Numerik gewöhnlicher Differentialgleichungen

Mitschrift[1] zur Vorlesung von Prof. März
Humboldt Universität zu Berlin

Sommersemester 2004

[1]Gefundene Fehler bitte an Stefan Vigerske (vigerske@mathematik.hu-berlin.de).

Inhaltsverzeichnis

Kapitel 1

Qualitatives Verhalten von Differentialgleichungslösungen und dessen Wiederspiegelung durch Approximation

1.1 Kontraktive und dissipative Systeme

Wir betrachten zunächst autonome Differentialgleichungen (DGL) der Form

$$\boxed{x'(t) = f(x(t)), \quad f \in C^1(\mathbb{R}^m).} \tag{1.1}$$

Die Funktion

$$x(\cdot, x_0) \tag{1.2}$$

sei Lösung der Anfangswertaufgabe (AWA)

$$\boxed{x'(t) = f(x(t)), \qquad x(0) = x_0.}$$

Definition 1.1 (Gleichgewichtslage, stationäre Lösung). Ein Punkt $c \in \mathbb{R}^m$ mit $f(c) = 0$ heißt *Gleichgewichtslage*. Die Funktion $x(\cdot, c) \equiv c$ heißt dann *stationäre Lösung*.

Definition 1.2 (dynamisches System). Sei $E \subseteq \mathbb{R}^m$ eine Menge. Die Differentialgleichung (1.1) definiert ein *dynamisches System* auf E, falls, für jedes $x_0 \in E$, die Anfangswertaufgabe

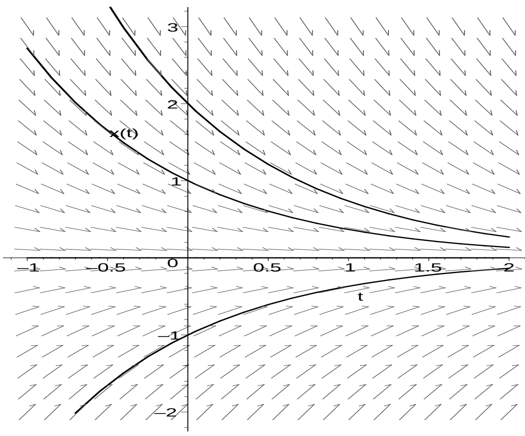
$$x'(t) = f(x(t)), \qquad x(0) = x_0,$$

eine eindeutige Lösung besitzt, welche für alle $t \in [0, \infty)$ definiert ist, und die für alle $t \in [0, \infty)$ in E verbleibt ($x(t) \in E$, $t \geq 0$).

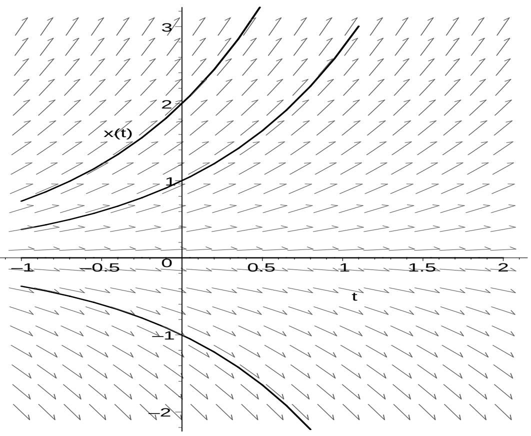
Beispiel. $m = 1$

1. Für die Differentialgleichung $x'(t) = a \cdot x(t)$ ist die Lösung gegeben durch

$$x(t) = e^{at} \cdot x_0, \qquad x_0 \in \mathbb{R}.$$



$x'(t) = -x(t)$ mit $x_0 \in \{-1, 0, 1, 2\}$.



$x'(t) = x(t)$ mit $x_0 \in \{-1, 0, 1, 2\}$.

2. Für die Differentialgleichung $x'(t) = x^2(t)$ ist die Lösung gegeben durch

$$x(t) = \frac{x_0}{1 - t \cdot x_0}, \qquad t \in \begin{cases} \left(\frac{1}{x_0}, \infty\right) & x_0 < 0, \\ (-\infty, \infty) & x_0 = 0, \\ \left(-\infty, \frac{1}{x_0}\right) & x_0 > 0. \end{cases}$$

   Die DGL definiert damit ein dynamisches System auf $E = \{x_0 \in \mathbb{R} | \, x_0 \leq 0\}$.

3. Für die Differentialgleichung $x'(t) = -x^3(t)$ ist die Lösung gegeben durch

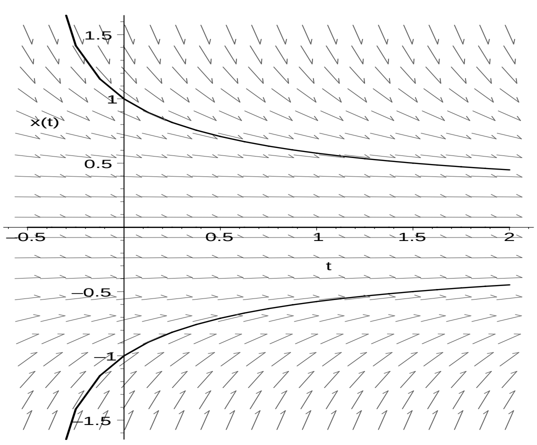
$$x(t) = \frac{x_0}{\sqrt{1 + 2tx_0^2}}, \qquad t \in \begin{cases} (-\infty, \infty) & x_0 = 0, \\ \left(-\frac{1}{2x_0^2}, \infty\right) & x_0 \neq 0. \end{cases}$$

   Die DGL definiert ein dynamisches System für $E = \mathbb{R}$, $E = \mathbb{R}^+$ oder $E = \mathbb{R}^-$.

   Betrachte weiterhin die Menge $B = (-a, a)$, $a > 0$. Es ist

$$\frac{d}{dt} |x(t)|^2 = 2 \cdot x(t) \cdot x'(t) = -2 \cdot |x(t)|^4 \leq 0,$$

   d.h. $|x(t)|$ ist nicht wachsend. Für $x_0 \in B$ ist also auch $x(t) \in B$ für alle $t \geq 0$, also definiert $E = B$ ein dynamisches System.



Lösungen der DGL $x'(t) = -x^3(t)$ für $x_0 \in \{-1, 0, 1\}$

Definition 1.3 (Evolutionshalbgruppe, Lösungsoperator). Sei durch (1.1) ein dynamisches System auf $E \subseteq \mathbb{R}^m$ gegeben. Die Familie von Operatoren

$$\boxed{S(t) : E \to E, \qquad S(t)\, x_0 := x(t, x_0), \quad x_0 \in E, t \geq 0,}$$

heißt *Evolutionshalbgruppe* der DGL. $S(t)$ heißt *Lösungsoperator* der DGL.

Bemerkung. Es gilt $S(t+\tau) = S(t)\, S(\tau) = S(\tau)\, S(t)$ und $S(0) = I$.

Definition 1.4 (einseitige Lipschitz-Bedingung). Sei $\langle \cdot, \cdot \rangle$ ein inneres Produkt auf $\mathbb{R}^m$ und $|\cdot|$ die zugehörige Norm. Die Ungleichung

$$\boxed{\langle f(x) - f(\bar{x}), x - \bar{x} \rangle \leq \gamma \cdot |x - \bar{x}|^2, \quad x, \bar{x} \in \mathbb{R}^m,} \tag{1.3}$$

mit einer Konstanten $\gamma \in \mathbb{R}$ heißt *einseitige Lipschitz-Bedingung*.

Beispiel. $m = 1$

1. $x'(t) = a \cdot x(t)$, d.h. $f(x) = a \cdot x$. Dann ist

$$\langle f(x) - f(\bar{x}), x - \bar{x} \rangle = a \cdot (x - \bar{x})^2,$$

d.h.. die einseitige Lipschitz-Bedingung (1.3) ist für $\gamma = a$ erfüllt.

2. Sei $f(x) = x^2$. Es ist

$$\langle f(x) - f(\bar{x}), x - \bar{x} \rangle = \left(x^2 - \bar{x}^2\right) \cdot (x - \bar{x}) = (x + \bar{x}) \cdot (x - \bar{x})^2.$$

Dann ist keine Bedingung (1.3) erreichbar.

3. Sei $f(x) = -x^3$. Es ist

$$\begin{aligned} \langle f(x) - f(\bar{x}), x - \bar{x} \rangle &= -\left(x^3 - \bar{x}^3\right) \cdot (x - \bar{x}) \\ &= -\left(x^2 + x\bar{x} + \bar{x}^2\right) \cdot (x - \bar{x})^2 \\ &= -\left(x^2 + \bar{x}^2\right)(x - \bar{x})^2 - x\bar{x}\,(x - \bar{x})^2 \end{aligned}$$

und weil $x^2 + 2x\bar{x} + \bar{x}^2 = (x + \bar{x})^2 \geq 0$, d.h. $\frac{1}{2}\left(x^2 + \bar{x}^2\right) \geq -x\bar{x}$ folgt weiter

$$\begin{aligned} &\leq -\left(x^2 + \bar{x}^2\right)(x - \bar{x})^2 + \frac{1}{2}\left(x^2 + \bar{x}^2\right)(x - \bar{x})^2 \\ &= -\frac{1}{2}\left(x^2 + \bar{x}^2\right)(x - \bar{x})^2 \\ &\leq 0, \end{aligned}$$

also ist die Bedingung (1.3) mit $\gamma = 0$ erfüllt. Die Funktion ist aber nicht Lipschitz-stetig!

Bemerkung. Ist f Lipschitz-stetig mit Lipschitz-Konstante L, so folgt mittels Cauchy-Schwarzscher-Ungleichung

$$\langle f(x) - f(\bar{x}), x - \bar{x} \rangle \leq |f(x) - f(\bar{x})| \cdot |x - \bar{x}| \leq L \cdot |x - \bar{x}|^2,$$

d.h. die einseitige Lipschitz-Bedingung (1.3) ist mit $\gamma = L$ erfüllt.

Der folgende Satz zeigt, wie die einseitige Lipschitz-Bedingung dem Grad, mit dem zwei Trajektorien des dynamischen Systems möglicherweise divergieren, eine Beschränkung auferlegt. Zuvor benötigen wir aber das Lemma von Gronwall.

Lemma 1.5 (Gronwall [SH96, S. 108]). *Für eine Funktion $z \in C^1([0,\infty))$ gelte*

$$z'(t) \leq a\, z(t) + b$$

mit Konstanten $a \neq 0$ und b. Dann ist

$$z(t) \leq e^{at}\, z(0) + \frac{b}{a}\left(e^{at} - 1\right), \qquad t \geq 0.$$

Satz 1.6. *Sei für die DGL (1.1) die einseitige Lipschitz-Bedingung (1.3) gegeben. Dann gilt*

1. *Alle Anfangswertaufgaben (1.2) sind eindeutig lösbar auf $[0,\infty)$, d.h. sie generieren ein dynamisches System auf $\mathbb{R}^m$.*
2. *Für je zwei Lösungen $x(\cdot)$, $\bar{x}(\cdot)$ gilt:*

$$\boxed{|x(t) - \bar{x}(t)| \leq e^{\gamma t}\,|x(0) - \bar{x}(0)|, \quad t \geq 0.}$$

3. *Ist $\gamma < 0$, so besitzt die Differentialgleichung genau eine Gleichgewichtslage $c \in \mathbb{R}^m$. Diese ist exponentiell attraktiv, d.h. es gilt*

$$|x(t) - c| \leq e^{\gamma t} \cdot |x(0) - c|, \quad t \geq 0,$$

für jede beliebige Lösung $x(\cdot)$.

Beweis. 1.,2.: Die Funktionen $x(\cdot)$ und $\bar{x}(\cdot)$ seien beliebige Lösungen der DGL (1.1). Sei x auf dem Intervall $[0,\tau)$ und $\bar{x}$ auf dem Intervall $[0,\bar{\tau})$ definiert. Für 1. genügt es zu zeigen, daß $\tau = \bar{\tau} = \infty$ gilt. O.B.d.A. sei $\tau \leq \bar{\tau}$.

Dann gilt auf dem Intervall $[0,\tau)$

$$\begin{aligned} \frac{d}{dt}|x(t) - \bar{x}(t)|^2 &= 2\langle x'(t) - \bar{x}'(t), x(t) - \bar{x}(t)\rangle \\ &= 2\langle f(x(t)) - f(\bar{x}(t)), x(t) - \bar{x}(t)\rangle \\ &\leq 2\gamma \cdot |x(t) - \bar{x}(t)|^2 \end{aligned}$$

Für die Funktion $z(t) := |x(t) - \bar{x}(t)|^2$, $t \in [0,\tau)$, gilt dann

$$z'(t) \leq 2\gamma\, z(t).$$

Die Anwendung des Gronwall-Lemmas 1.5 liefert dann

$$z(t) \leq e^{2\gamma t}\, z(0), \qquad t \in [0,\tau),$$

also

$$|x(t) - \bar{x}(t)| \leq e^{\gamma t}\,|x(0) - \bar{x}(0)|, \qquad t \in [0,\tau).$$

Dies zeigt 2. Weiterhin folgt

$$\begin{aligned} |\bar{x}(t)| &\leq |x(t)| + |x(t) - \bar{x}(t)| \leq |x(t)| + e^{\gamma t}\,|x(0) - \bar{x}(0)|, \qquad t \in [0,\tau), \\ |\bar{x}(t)| &\geq |x(t)| - |x(t) - \bar{x}(t)| \geq |x(t)| - e^{\gamma t}\,|x(0) - \bar{x}(0)|, \qquad t \in [0,\tau). \end{aligned}$$

Dies zeigt: Wenn x unbeschränkt wird zur endlichen Zeit $t = \tau$, so ist $\bar{x}$ beschränkt auf $[0,\tau)$ und wird unbeschränkt genau zur Zeit $t = \tau$.

Nun waren x und $\bar{x}$ beliebig gewählt. D.h., falls die DGL kein dynamisches System generiert, so müssen alle Lösungen zur gleichen Zeit unbeschränkt werden. Dies führt zu einem Widerspruch:

Angenommen, es gibt eine Zeit T_*, zu der keine Lösung fortsetzbar ist.

Zu einem fixierten $x_* \in \mathbb{R}^m$ löst $x(\cdot, x_*)$ die Anfangswertaufgabe mit $x(0) = x_*$ auf dem Intervall $[0, T_*)$. Wir wählen nun $x_{**} := x\left(\frac{T_*}{2}, x_*\right) = S\left(\frac{T_*}{2}\right) x_*$. Dann gilt

$$x(t, x_{**}) = S(t)\, x_{**} = S(t)\, S\left(\frac{T_*}{2}\right) x_* = S\left(t + \frac{T_*}{2}\right) x_* \xrightarrow{t \to \frac{T_*}{2}} \infty,$$

d.h. $x(\cdot, x_{**})$ wird schon zum Zeitpunkt $\frac{T_*}{2}$ unbeschränkt. Dies ist ein Widerspruch zur Beobachtung, daß alle Lösungen erst zur Zeit T_* unbeschränkt werden können (da $x(\cdot, x_{**})$ durch $x(\cdot, x_*)$ beschränkt wird).

3.: Sei $\gamma < 0$.

Eindeutigkeit: Seien $c, \bar{c} \in \mathbb{R}^m$ Gleichgewichtslagen, d.h. $f(c) = 0$, $f(\bar{c}) = 0$. Anwendung von 2. auf $x(t) \equiv c$, $\bar{x}(t) \equiv \bar{c}$ ergibt

$$|c - \bar{c}| = |x(t) - \bar{x}(t)| \le e^{\gamma t} |x(0) - \bar{x}(0)| = e^{\gamma t} |c - \bar{c}| \xrightarrow{t \to \infty} 0,$$

da $\gamma < 0$. Also ist $c = \bar{c}$.

Existenz: Sei $T > 0$ fest. Wegen $\gamma < 0$ ist $S(T)$ eine Kontraktion auf $\mathbb{R}^m$. Dann existiert ein eindeutiger Fixpunkt $x_T \in \mathbb{R}^m$ mit $S(T)\, x_T = x_T$. Siehe weiter in [SH96, S. 174]. □

Bemerkung. Ist die einseitige Lipschitz-Bedingung $\gamma < 0$ erfüllt, und ist $c \in \mathbb{R}^m$ Gleichgewichtslage der DGL, so folgt für eine beliebige Lösung $x(\cdot)$

nicht in der Vorlesung

$$\langle f(x(t)) - f(c), x(t) - c \rangle \le 0, \qquad \text{d.h. } \langle f(x(t)), c - x(t) \rangle \ge 0,$$

der Vektor $f(x(t))$ zeigt also in die Richtung der Gleichgewichtslage („Die Lösung $x(t)$ bewegt sich in Richtung c.") .

Beispiel. Die Voraussetzung $\gamma < 0$ in Teil 3 von Satz 1.3 ist notwendig, wie dieses Beispiel zeigt: Für $m = 3$ betrachte die DGL

$$\begin{aligned} x_1'(t) &= -x_2(t) \\ x_2'(t) &= x_1(t) \\ x_3'(t) &= 0. \end{aligned}$$

Dann ist

$$\begin{aligned} \langle f(x) - f(\bar{x}), x - \bar{x} \rangle &= \left\langle \begin{pmatrix} -x_2 + \bar{x}_2 \\ x_1 - \bar{x}_1 \\ 0 \end{pmatrix}, \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ x_3 - \bar{x}_3 \end{pmatrix} \right\rangle \\ &= (-x_2 + \bar{x}_2)(x_1 - \bar{x}_1) + (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \\ &= 0, \end{aligned}$$

d.h. die einseitige Lipschitz-Bedingung ist nur für $\gamma = 0$ erfüllt.

Gleichgewichtslagen sind alle Punkte $c = (0, 0, c_3)$, $c_3 \in \mathbb{R}$.

Satz 1.7. *Sei für die DGL* (1.1) *die Bedingung* (1.3) *mit* $\gamma = 0$ *erfüllt. Dann ist die Menge aller Gleichgewichtslagen*

$$\boxed{\mathcal{E} := \{c \in \mathbb{R}^m \,|\, f(c) = 0\}}$$

abgeschlossen und konvex.

Beweis. Für die Konvexität siehe [SH96, S. 176].

Abgeschlossenheit: Sei $c_n \in \mathcal{E}$ eine Folge von Gleichgewichtslagen, d.h. $f(c_n) = 0$ mit $c_n \xrightarrow{n \to \infty} c_*$. Aufgrund der Stetigkeit von f ist dann $f(c_*) = \lim_{n \to \infty} f(c_n) = 0$, also $c_* \in \mathcal{E}$. □

Definition 1.8 ((schwach) kontraktiv, Kontraktivitätsungleichung). Falls für die DGL (1.1) die Bedingung (1.3) mit $\gamma < 0$ ($\gamma = 0$) erfüllt ist, heißt die DGL *kontraktiv* (*schwach kontraktiv*). (1.3) heißt dann *Kontraktivitätsungleichung*.

Bisher haben wir nur nichtlineare autonome Differentialgleichungen betrachtet, für welche die Trajektorien asymptotisch gegen eine eindeutige Gleichgewichtslage konvergieren. Eine natürliche Verallgemeinerung ist es nun, Systeme zu betrachten, bei denen sich das asymptotische Verhalten auf ein beschränkte Menge beschränkt, wobei keine Einschränkungen bezüglich der Dynamik innerhalb dieser Menge gemacht werden. Dies motiviert die folgenden Definitionen:

Definition 1.9 ((positiv, negativ) invariant). Eine Menge $\mathcal{A} \subseteq \mathbb{R}^m$ heißt *positiv invariant* in Bezug auf (1.1), falls für jede beliebige Lösung $x(\cdot)$ aus $x(t_*) \in \mathcal{A}$ für ein $t_* \in I_*$, folgt, daß $x(t) \in \mathcal{A}$ für alle $t > t_*, t \in I_*$.

Analog: *Negativ invariant*.
$\mathcal{A}$ heißt *invariant* in Bezug auf (1.1), wenn es positiv und negativ invariant ist.

Definition 1.10 (dissipativ, absorbierende Menge). Ein dynamisches System auf $\mathbb{R}^m$ heißt *dissipativ*, falls es eine beschränkte positiv invariante Menge $\mathcal{A} \subset \mathbb{R}^m$ gibt und für jede beschränkte Menge $\mathcal{B} \subset \mathbb{R}^m$ ein $t_*(\mathcal{A},\mathcal{B})$ existiert mit der Eigenschaft, daß

$$\boxed{S(t)\mathcal{B} \subseteq \mathcal{A} \qquad \forall t > t_*(\mathcal{A},\mathcal{B})}$$

$\mathcal{A}$ heißt dann *absorbierende Menge*.

Beispiel. $m = 1$, $x'(t) = -x(t)$. Dann ist

$$S(t)\,x_0 = e^{-t}x_0, \qquad x_0 \in \mathbb{R}$$

Die Menge $\mathcal{A} = [-a,a]$ mit $a > 0$ ist beschränkt und positiv invariant:

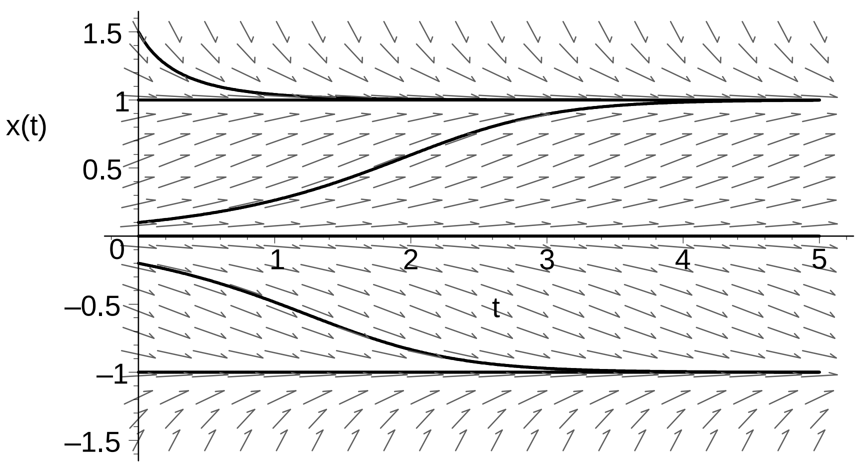
Sei $\mathcal{B} \subset \mathbb{R}$ beschränkt, dann existiert ein $r > 0$, so daß $\mathcal{B} \subset B(0,r)$. Wählen wir $t_*(\mathcal{A},\mathcal{B}) := t_*$ so, daß $e^{-t_*}r = a$ ist, d.h. $t_* = \ln\left(\frac{r}{a}\right)$. Dann ist $e^{-t}r \leq a$ für alle $t \geq t_*$, also $x(t) \in \mathcal{A}$ für alle $t > t_*(\mathcal{A},\mathcal{B})$. Auch die Menge $\mathcal{A} = (-a,a)$ ist positiv invariant.

nicht in der Vorlesung

Beispiel ([SH96, S. 181]). Für $m = 1$ betrachte die Differentialgleichung

$$x'(t) = x(t) - x^3(t).$$

Jede Menge $\mathcal{A} = [-a,a]$ mit $a > 1$ ist absorbierend.



Lösungen der Differentialgleichung für $x(0) \in \left\{-1, -\frac{1}{5}, 0, \frac{1}{10}, 1, \frac{3}{2}\right\}$.

Dissipativität entsteht häufig als direkte Konsequenz aus einer Strukturbedingung, die von f erfüllt wird. Existiert z.B. ein $\beta > 0$, so daß f die Bedingung

$$\langle f(x), x\rangle \leq -\beta |x|^2, \qquad x \in \mathbb{R}^n, \tag{1.4}$$

erfüllt, so folgt aus dem Gronwall-Lemma

$$\|x(t)\| \leq e^{-\beta t}\|x(0)\|$$

für jede Lösung $x(t)$ von (1.1), d.h. alle Trajektorien des Systems konvergieren gegen den Ursprung. Diese Systeme sind also dissipativ, wobei jede Kugel um 0 eine absorbierende Menge darstellt. Wir wollen die Ungleichung (1.4) nun verallgemeinern:

Definition 1.11 (Dissipativitätsungleichung). Seien $\alpha \geq 0$, $\beta > 0$ Konstanten. Die *Dissipativitätsungleichung* lautet:

$$\boxed{\langle f(x), x\rangle \leq \alpha - \beta \cdot |x|^2 \qquad \forall x \in \mathbb{R}^m.} \tag{1.5}$$

Satz 1.12. *Sei für die DGL (1.1) die Ungleichung (1.5) gegeben.*

Dann ist durch (1.1) ein dynamisches System auf $\mathbb{R}^m$ definiert, welches dissipativ ist. Alle Kugeln

$$\boxed{B\left(0, \sqrt{\frac{\alpha}{\beta}+\varepsilon}\right), \quad \varepsilon > 0,}$$

sind absorbierende Mengen.

Beweis. Sei $x \in C^1([0,\tau))$ Lösung. Es ist

$$\begin{aligned} \frac{d}{dt}|x(t)|^2 &= \frac{d}{dt}\langle x(t), x(t)\rangle \\ &= 2\langle f(x(t)), x(t)\rangle \\ &\leq 2\alpha - 2\beta|x(t)|^2, \quad t \in [0,\tau) \end{aligned}$$

Anwendung des Gronwall-Lemmas 1.5 auf $z(t) := |x(t)|^2$, $z'(t) \leq 2\alpha - 2\beta z(t)$, liefert

$$\begin{aligned} |x(t)|^2 &\leq e^{-2\beta t}|x(0)|^2 + \frac{2\alpha}{-2\beta}\left(e^{-2\beta t}-1\right) \\ &= e^{-2\beta t}|x(0)|^2 + \frac{\alpha}{\beta}\left(1-e^{-2\beta t}\right) \\ &= e^{-2\beta t}\left(|x(0)|^2 - \frac{\alpha}{\beta}\right) + \frac{\alpha}{\beta} \end{aligned}$$

Falls $|x(0)|^2 \leq \frac{\alpha}{\beta}$, so ist $|x(t)|^2 \leq \frac{\alpha}{\beta}$.

