_N\}$ ,  $N \in \mathbb{N}$ , erhält man in anschaulicher Form durch das *Approximationsschema*

$$\begin{array}{ccc} B & \xrightarrow{F(y) = 0} & B^0 & \text{Lösung: } y^* \in B \\ p_N \downarrow & & \downarrow p_N^0 & \\ B_N & \xrightarrow{F_N(u_N) = 0} & B_N^0 & \text{Lösung: } u_N^* \in B_N \end{array}$$

Dieses allgemeine Schema verdeutlicht zugleich das Ziel, für  $y \in B$  anstelle der vorgegebenen Gleichung  $F(y) = 0$  das diskrete Problem  $F_N(u_N) = 0$  zu lösen. Das gelingt offenbar, wenn der Operator  $F$  und die Diskretisierung  $p_N$  für  $N \rightarrow \infty$  miteinander vertauschbar sind. Am Randwertproblem erkennt man zudem, dass die endlich dimensionalen Ersatzräume  $B_N$  und  $B_N^0$  keine Unterräume der Originalräume  $B$  und  $B^0$  sind. Hierin unterscheiden sich Diskretisierungsverfahren von der Klasse der Projektionsverfahren.

Wie man an Randwertproblem (52) sieht, ist die Wahl der Operatoren  $F, p_N, p_N^0$  und der entsprechenden Banach-Räume meistens evident. Die grundlegende Frage eines Diskretisierungsverfahrens lautet dann: Wie ist zu vorgegebenem Operator  $F$  der diskrete Operator  $F_N$  zu konstruieren, damit er das mit  $F$  gestellte Problem „möglichst genau“ approximiert? Dazu sollten die Bilder  $F_N(p_N y)$  und  $p_N^0 F(y)$  für alle  $y$  in einer Umgebung der Lösung  $y^*$  ebenfalls nahe beieinander liegen. Wir kommen deshalb zu

**Definition 26 (Konsistenz)**

- (i) Das diskrete Problem  $\mathcal{P}_N = (B_N, B_N^0 F_N)$  heißt konsistent mit dem Ausgangsproblem  $\mathcal{P} = (B, B^0, F)$  im Punkt  $y \in B$ , falls

$$\lim_{N \rightarrow \infty} \|F_N(p_N y) - p_N^0 F(y)\|_{B_N^0} = 0, \quad N \in \mathbb{N} \quad (67)$$

gilt. Ist  $\mathcal{P}_N$  konsistent mit  $\mathcal{P}$  für alle  $y \in D \subset B$ , so ist das Diskretisierungsverfahren auf  $D$  konsistent.

- (ii)  $\mathcal{P}_N$  ist konsistent (in  $y \in B$ ) mit  $\mathcal{P}$  mit der Ordnung  $p \in \mathbb{N}$ , falls

$$\|F_N(p_N y) - p_N^0 F(y)\|_{B_N^0} = \mathcal{O}(N^{-p}), \quad N \in \mathbb{N} \quad (68)$$

ist, d.h. falls Konstanten  $N_0 \in \mathbb{N}$  und  $M > 0$  existieren, so dass für alle  $N \in \mathbb{N}$ ,  $N \geq N_0$ ,

$$\|F_N(p_N y) - p_N^0 F(y)\|_{B_N^0} \leq M N^{-p}$$

gilt.

Anschaulich bedeutet Konsistenz die asymptotische Vertauschbarkeit von  $F_N$  und  $p_N$  in obigem Approximationsschema. Um die Konsistenz eines Diskretisierungsverfahrens nachzuweisen und seine Konsistenzordnung zu bestimmen, entwickelt man in der Regel die Funktion  $y(t)$  an geeigneten Argumentstellen  $t_j$  in eine *Taylor-Reihe* in Termen von  $N^{-1}$ . Eine besondere Rolle unter allen Elementen  $y \in B$  spielt die vorausgesetzte exakte Lösung  $y^* \in D \subset B$  des Ausgangsproblems  $\mathcal{P}$ .

**Definition 27 (Lokaler Diskretisierungsfehler)**

$y^* \in D \subset B$  sei Lösung des Ausgangsproblems  $F(y) = 0$ . Dann heißt das Element  $\tau \in B_N^0$  mit

$$\tau = F_N(p_N y^*) - p_N^0 F(y^*) = F_N(p_N y^*), \quad N \in \mathbb{N} \quad (69)$$

lokaler Diskretisierungsfehler des Verfahrens  $\mathcal{P}_N$ .

**Beispiel 28** Für die Trapezregel (65) erhalten wir den lokalen Diskretisierungsfehler  $\tau_0 = g(y^*(a), y^*(b)) = 0$  und mittels Taylor-Entwicklung für  $j = 1(1)N$

$$\begin{aligned} \tau_j &= \{F_N(p_N y^*)\}_j = \frac{1}{h_j} [y^*(t_j) - y^*(t_{j-1})] - \frac{1}{2} [f(t_j, y^*(t_j)) + f(t_{j-1}, y^*(t_{j-1}))] \\ &= \frac{1}{h_j} [y^*(t_j) - y^*(t_{j-1})] - \frac{1}{2} [\dot{y}^*(t_j) + \dot{y}^*(t_{j-1})] = -\frac{1}{12} R(t_{j-1}, t_j) h_j^2 \end{aligned}$$

mit dem Restglied  $R(t_{j-1}, t_j)$ . Mit der Norm in  $B_N^0$  gewinnen wir daraus

$$\|\tau_j\|_{B_N^0} = \|F_N(p_N y^*) - p_N^0 F(y^*)\|_{B_N^0} \leq C \cdot h^2, \quad C = \text{const},$$

woraus die Konsistenzordnung 2 folgt, falls nur  $f \in \mathcal{C}^3(I \times \mathbb{R}^n)$  ist.  $\square$

