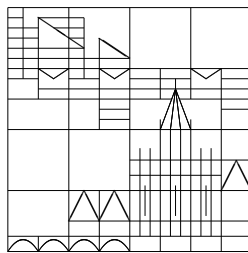


Skript zur Vorlesung

Numerik gewöhnlicher Differentialgleichungen

Sommersemester 2025

Johannes Schropp



Universität Konstanz

Fachbereich Mathematik und Statistik

Stand: 2. Juli 2025

Inhaltsverzeichnis

1	Beispiele und einiges über Differentialgleichungen	3
	a) Beispiele	3
	b) Einführung in die dynamischen Systeme	7
	c) Hamiltonsche dynamische Systeme	10
2	Allgemeine Einschrittverfahren	12
	a) Euler-Cauchy Verfahren	12
	b) Explizite Runge-Kutta Verfahren	14
	c) Implizite Runge-Kutta Formeln	19
	d) Systematische Bestimmung der Konsistenzordnung	23
	e) Klassifikation der impliziten Runge-Kutta Verfahren	25
	f) Das Newtonverfahren	27
3	Stabilität und Konvergenz von Einschrittverfahren	28
4	Asymptotische Entwicklung und Schrittweitensteuerung	31
	a) Verschärfte Konvergenzaussagen	31
	b) Schrittweitensteuerung und Fehlerschätzungen	33
5	Numerik Hamiltonischer Systeme	37
	a) Kollokationsverfahren	37
	b) Symplektische Systeme	38
6	Numerische Verfahren für Randwertaufgaben	45
	a) Erinnerung: Reaktions- Transportphänomene	45
	b) Die Idee des Schießverfahrens	46
7	Das klassische Differenzenverfahren	50
	a) Aufstellung des Verfahrens	50
	b) Allgemeine Zweipunkt-Randwertaufgabe	52
	c) Inversmonotone Matrizen und inversmonotone Randwertprobleme	53
	d) Nichtlineare Gleichungssysteme	55
	e) Anwendung auf Randwertaufgaben	58

8	Konvergenz des klassischen Differenzenverfahrens	60
	a) Konsistenz des Verfahrens	60
	b) Stabilität des klassischen Differenzen-Verfahrens	62

1. Beispiele und einiges über Differentialgleichungen

a) Beispiele

1.1 Beispiel. Fallschirmspringer

Das Grundgesetz der Mechanik lautet:

$$\text{Kraft} = \text{Masse} \cdot \text{Beschleunigung} \quad (\text{Newton})$$

$$K = m \cdot \ddot{x} \quad (1-1)$$

Durch Angabe der Abhängigkeiten der Kraft

$$K := K(x) \text{ oder } K := K(t, x, \dot{x}) \quad (1-2)$$

wird nun aus der Gleichung (1-1) eine Differentialgleichung, nämlich

$$m\ddot{x} = K(t, x, \dot{x}).$$

Beim Fallschirmspringer sieht das Kraftgesetz (1-2) so aus:

$$K = K(t, \dot{x}) = mg - k(t)\dot{x}^2,$$

wobei g die Erdbeschleunigung und m die Masse des Fallschirmspringers bezeichnet. Die Funktion $k(t)$ charakterisiert den Luftwiderstand

$$k(t) := \begin{cases} \Delta_0 & \text{für } t \leq \tau_0, \\ \frac{\Delta_1 - \Delta_0}{\tau_1 - \tau_0}(t - \tau_0) + \Delta_0 & \text{für } \tau_0 \leq t \leq \tau_1, \\ \Delta_1 & \text{für } t \geq \tau_1. \end{cases}$$

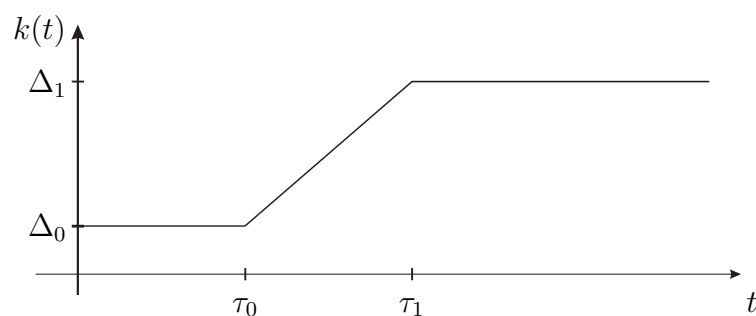


Abbildung 1: Beispiel für $k(t)$

Für unser Modell können wir die Reduktion der Ordnung der Gleichung durchführen. Setze dazu $v = \dot{x}$, $\dot{v} = \ddot{x}$, dann

$$m\dot{v} = mg - k(t)v^2 = K(t, v), \quad v(0) = v_0.$$

1.2 Beispiel. Das Dreikörperproblem

Problem: Beschreibe die Bewegung dreier Körper, welche sich aufgrund der Gravitation gegenseitig anziehen (als spezielles Beispiel können Sonne-Erde-Mond dienen).

Wir bezeichnen die Masse des i -ten Körpers mit m_i für $i = 1, 2, 3$. Der Vektor $q^i = (q_1^i(t), q_2^i(t), q_3^i(t)) \in \mathbb{R}^3$ bezeichnet die Position des i -ten Körpers zur Zeit t , $i = 1, 2, 3$. Die Bewegungsgleichung für drei Massenpunkte in einem Potential $U := U(q^1, q^2, q^3)$ lautet:

$$m_i \ddot{q}^i = -\frac{\partial U}{\partial q^i}(q^1, q^2, q^3), \quad i = 1, 2, 3. \quad (1-3)$$

Gemäß des Gravitationsgesetzes finden wir

$$U = U(q^1, q^2, q^3) = \frac{m_1 m_2}{\|q^1 - q^2\|_2} + \frac{m_2 m_3}{\|q^2 - q^3\|_2} + \frac{m_1 m_3}{\|q^1 - q^3\|_2}.$$

Dabei ist (1-3) ein System von 9 Gleichungen 2-ter Ordnung. Die entsprechenden Anfangswertbedingungen sind $q^i(0) = q_{i0}$, $\dot{q}^i(0) = v_{i0}$, $i = 1, 2, 3$. Jetzt können wir (1-3) auf ein System erster Ordnung reduzieren. Setze $p^i = m_i \dot{q}^i$, $i = 1, 2, 3$, dann

$$\begin{aligned} \dot{q}^i &= \frac{1}{m_i} p^i, \\ \dot{p}^i &= m_i \ddot{q}^i = -\frac{\partial U}{\partial q^i}(q^1, q^2, q^3), \\ q^i(0) &= q_{i0}, \\ p^i(0) &= p_{i0} := m_i v_{i0}, \quad i = 1, 2, 3. \end{aligned} \quad (1-4)$$

Das reduzierte System (1-4) besteht also aus 18 Gleichungen erster Ordnung.

Das System (1-4) hat eine Besonderheit. Sei

$$H = H(q^1, q^2, q^3, p^1, p^2, p^3) = \frac{1}{2} \sum_{i=1}^3 \frac{1}{m_i} \langle p^i, p^i \rangle + U(q^1, q^2, q^3).$$

Dann finden wir

$$\begin{aligned} \frac{\partial H}{\partial p^i} &= \frac{1}{m_i} p^i \\ \frac{\partial H}{\partial q^i} &= \frac{\partial U}{\partial q^i}, \quad i = 1, 2, 3, \end{aligned} \quad (1-5)$$

d.h. mit $q = (q^1, q^2, q^3) \in \mathbb{R}^9$, $p = (p^1, p^2, p^3) \in \mathbb{R}^9$ hat das System (1-4) die Gestalt

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p}(q, p), & q(0) &= q_0 = (q_{10}, q_{20}, q_{30}) \\ \dot{p} &= -\frac{\partial H}{\partial q}(q, p), & p(0) &= p_0 = (p_{10}, p_{20}, p_{30}) \end{aligned} \quad (1-6)$$

Man nennt (1-6) ein Hamiltonsches System und H die Hamiltonfunktion des Systems.

Bemerkung 1. Eine C^2 -Hamiltonfunktion H eines Systems der Form (1-6) ist längs Lösungen $(\bar{q}(t), \bar{p}(t))$, $t \in I$ von (1-6) konstant. Es gilt

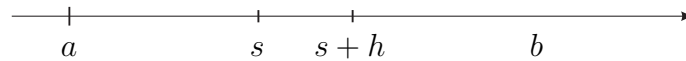
$$\frac{d}{dt}H(\bar{q}(t), \bar{p}(t)) = \frac{\partial H}{\partial q} \dot{\bar{q}}(t) + \frac{\partial H}{\partial p} \dot{\bar{p}}(t) = \frac{\partial H}{\partial q} \frac{\partial H}{\partial p} + \frac{\partial H}{\partial p} \left(-\frac{\partial H}{\partial q} \right) = 0 \quad t \in I.$$

Bemerkung 2. Physikalisch stellt H die Gesamtenergie dar, denn es gilt

$$H = \frac{1}{2} \sum_{i=1}^3 \frac{1}{m_i} \langle p_i, p_i \rangle + U(q^1, q^2, q^3) = \underbrace{\frac{1}{2} \sum_{i=1}^3 m_i \langle \dot{q}_i, \dot{q}_i \rangle}_{=E_{kin}} + \underbrace{U(q^1, q^2, q^3)}_{=E_{pot}} = E_{ges}$$

1.3 Beispiel. Ein Reaktions-Transport Problem

Sei $c = c(t, s)$ die Konzentration einer Substanz im Raumpunkt $s \in [a, b]$ zur Zeit $t \geq 0$. Die Funktion $J = J(t, s)$ beschreibe den Fluss der Konzentration c am Raumpunkt s und Zeit t . Schließlich werde die Substanz c kreiert oder verbraucht; $f = f(c)$ beschreibe diesen Vorgang.



Fluss $J(t, s)$ bzw. $J(t, s+h)$

Kreation $f(c(t, s))$ bzw. $f(c(t, s+h))$

Abbildung 2: Fluss- und Reaktionsbilanzen

Sei $h > 0$. Dann ist $\int_s^{s+h} c(t, \sigma) d\sigma$ die Masse im Raumstück $[s, s+h]$ zur Zeit t . Entsprechend ist $\int_s^{s+h} f(c(t, \sigma)) d\sigma$ die entstehende bzw. vergehende Masse pro Zeiteinheit im Raumstück $[s, s+h]$. Wir erhalten nun die folgende Massenbilanzgleichung in $[s, s+h]$:

$$\underbrace{\frac{d}{dt} \int_s^{s+h} c(t, \sigma) d\sigma}_{\text{Gesamtmassenveränderung}} = \underbrace{J(t, s) - J(t, s+h)}_{\text{Massenveränderung durch Transport}} + \underbrace{\int_s^{s+h} f(c(t, \sigma)) d\sigma}_{\text{Massenveränderung durch Kreation\Verbrauch}}$$

Dividieren durch h liefert

$$\frac{1}{h} \int_s^{s+h} \frac{\partial}{\partial t} c(t, \sigma) d\sigma = -\frac{J(t, s+h) - J(t, s)}{h} + \frac{1}{h} \int_s^{s+h} f(c(t, \sigma)) d\sigma.$$

Dabei ist zu beachten, dass die Vertauschung von Ableitung und Integral nur bei hinreichend glattem c erlaubt ist, z.B. $c \in C^1$. Mit Hilfe des Mittelwertsatzes der Integralrechnung

$$\int_a^b f(x) dx = f(x_0)(b-a), \quad x_0 \in [a, b]$$

liefert der Grenzübergang $h \rightarrow 0$ für

$$\frac{1}{h} \frac{\partial}{\partial t} c(t, s_h) h = - \frac{J(t, s+h) - J(t, s)}{h} + \frac{1}{h} f(c(t, \hat{s}_h)) h, \quad s_h, \hat{s}_h \in [s, s+h]$$

die Differentialgleichung

$$\underbrace{\frac{\partial}{\partial t} c(t, s)}_{\text{zeitliche Veränderung}} = \underbrace{-\frac{\partial}{\partial s} J(t, s)}_{\text{negativer Flussgradient}} + f(c(t, s)), \quad a \leq s \leq b, \quad t \geq 0.$$

Die Ausgestaltung von J bestimmt die Transportmöglichkeiten:

konvektiver Fluss

$$J(t, s) = v(s)c(t, s);$$

diffusiver Fluss

$$J(t, s) = -D(s) \frac{\partial}{\partial s} c(t, s); \quad D(s) > 0$$

allgemein

$$J(t, s) = J_{konv}(t, s) + J_{diff}(t, s) = v(s)c(t, s) - D(s) \frac{\partial}{\partial s} c(t, s).$$

Somit erhalten wir

$$\frac{\partial}{\partial t} c(t, s) = -\frac{\partial}{\partial s} (v(s)c(t, s)) + \frac{\partial}{\partial s} \left[D(s) \frac{\partial}{\partial s} c(t, s) \right] + f(c(t, s)) \quad (1-7)$$

für $a \leq s \leq b, \quad t \geq 0$.

Um die Eindeutigkeit der Lösung zu sichern, müssen Zusatzbedingungen gestellt werden:

Anfangsbedingungen

$$c(0, s) = c_0(s), \quad a \leq s \leq b.$$

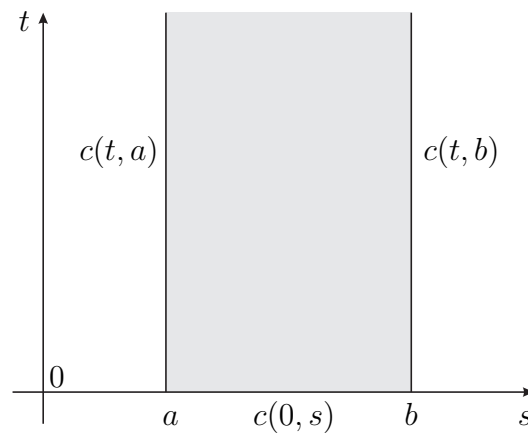
Randbedingungen

Hier sind folgende Bedingungen möglich:

- **Dirichlet-Randbedingungen**

$$c(t, a) = \gamma_a,$$

$$c(t, b) = \gamma_b.$$

Abbildung 3: Zusatzbedingungen für c

- Neumann-Randbedingungen (z.B. bei $s = b$)

$$\begin{aligned} c(t, a) &= \gamma_a, \\ \frac{\partial}{\partial s} c(t, b) &= \gamma_b. \end{aligned}$$

- allgemeiner: Zwei-Punkt-Randbedingungen

$$\alpha_\sigma c(t, \sigma) + \beta_\sigma \frac{\partial}{\partial s} c(t, \sigma) = \gamma_\sigma, \quad \sigma = a, b.$$

Das Problem (1-7) zusammen mit Anfangs- und Randbedingungen heißt *Anfangs-Randwertaufgabe*. Die Sonderlösungen der Reaktions-Transportgleichung sind:

Raumunabhängige Lösungen $c(t, s) = \gamma(t)$ führen zu einem **Anfangswertproblem**

$$(1-7) \xrightarrow{c(t,s)=\gamma(t)} \begin{cases} \gamma'(t) = f(\gamma(t)) \\ \gamma(0) = c_0 \end{cases}$$

Zeitunabhängige Lösungen $c(t, s) = u(s)$ führen zu einer **Randwertaufgabe zweiter Ordnung**

$$(1-7) \xrightarrow{c(t,s)=u(s)} \begin{cases} -(v(s)u(s))' + (D(s)u'(s))' + f(u(s)) = 0, & a \leq s \leq b, \\ \alpha_\sigma u(\sigma) + \beta_\sigma u'(\sigma) = \gamma_\sigma, & \sigma = a, b. \end{cases}$$

b) Einführung in die dynamischen Systeme

Vorgelegt sei die Aufgabe

$$\begin{aligned} x'(t) &= f(x(t)), \\ x(t_0) &= x_0 \end{aligned} \tag{1-8}$$

mit $f \in C^1(\Omega, \mathbb{R}^N)$, $\Omega \subset \mathbb{R}^N$ offen. Wie wir aus der Theorie gewöhnlicher Differentialgleichungen wissen (Satz von Picard-Lindelöf, Fortsetzungssatz), hat die Anfangswertaufgabe (1-8) eine eindeutige Lösung $\bar{x}(t)$ für t aus einem offenen maximalen Existenzintervall $J = J(x_0) \ni t_0$. Um die Abhängigkeit der Lösung von x_0 zu verdeutlichen, schreibt man auch $\bar{x}(t, t_0, x_0)$, $t \in J(x_0)$. Bei einem autonomen System kommt es nicht auf die Wahl des Zeitpunktes t_0 an, sondern nur auf die verschobene Zeit $t - t_0$. Es gilt

$$\bar{x}(t, t_0, x_0) = \bar{x}(t - t_0, 0, x_0), \quad t \in J(x_0)$$

denn $\bar{y}(t) := \bar{x}(t - t_0, 0, x_0)$ erfüllt

$$\begin{aligned} \dot{\bar{y}}(t) &= f(\bar{y}(t)), \\ \bar{y}(t_0) &= x_0, \end{aligned}$$

d.h. $\bar{y}(t) = \bar{x}(t)$ nach Existenz- und Eindeutigkeitssatz. Ohne Einschränkung wählt man deshalb $t_0 = 0$ und schreibt $\bar{x} = \bar{x}(t, x_0)$, $t \in J(x_0) \ni 0$.

Für eine allgemeine nichtautonome Anfangswertaufgabe $x'(t) = f(t, x(t))$, $x(t_0) = x_0$ schreibt man aber weiterhin $\bar{x}(t, t_0, x_0)$.

Man kann sich die Lösungen von Differentialgleichungen auf zwei verschiedene Weisen veranschaulichen: in Zeitbild und in Phasenbild.

- Im Zeitbild wird der Graph der Lösung $(t, \bar{x}(t, x_0))$, $t \in J(x_0)$ gezeichnet.

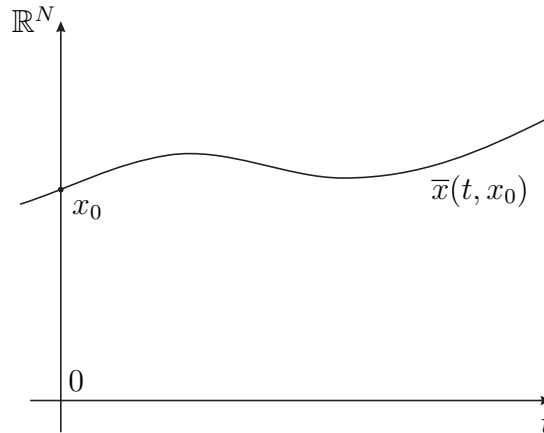


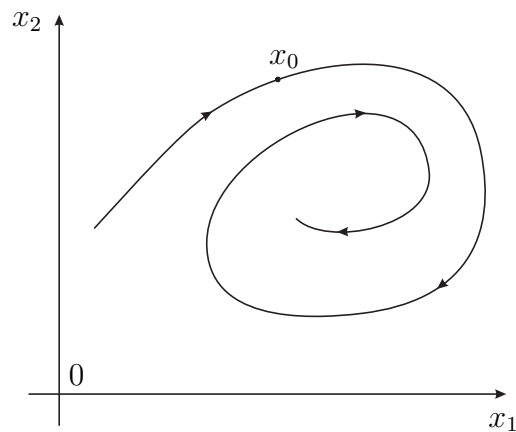
Abbildung 4: Zeitbild

- Im Phasenbild wird der zu x_0 gehörige Orbit, d.h. das Bild

$$\gamma(x_0) = \{\bar{x}(t, x_0) \mid t \in J(x_0)\}$$

der von der Lösung durchlaufenen Kurve in $\mathbb{R}^n$ dargestellt.

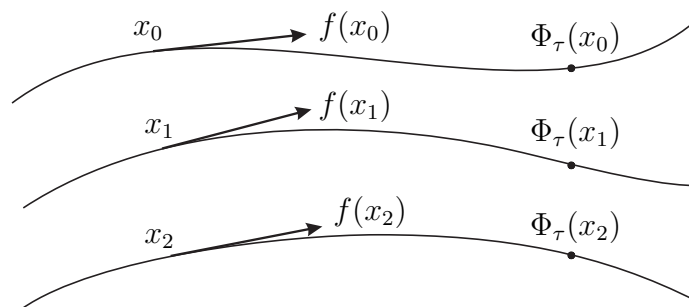
Die positive Zeitrichtung wird daher durch einen Pfad angedeutet. Ein Phasenbild wird umso aussagekräftiger, je mehr Orbits eingezeichnet werden. Dabei sind zwei Orbits $\gamma(x_0)$ und $\gamma(\hat{x}_0)$ entweder identisch oder disjunkt.

Abbildung 5: Phasenbild: $N = 2$

Man betrachtet auch oft für $\tau \in \mathbb{R}$ fest die Abbildung

$$\Phi_\tau : x_0 \mapsto \bar{x}(\tau, x_0),$$

den so genannten τ -Fluss.

Abbildung 6: Phasenbild: $N = 2$

Der τ -Fluss hat die Halbgruppeneigenschaft:

$$\Phi_t \circ \Phi_s = \Phi_{t+s}$$

für alle $t, s \in \mathbb{R}$, für die die linke Seite wohldefiniert ist. Allgemein bezeichnet man eine Familie von stetigen Abbildungen

$$\Phi_t : \mathbb{R}^N \rightarrow \mathbb{R}^N, \quad t \in \mathbb{R}$$

als ein globales dynamisches System (globaler Fluss) auf $\mathbb{R}^n$, wenn $\Phi_t \circ \Phi_s = \Phi_{t+s}$, $\forall t, s \in \mathbb{R}$ und $\Phi_0 = \text{id}_{\mathbb{R}}$ gilt.