Falls $|x(0)|^2 > \frac{\alpha}{\beta}$, so ist $|x(t)|^2 \leq e^{-2\beta t}|x(0)|^2 + |x(0)|^2\left(1-e^{-2\beta t}\right) = |x(0)|^2$.
Allgemein gilt also

$$|x(t)|^2 \leq \max\left\{|x(0)|^2, \frac{\alpha}{\beta}\right\}, \qquad t \in [0,\tau).$$

Dann ist die Lösung auf $[0,\infty)$ fortsetzbar und die DGL generiert ein dynamisches System auf $\mathbb{R}^m$.

Positive Invarianz von $B\left(0, \sqrt{\frac{\alpha}{\beta}+\varepsilon}\right)$: Sei $x_0 \in B\left(0, \sqrt{\frac{\alpha}{\beta}+\varepsilon}\right)$. Dann ist

$$|S(t)x_0| = |x(t,x_0)| \leq \max\left\{|x_0|, \sqrt{\frac{\alpha}{\beta}}\right\} \leq \sqrt{\frac{\alpha}{\beta}+\varepsilon}, \quad t \geq 0.$$

Dissipativität: Sei $\mathcal{B} \subset \mathbb{R}^m$ beschränkt, also $\mathcal{B} \subseteq B(0,r)$ für ein $r > 0$, und sei $x_0 \in \mathcal{B}$. Dann ist

$$|x(t,x_0)|^2 \leq e^{-2\beta t}\left(r^2 - \frac{\alpha}{\beta}\right) + \frac{\alpha}{\beta}.$$

Wir wählen nun t_* so, daß $e^{-2\beta t_*}\left(r^2 - \frac{\alpha}{\beta}\right) = \varepsilon$ gilt. Dann ist für $t > t_*$

$$|x(t,x_0)|^2 \leq e^{2\beta t_*}\left(r^2 - \frac{\alpha}{\beta}\right) + \frac{\alpha}{\beta} \leq \varepsilon + \frac{\alpha}{\beta}. \qquad \square$$

1.1.1 Nichtautonome Differentialgleichungen

Sei $f : \mathbb{R}^m \times I_f \to \mathbb{R}^m$ stetig, $I_f \subset \mathbb{R}$, $f_x : \mathbb{R}^m \times I_f \to L(\mathbb{R}^m)$ stetig. Für $t_0 \in I_f$, $x_0 \in \mathbb{R}^m$ mit $I_f \supseteq [t_0, \infty)$ betrachte die (nichtautonome) Differentialgleichung

$$x'(t) = f(x(t), t) \tag{1.6}$$

und die Anfangswertaufgabe

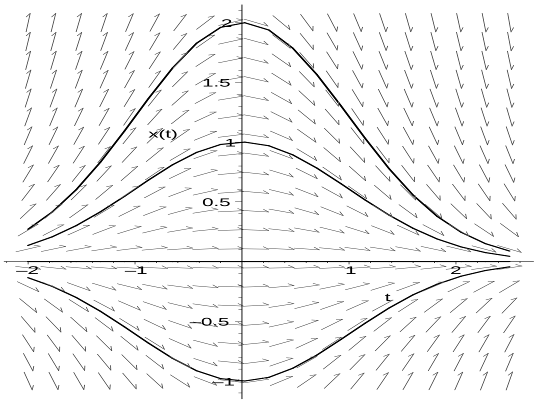
$$x'(t) = f(x(t), t), \qquad x(t_0) = x_0. \tag{1.7}$$

Definition 1.13 (Gleichgewichtslage, stationäre Lösung). Ein $c \in \mathbb{R}^m$ heißt *Gleichgewichtslage* von (1.6), falls $f(c,t) = 0$ für alle $t \in I_f$ gilt.

$x(t) \equiv c$, $t \in I_f$, heißt *stationäre Lösung*.

Beispiel. $m = 1$, $x'(t) = -t\,x(t)$, d.h. $f(x,t) = -tx$. Dann ist $c = 0$ Gleichgewichtslage. Die Lösung der DGL ist

$$x(t) = e^{-\frac{t^2}{2}} \cdot x_0.$$



Lösungen der Differentialgleichung $x'(t) = -t\,x(t)$ für $x_0 \in \{-1, 1, 2\}$.

Definition 1.14 (einseitige Lipschitz-Bedingung, Kontraktivitätsbedingung). Die Ungleichung

$$\boxed{\langle f(x,t) - f(\bar{x},t), x - \bar{x}\rangle \leq \gamma\,|x - \bar{x}|^2 \quad \text{für } x, \bar{x} \in \mathbb{R}^m, t \in I_f,} \tag{1.8}$$

mit Konstante $\gamma \in \mathbb{R}$ heißt *einseitige Lipschitzbedingung* für (1.6). Ist $\gamma \leq 0$, so heißt (1.8) *Kontraktivitätsbedingung.*

Definition 1.15 (Dissipativitätsungleichung). Die Ungleichung

$$\boxed{\langle f(x,t), x\rangle \leq \alpha - \beta\,|x|^2, \qquad x \in \mathbb{R}^m, t \in I_f,} \tag{1.9}$$

mit Konstanten $\alpha \geq 0$, $\beta > 0$, heißt *Dissipativitätsbedingung* für (1.6).

Die Aussagen 1 und 2 von Satz 1.6 und die Aussage von Satz 1.12 gelten auch für (1.6).

1.2 B-stabile Runge-Kutta Verfahren

Wir betrachten die Anfangswertaufgabe (1.7) und eine Diskretisierung von $[t_0, \infty)$:

$$t_0 < t_1 < \ldots < t_{n-1} < t_n < \ldots$$

mit Intervallängen

$$h_j = t_j - t_{j-1}, \qquad j = 1, 2, \ldots$$

Wir bezeichnen mit x_n die durch ein numerisches Verfahren bestimmte Approximation zu $x(t_n)$.

Wir suchen im folgenden Bedingungen für die numerischen Verfahren, so daß die approximierte Lösung das Langzeitverhalten der exakten Lösung von (1.7) unter geringen oder keinen Einschränkungen an die Schrittweite h_j wiedergibt. Speziell werden wir uns mit kontraktiven und dissipativen Systemen beschäftigen.

Falls die Differentialgleichung schwach kontraktiv ist, so gilt für je zwei Lösungen $x(\cdot)$, $\bar{x}(\cdot)$ zur AWA (1.7) mit Anfangswerten x_0, $\bar{x}_0$:

$$|x(t_n) - \bar{x}(t_n)| \leq |x(t_{n-1}) - \bar{x}(t_{n-1})|, \quad n \geq 1,$$

unabhängig von Schrittweiten. Wir möchten, daß sich diese Eigenschaft auch auf die durch ein numerisches Verfahren approximierte Lösung überträgt:

Definition 1.16 (B-stabil). Ein numerisches Integrationsverfahren heißt *B-stabil*, falls es bei Anwendung auf eine schwach kontraktive DGL (1.6) zu Anfangswerten x_0, $\bar{x}_0$ und beliebigen Zerlegungen $t_0 < t_1 < \ldots$ Sequenzen $\{x_n\}_{n\geq 0}$, $\{\bar{x}_n\}_{n\geq 0}$ erzeugt, für die die Ungleichung

$$\boxed{|x_n - \bar{x}_n| \leq |x_{n-1} - \bar{x}_{n-1}|, \quad n \geq 1,} \tag{1.10}$$

gilt.

Beispiel. Sei die DGL

$$x'(t) = f(x(t), t) := -10^5 x(t)$$

gegeben. Der Startwert $\bar{x}_0 = 0$ liefert die stationäre Lösung $\bar{x}_n = 0 \ \forall n$. Wegen $-10^5 < 0$ ist die DGL nach Lemma 3.5 kontraktiv. Wir betrachten nun einige numerische Integrationsverfahren:

1. Explizites Euler-Verfahren: Die Iterationsvorschrift lautet

$$x_n = x_{n-1} + h_n f(x_{n-1}, t_{n-1}).$$

Für dieses Beispiel ist dann

$$x_n = x_{n-1} + h_n\left(-10^5 x_{n-1}\right) = \left(1 - 10^5 h_n\right) x_{n-1}.$$

Die Ungleichung (1.10) lautet dann

$$|x_n - \bar{x}_n| = |x_n| = \left|1 - 10^5 h_n\right| |x_{n-1}| = \underbrace{\left|1 - 10^5 h_n\right|}_{\leq 1?} |x_{n-1} - \bar{x}_{n-1}| \overset{?}{\leq} |x_{n-1} - \bar{x}_{n-1}|.$$

Da dies aber eine Schrittweitenbeschränkung für h_n darstellt, ist dieses Verfahren nicht B-stabil.

2. Implizites Euler-Verfahren: Die Iterationsvorschrift lautet hier

$$x_n = x_{n-1} + h_n f(x_n, t_n).$$

Für dieses Beispiel also

$$\begin{aligned} x_n &= x_{n-1} + h_n \left(-10^5 x_n\right), \\ \text{d.h. } x_n &= \frac{1}{1+10^5 h_n} x_{n-1}. \end{aligned}$$

Wegen $\frac{1}{1+10^5 h_n} \leq 1$ ist die Ungleichung (1.10) für jede Wahl von h_n erfüllt:

$$|x_n - \bar{x}_n| = |x_n| \leq |x_{n-1}| = |x_{n-1} - \bar{x}_{n-1}|.$$

Dieses Verfahren ist B-stabil (zum Beweis siehe Satz 1.18 auf Seite 16).

3. Verbessertes Euler-Verfahren: Die Iterationsvorschrift dieses expliziten Verfahrens lautet

$$x_n = x_{n-1} + h_n f\left(x_{n-1} + \frac{h_n}{2} f\left(x_{n-1}, t_{n-1}\right), t_{n-1} + \frac{h_n}{2}\right).$$

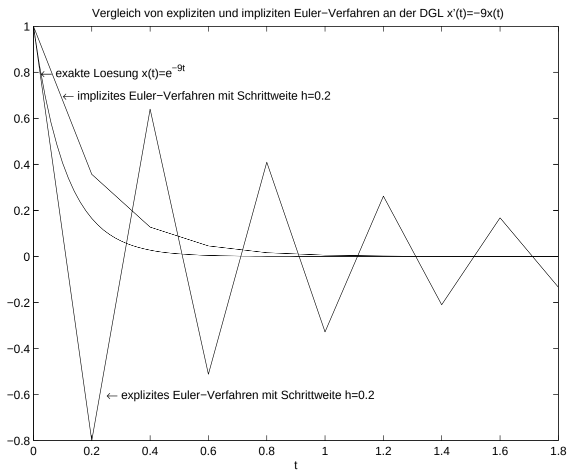
Für dieses Beispiel also

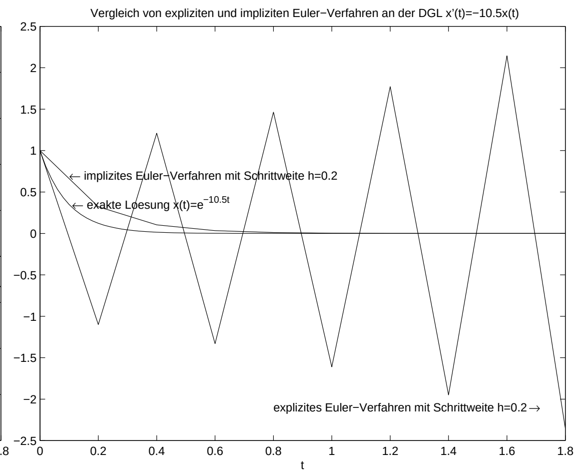
$$x_n = x_{n-1} + h_n \left(-10^5\right)\left(x_{n-1} + \frac{h_n}{2}\left(-10^5 x_{n-1}\right)\right) = \left(1 - 10^5 h_n + \frac{1}{2}\left(10^5 h_n\right)^2\right) x_{n-1}.$$

Für die Ungleichung (1.10) ergibt dies

$$|x_n| = \underbrace{\left|1 - 10^5 h_n + \frac{1}{2}\left(10^5 h_n\right)^2\right|}_{\leq 1?} |x_{n-1}| \overset{?}{\leq} |x_{n-1}|.$$

Dies ergibt wieder eine Einschränkung von h_n, so daß dieses Verfahren nicht B-stabil ist.





Allgemeines s-stufiges Runge-Kutta Verfahren:

Ein s-stufiges *Runge-Kutta-Verfahren* basiert auf den s *Zwischenschritten*

$$\boxed{t_{ni} := t_{n-1} + c_i h_n, \qquad i = 1, \ldots, s,}$$

zu *Knoten* $0 \leq c_1 \leq \ldots \leq c_s \leq 1$. Die Iterierten $\{x_n\}_{n\geq 0}$ werden vermöge der folgenden Vorschrift gebildet:

$$\boxed{\begin{aligned} x_n &= x_{n-1} + h_n \sum_{j=1}^{s} \beta_j X'_{nj}, \\ X'_{ni} &= f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{ij} X'_{nj}, t_{ni}\right), \qquad i = 1, \ldots, s, \end{aligned}} \tag{1.11}$$

mit *Gewichten* $\alpha_{i,j}$ und β_j, $i, j = 1, \ldots, s$. Die Parameter $c_1, \ldots, c_s$, $\beta_1, \ldots, \beta_s$ und $\alpha_{1,1}, \ldots, \alpha_{s,s}$ werden im *Butcher-Tableau* zusammengefasst (β ist ein Zeilenvektor!):

$$\boxed{\begin{array}{c|c} c & \mathcal{A} \\ \hline & \beta \end{array} = \begin{array}{c|ccc} c_1 & \alpha_{11} & \cdots & \alpha_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & \alpha_{s1} & \cdots & \alpha_{ss} \\ \hline & \beta_1 & \cdots & \beta_s \end{array}}$$

Die Bildungsvorschrift (1.11) liefert ein Gleichungssystem für $X'_{n1}, \ldots, X'_{ns}$. Bezeichne mit X_{ni} die *Zwischenstufenapproximationen*

$$\boxed{X_{ni} := x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{ij} X'_{nj}, \qquad i = 1, \ldots, s.} \tag{1.12}$$

Falls $\alpha_{i,j} = 0$ für $i \leq j$, so ergibt sich ein *explizites Runge-Kutta Verfahren*.

Beispiele für Runge-Kutta-Verfahren sind

1. *Explizites Euler-Verfahren*: Die Iterationsvorschrift lautet
$$x_n = x_{n-1} + h_n f\left(x_{n-1}, t_{n-1}\right).$$
$s = 1$, Butcher-Tableau: $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$
$$\begin{aligned} x_n &= x_{n-1} + h_n X'_{n1} \\ X'_{n1} &= f\left(x_{n-1}, t_{n-1}\right) \end{aligned}$$

2. *Implizites Euler-Verfahren*: Die Iterationsvorschrift lautet
$$x_n = x_{n-1} + h_n f\left(x_n, t_n\right).$$
$s = 1$, Butcher-Tableau: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$
$$\begin{aligned} x_n &= x_{n-1} + h_n X'_{n1} \\ X'_{n1} &= f\left(\underbrace{x_{n-1} + h_n X'_{n1}}_{x_n}, \underbrace{t_{n-1} + h_n}_{t_n}\right) \end{aligned}$$

3. *Implizite Mittelpunktregel*: Die Iterationsvorschrift lautet
$$x_n = x_{n-1} + h_n f\left(\frac{1}{2}\left(x_n + x_{n-1}\right), t_{n-1} + \frac{h_n}{2}\right).$$
$s = 1$, Butcher-Tableau: $\begin{array}{c|c} 0.5 & 0.5 \\ \hline & 1 \end{array}$
$$\begin{aligned} x_n &= x_{n-1} + h_n X'_{n1} \\ X'_{n1} &= f\left(x_{n-1} + \frac{h_n}{2} X'_{n1}, t_{n-1} + \frac{h_n}{2}\right) \end{aligned}$$

4. *Verbessertes Euler-Verfahren*: Die Iterationsvorschrift lautet
$$x_n = x_{n-1} + h_n f\left(x_{n-1} + \frac{h_n}{2} f\left(x_{n-1}, t_{n-1}\right), t_{n-1} + \frac{h_n}{2}\right).$$
$s = 2$, Butcher-Tableau: $\begin{array}{c|cc} 0 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ \hline & 0 & 1 \end{array}$
$$\begin{aligned} x_n &= x_{n-1} + h_n X'_{n2} \\ X'_{n1} &= f\left(x_{n-1}, t_{n-1}\right) \\ X'_{n2} &= f\left(x_{n-1} + \frac{h_n}{2} X'_{n1}, t_{n-1} + \frac{h_n}{2}\right) \end{aligned}$$

5. *Trapezregel*: Die Iterationsvorschrift lautet

$$x_n = x_{n-1} + \frac{h_n}{2}\left(f\left(x_{n-1}, t_{n-1}\right) + f\left(x_n, t_n\right)\right).$$

$s = 2$, Butcher-Tableau: $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline & 0.5 & 0.5 \end{array}$

$$\begin{aligned} x_n &= x_{n-1} + \frac{h_n}{2}X'_{n1} + \frac{h_n}{2}X'_{n2} \\ X'_{n1} &= f\left(x_{n-1}, t_{n-1}\right) \\ X'_{n2} &= f\left(\underbrace{x_{n-1} + \frac{h_n}{2}X'_{n1} + \frac{h_n}{2}X'_{n2}}_{x_n}, \underbrace{t_{n-1} + h_n}_{t_n}\right) \end{aligned}$$

Definition 1.17 (algebraisch stabil). Ein s-stufiges Runge-Kutta Verfahren heißt *algebraisch stabil*, falls $\beta_i \geq 0$, $i = 1, \ldots, s$, und die Matrix $M = (m_{i,j})$,

$$\boxed{m_{i,j} = \alpha_{i,j}\beta_i + \alpha_{j,i}\beta_j - \beta_i\beta_j, \quad i, j = 1, \ldots, s,}$$

positiv semidefinit ist.

$$\boxed{M = \mathcal{A}^T \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_s \end{pmatrix} + \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_s \end{pmatrix} \mathcal{A} - \beta^T\beta}$$

Satz 1.18 (M. Crouzeix, 1979). *Ein algebraisch stabiles Runge-Kutta Verfahren ist B-stabil.*

Beweis. Betrachte eine DGL $x'(t) = f(x(t), t)$, für die die Kontraktivitätsungleichung

$$\langle f(x,t) - f(\bar{x},t), x - \bar{x}\rangle \leq 0, \qquad x, \bar{x} \in \mathbb{R}^m, t \geq t_0,$$

gilt. Die Iterierten $\{x_n\}_{n\geq 0}$ des Runge-Kutta Verfahrens werden entsprechend (1.11) gebildet:

$$\begin{aligned} x_n &= x_{n-1} + h_n \sum_{j=1}^{s} \beta_j X'_{nj}, \\ X_{ni} &= x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{ij} X'_{nj}, \qquad i = 1, \ldots, s, \\ X'_{ni} &= f\left(X_{ni}, t_{ni}\right), \qquad i = 1, \ldots, s. \end{aligned}$$

Aufgrund der Kontraktivität ist dann

$$\left\langle X'_{ni} - \bar{X}'_{ni}, X_{ni} - \bar{X}_{ni}\right\rangle \leq 0 \qquad i = 1, \ldots, s;\ n \geq 1. \tag{1.13}$$

Wir haben zu zeigen: $|x_n - \bar{x}_n|^2 \leq |x_{n-1} - \bar{x}_{n-1}|^2$.

Nun ist

$$\begin{aligned} |x_n - \bar{x}_n|^2 &= \left\langle x_{n-1} - \bar{x}_{n-1} + h_n \sum_{j=1}^{s} \beta_j \left(X'_{nj} - \bar{X}'_{nj}\right), x_{n-1} - \bar{x}_{n-1} + h_n \sum_{j=1}^{s} \beta_j \left(X'_{nj} - \bar{X}'_{nj}\right)\right\rangle \\ &= |x_{n-1} - \bar{x}_{n-1}|^2 + A \\ \text{mit } A &:= 2h_n \left\langle \sum_{j=1}^{s} \beta_j \left(X'_{nj} - \bar{X}'_{nj}\right), x_{n-1} - \bar{x}_{n-1}\right\rangle + h_n^2 \sum_{j,i=1}^{s} \beta_j\beta_i \left\langle X'_{nj} - \bar{X}'_{nj}, X'_{ni} - \bar{X}'_{ni}\right\rangle \end{aligned}$$

Aus Formel (1.12) folgt

$$\begin{aligned} x_{n-1} - \bar{x}_{n-1} &= X_{nj} - h_n \sum_{l=1}^{s} \alpha_{jl} X'_{nl} - \left(\bar{X}_{nj} - h_n \sum_{l=1}^{s} \alpha_{jl} \bar{X}'_{nl} \right) \\ &= X_{nj} - \bar{X}_{nj} - h_n \sum_{l=1}^{s} \alpha_{jl} \left(X'_{nl} - \bar{X}'_{nl} \right), \qquad j = 1, \ldots, s, \end{aligned}$$

also

$$\begin{aligned} A &= h_n^2 \sum_{j,i=1}^{s} \beta_j \beta_i \left\langle X'_{nj} - \bar{X}'_{nj}, X'_{ni} - \bar{X}'_{ni} \right\rangle + 2h_n \sum_{j=1}^{s} \beta_j \left\langle X'_{nj} - \bar{X}'_{nj}, X_{nj} - \bar{X}_{nj} - h_n \sum_{l=1}^{s} \alpha_{jl} \left(X'_{nl} - \bar{X}'_{nl} \right) \right\rangle \\ &= h_n^2 \sum_{j,i=1}^{s} \beta_j \beta_i \left\langle X'_{nj} - \bar{X}'_{nj}, X'_{ni} - \bar{X}'_{ni} \right\rangle + h_n \sum_{j=1}^{s} \beta_j 2 \left\langle X'_{nj} - \bar{X}'_{nj}, X_{nj} - \bar{X}_{nj} \right\rangle \\ &\quad - h_n^2 \sum_{j=1}^{s} \beta_j 2 \left\langle X'_{nj} - \bar{X}'_{nj}, \sum_{l=1}^{s} \alpha_{jl} \left(X'_{nl} - \bar{X}'_{nl} \right) \right\rangle \\ &= h_n^2 \sum_{j,i=1}^{s} \beta_j \beta_i \left\langle X'_{nj} - \bar{X}'_{nj}, X'_{ni} - \bar{X}'_{ni} \right\rangle + h_n \sum_{j=1}^{s} \beta_j 2 \left\langle X'_{nj} - \bar{X}'_{nj}, X_{nj} - \bar{X}_{nj} \right\rangle \\ &\quad - h_n^2 \sum_{j=1}^{s} \beta_j \left(\left\langle X'_{nj} - \bar{X}'_{nj}, \sum_{l=1}^{s} \alpha_{jl} \left(X'_{nl} - \bar{X}'_{nl} \right) \right\rangle + \left\langle \sum_{l=1}^{s} \alpha_{jl} \left(X'_{nl} - \bar{X}'_{nl} \right), X'_{nj} - \bar{X}'_{nj} \right\rangle \right) \\ &= h_n \sum_{j=1}^{s} \beta_j \underbrace{2 \left\langle X'_{nj} - \bar{X}'_{nj}, X_{nj} - \bar{X}_{nj} \right\rangle}_{\leq 0 \quad \text{nach (1.13)}} - h_n^2 \sum_{j,i=1}^{s} \underbrace{\{ \beta_j \alpha_{ji} + \beta_i \alpha_{ij} - \beta_j \beta_i \}}_{m_{i,j} = m_{j,i}} \left\langle X'_{nj} - \bar{X}'_{nj}, X'_{ni} - \bar{X}'_{ni} \right\rangle. \end{aligned}$$

Sei nun

$$B := \sum_{j,i=1}^{s} m_{ij} \left\langle U_j, U_i \right\rangle, \qquad U_i \in \mathbb{R}^m, i = 1, \ldots s.$$

$M = (m_{i,j})$ ist eine symmetrische, positiv semidefinite Matrix und kann daher mit Cholesky zerlegt werden: $M = R^T R$ mit einer oberen Dreiecksmatrix R. Dann ist

$$m_{ij} = \sum_{l=1}^{s} r_{il}^T r_{lj} = \sum_{l=1}^{s} r_{li} r_{lj},$$

$$B = \sum_{j,i=1}^{s} \sum_{l=1}^{s} r_{li} r_{lj} \left\langle U_j, U_i \right\rangle = \sum_{l=1}^{s} \left\langle \sum_{j=1}^{s} r_{lj} U_j, \sum_{i=1}^{s} r_{li} U_i \right\rangle = \sum_{l=1}^{s} \left| \sum_{j=1}^{s} r_{lj} U_j \right|^2 \geq 0,$$

$$A = \underbrace{h_n \sum_{j=1}^{s} \beta_j 2 \left\langle X'_{nj} - \bar{X}'_{nj}, X_{nj} - \bar{X}_{nj} \right\rangle}_{\leq 0 \quad \text{nach (1.13)}} \underbrace{- h_n^2 \sum_{l=1}^{s} \left| \sum_{j=1}^{s} r_{lj} \left(X'_{nj} - \bar{X}'_{nj} \right) \right|^2}_{\leq 0} \leq 0. \tag{1.14}$$

Also gilt

$$\left| x_n - \bar{x}_n \right|^2 \leq \left| x_{n-1} - \bar{x}_{n-1} \right|^2 + A \leq \left| x_{n-1} - \bar{x}_{n-1} \right|^2. \qquad \square$$

Bemerkung 1.19. Was passiert bei kontraktiven Differentialgleichungen (mit $\gamma < 0$), d.h.

$$\left\langle f(x,t) - f(\bar{x},t), x - \bar{x} \right\rangle \leq \gamma \left| x - \bar{x} \right|^2 \qquad x, \bar{x} \in \mathbb{R}^m, t \geq t_0.$$

Formel (1.14) liefert nun

$$A \le h_n \sum_{j=1}^{s} \beta_j \cdot 2\gamma \left| X_{nj} - \bar{X}_{nj} \right|^2 \le 0.$$

Betrachte den Spezialfall des impliziten Euler-Verfahren: $s = 1$, $\alpha_{11} = \beta_1 = c_1 = 1$, $X_{n1} = x_n$. Dann ist

$$\begin{aligned} |x_n - \bar{x}_n|^2 &\le |x_{n-1} - \bar{x}_{n-1}|^2 + h_n \cdot 2\gamma \cdot |x_n - \bar{x}_n|^2 \\ \text{also } |x_n - \bar{x}_n|^2 &\le \underbrace{\frac{1}{1 - 2\gamma h_n}}_{\in(0,1)} |x_{n-1} - \bar{x}_{n-1}|^2 < |x_{n-1} - \bar{x}_{n-1}|^2 \end{aligned}$$

In [SH96, S. 387] wird (für autonome DGLn, $h_n = h$, $n \ge 1$) gezeigt

$$|x_n - \bar{x}_n|^2 \le \frac{(1 + hKL)^2}{(1 + hKL)^2 - 2\gamma h} |x_{n-1} - \bar{x}_{n-1}|^2, \quad n \ge 1,$$

für „Konstanten“ K und L, die nur von h, x_0 und $\bar{x}_0$ abhängen.