Die Konsistenz eines Diskretisierungsverfahrens bedeutet wegen (67), dass der diskrete Operator  $F_N$  den gegebenen Operator  $F$  in der Umgebung der Lösung  $y^*$  approximiert. Falls diese Grundbedingung nicht erfüllt ist, so kann im Allgemeinen nicht erwartet werden, dass die Näherungslösungen  $u^*$  – falls sie überhaupt existieren – die exakte Lösung  $y^*$  approximieren. In vielen Fällen ist die Konsistenz allein jedoch nicht hinreichend für die Konvergenz der Näherungslösung, d.h. für

$$\lim_{N \rightarrow \infty} \|u^* - p_N x^*\|_{B_N} = 0.$$

Kleine Störungen wie Rundungsfehler, Abschneidefehler etc. dürfen sich möglichst nicht verstärken und kumulieren. Das Verfahren muss „robust“ oder auch „stabil gegenüber Störungen“ sein. Ersetzt man die rechte Seite des diskreten Problems  $F_N(u) = 0$  durch kleine Änderungen  $\delta^1, \delta^2$ , so lauten nun die gestörten Probleme

$$F_N(u^1) = \delta^1, \quad F_N(u^2) = \delta^2$$

mit den vorerst als existent vorausgesetzten Lösungen  $u^1, u^2 \in B_N$ . Aus der Theorie der algebraischen Gleichungssysteme ist bekannt, dass die Abbildung  $F_N$  numerisch stabil ist, wenn kleine Störungen der rechten Seiten nicht zu extrem großen Störungen der Lösungen führen. Das ist offenbar garantiert, wenn eine Beziehung

$$\|u^1 - u^2\|_{B_N} \leq S \cdot \|\delta^1 - \delta^2\|_{B_N^0}$$

mit einer Konstanten  $S > 0$  nachgewiesen werden kann. Damit für  $N \rightarrow \infty$  diese Stabilitätseigenschaft nicht verloren geht, fordert man die Unabhängigkeit der Konstanten  $S$  von  $N$  und gelangt so zum Begriff der diskreten Stabilität.

### Definition 29 (Diskrete Stabilität)

$F_N$  sei stetig auf  $\mathcal{K}[u^0; r] = \{u \in B_N \mid \|u - u^0\| \leq r\}$ . Falls für alle  $u^1, u^2 \in \mathcal{K}[u^0; r]$

$$\|u^1 - u^2\|_{B_N} \leq S \cdot \|F_N(u^1) - F_N(u^2)\|_{B_N^0} \quad \forall N \in \mathbb{N} \quad (70)$$

mit Konstanten  $S > 0, r > 0$  gilt, so ist der Operator  $F_N$  stabil auf  $u^0$  mit der Stabilitätsgrenze  $S$  und der Stabilitätsschwelle  $s = r/S$ .

In [8] wird eine vollständige Stabilitätstheorie für Differenzenverfahren zur Lösung des Randwertproblems (52) gegeben, mit der die Stabilität der beiden Diskretisierungen (65) und (66) gezeigt wird. Kann man nachweisen, dass die Lösung  $y^* \in B$  regulär (isoliert) gemäß Definition 10 ist, d.h. dass das homogene Variationssystem

$$\begin{aligned} \dot{e}(t) - f_x(t, y^*(t))e(t) &= 0 \\ B_a e(a) + B_b e(b) &= 0 \end{aligned} \quad \text{mit} \quad \begin{cases} B_a := \left. \frac{\partial g(v, w)}{\partial v} \right|_{v=y^*(a), w=y^*(b)} \\ B_b := \left. \frac{\partial g(v, w)}{\partial w} \right|_{v=y^*(a), w=y^*(b)} \end{cases} \quad (71)$$

nur die Lösung  $e(t) \equiv 0$  in  $B$  hat, so erhält man folgende Konsistenz- und Stabilitätsaussage:

**Lemma 30** Sei  $f \in \mathcal{C}^3(I \times \mathbb{R}^n), g \in \mathcal{C}^3(\mathbb{R}^n \times \mathbb{R}^n)$  und  $y^* \in B$  eine reguläre Lösung von (52). Dann existieren Konstanten  $H > 0$  und  $R > 0$ , so dass für alle Gitter  $I_N$  mit  $0 < h \leq H$  gilt:

$$\|F_N p_N y^* - p_N^0 F y^*\|_{B_N^0} \leq C \cdot h^2 \quad \text{und} \quad (72)$$

$$\|u^1 - u^2\|_{B_N} \leq S \cdot \|F_N u^1 - F_N u^2\|_{B_N^0}, \quad u^1, u^2 \in S(p_N y^*, R) \quad (73)$$

mit Konstanten  $C > 0, S > 0$ , wobei  $F_N$  der Darstellung (65) oder (66) genügt und  $p_N^0$  gemäß (63) oder (64) definiert ist.

Beide Differenzenverfahren sind also konsistent mit Ordnung 2 und zudem stabil in den Räumen  $B_N$  und  $B_N^0$ . Um die Existenz der diskreten Lösungen  $u^*$  und deren diskrete Konvergenz gegen die exakte Lösung  $y^*$  nachzuweisen, führen wir die folgenden Begriffe ein.

**Definition 31 (Diskrete Konvergenz)**

$y^* \in D \subset B$  und  $u_N^* \in B_N$  seien Lösungen der Probleme  $\mathcal{P}$  bzw.  $\mathcal{P}_N$ .

(i) Die Differenz  $e_N := u_N^* - p_N y^*$  in  $B_N$  nennt man globalen Diskretisierungsfehler des Verfahrens  $\mathcal{P}_N$ .

(ii) Das Verfahren  $\mathcal{P}_N = (B_N, B_N^0, F_N)$  konvergiert diskret gegen  $\mathcal{P} = (B, B^0, F)$ , falls

$$\lim_{N \rightarrow \infty} \|e_N\|_{B_N} = \lim_{N \rightarrow \infty} \|u_N^* - p_N y^*\|_{B_N} = 0 \quad \text{gilt.} \quad (74)$$

(iii)  $\mathcal{P}_N$  konvergiert diskret gegen  $\mathcal{P}$  mit der Ordnung  $p \in \mathbb{N}$ , falls

$$\|u_N^* - p_N y^*\|_{B_N} = \mathcal{O}(N^{-p}), \quad N \in \mathbb{N}, \quad \text{ist.} \quad (75)$$

Beziehung (75) schreibt man dann häufig in der Form  $u_N^* = p_N y^* + \mathcal{O}(N^{-p})$  und sagt, dass  $u_N^*$  mit Ordnung  $p$  gegen  $y^*$  konvergiert. Kann man die Existenz einer Näherungslösung  $u_N^*$  des Problems  $\mathcal{P}_N$  voraussetzen, so lassen sich hinreichende Bedingungen für deren diskrete Konvergenz angeben.

**Satz 32 (Konvergenz)**

Mit der Konstanten  $P$  aus (62) seien folgende Voraussetzungen erfüllt:

(i)  $y^* \in D \subset B$  ist Lösung von  $F(y) = 0$ .

(ii) Für  $N \in \mathbb{N}$  existiert ein  $R > 0$ , so dass die diskrete Gleichung  $F_N(u) = 0$  in der Kugelmenge  $K[p_N y^*; PR]$  eine Lösung  $u_N^*$  besitzt.

(iii)  $F_N$  ist stabil auf  $p_N y^*$  in der Kugel  $K[p_N y^*; PR]$ .