Die Abhängigkeit der Lösung von Anfangswerten und Parametern beschreibt der folgende

1.4 Satz. Sei $\Omega \subset \mathbb{R}^N$ offen, $\Lambda \subset \mathbb{R}^p$ offen und sei $f \in C^k(\Omega \times \Lambda, \mathbb{R}^N)$, $k \geq 1$. Dann besitzt die AWA

$$\begin{aligned} x' &= f(x, \lambda) \\ x(0) &= \xi \end{aligned}$$

für jedes $(\xi, \lambda) \in \Omega \times \Lambda$ genau eine nicht fortsetzbare Lösung $\bar{x}(t, \xi, \lambda) \in \Omega$ für $t \in J(\xi, \lambda) =]t^-(\xi, \lambda), t^+(\xi, \lambda)[$. Der Definitionsbereich von $\bar{x}$

$$\mathcal{D}(\Lambda) = \{(t, \xi, \lambda) \in \mathbb{R} \times \Omega \times \Lambda, t \in J(\xi, \lambda)\}$$

ist offen, und es gilt $\bar{x} \in C^k(\mathcal{D}(\Lambda), \Omega)$.

Sei $\bar{x}(t, x_0)$, $t \in J(x_0)$ die Lösung von

$$\begin{aligned} x'(t) &= f(x), \\ x(0) &= x_0 \end{aligned}$$

mit $f \in C^1$. Bezeichne durch $\gamma(x_0) = \{\bar{x}(t, x_0), t \in J(x_0)\}$ den zu x_0 gehörigen Orbit. Die Größe $\partial \bar{x}(t, x_0) / \partial x_0$, $t \in J(x_0)$ bezeichnet man als die *Linearisierung* von $\bar{x}(t, x_0)$. Es gilt gemäß Satz 1.4

$$\frac{\partial}{\partial t} \left[\frac{\partial}{\partial x_0} \bar{x}(t, x_0) \right] = \frac{\partial}{\partial x_0} \frac{\partial}{\partial t} \bar{x}(t, x_0) = \frac{\partial}{\partial x_0} f(\bar{x}(t, x_0)) = \underbrace{Df(\bar{x}(t, x_0))}_{\in \mathbb{R}^{N,N}} \left[\frac{\partial}{\partial x_0} \bar{x}(t, x_0) \right]$$

und

$$\frac{\partial}{\partial x_0} \bar{x}(0, x_0) = \frac{\partial}{\partial x_0} x_0 = I_N.$$

Die Funktion $\frac{\partial}{\partial x_0} \bar{x}(\cdot, x_0) : J \rightarrow \mathbb{R}^{N,N}$ genügt also der linearen AWA

$$\begin{aligned} X'(t) &= Df(\bar{x}(t, x_0)) X(t), \\ X(0) &= I_N, \end{aligned}$$

d.h. $\frac{\partial}{\partial x_0} \bar{x}(\cdot, x_0)$ ist die *Hauptfundamentalmatrix* von $y' = Df(\bar{x}(t, x_0))y$.

c) Hamiltonsche dynamische Systeme

Es sei $\Omega \subset \mathbb{R}^{2N}$ offen und $H \in C^k(\Omega, \mathbb{R})$, $k \geq 2$. Mit $H = H(q, p)$ betrachten wir die AWA:

$$\begin{aligned} q'(t) &= \frac{\partial H}{\partial p}(q, p), & q(0) &= q_0, \\ p'(t) &= -\frac{\partial H}{\partial q}(q, p), & p(0) &= p_0. \end{aligned} \tag{1-9}$$

Die Lösung sei $\bar{x}(t, x_0)$, $x_0 = (q_0, p_0)$.

1.5 Bemerkung. Es gilt

$$H(\bar{x}(t, x_0)) = H(q_0, p_0) \quad \forall t \in J((q_0, p_0)).$$

1.6 Definition. Eine Abbildung $g \in C^1(\Omega, \mathbb{R}^{2N})$ heißt symplektisch, falls gilt

$$Dg(q, p)^\top J Dg(q, p) = J, \quad (q, p) \in \Omega$$

mit

$$J = \begin{pmatrix} 0 & I_N \\ -I_N & 0 \end{pmatrix} \in \mathbb{R}^{2N, 2N}.$$

1.7 Satz. (*Poincaré*)

Sei $H \in C^2(\Omega, \mathbb{R})$, und $\bar{x}(t, x_0)$, $x_0 = (q_0, p_0)$ sei der Lösungsfluss von (1-9). Dann ist für jedes feste $t \in \mathbb{R}$ der Fluss $\bar{x}(t, \cdot) : \Omega \rightarrow \mathbb{R}^{2N}$ eine symplektische Abbildung.

1.8 Bemerkung. Eine symplektische Abbildung erhält das Volumen, d.h.

$$\begin{aligned} \text{Vol}(K) &= \text{Vol}(\bar{x}(t, K)), \\ \bar{x}(t, K) &:= \{\bar{x}(t, x_0), x_0 \in K\}. \end{aligned}$$

2. Allgemeine Einschrittverfahren

a) Euler-Cauchy Verfahren

Betrachte die AWA

$$\begin{aligned} u'(t) &= f(t, u(t)) \\ u(t_0) &= \alpha, \end{aligned} \tag{2-1}$$

mit $t_0 \leq t \leq t_{\text{end}}$ und $f \in C^1([t_0, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N)$. Wir nehmen an, dass eine Lösung $\bar{u}(t, t_0, \alpha)$ für $t \in [t_0, t_{\text{end}}]$ existiert.

Zur numerischen Berechnung von $\bar{u}(t, t_0, \alpha)$, $t_0 \leq t \leq t_{\text{end}}$ wählen wir eine Schrittweite $h > 0$ und ein äquidistantes Gitter

$$\Omega_h = \{t_j = t_0 + jh, j = 0, \dots, \sigma(h)\}$$

mit $\sigma : [0, h_0[\rightarrow \mathbb{N}$ definiert durch

$$t_0 + (\sigma(h) - 1)h < t_{\text{end}} \leq t_0 + (\sigma(h))h.$$

Optimalerweise wählt man $h > 0$ in der Form $h := (t_{\text{end}} - t_0)/N$ für ein $N \in \mathbb{N}$.

Ein allgemeines Einschrittverfahren für (2-1) hat die Form

$$\begin{aligned} u(t_0) &= \alpha \\ h^{-1}(u(t_{j+1}) - u(t_j)) &= V(h, t_j, u(t_j)), \quad j = 0, \dots, \sigma(h) - 1 \end{aligned} \tag{2-2}$$

mit $V :]0, h_0] \times [t_0, t_{\text{end}}] \times \mathbb{R}^N \rightarrow \mathbb{R}^N$, $h_0 > 0$. Die Funktion V heißt Verfahrensfunktion des Einschrittverfahrens.

Die einfachste Wahl ist $V(h, t, u) = f(t, u)$. Man erhält dann das Euler-Cauchy Verfahren.

Die Lösung des Gleichungssystems (2-2) ist durch Rekursion gegeben:

$$\begin{aligned} u(t_0) &= \alpha \\ u(t_{j+1}) &= u(t_j) + hV(h, t_j, u(t_j)), \quad j = 0, \dots, \sigma(h) - 1. \end{aligned}$$

Statt eines äquidistanten Gitters können wir auch ein nicht äquidistantes Gitter betrachten

$$\Omega = \{t_0 < t_1 < \dots < t_\sigma = t_{\text{end}}\}$$

mit $t_{j+1} = t_j + h_j$, $h = (h_0, h_1, \dots, h_{\sigma-1})$. Dann lautet das Gleichungssystem (2-2)

$$\begin{aligned} u(t_0) &= \alpha \\ h_j^{-1}(u(t_{j+1}) - u(t_j)) &= V(h_j, t_j, u(t_j)), \quad j = 0, \dots, \sigma - 1. \end{aligned}$$

Wenn wir im Folgenden spezielle Verfahrensfunktionen angehen, so ist klar, wie bei nicht äquidistanten Gittern zu rechnen ist. Deshalb sei $h > 0$ o.E. konstant.

Wir schreiben die Rekursion (2-2) als Gleichungssystem $T^h(u) = 0$. Definiere dazu

$$\Omega_h = \{t_0 + jh \mid j = 0, \dots, \sigma(h)\}$$

und

$$w = (\underbrace{w(t_0)}_{\in \mathbb{R}^N}, \underbrace{w(t_1)}_{\in \mathbb{R}^N}, \dots, \underbrace{w(t_{\sigma(h)})}_{\in \mathbb{R}^N}) \in (\mathbb{R}^N)^{\Omega_h}.$$

Jetzt lässt sich $T^h : (\mathbb{R}^N)^{\Omega_h} \rightarrow (\mathbb{R}^N)^{\Omega_h}$ wie folgt definieren:

$$T^h(u) = (\underbrace{u(t_0) - \alpha}_{\in \mathbb{R}^N}, \underbrace{h^{-1}(u(t_{j+1}) - u(t_j)) - V(h, t_j, u(t_j))}_{\in \mathbb{R}^N}), \quad j = 0, \dots, \sigma(h) - 1).$$

Als Norm auf $(\mathbb{R}^N)^{\Omega_h}$ wählt man die klassische Maximumsnorm

$$\|u\|_\infty := \max\{|u_i(t_j)| \mid i = 1, \dots, N, j = 0, \dots, \sigma(h)\}$$

Wir bezeichnen mit $\bar{u}_h = \bar{u}(t, t_0, \alpha)|_{\Omega_h}$ die Restriktion der Lösung $\bar{u}(t, t_0, \alpha)$, $t \in [t_0, t_{end}]$ der AWA auf das Gitter Ω_h . Ferner definiere durch

$$\text{def}(u) := T^h(u), \quad u \in (\mathbb{R}^N)^{\Omega_h}$$

den Defektvektor. Der Defektvektor $\text{def}(\bar{u}_h) = T(\bar{u}_h)$ für $\bar{u}_h$ nennt man den Konsistenzfehler.

2.1 Definition. Das numerische Modell $T^h(u) = 0$ für $u \in (\mathbb{R}^N)^{\Omega_h}$ nennt man **W-konsistent**, falls für jede Lösung $\bar{u} \in W$ von der AWA gilt

$$\|T^h(\bar{u}_h)\|_\infty \rightarrow 0 \text{ für } h \rightarrow 0.$$

Ist sogar $\|T^h(\bar{u}_h)\|_\infty = O(h^p)$, so heißt das Modell **W-konsistent** der Ordnung p .

2.2 Definition. Das Modell $T^h(u) = 0$ heißt **W-konvergent** (der Ordnung p), falls es zu jeder Lösung $\bar{u} \in W$ der AWA ein $h_0 > 0$ gibt, so dass $T^h(u) = 0$, $0 < h \leq h_0$ eine Lösung $u^h \in (\mathbb{R}^N)^{\Omega_h}$ besitzt mit

$$\|\bar{u}_h - u^h\|_\infty \rightarrow 0 \quad (\|\bar{u}_h - u^h\|_\infty = O(h^p)).$$

2.3 Definition. Das Modell $T^h(u) = 0$, $u \in (\mathbb{R}^N)^{\Omega_h}$ heißt **stabil** bzgl. h , falls ein $h_0 > 0$ und ein $C > 0$ (unabhängig von h) existieren mit

$$\|u - v\|_\infty \leq C \|T^h(u) - T^h(v)\|_\infty \quad \forall u, v \in (\mathbb{R}^N)^{\Omega_h}, \quad 0 < h \leq h_0$$

2.4 Korollar. Sei das Modell W -konsistent (der Ordnung p) und stabil, und es existieren Lösungen $u^h \in (\mathbb{R}^N)^{\Omega_h}$ von $T^h(u) = 0$. Dann ist das Modell auch W -konvergent (der Ordnung p).

Beweis: Setze $u = u^h$ und $v = \bar{u}_h$ in die Stabilitätsungleichung ein und finde

$$\|u^h - \bar{u}_h\|_\infty \leq C \underbrace{\|T^h(u^h) - T^h(\bar{u}_h)\|_\infty}_{=0} = C \|T^h(\bar{u}_h)\|_\infty \rightarrow 0 \quad (= O(h^p))$$

□

SLOGAN:

Konsistenz (der Ordnung p) + Stabilität $\Rightarrow$ Konvergenz (der Ordnung p).

Die Aufstellung spezieller Einschrittverfahren geschieht in der Regel anhand des Konsistenzfehlers, d.h. man versucht V so zu bestimmen, dass die Ordnung p möglichst groß wird.

Wir untersuchen den Konsistenzfehler des Euler-Cauchy-Verfahrens. Schreibe $\bar{u}(t_j)$ für $\bar{u}(t_j, t_0, \alpha)$. Dann gilt

$$\begin{aligned} \|T^h(\bar{u}_h)\|_\infty &= \|\underbrace{\bar{u}(t_0) - \alpha}_{=0}, (h^{-1}(\bar{u}(t_{j+1}) - \bar{u}(t_j)) - \underbrace{f(t_j, \bar{u}(t_j))}_{\bar{u}'(t_j)})\|_\infty, \quad j = 0, \dots, \sigma(h) - 1 \\ &= \max_{j=0, \dots, \sigma(h)-1} |h^{-1}(\bar{u}(t_{j+1}) - \bar{u}(t_j)) - \bar{u}'(t_j)| \\ &= \max_{j=0, \dots, \sigma(h)-1} \left| \frac{h}{2} \bar{u}''(\eta_{t_j}) \right| \leq \frac{1}{2} h \|\bar{u}''(t)\|_\infty, \end{aligned}$$

falls $\bar{u} \in C^2$.

2.5 Satz. Das Euler-Cauchy Verfahren ist C^2 -konsistent der Ordnung 1.

b) Explizite Runge-Kutta Verfahren

2.6 Definition. Sei $s \in \mathbb{N}$ und sei $u_m = u(t_m)$, $t_m = t_0 + mh$. Ein Einschrittverfahren der Gestalt

$$u_{m+1} = u_m + h \underbrace{\sum_{j=1}^s b_j f(t_m + c_j h, U^j(h, t_m, u_m))}_{V(h, t_m, u_m)},$$

$$u_0 = \alpha$$

mit (Darstellung auf Basis der Stufenwerte)

$$U^1(h, t_m, u_m) = u_m, \quad c_1 = 0$$

$$U^i(h, t_m, u_m) = u_m + h \sum_{j=1}^{i-1} a_{ij} f(t_m + c_j h, U^j(h, t_m, u_m)) \quad i = 2, 3, \dots, s$$

heißt explizites s -stufiges Runge-Kutta Verfahren. Dabei bestimmen die geeignet gewählten Parameter a_{ij} , b_j , c_j das Verfahren vollständig.

Eine bequeme Darstellung liefert das sogenannte Runge-Kutta Tableau

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ c_3 & a_{31} & a_{32} & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss-1} \\ \hline & b_1 & b_2 & \dots & b_{s-1} & b_s \end{array}$$

Äquivalente Formulierung liefert die Darstellung auf Basis der Steigungswerte:

$$u_{m+1} = u_m + h \sum_{j=1}^s b_j K^j(h, t_m, u_m)$$

$$u_0 = \alpha$$

mit

$$K^1(h, t_m, u_m) = f(t_m, u_m), \quad c_1 = 0$$

$$K^i(h, t_m, u_m) = f(t_m + c_i h, u_m + h \sum_{j=1}^{i-1} a_{ij} K^j(h, t_m, u_m)), \quad i = 2, 3, \dots, s.$$

Dem Euler-Cauchy Verfahren ($s = 1$) entsprechen die Werte $b_1 = 1$, $c_1 = 0$. Wir betrachten nun den Fall $s = 2$, d.h. das Tableau

$$\begin{array}{c|cc} c_2 & a_{21} & \\ \hline & b_1 & b_2 \end{array}.$$

Sei $\bar{u}$ die Lösung, so gilt für $\bar{u} \in C^3$, $t = t_j = t_0 + jh$,

$$\bar{u}(t+h) = \bar{u}(t) + h\bar{u}'(t) + \bar{u}''(t)\frac{h^2}{2} + \frac{1}{6}\bar{u}'''(\eta_t)h^3, \quad \eta_t \in [t, t+h],$$

$$h^{-1}(\bar{u}(t+h) - \bar{u}(t)) = \bar{u}'(t) + \bar{u}''(t)\frac{h}{2} + \underbrace{\frac{h^2}{6}\bar{u}'''(\eta_t)}_{=: R_1(t) \text{ mit } \|R_1(t)\|_\infty \leq c_1 h^2}.$$

Um die Verfahrensfunktion zu bekommen, drücken wir $\bar{u}$ und seine Ableitungen mit

Hilfe von f aus:

$$\begin{aligned}
\bar{u}'(t) &= f(t, \bar{u}(t)), \\
\bar{u}''(t) &= \frac{d}{dt} \bar{u}'(t) = \frac{d}{dt} f(t, \bar{u}(t)) = \frac{\partial}{\partial t} f(t, \bar{u}(t)) + \frac{\partial}{\partial u} f(t, \bar{u}(t)) \underbrace{\bar{u}'(t)}_{f(t, \bar{u}(t))} \\
&= \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)), \\
\bar{u}'''(t) &= \frac{d}{dt} \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) = \left(\frac{\partial^2 f}{\partial t^2} + \frac{\partial^2 f}{\partial t \partial u} f + f \frac{d}{dt} \frac{\partial f}{\partial u} + \frac{\partial f}{\partial u} \frac{d}{dt} f \right) (t, \bar{u}(t)) \\
&= \left(\frac{\partial^2 f}{\partial t^2} + \frac{\partial^2 f}{\partial t \partial u} f + f \frac{\partial^2 f}{\partial u \partial t} + \frac{\partial^2 f}{\partial u^2} f^2 + \frac{\partial f}{\partial u} \frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) \\
&= \left(\frac{\partial^2 f}{\partial t^2} + 2 \frac{\partial^2 f}{\partial t \partial u} f + \frac{\partial^2 f}{\partial u^2} f^2 + \frac{\partial f}{\partial u} \frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)).
\end{aligned}$$

Sei nun

$\|f\|_{2,\infty} = \max \{ \|\cdot\|_\infty\text{-Normen aller partiellen Ableitungen von } f \text{ bis zur Ordnung } 2 \},$

dann gilt die Abschätzung

$$\max \{ \|u'''(t)\|_\infty, t_0 \leq t \leq t_{\text{end}} \} \leq C \|f\|_{2,\infty}.$$

Wir finden also

$$h^{-1}(\bar{u}(t+h) - \bar{u}(t)) = f(t, \bar{u}(t)) + \frac{h}{2} \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) + O(h^2).$$

Drücke nun V durch f aus:

$$\begin{aligned}
V(h, t, \bar{u}(t)) &= b_1 f(t, \bar{u}(t)) + b_2 f(t + c_2 h, \bar{u}(t) + h a_{21} f(t, \bar{u}(t))), \\
V(h, t, \bar{u}(t)) &= V(0, t, \bar{u}(t)) + h \frac{\partial}{\partial h} V(0, t, \bar{u}(t)) + O(h^2),
\end{aligned}$$

d.h.

$$\begin{aligned}
V(0, t, \bar{u}(t)) &= (b_1 + b_2) f(t, \bar{u}(t)), \\
\frac{\partial}{\partial h} V(0, t, \bar{u}(t)) &= b_2 \left(\frac{\partial f}{\partial t} c_2 + \frac{\partial f}{\partial u} a_{21} f \right) (t, \bar{u}(t)).
\end{aligned}$$

Wir erhalten dann

$$V(h, t, \bar{u}(t)) = (b_1 + b_2) f(t, \bar{u}(t)) + h b_2 \left(\frac{\partial f}{\partial t} c_2 + \frac{\partial f}{\partial u} a_{21} f \right) (t, \bar{u}(t)) + O(h^2).$$

Insgesamt folgt damit für den Konsistenzfehler

$$\begin{aligned}
 h^{-1}(\bar{u}(t+h) - \bar{u}(t)) - V(h, t, \bar{u}(t)) &= f(t, \bar{u}(t)) + \frac{h}{2} \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) \\
 &\quad - (b_1 + b_2) f(t, \bar{u}(t)) \\
 &\quad - hb_2 \left(\frac{\partial f}{\partial t} c_2 + \frac{\partial f}{\partial u} a_{21} f \right) (t, \bar{u}(t)) + O(h^2) \\
 &= (1 - b_1 - b_2) f(t, \bar{u}(t)) \\
 &\quad + \frac{h}{2} \left((1 - 2b_2 c_2) \frac{\partial f}{\partial t} + (1 - 2b_2 a_{21}) \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) \\
 &\quad + O(h^2).
 \end{aligned}$$

Als hinreichende Bedingungen für Konsistenz der Ordnung 2 erhalten wir folglich

$$\begin{aligned}
 1 &= b_1 + b_2 \\
 1 &= 2b_2 c_2 \\
 1 &= 2a_{21} b_2
 \end{aligned}$$

mit der Lösungsschar

$$\begin{aligned}
 b_1 &= 1 - b_2, \\
 c_2 &= \frac{1}{2b_2}, \\
 a_{21} &= \frac{1}{2b_2}, \quad b_2 \neq 0.
 \end{aligned}$$

Als spezielle Lösungen bekommt man

- $b_1 = b_2 = \frac{1}{2}$, $c_2 = a_{21} = 1$. Das ist das sog. verbesserte Euler-Cauchy Verfahren oder auch Verfahren von Heun.