1.2.1 Dissipativität

Im letzten Abschnitt haben wir dynamische Systeme betrachte, deren Trajektorien asymptotisch gegen eine eindeutige Gleichgewichtslage oder eine abgeschlossene konvexe Menge von Gleichgewichtslagen konvergierten, und haben das Verhalten der exakten Lösung mit der numerisch approximierten Lösung verglichen. In diesem Abschnitt wollen wir diese Theorie nun auf dynamische Systeme verallgemeinern, deren Trajektorien nicht zu einem festen Punkt konvergieren müssen, sondern eine absorbierende Menge in einer endlichen Zeit erreichen und dort verbleiben (vgl. Def. 1.10, Seite 10).

Betrachte eine DGL, für die die Dissipativitätsungleichung (1.9),

$$\langle f(x,t), x \rangle \le \alpha - \beta |x|^2, \qquad x \in \mathbb{R}^m, t \ge t_0,$$

gilt, mit Konstanten $\alpha \ge 0$ und $\beta > 0$. Dann ist

$$\langle X'_{ni}, X_{ni} \rangle \le \alpha - \beta |X_{ni}|^2, \qquad i = 1, \ldots, s, n \ge 1.$$

Nun ist

$$\begin{aligned} |x_n|^2 &= \left\langle x_{n-1} + h_n \sum_{j=1}^{s} \beta_j X'_{nj}, x_{n-1} + h_n \sum_{j=1}^{s} \beta_j X'_{nj} \right\rangle \\ &= |x_{n-1}|^2 + 2h_n \sum_{i=1}^{s} \beta_i \langle X'_{ni}, x_{n-1} \rangle + h_n^2 \sum_{i,j=1}^{s} \beta_i \beta_j \langle X'_{ni}, X'_{nj} \rangle \\ &\overset{(1.12)}{=} |x_{n-1}|^2 + 2h_n \sum_{i=1}^{s} \beta_i \left\langle X'_{ni}, X_{ni} - h_n \sum_{j=1}^{s} \alpha_{ij} X'_{nj} \right\rangle + h_n^2 \sum_{i,j=1}^{s} \beta_i \beta_j \langle X'_{ni}, X'_{nj} \rangle \\ &= |x_{n-1}|^2 + 2h_n \sum_{i=1}^{s} \beta_i \langle X'_{ni}, X_{ni} \rangle - h_n^2 \sum_{i,j=1}^{s} (2\beta_i \alpha_{ij} - \beta_i \beta_j) \langle X'_{ni}, X'_{nj} \rangle \\ &= |x_{n-1}|^2 + 2h_n \sum_{i=1}^{s} \beta_i \langle X'_{ni}, X_{ni} \rangle - h_n^2 \sum_{i,j=1}^{s} m_{i,j} \langle X'_{ni}, X'_{nj} \rangle, \\ &= |x_{n-1}|^2 + 2h_n \sum_{i=1}^{s} \beta_i \underbrace{\langle X'_{ni}, X_{ni} \rangle}_{\le \alpha - \beta |X_{ni}|^2} - \underbrace{h_n^2 \sum_{l=1}^{s} \left| \sum_{j=1}^{s} r_{lj} X'_{nj} \right|^2}_{\ge 0}, \end{aligned}$$

also folgt

$$|x_n|^2 \le |x_{n-1}|^2 + 2h_n \sum_{i=1}^{s} \beta_i \left(\alpha - \beta |X_{ni}|^2\right) \tag{1.15}$$

für ein algebraisch stabiles Runge-Kutta-Verfahren.

Betrachte nun weiterhin den Spezialfall des *impliziten Euler-Verfahrens*: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$.

Dieses Verfahren ist algebraisch stabil, denn es ist $m_{i,j} = m_{1,1} = 1 + 1 - 1 = 1 \ge 0$, d.h. M ist positiv semidefinit. Aus den Formeln (1.11) und (1.12) folgt daher

$$\begin{aligned} x_n &= x_{n-1} + h_n X'_{n1}, \\ X_{n1} &= x_{n-1} + h_n X'_{n1}, \end{aligned}$$

also

$$x_n = X_{n1}.$$

Formel (1.15) liefert dann

$$\begin{aligned} |x_n|^2 &\le |x_{n-1}|^2 + 2h_n \left(\alpha - \beta |x_n|^2\right) \\ \rightsquigarrow \underbrace{(1 + 2h_n\beta)}_{\ge 1} |x_n|^2 &\le |x_{n-1}|^2 + 2h_n\alpha \\ \rightsquigarrow |x_n|^2 &\le \underbrace{\frac{1}{1 + 2h_n\beta}}_{=:g(h_n)} |x_{n-1}|^2 + \underbrace{\frac{2h_n\alpha}{1 + 2h_n\beta}}_{=:d_n}. \end{aligned}$$

Angenommen, es existiert eine Konstante κ, so daß

$$g(h_n) \le \kappa < 1, \qquad n \ge 1,$$

gilt ($\Leftrightarrow h_n \ge \frac{1-\kappa}{2\beta\kappa}$) und sei die Schrittweite von unten beschränkt, d.h. $h_n \ge h_{\min} > 0$, $n \ge 1$. Weiterhin ist

$$d_n = \frac{\alpha}{\beta} \cdot \frac{2h_n\beta}{1 + 2h_n\beta} \le \frac{\alpha}{\beta} =: d.$$

Dann folgt

$$|x_n|^2 \le \kappa |x_{n-1}|^2 + d \le \kappa^n |x_0|^2 + \sum_{j=0}^{n-1} \kappa^j d = \kappa^n |x_0|^2 + d\frac{1 - \kappa^n}{1 - \kappa} \xrightarrow{n\to\infty} \frac{d}{1 - \kappa}, \tag{1.16}$$

also

$$\limsup_{n\to\infty} |x_n|^2 \le \frac{d}{1 - \kappa}.$$

Lemma 1.20. *Jede Menge*

$$\mathcal{A} := \bar{B}\left(0, \sqrt{\frac{d}{1 - \kappa} + \varepsilon}\right), \quad \varepsilon > 0,$$

ist eine absorbierende Menge für das implizite Euler-Verfahren.

Beweis. Positive Invarianz von $\mathcal{A}$: Für $x_0 \in \mathcal{A}$ ist zu zeigen: $x_n \in \mathcal{A}$ für alle $n \geq 1$.

Wegen (1.16) und $|x_0| \leq \sqrt{\frac{d}{1-\kappa} + \varepsilon}$ ist nun

$$|x_n|^2 \leq \kappa^n \cdot \left(\frac{d}{1-\kappa} + \varepsilon \right) + d \frac{1-\kappa^n}{1-\kappa} = \kappa^n \varepsilon + \frac{d}{1-\kappa} \leq \frac{d}{1-\kappa} + \varepsilon, \qquad n \geq 1,$$

d.h. $x_n \in \mathcal{A}$ für alle $n \geq 1$.

Absorptionseigenschaft von $\mathcal{A}$: Sei $\mathcal{B}$ eine beschränkte Menge, d.h. $\mathcal{B} \subseteq \bar{B}(0, r)$ für ein $r > 0$.

Zu zeigen: Es existiert ein $n_* := n_*(\mathcal{A}, \mathcal{B}) := n_*(\varepsilon, r)$, so daß bei Wahl von $x_0 \in \mathcal{B}$ folgt, daß $x_n \in \mathcal{A}$ für alle $n \geq n_*$ gilt.

Für $x_0 \in \mathcal{B}$ gilt wegen (1.16):

$$|x_n|^2 \leq \kappa^n r^2 + d \frac{1-\kappa^n}{1-\kappa} \xrightarrow{n \to \infty} \frac{d}{1-\kappa}.$$

Dann existiert zu $\varepsilon > 0$ ein $n_* = n_*(\varepsilon, r)$ mit der Eigenschaft, daß

$$\kappa^n r^2 + d \frac{1-\kappa^n}{1-\kappa} \leq \frac{d}{1-\kappa} + \varepsilon, \qquad n \geq n_*,$$

d.h. $|x_n|^2 \leq \frac{d}{1-\kappa} + \varepsilon$, $n \geq n_*$, also $x_n \in \mathcal{A}$ für alle $n \geq n_*$. □

Bemerkung 1.21. Speziell für äquidistante Zerlegungen $h_n = h$, $n \in \mathbb{N}$, gilt $\kappa = \frac{1}{1+2h\beta}$ und $d_n = \frac{\alpha}{\beta} \frac{2h\beta}{1+2h\beta} = \frac{\alpha}{\beta}(1-\kappa) =: d$, d.h. $\frac{d}{1-\kappa} = \frac{\alpha}{\beta}$. Also sind die Kugeln $\bar{B}\left(0, \sqrt{\frac{\alpha}{\beta} + \varepsilon}\right)$, $\varepsilon > 0$, absorbierende Mengen für das implizite Euler-Verfahren (vgl. mit Satz 1.12, Seite 11).

Zusammenfassend erhalten wir folgende Aussage:

Satz 1.22. *Sei für die DGL (1.7) die Dissipativitätsungleichung (1.9) erfüllt.*

Dann bildet das implizite Euler-Verfahren die Dissipativitätseigenschaft der DGL qualitativ richtig nach, und zwar ohne praktisch relevante Einschränkung der Schrittweiten.

Jede Kugel $\bar{B}\left(0, \sqrt{\frac{d}{1-\kappa} + \varepsilon}\right)$, $\varepsilon > 0$, *ist absorbierende Menge für die Approximationen.*

Weiterhin gilt folgender Satz:

Satz 1.23 ([SH96, S. 403]). *Sei für die DGL (1.7) die Dissipativitätsungleichung (1.9) erfüllt.*

Dann bildet ein s-stufiges, algebraisch stabiles Runge-Kutta-Verfahren mit $\beta_i > 0$, $i = 1, \ldots, s$, *bei äquidistanter Schrittweite* $h_n = h > 0$, $n \geq 1$, *die Dissipativität qualitativ richtig nach.*

Jede Kugel $B\left(0, \sqrt{\frac{\alpha}{\beta} + h \cdot c(h) + \varepsilon}\right)$, $\varepsilon > 0$, *ist absorbierende Menge* ($c(h)$ *ist eine von* h *abhängige Konstante).*

1.3 Stationäre Lösungen und Geisterlösungen

Sei durch die DGL $x'(t) = f(x(t), t)$ ein dynamisches System auf $\mathbb{R}^m$ gegeben (d.h. für jeden Anfangswert $x_0 \in \mathbb{R}^m$ ist die Lösung auf $[0, \infty)$ gegeben). Bezeichne mit $\mathcal{E}$ die Menge aller Gleichgewichtslagen der DGL:

$$\boxed{\mathcal{E} := \{x \in \mathbb{R}^m \,|\, f(x,t) = 0 \quad \forall t \in [0, \infty)\}}$$

Zu $x_0^* \in \mathcal{E}$ ist $x_*(\cdot) \equiv x_0^*$ stationäre Lösung. Bei äquidistanter Zerlegung $t_n := nh$, $n \in \mathbb{N}$, $t_0 = 0$, mit $h > 0$ ist $\{x_*(t_n)\}_{n \geq 0} = \{x_0^*\}_{n \geq 0}$ dann eine stationäre Folge.

Ein Integrationsverfahren liefert zu einem Anfangswert x_0 eine Folge $\{x_n\}_{n \geq 0}$. Ist auch die Folge $\{x_n\}_{n \geq 0}$ mit $x_0 = x_0^*$ stationär und liefert das Integrationsverfahren <u>nur</u> dann eine stationäre Folge, wenn es bzgl. einer Gleichgewichtslage gestartet wird? D.h. gilt für die Menge

$$\boxed{\mathcal{E}_{h,\text{Verfahren}} := \{x \in \mathbb{R}^m \,|\text{aus } x_0 = x \text{ folgt } x_n = x, \quad n \geq 1\}}$$

$$\mathcal{E} = \mathcal{E}_{h,\text{Verfahren}} \quad ?$$

Satz 1.24. *Für alle konsistenten und stabilen linearen Mehrschrittverfahren und alle Runge-Kutta-Methoden gilt*

$$\boxed{\mathcal{E} \subseteq \mathcal{E}_{h,\text{Verfahren}}}$$

Beweis. Zu $x_0^* \in \mathcal{E}$ ist zu zeigen, daß die Iteriertenfolge $\{x_n\}$ des Verfahrens bei Startwert x_0^* stationär ist: $x_n = x_0^*$, $n \geq 1$.

1. Betrachte ein konsistentes stabiles lineares Mehrschrittverfahren:

$$\sum_{j=0}^{k} \alpha_j x_{n-j} = h \sum_{j=0}^{k} \beta_j f(x_{n-j}, t_{n-j}), \quad n \geq k. \tag{1.17}$$

   Es gelten die Konsistenzbedingungen $\sum_{j=0}^{k} \alpha_j = 0$ und $\sum_{j=0}^{k} \alpha_j j + \beta_j = 0$ und das Dahlquistsche Wurzelkriterium.

   Wegen $x_0^* \in \mathcal{E}$ ist $f(x_0^*, t_n) = 0$ für alle $n \in \mathbb{N} \cup \{0\}$. Wir zeigen, daß die Folge $x_n := x_0^*$, $n \geq 1$, die Rekursion (1.17) erfüllt: Es ergibt sich die Gleichung

$$\sum_{j=0}^{k} \alpha_j x_0^* = h \sum_{j=0}^{k} \beta_j \underbrace{f(x_0^*, t_{n-j})}_{=0} = 0, \quad n \geq k,$$

   welche aufgrund der Konsistenzbedingung $\sum_{j=0}^{k} \alpha_j = 0$ erfüllt ist. Also ist $x_0^* \in \mathcal{E}_{h,\text{lin.MS.}}$.

2. Betrachte nun ein beliebiges s-stufiges Runge-Kutta Verfahren $\begin{array}{c|c} c & \mathcal{A} \\ \hline & \beta \end{array}$.

   Wir zeigen, daß die Folge $x_n := x_0^*$, $X'_{ni} := 0$, $i = 1, \ldots, s$, $n \geq 1$, die Runge-Kutta-Rekursion erfüllt: Es ist

$$x_n = x_{n-1} + h \sum_{j=1}^{s} \beta_j X'_{nj} = x_{n-1} = x_{n-2} + 0 = \ldots = x_0 = x_0^*, \qquad n \geq 1$$

$$X'_{ni} = f\left(x_{n-1} + h \sum_{j=1}^{s} \alpha_{ij} X'_{nj}, t_{ni}\right) = f(x_{n-1}, t_{ni}) = f(x_0^*, t_{ni}) = 0, \qquad i = 1, \ldots, s, n \geq 1.$$

   Also ist $x_0^* \in \mathcal{E}_{h,\text{RungeKutta}}$. □

Satz 1.25. *Für konsistente, stabile lineare Mehrschrittverfahren gilt bei Anwendung auf eine autonome DGL (1.1) auch*

$$\boxed{\mathcal{E}_{h,\text{lin.MS.}} \subseteq \mathcal{E}}$$

Beweis. Sei nun die autonome DGL $x'(t) = f(x(t))$ gegeben. Für eine stationäre Folge $\{x_n\}_{n\geq 0}$ von Iterierten, d.h. $x_n = x^* \in \mathcal{E}_{h,\text{lin.MS.}}$, $n \geq 1$, ist zu zeigen, daß $x^* \in \mathcal{E}$ gilt.

Aufgrund der Stabilitätseigenschaft des linearen Mehrschrittverfahrens gilt[1] $\sum_{j=0}^{k} \beta_j \neq 0$. Weiterhin haben wir die Konsistenzbedingung $\sum_{j=0}^{k} \alpha_j = 0$.
Einsetzen der stationären Folge in die erste Gleichung ergibt dann

$$h \underbrace{\sum_{j=0}^{k} \beta_j}_{\neq 0} f(x^*) = \underbrace{\sum_{j=0}^{k} \alpha_j}_{=0} x^*.$$

Also folgt $f(x^*) = 0$, d.h. $x^* \in \mathcal{E}$. □

Beispiel. Sei $m = 1$, $\lambda \neq 0$ und betrachte die DGL

$$x'(t) = -\lambda \frac{x(t)}{1 + x^2(t)},$$

$f(x,t) = -\lambda \frac{x}{1+x^2}$. Diese DGL besitzt nur eine stationäre Lösung, $\mathcal{E} = \{0\}$.

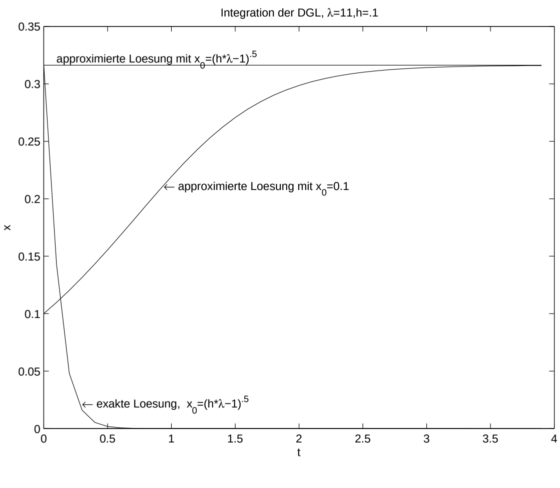
Betrachte weiterhin das Runge Kutta-Verfahren zum Butcher-Tableau $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 0 & 1 \end{array}$ und äquidistanten Schrittweiten $h_n = h$, $n \geq 1$. Die Iterationsvorschrift lautet also

$$\begin{aligned} x_n &= x_{n-1} + hX'_{n2} \\ X'_{n1} &= f(x_{n-1}, t_{n-1}) = -\lambda \frac{x_{n-1}}{1 + x_{n-1}^2} \\ X'_{n2} &= f(x_{n-1} + hX'_{n1}, t_{n1}) \end{aligned}$$

Dann ist

$$x_n = x_{n-1} - \lambda h \frac{x_{n-1} - \lambda h \frac{x_{n-1}}{1+x_{n-1}^2}}{1 + \left(x_{n-1} - \lambda h \frac{x_{n-1}}{1+x_{n-1}^2}\right)^2}$$

und die Folge $\{x_n\}_{n\geq 1}$ ist stationär, gdw. $x_n = 0$ oder



$$x_n = x_{n-1} \quad \Leftrightarrow \quad 1 - \lambda h \frac{1}{1 + x_{n-1}^2} = 0 \quad \Leftrightarrow \quad x_{n-1}^2 = h\lambda - 1 \quad \Leftrightarrow \quad x_{n-1} = \pm\sqrt{h\lambda - 1}.$$

Falls $h\lambda - 1 > 0$, so sind die Folgen $x_n = +\sqrt{h\lambda - 1}$, $n \geq 0$, und $x_n = -\sqrt{h\lambda - 1}$, $n \geq 0$, stationär. Es ist also

$$\mathcal{E}_{h,\text{Verfahren}} = \left\{\sqrt{h\lambda - 1}, 0, -\sqrt{h\lambda - 1}\right\} \supset \mathcal{E}.$$

[1] Annahme: $0 = -\sum_{j=0}^{k} \beta_j = \sum_{j=0}^{k} \alpha_j j$. Für das charakteristische Polynom des linearen Mehrschrittverfahrens $\rho(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^{k-j}$ ist dann $\rho(1) = 0$ und $\rho'(1) = 0$. D.h. 1 ist doppelte Nullstelle von ρ. Dies ist ein Widerspruch zum Dahlquistschem Wurzelkriterium.

Kapitel 2

Aufbau und Eigenschaften von Runge-Kutta Verfahren

Allgemeines

Wir behandeln die DGL

$$x'(t) = f(x(t), t). \tag{2.1}$$

Euler 1768:

$$x_n = x_{n-1} + h_n \cdot f(x_{n-1}, t_{n-1})$$

Runge begann 1895 die Verfahrensentwicklung:

$$x_n = x_{n-1} + h_n f\left(x_{n-1} + \frac{h_n}{2} f(x_{n-1}, t_{n-1}), t_{n-1} + \frac{h_n}{2}\right).$$

Kutta schrieb 1905 ein allgemeines „explizites Runge-Kutta Verfahren“ auf:

$$\begin{aligned} x_n &= x_{n-1} + h_n \sum_{i=1}^{s} \beta_i X'_{ni}, \\ X'_{ni} &= f\left(x_{n-1} + h_n \sum_{j=1}^{i-1} \alpha_{ij} X'_{nj}, t_{n-1} + c_i h_n\right), \qquad i = 2, \ldots, s, \\ X'_{n1} &= f(x_{n-1}, t_{n-1}), \end{aligned}$$

d.h. das Butcher-Tableau $\begin{array}{c|ccc} c_1 & \alpha_{11} & \cdots & \alpha_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & \alpha_{s1} & \cdots & \alpha_{ss} \\ \hline & \beta_1 & \cdots & \beta_s \end{array}$ mit $c_1 = 0$ und $\alpha_{ij} = 0$ für $j = i, \ldots, s$, $i = 1, \ldots, s$.

Ziel. Für die Runge-Kutta-Approximationen $\bar{x}_n$, die ausgehend von $x(t_{n-1})$ anstatt x_{n-1} gebildet werden, soll

$$\boxed{x(t_n) - \bar{x}_n = O\left(h_n^{p+1}\right)}$$

mit möglichst großen p gelten. p heißt *Konsistenzordnung*.

Welche Bedingungen müssen nun an die Parameter $\mathcal{A}, \beta, c$ des Verfahrens gestellt werden, damit eine möglichst hohe Konsistenzordnung erreicht werden kann?

Für äquidistante Schrittweiten $h_n = h$ wird $\bar{x}_n$ durch Lösung des folgenden Gleichungssystem berechnet:

$$\begin{aligned}
\bar{x}_n &= x\left(t_{n-1}\right) + h \sum_{i=1}^{s} \beta_i \bar{X}'_{ni}, \\
\bar{X}'_{ni} &= f\left(x\left(t_{n-1}\right) + h \sum_{j=1}^{i-1} \alpha_{ij} \bar{X}'_{nj}, t_{n-1} + c_i h\right), \qquad i = 2, \ldots, s, \\
\bar{X}'_{n1} &= f\left(x\left(t_{n-1}\right), t_{n-1}\right).
\end{aligned}$$

Taylorentwicklung um $\left(x\left(t_{n-1}\right), t_{n-1}\right)$ ergibt dann für die Approximation

$$\begin{aligned}
\bar{X}'_{n1} &= f\left(x\left(t_{n-1}\right), t_{n-1}\right), \\
\bar{X}'_{n2} &= f\left(x\left(t_{n-1}\right), t_{n-1}\right) + h\alpha_{21} f_x\left(x\left(t_{n-1}\right), t_{n-1}\right) \bar{X}'_{n1} + c_2 f_t\left(x\left(t_{n-1}\right), t_{n-1}\right) h + O\left(h^2\right) \\
&= f\left(x\left(t_{n-1}\right), t_{n-1}\right) + h\alpha_{21} f_x\left(x\left(t_{n-1}\right), t_{n-1}\right) f\left(x\left(t_{n-1}\right), t_{n-1}\right) + c_2 f_t\left(x\left(t_{n-1}\right), t_{n-1}\right) h + O\left(h^2\right) \\
&\ \vdots \\
\bar{X}'_{ni} &= f + h \sum_{j=1}^{i-1} \alpha_{ij} f_x\left(f + h \ldots\right) + h c_i f_t + O\left(h^2\right) \\
&= f + h \sum_{j=1}^{i-1} \alpha_{ij} f_x f + h c_i f_t + O\left(h^2\right) \qquad \text{da } h \sum \alpha_{ij} f_x\left(h \ldots\right) = O\left(h^2\right) \\
\bar{x}_n &= x\left(t_{n-1}\right) + h \sum_{i=1}^{s} \beta_i \left(f + h \sum_{j=1}^{i-1} \alpha_{ij} f_x f + h c_i f_t + O\left(h^2\right)\right),
\end{aligned}$$

während für die exakte Lösung

$$x\left(t_n\right) = x\left(t_{n-1}\right) + h x'\left(t_{n-1}\right) + \frac{1}{2} h^2 x''\left(t_{n-1}\right) + O\left(h^3\right).$$

Nun gilt

$$\begin{aligned}
x'(t) &= f\left(x(t), t\right), \\
x''(t) &= f_x\left(x(t), t\right) \underbrace{x'(t)}_{f(x(t),t)} + f_t\left(x(t), t\right),
\end{aligned}$$

also

$$x\left(t_n\right) = x\left(t_{n-1}\right) + h f + \frac{1}{2} h^2 \left(f_x f + f_t\right) + O\left(h^3\right).$$

Die Approximation macht also beim Schritt von der exakten Lösung $x\left(t_{n-1}\right)$ zum approximierten Punkt $\bar{x}_n$ einen Fehler von

$$x\left(t_n\right) - \bar{x}_n = h\left(1 - \sum_{i=1}^{s} \beta_i\right) f + h^2 \left(\frac{1}{2} - \sum_{i=1}^{s} \beta_i \sum_{j=1}^{i-1} \alpha_{ij}\right) f_x f + h^2 \left(\frac{1}{2} - \sum_{i=1}^{s} \beta_i c_i\right) f_t + O\left(h^3\right).$$

Hieraus kann man nun Bedingungen dafür ablesen, daß man $p = 1$ oder $p = 2$ erreicht:

$$\begin{aligned}
p = 1 \quad &: \quad \sum_{i=1}^{s} \beta_i = 1 \\
p = 2 \quad &: \quad \sum_{i=1}^{s} \beta_i = 1, \quad \sum_{i=1}^{s} \beta_i \sum_{j=1}^{i-1} \alpha_{ij} = \frac{1}{2}, \quad \sum_{i=1}^{s} \beta_i c_i = \frac{1}{2}.
\end{aligned}$$

Meist wird $c_i = \sum_{j=1}^{i-1} \alpha_{ij}$ gesetzt (vgl. auch Def. 2.7, Seite 29).