(iv)  $F_N$  ist konsistent mit  $F$  auf  $y^*$ .

Dann konvergiert  $u_N^*$  diskret gegen  $y^*$ .

BEWEIS: Nach Voraussetzung (iii) schätzt man mit den Lösungen  $u_N^*$  und  $y^*$  ab

$$\begin{aligned} \|u_N^* - p_N y^*\|_{B_N} &\leq S \cdot \|F_N(u_N^*) - F_N(p_N y^*)\|_{B_N^0} \\ &= S \cdot \|F_N(p_N y^*) - p_N^0 F(y^*)\|_{B_N^0}, \quad \text{d.h.} \\ \|e_N\|_{B_N} &\leq S \cdot \|\tau\|_{B_N^0}. \end{aligned} \quad (76)$$

Mit Voraussetzung (iv) folgt für  $N \rightarrow \infty$  hieraus die Konvergenz mit  $\|e_N\| \rightarrow 0$ .  $\square$

Die Aussage dieses grundlegenden Satzes der Numerischen Mathematik lässt sich in der griffigen Formel

Konsistenz + Stabilität  $\implies$  Konvergenz

zusammenfassen. Voraussetzung für deren Gültigkeit ist jedoch, dass alle betrachteten Lösungen existieren. Kann man zusätzlich nachweisen, dass der Operator  $F_N$  eine bestimmte Konsistenzordnung  $p \in \mathbb{N}$  hat, so folgt aus Ungleichung (76) mit einer Konstanten  $M > 0$  unmittelbar

$$\|e_N\|_{B_N} \leq S \cdot \|\tau\| \leq S \cdot M \cdot N^{-p},$$

so dass das Verfahren auch dieselbe *Konvergenzordnung*  $p$  besitzt.

Um auch die Frage positiv zu beantworten, ob das diskrete Problem  $\mathcal{P}_N$  für hinreichend großen Parameter  $N$  eine Lösung  $u^*$  besitzt, werden wir die Regularität der exakten Lösung  $y^*$  voraussetzen. Für reguläre Lösungen liefert der folgende Satz neben der Konvergenzaussage auch eine lokale Existenz- und Eindeutigkeitsaussage (vgl. [15]):

**Satz 33** Sei  $f \in \mathcal{C}^3(I \times \mathbb{R}^n)$ ,  $g \in \mathcal{C}^3(\mathbb{R}^n \times \mathbb{R}^n)$  und  $y^* \in B$  eine reguläre Lösung. Dann existieren Konstanten  $H > 0$  und  $R > 0$ , so dass für alle Gitter  $I_N$  mit  $0 < h \leq H$  gilt:

(i) Das finite Gleichungssystem

$$\begin{aligned} g(u_0, u_N) &= 0, & j = 0 \\ \frac{1}{h_j}(u_j - u_{j-1}) - \frac{1}{2}(f(t_j, u_j) + f(t_{j-1}, u_{j-1})) &= 0, & j = 1(1)N \end{aligned} \quad (77)$$

besitzt eine eindeutige Lösung  $u^*$  in der Menge  $\|u - p_N y^*\|_{B_N} \leq R$ .

(ii)  $u^*$  konvergiert diskret gegen  $y^*$  mit  $u^* - p_N y^* = \mathcal{O}(h^2)$ .

(iii) Das Newton-Verfahren konvergiert lokal  $Q$ -quadratisch gegen  $u^*$ .

Analog kann man nachweisen, dass die Behauptungen des Satzes 33 ebenfalls für die implizite Mittelpunkregel (44) erfüllt sind.

**Bemerkung 34** Die hier auf abstraktem Niveau eingeführten Grundbegriffe *Konsistenz*, *diskrete Stabilität* und *Konvergenz* erfassen auch die schon bei den Anfangswertproblemen in [17] benutzten Termini *Konsistenz*, *Nullstabilität* und *Konvergenz*.

**Beispiel 35** Wir lösen das Eigenwertproblem aus Beispiel 5

$$\ddot{x} + g(t, \lambda)x = 0, \quad x(0) = 0, \quad \dot{x}(1) = f(\lambda)x(1), \quad t \in [0, 1] \quad (78)$$

mit den Funktionen  $g(t, \lambda) = \frac{1}{\lambda}(t + 10) - \lambda$ ,  $f(\lambda) = -\lambda$  nach Transformation in die Standardform auf einem äquidistanten Gitter der Größe  $N = 80$ . In Tabelle 2 werden links die mit Algorithmus 21 berechneten Eigenwertnäherungen  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$  dargestellt.  $k_F$  bezeichnet wiederum die Zahl der ausgeführten Newton-Schritte. Die Konvergenzaussage (ii) des Satzes 33 kann wegen höherer Glattheit der Funktionen  $g$  und  $f$  weiter verbessert werden, indem eine *asymptotische Entwicklung*

$$u^* = p_N y^* + p_N e^* h^2 + \mathcal{O}(h^4) \quad (79)$$

$N = 80$				$N = 40$		
$\lambda_{start}$	$\lambda_i^F$	$k_F$	$error$	$\lambda_i^{FG}$	$k_G(k_F)$	$error$
1.60	1.634 857 804	7	8.15e-5	1.634 939 309	7(7)	-1.05e-10
0.40	0.447 051 777	10	2.44e-4	0.447 296 031	11(10)	5.43e-8
0.16	0.168 682 320	8	2.69e-4	0.168 950 987	8(9)	2.46e-7
0.08	0.086 408 504	14	2.72e-4	0.086 680 144	10(12)	5.11e-7

Tabelle 2: Ausgewählte Eigenwerte des Problems (78) mit FDM 2. Ordnung

mit einer von  $h$  unabhängigen Funktion  $e^*$  nachgewiesen wird (vgl. [15]). Ermitteln wir außer der „Feinrechnung“  $u_j^F$  mit  $N$  Teilintervallen der Weite  $h$  nun „Grobwerte“  $u_j^G$  auf  $N/2$  Intervallen der Weite  $2h$ , so können wir nach dem Extrapolationsprinzip (vgl. [17]) mittels (79) den asymptotischen Fehlerschätzer

$$error := \frac{1}{3}(u_j^F - u_j^G) \quad \text{auf dem Grobgitter}$$

gewinnen. Mit seiner Hilfe kann die Feinlösung  $u_j^F$  auf dem Grobgitter weiter zu

$$u_j^{FG} = u_j^F + \frac{1}{3}(u_j^F - u_j^G)$$

mit dem globalen Diskretisierungsfehler  $\mathcal{O}(h^4)$  verbessert werden. In Tabelle 2 sind rechts die mit  $N = 80$  und  $N/2 = 40$  erhaltenen Werte  $u_j^{FG}$  samt ihrem Fehler  $error$  dargestellt. Die Extrapolation führt hier zu einer deutlichen Verbesserung der Näherungswerte.  $\square$

### 4.3 Kollokationsverfahren

Die beiden eingeführten Finite-Differenzenverfahren sind nur von der Konvergenzordnung 2. Bei höherer Genauigkeitsforderung bieten sich *implizite Runge-Kutta-Verfahren* an, die

- eine hohe Konvergenzordnung garantieren,
- gute Stabilitätseigenschaften wie A-Stabilität besitzen können,
- keine speziellen Approximationen in Randnähe erfordern und zudem
- unproblematisch auf nichtäquidistanten Gittern sind.