Iterationsvorschrift:

$$\begin{aligned}
 U^1(h, t_m, u_m) &= u_m \\
 U^2(h, t_m, u_m) &= u_m + hf(t_m, u_m) \\
 V(h, t_m, u_m) &= \frac{1}{2} f(t_m, u_m) + \frac{1}{2} f(t_m + h, u_m + hf(t_m, u_m))
 \end{aligned}$$

Tableau:

$$\begin{array}{c|c}
 1 & 1 \\
 \hline
 & \frac{1}{2} \quad \frac{1}{2}
 \end{array}$$

- $b_2 = 1$, $b_1 = 0$, $c_2 = a_{21} = \frac{1}{2}$. Das ist das sog. verbesserte Polygonzug Verfahren.

Iterationsvorschrift:

$$U^1(h, t_m, u_m) = u_m$$

$$U^2(h, t_m, u_m) = u_m + \frac{h}{2} f(t_m, u_m)$$

$$V(h, t_m, u_m) = f\left(t_m + \frac{h}{2}, u_m + \frac{h}{2} f(t_m, u_m)\right)$$

Tableau:

$\frac{1}{2}$	$\frac{1}{2}$
0	1

Die beiden Verfahren besitzen die Konsistenzordnung 2, falls $\bar{u} \in C^3$. Als weitere Formeln sind zu erwähnen:

- klassisches Runge-Kutta Verfahren ($s = 4$):

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Dieses Verfahren besitzt die Konsistenzordnung 4, falls $\bar{u} \in C^5$ gilt.

- Verfahren von Lawson ($s = 6$):

0					
$\frac{1}{2}$	$\frac{1}{2}$				
$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{16}$			
$\frac{1}{2}$	0	0	$\frac{1}{2}$		
$\frac{3}{4}$	0	$-\frac{3}{16}$	$\frac{6}{16}$	$\frac{9}{16}$	
1	$\frac{1}{7}$	$\frac{4}{7}$	$\frac{6}{7}$	$-\frac{12}{7}$	$\frac{8}{7}$
	$\frac{7}{90}$	0	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$
				$\frac{7}{90}$	

Dieses Verfahren besitzt die Konsistenzordnung 5, falls $\bar{u} \in C^6$ gilt.

Bemerkung 3. Alle angegebenen Verfahren erfüllen die Bedingung $\sum_{i=1}^s b_i = 1$.

c) Implizite Runge-Kutta Formeln

2.7 Definition. Ein s -stufiges implizites Runge-Kutta Verfahren für die AWA

$$\begin{aligned} u'(t) &= f(t, u(t)), \\ u(t_0) &= \alpha \end{aligned}$$

ist gegeben durch

$$\begin{aligned} u(t_0) &= \alpha, \\ u(t_{n+1}) &= u(t_n) + h \underbrace{\sum_{i=1}^s b_i f(t_n + c_i h, U^i(h, t_n, u(t_n)))}_{V(h, t_n, u(t_n))}, \quad n = 0, \dots, \sigma(h) - 1 \end{aligned}$$

mit

$$U^i(h, t_n, u(t_n)) = u(t_n) + h \sum_{j=1}^s a_{ij} f(t_n + c_j h, U^j(h, t_n, u(t_n))), \quad i = 1, \dots, s$$

für $t_n = t_0 + nh$, $n = 0, \dots, \sigma(h)$. Die Stufenwerte $U^i(h, t_n, u(t_n))$, $i = 1, \dots, s$ sind nun implizit gegeben.

Die äquivalente Steigungsdarstellung lautet:

$$\begin{aligned} u(t_0) &= \alpha, \\ u(t_{n+1}) &= u(t_n) + h \sum_{i=1}^s b_i K^i(h, t_n, u(t_n)), \quad n = 0, \dots, \sigma(h) - 1, \\ K^i(h, t_n, u(t_n)) &= f(t_n + c_i h, u(t_n) + h \sum_{j=1}^s a_{ij} K^j(h, t_n, u(t_n))), \quad i = 1, \dots, s. \end{aligned}$$

Die Tableau-Beschreibung:

$$\begin{array}{c|cccc} c_1 & a_{11} & \dots & \dots & a_{1s} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ c_s & a_{s1} & \dots & \dots & a_{ss} \\ \hline & b_1 & \dots & \dots & b_s \end{array}$$

bzw. kürzer

$$\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$$

mit $b, c \in \mathbb{R}^s$, $A \in \mathbb{R}^{s,s}$.

Bemerkung 4.

- Die expliziten Runge-Kutta Verfahren erhalten wir also aus dem allgemeinen Ansatz, falls $a_{ij} = 0$ für $j \geq i$.
- Ein Nachteil dieser Vorgehensweise ist, dass an jedem Punkt $(t_n, u(t_n))$ ein nichtlineares Gleichungssystem zu lösen ist.
- Ein Vorteil des allgemeinen Zugangs besteht etwa darin, dass sich wegen der zusätzlichen freien Parameter mit kleinerer Stufe s eine hohe Konsistenzordnung erreichen lässt.

Wir betrachten zuerst den einfachsten Fall $s = 1$:

$$\begin{aligned} V(h, t, u) &= b_1 f(t + c_1 h, U^1(h, t, u)), \\ U^1(h, t, u) &= u + h a_{11} f(t + c_1 h, U^1(h, t, u)). \end{aligned}$$

Wir finden

$$\begin{aligned} U^1(0, t, \bar{u}(t)) &= \bar{u}(t), \\ U^1(h, t, \bar{u}(t)) &= \bar{u}(t) + O(h) \end{aligned}$$

und somit

$$\begin{aligned} f(t + c_1 h, U^1(h, t, \bar{u}(t))) &= f(t, \bar{u}(t)) \\ &+ h \left[\frac{\partial f}{\partial t}(t, \bar{u}(t)) c_1 + \frac{\partial f}{\partial u}(t, \bar{u}(t)) \frac{\partial U^1}{\partial h}(0, t, \bar{u}(t)) \right] + O(h^2). \end{aligned}$$

Nun bestimmen wir $\frac{\partial U^1}{\partial h}(0, t, \bar{u}(t))$:

$$\begin{aligned} \frac{\partial U^1}{\partial h}(h, t, u) &= \frac{\partial}{\partial h} (u + h a_{11} f(t + c_1 h, U^1(h, t, u))) \\ &= a_{11} f(t + c_1 h, U^1(h, t, u)) + h a_{11} \frac{\partial}{\partial h} f(t + c_1 h, U^1(h, t, u)), \\ &\Rightarrow \frac{\partial U^1}{\partial h}(0, t, \bar{u}(t)) = a_{11} f(t, \bar{u}(t)). \end{aligned}$$

Dies liefert

$$\begin{aligned} f(t + c_1 h, U^1(h, t, \bar{u}(t))) &= f(t, \bar{u}(t)) \\ &+ h \left[\frac{\partial f}{\partial t}(t, \bar{u}(t)) c_1 + \frac{\partial f}{\partial u}(t, \bar{u}(t)) a_{11} f(t, \bar{u}(t)) \right] + O(h^2). \end{aligned}$$

Andererseits gilt

$$h^{-1}(\bar{u}(t + h) - \bar{u}(t)) = f(t, \bar{u}(t)) + \frac{h}{2} \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) + O(h^2).$$

Also erhalten wir für den Konsistenzfehler

$$\begin{aligned}
 h^{-1}(\bar{u}(t+h) - \bar{u}(t)) - b_1 f(t + c_1 h, U^1(h, t, \bar{u}(t))) &= f(t, \bar{u}(t)) - b_1 f(t, \bar{u}(t)) + O(h^2) \\
 &+ \frac{h}{2} \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) \\
 &- h b_1 \left(c_1 \frac{\partial f}{\partial t} + a_{11} \frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) \\
 &= (1 - b_1) f(t, \bar{u}(t)) \\
 &+ \frac{h}{2} (1 - 2b_1 c_1) \frac{\partial f}{\partial t} (t, \bar{u}(t)) \\
 &+ \frac{h}{2} (1 - 2b_1 a_{11}) \left(\frac{\partial f}{\partial u} f \right) (t, \bar{u}(t)) \\
 &+ O(h^2).
 \end{aligned}$$

Das Verfahren ist dann von Konsistenzordnung 2, falls gilt

$$\begin{cases} 1 - b_1 = 0 \\ 1 - 2b_1 c_1 = 0 \\ 1 - 2b_1 a_{11} = 0 \\ \bar{u} \in C^3 \end{cases} \Rightarrow \begin{cases} b_1 = 1 \\ c_1 = 1/2 \\ a_{11} = 1/2 \\ \bar{u} \in C^3 \end{cases}$$

Das ist die Gauß-Form erster Stufe (Konsistenzordnung 2, falls $\bar{u} \in C^3$):

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Ein weiteres Verfahren der Ordnung 4 ist durch die Gauß-Form der Stufe 2 gegeben

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{3+\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Jetzt betrachten wir die Auflösbarkeit der Systeme der Form

$$\begin{aligned}
 U^1 &= u + h \sum_{j=1}^s a_{1j} f(t + c_j h, U^j) \\
 U^2 &= u + h \sum_{j=1}^s a_{2j} f(t + c_j h, U^j) \\
 &\dots \\
 U^s &= u + h \sum_{j=1}^s a_{sj} f(t + c_j h, U^j)
 \end{aligned} \tag{2-3}$$

für $(t, u) \in [t_0, t_{\text{end}}] \times \mathbb{R}^N$. Eine hinreichende Bedingung dafür liefert

2.8 Satz. Sei f Lipschitz-stetig mit einer Konstanten L_1 gleichmäßig in t , d.h.

$$\|f(t, v) - f(t, w)\|_\infty \leq L_1 \|v - w\|_\infty, \quad t_0 \leq t \leq t_{\text{end}} \quad v, w \in \mathbb{R}^N.$$

Dann besitzt das Gleichungssystem (2-3) für jedes Paar $(t, u) \in [t_0, t_{\text{end}}] \times \mathbb{R}^N$ und jede Schrittweite $h > 0$ mit

$$q := hL_1 \|A\|_\infty < 1$$

genau eine Lösung

$$U(h, t, u) = (U^1(h, t, u), U^2(h, t, u), \dots, U^s(h, t, u)) \in \mathbb{R}^{sN}.$$

Dabei versteht man unter $\|\cdot\|_\infty$ die Zeilensummennorm:

$$\|A\|_\infty := \max\left\{\sum_{j=1}^s |a_{ij}|, i = 1, \dots, s\right\}$$

Beweis: Wir zeigen, dass die Abbildung $T : \mathbb{R}^{Ns} \rightarrow \mathbb{R}^{Ns}$

$$T(U) = T(U^1, U^2, \dots, U^s) := \begin{pmatrix} u + h \sum_{j=1}^s a_{1j} f(t + c_j h, U^j) \\ \vdots \\ u + h \sum_{j=1}^s a_{sj} f(t + c_j h, U^j) \end{pmatrix}$$

eine Kontraktion ist. Dann folgt die Behauptung aus dem BANACHschen Fixpunktsatz. Es gilt

$$\begin{aligned} \|T(U) - T(\tilde{U})\|_\infty &= \left\| \begin{pmatrix} h \sum_{j=1}^s a_{1j} (f(t + c_j h, U^j) - f(t + c_j h, \tilde{U}^j)) \\ \vdots \\ h \sum_{j=1}^s a_{sj} (f(t + c_j h, U^j) - f(t + c_j h, \tilde{U}^j)) \end{pmatrix} \right\|_\infty \\ &= h \left\| (A \otimes I_N) \begin{pmatrix} f(t + c_1 h, U^1) - f(t + c_1 h, \tilde{U}^1) \\ \vdots \\ f(t + c_s h, U^s) - f(t + c_s h, \tilde{U}^s) \end{pmatrix} \right\|_\infty \\ &\leq h \underbrace{\|A \otimes I_N\|_\infty}_{=\|A\|_\infty} \max\{|f(t + c_i h, U^i) - f(t + c_i h, \tilde{U}^i)|, i = 1, \dots, s\} \\ &\leq h \|A\|_\infty L_1 \max\{|U^i - \tilde{U}^i|, i = 1, \dots, s\} \\ &= q \|U - \tilde{U}\|_\infty. \end{aligned}$$

□

Zur Lösung des Gleichungssystems (2-3) steht z.B. das Newtonverfahren zur Verfügung. Im speziellen Fall des Gleichungssystems (2-3) liefert uns aber auch der Fixpunktsatz eine Lösungsmethode.

2.9 Satz. (Fixpunktsatz) Es sei D eine nichtleere abgeschlossene Teilmenge eines Banachraumes B . Der Operator T sei auf D erklärt und bilde D in sich ab, d.h. $T(D) \subset D$. Er genüge in D einer Lipschitzbedingung mit einer Konstanten $q < 1$

$$\|T(x) - T(y)\| \leq q\|x - y\|, \quad x, y \in D.$$

Dann hat die Gleichung $x = T(x)$ in D genau eine Lösung $\bar{x}$. Bildet man ausgehend von $x_0 \in D$ die Folge $x_{n+1} = T(x_n)$, so gilt

$$\|\bar{x} - x_n\| \leq \frac{q^n}{1 - q} \|x_1 - x_0\|.$$

2.10 Korollar. Es gelten die Voraussetzungen von Satz 2.8. Dann konvergiert die Folge $U^{(n+1)} = T(U^{(n)})$ für $U^{(0)} \in \mathbb{R}^{Ns}$ beliebig gegen die Lösung $\bar{U}$ von $T(U) = U$, und es gilt

$$\|\bar{U} - U^{(n)}\|_\infty \leq \frac{q^n}{1 - q} \|U^1 - U^0\|_\infty$$

Bemerkung 5. Da sich

$$q = hL_1\|A\|_\infty$$

wie $O(h)$ verhält, folgt $q^n = O(h^n)$ und wir gewinnen mit jedem Iterationsschritt eine Ordnung von h . Daher empfiehlt es sich zumindest für kleine Schrittweiten h , das Iterationsverfahren so oft auszuführen, wie die Ordnung des Runge-Kutta Verfahrens angibt. Als Startwert nimmt man etwa $U^{(0)} = (u, u, \dots, u)$, denn $U^{(0)}$ löst das System (2-3) für $h = 0$.

d) Systematische Bestimmung der Konsistenzordnung

Vorgelegt sei

$$\begin{aligned} u'(t) &= f(t, u(t)), \\ u(t_0) &= \alpha. \end{aligned}$$

mit Lösung $\bar{u}(t, t_0, \alpha)$. Das implizite Runge-Kutta Verfahren lautet:

$$\begin{aligned} u_0 &= \alpha, \\ u_{n+1} &= u_n + h \sum_{i=1}^s b_i f(t_n + c_i h, U^i), \\ U^i &= u_n + h \sum_{j=1}^s a_{ij} f(t_n + c_j h, U^j), \quad i = 1, \dots, s \end{aligned}$$

mit dem Koeffizienten-Tableau

$$\begin{array}{c|c} c & A \\ \hline & b^\top. \end{array}$$

Gesucht sind systematische Bedingungen an A , b , c , so dass das zugehörige Runge-Kutta Verfahren eine gewisse vorgegebene Konsistenzordnung besitzt. Ein wesentliches Hilfsmittel stellen die vereinfachenden Bedingungen von Butcher dar:

$$\begin{aligned} B(p) : \sum_{i=1}^s b_i c_i^{k-1} &= \frac{1}{k}, \quad k = 1, \dots, p, \\ C(q) : \sum_{j=1}^s a_{ij} c_j^{k-1} &= \frac{1}{k} c_i^k, \quad i = 1, \dots, s, \quad k = 1, \dots, q, \\ D(m) : \sum_{i=1}^s b_i c_i^{k-1} a_{ij} &= \frac{1}{k} b_j (1 - c_j^k), \quad j = 1, \dots, s \quad k = 1, \dots, m. \end{aligned}$$

Die Bedingung $B(p)$ lässt sich auf folgende Weise interpretieren. Betrachte das spezielle Problem

$$u'(t) = f(t), \quad u(t_m) = 0. \quad (2-4)$$

Dann gilt mit $\tau = t_m + \vartheta(t - t_m)$, $d\tau = (t - t_m)d\vartheta$ sofort

$$\bar{u}(t, t_m, 0) = \underbrace{u(t_m)}_{=0} + \int_{t_m}^t f(\tau) d\tau = \int_0^1 f(t_m + \vartheta(t - t_m))(t - t_m) d\vartheta.$$

Für $t = t_m + h$ bekommt man

$$\bar{u}(t_m + h, t_m, 0) = h \int_0^1 f(t_m + \vartheta h) d\vartheta.$$

Wende ein s -stufiges Runge-Kutta Verfahren auf (2-4) an und finde:

$$u_{m+1} = \underbrace{u_m}_{=0} + h \sum_{i=1}^s b_i f(t_m + c_i h) = h \sum_{i=1}^s b_i f(t_m + c_i h).$$

Sei nun speziell $f(t) = (t - t_m)^{k-1}$. Dann finden wir

$$\bar{u}(t_m + h, t_m, 0) = h \int_0^1 (t_m + \vartheta h - t_m)^{k-1} d\vartheta = h^k \int_0^1 \vartheta^{k-1} d\vartheta = \frac{h^k}{k}.$$

Entsprechend liefert das RKV

$$u_{m+1} = h \sum_{i=1}^s b_i (t_m + c_i h - t_m)^{k-1} = h^k \sum_{i=1}^s b_i c_i^{k-1}.$$

Nun folgt

$$\bar{u}(t_m + h, t_m, 0) - u_{m+1} = h^k \left(\frac{1}{k} - \sum_{i=1}^s b_i c_i^{k-1} \right).$$

Die Bedingung $B(p)$ bedeutet daher, dass die dem RKV zugrunde liegende Quadraturmethode (Quadraturformel)

$$\int_0^1 g(t) dt = \sum_{i=1}^s b_i g(c_i) + R(g), \quad g \in C([0, 1])$$

für Polynome q mit $\deg(q) \leq p-1$ exakt ist (d.h. Quadraturfehler $R(q) \equiv 0$). Dabei sind b_i , $i = 1, \dots, s$ die Gewichte und c_i , $i = 1, \dots, s$ die Knoten (Stützstellen) der Quadraturformel. Analog lässt sich zeigen, dass die Bedingung $C(q)$ Folgendes bedeutet

$$\bar{u}(t_0 + c_i h, t_0, \alpha) - U^i(h, t_0, \alpha) = O(h^{q+1}), \quad i = 1, \dots, s,$$

d.h. die Stufenwerte $U^i(h, t_0, \alpha)$ approximieren die Lösungen zum Zeitpunkt $t_0 + c_i h$ bis auf $O(h^{q+1})$.

2.11 Satz. (*Butcher, ohne Beweis*)

Genügen die Koeffizienten b_i , c_i , a_{ij} eines s -stufigen Runge-Kutta Verfahren den vereinfachenden Bedingungen $B(p)$, $C(q)$ und $D(m)$ mit

$$\begin{aligned} p &\leq q + m + 1, \\ p &\leq 2q + 2 \end{aligned}$$

so besitzt das RKV die Konsistenzordnung p .

2.12 Satz. (*Butcher, ohne Beweis*)

Die Konsistenzordnung eines s -stufigen Runge-Kutta Verfahren kann $p = 2s$ nicht überschreiten.

e) Klassifikation der impliziten Runge-Kutta Verfahren

Wir betrachten zwei wichtige Klassen von impliziten Runge-Kutta Verfahren.

a) Gauß-Verfahren

Diese Verfahren beruhen auf der GAUSS'schen Quadraturmethode. Die Knoten c_i erfüllen $c_i \in [0, 1]$ und sind symmetrisch: $c_{s+1-i} = 1 - c_i$, $i = 1, \dots, s$. Die entsprechenden Methoden erfüllen $B(2s)$, $C(s)$, $D(s)$. Der Satz 2.11 impliziert

die Konsistenzordnung $p = 2s$.

$s = 1$	$\frac{1}{2}$	$\frac{1}{2}$
$p = 2$	1	
$s = 2$	$\frac{3-\sqrt{3}}{6}$	$\frac{1}{4}$
$p = 4$	$\frac{3+\sqrt{3}}{6}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

b) Radau-Verfahren

Diese Verfahren beruhen auf der RADAU-Quadraturmethode. Es gilt $c_i \in [0, 1]$, $i = 1, \dots, s$. Es gibt die linksseitige ($c_1 = 0$) und die rechtsseitige ($c_s = 1$) Quadraturmethode. Die zugehörigen Verfahren werden als Radau IA und Radau IIA bezeichnet. Dabei erfüllen

- Radau IA Verfahren die Bedingungen $B(2s - 1)$, $C(s - 1)$ und $D(s)$;
- Radau IIA Verfahren die Bedingungen $B(2s - 1)$, $C(s)$ und $D(s - 1)$.

Daraus folgt die Konsistenzordnung $2s - 1$, d.h. Polynome bis zum Grad $2s - 2$ werden exakt integriert.

Radau IA:

$s = 1$	0	1	
$p = 1$		1	
	0	$\frac{1}{4}$	$-\frac{1}{4}$
$s = 2$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
$p = 3$		$\frac{1}{4}$	$\frac{3}{4}$

Radau IIA:

$s = 1$	1	1	implizit Euler-Cauchy
$p = 1$		1	
	$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
$s = 2$	1	$\frac{3}{4}$	$\frac{1}{4}$
$p = 3$		$\frac{3}{4}$	$\frac{1}{4}$

f) Das Newtonverfahren

Wir betrachten ein nichtlineares Gleichungssystem $T(u) = 0$ für $T : U \rightarrow \mathbb{R}^N$ mit $U \subset \mathbb{R}^N$ offen und $T \in C^1(U, \mathbb{R}^N)$. Das Newtonverfahren besteht in der Iterationsvorschrift

$$u^0 \in U, \quad u^{n+1} := u^n - (T'(u^n))^{-1}T(u^n), \quad n = 0, 1, \dots,$$

falls diese durchführbar ist.

