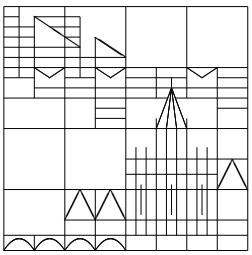
Skript zur Vorlesung

Numerik gewöhnlicher Differentialgleichungen

Sommersemester 2025

Johannes Schropp



Universität Konstanz

Fachbereich Mathematik und Statistik

Stand: 13. Mai 2025

Inhaltsverzeichnis

1. Beispiele und einiges über Differentialgleichungen

a) Beispiele

1.1 Beispiel. Fallschirmspringer
Das Grundgesetz der Mechanik lautet:

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

$$K = m \cdot \ddot{x} \tag{1-1}$$

Durch Angabe der Abhängigkeiten der Kraft

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

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

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

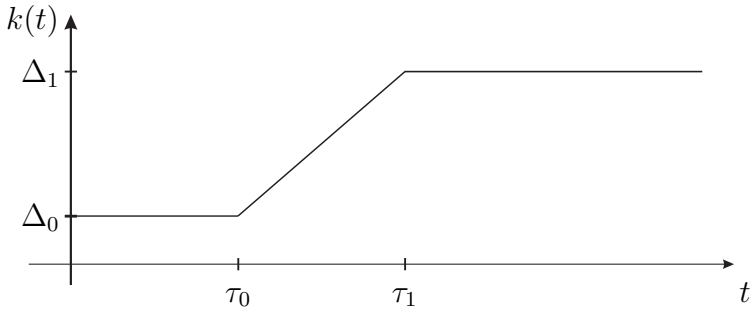


Abbildung 1: Beispiel für $k(t)$

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:

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

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

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

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

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

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

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

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 $(\overline{q}(t), \overline{p}(t))$, $t \in I$ von (1-6) konstant. Es gilt*

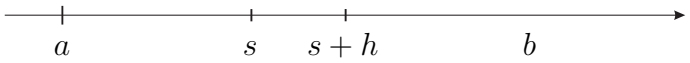
$$\frac{\mathrm{d}}{\mathrm{d}t} H(\overline{q}(t), \overline{p}(t)) = \frac{\partial H}{\partial q}\dot{\overline{q}}(t) + \frac{\partial H}{\partial p}\dot{\overline{p}}(t) = \frac{\partial H}{\partial q}\frac{\partial H}{\partial p} + \frac{\partial H}{\partial p}\left(-\frac{\partial H}{\partial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

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

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

Dividieren durch h liefert

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

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

$$\int_a^b f(x)\mathrm{d}x = f(x_0)(b - a), \quad x_0 \in [a, b]$$

liefert der Grenzübergang $h \to 0$ für

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

die Differentialgleichung

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

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

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

für $a \le s \le b,\ t \ge 0$.

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

Anfangsbedingungen

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

Randbedingungen

Hier sind folgende Bedingungen möglich:

- **Dirichlet-Randbedingungen**

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

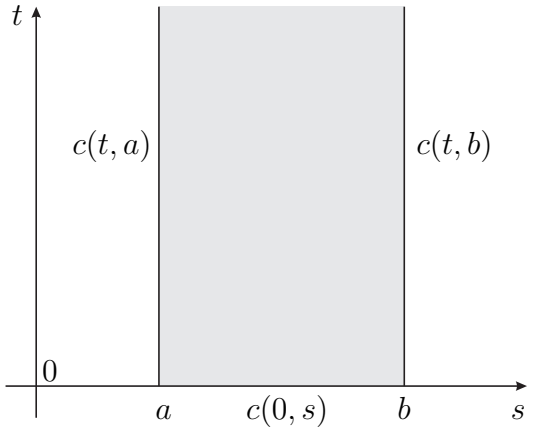


Abbildung 3: Zusatzbedingungen für c

- **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**

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

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

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

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 $\overline{x}(t)$ 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 $\overline{x}(t, t_0, x_0)$, $t \in J(x_0)$. 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

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

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

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

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

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

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

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

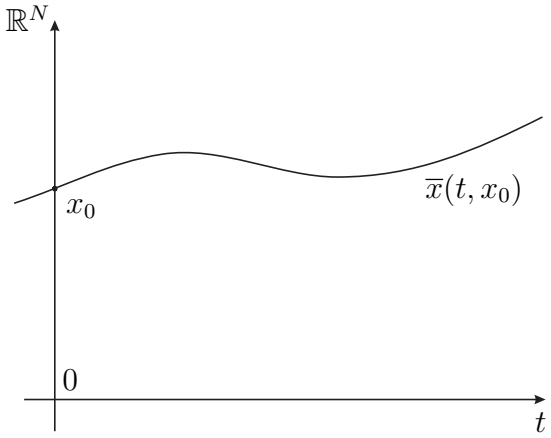


Abbildung 4: Zeitbild

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

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

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.