Da das Randwertproblem (40) inhärent implizit ist, verlieren die für Anfangswertprobleme effizienten expliziten Verfahren ihren Vorteil. Wir setzen deshalb auf dem Gitter

$$I_N = \{ t_i \mid t_i = t_{i-1} + h_i, h_i > 0, i = 1(1)N, t_0 = a, t_N = b \} \quad (80)$$

sofort  $k$ -stufige implizite Runge-Kutta-Verfahren in der Notation

$$u_i = u_{i-1} + h_i \sum_{j=1}^k b_j f_{ij} \quad \text{mit den Steigungswerten} \quad (81)$$

$$f_{ij} = f(t_{i-1} + h_i c_j, u_{i-1} + h_i \sum_{l=1}^k a_{jl} f_{il}), \quad j = 1(1)k$$

an. Das zugehörige Parameterschema mit den *kanonischen Punkten*  $c_j$  liefert die so genannten *Kollokationspunkte*

$$t_{ij} := t_{i-1} + h_i c_j \quad (82)$$

und die Approximationen von  $x(t_{ij})$

$$u_{ij} := u_{i-1} + h_i \sum_{l=1}^k a_{jl} f_{il}, \quad j = 1(1)k.$$

$c_1$	$a_{11}$	$a_{12}$	$\cdots$	$a_{1k}$
$c_2$	$a_{21}$	$a_{22}$	$\cdots$	$a_{2k}$
$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
$c_k$	$a_{k1}$	$a_{k2}$	$\cdots$	$a_{kk}$
	$b_1$	$b_2$	$\cdots$	$b_k$

Mit dem Index  $i$  wird die Abhängigkeit der Werte  $f_{ij} = f(t_{ij}, u_{ij})$  von den  $u_{ij}$  gekennzeichnet. Wir wollen nur solche Runge-Kutta-Verfahren betrachten, die folgende Voraussetzung erfüllen:

### Voraussetzung 36

- (i) Die kanonischen Punkte  $c_j$  genügen der Bedingung  $0 \leq c_1 < c_2 < \dots < c_k \leq 1$ .  
(ii) Die Quadraturformel zu (81) hat die Ordnung  $p$  mit  $p > k$ , d.h.

$$\sum_{l=1}^k b_l c_l^{q-1} = \frac{1}{q}, \quad q = 1(1)p.$$

- (iii) Die Ordnung aller zu den Stufenformeln (81) gehörenden Quadraturformeln sei gleich  $k$ , d.h.

$$\sum_{l=1}^k a_{jl} c_l^{q-1} = \frac{c_j^q}{q}, \quad j = 1(1)k, \quad q = 1(1)k.$$

Für  $q = 1$  sind damit insbesondere die Konsistenzbedingung für Runge-Kutta-Verfahren aus [17] und die Knotenbedingungen

$$\sum_{j=1}^k b_j = 1 \quad \text{und} \quad \sum_{l=1}^s a_{jl} = c_j, \quad j = 1(1)k$$

erfüllt. Unter den Voraussetzungen 36 hat das implizite Runge-Kutta-Verfahren die Ordnung  $p$  (vgl. [3]) und die Quadraturgewichte  $b_j$  und  $a_{jl}$  sind eindeutig bestimmt. Um sie formelmäßig zu erhalten, approximieren wir  $\dot{x}(t) = f(t, x(t))$  auf jedem Teilintervall  $[t_{i-1}, t_i]$  durch das Lagrangesche Interpolationspolynom mit den Knoten  $t_{ij}$ ,  $j = 1(1)k$ . Damit lässt sich die Funktion  $\dot{x}(t)$  als Summe ihrer Lagrange-Interpolierenden vom Grad  $k-1$  und einem Restterm, dem Interpolationsfehler, notieren:

$$\dot{x}(t) = \sum_{l=1}^k \dot{x}(t_{il}) L_l \left( \frac{t - t_{i-1}}{h_i} \right) + R_k(t) \quad \text{mit} \quad t_{i-1} \leq t \leq t_i. \quad (83)$$

Darin sind die auf  $[0, 1]$  normierten Lagrange-Basisfunktionen

$$L_l(s) = \frac{(s - c_1) \cdots (s - c_{l-1})(s - c_{l+1}) \cdots (s - c_k)}{(c_l - c_1) \cdots (c_l - c_{l-1})(c_l - c_{l+1}) \cdots (c_l - c_k)}$$

und der Interpolationsfehler wie in [5, Kapitel 22] definiert:

$$R_k(t) = \dot{x}[t_{i1}, \dots, t_{ik}, t] \prod_{l=1}^k (t - t_{il}).$$



2. *Radau-III-Verfahren*: Der Endpunkt  $c_k = 1$  wird hier festgehalten. In einem  $k$ -stufigen Verfahren ist eine Genauigkeit von  $\mathcal{O}(h^{2k-1})$  zu erreichen (vgl. [1]). Die Parameterschemata der bekanntesten Formeln der Ordnungen 1, 3, 5 haben die folgende Form:

$$\begin{array}{l} \text{Euler/implizit der Ordnung 1:} \end{array} \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \quad \text{Ordnung 3:} \quad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ \hline 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

$$\text{Ordnung 5:} \quad \begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \hline \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ \hline 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