1978 berechnete E. Hairer, wieviele Stufen und Gleichungen nötig sind, um bestimmte Konsistenzordnungen zu erreichen. Dies führt zur Theorie der Butcher-Reihen.

p	1	2	3	5	10
zu lösende Gleichungen	1	2	4	17	1205
notwendige Stufenzahl	1	2	3	6	≤ 17

Definition 2.1 (Runge-Kutta-Verfahren). Ein allgemeines *Runge-Kutta-Verfahren* (IRK) hat die Form

$$\boxed{\begin{aligned} x_n &= x_{n-1} + h_n \sum_{i=1}^{s} \beta_i X'_{ni} \\ X'_{ni} &= f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{ij} X'_{nj}, t_{n-1} + c_i h_n\right), \quad i = 1, \ldots, s. \end{aligned}}$$

Spezielle Runge-Kutta-Verfahren ergeben sich durch Einschränkung der Parametermenge:

Explizite Runge-Kutta-Verfahren (ERK): $c_1 = 0$, $\alpha_{ij} = 0$, $j = i, \ldots, s$, $i = 1, \ldots, s$

Diagonal implizite Runge-Kutta-Verfahren (DIRK): $\alpha_{ij} = 0$, $j = i+1, \ldots, s$, $i = 1, \ldots, s$, $\alpha_{ii} \neq 0$ für mindestens ein i

singly diagonal implicit Runge-Kutta-methods (SDIRK): $\alpha_{ij} = 0$, $j = i+1, \ldots, s$, $\alpha_{ii} = \gamma \neq 0$, $i = 1, \ldots, s$

Beispiel ("DAS Runge-Kutta-Verfahren"). Das klassische Runge-Kutta-Verfahren ist eine Verallgemeinerung der Simpson-Regel und hat Konsistenzordnung $p = 4$:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 1/2 & 1/2 & 0 & & \\ 1/2 & 0 & 1/2 & 0 & \\ 0 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array}$$

Definition 2.2 (Einschrittverfahren, Inkrementfunktion). Ein allgemeines *Einschrittverfahren* hat die Form

$$\boxed{x_n = x_{n-1} + h_n \varphi\left(x_{n-1}, t_{n-1}, h_n\right)} \tag{2.2}$$

Die Funktion φ heißt dabei *Verfahrensfunktion* bzw. *Inkrementfunktion*.

Beispiel. Für Runge-Kutta-Verfahren ist

$$\varphi\left(x_{n-1}, t_{n-1}, h_n\right) = \sum_{i=1}^{s} \beta_i X'_{ni}$$

mit $X'_{n1}, \ldots, X'_{ns}$ aus dem Gleichungssystem

$$\begin{pmatrix} X'_{n1} \\ \vdots \\ X'_{ns} \end{pmatrix} = \begin{pmatrix} f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{1j} X'_{nj}, t_{n1}\right) \\ \vdots \\ f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{sj} X'_{nj}, t_{ns}\right) \end{pmatrix} =: F\left(X'_{n1}, \ldots, X'_{ns}\right), \tag{2.3}$$

für $t_{ni} = t_{n-1} + c_i h_n$, $i = 1, \ldots, s$.

φ ist wohldefiniert, wenn die Gleichung (2.3) eindeutig lösbar ist.

Ist F eine Kontraktion? Sei dazu $|f(x,t) - f(\bar{x},t)| \le L_f |x - \bar{x}|$, $x, \bar{x} \in \mathbb{R}^m$. Dann ist

$$\begin{aligned}
&\left|F(Z_{n1},\dots,Z_{ns}) - F\left(\bar{Z}_{n1},\dots,\bar{Z}_{ns}\right)\right|_{\mathbb{R}^{ms}} \\
= &\left|\left[\begin{array}{c} f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{1j} Z_{nj}, t_{n1}\right) - f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{1j} \bar{Z}_{nj}, t_{n1}\right) \\ \vdots \\ f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{sj} Z_{nj}, t_{ns}\right) - f\left(x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{sj} \bar{Z}_{nj}, t_{ns}\right) \end{array}\right]\right|_{\mathbb{R}^{ms}} \\
\le &\ \gamma \max_{i=1,\dots,s} |f(\dots,t_{ni}) - f(\dots,t_{ni})|_{\mathbb{R}^m} \qquad \text{mit Normumwandlungskonstante } \gamma \\
\le &\ \gamma L_f \max_{i=1,\dots,s} \left| h_n \sum_{j=1}^{s} \alpha_{ij} \left(Z_{nj} - \bar{Z}_{nj}\right) \right|_{\mathbb{R}^m} \\
\le &\ \gamma L_f h_n \max_{i=1,\dots,s} \sum_{j=1}^{s} |\alpha_{ij}| \max_{j=1,\dots,s} \left|Z_{nj} - \bar{Z}_{nj}\right|_{\mathbb{R}^m} \\
\le &\ \underbrace{\gamma \bar{\gamma} L_f h_n \max_{i=1,\dots,s} \sum_{j=1}^{s} |\alpha_{ij}|}_{<1 \text{ für hinreichend kleine } h_n} \left|\left[\begin{array}{c} Z_{n1} - \bar{Z}_{n1} \\ \vdots \\ Z_{ns} - \bar{Z}_{ns} \end{array}\right]\right|_{\mathbb{R}^{ms}} \qquad \text{mit weiterer Normumwandlungskonst. } \bar{\gamma}.
\end{aligned}$$

Nach dem Banachschen Fixpunktsatz existiert dann genau eine Lösung der (Fixpunkt-)Gleichung (2.3).

Definition 2.3 (lokaler Diskretisierungsfaktor, lokaler Fehler). Die Größe

$$\boxed{\tau_n := \frac{1}{h_n}\left(x(t_n) - x(t_{n-1})\right) - \varphi\left(x(t_{n-1}), t_{n-1}, h_n\right)}$$

heißt *lokaler Diskretisierungsfaktor*.

τ_n beschreibt also den Defekt der Lösung der DGL in der Näherungsformel (2.2):

$$\tau_n = \frac{1}{h_n} \underbrace{\left(x(t_n) - \bar{x}_n\right)}_{\text{lokaler Fehler}}.$$

$x(t_n) - \bar{x}_n$ heißt *lokaler Fehler*.

Definition 2.4 (Konsistenzordnung). Ein Runge-Kutta-Verfahren hat die *Konsistenzordnung* $p \in \mathbb{N}$, falls

$$\boxed{|x(t_n) - \bar{x}_n| \le c\, h_n^{p+1} \quad \text{bzw.} \quad |\tau_n| \le c\, h_n^p}$$

für eine Konstante $c \in \mathbb{R}$ gilt.

2.1 Die vereinfachenden Konsistenzbedingungen von John Butcher

Bei der Herleitung der Bedingungen für Konsistenzordnung $p = 1, 2$ (Seite 24) haben wir gesehen, dass die Bestimmung der Taylorentwicklung für die Zwischenstufenapproximationen schnell unübersichtlich wird. Wir werden uns daher im folgenden mit Differentialgleichungen der Form

$$x'(t) = f(t), \qquad x(t_{n-1}) = 0, \qquad h_n =: h.$$

auseinandersetzen. Wir werden dann die aus der Numerik I bekannten Resultate über Quadraturformeln auf die Konstruktion von Runge-Kutta-Verfahren übertragen.

Gegeben sei ein s-stufiges Runge-Kutta-Verfahren $\begin{array}{c|c} c & \mathcal{A} \\ \hline & \beta \end{array}$.

Ziel ist es, eine „gute“ Quadraturformel

$$x(t_n) = \int_{t_{n-1}}^{t_n} f(t)\,dt \approx \sum_{j=1}^{s} \beta_j f(t_{nj})$$

zu erhalten. Das Runge-Kutta-Verfahren liefert

$$\bar{x}_n = h \sum_{j=1}^{s} \beta_j \bar{X}'_{nj}, \qquad \bar{X}'_{ni} = f(t_{ni}), \quad i = 1, \ldots, s.$$

also

$$\bar{x}_n = h \sum_{j=1}^{s} \beta_j f(t_{nj}) = h \sum_{j=1}^{s} \beta_j f(t_{n-1} + c_j h).$$

Für $f(t) = (t - t_{n-1})^{k-1}$ ist dann:

$$\begin{aligned} x(t_n) &= \int_{t_{n-1}}^{t_n} (t - t_{n-1})^{k-1}\,dt = \frac{1}{k} h^k \\ \text{und } \bar{x}_n &= h \sum_{j=1}^{s} \beta_j c_j^{k-1} h^{k-1}. \end{aligned}$$

Der lokale Fehler beträgt dann

$$x(t_n) - \bar{x}_n = \left(\frac{1}{k} - \sum_{j=1}^{s} \beta_j c_j^{k-1} \right) h^k.$$

Möchten wir diesen lokalen Fehler eliminieren, so führt dies zu folgender Bedingung:

Definition 2.5 ($B(p)$). Die *Butcher-Bedingung* $B(p)$ lautet:

$$\boxed{\sum_{j=1}^{s} \beta_j c_j^{k-1} = \frac{1}{k} \qquad k = 1, \ldots, p.} \tag{$B(p)$}$$

Speziell ist

$$\begin{aligned} B(1): &\quad \sum_{j=1}^{s} \beta_j = 1 \\ B(2): &\quad \sum_{j=1}^{s} \beta_j = 1, \quad \sum_{j=1}^{s} \beta_j c_j = \frac{1}{2}. \end{aligned}$$

Lemma 2.6. 1. *Die Butcher-Bedingung $B(p)$ ist notwendig für die Konsistenzordnung p.*

2. *Für ein s-stufiges Runge-Kutta-Verfahren mit Konsistenzordnung p muß $p \leq 2s$ gelten.*

Beweis. 1. klar.

2. Wir zeigen indirekt, daß $p = 2s + 1$ nicht gelten kann.
   Annahme: $p = 2s + 1$. Wegen 1. gilt dann $B(2s+1)$, d.h. die Quadraturformel

$$\int_{t_{n-1}}^{t_n} f(t)\,dt \approx h_n \sum_{j=1}^{s} \beta_j f(t_{ni})$$

   ist genau für alle Polynome mit Grad $\leq 2s$.
   Wir wählen

$$f(t) = \prod_{i=1}^{s} (t - t_{ni})^2 .$$

   Es gilt

$$\int_{t_{n-1}}^{t_n} f(t)\,dt > 0, \quad \text{aber} \quad \sum_{j=1}^{s} \beta_j f(t_{ni}) = 0.$$

   Widerspruch. □

Wir betrachten weiterhin die Zwischenstufen

$$\begin{aligned} X_{ni} &= h \sum_{j=1}^{s} \alpha_{ij} X'_{nj}, \qquad X'_{nj} = f(t_{n-1} + c_j h) \\ \rightsquigarrow X_{ni} &= h \sum_{j=1}^{s} \alpha_{ij} f(t_{n-1} + c_j h). \end{aligned}$$

Das Ziel ist es nun, eine „gute" Quadraturformel für die Zwischenstufen zu erhalten:

$$x(t_{ni}) = \int_{t_{n-1}}^{t_{ni}} f(t)\,dt \approx h \sum_{j=1}^{s} \alpha_{ij} f(t_{nj}).$$

Für $f(t) = (t - t_{n-1})^{k-1}$ liefert das Runge-Kutta-Verfahren

$$X_{ni} = h \sum_{j=1}^{s} \alpha_{ij} c_j^{k-1} h^{k-1},$$

während die exakte Lösung

$$x(t_{ni}) = \int_{t_{n-1}}^{t_{ni}} f(t)\,dt = \frac{1}{k} c_i^k h^k$$

beträgt. Der lokale Fehler ist also

$$x(t_{ni}) - X_{ni} = \left(\frac{1}{k} c_i^k - \sum_{j=1}^{s} \alpha_{ij} c_j^{k-1} \right) h^k.$$

Dies führt zu folgender Bedingung:

Definition 2.7 ($C(l)$). Die Bedingung $C(l)$ lautet

$$\boxed{\sum_{j=1}^{s} \alpha_{ij} c_j^{k-1} = \frac{1}{k} c_i^k \qquad i = 1, \ldots, s,\ k = 1, \ldots, l.} \tag{$C(l)$}$$

Speziell ist

$$C(1): \qquad \sum_{j=1}^{s} \alpha_{ij} = c_i \qquad i = 1, \ldots, s.$$

Weiterhin gelte folgende Bedingung:

Definition 2.8 ($D(q)$). Die Bedingung $D(q)$ lautet

$$\boxed{\sum_{i=1}^{s} \beta_i c_i^{k-1} \alpha_{i,j} = \frac{1}{k} \beta_j \left(1 - c_j^k\right) \qquad j = 1, \ldots, s,\ k = 1, \ldots, q.} \tag{$D(q)$}$$

$B(p), C(l), D(q)$ heißen *vereinfachte Konsistenzbedingungen* nach J. Butcher.

Satz 2.9 (Butcher, 1964). *Seien $B(p)$, $C(l)$, $D(q)$, $p \le l + q + 1$, $p \le 2l + 2$, erfüllt. Dann hat das Runge-Kutta-Verfahren die Konsistenzordnung p.*

Lemma 2.10. *Für ein s-stufiges Runge-Kutta-Verfahren gelte $\beta_i > 0$, $i = 1, \ldots, s$, und die $c_1, \ldots, c_s$ seien paarweise voneinander verschieden. Dann gilt:*

1. $B(s+m)$, $C(s) \Rightarrow D(m)$
2. $B(s+l)$, $D(s) \Rightarrow C(l)$

Beweis. 1. Für $j = 1, \ldots, s$, $k = 1, \ldots, m$ sei

$$r_{jk} := \sum_{i=1}^{s} \beta_i \alpha_{i,j} c_i^{k-1} - \frac{1}{k} \beta_j \left(1 - c_j^k\right).$$

Es ist zu zeigen: $r_{jk} = 0$. Wir zeigen:

$$\sum_{j=1}^{s} r_{jk} c_j^{\nu-1} = 0 \qquad \nu = 1, \ldots, s,$$

$$\text{d.h.} \begin{pmatrix} c_1^0 & \cdots & c_s^0 \\ c_1 & \cdots & c_s \\ \vdots & & \vdots \\ c_1^{s-1} & \cdots & c_s^{s-1} \end{pmatrix} \begin{pmatrix} r_{1k} \\ \vdots \\ r_{sk} \end{pmatrix} = 0.$$

Da die $c_1, \ldots, c_s$ paarweise voneinander verschieden sind, ist die Matrix auf der linken Seite regulär, so daß daraus $r_{1k} = \ldots = r_{sk} = 0$ folgt. Für $\nu = 1, \ldots, s$ ist nun

$$\begin{aligned} \sum_{j=1}^{s} r_{jk} c_j^{\nu-1} &= \sum_{j=1}^{s} \sum_{i=1}^{s} \beta_i \alpha_{ij} c_i^{k-1} c_j^{\nu-1} - \sum_{j=1}^{s} \frac{1}{k} \beta_j \left(1 - c_j^k\right) c_j^{\nu-1} \\ &= \sum_{i=1}^{s} \beta_i c_i^{k-1} \underbrace{\sum_{j=1}^{s} \alpha_{ij} c_j^{\nu-1}}_{C(s):\ \frac{1}{\nu} c_i^\nu} - \frac{1}{k} \underbrace{\sum_{j=1}^{s} \beta_j c_j^{\nu-1}}_{B(s+m):\ \frac{1}{\nu}} + \frac{1}{k} \underbrace{\sum_{j=1}^{s} \beta_j c_j^{k+\nu-1}}_{B(s+m):\ \frac{1}{k+\nu}} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\nu}\underbrace{\sum_{i=1}^{s}\beta_i c_i^{\nu+k-1}}_{B(s+m):\ \frac{1}{k+\nu}} - \frac{1}{k}\cdot\frac{1}{\nu} + \frac{1}{k}\cdot\frac{1}{k+\nu} \\
&= \frac{1}{\nu}\cdot\frac{1}{k+\nu} - \frac{1}{k}\cdot\frac{1}{\nu} + \frac{1}{k}\cdot\frac{1}{k+\nu} \\
&= \frac{k-(k+\nu)+\nu}{\nu k\,(k+\nu)} \\
&= 0
\end{aligned}$$

2. analog □

Folgerung 2.11.

1. $B(2s)$, $C(s) \Rightarrow B(2s)$, $C(s)$, $D(s)$. Nach Satz 2.9 folgt dann $p = 2s$.
2. $B(2s-1)$, $C(s) \Rightarrow B(2s-1)$, $C(s)$, $D(s-1)$. Nach Satz 2.9 folgt dann $p = 2s-1$.

2.1.1 Aufbau von Runge-Kutta-Verfahren

Wir benötigen folgende „Hilfsmatrizen“:

$$\begin{aligned}
V_s &:= \begin{pmatrix} 1 & c_1 & \cdots & c_1^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \cdots & c_s^{s-1} \end{pmatrix} \in L(\mathbb{R}^s) \\
e_H &:= \begin{pmatrix} 1 \\ 1/2 \\ \vdots \\ 1/s \end{pmatrix} \in \mathbb{R}^s \\
C_s &:= \begin{pmatrix} c_1 & \frac{c_1^2}{2} & \cdots & \frac{c_1^s}{s} \\ \vdots & \vdots & & \vdots \\ c_s & \frac{c_s^2}{2} & \cdots & \frac{c_s^s}{s} \end{pmatrix} \in L(\mathbb{R}^s) \\
B &:= \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_s \end{pmatrix} \in L(\mathbb{R}^s) \\
N_s &:= \begin{pmatrix} 1 & 1/2 & \cdots & 1/s \\ \vdots & \vdots & & \vdots \\ 1 & 1/2 & \cdots & 1/s \end{pmatrix} \in L(\mathbb{R}^s)
\end{aligned}$$

Dann lassen sich die vereinfachten Konsistenzbedingungen in folgender Form ausdrücken:

$$\boxed{\begin{aligned}
B(s) &\Leftrightarrow \beta V_s = e_H^T \\
C(s) &\Leftrightarrow \mathcal{A} V_s = C_s \\
D(s) &\Leftrightarrow V_s^T B \mathcal{A} = (N_s - C_s)\,B
\end{aligned}}$$

Ein Runge-Kutta-Verfahren kann man dann auf folgende Weise konstruieren:

1. Legen $0 \le c_1 < c_2 < \ldots < c_s \le 1$ fest. Dann ist V_s regulär.
2. garantieren $B(s)$ durch $\beta := e_H^T V_s^{-1}$.
3. Entweder $\mathcal{A} := C_s V_s^{-1}$, d.h. $C(s)$ erfüllt,
   oder, falls $\beta_i \neq 0$, $i = 1, \ldots, s$, $\mathcal{A} := \left(V_s^T\right)^{-1} B^{-1} \left(N_s - C_s^T\right) B$, d.h. $D(s)$ erfüllt, oder ...

Die wichtigsten Verfahrensklassen werden im folgenden vorgestellt.

2.1.1.1 Gauß-Methoden

Die Parameter $c_1, \ldots, c_s$ werden hier als Nullstellen der orthogonalen Polynome gewählt. Dadurch erhält man die Bedingung $B(2s)$. Dann wird $\beta := e_H^T V_s^{-1}$ gesetzt. Es ist $0 < c_1 < \ldots < c_s < 1$ und $\beta_i > 0$, $i = 1, \ldots, s$. Durch $\mathcal{A} := C_s V_s^{-1}$ wird zusätzlich die Bedingung $C(s)$ erfüllt.

$$\begin{array}{c|c} c & C_s V_s^{-1} \\ \hline & e_H^T V_s^{-1} \end{array}$$

Wie in Folgerung 2.11.1 folgt dann $p = 2s$.

Beispiel (implizite Mittelpunktregel). Dieses Verfahren ist einstufig, $s = 1$, und hat die Konsistenzordnung $p = 2$.

1. $c_1 = \frac{1}{2}$ liefert $C_s = \frac{1}{2}$ und $V_s = 1$.
2. Zur Erfüllung von $B(2s)$ folgt $\beta = 1$.
3. Zur Erfüllung von $C(s)$ folgt $\mathcal{A} = \frac{1}{2}$.

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

Beispiel. Als zweistufiges Verfahren erhält man

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

Wie wählt man nun die Knoten $c_1, \ldots, c_s$ „optimal", so daß wir die Bedingung $B(2s)$ erreichen können?

Dazu erinnern wir uns an die Interpolationsquadratur aus der Numerik I:

Für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ wollen wir das Integral $\int_{t_{n-1}}^{t_n} f(t)\,dt$ approximieren. Dazu interpolieren wir die Funktion an den Stützstellen $t_{ni} = t_{n-1} + c_i h_n$, $i = 1, \ldots, s$, mittels des Interpolationspolynoms von Lagrange:

$$L(t) = \sum_{i=1}^{s} L_i(t)\, f(t_{ni}) \qquad \text{mit} \qquad L_i(t) = \prod_{\substack{j=1 \\ j\neq i}}^{s} \frac{t - t_{nj}}{t_{ni} - t_{nj}}.$$

Es ist offensichtlich $L_i(t_{nj}) = \delta_{i,j} = \begin{cases} 1 & i = j, \\ 0 & i \neq j, \end{cases}$ und $\mathrm{Grad} L_i = s-1$. Also ist $\mathrm{Grad} L = s-1$.

Die Quadraturformel

$$\int\limits_{t_{n-1}}^{t_n} f(t)\,dt \approx h_n \sum_{i=1}^{s} \beta_i f(t_{ni}) \qquad \beta_i := \frac{1}{h_n} \int\limits_{t_{n-1}}^{t_n} L_i(t)\,dt, \quad i = 1,\dots,s, \tag{2.4}$$

ist genau, falls f ein Polynom mit $\mathrm{Grad} f \leq s-1$ ist, da dann $L \equiv f$ gilt.

Sei

$$\omega(t) = \prod_{i=1}^{s} (t - t_{ni}) = (t - t_{n1}) \cdots (t - t_{ns})$$

das Stützstellenpolynom mit $\mathrm{Grad}\omega = s$.

Satz 2.12. *Die Quadraturformel* (2.4) *ist genau für alle Polynome mit* $\mathrm{Grad} \leq 2s-1$, *gdw.*

$$\int\limits_{t_{n-1}}^{t_n} \omega(t)\, q(t)\,dt = 0$$

für alle Polynome q mit $\mathrm{Grad} q \leq s-1$.

Beweis. ($\Rightarrow$) Sei die Quadraturformel (2.4) genau für Polynome mit Grad $\leq 2s-1$. Sei q ein Polynom mit $\mathrm{Grad} q \leq s-1$. Es ist $\int_{t_{n-1}}^{t_n} \omega(t)\, q(t) = 0$ zu zeigen.

Wegen $\mathrm{Grad}(\omega q) \leq 2s-1$ ist die Formel (2.4) genau für ωq, so daß folgt

$$\int\limits_{t_{n-1}}^{t_n} \omega(t)\, q(t)\,dt = h_n \sum_{i=1}^{s} \beta_i \underbrace{\omega(t_{ni})}_{0} q(t_{ni}) = 0$$

($\Leftarrow$) Sei $\int_{t_{n-1}}^{t_n} \omega(t)\, q(t) = 0$ für alle Polynome q mit $\mathrm{Grad} q \leq s-1$.

Sei f ein Polynom mit $\mathrm{Grad} f \leq 2s-1$. Dann ist $\int_{t_{n-1}}^{t_n} f(t)\,dt = h_n \sum_{i=1}^{s} \beta_i f(t_{ni})$ zu zeigen.