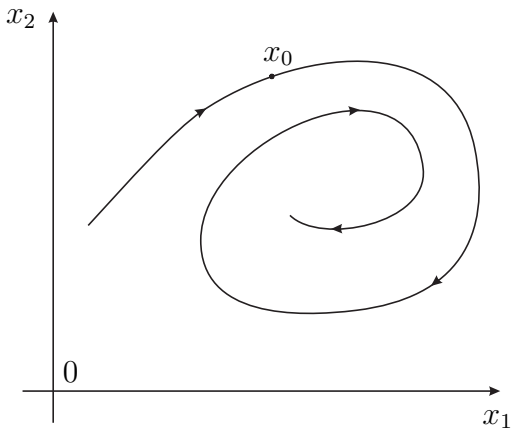


Abbildung 5: Phasenbild: $N = 2$

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

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

den so genannten τ-Fluss.

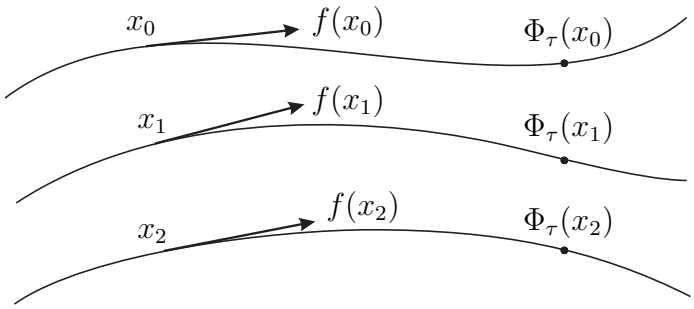


Abbildung 6: Phasenbild: $N = 2$

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

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

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 $\Phi_0 = \mathrm{id}_{\mathbb{R}}$ 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 $\overline{x}(t, \xi, \lambda) \in \Omega$ für $t \in J(\xi, \lambda) =]t^-(\xi, \lambda), t^+(\xi, \lambda)[$. Der Definitionsbereich von $\overline{x}$

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

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

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

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

mit $f \in C^1$. Bezeichne durch $\gamma(x_0) = \{\overline{x}(t, x_0), t \in Jx_0\}$ den zu x_0 gehörigen Orbit. Die Größe $\partial \overline{x}(t, x_0)/\partial x_0$, $t \in J(x_0)$ bezeichnet man als die *Linearisierung* von $\overline{x}(t, x_0)$. Es gilt gemäß Satz 1.4

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

und

$$\frac{\partial}{\partial x_0}\overline{x}(0, x_0) = \frac{\partial}{\partial x_0} x_0 = \mathrm{I}_N\,.$$

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

$$\begin{aligned} X'(t) &= Df(\overline{x}(t, x_0))\, X(t), \\ X(0) &= \mathrm{I}_N, \end{aligned}$$

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

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:

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

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

1.5 Bemerkung. Es gilt

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

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

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

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

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 $t_0 \leq t \leq t_{end}$ und $f \in C^1([t_0, t_{end}] \times \mathbb{R}^N, \mathbb{R}^N)$. Wir nehmen an, dass eine Lösung $\overline{u}(t, t_0, \alpha)$ für $t \in [t_0, t_{end}]$ existiert.
Zur numerischen Berechnung von $\overline{u}(t, t_0, \alpha)$, $t_0 \leq t \leq t_{end}$ wählen wir eine Schrittweite $h > 0$ und ein äquidistantes Gitter

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

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

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

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

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

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

mit $V :]0, h_0] \times [t_0, t_{end}] \times \mathbb{R}^N \to \mathbb{R}^N$, $h_0 > 0$. 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:

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

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

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

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

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

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

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

und

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

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

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

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

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

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

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

den Defektvektor. Der Defektvektor $\mathrm{def}(\overline{u}_h) = T(\overline{u}_h)$ für $\overline{u}_h$ 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 $\overline{u} \in W$ von der AWA gilt

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

Ist sogar $\|T^h(\overline{u}_h)\|_\infty = O(h^p)$, 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 $\overline{u} \in W$ der AWA ein $h_0 > 0$ gibt, so dass $T^h(u) = 0$, $0 < h \le h_0$ eine Lösung $u^h \in (\mathbb{R}^N)^{\Omega_h}$ besitzt mit

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

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

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

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 $v = \overline{u}_h$ in die Stabilitätsungleichung ein und finde

$$\|u^h - \overline{u}_h\|_\infty \le C\|\underbrace{T^h(u^h)}_{=0} - T^h(\overline{u}_h)\|_\infty = C\|T^h(\overline{u}_h)\|_\infty \to 0 \quad (= O(h^p))$$

□

Slogan:
Konsistenz (der Ordnung p) + Stabilität ⇒ 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 $\overline{u}(t_j)$ für $\overline{u}(t_j, t_0, \alpha)$. Dann gilt

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

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

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

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

Äquivalente Formulierung liefert die Darstellung auf Basis der Steigungswerte:

$$\begin{aligned} u_{m+1} &= u_m + h\sum_{j=1}^{s} b_j K^j(h, t_m, u_m) \\ u_0 &= \alpha \end{aligned}$$

mit

$$\begin{aligned} K^1(h, t_m, u_m) &= f(t_m, u_m), \quad c_1 = 0 \\ K^i(h, t_m, u_m) &= f(t_m + c_i h, u_m + h\sum_{j=1}^{i-1} a_{ij} K^j(h, t_m, u_m)), \quad i = 2, 3, \dots, s. \end{aligned}$$

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

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

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

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

Hilfe von f aus:

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

Sei nun

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

dann gilt die Abschätzung

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

Wir finden also

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

Drücke nun V durch f aus:

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

d.h.

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

Wir erhalten dann

$$V(h,t,\overline{u}(t)) = (b_1+b_2)f(t,\overline{u}(t)) + hb_2\left(\frac{\partial f}{\partial t}c_2 + \frac{\partial f}{\partial u}a_{21}f\right)(t,\overline{u}(t)) + O(h^2).$$

Insgesamt folgt damit für den Konsistenzfehler

$$\begin{aligned}
h^{-1}(\overline{u}(t+h)-\overline{u}(t))-V(h,t,\overline{u}(t)) &= f(t,\overline{u}(t))+\frac{h}{2}\left(\frac{\partial f}{\partial t}+\frac{\partial f}{\partial u}f\right)(t,\overline{u}(t)) \\
&\quad -(b_1+b_2)f(t,\overline{u}(t)) \\
&\quad -hb_2\left(\frac{\partial f}{\partial t}c_2+\frac{\partial f}{\partial u}a_{21}f\right)(t,\overline{u}(t))+O(h^2) \\
&= (1-b_1-b_2)f(t,\overline{u}(t)) \\
&\quad +\frac{h}{2}\left((1-2b_2c_2)\frac{\partial f}{\partial t}+(1-2b_2a_{21})\frac{\partial f}{\partial u}f\right)(t,\overline{u}(t)) \\
&\quad +O(h^2).
\end{aligned}$$

Als hinreichende Bedingungen für Konsistenz der Ordnung 2 erhalten wir folglich

$$\begin{aligned}
1 &= b_1+b_2 \\
1 &= 2b_2c_2 \\
1 &= 2a_{21}b_2
\end{aligned}$$

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:

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

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

Iterationsvorschrift:

$$U^1(h, t_m, u_m) = u_m$$
$$U^2(h, t_m, u_m) = u_m + \frac{h}{2} f(t_m, u_m)$$
$$V(h, t_m, u_m) = f(t_m + \frac{h}{2}, u_m + \frac{h}{2} f(t_m, u_m))$$

Tableau:

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

Die beiden Verfahren besitzen die Konsistenzordnung 2, falls $\overline{u} \in C^3$. 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 $\overline{u} \in C^5$ 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 $\overline{u} \in C^6$ 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