Bemerkung 6. Bei der praktischen Durchführung löst man das Gleichungssystem und invertiert niemals die Matrix:

$$u_0 \in U, \quad T'(u^n) d^n = -T(u^n), \quad u^{n+1} := u^n + d^n \quad n = 0, 1, \dots$$

Das Newtonverfahren wird im Erfolgsfall abgebrochen an einer Iterierten u^{n_0} , welche $\|T(u^{n_0})\|_\infty < \varepsilon$, $\varepsilon > 0$ vorgegeben, erfüllt. Für einen Rechner mit einer Arithmetik von 16 Stellen ist z.B. je nach Problem $\varepsilon = 10^{-j}$, $j = 8, 10, 12, 14$ vernünftig.

Satz 2.1. (lokaler Konvergenzsatz, ohne Beweis)

Sei $U \subset \mathbb{R}^N$ offen und sei $T : U \rightarrow \mathbb{R}^N$ zweimal stetig differenzierbar. Ferner existiere eine Lösung $\bar{u} \in U$ von $T(u) = 0$ und $T'(\bar{u})$ sei invertierbar.

Dann gibt es eine Kugel $K_\rho(\bar{u}) = \{u \in \mathbb{R}^N \mid \|u - \bar{u}\|_\infty \leq \rho\} \subset U$, $\rho > 0$, so dass $T(u) = 0$ keine weitere Lösung in $K_\rho(\bar{u})$ besitzt. Für jedes $u^0 \in K_\rho(\bar{u})$ existiert die Newtonfolge

$$u^{n+1} := u^n - (T'(u^n))^{-1}T(u^n), \quad n = 0, 1, \dots,$$

die in $K_\rho(\bar{u})$ liegt und gegen $\bar{u}$ konvergiert. Überdies gibt es ein $C > 0$ mit

$$\|u^{n+1} - \bar{u}\|_\infty \leq C \|u^n - \bar{u}\|_\infty^2 \quad \forall n \geq 0, u^0 \in K_\rho(\bar{u}). \quad (2-5)$$

Formelzeile (2-5) besagt gerade, dass das Newtonverfahren lokal quadratisch konvergiert.

Bemerkung 7. Lösungen $\bar{u}$ von $T(u) = 0$ mit $T'(\bar{u})$ invertierbar heißen auch reguläre Lösungen.

SLOGAN: Das Newton-Verfahren konvergiert lokal quadratisch an einer regulären Lösung.

3. Stabilität und Konvergenz von Einschrittverfahren

Vorgelegt sei die AWA

$$\begin{aligned} u'(t) &= f(t, u(t)) \\ u(t_0) &= \alpha \end{aligned}$$

mit $f \in C^1([t, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N)$. Es existiere eine Lösung $\bar{u}(t) \in C^1([t_0, t_{\text{end}}], \mathbb{R}^N)$. Zur Lösung dieser AWA betrachten wir wie in Kapitel 2 ein allgemeines Einschrittverfahren der Form

$$\begin{aligned} T^h(u) &= (u(t_0) - \alpha, (h^{-1}(u(t_{j+1}) - u(t_j)) - V(h, t_j, u(t_j))), j = 0, \dots, \sigma(h) - 1) \\ &= 0, \quad u \in (\mathbb{R}^N)^{\Omega_h} \end{aligned} \quad (3-1)$$

Es gilt der folgende Stabilitätssatz:

3.1 Satz. (Stabilitätssatz)

Es existiere ein $\bar{h} > 0$ und ein $L > 0$, so dass die Verfahrensfunktion V einer LIPSCHITZbedingung

$$\|V(h, t, v) - V(h, t, w)\|_\infty \leq L\|v - w\|_\infty, \quad 0 < h \leq \bar{h}, \quad t \in [t_0, t_{\text{end}}]$$

genügt. Dann ist das Einschrittverfahren (3-1) stabil, d.h. es existiert ein $C > 0$ mit

$$\|u_1 - u_2\|_\infty \leq C\|T^h(u_1) - T^h(u_2)\|_\infty, \quad \forall u_1, u_2 \in (\mathbb{R}^N)^{\Omega_h} \quad 0 < h \leq \bar{h}.$$

Beweis: Seien $u_1, u_2 \in (\mathbb{R}^N)^{\Omega_h}$. Wir setzen $z(t) = u_1(t) - u_2(t)$ für $t = t_0 + jh$, $j = 0, \dots, \sigma(h) - 1$ und erhalten

$$\begin{aligned} h^{-1}(z(t+h) - z(t)) &= [h^{-1}(u_1(t+h) - u_1(t)) - V(h, t, u_1(t))] \\ &\quad - [h^{-1}(u_2(t+h) - u_2(t)) - V(h, t, u_2(t))] \\ &\quad + [V(h, t, u_1(t)) - V(h, t, u_2(t))] \\ &= T^h(u_1)(t+h) - T^h(u_2)(t+h) + V(h, t, u_1(t)) - V(h, t, u_2(t)) \end{aligned}$$

und schließlich

$$z(t+h) = z(t) + h(T^h(u_1)(t+h) - T^h(u_2)(t+h) + V(h, t, u_1(t)) - V(h, t, u_2(t))).$$

Mit $\rho := \|T^h(u_1) - T^h(u_2)\|_\infty$ finden wir

$$\begin{aligned} \|z(t+h)\|_\infty &\leq \|z(t)\|_\infty + h \overbrace{\|T^h(u_1)(t+h) - T^h(u_2)(t+h)\|_\infty}^{\leq \rho} \\ &\quad + h \underbrace{\|V(h, t, u_1(t)) - V(h, t, u_2(t))\|_\infty}_{\leq L\|z(t)\|_\infty} \\ &\leq (1 + hL)\|z(t)\|_\infty + h\rho. \end{aligned}$$

Wir zeigen nun die Abschätzung

$$\|z(t_j)\|_\infty \leq \left[(1 + hL)^j + h \sum_{i=0}^{j-1} (1 + hL)^i \right] \rho, \quad j = 0, \dots, \sigma(h).$$

Induktionsanfang ($j = 0$):

$$\begin{aligned} \|z(t_0)\|_\infty &= \|u_1(t_0) - u_2(t_0)\|_\infty = \|T^h(u_1)(t_0) - T^h(u_2)(t_0)\|_\infty \\ &\leq \|T^h(u_1) - T^h(u_2)\|_\infty = \rho \end{aligned}$$

Induktionsschritt ($j \rightarrow j + 1$):

$$\begin{aligned} \|z(t_{j+1})\|_\infty &\leq (1 + hL)\|z(t_j)\|_\infty + h\rho \\ &\leq (1 + hL) \left[(1 + hL)^j + h \sum_{i=0}^{j-1} (1 + hL)^i \right] \rho + h\rho \\ &= \left[(1 + hL)^{j+1} + h \sum_{i=1}^j (1 + hL)^i \right] \rho + h\rho \\ &= \left[(1 + hL)^{j+1} + h \sum_{i=0}^j (1 + hL)^i \right] \rho \end{aligned}$$

Wir benutzen

$$\sum_{i=0}^n q^i = \frac{1 - q^{n+1}}{1 - q}$$

und die Ungleichung

$$1 + hL \leq \exp(hL) \Rightarrow (1 + hL)^j \leq \exp(hL)^j = \exp(hLj)$$

um weiter abschätzen zu können:

$$\begin{aligned} \|z(t_j)\|_\infty &\leq \rho \left[(1 + hL)^j + h \sum_{i=0}^{j-1} (1 + hL)^i \right] \\ &= \rho \left[(1 + hL)^j + h \frac{1 - (1 + hL)^j}{1 - (1 + hL)} \right] \\ &= \rho \left[(1 + hL)^j + \frac{(1 + hL)^j - 1}{L} \right] \\ &\leq \rho \left[\exp(jhL) + \frac{\exp(jhL) - 1}{L} \right], \quad j = 0, \dots, \sigma(h) \end{aligned}$$

Wegen der Monotonie der Abschätzung gilt auch

$$\begin{aligned} \|z(t_j)\|_\infty &\leq \rho \left[\exp(\sigma(h)hL) + \frac{\exp(\sigma(h)hL) - 1}{L} \right] \\ &\leq \rho \underbrace{\left[\exp((t_{\text{end}} - t_0 + \bar{h})L) + \frac{\exp((t_{\text{end}} - t_0 + \bar{h})L) - 1}{L} \right]}_{=: C = \text{const}} \end{aligned}$$

für jedes $j = 0, \dots, \sigma(h)$. Somit erhält man

$$\max_{j=0, \dots, \sigma(h)} \{\|u_1(t_j) - u_2(t_j)\|_\infty\} = \max_{j=0, \dots, \sigma(h)} \{\|z(t_j)\|_\infty\} \leq C\rho = C\|T^h(u_1) - T^h(u_2)\|_\infty.$$

□

3.2 Korollar. *Ein konsistentes Einschrittverfahren (der Ordnung p) dessen Verfahrensfunktion V die Lipschitz-Bedingung*

$$\|V(h, t, v) - V(h, t, w)\|_\infty \leq L\|v - w\|_\infty, \quad v, w \in \mathbb{R}^N, \quad 0 < h \leq \bar{h}, \quad t_0 \leq t \leq t_{\text{end}}$$

erfüllt, ist konvergent (der Ordnung p), d.h.

$$\|\bar{u}_h - u^h\|_\infty = \max\{\|\bar{u}(t) - u^h(t)\|_\infty, t = t_0 + jh, j = 0, \dots, \sigma(h)\} \rightarrow 0 (= O(h^p)),$$

wobei u^h die Lösung von $T^h(u) = 0$ ist.

Im Fall der Runge-Kutta Verfahren gilt nun

3.3 Lemma. (*Runge-Kutta-Formel, ohne Beweis*)

Vorgelegt sei $u'(t) = f(t, u(t))$, $t \in [t_0, t_{\text{end}}]$, $u(t_0) = \alpha$ mit $f : [t_0, t_{\text{end}}] \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ global Lipschitz-stetig bezüglich u mit Konstante L . Sei V die Verfahrensfunktion eines Runge-Kutta Verfahrens mit Tableau $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$, und es gelte

$$q := \bar{h}L\|A\|_\infty < 1.$$

Dann ist die Verfahrensfunktion $V = V(h, t, v)$ des Runge-Kutta Verfahrens global Lipschitz stetig bzgl. v mit

$$L_1 = \sum_{i=1}^s |b_i| \frac{L}{1 - q}, \quad t \in [t_0, t_{\text{end}}], \quad 0 < h \leq \bar{h}.$$

4. Asymptotische Entwicklung und Schrittweitensteuerung

a) Verschärfte Konvergenzaussagen

Wir betrachten erneut ein Einschrittverfahren der Form mit

$$\begin{aligned} T^h(u) &= (u(t_0) - \alpha, (h^{-1}(u(t_{j+1}) + u(t_j)) - V(h, t_j, u(t_j))), j = 0, \dots, \sigma(h) - 1) \\ &= 0, \quad u \in (\mathbb{R}^N)^{\Omega_h} \end{aligned}$$

mit einer Verfahrensfunktion $V \in C([0, h_0] \times [t_0, t_{end}] \times \mathbb{R}^N, \mathbb{R}^N)$.

Unsere bisherigen Ergebnisse liefern Konvergenzaussagen der Form $u^h = \bar{u}_h + O(h^p)$, wobei $T^h(u^h) = 0$ und $O(h^p)$ gleichmäßig in t gilt. Wir wollen diese Aussagen verschärfen zu einer asymptotischen Entwicklung des Diskretisierungsfehlers $u^h - \bar{u}_h$ in der Form

$$u^h(t) = \bar{u}(t) + \sum_{i=0}^{r-1} h^{p+i} e_i(t) + O(h^{r+p}), \quad t \in \Omega_h$$

für $r \geq 0$, wobei die $e_i \in C([t_0, t_{end}], \mathbb{R}^N)$ von h unabhängige Funktionen sein sollen. Die letzte Formel lässt sich schreiben als

$$\|u^h - \left(\bar{u} + \sum_{i=0}^{r-1} h^{p+i} e_i \right)_h\|_{\infty} = O(h^{r+p}).$$

Erfüllt unser Einschrittverfahren die Stabilitätsungleichung (Definition 2.3), so folgt

$$\begin{aligned} \|u^h - \left(\bar{u} + \sum_{i=0}^{r-1} h^{p+i} e_i \right)_h\|_{\infty} &\leq C \|\overbrace{T^h(u^h)}^{=0} - T^h \left(\bar{u} + \sum_{i=0}^{r-1} h^{p+i} e_i \right)_h\|_{\infty} \\ &= C \|T^h \left(\bar{u} + \sum_{i=0}^{r-1} h^{p+i} e_i \right)_h\|_{\infty} \end{aligned}$$

Werden nun die Funktionen e_i so bestimmt, dass dieser letzte Term sich wie $O(h^{r+p})$ verhält, so erhalten wir die gewünschte Aussage. Aus Komplexitätsgründen wollen wir uns dabei auf den Fall $r = 1$, d.h.

$$u^h(t) = \bar{u}(t) + \underbrace{h^p e_0(t)}_{=O(h^p)} + O(h^{p+1}), \quad t \in \Omega_h$$

beschränken. Zunächst betrachten wir das Euler-Cauchy Verfahren ($p = 1$):

$$T^h((\bar{u} + h e_0)_h)(t_0) = (\bar{u} + h e_0)(t_0) - \alpha = \bar{u}(t_0) + h e_0(t_0) - \alpha = h e_0(t_0).$$

Sei $t \in \{t_0 + jh \mid j = 0, \dots, \sigma(h) - 1\}$.

$$\begin{aligned}
T^h((\bar{u} + he_0)_h)(t+h) &= \frac{(\bar{u} + he_0)(t+h) - (\bar{u} + he_0)(t)}{h} - f(t, (\bar{u} + he_0)(t)) \\
&= \frac{\bar{u}(t+h) - \bar{u}(t)}{h} + e_0(t+h) - e_0(t) - f(t, (\bar{u} + he_0)(t)) \\
&= \bar{u}'(t) + \frac{h}{2}\bar{u}''(t) + O(h^2) + he_0'(t) + O(h^2) \\
&\quad - \underbrace{f(t, \bar{u}(t))}_{=\bar{u}'(t)} - \frac{\partial}{\partial u}f(t, \bar{u}(t))he_0(t) + O(h^2) \\
&= h \left[\frac{1}{2}\bar{u}''(t) + e_0'(t) - \frac{\partial}{\partial u}f(t, \bar{u}(t))e_0(t) \right] + O(h^2)
\end{aligned}$$

Wir erhalten also

$$\|T^h((\bar{u} + he_0)_h)\|_\infty = O(h^2),$$

falls e_0 durch die lineare Anfangswertaufgabe

$$\begin{aligned}
e_0(t_0) &= 0 \\
e_0'(t) &= \frac{\partial}{\partial u}f(t, \bar{u}(t))e_0(t) - \frac{1}{2}\bar{u}''(t), \quad t \in [t_0, t_{\text{end}}]
\end{aligned} \tag{4-1}$$

definiert wird. Ist $f \in C^2([t_0, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N)$, so gilt $\bar{u} \in C^3([t_0, t_{\text{end}}], \mathbb{R}^N)$, $e_0 \in C^2([t_0, t_{\text{end}}], \mathbb{R}^N)$, womit die $O(h^2)$ -Terme gleichmäßig in t sind.

Eine Verallgemeinerung dieser Untersuchungen ist der folgende

4.1 Satz. (ohne Beweis)

Sei $\bar{u} \in C^{p+r+1}([t_0, t_{\text{end}}], \mathbb{R}^N)$ Lösung der AWA $u'(t) = f(t, u(t))$, $u(t_0) = \alpha$. Ferner sei ein bei $\bar{u}$ konsistentes Einschnittverfahren $T^h(u) = 0$ der Ordnung p gegeben mit $V \in C^{p+r}([0, h_0] \times [t_0, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N)$ und $V = V(h, t, u)$ global Lipschitz-stetig in u . Dann existieren Funktionen $e_i \in C^{r+1-i}([t_0, t_{\text{end}}], \mathbb{R}^N)$, $e_i(t_0) = 0$, $i = 0, \dots, r-1$ mit

$$\|u^h - \left(\bar{u} + \sum_{i=0}^{r-1} h^{p+i} e_i \right)_h\|_\infty = O(h^{p+r}).$$

Wir wenden nun dieses Ergebnis auf Runge-Kutta Verfahren an.

Bemerkung 8. Für Runge-Kutta Verfahren und $q \in \mathbb{N}$ lässt sich $V \in C^q([0, h_0] \times [t_0, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N)$, V global lipschitz-stetig bezgl. u folgern, falls $f \in C^q([t_0, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N)$ und f global lipschitz-stetig bezgl. u ist.

b) Schrittweitensteuerung und Fehlerschätzungen

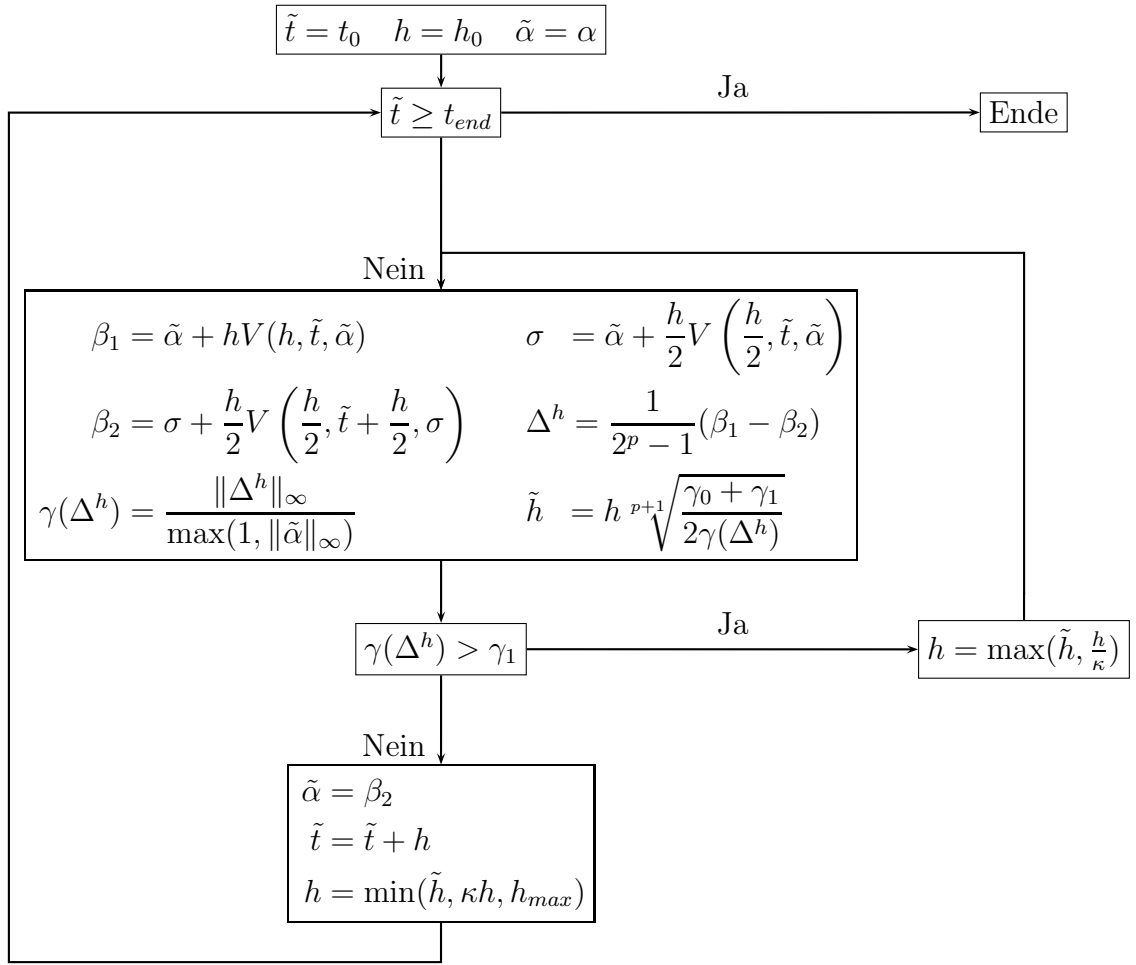
Vorgelegt sei

$$\begin{aligned}u'(t) &= f(t, u(t)), \quad t_0 \leq t \leq t_{\text{end}}, \\u(t_0) &= \alpha.\end{aligned}$$

Wir führen folgende Parameter ein:

$h_0 > 0$	Startschrittweite
$h_{\max} \geq h_0$	max. Schrittweite
$\kappa > 1$	Vergrößerungs- bzw. Verkleinerungsfaktor
$[\gamma_0, \gamma_1]$	Toleranzintervall für den geschätzten Fehler

Üblicherweise wählt man $\kappa = 2$ als Vergrößerungs- bzw. $\kappa^{-1} = 1/2$ Verkleinerungsfaktor, die so genannte $(h - \frac{h}{2})$ -Steuerung