Dazu zerlegen wir f in der Form

$$f = \omega q + r,$$

so daß $\mathrm{Grad} q \leq s-1$ und $\mathrm{Grad} r \leq s-1$. Dann ist

$$f(t_{ni}) = \underbrace{\omega(t_{ni})}_{=0} q(t_{ni}) + r(t_{ni}) = r(t_{ni}), \qquad i = 1,\dots,s,$$

und

$$\int\limits_{t_{n-1}}^{t_n} f(t)\,dt = \underbrace{\int\limits_{t_{n-1}}^{t_n} \omega(t)\, q(t)\,dt}_{=0 \text{ nach Vor.}} + \int\limits_{t_{n-1}}^{t_n} r(t)\,dt$$

Da aber $\mathrm{Grad} r \leq s-1$ ist und solche Polynome exakt integriert werden, folgt

$$\int\limits_{t_{n-1}}^{t_n} f(t)\,dt = \int\limits_{t_{n-1}}^{t_n} r(t)\,dt = h_n \sum_{i=1}^{s} \beta_i r(t_{ni}) = h_n \sum_{i=1}^{s} \beta_i f(t_{ni}),$$

also ist die Formel (2.4) genau für alle Polynome mit $\mathrm{Grad} \leq 2s-1$. $\square$

Man wählt nun die Stützstellen $c_1, \ldots, c_s$ so, daß für das Stützstellenpolynom $\tilde{\omega}(\tau) := \prod_{i=1}^{s}(\tau - c_i)$ die Bedingung

$$\int_0^1 \tilde{\omega}(\tau)\, \tilde{q}(\tau)\, d\tau = 0$$

für alle Polynome $\tilde{q}$ mit Grad$\tilde{q} \leq s-1$ erfüllt ist (Ansatz $\tilde{\omega}(\tau) = \tau^s + \ldots$), d.h. wir erhalten eine Quadraturformel für das Intervall $[0,1]$, welche genau ist für alle Polynome mit Grad $\leq 2s-1$.

Für $[t_{n-1}, t_n]$ folgt dann mit $\omega(t) = \prod_{i=1}^{s}(t - t_{ni}) = \prod_{i=1}^{s}(t - (t_{n-1} + c_i h_n))$, daß

$$\begin{aligned}
\int_{t_{n-1}}^{t_n} \omega(t)\, q(t)\, dt &= h_n \int_0^1 \omega(t_{n-1} + \tau h_n)\, q(t_{n-1} + \tau h_n)\, d\tau \\
&= h_n \int_0^1 \prod_{i=1}^{s} (\tau h_n - c_i h_n)\, q(t_{n-1} + \tau h_n)\, d\tau \\
&= h_n \int_0^1 h_n^s \underbrace{\prod_{i=1}^{s} (\tau - c_i)}_{=\tilde{\omega}(\tau)} \underbrace{q(t_{n-1} + \tau h_n)}_{=:\tilde{q}(\tau)}\, d\tau \\
&= h_n^{s+1} \int_0^1 \tilde{\omega}(\tau)\, \tilde{q}(\tau)\, d\tau \\
&= 0
\end{aligned}$$

ist für alle Polynome q mit Grad$q \leq s-1$. Wegen Satz 2.12 ist die Quadraturformel (2.4) also genau für alle Polynome f mit Grad$f \leq 2s-1$.

Folgt man nun der Argumentation vom Anfang dieses Abschnittes, so erkennt man, daß für diese Wahl der $c_1, \ldots, c_s$ die Bedingung $B(2s)$ erfüllt ist.

Noch zu verifizieren ist, daß die Knoten c_i die Bedingung $c_i \in (0,1)$, $c_i \neq c_j$, $i \neq j$, erfüllen und $\beta_i > 0$, $i = 1, \ldots, s$, gilt.

Lemma 2.13. *Sei p ein Polynom mit Grad s. Es gelte $\int_0^1 p(t)\, q(t)\, dt = 0$ für alle Polynome q mit* Grad$q \leq s-1$. *Dann hat p genau s paarweise verschiedene Nullstellen aus* $(0,1)$.

Beweis. Bezeichne mit t_j, $j = 1, \ldots, m$, $m \leq s$, die Nullstellen von p ohne Wiederholung, und seien ν_j deren Vielfachheiten. Dann ist

$$p(t) = c \prod_{j=1}^{m} (t - t_j)^{\nu_j}$$

für eine Konstante c. Sei

$$M := \{\xi_1, \ldots, \xi_l\} := \{t_j \in (0,1) \,|\, \nu_j \text{ ist ungerade}\}, \qquad l \leq s,$$

und

$$r(t) := \begin{cases} \prod_{j=1}^{l} (t - \xi_j) & \text{falls } M \neq \emptyset, \\ 1 & \text{sonst.} \end{cases}$$

Dann haben die Nullstellen des Polynoms $p \cdot r$ auf $(0,1)$ nur gerade Vielfachheit. Also kann $p \cdot r$ nicht das Vorzeichen auf $(0,1)$ wechseln, so daß

$$\int_0^1 p(t)\, r(t)\, dt \neq 0$$

gilt. Nach Voraussetzung ist aber $\int_0^1 p(t)\,q(t)\,dt = 0$ für alle Polynome q mit $\mathrm{Grad}q \leq s-1$. Also muß

$$\mathrm{Grad}r \geq s$$

gelten. Wegen

$$\mathrm{Grad}r \leq l \leq s$$

folgt also $l = s$, d.h. es gibt s paarweise verschiedene Nullstellen auf $(0,1)$. Da es aber nur genau s Nullstellen gibt, kann jede Nullstelle nur die Vielfachheit 1 besitzen. □

Lemma 2.14. *Ist die Quadraturformel*

$$\int_{t_{n-1}}^{t_n} f(t)\,dt \approx h_n \sum_{i=1}^{s} \beta_i f(t_{ni})$$

genau für alle Polynome f mit $\mathrm{Grad}f \leq 2s-2$, *so gilt* $\beta_i > 0$, $i = 1, \ldots, s$.

Beweis. Für $i = 1, \ldots, s$ hat das Polynom

$$L_i^2(t) = \left(\prod_{\substack{j=1\\ j\neq i}}^{s} \frac{t - t_{nj}}{t_{ni} - t_{nj}} \right)^2$$

Grad $2s-2$. Dann ist

$$0 < \int_{t_{n-1}}^{t_n} L_i^2(t)\,dt = h_n \sum_{j=1}^{s} \beta_j \underbrace{L_i^2(t_{nj})}_{=\delta_{ij}} = h_n \beta_i,$$

also ist $\beta_i > 0$. □

2.1.1.2 Radau-Methoden

Bei den Radau-Methoden wird entweder $c_1 = 0$ oder $c_s = 1$ fixiert. Die restlichen Parameter $c_2, \ldots, c_s$ bzw. $c_1, \ldots, c_{s-1}$ werden dann optimal gewählt: Bei $c_1 = 0$ erhält das Stützstellenpolynom ω die Form $\omega(t) = (t - t_{n-1})\,\bar{\omega}(t)$. Wir können nun die $c_2, \ldots, c_s$ so wählen, daß $\int_{t_{n-1}}^{t_n} \omega(t)\,q(t) = 0$ ist für alle Polynome q mit $\mathrm{Grad}q \leq s-2$.

- Radau I (wenig gebräuchlich): $c_1 = 0$; $c_2, \ldots, c_s$ „optimal". Dann ist die Bedingung $B(2s-1)$ erfüllt.
- Radau II: $c_s = 1$; $c_1, \ldots, c_{s-1}$ „optimal". Dann ist die Bedingung $B(2s-1)$ erfüllt.
  Wir setzen wieder $\beta = e_H^T V_s^{-1}$.
  Radau IIA: Setze $\mathcal{A} = C_s V_s^{-1}$, so daß $C(s)$ erfüllt ist.
  Folgerung 2.11 liefert nun die Konsistenzordnung $p = 2s-1$.
  Wegen $c_s = 1$ ist

$$C_s = \begin{pmatrix} c_1 & \frac{c_1^2}{2} & \cdots & \frac{c_1^s}{s} \\ \vdots & \vdots & & \vdots \\ c_{s-1} & \frac{c_{s-1}^2}{2} & \cdots & \frac{c_{s-1}^s}{s} \\ 1 & 1/2 & \cdots & 1/s \end{pmatrix} = \begin{pmatrix} c_1 & \cdots & \frac{c_1^s}{s} \\ \vdots & & \vdots \\ c_{s-1} & \cdots & \frac{c_{s-1}^s}{s} \\ & e_H^T & \end{pmatrix},$$

  also folgt $\alpha_{si} = \beta_i$, $i = 1, \ldots, s$, d.h. $x_n = X_{ns}$ und $t_n = t_{ns}$.

Beispiel (implizites Eulerverfahren). Dieses Verfahren ist einstufig, $s = 1$. $c_s = 1$ liefert dann $\beta = 1$ und $\mathcal{A} = 1$: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ Dieses Verfahren besitzt dann eine Konsistenzordnung von $p = 1$.

Beispiel. Als zweistufiges Verfahren erhält man

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}$$

Dieses Verfahren besitzt eine Konsistenzordnung von $p = 3$.

2.1.1.3 Lobatto III-Methoden

Hier werden $c_1 = 0$ und $c_s = 1$ fixiert und die restlichen Parameter $c_2, \ldots, c_{s-1}$ „optimal" gewählt, so daß die Bedingung $B(2s-2)$ erfüllt wird. Es wird wieder $\beta = e_H^T V_s^{-1}$ gesetzt.

- Lobatto IIIA: Setze $\mathcal{A}$, so daß $C(s)$ erfüllt ist.
- Lobatto IIIB: Setze $\mathcal{A}$, so daß $D(s)$ erfüllt ist.
- Lobatto IIIC: Setze $\mathcal{A}$, so daß $C(s-1)$ erfüllt ist und setze $\alpha_{i1} = \beta_1$, $i = 1, \ldots, s$.

Mit Lemma 2.10 und Satz 2.9 folgt dann aus der Gültigkeit von $B(2s-2)$ und $C(s)$ bzw. $C(s-1)$ bzw. $D(s)$ die Konsistenzordnung $p = 2s - 2$.

Beispiel (Trapezregel). Dieses Verfahren ist zweistufig, $s = 2$. Es ist also $c_1 = 0$ und $c_2 = 1$. Die entsprechende Lobatto IIIA Methode hat dann die Form

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

und erreicht die Konsistenzordnung $p = 2$.

Beispiel. Die Lobatto IIIC Methode für $s = 2$ hat hingegen die Form

$$\begin{array}{c|cc} 0 & 1/2 & -1/2 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

und erreicht ebenfalls die Konsistenzordnung $p = 2$.

2.1.2 Algebraische Stabilität

Satz 2.15. *Die Gauß-, Radau IIA- und Lobatto IIIC-Methoden sind algebraisch stabil.*

Beweis. Nach Definition 1.17 (Seite 16), ist ein Runge-Kutta-Verfahren algebraisch stabil, falls $M = (m_{i,j})$,

$$m_{i,j} := \beta_i \alpha_{i,j} + \beta_j \alpha_{ji} - \beta_i \beta_j, \qquad i, j = 1, \ldots, s,$$

positiv semidefinit ist und $\beta_i \geq 0$, $i = 1, \ldots, s$, gilt.

Die Bedingung $\beta_i \geq 0$, $i = 1, \ldots, s$, folgt aus der Wahl der $c_1, \ldots, c_s$ und Lemma 2.14. Sei

$$Q := V_s^T M V_s.$$

Dann ist $\langle Qz, z\rangle = \left\langle V_s^T M V_s z, z\right\rangle = \langle M V_s z, V_s z\rangle$, $z \in \mathbb{R}^s$. Da V_s regulär ist, ist dann Q positiv semidefinit, gdw. M positiv semidefinit ist.

$$\begin{aligned}
Q &= \begin{pmatrix} 1 & \cdots & 1 \\ c_1 & \cdots & c_s \\ \vdots & & \vdots \\ c_1^{s-1} & \cdots & c_s^{s-1} \end{pmatrix} \begin{pmatrix} m_{11} & \cdots & m_{1s} \\ \vdots & \ddots & \vdots \\ m_{s1} & \cdots & m_{ss} \end{pmatrix} \begin{pmatrix} 1 & c_1 & \cdots & c_1^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \cdots & c_s^{s-1} \end{pmatrix} \\
&= \begin{pmatrix} \sum\limits_{j=1}^{s} m_{j1} & \cdots & \sum\limits_{j=1}^{s} m_{js} \\ \vdots & & \vdots \\ \sum\limits_{j=1}^{s} c_j^{s-1} m_{j1} & \cdots & \sum\limits_{j=1}^{s} c_j^{s-1} m_{js} \end{pmatrix} \begin{pmatrix} 1 & c_1 & \cdots & c_1^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \cdots & c_s^{s-1} \end{pmatrix} \\
q_{l,k} &= \sum_{i=1}^{s} \sum_{j=1}^{s} c_j^{l-1} m_{ji} c_i^{k-1} \\
&= \sum_{i,j=1}^{s} \left(\beta_i \alpha_{ij} + \beta_j \alpha_{ji} - \beta_i \beta_j\right) c_j^{l-1} c_i^{k-1}
\end{aligned}$$

1. Gauß-Methoden: Es sind $B\,(2s)$ und $C\,(s)$ erfüllt, d.h.

$$\begin{aligned}
\sum_{j=1}^{s} \beta_j c_j^{k-1} &= \frac{1}{k}, \qquad k = 1, \ldots, 2s \\
\sum_{j=1}^{s} \alpha_{i,j} c_j^{k-1} &= \frac{1}{k} c_i^k, \qquad k = 1, \ldots, s, i = 1, \ldots, s
\end{aligned}$$

   Dann ist

$$\begin{aligned}
\sum_{i,j=1}^{s} \beta_i \beta_j c_j^{l-1} c_i^{k-1} &= \underbrace{\sum_{i=1}^{s} \beta_i c_i^{k-1}}_{1/k} \underbrace{\sum_{j=1}^{s} \beta_j c_j^{l-1}}_{1/l} = \frac{1}{kl} \\
\sum_{i,j=1}^{s} \beta_i \alpha_{i,j} c_j^{l-1} c_i^{k-1} &= \sum_{i=1}^{s} \beta_i c_i^{k-1} \underbrace{\sum_{j=1}^{s} \alpha_{i,j} c_j^{l-1}}_{\frac{1}{l} c_i^l} = \frac{1}{l} \sum_{i=1}^{s} \beta_i c_i^{l+k-1} = \frac{1}{l} \cdot \frac{1}{l+k}.
\end{aligned}$$

   Also folgt

$$q_{k,l} = \frac{1}{l} \cdot \frac{1}{l+k} + \frac{1}{k} \cdot \frac{1}{l+k} - \frac{1}{kl} = 0, \qquad k, l = 1, \ldots, s,$$

   d.h. $Q = 0$ ist positiv semidefinit. Dann ist auch M positiv semidefinit.

2. Radau IIA: Es sind $B\,(2s-1)$ und $C\,(s)$ erfüllt.

   Analog wie für Gauß-Verfahren ist $q_{lk} = 0$ für $l + k < 2s$. Es ist also

$$Q = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & q_{ss} \end{pmatrix}.$$

Q ist positiv semidefinit, wenn $q_{ss} \geq 0$ gilt.

$$
\begin{aligned}
q_{ss} &= \sum_{i,j=1}^{s} \left(\beta_i \alpha_{i,j} + \beta_j \alpha_{j,i} - \beta_i \beta_j\right) c_j^{s-1} c_i^{s-1} \\
&= 2 \sum_{i=1}^{s} \beta_i c_i^{s-1} \underbrace{\sum_{j=1}^{s} \alpha_{i,j} c_j^{s-1}}_{C(s):\ \frac{1}{s} c_i^s} - \underbrace{\sum_{i=1}^{s} \beta_i c_i^{s-1}}_{B(2s-1):\ 1/s} \underbrace{\sum_{j=1}^{s} \beta_j c_j^{s-1}}_{B(2s-1):\ 1/s} \\
&= \frac{2}{s} \left(\sum_{i=1}^{s} \beta_i c_i^{2s-1} - \frac{1}{2s} \right).
\end{aligned}
$$

Betrachte nun das Polynom

$$P(t) = (t - t_{n1})^2 \cdots (t - t_{n,s-1})^2 (t - t_n)$$

mit $\mathrm{Grad} P = 2s - 1$ und Stützstellen $t_{ni} = t_{n-1} + c_i h_n$, $c_s = 1$.
Dann ist $P(t_{ni}) = 0$, $i = 1 \ldots, s$, und

$$\int_{t_{n-1}}^{t_n} P(t)\, dt < 0.$$

Wir stellen P in der Form

$$P(t) = (t - t_{n-1})^{2s-1} + P_* (t - t_{n-1})$$

mit einem Polynom P_* mit $\mathrm{Grad} P_* \leq 2s - 2$ dar. Dann ist

$$0 = P(t_{ni}) = (c_i h_n)^{2s-1} + P_* (c_i h_n), \qquad \text{d.h. } P_* (c_i h_n) = -(c_i h_n)^{2s-1}.$$

Wegen $\mathrm{Grad} P_* \leq 2s - 2$ ist die Quadraturformel (2.4) genau, d.h. es ist

$$
\begin{aligned}
\int_{t_{n-1}}^{t_n} P_* (t - t_{n-1})\, dt &= h_n \sum_{i=1}^{s} \beta_i P_* (t_{ni} - t_{n-1}) \\
&= h_n \sum_{i=1}^{s} \beta_i P_* (c_i h_n) \\
&= -h_n \sum_{i=1}^{s} \beta_i (c_i h_n)^{2s-1}
\end{aligned}
$$

also

$$
\begin{aligned}
0 > \int_{t_{n-1}}^{t_n} P(t)\, dt &= \int_{t_{n-1}}^{t_n} (t - t_{n-1})^{2s-1}\, dt + \int_{t_{n-1}}^{t_n} P_* (t - t_{n-1})\, dt \\
&= \frac{1}{2s} h_n^{2s} - h_n \sum_{i=1}^{s} \beta_i (c_i h_n)^{2s-1} \\
&= h_n^{2s} \left(\frac{1}{2s} - \sum_{i=1}^{s} \beta_i c_i^{2s-1} \right) \\
&= -\frac{s}{2} q_{ss} h_n^{2s}.
\end{aligned}
$$

Damit folgt $q_{ss} > 0$.

3. Lobatto IIIC: Siehe [SW95, S. 6.2.11] □

Satz 2.16. *Kein konsistentes explizites Runge-Kutta-Verfahren ist algebraisch stabil.*

Beweis. Für ein explizites Verfahren ist $\mathcal{A} = \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ * & & 0 \end{pmatrix}$, also

$$\begin{aligned} M &= \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ * & & 0 \end{pmatrix} \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_s \end{pmatrix} + \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_s \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ * & & 0 \end{pmatrix} - \beta^T \beta \\ &= \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ * & & 0 \end{pmatrix} + \begin{pmatrix} 0 & & * \\ \vdots & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} - \begin{pmatrix} \beta_1^2 & & * \\ & \ddots & \\ * & & \beta_s^2 \end{pmatrix}. \end{aligned}$$

Dann ist

$$\langle M e_i, e_i \rangle = -\beta_i^2, \qquad i = 1, \ldots, s.$$

Die Konsistenzbedingung $B(1)$ liefert $\sum_{i=1}^{s} \beta_i = 1$, also existiert ein Index i_* mit $\beta_{i_*} \neq 0$. Dann ist

$$\langle M e_{i_*}, e_{i_*} \rangle = -\beta_{i_*}^2 < 0,$$

also kann M nicht positiv semidefinit sein. □

2.2 Stabilitätsfunktion

Wenn man über ein großes Zeitintervall $[0, T]$ integriert, so ist es wichtig, daß auftretende Approximationsfehler nicht verstärkt werden, daß Integrationsverfahren also stabil bezüglich diesen Fehlern ist. Wir werden feststellen, daß einige Methoden mit relativ kleiner Konsistenzordnung sich besser verhalten als andere Verfahren, die zwar eine höhere Konsistenzordnung besitzen, aber keine Dämpfung von Approximationsfehlern erlauben. Dazu wenden wir das s-stufige Runge-Kutta-Verfahren mit Butcher-Tableau $\begin{array}{c|c} c & \mathcal{A} \\ \hline & \beta \end{array}$ auf die skalare Testgleichung

$$\boxed{x'(t) = \lambda x(t), \qquad \lambda \in \mathbb{C},}$$

an. Die Runge-Kutta-Approximation für $h_n := h$ liefert

$$\begin{aligned} x_n &= x_{n-1} + h_n \sum_{j=1}^{s} \beta_j X'_{nj}, \\ X'_{ni} = f(X_{ni}) &= \lambda X_{ni} = \lambda \left(x_{n-1} + h \sum_{j=1}^{s} \alpha_{i,j} X'_{nj} \right), \qquad i = 1, \ldots, s, \end{aligned}$$

also

$$\begin{aligned} x_n &= x_{n-1} + h\lambda \sum_{j=1}^{s} \beta_j X_{nj}, \\ X_{ni} &= x_{n-1} + h\lambda \sum_{j=1}^{s} \alpha_{i,j} X_{nj}, \qquad i = 1, \ldots, s, \end{aligned}$$

d.h.

$$X_n := \begin{pmatrix} X_{n1} \\ \vdots \\ X_{ns} \end{pmatrix} = \begin{pmatrix} x_{n-1} \\ \vdots \\ x_{n-1} \end{pmatrix} + h\lambda\mathcal{A} \begin{pmatrix} X_{n1} \\ \vdots \\ X_{ns} \end{pmatrix}, \qquad X_n \in \mathbb{C}^s,$$

bzw.

$$(I_s - h\lambda\mathcal{A})\, X_n = x_{n-1}\mathbf{1}_s \qquad \text{mit } \mathbf{1}_s := \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T \in \mathbb{R}^s.$$

Falls $(I_s - h\lambda\mathcal{A})$ regulär ist, so folgt

$$X_n = (I_s - h\lambda\mathcal{A})^{-1}\, x_{n-1}\mathbf{1}_s$$

und

$$x_n = x_{n-1} + h\lambda\beta X_n = x_{n-1} + h\lambda\beta\, (I_s - h\lambda\mathcal{A})^{-1}\, x_{n-1}\mathbf{1}_s,$$

d.h. die Runge-Kutta-Iterierten berechnen sich durch die Vorschrift

$$\boxed{x_n = \left\{1 + h\lambda\beta\, (I_s - h\lambda\mathcal{A})^{-1}\, \mathbf{1}_s\right\} x_{n-1}.}$$

Für die exakte Lösung ist

$$x\,(t_n) = e^{h\lambda} x\,(t_{n-1})\,.$$

Definition 2.17 (Stabilitätsfunktion). Die komplexe Funktion $R : \mathbb{C} \to \mathbb{C}$,

$$\boxed{R\,(z) := 1 + z\beta\, (I_s - z\mathcal{A})^{-1}\, \mathbf{1}_s,}$$

heißt *Stabilitätsfunktion* des Runge-Kutta-Verfahrens.

Satz 2.18. *Die Stabilitätsfunktion eines s-stufigen Runge-Kutta-Verfahrens ist eine rationale Funktion, deren Zähler und Nenner einen Grad $\leq s$ haben.*

Beweis. Wie oben gezeigt ist

$$\begin{aligned} (I_s - z\mathcal{A})\, X_n &= \mathbf{1}_s x_{n-1} \\ x_n - z\beta X_n &= x_{n-1}, \end{aligned}$$

d.h.

$$\begin{pmatrix} I_s - z\mathcal{A} & 0 \\ -z\beta & 1 \end{pmatrix} \begin{pmatrix} X_n \\ x_n \end{pmatrix} = \begin{pmatrix} \mathbf{1}_s \\ 1 \end{pmatrix} x_{n-1}.$$

Nach der Cramerschen Regel gilt

$$x_n = \frac{\det \begin{pmatrix} I_s - z\mathcal{A} & \mathbf{1}_s x_{n-1} \\ -z\beta & x_{n-1} \end{pmatrix}}{\det\,(I_s - z\mathcal{A})} = \frac{\det \begin{pmatrix} I_s - z\mathcal{A} & \mathbf{1}_s \\ -z\beta & 1 \end{pmatrix}}{\det\,(I_s - z\mathcal{A})} \cdot x_{n-1}.$$

Wegen $x_n = R\,(z) \cdot x_{n-1}$ folgt dann

$$R\,(z) = \frac{\det \begin{pmatrix} I_s - z\mathcal{A} & \mathbf{1}_s \\ -z\beta & 1 \end{pmatrix}}{\det\,(I_s - z\mathcal{A})}.$$

□

Folgerung 2.19. Die Stabilitätsfunktion eines expliziten Runge-Kutta-Verfahrens ist ein Polynom.