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

mit

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

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

Die äquivalente Steigungsdarstellung lautet:

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

Die Tableau-Beschreibung:

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

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

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

und somit

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

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

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

Dies liefert

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

Andererseits gilt

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

Also erhalten wir für den Konsistenzfehler

$$\begin{aligned}
h^{-1}(\overline{u}(t+h)-\overline{u}(t)) - b_1 f(t+c_1 h, U^1(h,t,\overline{u}(t))) &= f(t,\overline{u}(t)) - b_1 f(t,\overline{u}(t)) + O(h^2)\\
&\quad + \frac{h}{2}\left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} f\right)(t,\overline{u}(t))\\
&\quad - h b_1 \left(c_1 \frac{\partial f}{\partial t} + a_{11}\frac{\partial f}{\partial u} f\right)(t,\overline{u}(t))\\
&= (1-b_1) f(t,\overline{u}(t))\\
&\quad + \frac{h}{2}(1-2b_1c_1)\frac{\partial f}{\partial t}(t,\overline{u}(t))\\
&\quad + \frac{h}{2}(1-2b_1a_{11})\left(\frac{\partial f}{\partial u} f\right)(t,\overline{u}(t))\\
&\quad + O(h^2).
\end{aligned}$$

Das Verfahren ist dann von Konsistenzordnung 2, falls gilt

$$\begin{cases} 1-b_1 = 0 \\ 1-2b_1c_1 = 0 \\ 1-2b_1a_{11} = 0 \\ \overline{u} \in C^3 \end{cases} \Rightarrow \begin{cases} b_1 = 1 \\ c_1 = 1/2 \\ a_{11} = 1/2 \\ \overline{u} \in C^3 \end{cases}$$

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

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

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

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

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

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

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

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

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

genau eine Lösung

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

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

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

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

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

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

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

□

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 $\overline{x}$. Bildet man ausgehend von $x_0 \in D$ die Folge $x_{n+1} = T(x_n)$, so gilt

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

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 $\overline{U}$ von $T(U) = U$, und es gilt*

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

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 $U^{(0)} = (u, u, \ldots, u)$, denn $U^{(0)}$ löst das System (2-3) für $h = 0$.

d) Systematische Bestimmung der Konsistenzordnung

Vorgelegt sei

$$u'(t) = f(t, u(t)),$$
$$u(t_0) = \alpha.$$

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

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

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:

$$B(p): \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \dots, p,$$

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

$$D(m): \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j (1 - c_j^k), \quad j = 1, \dots, s \quad k = 1, \dots, m.$$

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

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

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

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

Für $t = t_m + h$ bekommt man

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

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

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

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

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

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

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

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

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

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

$$p \leq q + m + 1,$$
$$p \leq 2q + 2$$

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: $c_{s+1-i} = 1 - c_i$, $i = 1, \ldots, s$. Die entsprechenden Methoden erfüllen $B(2s)$, $C(s)$, $D(s)$. Der Satz 2.11 impliziert

die Konsistenzordnung $p = 2s$.

$s = 1$, $p = 2$

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

$s = 2$, $p = 4$

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

b) **Radau-Verfahren**
Diese Verfahren beruhen auf der RADAU-Quadraturmethode. Es gilt $c_i \in [0, 1]$, $i = 1, \ldots, s$. 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:

$s = 1$, $p = 1$

0	1
	1

$s = 2$, $p = 3$

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

Radau IIA:

$s = 1$, $p = 1$ implizit Euler-Cauchy

1	1
	1

$s = 2$, $p = 3$

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

f) Das Newtonverfahren

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

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

falls diese durchführbar ist.

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

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

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 $U \subset R^N$ offen und sei $T : U \to \mathbb{R}^N$ zweimal stetig differenzierbar. Ferner existiere eine Lösung $\overline{u} \in U$ von $T(u) = 0$ und $T'(\overline{u})$ sei invertierbar.
Dann gibt es eine Kugel $K_\rho(\overline{u}) = \{u \in \mathbb{R}^N \mid \|u - \overline{u}\|_\infty \leq \rho\} \subset U$, $\rho > 0$, so dass $T(u) = 0$ keine weitere Lösung in $K_\rho(\overline{u})$ besitzt. Für jedes $u^0 \in K_\rho(\overline{u})$ existiert die Newtonfolge