Um dieses Verfahren zur Schrittweitensteuerung zu erstellen, benutzen wir die Extrapolationsergebnisse. Sei $\tilde{\alpha}$ Näherungswert für $\bar{u}(\tilde{t})$, und sei $\tilde{u}(t)$ die Lösung von

$$\begin{aligned} u'(t) &= f(t, u(t)), \\ u(\tilde{t}) &= \tilde{\alpha}, \end{aligned}$$

so lässt sich der Gesamtfehler bei $\tilde{t} + h$ so darstellen:

$$\begin{aligned} u^h(\tilde{t} + h) - \bar{u}(\tilde{t} + h) &= \overbrace{u^h(\tilde{t} + h) - \tilde{u}(\tilde{t} + h)}^{\text{neu hinzukommender Fehler}} \\ &+ \underbrace{\tilde{u}(\tilde{t} + h) - \bar{u}(\tilde{t} + h)}_{\substack{\text{durch Dgl. fortgeplanter Fehler} \\ \text{(nicht mehr kontrollierbar)}}} \end{aligned}$$

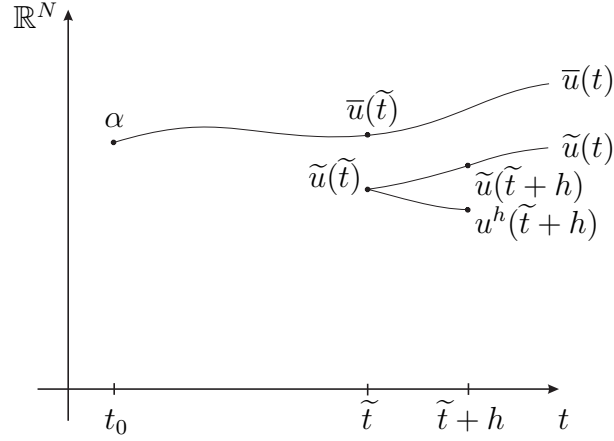


Abbildung 7: Erklärung der Schrittweitensteuerung

Für den neu hinzukommenden Fehler gilt nach Satz 4.1 mit $r = 2$

$$\begin{aligned}
 u^h(\tilde{t} + h) - \tilde{u}(\tilde{t} + h) &= h^p e_0(\tilde{t} + h) + h^{p+1} e_1(\tilde{t} + h) + O(h^{p+2}) \\
 &= h^p (\underbrace{e_0(\tilde{t})}_{=0} + h e'_0(\tilde{t}) + O(h^2)) + h^{p+1} (\underbrace{e_1(\tilde{t})}_{=0} + O(h)) + O(h^{p+2}) \\
 &= h^{p+1} e'_0(\tilde{t}) + O(h^{p+2})
 \end{aligned}$$

Analog bekommt man

$$\begin{aligned}
 u^{h/2}(\tilde{t} + h) - \tilde{u}(\tilde{t} + h) &= \left(\frac{h}{2}\right)^p e_0(\tilde{t} + h) + \left(\frac{h}{2}\right)^{p+1} e_1(\tilde{t} + h) + O(h^{p+2}) \\
 &= \left(\frac{h}{2}\right)^p \left(\underbrace{e_0(\tilde{t})}_{=0} + e'_0(\tilde{t})h + O(h^2) \right) \\
 &\quad + \left(\frac{h}{2}\right)^{p+1} \left(\underbrace{e_1(\tilde{t})}_{=0} + O(h) \right) + O(h^{p+2}) \\
 &= \underbrace{\left(\frac{h}{2}\right)^p h e'_0(\tilde{t})}_{\text{Hauptfehlerterm}} + O(h^{p+2}).
 \end{aligned}$$

Zur Schätzung des Hauptfehlerterms berechnen wir

$$\begin{aligned}
 u^h(\tilde{t} + h) - u^{h/2}(\tilde{t} + h) &= - \left(\frac{h}{2}\right)^p h e'_0(\tilde{t}) + h^{p+1} e'_0(\tilde{t}) + O(h^{p+2}) \\
 &= \left(\frac{h}{2}\right)^p h e'_0(\tilde{t}) (2^p - 1) + O(h^{p+2}) \\
 \Rightarrow \left(\frac{h}{2}\right)^p h e'_0(\tilde{t}) &= \frac{u^h(\tilde{t} + h) - u^{h/2}(\tilde{t} + h)}{2^p - 1} + O(h^{p+2}).
 \end{aligned}$$

Mit den eingeführten Bezeichnungen bekommt man

$$\left(\frac{h}{2}\right)^p h e'_0(\tilde{t}) = \frac{\beta_1 - \beta_2}{2^p - 1} + O(h^{p+2}) = \Delta^h + O(h^{p+2}). \quad (4-2)$$

Dies ist eine $O(h^{p+2})$ Schätzung des Fehlers. Die Größe Δ^h heißt deshalb auch Fehler-schätzer. Benutze das Fehlermaß:

$$\gamma(\Delta^h) = \frac{\|\Delta^h\|_\infty}{\max(1, \|\tilde{u}(\tilde{t})\|_\infty)}$$

Wir testen nun, ob $\gamma(\Delta^h)$ in $[\gamma_0, \gamma_1]$ liegt. Im Falle $\gamma(\Delta^h) > \gamma_1$ versuche ein $\tilde{h}$ zu finden, mit $\gamma(\Delta^{\tilde{h}}) \approx \frac{1}{2}(\gamma_0 + \gamma_1)$. Nun gilt mit (4-2) für diese Schrittweite $\tilde{h}$

$$\begin{aligned} \Delta^{\tilde{h}} &= \left(\frac{\tilde{h}}{2}\right)^p \tilde{h} e'_0(\tilde{t}) + O(\tilde{h}^{p+2}) \\ &= \left(\frac{\tilde{h}}{h}\right)^{p+1} \underbrace{\left(\frac{h}{2}\right)^p h e'_0(\tilde{t})}_{=\Delta^h + O(h^{p+2})} + O(\tilde{h}^{p+2}) \\ &\Rightarrow \Delta^{\tilde{h}} = \left(\frac{\tilde{h}}{h}\right)^{p+1} \Delta^h + O(h^{p+2} + \tilde{h}^{p+2}). \end{aligned}$$

Näherungsweise verlangen wir also

$$\underbrace{\gamma(\Delta^{\tilde{h}})}_{=\frac{1}{2}(\gamma_0 + \gamma_1)} = \left(\frac{\tilde{h}}{h}\right)^{p+1} \gamma(\Delta^h) \Leftrightarrow \tilde{h}^{p+1} = \frac{\gamma_0 + \gamma_1}{2\gamma(\Delta^h)} h^{p+1} \Leftrightarrow \tilde{h} = h \sqrt[p+1]{\frac{\gamma_0 + \gamma_1}{2\gamma(\Delta^h)}}.$$

Damit ist $\tilde{h}$ genau dann größer (kleiner) als h , wenn $\gamma(\Delta^h)$ kleiner (größer) als $\frac{\gamma_0 + \gamma_1}{2}$ ist. Im Fall $\gamma(\Delta^h) \leq \gamma_1$ akzeptiere $\tilde{u}^{h/2}(\tilde{t} + h)$ und benutze $\tilde{h}$ als neue Schrittweiten-schätzung.

5. Numerik Hamiltonischer Systeme

a) Kollokationsverfahren

Vorgelegt sei die Anfangswertaufgabe

$$\begin{aligned} u'(t) &= f(t, u(t)), & f &\in C^1([t_0, t_{\text{end}}] \times \mathbb{R}^N, \mathbb{R}^N), \\ u(t_0) &= \alpha \in \mathbb{R}^N \end{aligned} \quad (5-1)$$

mit einer Lösung $\bar{u}(t, t_0, \alpha)$. Sei $s \in \mathbb{N}$ und seien $c_1, \dots, c_s \in [0, 1]$ paarweise verschieden. Ein Kollokationspolynom $v(t)$, $v(t) = v(t, t_0, \alpha)$ vom Grad $s \in \mathbb{N}$ zu (5-1) ist definiert durch

$$\begin{aligned} v(t_0) &= \alpha \\ v'(t_0 + c_i h) &= f(t_0 + c_i h, v(t_0 + c_i h)), \quad i = 1, \dots, s. \end{aligned}$$

Als neue Näherung für $\bar{u}(t_0 + h, t_0, \alpha)$ wählt man

$$u_1 = v(t_0 + h).$$

Ein solches Verfahren heißt Kollokationsverfahren für (5-1).

5.1 Bemerkung. Das Polynom v mit $\deg(v) = s$ erfüllt an der Stelle t_0 die Anfangsbedingung und an den Kollokationspunkten $t_0 + c_i h$, $i = 1, \dots, s$ die Differentialgleichung (5-1).

Zwischen den Kollokationsmethoden und den Runge-Kutta Verfahren besteht ein enger Zusammenhang.

5.2 Satz. (ohne Beweis) Vorgelegt sei die Kollokationsmethode

$$\begin{aligned} v(t_0) &= \alpha \\ v'(t_0 + c_i h) &= f(t_0 + c_i h, v(t_0 + c_i h)), \quad i = 1, \dots, s, \\ u_1 &= v(t_0 + h) \end{aligned}$$

zur Lösung der Anfangswertaufgabe (5-1). Ferner seien die $c_i \in [0, 1]$ paarweise verschieden. Dann ist die Kollokationsmethode äquivalent zu einem s -stufigen Runge-Kutta Verfahren mit den Koeffizienten $c_1, \dots, c_s$,

$$\begin{aligned} a_{ij} &= \int_0^{c_i} L_j(\vartheta) \, d\vartheta, \quad 1 \leq i, j \leq s, & b_j &= \int_0^1 L_j(\vartheta) \, d\vartheta, \quad 1 \leq j \leq s \text{ und} \\ L_j(\vartheta) &= \prod_{\substack{l=1 \\ l \neq j}}^s \frac{\vartheta - c_l}{c_j - c_l}, \quad j = 1, \dots, s. \end{aligned}$$

$L_j(\vartheta)$ sind die Lagrangepolynome vom Grad $s - 1$. Sie haben die Eigenschaft $L_j(c_i) = \delta_{ij}$, $1 \leq i, j \leq s$.

Andererseits gilt:

5.3 Satz (ohne Beweis). *Ein s -stufiges implizites Runge-Kutta Verfahren mit paarweise verschiedenen Knoten $c_i \in [0, 1]$, $i = 1, \dots, s$ und einer Ordnung $p \geq s$ ist genau dann ein Kollokationsverfahren, wenn die vereinfachenden Bedingungen $C(q)$*

$$\sum_{j=1}^s a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, \dots, s, \quad k = 1, \dots, q$$

mit $q = s$ gelten.

5.4 Korollar. *Gemäß Satz 5.3 sind die Gauß-Verfahren und die Radau IIA Verfahren Kollokationsverfahren.*

b) Symplektische Systeme

Betrachte nun Hamiltonsche dynamische Systeme

$$\begin{aligned} x' &= \frac{\partial H}{\partial y}(x, y), & x(t_0) &= x_0 \\ y' &= -\frac{\partial H}{\partial x}(x, y), & y(t_0) &= y_0 \end{aligned} \quad (5-2)$$

mit $H \in C^2(\Omega, \mathbb{R})$, $\Omega \subset \mathbb{R}^{2N}$ offen. Die Lösung von (5-2) sei $\varphi(t, (x_0, y_0))$ für $t \in J((x_0, y_0))$. Es gilt

(i) $H(\varphi(t, (x_0, y_0))) = H(x_0, y_0)$, $t \in J((x_0, y_0))$.

(ii) Die Abbildung $\varphi(t, \cdot)$ ist symplektisch, d.h.

$$\left(\frac{\partial}{\partial(x, y)} \varphi(t, (x_0, y_0)) \right)^\top \cdot J \cdot \frac{\partial}{\partial(x, y)} \varphi(t, (x_0, y_0)) = J := \begin{pmatrix} 0 & I_N \\ -I_N & 0 \end{pmatrix} \in \mathbb{R}^{2N, 2N}.$$

Wende nun ein Runge-Kutta Verfahren mit Tableau $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ auf (5-2) an. Sei $h > 0$,

$t_j = t_0 + jh$, $j = 0, \dots, \sigma(h)$. Dann erhalten wir die diskrete Iteration

$$x_{n+1} = x_n + h \sum_{i=1}^s b_i \frac{\partial H}{\partial y}(X^i, Y^i)$$

$$y_{n+1} = y_n - h \sum_{i=1}^s b_i \frac{\partial H}{\partial x}(X^i, Y^i) \quad (5-3)$$

mit $(X^1, Y^1), \dots, (X^s, Y^s)$ Lösung von

$$\begin{pmatrix} X^i \\ Y^i \end{pmatrix} = \begin{pmatrix} x_n + h \sum_{j=1}^s a_{ij} \frac{\partial H}{\partial y}(X^j, Y^j) \\ y_n - h \sum_{j=1}^s a_{ij} \frac{\partial H}{\partial x}(X^j, Y^j) \end{pmatrix}, \quad i = 1, \dots, s. \quad (5-4)$$

Genügt nun $\frac{\partial H}{\partial y}, \frac{\partial H}{\partial x}$ einer globalen Lipschitzbedingung, so ist (5-4) für $h > 0$ hinreichend klein lösbar, und wir erhalten $X^i = X^i(h, x_n, y_n, A)$, $Y^i = Y^i(h, x_n, y_n, A)$, $i = 1, \dots, s$. Einsetzen in (5-3) liefert

$$\begin{aligned} x_{n+1} &= x_n + h \sum_{i=1}^s b_i \frac{\partial H}{\partial y}(X^i(h, x_n, y_n, A), Y^i(h, x_n, y_n, A)) =: G_1(h, x_n, y_n, b, A), \\ y_{n+1} &= y_n - h \sum_{i=1}^s b_i \frac{\partial H}{\partial x}(X^i(h, x_n, y_n, A), Y^i(h, x_n, y_n, A)) =: G_2(h, x_n, y_n, b, A). \end{aligned}$$

Dies impliziert

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = G(h, x_n, y_n, A, b) = \begin{pmatrix} G_1(h, x_n, y_n, b, A) \\ G_2(h, x_n, y_n, b, A) \end{pmatrix}.$$

Man nennt $G(h, \cdot, \cdot, b, A)$ die Zeit h -Abbildung des Runge-Kutta Verfahrens oder die diskrete Zeit h -Abbildung.

5.5 Definition. Ein Runge-Kutta Verfahren mit Tableau $\frac{c}{b^\top} \begin{array}{c|c} A \\ \hline b^\top \end{array}$ heißt symplektisch, falls ein $\bar{h} > 0$ derart existiert, so dass die Abbildung $G(h, \cdot, \cdot, b, A)$ symplektisch für Schrittweiten $0 < h \leq \bar{h}$ ist, d.h.

$$\frac{\partial}{\partial(x, y)} G(h, x, y, b, A)^\top \cdot J \cdot \frac{\partial}{\partial(x, y)} G(h, x, y, b, A) = J, \quad 0 < h \leq \bar{h}.$$

Der folgende Satz charakterisiert nun die symplektischen Runge-Kutta Verfahren.

5.6 Satz (ohne Beweis). Ein Runge-Kutta Verfahren s -ter Stufe mit Tableau $\frac{c}{b^\top} \begin{array}{c|c} A \\ \hline b^\top \end{array}$ ist symplektisch genau dann, wenn mit $B := \text{diag}(b_1, \dots, b_s) \in \mathbb{R}^{s,s}$ gilt

$$M := BA + A^\top B - bb^\top = 0.$$

5.7 Bemerkung. • Explizite Runge-Kutta Verfahren ($A_{ij} = 0$ für $j \geq i$) mit $B(1)$, d.h. $\sum_{i=1}^s b_i = 1$ sind niemals symplektisch, denn es gilt

$$M_{ii} = (BA)_{ii} + (A^T B)_{ii} - (bb^T)_{ii} = -(bb^T)_{ii} = -b_i^2, \quad i = 1, \dots, s$$

und wegen $\sum_{i=1}^s b_i = 1$ ist mindestens ein $b_i \neq 0$.

- Bei den impliziten Runge-Kutta Verfahren muss jede Gruppe von Verfahren separat analysiert werden. Man findet $M \neq 0$ für die Radau-Verfahren und $M = 0$ für die Gauß-Verfahren, d.h. nach Satz 5.6 sind die Gauß-Verfahren symplektisch.

Es bleibt zu analysieren, ob ein symplektisches Verfahren angewandt auf eine Hamiltonsche Gleichung die Erhaltung der Hamiltonfunktion im Diskreten garantiert.

Wir betrachten eine beliebige autonome Differentialgleichung

$$u'(t) = f(u(t)) \in \mathbb{R}^N$$

mit Lösung $\bar{u}(t, u)$, $t \in J(u)$. Diese hat ein quadratisches erstes Integral, falls für $\tilde{Q} = u^\top C u$, $C \in \mathbb{R}^{N,N}$ symmetrisch, gilt

$$\bar{u}(t, u)^\top C \bar{u}(t, u) = \tilde{Q}(\bar{u}(t, u)) \stackrel{!}{=} \tilde{Q}(u) = u^\top C u$$

Aus $u^\top C u = \bar{u}(t, u)^\top C \bar{u}(t, u)$ folgt durch Differenzieren

$$\frac{d}{dt} \tilde{Q}(\bar{u}(t, u)) = 2\bar{u}(t, u)^\top C \frac{\partial \bar{u}}{\partial t}(t, u) = 2\bar{u}(t, u)^\top C f(\bar{u}(t, u)) = 0,$$

d.h. $\tilde{Q} = u^\top C u$ ist eine quadratische Invariante, falls $2u^\top C f(u) = 0$, $u \in \mathbb{R}^N$ gilt.

5.8 Satz. Die Runge-Kutta Verfahren vom Gauß-Typ erhalten quadratische Invarianten.

Beweis: Wende die Gauß-Verfahren als Kollokationsmethode auf $u'(t) = f(u(t))$, $u(t_0) = \alpha$ an. Sei $v(t)$ das Kollokationspolynom vom Grad s und

$$\begin{aligned} v(t_0) &= \alpha \\ v'(t_0 + c_i h) &= f(v(t_0 + c_i h)), \quad i = 1, \dots, s \\ u(t_1) &= v(t_0 + h) \end{aligned}$$

Sei nun $\tilde{Q}(u) = u^\top C u$, $C \in \mathbb{R}^{N,N}$ symmetrisch. Für das Kollokationspolynom $v(t)$ gilt

$$\frac{d}{dt} \tilde{Q}(v(t)) = 2v(t)^\top C v'(t).$$

Mit $v(t_0) = \alpha = u(t_0)$ und $v(t_0 + h) = u(t_1)$ folgt

$$\begin{aligned}
 u(t_1)^\top C u(t_1) - u(t_0)^\top C u(t_0) &= v(t_0 + h)^\top C v(t_0 + h) - v(t_0)^\top C v(t_0) \\
 &= \tilde{Q}(v(t_0 + h)) - \tilde{Q}(v(t_0)) \\
 &= \int_{t_0}^{t_0+h} \frac{d}{d\tau} \tilde{Q}(v(\tau)) d\tau \\
 &= \int_{t_0}^{t_0+h} \frac{d}{d\tau} v(\tau)^\top C v(\tau) d\tau \\
 &= \int_{t_0}^{t_0+h} 2 \underbrace{v(\tau)^\top C v'(\tau)}_{\substack{\text{ein Polynom} \\ \text{vom Grad } 2s-1}} d\tau.
 \end{aligned}$$

Die durch die Gewichte $b_1, \dots, b_s$ und die Knoten $c_1, \dots, c_s$ der Gauß-Form definierte Quadraturmethode ist exakt vom Grad $2s - 1$, d.h.

$$\begin{aligned}
 \int_{t_0}^{t_0+h} 2v(\tau)^\top C v'(\tau) d\tau &= 2 \sum_{i=1}^s b_i v(t_0 + c_i h)^\top C v'(t_0 + c_i h) \\
 \Rightarrow u(t_1)^\top C u(t_1) - u(t_0)^\top C u(t_0) &= 2 \sum_{i=1}^s b_i v(t_0 + c_i h)^\top C \underbrace{v'(t_0 + c_i h)}_{f(v(t_0 + c_i h))} \\
 &= 2 \sum_{i=1}^s b_i \underbrace{v(t_0 + c_i h)^\top}_{y^\top} C \underbrace{f(v(t_0 + c_i h))}_y = 0.
 \end{aligned}$$

Beachte: $y^\top C f(y) = 0 \forall y$, da $\tilde{Q}$ ein erstes Integral ist. Somit folgt: $u(t_1)^\top C u(t_1) = u(t_0)^\top C u(t_0)$. Induktiv erhalten wir $u(t_i)^\top C u(t_i) = u(t_0)^\top C u(t_0)$, $i = 1, 2, 3 \dots$ und der Satz ist gezeigt. $\square$

5.9 Bemerkung. Allgemeiner als in Satz 5.8 lässt sich zeigen, dass jedes symplektische Runge-Kutta Verfahren quadratische Invarianten erhält.

Symplektische Runge-Kutta Methoden sind also besonders geeignet zur Integration Hamiltonischer dynamischer Systeme und von Systemen mit quadratischen Invarianten.