Beweis. Aus $\mathcal{A} = \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ * & & 0 \end{pmatrix}$ folgt $\det\,(I_s - z\mathcal{A}) \equiv 1$, also $R\,(z) = \det \begin{pmatrix} I_s - z\mathcal{A} & \mathbf{1}_s \\ -z\beta & 1 \end{pmatrix}$. □

Beispiel. 1. Explizites Euler-Verfahren: $s = 1$, $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$. Die Stabilitätsfunktion ist

$$R(z) = 1 + z.$$

2. Implizites Euler-Verfahren: $s = 1$, $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$. Die Stabilitätsfunktion ist

$$R(z) = 1 + z(1 - z)^{-1} = \frac{1}{1 - z}.$$

3. Implizite Mittelpunktregel: $s = 1$, $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$. Die Stabilitätsfunktion ist

$$R(z) = 1 + z\left(1 - \frac{1}{2}z\right)^{-1} = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}.$$

Für kleine z ist $R(z)$ eine gute Approximation von e^z aufgrund der Konsistenzbedingungen: Für $z = 0$ ist $R(0) = 1$. Für kleine z ist

$$e^z = 1 + z + \frac{1}{2}z^2 + \ldots$$

Für die numerische Approximation ist

$$(I_s - z\mathcal{A})^{-1} = I_s + z\mathcal{A} + z^2\mathcal{A}^2 + \ldots,$$

$$\begin{aligned} R(z) &= 1 + z\beta\left(I_s + z\mathcal{A} + z^2\mathcal{A}^2 + \ldots\right)\mathbf{1}_s \\ &= 1 + z\sum_{i=1}^{s}\beta_i + z^2\beta\mathcal{A}\mathbf{1}_s + \ldots \end{aligned}$$

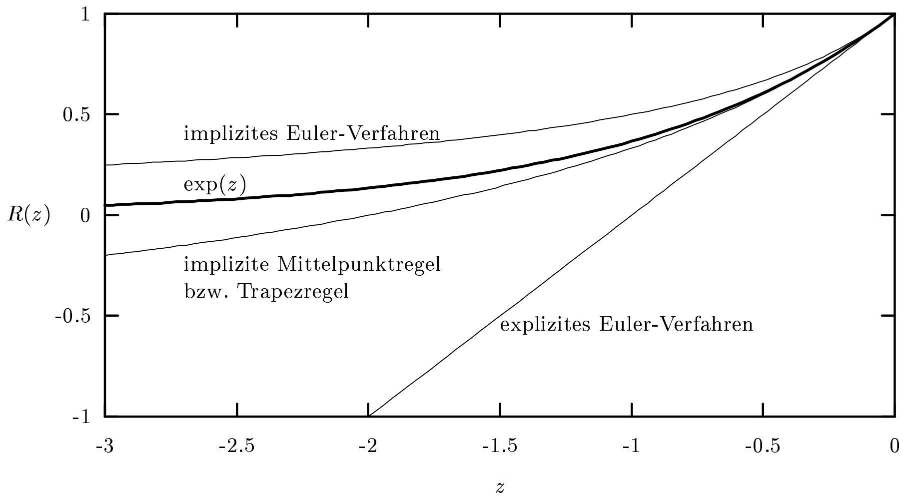
wobei $\sum_{i=1}^{s}\beta_i = 1$ wegen $B(1)$ und

$$\beta\mathcal{A}\mathbf{1}_s = \beta\begin{pmatrix} \sum_{i=1}^{s}\alpha_{1i} \\ \vdots \\ \sum_{i=1}^{s}\alpha_{si} \end{pmatrix} \overset{C(1)}{=} \begin{pmatrix} \beta_1 & \cdots & \beta_s \end{pmatrix}\begin{pmatrix} c_1 \\ \vdots \\ c_s \end{pmatrix} = \sum_{i=1}^{s}\beta_i c_i \overset{B(2)}{=} \frac{1}{2}.$$

Damit ist

$$R(z) = 1 + z + \frac{1}{2}z^2 + \ldots$$

Stabilitätsfunktion verschiedener Runge-Kutta-Verfahren im Vergleich zur Exponentialfunktion für $z \in \mathbb{R}$



Definition 2.20 (Stabilitätsgebiet). Die Menge

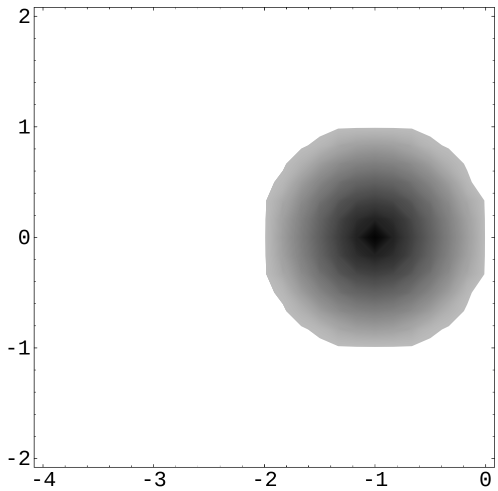
$$\boxed{S := \{z \in \mathbb{C} \,|\, |R(z)| \leq 1\}}$$

heißt *Stabilitätsgebiet* (Bereich der absoluten Stabilität).

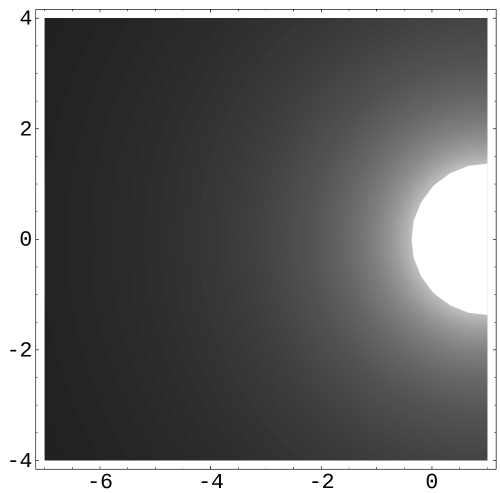
Bemerkung (Mehrschrittverfahren). In der Grundvorlesung wurde für Mehrschrittverfahren $\mathcal{G} := \{z \in \mathbb{C} \,|\, |R(z)| \leq 1\}$ mit einer entsprechenden Stabilitätsfunktion R definiert.

Bemerkung (Aspekt zur Konstruktion von Runge-Kutta-Verfahren). Gute Approximation von e^z durch $R(z)$ für große Bereiche aus $\mathbb{C}^-$ (Padé-Approximation). (Für $\mathrm{Re}\lambda \leq 0$ ist $\left|e^{h\lambda}\right| \leq 1$. Dann ist die Eigenschaft $|R(h\lambda)| \leq 1$ für (fast) beliebige Werte von h wünschenswert, d.h. $S \cap \mathbb{C}^-$ sollte groß sein.)

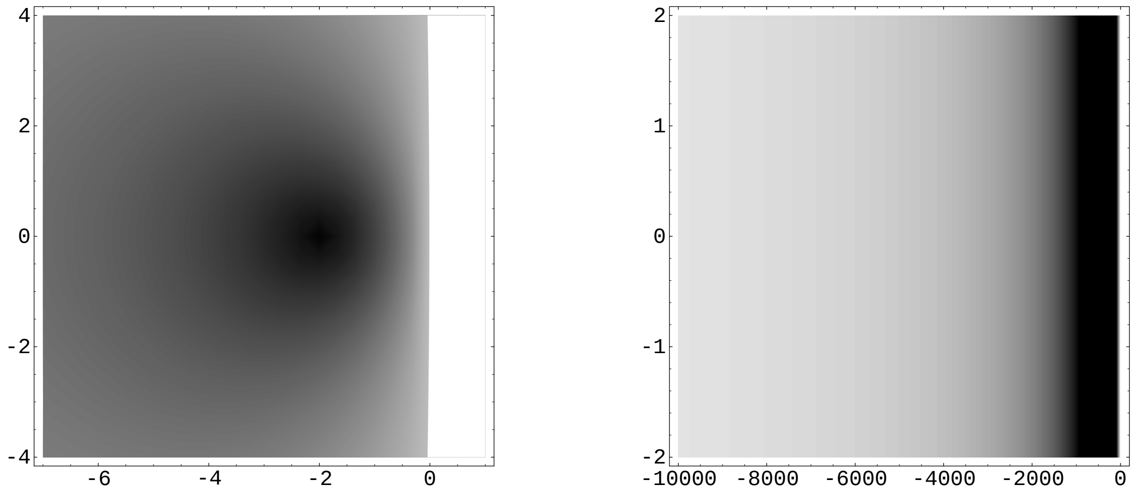
Beispiel. Das Stabilitätsgebiet S für verschiedene Runge-Kutta-Verfahren (Plot der Funktion $\min(1, |R(z)|)$ für $z = a + bi \in [-8, 1] + i[-4, 4]$; 1 ist weiß, Werte < 1 sind grau bis schwarz)



Explizites Eulerverfahren: $R(z) = 1 + z$



Implizites Eulerverfahren: $R(z) = \frac{1}{1-z}$



Trapezregel bzw. implizite Mittelpunktregel (für verschiedene Wertebereiche): $R(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$

Wir verallgemeinern nun die Testgleichung $x'(t) = \lambda x(t)$ auf die lineare homogene DGL

$$\boxed{x'(t) = Bx(t) \qquad \text{mit} \quad B = T^{-1}JT, \quad J = \operatorname{diag}(\sigma_1, \ldots, \sigma_m).}$$

Dazu transformieren wir

$$\tilde{x}(t) := Tx(t).$$

Dann ist für die exakte Lösung

$$\begin{aligned} \tilde{x}'(t) &= J\tilde{x}(t), \\ \text{also } \tilde{x}_k'(t) &= \sigma_k \tilde{x}_k(t), \qquad k = 1, \ldots, m. \end{aligned}$$

Für die numerische Approximation

$$\begin{aligned} x_n &= x_{n-1} + h\sum_{i=1}^{s} \beta_i X_{ni}' \\ X_{ni}' &= B\left(x_{n-1} + h\sum_{j=1}^{s} \alpha_{ij} X_{nj}'\right), \quad i = 1, \ldots, s \end{aligned}$$

folgt mit

$$\tilde{x}_n := Tx_n, \qquad \tilde{X}_{ni} := TX_{ni}, \qquad \tilde{X}_{ni}' := TX_{ni}',$$

daß

$$\begin{aligned} \tilde{x}_n &= \tilde{x}_{n-1} + h\sum_{i=1}^{s} \beta_i \tilde{X}_{ni}' \\ \tilde{X}_{ni}' &= J\left(\tilde{x}_{n-1} + h\sum_{j=1}^{s} \alpha_{ij} \tilde{X}_{nj}'\right), \quad i = 1, \ldots, s, \end{aligned}$$

also

$$\begin{aligned} \tilde{x}_{n,k} &= \tilde{x}_{n-1,k} + h\sum_{i=1}^{s} \beta_i \tilde{X}_{ni,k}' \\ \tilde{X}_{ni,k}' &= \sigma_k\left(\tilde{x}_{n-1,k} + h\sum_{j=1}^{s} \alpha_{i,j} \tilde{X}_{nj,k}'\right), \quad i = 1, \ldots, s. \end{aligned}$$

Mit der oben definierten Stabilitätsfunktion folgt dann also

$$\boxed{\tilde{x}_{n,k} = R(\sigma_k h)\,\tilde{x}_{n-1,k}.}$$

Sei nun

$$\operatorname{Re}\sigma_k < 0, \qquad k = 1, \ldots, m.$$

Dann gilt für alle Lösungen $x(\cdot)$

$$x(t_n) \xrightarrow{n\to\infty} 0.$$

Gilt auch $x_n \to 0$ für die numerische Approximation?
Falls $h\sigma_k \in \operatorname{Int}S$, d.h. $|R(h\sigma_k)| < 1$, $k = 1, \ldots, m$, so folgt

$$|\tilde{x}_{n,k}| = |R(h\sigma_k)|\,|\tilde{x}_{n-1,k}|,$$

also $\tilde{x}_{n,k} \xrightarrow{n\to\infty} 0$, $k = 1, \ldots, m$, d.h. $\tilde{x}_n \xrightarrow{n\to\infty} 0$. Dann folgt auch

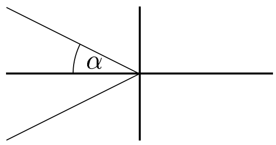
$$x_n = T^{-1}\tilde{x}_n \xrightarrow{n\to\infty} 0.$$

Da die Eigenwerte $\sigma_1, \ldots, \sigma_m$ von B im allgemeinen im Voraus nicht bekannt sind, ist es wichtig, Methoden zu haben, für die der Stabilitätsbereich die gesamte linke komplexe Halbebene umfasst. Dies motiviert die Definition der A-Stabilität. Da dies aber für manche Anwendungen wiederum auch eine zu starke Bedingung ist, schränkt man die linke komplexe Halbebene auch auf einen Keil ein, der komplett in der linken Halbebene liegt und einen Winkel α mit der reellen Achse aufspannt.

Die Definition der L-Stabilität ist motiviert durch die Beobachtung, daß sich die exponentielle Dämpfung für Eigenwerte $\lambda \to -\infty$ immer stärker beschleunigt. L-Stabilität fordert nun von der Integrationsmethode, daß sie dieses Verhalten auch auf die approximierte Lösung überträgt.

Definition 2.21 (stark A-, A-, L-, A_0, $A(\alpha)$-stabil).

1. Ein Runge-Kutta-Verfahren heißt A-*stabil*, falls
$$\boxed{|R(z)| \leq 1 \qquad \text{für alle } z \in \mathbb{C} \text{ mit } \mathrm{Re} z \leq 0.}$$
2. Ein Runge-Kutta-Verfahren heißt L-*stabil*, falls es A-stabil ist und
$$\boxed{\lim_{\mathrm{Re} z \to -\infty} R(z) = 0}$$
3. Ein Runge-Kutta-Verfahren heißt *stark* A-*stabil*, falls es A-stabil ist und
$$\boxed{\lim_{\mathrm{Re} z \to -\infty} |R(z)| < 1}$$
4. Ein Runge-Kutta-Verfahren heißt A_0-*stabil*, falls
$$\boxed{|R(z)| \leq 1 \qquad \text{für alle } z \in \mathbb{R} \text{ mit } z \leq 0.}$$
5. Ein Runge-Kutta-Verfahren heißt für $\alpha \in \left(0, \frac{\pi}{2}\right)$ $A(\alpha)$-*stabil*, falls
$$\boxed{|R(z)| \leq 1 \qquad \forall\, z \in \mathbb{C},\ \mathrm{Re} z \leq 0,\ |\arg(z) - \pi| \leq \alpha.}$$



Beispiel (Trapezregel). Butcher-Tableau:
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

Die Stabilitätsfunktion ist
$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$$

Dieses Verfahren ist zwar A-stabil, da
$$\begin{aligned}
|R(z)| &\leq 1 \\
\Leftrightarrow \left|1 + \frac{1}{2}z\right| &\leq \left|1 - \frac{1}{2}z\right| \\
\Leftrightarrow \mathrm{Re}\left(1 + \frac{1}{2}z\right)^2 + \mathrm{Im}\left(1 + \frac{1}{2}z\right)^2 &\leq \mathrm{Re}\left(1 - \frac{1}{2}z\right)^2 + \mathrm{Im}\left(1 - \frac{1}{2}z\right)^2 \\
\Leftrightarrow \left(1 + \frac{1}{2}\mathrm{Re} z\right)^2 + \left(\frac{1}{2}\mathrm{Im} z\right)^2 &\leq \left(1 - \frac{1}{2}\mathrm{Re} z\right)^2 + \left(\frac{1}{2}\mathrm{Im} z\right)^2 \\
\Leftrightarrow 1 + \mathrm{Re} z + \left(\frac{1}{2}\mathrm{Re} z\right)^2 &\leq 1 - \mathrm{Re} z + \left(\frac{1}{2}\mathrm{Re} z\right)^2 \\
\Leftrightarrow 2\mathrm{Re} z &\leq 0.
\end{aligned}$$

Es ist aber nicht stark A-stabil, da

$$\lim_{\mathrm{Re}z\to-\infty}|R(z)|=1.$$

Siehe dazu auch die Abbildungen auf Seite 41.

Satz 2.22. *Kein konsistentes explizites Runge-Kutta-Verfahren ist A-stabil.*

Für Verfahren mit regulärer Matrix $\mathcal{A}$ ist

$$\begin{aligned} R(z) &= 1+z\beta\left(I_s-z\mathcal{A}\right)^{-1}\mathbf{1}_s \\ &= 1+\beta\left(\frac{1}{z}I_s-\mathcal{A}\right)^{-1}\mathbf{1}_s \\ &= 1-\beta\left(\mathcal{A}-\frac{1}{z}I_s\right)^{-1}\mathbf{1}_s. \end{aligned}$$

Sei dann

$$\boxed{R(\infty):=\lim_{\mathrm{Re}z\to-\infty}R(z)=1-\beta\mathcal{A}^{-1}\mathbf{1}_s.}$$

Satz 2.23. *Radau IIA und Lobatto IIIC-Verfahren sind L-stabil.*

Die Gauß-Methoden sind A-stabil.

Beweis. Zur L-Stabilität der Radau IIA-Methoden: Hier ist $\beta=\begin{pmatrix}\alpha_{s1} & \cdots & \alpha_{ss}\end{pmatrix}=e_s^T\mathcal{A}$, $e_s^T=\begin{pmatrix}0 & \cdots & 0 & 1\end{pmatrix}$, also ist $\beta\mathcal{A}^{-1}=e_s^T$. Dann ist

$$R(\infty)=1-e_s^T\mathbf{1}_s=1-1=0,$$

diese Verfahren sind also L-stabil.

Für die Lobatto IIIC-Methoden: Hier ist $\alpha_{i1}=\beta_1$, $i=1,\ldots,s$, d.h. die erste Spalte von $\mathcal{A}$ ist $\beta_1\mathbf{1}_s$ ($\mathcal{A}e_1=\beta_1\mathbf{1}_s$, $e_1^T=\begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}$). Da $\mathcal{A}$ regulär ist, muß $\beta_1\neq 0$ sein. Dann ist

$$\begin{aligned} R(\infty) &= 1-\beta\mathcal{A}^{-1}\mathbf{1}_s \\ &= 1-\beta\mathcal{A}^{-1}\left(\frac{1}{\beta_1}\mathcal{A}\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}\right) \\ &= 1-\frac{1}{\beta_1}\beta\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} \\ &= 1-\frac{1}{\beta_1}\beta_1 \\ &= 0, \end{aligned}$$

also sind diese Verfahren auch L-stabil.

Zur A-Stabilität der Gauß-Methoden siehe das Buch [SW95]. □

2.3 Globale Fehlerabschätzung, tatsächlicher Fehler

Wir erinnern uns an die Definition 2.3 des lokalen Diskretisierungsfaktor auf Seite 26: Für eine Inkrementfunktion φ (Def. 2.2, Seite 25) ist

$$\tau_n = \frac{1}{h_n}\left(x\left(t_n\right) - x\left(t_{n-1}\right)\right) - \varphi\left(x\left(t_{n-1}\right), t_{n-1}, h_n\right).$$

Das Einschrittverfahren ist definiert mittels

$$x_n = x_{n-1} + h_n \varphi\left(x_{n-1}, t_{n-1}, h_n\right).$$

Zur Bestimmung des lokalen Fehlers betrachten wir den n-ten Iterationsschritt ausgehend von der exakten Lösung $x\left(t_{n-1}\right)$,

$$\bar{x}_n = x\left(t_{n-1}\right) + h_n \varphi\left(x\left(t_{n-1}\right), t_{n-1}, h_n\right),$$

so daß für den lokalen Fehler gilt

$$x\left(t_n\right) - \bar{x}_n = h_n \tau_n.$$

Definition 2.24 (globaler, tatsächlicher Fehler). Der *globale Fehler* an der Stelle t_n ist definiert als

$$\boxed{\varepsilon_n := x\left(t_n\right) - x_n.}$$

Aus der näherungsweisen Lösung der nichtlinearen Gleichung (z.B. für X'_{ni}) entstehen noch zusätzliche Fehler δ_n, so daß wir tatsächlich nur

$$\boxed{\tilde{\varepsilon}_n := x\left(t_n\right) - \tilde{x}_n, \qquad \tilde{x}_n = \tilde{x}_{n-1} + h_n\left(\varphi\left(\tilde{x}_{n-1}, t_{n-1}, h_n\right) + \delta_n\right), \quad \tilde{x}_0 = x_0} \tag{2.5}$$

erhalten. $\tilde{\varepsilon}_n$ heißt *tatsächlicher Fehler*.

Wir betrachten nun das Intervall $[t_0, T]$ mit einer Diskretisierung

$$t_0 < t_1 < \ldots < t_N = T, \qquad h := \max_{j=1,\ldots,N} h_j.$$

Aus der Definition des lokalen Diskretisierungsfaktors erhalten wir

$$x\left(t_n\right) = x\left(t_{n-1}\right) + h_n \varphi\left(x\left(t_{n-1}\right), t_{n-1}, h_n\right) + h_n \tau_n, \qquad n = 1, \ldots, N. \tag{2.6}$$

Die Formeln (2.5) und (2.6) ergeben nun

$$\begin{aligned}
\tilde{\varepsilon}_n &= x\left(t_n\right) - \tilde{x}_n \\
&= \underbrace{x\left(t_{n-1}\right) - \tilde{x}_{n-1}}_{\tilde{\varepsilon}_{n-1}} + h_n\left(\varphi\left(x\left(t_{n-1}\right), t_{n-1}, h_n\right) - \varphi\left(\tilde{x}_{n-1}, t_{n-1}, h_n\right)\right) + h_n\left(\tau_n - \delta_n\right)
\end{aligned}$$

Unter entsprechenden Voraussetzungen an f ist φ Lipschitz-stetig bzgl. des ersten Arguments. Dann ist

$$\begin{aligned}
\left|\tilde{\varepsilon}_n\right| &\leq \left|\tilde{\varepsilon}_{n-1}\right| + h_n L_\varphi \underbrace{\left|x\left(t_{n-1}\right) - \tilde{x}_{n-1}\right|}_{\left|\tilde{\varepsilon}_{n-1}\right|} + h_n\left|\tau_n - \delta_n\right| \\
&\leq \left(1 + L_\varphi h_n\right)\left|\tilde{\varepsilon}_{n-1}\right| + h_n\left|\tau_n - \delta_n\right| \\
&\vdots \\
&\leq \left(1 + L_\varphi h\right)^n \underbrace{\left|\tilde{\varepsilon}_0\right|}_{\left|\tilde{x}_0 - x_0\right| = 0} + h \sum_{j=1}^{n-1}\left(1 + L_\varphi h\right)^j \cdot \max_{l=1,\ldots,n}\left|\tau_l - \delta_l\right| \\
&= h\frac{\left(1 + L_\varphi h\right)^n - 1}{\left(1 + L_\varphi h\right) - 1} \cdot \max_{l=1,\ldots,n}\left|\tau_l - \delta_l\right| \\
&\leq \left(1 + L_\varphi h\right)^n \cdot \frac{1}{L_\varphi} \cdot \max_{l=1,\ldots,n}\left|\tau_l - \delta_l\right|, \qquad n = 1, \ldots, N.
\end{aligned}$$

und mit $1 + L_\varphi h \leq e^{L_\varphi h}$ folgt nun

$$\boxed{|\tilde{\varepsilon}_n| \leq \frac{1}{L_\varphi} \cdot e^{L_\varphi hn} \cdot \max_{l=1,\ldots,n} |\tau_l - \delta_l| \,, \qquad n = 1, \ldots, N.}$$

Für äquidistante Zerlegungen ist $nh = t_n - t_0$, ansonsten muß für die Konvergenz $Nh \leq \text{const}$ gelten.

Für den globalen Fehler erhalten wir

$$|\varepsilon_n| \leq \frac{1}{L_\varphi} e^{L_\varphi hn} \cdot \max_{l=1,\ldots,n} |\tau_l| \,, \qquad n = 1, \ldots, N,$$

also

$$\boxed{\max_{n=0,\ldots,N} |\varepsilon_n| \leq \frac{1}{L_\varphi} e^{L_\varphi \text{const}} \cdot \max_{l=1,\ldots,N} |\tau_l| = O\left(h^p\right).}$$

D.h. **ein Runge-Kutta-Verfahren mit der Konsistenzordnung p hat auch die Konvergenzordnung p.** Aus

$$\max_{l=1,\ldots,N} |\tau_l| \xrightarrow{h \to 0} 0 \qquad \text{folgt} \quad \max_{l=0,\ldots,N} |\varepsilon_l| \xrightarrow{h \to 0} 0.$$

2.4 Praktische Umsetzung von Runge-Kutta-Verfahren

2.4.1 Schrittweitensteuerung

Wie können wir die Schrittweite so steuern, daß wir $|\tau_n| = \text{TOL}$ für eine vorgegebene Toleranz TOL erreichen?

Da wir den lokalen Fehler $x(t_n) - x_n$ nicht kennen, benutzen wir stattdessen $\hat{x}_n - \tilde{x}_n$, wobei $\hat{x}_n$ eine zusätzliche Approximation ist, die aus einem Verfahren mit Ordnung $p + 1$ oder manchmal auch $p - 1$ entsteht, z.B.

- Berechnung von $\hat{x}_n$ aus einem „eingebetteten" Runge-Kutta-Verfahren

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & \beta \\ & \bar{\beta} \end{array}$$

- Berechnung von $\hat{x}_n$ ausgehend von $\tilde{x}_{n-1}$ mit 2 Schritten der Länge $\frac{h_n}{2}$.

  Dies ist viel Aufwand, aber sinnvoll in Verbindung mit Extrapolationsverfahren.

Nun approximieren wir den lokalen Diskretisierungfaktor mittels

$$\hat{\tau}_n = \frac{1}{h_n} \left(\hat{x}_n - \tilde{x}_n\right).$$

Benutzen wir ein Verfahren der Konsistenzordnung p, so ist $|\tau_n| \approx c_n h_n^p$ für eine Konstante c_n. Unter der Annahme $c_n = c_{n+1}$ ist dann

$$\frac{|\tau_n|}{|\tau_{n+1}|} = \frac{h_n^p}{h_{n+1}^p}.$$

h_{n+1} soll nun so bemessen sein, daß $|\hat{\tau}_{n+1}| \approx \mathrm{TOL}$ gilt, d.h. wir müssen h_{n+1} so wählen, daß

$$\frac{|\hat{\tau}_n|}{\mathrm{TOL}} = \frac{h_n^p}{h_{n+1}^p}$$

gilt, d.h.

$$\boxed{h_{n+1} := h_n \sqrt[p]{\frac{\mathrm{TOL}}{|\hat{\tau}_n|}} = h_n \cdot \sqrt[p]{\frac{\mathrm{TOL}}{|\hat{x}_n - \tilde{x}_n|}}.}$$

2.4.2 Bestimmung von X'_{nj}

Der größte Arbeitsanteil im Runge-Kutta-Verfahren ist die Lösung der nichtlinearen Gleichungen zur Bestimmung der X'_{nj},

$$\begin{aligned} X_{ni} &= x_{n-1} + h_n \sum_{j=1}^{s} \alpha_{ij} f\left(X_{nj}, t_{nj}\right), \\ x_n &= x_{n-1} + h_n \sum_{j=1}^{s} \beta_j f\left(X_{nj}, t_{nj}\right). \end{aligned}$$

Für das implizite Eulerverfahren ist $X_{n1} = x_{n-1} + h_n f\left(X_{n1}, t_n\right)$. Ist $f(x,t) = Bx$, so ist

$$X_{n1} = x_{n-1} + h_n B X_{n1}.$$

Möchten wir die Fixpunktiteration

$$\begin{aligned} X_{n1}^{[0]} &:= x_{n-1} \\ X_{n1}^{[i+1]} &:= x_{n-1} + h_n B X_{n1}^{[i-1]}, \quad i \geq 1, \end{aligned}$$

benutzen, um das Gleichungssystem zu lösen, so können wir Konvergenz gegen eine Lösung (d.h. einen Fixpunkt) nur erwarten, wenn die Abbildung

$$\mathcal{B} : \mathbb{R}^m \to \mathbb{R}^m, \qquad \mathcal{B}X := x_{n-1} + h_n B X,$$

kontraktiv ist. Dies ist nur der Fall, falls $|h_n B| < 1$, d.h. $|h_n \lambda_B| < 1$ für die Eigenwerte λ_B von B, ist. Dies ergibt wieder eine Schrittweiteneinschränkung.