3. *Lobatto-III-Verfahren*: Hierbei werden  $c_1 = 0$  und  $c_k = 1$  gesetzt. Bei einem  $k$ -stufigen Verfahren ist der Fehler von der Ordnung  $\mathcal{O}(h^{2k-2})$ . Das 2-stufige Lobatto-Verfahren ist die Trapezregel (43) mit dem Parameterschema

$$\begin{array}{l} \text{der Ordnung 2:} \end{array} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ \hline 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \text{sowie das Verfahren} \\ \text{der Ordnung 4:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \hline \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ \hline 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Wir wollen nun die Beziehung zu den so genannten Kollokationsverfahren begründen. Mit gegebenem Runge-Kutta-Verfahren (81) seien die Vektoren  $u_i$  für  $i = 1, \dots, N + 1$  sowie  $u_{ij}$  für  $i = 1, \dots, N$  und  $j = 1, \dots, k$  bestimmt. Sei nun  $x_\pi(t)$  ein Polynom vom Grad  $\leq k$  auf  $[t_{i-1}, t_i]$ , das den  $k + 1$  Interpolationsbedingungen

$$x_\pi(t_{i-1}) = u_{i-1} \quad \text{und} \quad \dot{x}_\pi(t_{ij}) = f(t_{ij}, u_{ij}), \quad j = 1(1)k \quad (86)$$

genügt. Dieses Polynom  $x_\pi(t)$  ist mit diesen Eigenschaften wohl definiert. Setzen wir  $x_\pi(t)$  auf dem nächsten Intervall  $[t_i, t_{i+1}]$  fort, so werden die zwei Teilpolynome stetig verbunden. Wiederholung für alle Teilintervalle ergibt eine stetige, stückweise polynomiale Funktion  $x_\pi(t)$  mit maximalem Grad  $k$ , die die Differentialgleichung an den Kollokationspunkten  $t_{ij}$  erfüllt. Damit kommen wir zu folgender

**Definition 38 (Kollokationslösung)** Eine stetige, stückweise polynomiale Funktion  $x_\pi(t)$  heißt Kollokationslösung für das Randwertproblem (2), falls sie stetig und stückweise polynomial ist, wobei der Grad auf jedem Teilintervall  $[t_{i-1}, t_i]$  maximal  $k$  ist und die Gleichungen

$$\dot{x}_\pi(t_{ij}) = f(t_{ij}, x_\pi(t_{ij})), \quad j = 1(1)k \quad \text{und} \quad g(x_\pi(a), x_\pi(b)) = 0 \quad (87)$$

erfüllt sind.

Im Unterschied zu den *Gitterpunkten*  $t_i$  wird gefordert, dass an den *Kollokationspunkten*  $t_{ij}$  die DGL exakt erfüllt ist. Durch unsere Herleitung haben wir die folgende Äquivalenzaussage bewiesen:

**Satz 39 (Äquivalenz des Kollokationsverfahrens)** *Mit Voraussetzung 36 an die Parameter  $(a_{jl}, b_l, c_j)$  ist das Runge-Kutta-Verfahren (81), (82) zuzüglich der Randbedingungen mit den Näherungen  $u_i$  und  $u_{ij}$  äquivalent dem Kollokationsverfahren (86), (87). Es gilt  $x_\pi(t_i) = u_i$ ,  $x_\pi(t_{ij}) = u_{ij}$ ,  $i = 1(1)N$ ,  $j = 1(1)k$ .*

In der Runge-Kutta-Formulierung sind die  $N + 1$  Vektoren  $u_i$ ,  $i = 0(1)N$  simultan auf ganz  $I$  zu bestimmen, weshalb sie auch als *globale Variablen* bezeichnet werden. Die Notation der  $N$  Runge-Kutta-Schritte (81) zusammen mit den Randbedingungen verallgemeinert die Darstellung der Trapezregel (43) in der Form des Gleichungssystems

$$F(u) := \begin{pmatrix} F_0(u_0, u_N) \\ F_1(u_0, u_1) \\ \vdots \\ F_N(u_{N-1}, u_N) \end{pmatrix} = \begin{pmatrix} g(u_0, u_N) \\ \frac{1}{h_1}(u_1 - u_0) - \sum_{j=1}^k b_j f_{1j} \\ \vdots \\ \frac{1}{h_N}(u_N - u_{N-1}) - \sum_{j=1}^k b_j f_{Nj} \end{pmatrix} = 0 \quad (88)$$

für die Gitterfunktion  $u = (u_0, u_1, \dots, u_N)^T$ . Wie bei den einfachen Finite-Differenzenverfahren notieren wir zur Lösung dieses großen Systems  $F(u) = 0$  das Newton-Verfahren in der praktikablen Form

$$F'(u^{(\nu)})(u^{(\nu+1)} - u^{(\nu)}) = -F(u^{(\nu)}), \quad \nu = 0, 1, 2, \dots \quad (89)$$

Die Hilfswerte  $f_{ij}$  sind *lokale Unbekannte* im Teilintervall  $[t_{i-1}, t_i]$  und werden durch einen aufwändigen Kondensationsprozess eliminiert. Im  $\nu$ -ten Newton-Schritt ist schließlich auch hier ein Gleichungssystem mit bidiagonaler Blockstruktur

$$\begin{pmatrix} B_a & 0 & 0 & \cdots & B_b \\ S_1 & R_1 & 0 & \cdots & 0 \\ 0 & S_2 & R_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & S_N & R_N \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix} \quad \text{mit} \quad \begin{cases} B_a := \frac{\partial g(v, w)}{\partial v} \Big|_{v=u_0, w=u_N} \\ B_b := \frac{\partial g(v, w)}{\partial w} \Big|_{v=u_0, w=u_N} \end{cases}$$

zu lösen, wofür Algorithmus 21 mit den dortigen Lösungsansätzen genutzt werden kann. Eine vollständige Verfahrensbeschreibung findet man in [1, S. 222f].

Betrachtet man das Kollokationsverfahren als Einschrittverfahren mit den Gitterpunkten  $t_i$ , so erhält man eine Konvergenz der Ordnung  $p$ , wogegen an den Kollokationspunkten  $t_{ij}$  wegen  $p > k$  lediglich die Ordnung  $k$  erreicht wird. Bei  $p > k + 1$  ergibt sich so eine *Superkonvergenz* an den Gitterpunkten. Wir fassen die Grundaussagen in folgendem Satz aus [1] zusammen:

**Satz 40 (Existenz, Konvergenz)** *Die Lösung  $x^*(t)$  nach Voraussetzung 7 sei regulär und hinreichend glatt. Die Gitterfolge  $\{I_N\}$ ,  $N \in \mathbb{N}$  konvergiere linear und das  $k$ -stufige Kollokationsverfahren erfülle Voraussetzung 36. Dann existieren Konstanten  $\varrho > 0$  und  $H > 0$ , so dass für alle Schrittweiten  $0 < h \leq H$  gilt:*