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

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

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

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

Bemerkung 7. *Lösungen $\overline{u}$ von $T(u) = 0$ mit $T'(\overline{u})$ 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

$$u'(t) = f(t, u(t))$$
$$u(t_0) = \alpha$$

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

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

Es gilt der folgende Stabilitätssatz:

3.1 Satz. ***(Stabilitätssatz)***
Es existiere ein $\overline{h} > 0$ und ein $L > 0$, so dass die Verfahrensfunktion V einer Lipschitz*bedingung*

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

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

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

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

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

und schließlich

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

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

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

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 ($j \to j+1$):

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

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:

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

Wegen der Monotonie der Abschätzung gilt auch

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

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

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

□

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

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

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

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

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 $u'(t) = f(t, u(t))$, $t \in [t_0, t_{end}]$, $u(t_0) = \alpha$ mit $f : [t_0, t_{end}] \times \mathbb{R}^N \to \mathbb{R}^N$ 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*

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

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

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

4. Asymptotische Entwicklung und Schrittweitensteuerung

a) Verschärfte Konvergenzaussagen

Wir betrachten erneut ein Einschrittverfahren der Form mit

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

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 $u^h = \overline{u}_h + O(h^p)$, 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 $u^h - \overline{u}_h$ in der Form

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

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

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

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

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

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.