Besser geeignet ist das Newton-Verfahren mit Startwerten $X_{ni}^{[0]} := X_{n-1,i}$.

Kapitel 3

Stabilität im Sinne von Lyapunov

Wir betrachten die DGL

$$x'(t) = f(x(t), t), \tag{3.1}$$

für eine stetige Funktion $f : \mathbb{R}^m \times [0, \infty) \to \mathbb{R}^m$ mit stetiger Ableitung f_x nach x.

Dazu betrachten wir die AWA

$$x'(t) = f(x(t), t), \qquad x(0) = x_0. \tag{3.2}$$

Definition 3.1 ((asymptotisch) stabil im Sinne von Lyapunov, instabil).

1. Eine Lösung $x_* \in C^1([0, \infty), \mathbb{R}^m)$ der DGL (3.1) heißt *stabil im Sinne von Lyapunov*, falls für ein $\tau > 0$ alle Anfangswertaufgaben (3.2) mit $|x_0 - x_*(0)| \leq \tau$ Lösungen auf $[0, \infty)$ haben und zu jedem $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$ existiert, derart daß folgende Implikation gilt:

$$\boxed{|x_0 - x_*(0)| \leq \delta(\varepsilon) \quad \Rightarrow \quad |x(t, x_0) - x_*(t)| \leq \varepsilon \quad \forall t \geq 0}$$

2. x_* heißt *asymptotisch stabil im Sinne von Lyapunov*, falls zusätzlich ein $\tilde{\tau} > 0$ existiert, so daß folgende Implikation gilt:

$$\boxed{|x_0 - x_*(0)| \leq \tilde{\tau} \quad \Rightarrow \quad |x(t, x_0) - x_*(t)| \xrightarrow{t \to \infty} 0.}$$

3. Eine Lösung x_* heißt *instabil*, falls sie nicht stabil ist.

Bemerkung ([SH96, S. 126]). Eine Lösung $x_*(t)$ ist also stabil im Sinne von Lyapunov, wenn die Lösungen von (3.2) durch geeignete Wahl der Anfangswerte dazu gezwungen werden können, in einer Nachbarschaft von $x_*(t)$ zu verbleiben. nicht in der Vorlesung

$x_*(t)$ ist asymptotisch stabil im Sinne von Lyapunov, wenn zusätzlich jede Lösung, die in der Nähe von $x_*(t_0)$ startet, gegen $x_*(t)$ konvergiert, für $t \to \infty$.

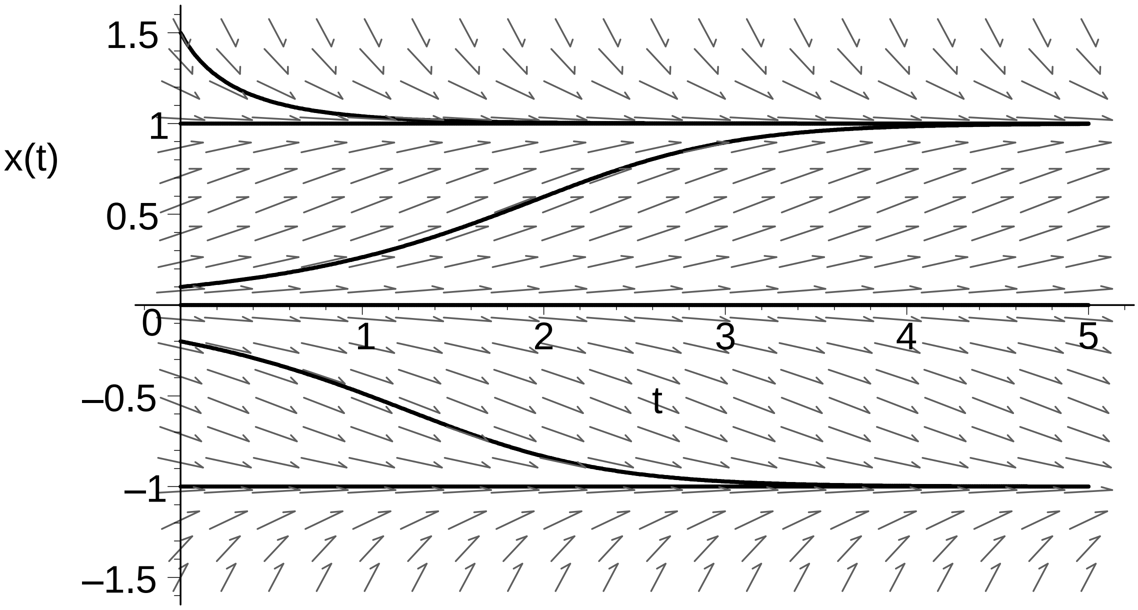
Beispiel ([SH96, S. 126]). Das Anfangswertproblem nicht in der Vorlesung

$$x'(t) = x(t) - x^3(t), \quad t \in [0, \infty), \qquad x(0) = x_0,$$

hat die exakte Lösung

$$x(t) = \frac{x_0 e^t}{\sqrt{1 + x_0^2 (e^{2t} - 1)}}.$$

Die Gleichgewichtslagen sind 0, 1 und -1. 1 und -1 sind asymptotisch stabil, 0 ist nicht stabil.



Lösungen der Anfangswertaufgabe für $x_0 \in \left\{-1, -\frac{1}{5}, 0, \frac{1}{10}, 1, \frac{3}{2}\right\}$.

3.1 Lineare Gleichung mit konstanten Koeffizienten

Wir betrachten lineare inhomogene Differentialgleichungen der Form

$$x'(t) = Bx(t) + q(t), \tag{3.3}$$

und die zugehörige homogene DGL

$$x'(t) = Bx(t). \tag{3.4}$$

Ist $x(\cdot, x_0)$ Lösung der AWA zur DGL (3.3) mit $x(0) = x_0$ und $x_*(\cdot)$ eine Referenzlösung, so ist $x(\cdot, x_0) - x_*(\cdot)$ Lösung der homogenen Differentialgleichung (3.4).

Also ist x_* eine stabile (asymptotisch stabile) Lösung der inhomogenen DGL (3.3), gdw. die Null-Lösung ($x(\cdot) \equiv 0$) eine stabile (asymptotisch stabile) Lösung der homogenen DGL (3.4) ist.

Daher betrachten wir in diesem Abschnitt nur noch die homogene DGL (3.4):

$$x'(t) = Bx(t).$$

Seien $\sigma_1, \ldots, \sigma_m$ die Eigenwerte von B.

Wir möchten zeigen:

- Falls $\mathrm{Re}\sigma_i < 0$, $i = 1, \ldots, m$, so ist die Null-Lösung asymptotisch stabil.
- Falls $\mathrm{Re}\sigma_i \leq 0$, $i = 1, \ldots, m$, und zu den rein imaginären Eigenwerten ($\mathrm{Re}\sigma_i = 0$) gehören ausschließlich Jordanblöcke der Ordnung 1, so ist die Null-Lösung stabil.

Definition 3.2 (Spektralabzisse). Sei $B \in L(\mathbb{C}^m)$,

$$\boxed{\alpha(B) := \max\{\mathrm{Re}\lambda \mid \det(\lambda I - B) = 0\}}$$

heißt *Spektralabzisse* von B.

Lemma 3.3. *Sei $B \in L(\mathbb{R}^m)$.*

1. *Ist $\alpha(B) < 0$, so gibt es eine reguläre Matrix $C \in L(\mathbb{C}^m)$ und ein $\gamma < 0$ derart, daß*[1]

$$\operatorname{Re}\langle C^* C B x, x\rangle_2 \leq \gamma \cdot |Cx|_2^2 \qquad \text{für alle } x \in \mathbb{R}^m.$$

2. *Ist $\alpha(B) = 0$ und gehören zu den rein imaginären Eigenwerten von B ausschließlich Jordanblöcke der Ordnung 1, so gibt es eine reguläre Matrix $C \in L(\mathbb{C}^m)$ mit*

$$\operatorname{Re}\langle C^* C B x, x\rangle_2 \leq 0 \qquad \text{für alle } x \in \mathbb{R}^m.$$

Erinnerung:

1. Sei $A \in L(\mathbb{C}^m)$ eine hermitische Matrix, d.h. $A = A^*$.
   Dann existiert in $\mathbb{C}^m$ eine Orthonormalbasis (ONB) von Eigenvektoren $v_1, \ldots, v_m$, und die Eigenwerte $\lambda_1, \ldots, \lambda_m$ sind reell. Sei

$$x = \sum_{j=1}^m c_j v_j, \qquad |x|_2^2 = \sum_{j=1}^m |c_j|^2.$$

   Dann ist

$$\langle Ax, x\rangle_2 = \left\langle \sum_{j=1}^m c_j \lambda_j v_j, \sum_{i=1}^m c_i v_i \right\rangle_2 = \sum_{j=1}^m |c_j|^2 \lambda_j \leq \underbrace{\max_{j=1,\ldots,m} \lambda_j}_{=:\lambda_{\max}(A)} \cdot |x|_2^2.$$

2. Für eine allgemeine Matrix $A \in L(\mathbb{C}^m)$ ist

$$\langle A^* x, x\rangle_2 = \langle x, (A^*)^* x\rangle_2 = \langle x, Ax\rangle_2 = \overline{\langle Ax, x\rangle_2}.$$

   Dann ist für $x \in \mathbb{C}^m$

$$\operatorname{Re}\langle Ax, x\rangle_2 = \frac{1}{2}\langle Ax, x\rangle_2 + \frac{1}{2}\overline{\langle Ax, x\rangle_2} = \left\langle \frac{1}{2}\underbrace{(A + A^*)}_{\text{hermitisch}} x, x \right\rangle_2 \leq \lambda_{\max}\left(\frac{1}{2}(A + A^*)\right) \cdot |x|_2^2.$$

Beweis. Sei $B \in L(\mathbb{R}^m)$. Wir stellen B dar als

$$B = TJT^{-1} \quad \text{mit} \quad J = T^{-1}BT = \begin{pmatrix} \sigma_1 & \delta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \delta_{m-1} \\ & & & \sigma_m \end{pmatrix}, \quad \delta_i \in \{0, 1\}.$$

Definiere weiter für $\varepsilon > 0$

$$D_\varepsilon := \begin{pmatrix} \varepsilon^1 & & \\ & \ddots & \\ & & \varepsilon^m \end{pmatrix}$$

und

$$J_\varepsilon := D_\varepsilon^{-1} J D_\varepsilon = \begin{pmatrix} \sigma_1 & \varepsilon\delta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon\delta_{m-1} \\ & & & \sigma_m \end{pmatrix} = \underbrace{D_\varepsilon^{-1} T^{-1}}_{=:C_\varepsilon} B T D_\varepsilon = C_\varepsilon B C_\varepsilon^{-1}.$$

[1] Für eine Matrix $A \in L(\mathbb{C}^m)$ definiere $A^* := \bar{A}^T$.

Für $x, y \in \mathbb{C}^m$ definiere dann ein inneres Produkt auf $\mathbb{C}^m$ mittels

$$\langle x, y\rangle_\varepsilon := \langle C_\varepsilon^* C_\varepsilon x, y\rangle_2 .$$

Die zugehörige Norm ist

$$|x|_\varepsilon = |C_\varepsilon x|_2 .$$

Für $x \in \mathbb{R}^m$ berechnen wir:

$$\begin{aligned}
\operatorname{Re}\langle Bx, x\rangle_\varepsilon &= \operatorname{Re}\langle C_\varepsilon^* C_\varepsilon Bx, x\rangle_2 \\
&= \operatorname{Re}\left\langle C_\varepsilon B\left(C_\varepsilon^{-1} C_\varepsilon\right) x, C_\varepsilon x\right\rangle_2 \\
&= \operatorname{Re}\langle J_\varepsilon (C_\varepsilon x), (C_\varepsilon x)\rangle_2 \\
&\leq \lambda_{\max}\left(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)\right) \cdot |C_\varepsilon x|_2^2
\end{aligned}$$

Es ist nun

$$\frac{1}{2}(J_\varepsilon + J_\varepsilon^*) = \begin{pmatrix} \operatorname{Re}\sigma_1 & \frac{\varepsilon}{2}\delta_1 & & \\ \frac{\varepsilon}{2}\delta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \frac{\varepsilon}{2}\delta_{m-1} \\ & & \frac{\varepsilon}{2}\delta_{m-1} & \operatorname{Re}\sigma_m \end{pmatrix} \xrightarrow{\varepsilon \to 0} \begin{pmatrix} \operatorname{Re}\sigma_1 & & \\ & \ddots & \\ & & \operatorname{Re}\sigma_m \end{pmatrix},$$

also

$$\lambda_{\max}\left(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)\right) \xrightarrow{\varepsilon \to 0} \max_{i=1,\ldots,m} \operatorname{Re}\sigma_i = \alpha(B).$$

zu 1. Wähle nun γ so, daß $\alpha(B) < \gamma < 0$ gilt. Für hinreichend kleine $\varepsilon > 0$ gilt dann

$$\lambda_{\max}\left(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)\right) \leq \gamma,$$

Fixiere solch ein $\varepsilon > 0$ und setze $C := C_\varepsilon$. Dann folgt also

$$\operatorname{Re}\langle C^* CBx, x\rangle_2 \leq \lambda_{\max}\left(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)\right) \cdot |Cx|_2^2 \leq \gamma \cdot |Cx|_2^2 .$$

zu 2. Seien $\sigma_1, \ldots, \sigma_s$ die rein imaginären Eigenwerte (d.h. $\operatorname{Re}\sigma_i = 0$, $i = 1, \ldots, s$). Dann ist

$$J = \begin{pmatrix} \sigma_1 & & & & & & \\ & \ddots & & & & & \\ & & \sigma_s & & & & \\ & & & \sigma_{s+1} & \delta_{s+1} & & \\ & & & & \ddots & \ddots & \\ & & & & & \ddots & \delta_{m-1} \\ & & & & & & \sigma_m \end{pmatrix},$$

$$\frac{1}{2}(J_\varepsilon + J_\varepsilon^*) = \begin{pmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & \operatorname{Re}\sigma_{s+1} & \frac{\varepsilon}{2}\delta_{s+1} & & \\ & & & \frac{\varepsilon}{2}\delta_{s+1} & \ddots & \ddots & \\ & & & & \ddots & \ddots & \frac{\varepsilon}{2}\delta_{m-1} \\ & & & & & \frac{\varepsilon}{2}\delta_{m-1} & \operatorname{Re}\sigma_m \end{pmatrix}$$

und nach Voraussetzung ist

$$\tilde{\alpha} := \max_{i=s+1,\ldots,m} \mathrm{Re}\sigma_i < 0.$$

Also ist für hinreichend kleines $\varepsilon > 0$

$$\lambda_{\max}\left(\frac{1}{2}\left(J_\varepsilon + J_\varepsilon^*\right)\right) \leq 0.$$

Fixiere ein solches $\varepsilon > 0$ und setze $C := C_\varepsilon$. Dann folgt

$$\mathrm{Re}\left\langle C^* C x, x\right\rangle_2 \leq \lambda_{\max}\left(\frac{1}{2}\left(J_\varepsilon + J_\varepsilon^*\right)\right) \cdot |Cx|_2^2 \leq 0.$$ □

Lemma 3.4. *Sei $C \in L(\mathbb{C}^m)$ regulär. Dann definiert*

$$b : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \qquad b(x,y) := \mathrm{Re}\left\langle C^* C x, y\right\rangle_2, \quad x, y \in \mathbb{R}^m,$$

ein inneres Produkt auf $\mathbb{R}^m$.

Beweis. Es ist die Symmetrie, Linearität und positive Definitheit von b zu zeigen:
Zur Symmetrie: Für $x, y \in \mathbb{R}^m$ ist

$$b(x,y) = \mathrm{Re}\left\langle C^* C x, y\right\rangle_2 = \mathrm{Re}\left\langle x, C^* C y\right\rangle_2 = \mathrm{Re}\overline{\left\langle x, C^* C y\right\rangle_2} = \mathrm{Re}\left\langle C^* C y, x\right\rangle_2 = b(y,x).$$

Zur Linearität: Für $x, y, z \in \mathbb{R}^m$ und $\alpha, \beta \in \mathbb{R}$ gilt

$$b(x, \alpha y + \beta z) = \mathrm{Re}\left\langle C^* C x, \alpha y + \beta z\right\rangle_2 = \alpha b(x,y) + \beta b(x,z).$$

Zur positiven Definitheit: Für $x \in \mathbb{R}^m$ ist

$$b(x,x) = \mathrm{Re}\left\langle C^* C x, x\right\rangle_2 = |Cx|_2^2 \geq 0$$

und es ist $b(x,x) = |Cx|_2^2 > 0$ gdw. $x \neq 0$ (da C regulär). □

Aus den letzten beiden Lemmas folgt dann

Satz 3.5. *Sei $B \in L(\mathbb{R}^m)$.*

1. *Ist $\alpha(B) < 0$, so existiert ein inneres Produkt $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^m$ und ein $\gamma < 0$ mit:*

$$\langle Bx, x\rangle \leq \gamma |x|^2, \quad x \in \mathbb{R}^m.$$

2. *Ist $\alpha(B) = 0$ und gehören zu den rein imaginären Eigenwerten von B ausschließlich Jordanblöcke der Ordnung 1, so existiert ein inneres Produkt $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^m$ mit*

$$\langle Bx, x\rangle \leq 0, \quad x \in \mathbb{R}^m.$$

Korollar 3.6 (Stabilität für lineare DGLn, [SH96, S. 127]). *Sei $B \in L(\mathbb{R}^m)$ und betrachte die DGL (3.4): $x'(t) = Bx(t)$. Dann gilt* nicht in der Vorlesung

1. *Die Null-Lösung $x_*(\cdot) \equiv 0$ ist genau dann asymptotisch stabil, wenn $\alpha(B) < 0$ gilt.*
2. *Ist $\alpha(B) \leq 0$ und gehören zu den rein imaginären Eigenwerten von B ausschließlich Jordanblöcke der Ordnung* 1, *so ist die Null-Lösung $x_*(\cdot) \equiv 0$ stabil.*

Beweisidee. Mit dem inneren Produkt aus Satz 3.5 folgt, daß die einseitige Lipschitz-Bedingung (1.3) (Seite 7) erfüllt ist. Wende nun Satz 1.6 (Seite 8) an. □

3.2 Nichtlineare Differentialgleichungen

Wir möchten nun die Aussage von Korollar 3.6 auf nichtlineare Differentialgleichungen übertragen.

Satz 3.7 (Perron). *Gegeben sei die Differentialgleichung*

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

$$f(x,t) = Bx + h(x,t), \qquad x \in \mathbb{R}^m, \quad t \in [0,\infty).$$

Sei $\alpha(B) < 0$ und $h(0,t) = 0$ für alle $t \geq 0$. Zu jedem $\varepsilon > 0$ existiere ein $\rho(\varepsilon) > 0$ derart, daß

$$|h(x,t)|_2 \leq \varepsilon |x|_2, \qquad \text{für alle } x \in \mathbb{R}^m \text{ mit } |x|_2 \leq \rho(\varepsilon), \quad t \geq 0.$$

Dann ist die triviale Lösung $x_*(\cdot) \equiv 0$ asymptotisch stabil im Sinne von Lyapunov.

Beweis. Nach Satz 3.5, Teil 1, existiert ein inneres Produkt $\langle \cdot, \cdot \rangle$ auf $\mathbb{R}^m$ und ein $\gamma < 0$, so daß

$$\langle Bx, x \rangle \leq \gamma |x|^2, \qquad x \in \mathbb{R}^m.$$

Sei $\varepsilon > 0$, $\rho(\varepsilon) > 0$. Zu einem Anfangswert $x_0 \in B(0, \rho(\varepsilon))$ lösen wir die Anfangswertaufgabe für (3.5) mit $x(0) = x_0$. Die Lösung $x(t) := x(t, x_0)$ sei auf $[0, T)$ definiert. Dann ist

$$\begin{aligned} \frac{d}{dt}|x(t)|^2 &= 2\langle x'(t), x(t) \rangle \\ &= 2\langle Bx(t) + h(x(t), t), x(t) \rangle \\ &= 2\langle Bx(t), x(t) \rangle + 2\langle h(x(t), t), x(t) \rangle \\ &\leq 2\gamma |x(t)|^2 + 2\langle h(x(t), t), x(t) \rangle \end{aligned}$$

Auf $\mathbb{R}^m$ sind die Normen $|\cdot|$ und $|\cdot|_2$ äquivalent, d.h. es existieren Konstanten η_1, η_2, so daß

$$\eta_1 |x|_2 \leq |x| \leq \eta_2 |x|_2$$

für alle $x \in \mathbb{R}^m$. Dann folgt

$$|h(x(t), t)| \leq \eta_2 |h(x(t), t)|_2 \leq \eta_2 \cdot \varepsilon |x(t)|_2 \leq \frac{\eta_2}{\eta_1} \cdot \varepsilon |x(t)|,$$

also

$$\frac{d}{dt}|x(t)|^2 \leq 2\gamma |x(t)|^2 + 2\frac{\eta_2}{\eta_1} \cdot \varepsilon |x(t)|^2 = 2\underbrace{\left(\gamma + \varepsilon \frac{\eta_2}{\eta_1}\right)}_{=:\kappa=\kappa(\varepsilon)} \cdot |x(t)|^2.$$

Für hinreichend kleine $\varepsilon > 0$ ist $\kappa(\varepsilon) < 0$, da $\gamma < 0$. Wir fixieren ein solches ε mit $\kappa(\varepsilon) < 0$. Dann folgt mit dem Lemma von Gronwall (Lemma 1.5, Seite 8, mit $z(\cdot) := |x(\cdot)|^2$, $a := 2\kappa$, $b := 0$), daß

$$|x(t)| \leq e^{\kappa t} |x(0)| = e^{\kappa t} |x_0| \qquad \text{für alle } t \in [0, T).$$

Dann ist die Lösung $x(\cdot, x_0)$ auf $[0, \infty)$ fortsetzbar. Zusätzlich gilt

$$x(t) \xrightarrow{t \to \infty} 0.$$

Wegen $x_* \equiv 0$ folgt daraus die asymptotische Stabilität. □

Bemerkung. Falls h glatt ist, so ist $h(x,t) = \underbrace{h(0,t)}_{=0} + h_x(0,t)\, x + O\left(|x|^2\right)$.

Satz 3.8 (Lyapunov: Lineare Stabilität impliziert Stabilität). *Für eine Funktion* $f \in C^2(\mathbb{R}^m, \mathbb{R}^m)$ *sei die autonome Differentialgleichung*

$$x'(t) = f(x(t))$$

gegeben. Sei $c \in \mathbb{R}^m$ *mit* $f(c) = 0$ *eine Gleichgewichtslage und sei* $B := f'(c)$.

Gilt $\alpha(B) < 0$, *so ist die stationäre Lösung* $x_*(\cdot) :\equiv c$ *asymptotisch stabil.*

Beweis. Sei zunächst $c = 0$. Dann ist

$$f(x) = \underbrace{f(0)}_{=0} + \underbrace{f'(0)}_{=:B} x + h(x) \quad \text{mit } |h(x)|_2 \leq \kappa |x|_2^2 \qquad \text{für } x \in \overline{B(0,R)}.$$

Die DGL $x'(t) = f(x(t))$ hat dann die Form

$$x'(t) = Bx(t) + h(x(t)).$$

Die Aussage folgt nun aus dem Satz 3.7 von Perron.

Sei nun $c \neq 0$. Wie führen folgende Translation aus

$$\tilde{x}(t) := x(t) - c, \qquad \tilde{f}(\tilde{x}) := f(\tilde{x} + c).$$

Dann ist

$$\begin{aligned} \tilde{x}'(t) &= f(\tilde{x}(t) + c) = \tilde{f}(\tilde{x}(t)), \\ \tilde{f}(0) &= f(c) = 0, \\ \tilde{f}'(0) &= f'(c) = B. \end{aligned}$$

Nach dem eben gezeigten ist dann $\tilde{x}_*(\cdot) := 0$ eine asymptotisch stabile Lösung der DGL $\tilde{x}'(t) = \tilde{f}(\tilde{x}(t))$, also $x_* = \tilde{x}_* + c$ eine asymptotisch stabile Lösung der DGL $x'(t) = f(x(t))$. □

Beispiel ([SH96, S. 132]). Für $m = 2$ betrachte die Differentialgleichung

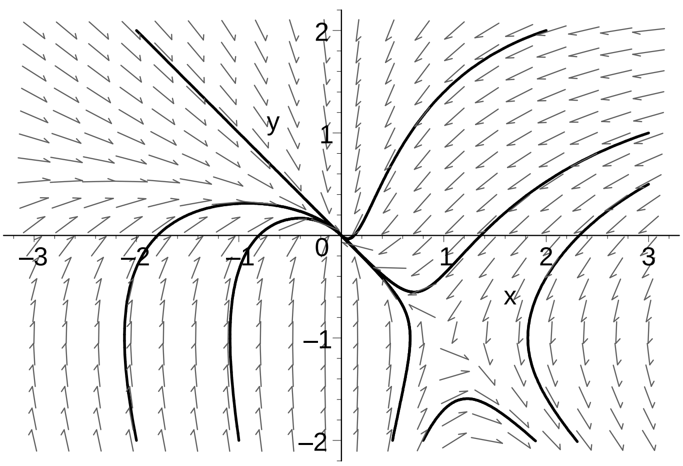
nicht in der Vorlesung

$$\begin{aligned} x_1'(t) &= -x_1(t) - x_1(t)\,x_2(t), \\ x_2'(t) &= -x_1(t) - 2x_2(t) + x_1(t)\,x_2(t). \end{aligned}$$

Die Jacobi-Matrix in der Gleichgewichtslage $(0,0)$ ist gegeben durch

$$\begin{pmatrix} -1 & 0 \\ -1 & -2 \end{pmatrix}.$$

Diese Matrix hat die Eigenwerte -1 und -2. Nach Satz 3.8 ist die stationäre Lösung $x_*(\cdot) \equiv 0$ also asymptotisch stabil.