- (i) In einem Schlauch um  $x^*(t)$  mit Radius  $\varrho > 0$  existiert eine eindeutige Kollokationslösung  $x_\pi(t)$ , die (87) erfüllt.
- (ii)  $x_\pi(t)$  kann mit dem Newton-Verfahren bestimmt werden, das  $Q$ -quadratisch konvergiert, falls die Startnäherung  $x_\pi^0(t)$  hinreichend nahe bei  $x^*(t)$  liegt.
- (iii) Die folgenden Fehlerschätzungen sind erfüllt:

$$\begin{aligned} \|u_i - x^*(t_i)\| &= \mathcal{O}(h^p), & u_i &= x_\pi(t_i), \quad i = 0(1)N \\ \|x_\pi(t) - x^*(t)\| &= \mathcal{O}(h_i^{k+1}) + \mathcal{O}(h^p), & t_{i-1} \leq t \leq t_i, \quad i &= 1(1)N. \end{aligned} \quad (90)$$

**MATLAB** kann parameterabhängige Systeme 1. Ordnung in Standardform

$$\frac{dy}{dx} = f(x, y, p), \quad g(y(a), y(b), p) = 0 \quad (91)$$

mit  $f: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  und  $g: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  lösen.  $p \in \mathbb{R}^m$  ist ein vorgegebener Parametervektor. Der Kollokationscode `bvp4c` implementiert eine dreistufige Lobatto-IIIa-Formel. Die stetig differenzierbare Näherungslösung der Konvergenzordnung 4 wird durch Gitteranpassung und Fehlerkontrolle bis auf eine vorgegebene absolute und relative Genauigkeit `AbsTol` und `RelTol` geliefert. Die Aufruf-Syntax lautet in 3 Varianten:

```
<sol> = bvp4c(<odefun>, <bcfun>, <solinit>)
<sol> = bvp4c(<odefun>, <bcfun>, <solinit>, <options>)
<sol> = bvp4c(<odefun>, <bcfun>, <solinit>, <options>, <p1>, <p2>, ..., <pm>),
```

wobei die syntaktischen Begriffe folgende Bedeutung haben:

- `<odefun>` – Funktionshandle der rechten DGL-Seiten
- `<bcfun>` – Funktionshandle der Randbedingungen
- `<solinit>` – Struktur der Anfangslösung
- `<options>` – Optionsstruktur
- `<p1>, ...` – bekannte Parameter zur Übergabe an `<odefun>` und `<bcfun>`
- `<sol>` – Struktur der Lösung

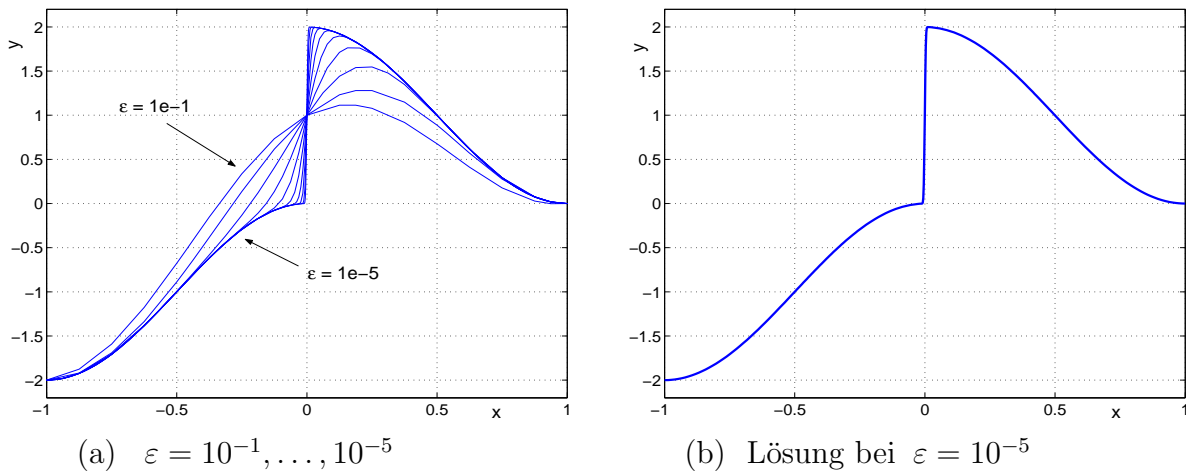
Die Struktur der Anfangslösung `<solinit>` besteht aus dem Feld `solinit.x` der geordneten Knoten des Anfangsgitters (Endpunkte sind `a=solinit.x(1)` und `b=solinit.x(end)`) sowie der Matrix `solinit.y` der Anfangswerte mit Näherungswerten `solinit.y(:,i)` am Knoten `solinit.x(i)`.

Die Struktur der Lösung `<sol>` setzt sich zusammen aus dem Feld `sol.x` der geordneten Knoten des Endgitters, der Matrix `sol.y` der gefundenen Lösungswerte `sol.y(:,i)` am Knoten `sol.x(i)`, der Matrix `sol.ypr` der Approximationen von  $y'(x)$  mit `sol.ypr(:,i)` am Knoten `sol.x(i)` und dem Feld `sol.parameters` der ermittelten unbekannt Parameter, falls solche in der DGL vorhanden sind.

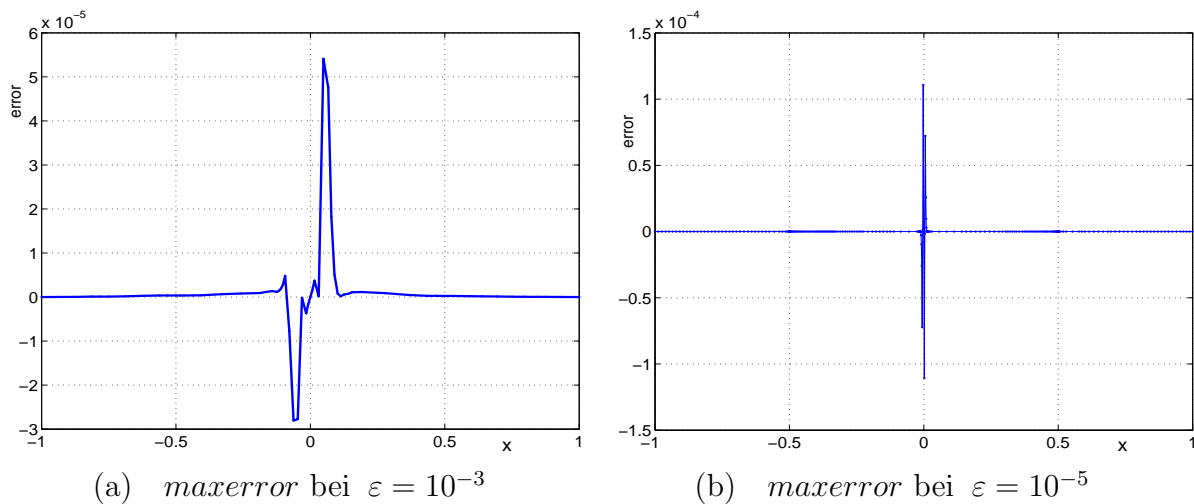
**Beispiel 41** Das lineare Randwertproblem 2. Ordnung für  $y(x)$  aus [11]

$$\varepsilon y'' + xy' = -\varepsilon \pi^2 \cos(\pi x) - \pi x \sin(\pi x), \quad y(-1) = -2, \quad y(1) = 0, \quad \varepsilon > 0 \quad (92)$$

besitzt für  $\varepsilon \rightarrow 0$  eine freie *Grenzschicht* (engl.: *shock layer*) im Intervallinnern. Gesucht ist eine Lösung zu  $\varepsilon = 10^{-5}$ . Die schmale Grenzschicht bei  $x = 0$  stellt hohe Anforderungen