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

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

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

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

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

Wir erhalten also

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

falls e_0 durch die lineare Anfangswertaufgabe

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

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

Eine Verallgemeinerung dieser Untersuchungen ist der folgende

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

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

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

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

b) Schrittweitensteuerung und Fehlerschätzungen

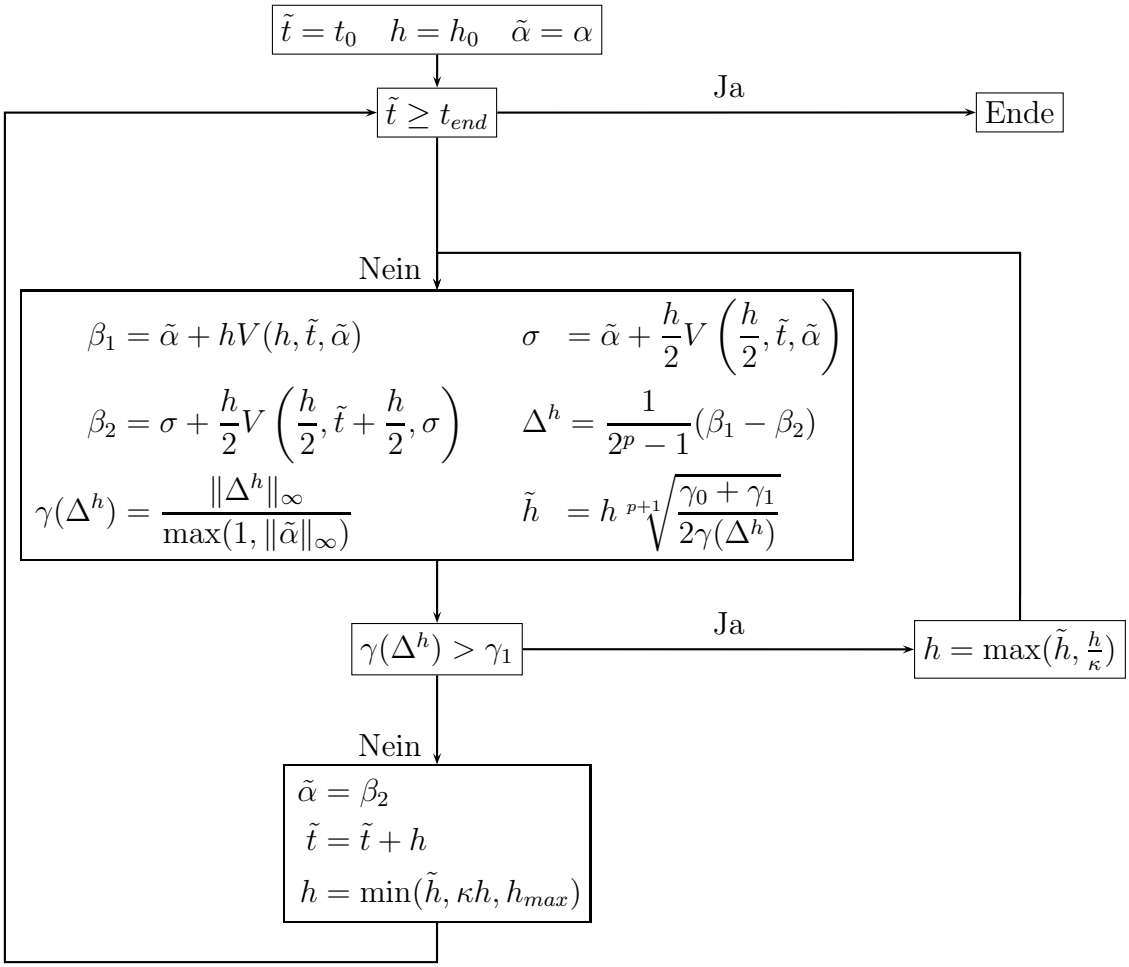
Vorgelegt sei

$$u'(t) = f(t, u(t)), \quad t_0 \le t \le t_{end},$$
$$u(t_0) = \alpha.$$

Wir führen folgende Parameter ein:

$h_0 > 0$	Startschrittweite
$h_{max} \ge h_0$	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 $\widetilde{\alpha}$ Näherungswert für $\overline{u}(\widetilde{t})$, und sei $\widetilde{u}(t)$ die Lösung von

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

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

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

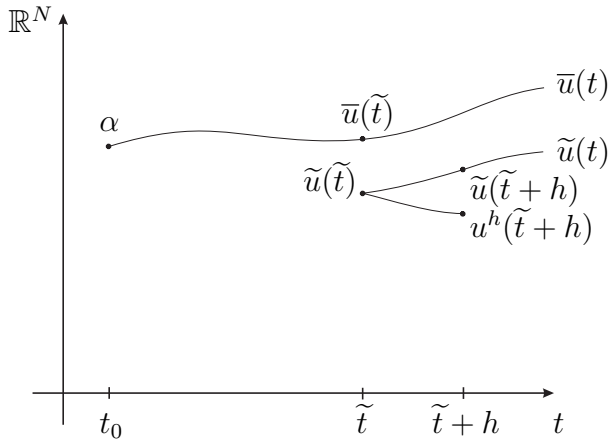


Abbildung 7: Erklärung der Schrittweitensteuerung

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

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

Analog bekommt man

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

Zur Schätzung des Hauptfehlerterms berechnen wir

$$\begin{aligned} u^h(\widetilde{t}+h) - u^{h/2}(\widetilde{t}+h) &= -\left(\frac{h}{2}\right)^p h e_0'(\widetilde{t}) + h^{p+1} e_0'(\widetilde{t}) + O(h^{p+2}) \\ &= \left(\frac{h}{2}\right)^p h e_0'(\widetilde{t})(2^p - 1) + O(h^{p+2}) \\ \Rightarrow \left(\frac{h}{2}\right)^p h e_0'(\widetilde{t}) &= \frac{u^h(\widetilde{t}+h) - u^{h/2}(\widetilde{t}+h)}{2^p - 1} + O(h^{p+2}). \end{aligned}$$

Mit den eingeführten Bezeichnungen bekommt man

$$\left(\frac{h}{2}\right)^p he_0'(\widetilde{t}) \;=\; \frac{\beta_1-\beta_2}{2^p-1}+O(h^{p+2}) \;=\; \Delta^h+O(h^{p+2}). \tag{4-2}$$

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

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

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

$$\begin{aligned}
\Delta^{\widetilde{h}} &= \left(\frac{\widetilde{h}}{2}\right)^p \widetilde{h}e_0'(\widetilde{t})+O(\widetilde{h}^{p+2})\\
&= \left(\frac{\widetilde{h}}{h}\right)^{p+1} \underbrace{\left(\frac{h}{2}\right)^p he_0'(\widetilde{t})}_{=\Delta^h+O(h^{p+2})}+O(\widetilde{h}^{p+2})\\
&\Rightarrow \Delta^{\widetilde{h}} = \left(\frac{\widetilde{h}}{h}\right)^{p+1}\Delta^h+O(h^{p+2}+\widetilde{h}^{p+2}).
\end{aligned}$$

Näherungsweise verlangen wir also

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

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