5.10 Beispiel. Starrer Körper

Die Bewegungsgleichung eines rotierenden starren Körpers, dessen Massenzentrum

im Ursprung liegt, ist gegeben durch die Euler-Gleichung. Diese lautet:

$$\begin{cases} y_1' = a_1 y_2 y_3, & a_1 = \frac{I_2 - I_3}{I_2 I_3} \\ y_2' = a_2 y_1 y_3, & a_2 = \frac{I_3 - I_1}{I_1 I_3} \\ y_3' = a_3 y_1 y_2, & a_3 = \frac{I_1 - I_2}{I_1 I_2} \end{cases}$$

Dabei bezeichnet $y(t) = (y_1(t), y_2(t), y_3(t))$ den Drehimpuls zur Zeit t und I_1, I_2, I_3 die Hauptträgheitsmomente. Als Invarianten dienen hier die Energie des Systems

$$H(y) = \frac{1}{2} \left(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3} \right)$$

und das Normquadrat des Drehimpulses

$$L(y) = y_1^2 + y_2^2 + y_3^2.$$

Wir zeigen nun, dass $L(y)$ erhalten bleibt (für $H(y)$ zeigt man analog):

$$\begin{aligned} \frac{d}{dt} L(\bar{y}(t)) &= 2\bar{y}_1(t)\bar{y}_1'(t) + 2\bar{y}_2(t)\bar{y}_2'(t) + 2\bar{y}_3(t)\bar{y}_3'(t) \\ &= 2\bar{y}_1(t)a_1\bar{y}_2(t)\bar{y}_3(t) + 2\bar{y}_2(t)a_2\bar{y}_1(t)\bar{y}_3(t) + 2\bar{y}_3(t)a_3\bar{y}_1(t)\bar{y}_2(t) \\ &= 2\bar{y}_1(t)\bar{y}_2(t)\bar{y}_3(t)(a_1 + a_2 + a_3) \\ &= 2\bar{y}_1(t)\bar{y}_2(t)\bar{y}_3(t) \left(\frac{I_2 - I_3}{I_2 I_3} + \frac{I_3 - I_1}{I_1 I_3} + \frac{I_1 - I_2}{I_1 I_2} \right) \\ &= 2\bar{y}_1(t)\bar{y}_2(t)\bar{y}_3(t) \left(\frac{I_1 I_2 - I_1 I_3 + I_2 I_3 - I_1 I_2 + I_1 I_3 - I_2 I_3}{I_1 I_2 I_3} \right) \\ &= 0 \Rightarrow L(\bar{y}(t)) = L(\bar{y}(0)), \quad t \geq 0. \end{aligned}$$

Die Lösungen $\bar{y}(t)$ des Systems liegen also auf dem Schnitt der Sphäre

$$y_1^2 + y_2^2 + y_3^2 = L(\bar{y}(0))$$

mit dem Ellipsoid gegeben durch

$$\frac{1}{2} \left(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3} \right) = H(\bar{y}(0)).$$

Da sowohl H als auch L quadratische Invarianten sind, werden diese durch Diskretisierung mit symplektischen Runge-Kutta Verfahren korrekt wiedergegeben. Wir erwarten deshalb qualitativ gute Ergebnisse bei symplektischen Verfahren, denn auch die diskreten Orbits liegen auf dem Schnitt der Sphäre und dem Ellipsoid.

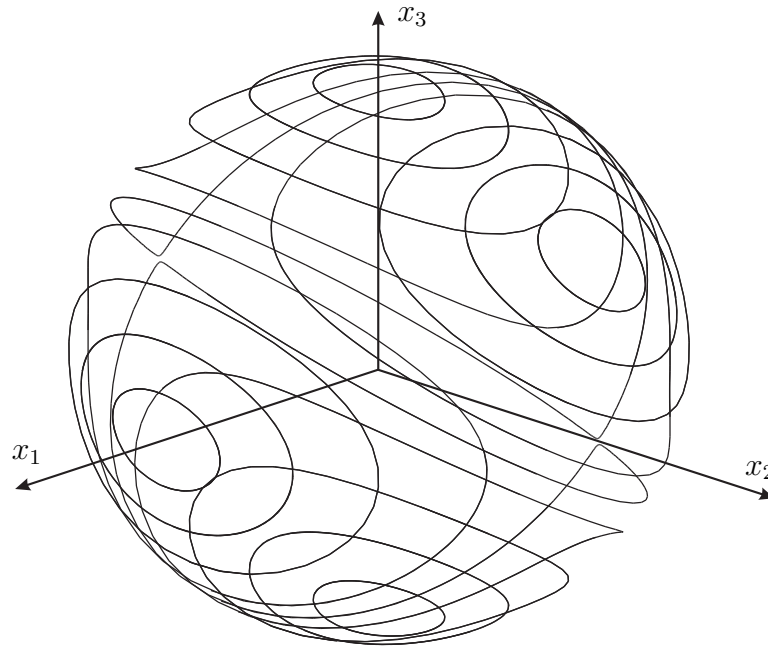


Abbildung 8: Phasenbild des starren Körpers

5.11 Beispiel. Restringsiertes Drei-Körper-Problem

Bewegung eines kleinen Körpers (z.B. eines Satelliten) im Kraftfeld zweier großer Körper (z.B. Erde und Mond), d.h. $m_3 \ll \min\{m_1, m_2\}$.

Annahme: Die Bewegung aller Körper verläuft in einer Ebene. Idealisiert nimmt man überdies $m_3 = 0$ an.

In einem mitrotierenden Koordinaten-System, in welchem die beiden großen Körper als ruhend erscheinen, wird die Bahn $(x_1^*(t), x_2^*(t))$ des kleinen Körpers durch die Differentialgleichung

$$\begin{aligned} x_1'' &= x_1 + 2x_2' - (1 - \mu) \frac{x_1 + \mu}{r_1^3} - \mu \frac{x_1 - 1 + \mu}{r_2^3}, \\ x_2'' &= x_2 - 2x_1' - (1 - \mu) \frac{x_2}{r_1^3} - \mu \frac{x_2}{r_2^3} \end{aligned}$$

beschrieben, wobei $r_1 = \sqrt{(x_1 + \mu)^2 + x_2^2}$, $r_2 = \sqrt{(x_1 - 1 + \mu)^2 + x_2^2}$ und $\mu = m_2/(m_1 + m_2)$ für das Massenverhältnis der beiden großen Körper steht. Dem Massenverhältnis von Erde und Mond entspricht der Wert $\mu = \frac{1}{82.45}$.

Die Energie des Systems ist gegeben durch

$$\begin{aligned} H(x_1, x_2, x_1', x_2') &= \frac{1}{2}((x_1')^2 + (x_2')^2) - \frac{1}{2}(x_1^2 + x_2^2) \\ &\quad - \frac{1 - \mu}{\sqrt{(x_1 + \mu)^2 + x_2^2}} - \frac{\mu}{\sqrt{(x_1 - (1 - \mu))^2 + x_2^2}} \end{aligned}$$

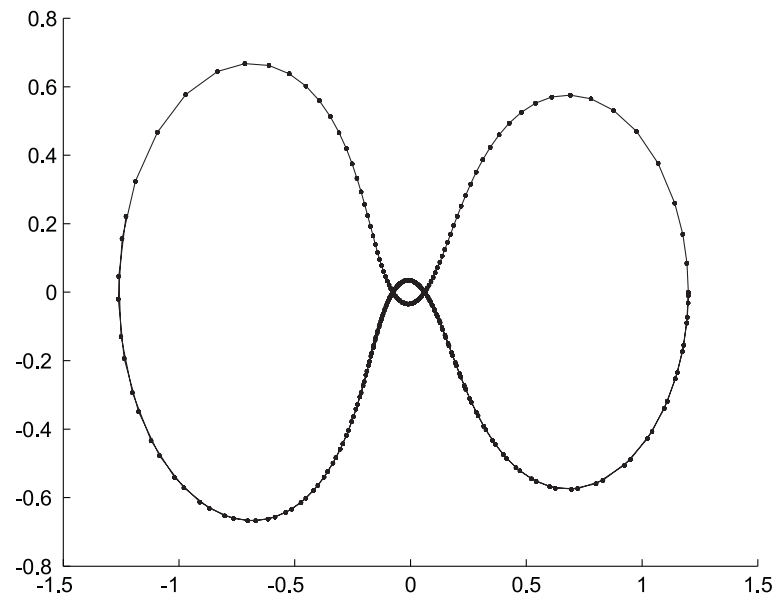


Abbildung 9: Phasenbild des restringierten Dreikörperproblems

Man bringt nun die Gleichungen auf ein System erster Ordnung und sucht periodische Lösungen (d.h. geschlossene Satellitenbahnen). Zur Untersuchung geeignet erscheinen symplektische Runge-Kutta Verfahren. Es ist spannend zu sehen, wie sich explizite Runge-Kutta Verfahren oder implizite, nicht-symplektische Verfahren verhalten.

6. Numerische Verfahren für Randwertaufgaben

a) Erinnerung: Reaktions- Transportphänomene

Sei $c(t, s)$ die Konzentration eines Stoffes c zur Zeit t am Orte $s \in [a, b]$, welche einem Fluss unterliegt und welche durch eine Reaktion entsteht oder vergeht. Dann lautet die Reaktions- Transport Gleichung wie folgt

$$\begin{aligned} \frac{\partial}{\partial t} c(t, s) &= - \overbrace{\frac{\partial}{\partial s} J(t, s)}^{\text{Fluss}} + \overbrace{f(c(t, s))}^{\text{Reaktion}} \\ &= - \underbrace{\frac{\partial}{\partial s} (v(s)c(t, s))}_{\text{konv. Fluss}} + \underbrace{\frac{\partial}{\partial s} \left[D(s) \frac{\partial}{\partial s} c(t, s) \right]}_{\text{diff. Fluss}} + f(c(t, s)). \end{aligned}$$

Die zugehörigen Anfangs- und Randbedingungen sind

$$\begin{aligned} \alpha_\sigma c(t, \sigma) + \beta_\sigma \frac{\partial}{\partial s} c(t, \sigma) &= \gamma_\sigma, & \sigma &= a, b, \\ c(0, s) &= c_0(s), & a &\leq s \leq b. \end{aligned}$$

Wir suchen nach einer zeitunabhängigen Lösung $c(t, s) = u(s)$. In diesem Fall bekommt man

$$\begin{aligned} -(v(s)u(s))' + (D(s)u'(s))' + f(u(s)) &= 0, & a &\leq s \leq b, \\ \alpha_\sigma u(\sigma) + \beta_\sigma u'(\sigma) &= \gamma_\sigma, & \sigma &= a, b. \end{aligned}$$

Das ist ein Beispiel für eine Randwertaufgabe (RWA) zweiter Ordnung.

Betrachte jetzt eine RWA der Form

$$\begin{aligned} -u''(s) &= f(s, u(s)), & a &\leq s \leq b, \\ \alpha_\sigma u(\sigma) + \beta_\sigma u'(\sigma) &= \gamma_\sigma, & \sigma &= a, b. \end{aligned} \tag{6-1}$$

Die Lösungsverfahren für (6-1) unterteilen sich in 3 Gruppen

Differenzenverfahren: Hierbei werden Ableitungen durch Differenzenquotienten in einem numerischen Gitter ersetzt.

Schießverfahren: Bei diesen Methoden wird die Lösung von (6-1) auf das Lösen von AWA zurückgeführt.

Variationsmethoden: Bei den Variationsmethoden macht man einen Lösungsansatz mit gegebenen Ansatzfunktionen $\psi, \varphi_1, \dots, \varphi_N$, z.B.

$$u(s) = \psi(s) + \sum_{i=1}^N c_i \varphi_i(s), \quad c_1, \dots, c_N \in \mathbb{R}$$

Man bestimmt dann $c_1, \dots, c_N$ so, dass (6-1) möglichst gut erfüllt wird.

b) Die Idee des Schießverfahrens

Betrachte die RWA (Dirichlet Randwertaufgabe)

$$\begin{aligned} -u''(s) &= f(s, u(s)), & a \leq s \leq b, \\ u(a) &= \gamma_a, \\ u(b) &= \gamma_b. \end{aligned} \tag{6-2}$$

Es sei $\bar{u}(s, \sigma)$, $a \leq s \leq b$, $\sigma \in \mathbb{R}$ Lösung der Anfangswertaufgabe

$$\begin{aligned} -u''(s) &= f(s, u(s)), & a \leq s \leq b, \\ u(a) &= \gamma_a, \\ u'(a) &= \sigma. \end{aligned} \tag{6-3}$$

Idee: Bestimme jetzt $\bar{\sigma} \in \mathbb{R}$ derart, dass $\bar{u}(b, \bar{\sigma}) = \gamma_b$

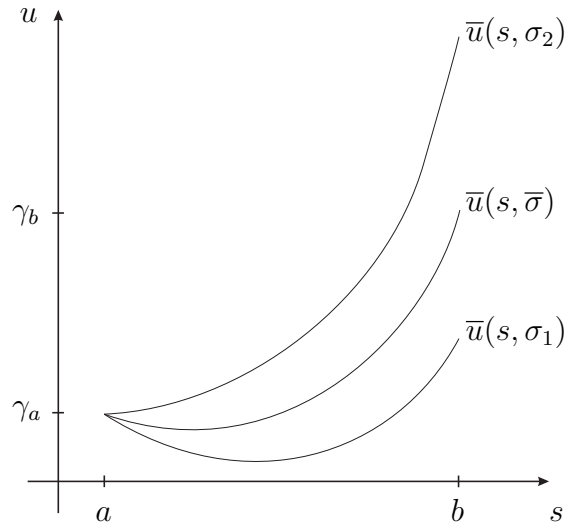


Abbildung 10: Schießverfahren

Offensichtlich gilt für

$$g(\sigma) := \bar{u}(b, \sigma) - \gamma_b$$

die Äquivalenz

$$g(\bar{\sigma}) = 0 \Leftrightarrow \bar{u}(s, \bar{\sigma}), a \leq s \leq b \text{ ist Lösung der RWA (6-2)}$$

Zur Evaluation von $g(\sigma)$ für ein $\sigma \in \mathbb{R}$ muss die AWA (6-3) numerisch gelöst werden. Die Bestimmung einer Nullstelle $\bar{\sigma}$ von g erfolgt mit den aus Numerik I bekannten Verfahren, z.B. das Bisektions- oder das Newtonverfahren. Beim Newtonverfahren

$$\sigma_{n+1} = \sigma_n - \frac{g(\sigma_n)}{g'(\sigma_n)}, \quad n = 0, 1, 2, \dots$$

zur Lösung von $g(\sigma) = 0$ muss zuerst $g'(\sigma) = \frac{\partial}{\partial \sigma} \bar{u}(b, \sigma)$ berechnet werden. Es gilt

$$-\bar{u}''(s, \sigma) = f(s, \bar{u}(s, \sigma)), \quad a \leq s \leq b.$$

Somit folgt durch Differentiation

$$\begin{aligned} -\frac{\partial}{\partial \sigma} \bar{u}''(s, \sigma) &= \frac{\partial}{\partial \sigma} f(s, \bar{u}(s, \sigma)) = \frac{\partial}{\partial u} f(s, \bar{u}(s, \sigma)) \frac{\partial}{\partial \sigma} \bar{u}(s, \sigma), \\ \frac{\partial}{\partial \sigma} \bar{u}(a, \sigma) &= \frac{\partial}{\partial \sigma} \gamma_a = 0, \\ \frac{\partial}{\partial \sigma} \bar{u}'(a, \sigma) &= \frac{\partial}{\partial \sigma} \sigma = 1. \end{aligned}$$

Also genügt $v(s, \sigma) = \frac{\partial}{\partial \sigma} \bar{u}(s, \sigma)$ der linearen AWA

$$\begin{aligned} -v''(s, \sigma) &= \frac{\partial}{\partial u} f(s, \bar{u}(s, \sigma)) v(s, \sigma), \\ v(a, \sigma) &= 0, \\ v'(a, \sigma) &= 1. \end{aligned} \tag{6-4}$$

Zur Auswertung von $g(\sigma)$ und $g'(\sigma)$ ist also das gekoppelte System (6-3), (6-4) zu lösen:

$$\begin{aligned} -u''(s) &= f(s, u(s)), & u(a) &= \gamma_a, & u'(a) &= \sigma, \\ -v''(s) &= \frac{\partial f}{\partial u}(s, u(s)) v(s), & v(a) &= 0, & v'(a) &= 1. \end{aligned}$$

Dafür konvergiert das Newton-Verfahren unter geeigneten Voraussetzungen lokal quadratisch, d.h.

$$|\bar{\sigma} - \sigma_{n+1}| \leq c |\bar{\sigma} - \sigma_n|^2, \quad n = 0, 1, 2$$

Wir betrachten nun den Fall allgemeiner Zweipunkt-Randbedingungen

$$\begin{aligned} -u''(s) &= f(s, u(s)), & a \leq s \leq b, \\ \alpha_\eta u(\eta) + \beta_\eta u'(\eta) &= \gamma_\eta, & \eta = a, b. \end{aligned}$$

Sei $\bar{u}(s, \sigma_1, \sigma_2)$, $a \leq s \leq b$ die Lösung der AWA

$$\begin{aligned} -u''(s) &= f(s, u(s)), & a \leq s \leq b \\ u(a) &= \sigma_1, \\ u'(a) &= \sigma_2, \end{aligned}$$

bzw. $(\bar{u}(s, \sigma_1, \sigma_2), \bar{u}'(s, \sigma_1, \sigma_2))$ ist Lösung des Systems

$$\begin{aligned} w_1'(s) &= w_2(s), \\ w_2'(s) &= -f(s, w_1(s)), \\ w_1(a) &= \sigma_1, \quad w_2(a) = \sigma_2. \end{aligned} \tag{6-5}$$

Bestimme nun $(\bar{\sigma}_1, \bar{\sigma}_2) \in \mathbb{R}^2$ mit

$$\begin{aligned}\alpha_a \bar{u}(a, \bar{\sigma}_1, \bar{\sigma}_2) + \beta_a \bar{u}'(a, \bar{\sigma}_1, \bar{\sigma}_2) &= \alpha_a \bar{\sigma}_1 + \beta_a \bar{\sigma}_2 = \gamma_a, \\ \alpha_b \bar{u}(b, \bar{\sigma}_1, \bar{\sigma}_2) + \beta_b \bar{u}'(b, \bar{\sigma}_1, \bar{\sigma}_2) &= \gamma_b,\end{aligned}$$

d.h. $(\bar{\sigma}_1, \bar{\sigma}_2)$ ist die Lösung von

$$g(\sigma_1, \sigma_2) = \begin{pmatrix} \alpha_a \sigma_1 + \beta_a \sigma_2 - \gamma_a \\ \alpha_b \bar{u}(b, \sigma_1, \sigma_2) + \beta_b \bar{u}'(b, \sigma_1, \sigma_2) - \gamma_b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Löse $g(\sigma_1, \sigma_2) = 0$ mit dem Newton-Verfahren. Bestimme dazu die zur Berechnung von $Dg(\sigma_1, \sigma_2)$ notwendigen Ableitungen $\frac{\partial}{\partial(\sigma_1, \sigma_2)} \bar{u}(b, \sigma_1, \sigma_2)$, $\frac{\partial}{\partial(\sigma_1, \sigma_2)} \bar{u}'(b, \sigma_1, \sigma_2)$ wie folgt. Sei

$$v(s, \sigma_1, \sigma_2) = \left(\frac{\partial}{\partial(\sigma_1, \sigma_2)} \bar{u}(s, \sigma_1, \sigma_2), \frac{\partial}{\partial(\sigma_1, \sigma_2)} \bar{u}'(s, \sigma_1, \sigma_2) \right), \quad (6-6)$$

und sei

$$F(s, w_1, w_2) = (w_2, -f(s, w_1))$$

das Vektorfeld der als System geschriebenen Differentialgleichung $-u''(s) = f(s, u(s))$ (siehe (6-5)). Durch Differentiation von (6-6) findet man, dass v die Anfangswertaufgabe

$$\begin{aligned}v'(s, \sigma_1, \sigma_2) &= \frac{\partial}{\partial(w_1, w_2)} F(s, \bar{u}(s, \sigma_1, \sigma_2), \bar{u}'(s, \sigma_1, \sigma_2)) \cdot v(s, \sigma_1, \sigma_2), \\ v(a, \sigma_1, \sigma_2) &= \frac{\partial}{\partial(\sigma_1, \sigma_2)} \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix} = I_2\end{aligned}$$

löst.

Bewertung des Schießverfahrens:

Vorteile: Es sind lediglich Anfangswertaufgaben zu lösen.

Probleme: Die für das Schießverfahren zu lösenden AWAs können schlecht konditioniert sein, selbst wenn die zu lösende Randwertaufgabe gutartig ist.