Abbildung 10: Lösungen des Randwertproblems (92) mit `bvp4c`

an Gitterwahl und Startnäherung. Wir führen deshalb eine Lösungsfortsetzung ein: Mit dem moderaten Parameterwert  $\varepsilon_0 = 1.0$  gewinnen wir eine Startnäherung und lösen das Randwertproblem für die Parameterfolge  $(\varepsilon_i)$ ,  $i = 0, 1, 2, \dots$  mit  $\varepsilon_i = \varepsilon_{i-1}/\sqrt{10}$ , wobei die erhaltene Lösung stets als Startnäherung des folgenden  $\varepsilon_i$ -Wertes genutzt wird. Nach 10 Divisionen wird so  $\varepsilon = 10^{-5}$  erreicht.

Abbildung 11: Maximaler absoluter Fehler  $\max |u_i - y(x_i)|$  von (92)

Einführung von  $y_1 = y$ ,  $y_2 = y'$  liefert das parameterabhängige System der Form (91)

$$\begin{aligned} y_1' &= y_2, & y_1(-1) + 2 &= 0 \\ y_2' &= -xy_2/\varepsilon - \pi^2 \cos(\pi x) - \pi x \sin(\pi x)/\varepsilon, & y_1(1) &= 0, \end{aligned}$$

das in [11] als Function-Files `shockode.m` und `shockbc.m` dargestellt wird:

```
function dydx = shockODE(x,y,e)
pix = pi*x;
dydx = [ y(2); -x/e*y(2) - pi^2*cos(pix) - pix/e*sin(pix) ];
```

```
function res = shockBC(ya,yb,e)
res = [ ya(1)+2; yb(1) ];
```

Eine Anfangslösung für `bvp4c` lässt sich mit der Funktion `bvpinit` durch den Aufruf

```
<solinit> = bvpinit(<x>,<y>)
<solinit> = bvpinit(<x>,<y>,<parameters>)
```

gewinnen. Dabei ist `<x>` der Vektor des Anfangsgitters, `<y>(i)` ein konstanter Wert für die  $i$ -te Komponente `<y>(i,:)` der Lösung an allen Gitterpunkten in  $x$ . Mit der Sequenz

```
sol = bvpinit([-1 -0.5 0 0.5 1],[1 0]); e = 1.0;
for i = 1 : 10,
    e = e*3.162277660168379e-001
    sol = bvp4c(@shockODE,@shockBC,sol,[],e);
    plot(sol.x,sol.y(1,:)); hold on
end
```

erhalten wir alle berechneten Lösungen in Abb. 10 (a) und die gesuchte Lösung für den Parameterwert  $\varepsilon = 10^{-5}$  in Abb. 10 (b). Durch Vergleich mit der exakten Lösung

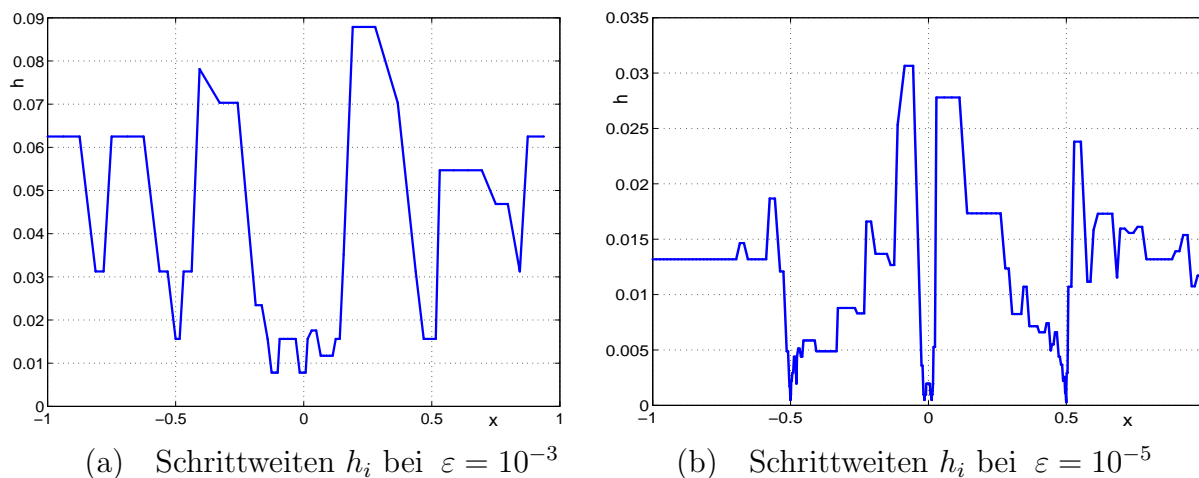


Abbildung 12: Interne Schrittweiten  $h_i$  über  $x$  von (92)

$$y(x) = \cos(\pi x) + \operatorname{erf}(0.5 x \sqrt{2/\varepsilon}) / \operatorname{erf}(0.5 \sqrt{2/\varepsilon})$$

gewinnen wir auf dem erhaltenen Gitter den maximalen absoluten Fehler  $MaxError$ , dessen Verteilung in den Abbildungen 11 (a) und 11 (b) gut erkennbar wird. Die Gitteranpassung durch Einlagerung von Knoten führt auf die internen Endgitter in den Abbildungen 12 (a) und 12 (b).

Die Rechnung mit den Standardtoleranzen  $AbsTol = 10^{-6}$  und  $RelTol = 10^{-3}$  liefert in Tabelle 3 das maximale Residuum  $MaxRes$ . Dieses eignet sich gut zur Fehlerschätzung anstelle des nicht verfügbaren Fehlers  $MaxError$ , der in der 3. Spalte dargestellt wird. Die Größe  $N_{end}$  des Endgitters und die Gesamtzahl von Funktionsaufrufen  $ODEcalls$  der DGL nehmen mit kleinerem  $\varepsilon$  rasch zu. Die Vergleichsrechnung mit erhöhter Genauigkeit  $RelTol = 10^{-6}$  benötigt 31 041 Funktionsaufrufe und ein Gitter mit 712 Punkten.  $\square$

$RelTol$	$\varepsilon$	$MaxRes$	$MaxError$	$N_{end}$	$ODEcalls$
1e-3	1e-1	9.586e-4	2.528e-5	23	188
	1e-3	9.577e-4	5.412e-5	62	1957
	1e-5	9.774e-4	1.106e-4	225	12764
1e-6	1e-1	9.972e-7	1.485e-8	131	2076
	1e-3	9.987e-7	2.804e-8	300	7648
	1e-5	9.925e-7	1.105e-7	712	31041

Tabelle 3: Verfahrensparameter von `bvp4c` für Problem (92)

## 5 Fazit

Randwertprobleme treten in vielen Formen auf, z.B. als *Schwingungsprobleme* mit Periodizitätsbedingungen, als *Eigenwertprobleme* oder als Probleme mit Mehrpunkt- oder Funktionalnebenbedingungen. Durch geeignete Problemerkweiterung bzw. a-priori-Transformation sollte stets die *Standardform für Systeme 1.Ordnung* hergestellt werden, um die dafür existierenden leistungsfähigen numerischen Verfahren einzusetzen.

Mittels Zurückführung des Randwertproblems auf eine Folge von Anfangswertproblemen lassen sich leistungsfähige Anfangswertlöser anwenden. Die so entstehenden *Schießverfahren* reduzieren das unendlich dimensionale Problem auf ein Nullstellenproblem in  $\mathbb{R}^n$ . Die unbekanntes Anfangswerte lassen sich numerisch mittels Newton-ähnlicher Verfahren unter Nutzung des *Variationssystems* oder alternativ dazu durch Differenzenapproximation der Newton-Matrix berechnen. Das *Mehrfach-Schießverfahren* (*auch: Mehrzielmethode*) sollte eingesetzt werden, wenn die Konvergenz des Einfach-Schießverfahrens nicht erreicht wird. Dazu ist eine geeignete Segmentierung des Gesamtintervalles so vorzunehmen, dass die Anfangswertprobleme auf jedem Teilintervall lösbar werden. Die Randbedingungen führen dann zusammen mit Stetigkeitsbedingungen auf ein großdimensionales, strukturiertes Gleichungssystem, das mit einem angepassten stabilen Verfahren gelöst werden muss.

*Finite-Differenzenverfahren (FDM)* gehen unmittelbar von einer diskretisierten Lösung, einer so genannten *Gitterfunktion*, auf einem vorgegebenen  $t$ -Gitter der maximalen Schrittweite  $h$  aus. Die bekanntesten FDM sind die *Trapezregel* und die *implizite Mittelpunkregel*. Anders als bei allgemeineren linearen Mehrschrittverfahren sind bei diesen Einschrittverfahren keine speziellen Approximationen in Randnähe erforderlich. In *Kollokationsverfahren* wird die Lösung auf den Teilintervallen eines vorgegebenen Gitters stückweise durch Polynome des Grades  $k$  approximiert. Deren unbekannte Koeffizienten sind mittels der Randbedingungen und der *Kollokationsbedingungen* zu bestimmen. An speziellen Kollokationspunkten wird dazu gefordert, dass die Ansatzfunktionen die DGL exakt erfüllen. Kollokationsverfahren lassen sich mit Hilfe impliziter Runge-Kutta-Verfahren gewinnen, wobei insbesondere Gauß-Legendre-, Radau-IIA- und Lobatto-IIIA-Formeln wegen ihrer guten Konvergenz- und Stabilitätseigenschaften genutzt werden.

Allgemeine *lineare Eigenwertprobleme* zur Bestimmung reeller *Eigenwerte und Eigenfunktionen* benötigen keine speziellen Lösungsansätze. Durch Einführung von Normierungsbedingungen für die gesuchten Eigenfunktionen können sie auf die Standardform des Zweipunkt-Randwertproblems transformiert werden. Auch allgemeine nichtlineare Verzweigungsprobleme lassen sich mit derartigen Techniken in die Standardform überführen.

## Literatur

- [1] Ascher, U. M.; Mattheij, R. M.; Russell, R. D.: *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, SIAM Philadelphia 1995
- [2] Ascher, U. M.; Petzold, L. R.: *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM Philadelphia 1998
- [3] Hairer, E.; Norsett, S. P.; Wanner, G.: *Solving Ordinary Differential Equations*, Band 1, Springer-Verlag Berlin 1987
- [4] Hermann, M.: *Numerik gewöhnlicher Differentialgleichungen*, Oldenbourg Verlag München 2004
- [5] Hoffmann, A.; Marx, B.; Vogt, W.: *Mathematik für Ingenieure 1. Lineare Algebra, Analysis – Theorie und Numerik*, Pearson Studium München 2005
- [6] Hofmann, W.; Voss, H.: *Shooting Verfahren für nichtlineare Eigenwertprobleme*, ISNM, Band 31, 1976, S. 79-89
- [7] Keller, H. B.: *Numerical Methods for Two-Point Boundary Value Problems*, Blaisdell Publishing Company Waltham 1968
- [8] Keller, H. B.: *Numerical Solution of Two-Point BVP's*, SIAM Publications Philadelphia 1976
- [9] Kuznetsov, Y. A.: *Elements of Applied Bifurcation Theory*, Springer-Verlag New York 1995
- [10] Mangoldt, H. v.; Lösch, F.: *Einführung in die Höhere Mathematik*, Bd. IV, S. Hirzel Verlag Leipzig 1973
- [11] MATLAB – The Language of Technical Computing. Using MATLAB, Version 6, The Math Works Inc., 2000  
<http://www.mathworks.com>
- [12] Perko, L.: *Differential Equations and Dynamical Systems*, Springer-Verlag New York 1996
- [13] Philippow, E. S.; Büntig, W. G.: *Analyse nichtlinearer dynamischer Systeme der Elektrotechnik*, Carl Hanser Verlag München 1992
- [14] Seydel, R.: *Practical Bifurcation and Stability Analysis. From Equilibrium to Chaos*, Springer-Verlag New York 1994
- [15] Stetter, H. J.: *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag Berlin 1973
- [16] Stoer, J.; Bulirsch, R.: *Numerische Mathematik 2*, 3. Auflage, Springer-Verlag Berlin 1990
- [17] Vogt, W.: *Zur Numerik gewöhnlicher Differentialgleichungen*, Inst. f. Mathematik, Preprint No. M 09/02, TU Ilmenau 2002