Beispiel:

$$\begin{aligned}-u''(s) &= -100u(s), \quad 0 \leq s \leq 3, \\ u(0) &= 1, \\ u(3) &= \exp(-30).\end{aligned}$$

Zu lösen sind Anfangswertaufgaben der Form

$$\begin{aligned}-u''(s) &= -100u(s), \\ u(0) &= 1, \\ u'(0) &= \sigma.\end{aligned}$$

Diese Familie von Anfangswertaufgaben hat die allgemeine Lösung

$$\bar{u}(s, \sigma) = \frac{1}{20}((10 - \sigma) \exp(-10s) + (10 + \sigma) \exp(10s)).$$

Man erhält ferner

$$\bar{u}(3, -10) = \frac{1}{20} 20 \exp(-30) = \exp(-30),$$

d.h. $\bar{\sigma} = -10$ ist Lösung von $g(\sigma) = \bar{u}(3, \sigma) - \exp(-30) = 0$. Berechne

$$\begin{aligned} \bar{u}(3, \bar{\sigma}(1 + \varepsilon)) &= \frac{1}{20} [(10 + 10(1 + \varepsilon)) \exp(-30) + (10 - 10(1 + \varepsilon)) \exp(30)] \\ &= \frac{1}{2} ((2 + \varepsilon) \exp(-30) - \varepsilon \exp(30)) \\ &= \exp(-30) - \frac{\varepsilon}{2} (\exp(30) - \exp(-30)). \end{aligned}$$

Relativer Fehler

$$\begin{aligned} \left| \frac{\bar{u}(3, \bar{\sigma}(1 + \varepsilon)) - \bar{u}(3, \bar{\sigma})}{\bar{u}(3, \bar{\sigma})} \right| &= \left| \frac{-\frac{\varepsilon}{2} (\exp(30) - \exp(-30))}{\exp(-30)} \right| \\ &= \frac{\varepsilon}{2} (\exp(60) - 1) \approx \varepsilon \cdot 6 \cdot 10^{25}. \end{aligned}$$

Mit $\varepsilon = 10^{-16}$ (Rechnerarithmetik mit 16 Stellen) folgt

$$\left| \frac{\bar{u}(3, \bar{\sigma}(1 + \varepsilon)) - \bar{u}(3, \bar{\sigma})}{\bar{u}(3, \bar{\sigma})} \right| \approx 6 \cdot 10^9.$$

7. Das klassische Differenzenverfahren

a) Aufstellung des Verfahrens

Betrachte die Dirichlet'sche Randwertaufgabe

$$\begin{aligned} -u''(s) &= f(s, u(s), u'(s)), \quad a \leq s \leq b, \\ u(a) &= \gamma_a, \\ u(b) &= \gamma_b. \end{aligned}$$

Wähle $h > 0$,

$$\Omega_h = \{s_j = a + jh, j = 0, \dots, \sigma(h)\}$$

mit $\sigma : \mathbb{R}_+ \rightarrow \mathbb{N}$ definiert durch

$$a + (\sigma(h) - 1)h < b \leq a + \sigma(h)h = b_h$$

Ohne Einschränkung kann man immer $b_h = b$ annehmen. Unsere Randwertaufgabe impliziert

$$-u''(s_j) = f(s_j, u(s_j), u'(s_j)), \quad s_j = a + jh, \quad j = 0, \dots, \sigma(h)$$

Wir benötigen also Approximationen für $-u''(s_j)$, $u'(s_j)$, $s_j \in \Omega_h$. Sei $s \in (a, b)$, $h > 0$ mit $s \pm h \in [a, b]$, dann gilt

$$u(s \pm h) = u(s) + u'(s)(\pm h) + \frac{1}{2}u''(s)h^2 + O(h^3)$$

falls $u \in C^3$. Daraus folgt

$$\begin{aligned} \frac{u(s+h) - u(s-h)}{2h} &= \frac{1}{2h}[u(s) + u'(s)h + \frac{1}{2}u''(s)h^2 - u(s) + u'(s)h \\ &\quad - \frac{1}{2}u''(s)h^2 + O(h^3)] \\ &= u'(s) + O(h^2) \end{aligned}$$

Analog bekommt man

$$\begin{aligned} \frac{-u(s+h) + 2u(s) - u(s-h))}{h^2} &= \frac{1}{h^2}[-u(s) - u'(s)h - \frac{1}{2}u''(s)h^2 - \frac{1}{6}u'''(s)h^3 \\ &\quad + 2u(s) \\ &\quad - u(s) + u'(s)h - \frac{1}{2}u''(s)h^2 + \frac{1}{6}u'''(s)h^3 + O(h^4)] \\ &= -u''(s) + O(h^2) \end{aligned}$$

Benutze dies, um $-u''(s_j)$ und $u'(s_j)$ zu approximieren. Der Ersetzungsprozess liefert

$$h^{-2}(-u(s_{j-1}) + 2u(s_j) - u(s_{j+1})) = f(s_j, u(s_j), \frac{1}{2h}(u(s_{j+1}) - u(s_{j-1}))),$$

$$\begin{aligned}
s_j &= a + jh, \quad j = 1, \dots, \sigma(h) - 1, \\
u(a) &= \gamma_a \\
u(b) &= u(s_{\sigma(h)}) = \gamma_b.
\end{aligned} \tag{7-1}$$

Das ist das klassische Differenzenverfahren. Es besteht aus $2 + \sigma(h) - 1 = \sigma(h) + 1$ Gleichungen für die $\sigma(h) + 1$ Unbekannten $u(s_j)$, $j = 0, \dots, \sigma(h)$.

Gesucht ist also ein Vektor

$$(u(a), u(a+h), \dots, u(b-h), u(b)) \in (\mathbb{R})^{\Omega_h},$$

welcher das Gleichungssystem (7-1) löst. Kompakt geschrieben hat (7-1) die Form

$$A^h u = G^h(u), \quad u \in (\mathbb{R})^{\Omega_h}$$

mit tridiagonaler Matrix

$$A^h = h^{-2} \begin{pmatrix} h^2 & & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & h^2 \end{pmatrix}$$

Schreibe hierfür kurz

$$A^h = h^{-2} \begin{pmatrix} h^2 & & \\ -1 & 2 & -1 \\ & & h^2 \end{pmatrix}.$$

Für

$$G^h(u) = \begin{pmatrix} f(a+h, u(a+h), \frac{\gamma_a}{2h}(u(a+2h) - u(a))) \\ \vdots \\ f(b-h, u(b-h), \frac{1}{2h}(u(b) - u(b-2h))) \\ \gamma_b \end{pmatrix}$$

benutzen wir die kurze Schreibweise

$$G^h(u) = \begin{pmatrix} \gamma_a \\ f(s, u(s), \frac{1}{2h}(u(s+h) - u(s-h))) \\ \gamma_b \end{pmatrix}, \quad s = a+h, \dots, a+\sigma(h)h-h.$$

Üblicherweise eliminiert man die Variablen $u(a)$, $u(b)$ und die Gleichungen $u(a) = \gamma_a$, $u(b) = \gamma_b$ aus dem System $A^h u = G^h(u)$. Für $s = a+h$ und $s = b-h$ folgt

$$\begin{aligned}
h^{-2}(-\gamma_a + 2u(a+h) - u(a+2h)) &= f(a+h, u(a+h), \frac{1}{2h}(u(a+2h) - \gamma_a)) \\
h^{-2}(-u(b-2h) + 2u(b-h) - \gamma_b) &= f(b-h, u(b-h), \frac{1}{2h}(\gamma_b - u(b-2h)))
\end{aligned}$$

und man erhält die eliminierte Form

$$\tilde{A}^h u = \tilde{G}^h(u)$$

mit

$$\tilde{A}^h = h^{-2} \begin{pmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -1 & 2 \end{pmatrix} \in \mathbb{R}^{\sigma(h)-1, \sigma(h)-1}$$

und

$$\tilde{G}^h(u) = \begin{pmatrix} f(a+h, u(a+h), \frac{1}{2h}(u(a+2h) - \gamma_a)) + \frac{\gamma_a}{h^2} \\ f(s, u(s), \frac{1}{2h}(u(s+h) - u(s-h))) \\ f(b-h, u(b-h), \frac{1}{2h}(\gamma_b - u(b-2h))) + \frac{\gamma_b}{h^2} \end{pmatrix}, \quad s = a+2h, \dots, b-2h.$$

b) Allgemeine Zweipunkt-Randwertaufgabe

Wir betrachten nun die Differentialgleichung

$$-u''(s) = f(s, u(s), u'(s)), \quad a \leq s \leq b$$

mit den allgemeinen Zweipunkt- oder ROBINSchen Randbedingungen

$$\begin{aligned} \alpha_a u(a) - \beta_a u'(a) &= \gamma_a, & \beta_a &\neq 0 \\ \alpha_b u(b) + \beta_b u'(b) &= \gamma_b, & \beta_b &\neq 0 \end{aligned}$$

Sei $h > 0$. Wir definieren ein Gitter

$$\Omega_h = \{s = a + jh, j = -1, 0, \dots, \sigma(h) + 1\},$$

d.h. $a-h, a, \dots, b, b+h \in \Omega_h$. Nun hat das Differenzenverfahren die Form

$$\begin{aligned} \alpha_a u(a) - \beta_a \frac{u(a+h) - u(a-h)}{2h} &= \gamma_a \\ h^{-2}(-u(s-h) + 2u(s) - u(s+h)) &= f(s, u(s), \frac{1}{2h}(u(s+h) - u(s-h))) \\ s &= a + jh, j = 0, \dots, \sigma(h) \\ \alpha_b u(b) + \beta_b \frac{u(b+h) - u(b-h)}{2h} &= \gamma_b \end{aligned} \quad (7-2)$$

d.h. $(\sigma(h)+3)$ Gleichungen für $(\sigma(h)+3)$ Unbekannte $(u(a-h), u(a), \dots, u(b), u(b+h))$. Damit sieht das zu lösende Gleichungssystem so aus:

$$A^h u = G^h(u)$$

mit

$$A^h = \begin{pmatrix} \frac{\beta_a}{2h} & \alpha_a & -\frac{\beta_a}{2h} & & \\ -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} \\ & & -\frac{\beta_b}{2h} & \alpha_b & \frac{\beta_b}{2h} \end{pmatrix}$$

$$G^h(u) = \begin{pmatrix} f(s, u(s), \frac{\gamma_a}{\gamma_b} \frac{u(s+h)-u(s-h)}{2h}) \\ \gamma_b \end{pmatrix}, \quad s = a + jh, \quad j = 0, \dots, \sigma(h).$$

Um die eliminierte Form zu bekommen, löst man die erste Zeile in (7-2) nach $u(a-h)$ und die letzte Zeile in (7-2) nach $u(b+h)$ auf. Setze dann $u(a-h)$ in die mittlere Zeile für von (7-2) für $j = 0$ und $u(b+h)$ für $j = \sigma(h)$ ein. Dies liefert ein System

$$\tilde{A}^h u = \tilde{G}^h(u)$$

mit $(\sigma(h) + 1)$ Gleichungen für $(\sigma(h) + 1)$ Unbekannte.

c) Inversmonotone Matrizen und inversmonotone Randwertprobleme

Ziel: Überprüfe, ob die qualitativen Eigenschaften der RWA durch die Diskretisierung mit dem klassischen Differenzenverfahren korrekt wiedergegeben werden.

Sei $D \subset \mathbb{R}$ nichtleer, und seien $u, v \in \mathbb{R}^D$ (d.h. Abbildungen $D \rightarrow \mathbb{R}$).

Schreibe $u \leq v$ (in D), falls $u(s) \leq v(s)$, $s \in D$ und $u < v$ (in D), falls $u(s) < v(s)$, $s \in D$. Die so erklärte Relation " $\leq$ " bzw. " $<$ " hat die üblichen Eigenschaften einer Halbordnung. Dann ist " $\leq$ " und " $<$ " auf $\mathbb{R}^{\sigma(h)+1}$ und $C([a, b])$ erklärt.

Betrachte die linearen Operatoren

$$L : C^2([a, b]) \rightarrow C([a, b]) \quad u \mapsto L(u) = -u'' + pu' + qu, \quad p, q \in C([a, b])$$

$$R : C^1([a, b]) \rightarrow \mathbb{R}^2 \quad u \mapsto R(u) = \begin{pmatrix} \alpha_a u(a) - \beta_a u'(a) \\ \alpha_b u(b) + \beta_b u'(b) \end{pmatrix}$$

Mithilfe von L und R lässt sich die RWA

$$\begin{aligned} -u''(s) &= f(s, u(s), u'(s)) = r(s) - q(s)u(s) - p(s)u'(s), \quad r, p, q \in C([a, b]) \\ \alpha_a u(a) - \beta_a u'(a) &= \gamma_a \\ \alpha_b u(b) + \beta_b u'(b) &= \gamma_b \end{aligned}$$

umschreiben als

$$\begin{aligned} Lu &= r \\ Ru &= \gamma, \quad \gamma := (\gamma_a, \gamma_b) \end{aligned}$$

7.1 Definition. Das Paar (L, R) heißt **invers monoton**, wenn die RWA

$$\begin{aligned} Lu &= r \\ Ru &= \gamma, \end{aligned}$$

für jedes $r \in C([a, b])$, $\gamma \in \mathbb{R}^2$ genau eine Lösung $\bar{u}$ besitzt, welche zusätzlich $\bar{u} \geq 0$ erfüllt, falls $r \geq 0$, und $\gamma \geq 0$. (d.h. $r(s) \geq 0$, $s \in [a, b]$, $\gamma_1 \geq 0$, $\gamma_2 \geq 0$)

Wir setzen die Bedingung

$$\beta_\sigma > 0 \text{ oder } (\beta_\sigma = 0, \alpha_\sigma > 0), \quad \sigma = a, b \quad (7-3)$$

voraus.

7.2 Satz. (ohne Beweis) Vorgelegt sei die RWA $Lu = r$, $Ru = \gamma$ mit $r \in C[a, b]$, und es gelte (7-3). Dann sind äquivalent

- (i) (L, R) ist invers monoton;
- (ii) Es gibt ein $e \in C^2([a, b])$ mit $e \geq 0$, $Le \geq 0$, $Re \geq 0$, aber nicht $Le = 0$, $Re = 0$. (e heißt majorisierendes Element)

7.3 Korollar. Sei $q \geq 0$ und es gelte

$$\begin{aligned} \alpha_\sigma, \beta_\sigma &\geq 0, \quad \alpha_\sigma + \beta_\sigma > 0, \quad \sigma = a, b, \\ \alpha_a + \alpha_b &> 0. \end{aligned} \quad (7-4)$$

Dann ist (L, R) invers monoton.

Beweis: Konstruiere ein majorisierendes Element durch $e(s) \equiv 1$, $a \leq s \leq b$, $e \in C^2([a, b])$. Dann gilt

$$\begin{aligned} Le &= -\underbrace{e''}_{=0} + p \underbrace{e'}_{=0} + qe = q \geq 0 \\ Re &= \begin{pmatrix} \alpha_a e(a) - \beta_a e'(a) \\ \alpha_b e(b) + \beta_b e'(b) \end{pmatrix} = \begin{pmatrix} \alpha_a \\ \alpha_b \end{pmatrix} \geq 0 \end{aligned}$$

□

Wir nehmen die zusammengefasste Theorie zur Behandlung linearer Randwertaufgaben als Richtschnur für die Behandlung der beim klassischen Differenzenverfahren auftretenden Gleichungssysteme.

7.4 Definition. Eine Matrix $A \in \mathbb{R}^{N,N}$ heißt nichtnegativ oder monoton, falls $A_{ij} \geq 0$ für $1 \leq i, j \leq N$.

Wir schreiben hierfür einfach $A \geq 0$. Für $A, B \in \mathbb{R}^{N,N}$ setzt man $A \leq B$ genau dann, wenn $B - A \geq 0$.

7.5 Definition. • $A \in \mathbb{R}^{N,N}$ heißt invers monoton, falls A^{-1} existiert und $A^{-1} \geq 0$ gilt.

- A heißt L_0 -Matrix, falls $A_{ij} \leq 0$ für $1 \leq i, j \leq N$ mit $i \neq j$.
- A heißt M-Matrix, falls A eine invers monotone L_0 -Matrix ist.

Wir setzen $\mathbb{R}_+^{N,N} := \{A \in \mathbb{R}^{N,N}; A \geq 0\}$.

7.6 Satz (M-Kriterium, ohne Beweis). Eine L_0 -Matrix $A \in \mathbb{R}^{N,N}$ ist genau dann eine M-Matrix, wenn ein $e \in \mathbb{R}^N$, $e > 0$ existiert mit $Ae \geq 0$ und der Eigenschaft

Zu jedem $i_0 \in \{1, \dots, N\}$ mit $(Ae)_{i_0} = 0$ gibt es eine Kette $i_0, i_1 \dots i_r \in \{1, \dots, N\}$ mit $(Ae)_{i_r} > 0$, $A_{i_{j-1}, i_j} \neq 0$, $j = 1, \dots, r$.
Verbindungseigenschaft (VE)

Ein $e > 0$ mit $Ae \geq 0$ und (VE) heißt majorisierendes Element für A .

7.7 Bemerkung. Zu einer L_0 -Matrix $A \in \mathbb{R}^{N,N}$ gibt es genau dann ein majorisierendes Element e , wenn es ein $e > 0$ gibt mit $Ae > 0$.

d) Nichtlineare Gleichungssysteme

Die Diskretisierung einer Randwertaufgabe

$$\begin{aligned} -u''(s) &= f(s, u(s), u'(s)), & a \leq s \leq b, \\ \alpha_a u(a) - \beta_a u'(a) &= \gamma_a, & \alpha_b u(b) + \beta_b u'(b) = \gamma_b \end{aligned}$$

mit dem klassischen Differenzenverfahren führt auf nichtlineare Gleichungssysteme der Form $Au = G(u)$. Wir untersuchen deshalb

$$Au = G(u), \quad A \in \mathbb{R}^{N,N}, \quad G: \mathbb{R}^N \rightarrow \mathbb{R}^N$$

auf Existenz bzw. Existenz und Eindeutigkeit von Lösungen.

7.8 Definition. • Für $u \in \mathbb{R}^N$ sei $|u| \in \mathbb{R}^N$ erklärt durch $|u|_i = |u_i|$, $i = 1, \dots, N$.

- Eine Abbildung $T: \mathbb{R}^N \rightarrow \mathbb{R}^N$ heißt P-beschränkt, falls $P \in \mathbb{R}_+^{N,N}$ und

$$|T(u) - T(v)| \leq P |u - v|, \quad \forall u, v \in \mathbb{R}^N.$$

- Zu $e > 0$, $e \in \mathbb{R}^N$ fest wird durch $\|u\|_e := \max\{\frac{|u_i|}{e_i}; i = 1, \dots, N\}$ eine Norm auf $\mathbb{R}^N$ definiert.

7.9 Bemerkung. Für die zur e -Norm gehörige Operatornorm ist folgende Aussage von Bedeutung. Für $P \in \mathbb{R}_+^{N,N}$ gilt $\|P\|_e = \|Pe\|_e$.

7.10 Korollar. Ist $T : \mathbb{R}^N \rightarrow \mathbb{R}^N$ P -beschränkt und ist $e > 0$, so gilt

$$\|T(u) - T(v)\|_e \leq \|P\|_e \|u - v\|_e, \quad \forall u, v \in \mathbb{R}^N.$$

Beweis. Nach Definition der P -Beschränktheit folgt

$$\frac{|(T(u))_i - (T(v))_i|}{e_i} \leq \frac{(P|u - v|)_i}{e_i}, \quad i = 1, \dots, N$$

und wir erhalten

$$\|T(u) - T(v)\|_e \leq \|P|u - v|\|_e \leq \|P\|_e \cdot \| |u - v| \|_e = \|P\|_e \cdot \|u - v\|_e.$$

□

7.11 Satz. Vorgelegt sei das Gleichungssystem $Au = G(u)$ mit einer invers monotonen L_0 -Matrix $A \in \mathbb{R}^{N,N}$ und einer P -beschränkten Abbildung $G : \mathbb{R}^N \rightarrow \mathbb{R}^N$. Ist dann $A - P$ invers monoton, so besitzt $Au = G(u)$ genau eine Lösung $\bar{u} \in \mathbb{R}^N$. Für die direkte Iteration

$$Au^{n+1} = G(u^n), \quad u^0 \in \mathbb{R}^N \text{ beliebig}$$

gilt $\lim_{n \rightarrow \infty} u^n = \bar{u}$. Überdies gilt für den Operator $T := A - G$ die Ungleichung

$$|u - v| \leq (A - P)^{-1} |T(u) - T(v)| \quad \forall u, v \in \mathbb{R}^N. \quad (7-5)$$

7.12 Bemerkung. Die Ungleichung (7-5) ist eine vektorwertige Stabilitätsungleichung. (7-5) impliziert

$$\|u - v\|_e \leq \|(A - P)^{-1}\|_e \cdot \|T(u) - T(v)\|_e \quad \forall u, v \in \mathbb{R}^N.$$

Beweis von Satz 7.11. Wir setzen $H(u) := A^{-1}G(u)$. H ist $A^{-1}P$ -beschränkt, denn

$$\begin{aligned} |H(u) - H(v)| &= |A^{-1}(G(u) - G(v))| \leq A^{-1} |G(u) - G(v)| \\ &\leq A^{-1}P |u - v|. \end{aligned} \quad (7-6)$$

Nach Korollar 7.10 gilt also

$$\|H(u) - H(v)\|_e \leq \|A^{-1}P\|_e \cdot \|u - v\|_e \quad u, v \in \mathbb{R}^N$$

für jedes $e > 0$. Wir möchten nun den Kontraktionssatz auf T anwenden. Dazu ist zu zeigen, dass ein $e > 0$ existiert mit $\|A^{-1}P\|_e < 1$.

Wir setzen $e := (A - P)^{-1}\mathbb{I}$ und finden $e > 0$, da $A - P$ invers monoton. Ferner gilt

$$e - A^{-1}Pe = (I - A^{-1}P)e = A^{-1}(A - P)e = A^{-1}\mathbb{I} > 0, \text{ d.h. } A^{-1}Pe < e.$$

Dies impliziert mit Bemerkung 7.9

$$\|A^{-1}P\|_e = \|A^{-1}Pe\|_e < \|e\|_e = 1.$$

Also ist der Kontraktionssatz anwendbar und liefert die eindeutige Existenz einer Lösung $\bar{u}$ von $u = H(u)$, d.h. $Au = G(u)$. Ferner gilt für die Iteration

$$u^{n+1} = H(u^n) = A^{-1}G(u^n) \longleftrightarrow Au^{n+1} = G(u^n)$$

überdies $\lim_{n \rightarrow \infty} u^n = \bar{u}$.

Noch zu zeigen bleibt die Ungleichung (7-5). Mit $T(u) := Au - G(u)$ und (7-6) folgt

$$\begin{aligned} |u - v - A^{-1}(G(u) - G(v))| &\geq |u - v| - |A^{-1}(G(u) - G(v))| \\ &\geq |u - v| - A^{-1}P |u - v| = (I - A^{-1}P) |u - v| \end{aligned} \quad (7-7)$$

sowie

$$\begin{aligned} |u - v - A^{-1}(G(u) - G(v))| &= |A^{-1}(A(u - v) - (G(u) - G(v)))| \\ &= |A^{-1}(T(u) - T(v))| \leq A^{-1} |T(u) - T(v)|. \end{aligned} \quad (7-8)$$

(7-7) und (7-8) liefern

$$A^{-1} |T(u) - T(v)| \geq (I - A^{-1}P) |u - v| = A^{-1}(A - P) |u - v|. \quad (7-9)$$

Wir multiplizieren (7-9) mit $(A - P)^{-1}A = (A - P)^{-1}(A - P + P) = I + (A - P)^{-1}P \geq 0$ und finden

$$(A - P)^{-1} |T(u) - T(v)| \geq (A - P)^{-1}AA^{-1}(A - P) |u - v| = |u - v|.$$

□

7.13 Definition. $F : \mathbb{R}^N \rightarrow \mathbb{R}^N$ heißt monoton, falls gilt

$$F(u) \leq F(v) \quad \forall u, v \in \mathbb{R}^N \text{ mit } u \leq v.$$

Sind $F, G : \mathbb{R}^N \rightarrow \mathbb{R}^N$ gegeben, und ist $G - F$ monoton, so schreiben wir $F \leq G$.

Diese Definition erweitert den Begriff der monotonen linearen Abbildung von $\mathbb{R}^{N,N}$ auf den Raum der nichtlinearen Abbildungen $\mathbb{R}^N \rightarrow \mathbb{R}^N$.

Sei nun $G : \mathbb{R}^N \rightarrow \mathbb{R}^N$ und $Q, P \in \mathbb{R}^{N,N}$ mit $Q \leq G \leq P$, d.h.

$$Q(v - u) \leq G(v) - G(u) \leq P(v - u) \quad \forall u, v \in \mathbb{R}^N \text{ mit } u \leq v.$$

Analog zu Satz 7.11 kann man in dieser Situation zeigen

7.14 Satz (ohne Beweis). Sei A eine L_0 -Matrix, $G : \mathbb{R}^N \rightarrow \mathbb{R}^N$ erfülle $Q \leq G \leq P$ für $Q, P \in \mathbb{R}^{N,N}$. Ferner sei $R := \frac{1}{2}(P+Q)$ und $A-R$, $A-P$ seien invers monoton. Dann besitzt die Gleichung $Au = G(u)$ genau eine Lösung $\bar{u} \in \mathbb{R}^N$. Überdies gilt für den Operator $T := A - G$ die Ungleichung

$$|u - v| \leq (A - P)^{-1} |T(u) - T(v)| \quad \forall u, v \in \mathbb{R}^N. \quad (7-10)$$

e) Anwendung auf Randwertaufgaben

Vorgelegt sei wieder die Differentialgleichung

$$-u''(s) = f(s, u(s)), \quad a \leq s \leq b \quad (7-11)$$

mit den allgemeinen Zweipunkt-Randbedingungen

$$\begin{aligned} \alpha_a u(a) - \beta_a u'(a) &= \gamma_a, \\ \alpha_b u(b) + \beta_b u'(b) &= \gamma_b, \end{aligned} \quad (7-12)$$

und es sei $A^h u = F^h(u)$ das durch Diskretisierung von (7-11), (7-12) mit dem klassischen Differenzenverfahren in eliminerter Form entstehende nichtlineare Gleichungssystem. Dann findet man mit $b_h = a + \sigma(h)h$

- im Fall von Dirichlet-Randbedingungen ($\alpha_a = \alpha_b = 1, \beta_a = \beta_b = 0$)

$$\begin{aligned} A^h &= \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \\ F^h(u) &= \begin{pmatrix} f(a+h, u(a+h)) + \frac{\gamma_a}{h^2} \\ f(s, u(s)) \\ f(b_h - h, u(b_h - h)) + \frac{\gamma_b}{h^2} \end{pmatrix}, \quad s = a + jh, \quad j = 2, \dots, \sigma(h) - 2. \end{aligned}$$

- bei allgemeinen Zweipunkt-Randbedingungen ($\beta_a \neq 0, \beta_b \neq 0$)

$$A^h = \frac{1}{h^2} \begin{pmatrix} 2 + \frac{2\alpha_a h}{\beta_a} & -2 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -2 & 2 + \frac{2\alpha_b h}{\beta_b} \end{pmatrix}$$

$$F^h(u) = \begin{pmatrix} f(a, u(a)) + \frac{2\gamma_a}{h\beta_a} \\ f(s, u(s)) \\ f(b_h, u(b_h)) + \frac{2\gamma_b}{h\beta_b} \end{pmatrix}, \quad s = a + jh, \quad j = 1, \dots, \sigma(h) - 1$$

sowie entsprechend in den gemischten Fällen.

Für unsere Analyse setzen wir die Vorzeichenbedingung

$$\alpha_a, \beta_a, \alpha_b, \beta_b \geq 0, \quad \alpha_a + \alpha_b > 0, \quad \alpha_a + \beta_a > 0, \quad \alpha_b + \beta_b > 0 \quad (7-13)$$

voraus. Zusätzlich sei jetzt eine Lipschitzbedingung angenommen.

Es gibt $q, \mu \in C([a, b])$ mit

$$q(s)(u_1 - u_2) \leq f(s, u_1) - f(s, u_2) \leq \mu(s)(u_1 - u_2), \quad a \leq s \leq b, \quad u_1 \geq u_2. \\ (LIP)$$

Wir setzen

$$Q^h = \text{diag}(q(s); s \in \Omega_h), \quad P^h = \text{diag}(\mu(s); s \in \Omega_h).$$

Für je zwei Vektoren $u, v \in \mathbb{R}^{\Omega_h}$ mit $v \geq u$ gilt dann

$$(Q^h(v - u))(s) = q(s)((v(s) - u(s)) \leq f(s, v(s)) - f(s, u(s)) = \\ (F^h(v))(s) - (F^h(u))(s) \leq \mu(s)(v(s) - u(s)) = P^h(v - u)(s), \quad s \in \Omega_h,$$

d.h. $Q^h \leq F^h \leq P^h$.

7.15 Korollar. *Unter den Voraussetzungen (7-13) und (LIP) besitzt das klassische Differenzenverfahren zu (7-11), (7-12) genau eine Lösung $\bar{u} \in \mathbb{R}^{\Omega_h}$, falls die Matrix $A^h - P^h$ invers monoton ist.*

Beweis. Da A^h eine L_0 -Matrix und $A^h - P^h$ invers monoton ist existiert nach dem M-Kriterium und Bemerkung 7.7 ein $e > 0$ mit $(A^h - P^h)e > 0$. Mit $R^h = \frac{1}{2}(P^h + Q^h)$ folgt $R^h \leq P^h$ und somit

$$(A^h - R^h)e \geq (A^h - P^h)e > 0,$$

d.h. ist nach dem M-Kriterium ebenfalls invers monoton. Also zeigt Satz 7.14 die Behauptung. $\square$

8. Konvergenz des klassischen Differenzenverfahrens

a) Konsistenz des Verfahrens

Betrachte die Dirichlet'sche Randwertaufgabe

$$\begin{aligned} -u''(s) &= f(s, u(s), u'(s)), & a \leq s \leq b, \\ u(a) &= \gamma_a, \\ u(b) &= \gamma_b \end{aligned}$$

mit der Lösung $\bar{u} \in C^2[a, b]$. Es sei $h > 0$,

$$\Omega_h = \{s_j = a + jh, j = 0, \dots, \sigma(h)\}$$

mit $\sigma : \mathbb{R}_+ \rightarrow \mathbb{N}$ definiert durch

$$a + (\sigma(h) - 1)h < b \leq a + \sigma(h)h = b_h.$$

Ohne Einschränkung kann man immer $b_h = b$ annehmen.

Nun gilt in nicht eliminierter Form $T^h(u) := A^h u - G^h(u) = 0$ mit

$$A^h = h^{-2} \begin{pmatrix} h^2 & & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & & h^2 \end{pmatrix}$$

und

$$G^h(u) = \begin{pmatrix} \gamma(a) \\ f(a+h, u(a+h), \frac{1}{2h}(u(a+2h) - u(a))) \\ \vdots \\ f(b-h, u(b-h), \frac{1}{2h}(u(b) - u(b-2h))) \\ \gamma(b) \end{pmatrix}.$$

8.1 Satz. Sei f lokal lipschitz stetig in der dritten Variablen, d.h. zu beliebig vorgegebenen $K_1, K_2 > 0$ existiert ein $L = L(K_1, K_2) > 0$ mit

$$\begin{aligned} |f(s, u, v_1) - f(s, u, v_2)| &\leq L|v_1 - v_2|, \\ a \leq s \leq b, |u| &\leq K_1, |v_1|, |v_2| \leq K_2. \end{aligned}$$

Das nicht eliminierte klassische Differenzenverfahren $T^h(u) = 0$ ist C^4 -konsistent der Ordnung 2 für Nullfolgen $(h_n)_{n \in \mathbb{N}}$ mit $h_n = (b-a)/M_n$.

Beweis. Wir berechnen den Konsistenzfehler. Dabei setzen wir $\bar{u}_h := \bar{u}|_{\Omega_h}$.

Sei $s \in \{a, b\}$, so gilt

$$T^h(\bar{u}_h)(s) := \bar{u}(s) - \gamma_s = 0.$$

Im Fall $s \in \{a + h, a + 2h, \dots, b - h\}$ findet man

$$\begin{aligned} |T^h(\bar{u}_h)(s)| &= \left| \frac{1}{h^2} (-\bar{u}(s-h) + 2\bar{u}(s) - \bar{u}(s+h)) + \underbrace{\bar{u}''(s)}_{=-f(s, \bar{u}(s), \bar{u}'(s))} \right. \\ &\quad \left. + f(s, \bar{u}(s), \bar{u}'(s)) - f(s, \bar{u}(s), \frac{1}{2h}(\bar{u}(s+h) - \bar{u}(s-h))) \right| \\ &\leq \left| \frac{1}{h^2} (-\bar{u}(s-h) + 2\bar{u}(s) - \bar{u}(s+h)) + \bar{u}''(s) \right| + \\ &\quad + \left| f(s, \bar{u}(s), \bar{u}'(s)) - f(s, \bar{u}(s), \frac{1}{2h}(\bar{u}(s+h) - \bar{u}(s-h))) \right| \\ &\leq \left| \frac{1}{h^2} (-\bar{u}(s-h) + 2\bar{u}(s) - \bar{u}(s+h)) + \bar{u}''(s) \right| + \\ &\quad + L \left| \bar{u}'(s) - \frac{1}{2h} (\bar{u}(s+h) - \bar{u}(s-h)) \right|. \end{aligned} \quad (8-1)$$

Um die C^4 -Konsistenz der Ordnung 2 in (8-1) zu zeigen, benötigen wir

$$\left| \frac{1}{h^2} (-\bar{u}(s-h) + 2\bar{u}(s) - \bar{u}(s+h)) + \bar{u}''(s) \right| \leq C_1 h^2, \quad 0 < h \leq h_0, \quad (8-2)$$

und

$$\left| \frac{1}{2h} (\bar{u}(s+h) - \bar{u}(s-h)) - \bar{u}'(s) \right| \leq C_2 h^2, \quad 0 < h \leq h_0. \quad (8-3)$$

Unterstellen wir (8-2), (8-3), so folgt

$$|T^h(\bar{u}_h)(s)| \leq C_1 h^2 + LC_2 h^2, \quad 0 < h \leq h_0, \quad s = a + h, \dots, b - h,$$

d.h. C^4 -Konsistenz der Ordnung 2.

Noch zu zeigen bleiben (8-2), (8-3). Nach Taylor gilt für $\bar{u} \in C^4$:

$$\begin{aligned} \bar{u}(s \pm h) &= \bar{u}(s) \pm h\bar{u}'(s) + \frac{h^2}{2}\bar{u}''(s) \pm \frac{h^3}{6}\bar{u}'''(s) \\ &\quad + \frac{h^4}{24}u''''(\eta_{\pm}), \quad \eta_- \in [s-h, s], \quad \eta_+ \in [s, s+h]. \end{aligned} \quad (8-4)$$

Somit folgt mit (8-4)

$$\left| \frac{1}{2h} (\bar{u}(s+h) - \bar{u}(s-h)) - \bar{u}'(s) \right|$$

$$\begin{aligned}
&= \left| \frac{1}{2h}(\bar{u}(s) + \bar{u}'(s)h + \frac{h^2}{2}\bar{u}''(s) + \frac{h^3}{6}\bar{u}'''(s) + \frac{h^4}{24}\bar{u}''''(\eta_+) \right. \\
&\quad \left. - \bar{u}(s) + h\bar{u}'(s) - \frac{h^2}{2}\bar{u}''(s) + \frac{h^3}{6}\bar{u}'''(s) - \frac{h^4}{24}\bar{u}''''(\eta_-) \right. \\
&\quad \left. - \bar{u}'(s) \right| \\
&= \frac{h^2}{6} \left| \bar{u}'''(s) + \frac{h}{8}(\bar{u}''''(\eta_+) - \bar{u}''''(\eta_-)) \right| \leq C_2 h^2
\end{aligned}$$

mit $C_2 > 0$ geeignet.

Mit (8-4) finden wir überdies

$$\begin{aligned}
&\left| \frac{1}{h^2}(-\bar{u}(s-h) + 2\bar{u}(s) - \bar{u}(s+h)) + \bar{u}''(s) \right| \\
&= \left| h^{-2} \left(-\bar{u}(s) + h\bar{u}'(s) - \frac{h^2}{2}\bar{u}''(s) + \frac{h^3}{6}\bar{u}'''(s) - \frac{h^4}{24}\bar{u}''''(\eta_-) \right. \right. \\
&\quad \left. \left. + 2\bar{u}(s) \right. \right. \\
&\quad \left. \left. - \bar{u}(s) - h\bar{u}'(s) - \frac{h^2}{2}\bar{u}''(s) - \frac{h^3}{6}\bar{u}'''(s) - \frac{h^4}{24}\bar{u}''''(\eta_-) \right) + \bar{u}''(s) \right| \\
&= \frac{h^2}{24} |\bar{u}''''(\eta_-) + \bar{u}''''(\eta_+)| \leq C_1 h^2.
\end{aligned}$$

□

8.2 Bemerkung. Im Fall $\bar{u} \in C^2$ erhält man lediglich pure Konsistenz, d.h. $T^h(u) = 0$ ist C^2 -konsistent.

b) Stabilität des klassischen Differenzen-Verfahrens

Vorgelegt sei die Randwertaufgabe

$$\begin{aligned}
-u''(s) + pu'(s) &= f(s, u(s)), \quad a \leq s \leq b, \\
u(a) &= \gamma_a, \\
u(b) &= \gamma_b.
\end{aligned} \tag{8-5}$$

Es gelte $p \in C[a, b]$, $f \in C([a, b] \times \mathbb{R})$. Ferner gelte für zwei Funktionen $q, \mu \in C[a, b]$

$$\begin{aligned}
q(s)(u_1 - u_2) &\leq f(s, u_1) - f(s, u_2) \leq \mu(s)(u_1 - u_2), \\
a \leq s \leq b, \quad u_1, u_2 &\in \mathbb{R}, \quad u_1 \geq u_2.
\end{aligned} \tag{LIP}$$

Desweiteren sei mit μ aus (LIP) das Paar (L, R) gegeben durch

$$Lu := -u'' + pu' - \mu u, \quad Ru := (u(a), u(b)) \tag{8-6}$$

8.3 Satz (ohne Beweis). Betrachtet werde die Randwertaufgabe (8-5). Es gelte (LIP) und das Paar (L, R) aus (8-6) sei invers monoton. Dann hat die nichtlineare Randwertaufgabe (8-5) genau eine Lösung $\bar{u} \in C^2([a, b])$.

8.4 Satz (Konvergenzsatz). Vorgelegt sei die Randwertaufgabe (8-5) unter den Voraussetzungen von Satz 8.3. Dann besitzt das klassische Differenzenverfahren zu (8-5) für hinreichend kleines h genau eine Lösung u^h . Das Verfahren ist C^4 -konvergent der Ordnung 2 für Nullfolgen der Form $(h_n)_{n \in \mathbb{N}}$, $h_n = (b - a)/M_n$ gegen die eindeutige Lösung $\bar{u}$ von (8-5).

Beweis. In dieser Situation haben wir $T^h(u) = A^h u - F^h(u)$ mit der Gitterfunktion $u = (u(a), u(a + h), \dots, u(b - h), u(b))$ und

$$A^h = \begin{pmatrix} 1 & & & \\ -h^{-2} - \frac{p(s)}{2h} & 2h^{-2} & -h^{-2} + \frac{p(s)}{2h} & \\ & & & 1 \end{pmatrix},$$

$$F^h(u) = \begin{pmatrix} \gamma_a \\ f(s, u(s)) \\ \gamma_b \end{pmatrix}, \quad s = a + h, \dots, b - h.$$

Die Konsistenz des Systems ist dabei nach Satz 8.1 klar.

Um die Stabilität einzusehen, gehen wir wie folgt vor. Mit (LIP) und

$$Q^h = \text{diag}(0, q(a + h), \dots, q(b - h), 0),$$

$$P^h = \text{diag}(0, \mu(a + h), \dots, \mu(b - h), 0)$$

folgt $Q^h \leq F^h \leq P^h$. Ferner ist A^h eine L_0 -Matrix, falls $|p(s)| < \frac{2}{h}$ für $a \leq s \leq b$. Dies ist für $h > 0$ hinreichend klein immer der Fall.

Da das Paar (L, R) invers monoton ist, existiert nun eine nicht negative Lösung der Aufgabe

$$-e'' + pe' - \mu e = 1, \quad e(a) = 1, \quad e(b) = 1. \quad (8-7)$$

Für die Lösung e aus (8-7) zeigen wir im Folgenden:

- i) Es gilt $0 < e(s)$, $a \leq s \leq b$.
- ii) Es existiert ein $C^* > 0$ mit $(A^h - P^h)e_h \geq C^* \mathbb{I}$ für $0 < h \leq h_0$. Dabei bezeichnet $e_h := e|_{\Omega_h}$.

Wir unterstellen zunächst i), ii) und schließen weiter. ii) stellt sicher, dass $A^h - P^h$ eine M-Matrix ist. Ferner gilt mit $R^h := \frac{1}{2}(Q^h + P^h) \leq P^h$ sofort

$$(A^h - R^h)e_h \geq (A^h - P^h)e_h \geq C^* \mathbb{I},$$

d.h. $A^h - R^h$ ist nach Bemerkung 7.7 ebenfalls eine M-Matrix.

Damit sind die Voraussetzungen von Satz 7.14 erfüllt. Das System $T^h(u) = A^h u - F^h(u) = 0$ hat also für $h > 0$ hinreichend klein genau eine Lösung u^h , und es gilt die Stabilitätsungleichung

$$\|u - v\|_\infty \leq \|(A^h - P^h)^{-1}\|_\infty \cdot \|T^h(u) - T^h(v)\|_\infty, \quad u, v \in \mathbb{R}^{\Omega_h}, \quad 0 < h \leq h_0.$$

Ferner folgt aus ii) sofort $(A^h - P^h)^{-1}\mathbb{I} \leq \frac{1}{C^*}e_h$, und somit erhalten wir

$$\begin{aligned} \|(A^h - P^h)^{-1}\|_\infty &= \|(A^h - P^h)^{-1}\mathbb{I}\|_\infty \leq \frac{1}{C^*}\|e_h\|_\infty \\ &\leq \frac{1}{C^*} \max\{|e(s)|; a \leq s \leq b\}. \end{aligned}$$

Damit ist die Stabilität dann gezeigt.

Noch nachzutragen bleibt der Beweis von i) und ii).

Beweis von i): Wegen $e(a) = e(b) = 1 > 0$ kann $e(\bar{s}) = 0$ nur für ein $\bar{s} \in]a, b[$ auftreten. Wegen $e(s) \geq 0$ auf $[a, b]$ folgt dann aber $e'(\bar{s}) = 0$ und $e''(\bar{s}) \geq 0$. Dies steht im Widerspruch zu

$$1 = (Le)(\bar{s}) = (-e'' + pe' - \mu e)(\bar{s}) = -e''(\bar{s}).$$

Beweis von ii): Das klassische Differenzenverfahren zu $Le = 1$, $Re = (1, 1)$ lautet $B^h u = M^h u + \mathbb{I}$ mit

$$B^h = \begin{pmatrix} 1 & & \\ -h^{-2} - \frac{p(s)}{2h} & 2h^{-2} & -h^{-2} + \frac{p(s)}{2h} \\ & 1 & \end{pmatrix}, \quad s = a + h, \dots, b - h$$

und

$$M^h = \text{diag}(0, \mu(a + h), \dots, \mu(b - h), 0).$$

Dies liefert $B^h = A^h$ und $M^h = P^h$. Wir erhalten mit $e_h := e|_{\Omega_h}$ sofort

$$(A^h - P^h)e_h = \mathbb{I} + (A^h e_h - P^h e_h - \mathbb{I}) \geq \mathbb{I} - \underbrace{\|(A^h - P^h)e_h - \mathbb{I}\|_\infty}_{\rightarrow 0 \text{ für } h \rightarrow 0} \mathbb{I} \geq \frac{1}{2}\mathbb{I},$$

für $h > 0$ hinreichend klein, da das klassische Differenzenverfahren C^2 -konsistent ist und ii) ist gezeigt. $\square$

8.5 Bemerkung. Werden bei der Randwertaufgabe (8-5) die Dirichlet'schen durch allgemeine Zweipunkt-Randbedingungen ersetzt und gelten die Vorzeichenvoraussetzungen (7-13), so lässt für das klassische Differenzenverfahren zu dieser Aufgabe ebenfalls C^4 -Konvergenz der Ordnung 2 sicherstellen.