Lösungen der Differentialgleichung für Startwerte $x(0)$ aus $\{(-2,2),\ (-2,-2),\ (-1,-2),\ (2,2),\ (3,1),\ (3,\frac{1}{2}),\ (\frac{4}{5},-2),\ (\frac{1}{2},-2)\}$.

Beispiel ([SH96, S. 132]). Für $m = 2$ betrachte die Differentialgleichung

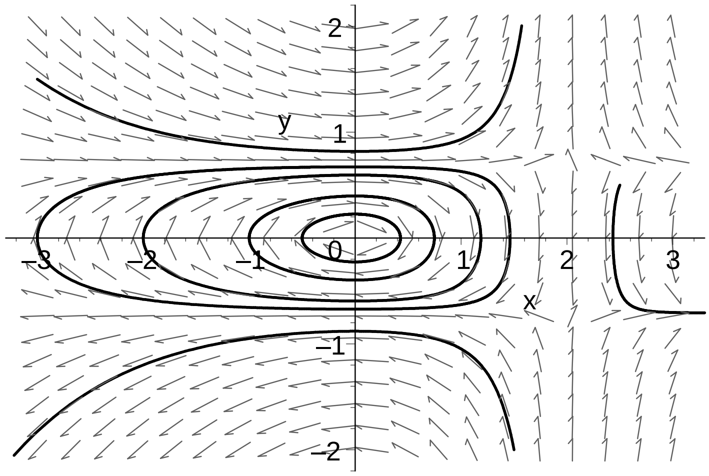
nicht in der Vorlesung

$$\begin{aligned} x_1'(t) &= 2x_2(t) - x_1(t)\, x_2(t), \\ x_2'(t) &= -\frac{1}{2}x_1(t) + x_1(t)\, x_2(t)^2. \end{aligned}$$

Die Jacobi-Matrix in der Gleichgewichtslage $(0, 0)$ ist gegeben durch

$$\begin{pmatrix} 0 & 2 \\ -\frac{1}{2} & 0 \end{pmatrix}.$$

Diese Matrix hat die Eigenwerte $\pm i$. Dann läßt sich Satz 3.8 nicht anwenden, um die Stabilität der stationären Lösung $x_*(\cdot) \equiv 0$ zu beurteilen.



Lösungen der Differentialgleichung für Startwerte
$x(0) \in \left\{ \left(-3, \frac{3}{2}\right), (-3, 0), (-2, 0), (-1, 0), \left(-\frac{1}{2}, 0\right), \left(\frac{3}{2}, -2\right), \left(\frac{5}{2}, \frac{1}{2}\right) \right\}$.

Beispiel (Routh (1877)). 1. Betrachten wir die Differentialgleichung

$$\begin{aligned} x_1'(t) &= -x_2(t) + x_1^3(t), \\ x_2'(t) &= x_1(t). \end{aligned}$$

mit Gleichgewichtslage $c = (0, 0)$. Dann ist

$$B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad h(x) = \begin{pmatrix} x_1^3 \\ 0 \end{pmatrix},$$

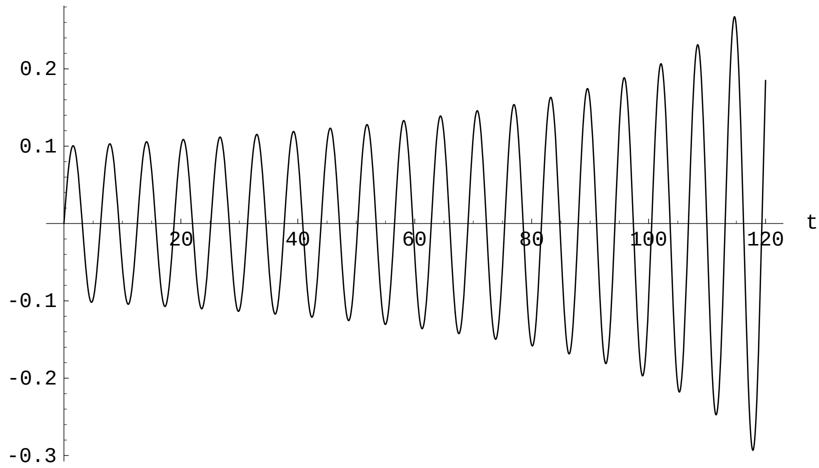
also

$$x'(t) = B \cdot x(t) + h(x(t)).$$

Die Eigenwerte von B sind $\sigma_{1,2} = \pm i$, es ist also $\alpha(B) = 0$, so daß wir Satz 3.8 nicht anwenden können. Nun ist

$$\begin{aligned} \frac{d}{dt}|x(t)|_2^2 &= 2\langle x'(t), x(t)\rangle_2 \\ &= 2\left\langle \begin{pmatrix} -x_2 + x_1^3 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\rangle \\ &= 2\left(-x_2x_1 + x_1^4 + x_1x_2\right) \\ &= 2 \cdot x_1^4(t) \\ &\geq 0, \end{aligned}$$

d.h. die Lösungen wachsen. Die stationäre Lösung $x_*(\cdot) \equiv 0$ kann also nicht asymptotisch stabil sein.



Darstellung der Komponente $x_2(t)$ für $x_1(0) = \frac{1}{10}$ und $x_2(0) = 0$.

2. Betrachte nun die Differentialgleichung

$$\begin{aligned} x_1'(t) &= -x_2(t) - x_1^3(t), \\ x_2'(t) &= x_1(t). \end{aligned}$$

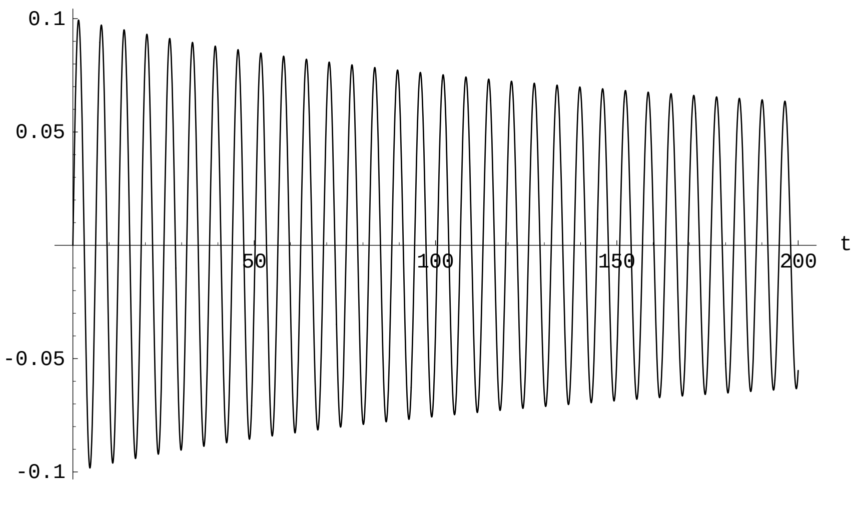
mit Gleichgewichtslage $c = (0, 0)$. Dann ist

$$B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad h(x) = \begin{pmatrix} -x_1^3 \\ 0 \end{pmatrix},$$

die Eigenwerte von B sind wieder $\sigma_{1,2} = \pm i$ und es ist $\alpha(B) = 0$. Nun ist

$$\frac{d}{dt} |x(t)|_2^2 = 2 \langle x'(t), x(t) \rangle = 2 \left(-x_2 x_1 - x_1^4 + x_1 x_2 \right) = -2x_1^4(t) \leq 0,$$

d.h. die Lösungen fallen hier. Die Null-Lösung $x_*(\cdot) \equiv 0$ ist stabil im Sinne von Lyapunov.



Darstellung der Komponente $x_2(t)$ für $x_1(0) = \frac{1}{10}$ und $x_2(0) = 0$.

Für diese Differentialgleichung gilt die Ungleichung der schwachen Kontraktivität mit $\langle \cdot, \cdot \rangle_2$ (vgl. Beispiel 1.1 (Seite 7), Punkt 3). Nach Satz 1.6 (Seite 8), Punkt 2, folgt dann

$$|x(t)|_2 \leq |x(0)|_2, \qquad t \geq 0.$$

Um auch im Fall $\alpha(f'(c)) = 0$ eine Aussage über die Stabilität der stationären Lösung zu machen, benötigt man den Begriff der Lyapunov-Funktion:

Definition 3.9 (Lyapunov-Funktion). Die AWA (3.1) habe $x_*(\cdot) \equiv 0$ als Lösung. Eine Funktion $V \in C^1(\mathbb{R}^m, \mathbb{R})$ heißt *Lyapunov-Funktion* zur DGL (3.1) falls

$$\boxed{V(y) \geq 0 \quad \forall y \in \mathbb{R}^m \quad \text{und} \quad V(y) = 0 \Leftrightarrow y = 0}$$

und für alle Lösungen $x(t)$ der DGL gilt:

$$\boxed{\frac{d}{dt} V(x(t)) \leq 0 \qquad \forall t > 0}$$

Bemerkung. Falls es eine Lyapunov-Funktion zu (3.1) gibt, so gilt

$$V(x(t)) \leq V(x(t_0)),$$

also sind alle Lösungen beschränkt, und daher fortsetzbar auf $[0, \infty)$. x_* ist stabil im Sinne von Lyapunov.

Kapitel 4

Randwertaufgaben und periodische Lösungen

4.1 Korrekt formulierte lineare Randwertaufgaben (RWAn) und Greensche Funktion

Wir betrachten lineare Differentialgleichungen der Form

$$\boxed{x'(t) = A(t)\,x(t) + q(t) \qquad t \in [t_0, T]} \tag{4.1}$$

mit $A \in C([t_0, T], L(\mathbb{R}^m))$, $q \in C([t_0, T], \mathbb{R}^m)$ und *Randbedingungen*

$$\boxed{B_0\,x(t_0) + B_T\,x(T) = d} \tag{4.2}$$

mit $B_0, B_T \in L(\mathbb{R}^m)$, und $d \in \mathbb{R}^m$.

Spezialfälle von Randbedingungen (4.2) sind:

- $B_0 = I$, $B_T = 0$, also $x(t_0) = d$. Dann liegt eine Anfangswertaufgabe vor.
- $B_0 = \begin{pmatrix} I & \\ & 0 \end{pmatrix}$, $B_T = \begin{pmatrix} 0 & \\ & I \end{pmatrix}$
- $B_0 = -B_T$, also $B_0\,(x(t_0) - x(T)) = d$
- $B_0 = I$, $B_T = -I$, $d = 0$, also $x(t_0) = x(T)$ (Periodizität)

Beispiel 4.1. Sei $m = 2$, und betrachte das Intervall $[0, T]$. Seien die DGL

$$\begin{aligned} x_1'(t) &= x_2(t), \\ x_2'(t) &= -x_1(t), \end{aligned}$$

und die Randbedingungen

$$x_1(0) = 0, \qquad x_1(T) = \beta$$

gegeben. Dann ist

$$B_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B_T = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad d = \begin{pmatrix} 0 \\ \beta \end{pmatrix}.$$

Die allgemeine Lösung der DGL lautet:

$$x(t) = c_1 z_1(t) + c_2 z_2(t) = c_1 \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix} + c_2 \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} \qquad c_1, c_2 \in \mathbb{R}.$$

Die Randbedingung $x_1(0) = 0$ ergibt

$$0 = c_1 \cos 0 + c_2 \sin 0 = c_1,$$

und aus $x_1(T) = \beta$ folgt

$$\beta = c_1 \cos T + c_2 \sin T = c_2 \sin T.$$

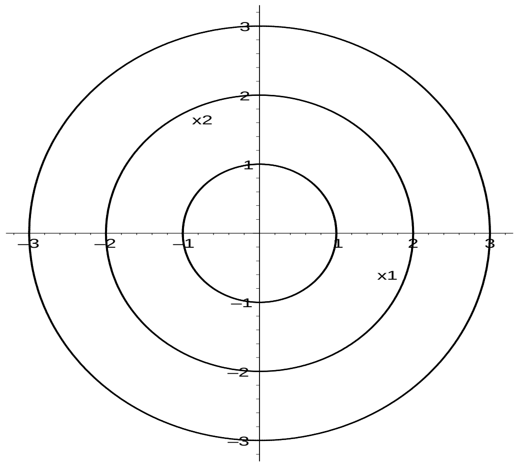
Ist nun $T \neq \pi k$ für alle $k \in \mathbb{Z}$, so erhalten wir $c_2 = \frac{\beta}{\sin T}$, also die eindeutige Lösung

$$x(t) = \frac{\beta}{\sin T} \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}.$$

Ist $T = \pi k$ für ein $k \in \mathbb{Z}$, so liefert die 2. Randbedingung die Bedingung $\beta = 0$. Ist also $\beta = 0$, so lösen alle Funktionen

$$x(t) = c_2 \cdot \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}, \qquad c_2 \in \mathbb{R},$$

die Randwertaufgabe. Ist $\beta \neq 0$, so existiert keine Lösung.



Lösungen ($c_2 = 1, 2, 3$) der Randwertaufgabe für $T = 2\pi$ und $\beta = 0$ (d.h. die Randbedingungen sind $x_1(0) = 0$ und $x_1(2\pi) = 0$)

4.1.1 Lösbarkeitsbedingung

Welche Bedingung müssen wir also an die Randbedingungen (4.2) stellen, damit die Randwertaufgabe eindeutig lösbar ist?

Die lineare Anfangswertaufgabe

$$x'(t) = A(t)\, x(t) + q(t), \quad t \in [t_0, T], \qquad x(t_0) = x_0, \quad x_0 \in \mathbb{R}^m \tag{4.3}$$

besitzt eine Lösung $x(\cdot, x_0) \in C^1([t_0, T], \mathbb{R}^m)$. Diese hat die Form

$$x(t, x_0) = X(t)\, x_0 + x_q(t) \qquad t \in [t_0, T],$$

mit Fundamentalmatrix $X \in C^1([t_0, T], L(\mathbb{R}^m))$, so daß

$$X'(t) = A(t)\, X(t), \quad t \in [t_0, T], \qquad X(t_0) = I,$$

und partikulärer Lösung $x_q \in C^1([t_0, T], \mathbb{R}^m)$, so daß

$$x_q'(t) = A(t)\, x_q(t) + q(t), \qquad x_q(t_0) = 0,$$

d.h. $x_q(t) = X(t) \int_{t_0}^t X^{-1}(s)\, q(s)\, ds$.

Durch Einsetzen der Lösungen in die Randbedingungen (4.2) erhält man

$$B_0 X(t_0)\, x_0 + B_T \left(X(T)\, x_0 + x_q(T) \right) = d,$$

d.h.

$$\left(B_0 X(t_0) + B_T X(T) \right) x_0 = d - B_T x_q(T).$$

Die Matrix

$$\boxed{M := B_0 X(t_0) + B_T X(T)}$$

heißt *Lösbarkeitsmatrix*. Wenn M regulär ist, so ist die lineare Randwertaufgabe (4.1), (4.2) eindeutig lösbar für beliebige $d \in \mathbb{R}^m$ und $q \in C([t_0, T], \mathbb{R}^m)$. Die Lösung der Randwertaufgabe entspricht dann der Lösung der linearen Anfangswertaufgabe (4.3) mit

$$x_0 = M^{-1}(d - B_T x_q(T)).$$

Für die Regularität von

$$M = \begin{pmatrix} B_0 & B_T \end{pmatrix} \cdot \begin{pmatrix} X(t_0) \\ X(T) \end{pmatrix}$$

ist die Bedingung Rang $\begin{pmatrix} B_0 & B_T \end{pmatrix} = m$ notwendig.

Beispiel. Betrachten wir nochmals die Randwertaufgabe aus Beispiel 4.1:

$$\begin{aligned} x_1'(t) &= x_2(t), \\ x_2'(t) &= -x_1(t), \end{aligned}$$

$$B_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B_T = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad d = \begin{pmatrix} 0 \\ \beta \end{pmatrix}.$$

Die Fundamentalmatrix ist dann

$$X(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix},$$

so daß man folgende Lösbarkeitsmatrix erhält

$$M = B_0 X(0) + B_T X(T) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \cos T & \sin T \\ -\sin T & \cos T \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \cos T & \sin T \end{pmatrix}.$$

Diese Matrix ist genau dann regulär, falls $T \neq \pi k$ für alle $k \in \mathbb{Z}$ gilt.

4.1.2 Darstellung der Lösung $x(\cdot)$

Sei M regulär. Die Lösung der RWA (4.1), (4.2) ist dann

$$\begin{aligned} x(t) &= X(t) \cdot M^{-1}(d - B_T x_q(T)) + x_q(t) \\ &= X(t) M^{-1} d - X(t) M^{-1} B_T X(T) \int_{t_0}^{T} X(s)^{-1} q(s)\, ds + X(t) \int_{t_0}^{t} X(s)^{-1} q(s)\, ds \\ &= X(t) M^{-1} d + \int_{t_0}^{t} X(t)\left(-M^{-1} B_T X(T) + I\right) X(s)^{-1} q(s)\, ds \\ &\quad - \int_{t}^{T} X(t) M^{-1} B_T X(T) X(s)^{-1} q(s)\, ds \end{aligned}$$

und mit $I = M^{-1} B_0 X(t_0) + M^{-1} B_T X(T)$ erhalten wir weiter

$$= X(t) M^{-1} d + \int_{t_0}^{t} X(t) M^{-1} B_0 X(t_0) X(s)^{-1} q(s)\, ds - \int_{t}^{T} X(t) M^{-1} B_T X(T) X(s)^{-1} q(s)\, ds.$$

Also

$$\boxed{x(t) = X(t)\, M^{-1} d + \int_{t_0}^{T} \mathcal{G}(t,s)\, q(s)\, ds, \qquad t \in [t_0, T],} \tag{4.4}$$

mit der *Greenschen Funktion*

$$\boxed{\mathcal{G}(t,s) := \begin{cases} X(t)\, M^{-1} B_0 X(t_0)\, X(s)^{-1} & s \leq t, \\ -X(t)\, M^{-1} B_T X(T)\, X(s)^{-1} & s > t. \end{cases}} \tag{4.5}$$

Die Zahlen

$$\boxed{\begin{aligned} \kappa_1 &:= \max_{t\in[t_0,T]} \left\| X(t)\, M^{-1} \right\|, \\ \kappa_2 &:= \sup_{t,s\in[t_0,T]} \left\| \mathcal{G}(t,s) \right\| \end{aligned}} \tag{4.6}$$

heißen *Konditionszahlen* der Randwertaufgabe.

Zu $q \in C([t_0,T], \mathbb{R}^m)$ und $d \in \mathbb{R}^m$ erhalten wir aus (4.4) für die Lösung $x \in C^1([t_0,T], \mathbb{R}^m)$ die Ungleichung

$$\boxed{\|x\|_\infty \leq \kappa_1 |d| + \kappa_2 (T - t_0) \|q\|_\infty .}$$

Wiederholung:

Definition 4.2 (korrekt gestellt). Sei $L : X \to Y$ eine lineare Abbildung in normierten Räumen X und Y. Die Gleichung

$$Lx = y$$

ist *korrekt gestellt* falls

- die Gleichung lösbar ist für beliebige $y \in Y$,
- die Lösung eindeutig ist,
- und die Lösung x stetig von y abhängt.

Ist L eine Bijektion und L^{-1} beschränkt, so ist die Gleichung $Lx = y$ korrekt gestellt.

Sei nun $X = C^1([t_0,T], \mathbb{R}^m)$ und $Y = C([t_0,T], \mathbb{R}^m) \times \mathbb{R}^m$. Für $x \in X$ sei

$$Lx := \begin{pmatrix} x' - Ax \\ B_0 x(t_0) + B_T x(T) \end{pmatrix}.$$

Die Gleichung

$$Lx = \begin{pmatrix} q \\ d \end{pmatrix} =: y \in Y$$

hat dann die Lösung

$$x = L^{-1} \begin{pmatrix} q \\ d \end{pmatrix}$$

und es ist

$$\|x\|_{C^1} := \|x\|_\infty + \|x'\|_\infty \leq \kappa_1 |d| + \kappa_2 \|q\|_\infty .$$

Für reguläres M ist die Aufgabe $Lx = y$ also korrekt gestellt.

4.2 Über die Kondition von linearen Randwertaufgaben

Definition 4.3 (skaliert). Die Randbedingung (4.2) heißt *skaliert*, falls orthogonale Matrizen $Q_0, Q_T \in L(\mathbb{R}^m)$ und Diagonalmatrizen $D_0, D_T \in L(\mathbb{R}^m)$ mit

$$D_0 = \begin{pmatrix} \Delta_0 & 0 \\ \underbrace{0}_{m-k} & \underbrace{I}_{k} \end{pmatrix}, \qquad D_T = \begin{pmatrix} I & 0 \\ \underbrace{0}_{m-k} & \underbrace{\Delta_T}_{k} \end{pmatrix},$$

wobei die Einträge von Δ_0 und Δ_T aus $[0,1]$ stammen, existieren, so daß

$$\boxed{B_0 = D_0 Q_0^T \qquad \text{und} \qquad B_T = D_T Q_T^T.}$$

Bemerkung. Falls $\operatorname{Rang}\begin{pmatrix} B_0 & B_T \end{pmatrix} = m$, so kann die Randbedingung (4.2) in skalierte Form überführt werden.

Sei

$$\boxed{\Phi(t) := X(t)\, M^{-1}, \qquad t \in [t_0, T].}$$

Dann ist

$$B_0 \Phi(t_0) + B_T \Phi(T) = (B_0 X(t_0) + B_T X(T))\, M^{-1} = I$$

und

$$\mathcal{G}(t,s) = \begin{cases} \Phi(t)\, B_0 \Phi(t_0)\, \Phi^{-1}(s) & s \le t, \\ -\Phi(t)\, B_T \Phi(T)\, \Phi^{-1}(s) & s > t. \end{cases}$$

Lemma 4.4. *M sei regulär, die Randbedingungen (4.2) seien skaliert. Dann ist*

$$\|\Phi(t)\|_2 \le 2 \sup_{s \in [t_0,T]} \|\mathcal{G}(t,s)\|_2, \qquad t \in [t_0, T],$$

und

$$\kappa_1 \le 2\kappa_2 \qquad (\text{bzgl. } \|\cdot\|_2).$$

Beweis. Sei $B_0 = D_0 Q_0^T$ und $B_T = D_T Q_T^T$ mit

$$D_0 + D_T = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \begin{pmatrix} \Delta_0 & 0 \\ 0 & \Delta_T \end{pmatrix}$$

Da die Einträge von Δ_0 und Δ_T aus $[0,1]$ stammen, ist $D_0 + D_T$ invertierbar und

$$\left\|(D_0 + D_T)^{-1}\right\|_2 \le 1.$$

Dann ist

$$\begin{aligned}
\|\Phi(t)\|_2 &= \left\|\Phi(t)(D_0 + D_T)(D_0 + D_T)^{-1}\right\|_2 \\
&\le \|\Phi(t)(D_0 + D_T)\|_2 \cdot \underbrace{\left\|(D_0 + D_T)^{-1}\right\|_2}_{\le 1} \\
&\le \|\Phi(t) D_0\|_2 + \|\Phi(t) D_T\|_2 \\
&= \left\|\Phi(t) D_0 Q_0^T\right\|_2 + \left\|\Phi(t) D_T Q_T^T\right\|_2, \qquad \text{da } Q_0, Q_T \text{ orthogonal} \\
&= \left\|\Phi(t) B_0 \cdot \Phi(t_0)\, \Phi^{-1}(t_0)\right\|_2 + \left\|\Phi(t) B_T \cdot \Phi(T)\, \Phi^{-1}(T)\right\|_2 \\
&= \|\mathcal{G}(t,t_0)\|_2 + \|\mathcal{G}(t,T)\|_2 \\
&\le 2 \cdot \sup_{s \in [t_0,T]} \|\mathcal{G}(t,s)\|_2
\end{aligned}$$

Wegen

$$\begin{aligned}\kappa_1 &= \max_{t\in[t_0,T]} \left\|X(t)\,M^{-1}\right\|_2 = \max_{t\in[t_0,T]} \|\Phi(t)\|_2\,,\\ \kappa_2 &= \sup_{t,s\in[t_0,T]} \|\mathcal{G}(t,s)\|_2\end{aligned}$$

folgt dann die zweite Behauptung. □

Definition 4.5 (dichotomisch). Die DGL

$$x'(t) = A(t)\,x(t)\,, \qquad t\in[t_0,\infty)\,,$$

mit $A\in C([t_0,\infty), L(\mathbb{R}^m))$ heißt *dichotomisch*, falls es einen Projektor $P\in L(\mathbb{R}^m)$, also $P^2=P$, und Konstanten $K>0$, $\alpha,\beta\geq 0$ gibt, mit der Eigenschaft

$$\begin{aligned}\left\|X(t)\,PX(s)^{-1}\right\| &\leq K\cdot e^{-\alpha(t-s)} && \text{für } s\leq t, \quad s,t\in[t_0,\infty)\,,\\ \left\|X(t)\,(I-P)\,X(s)^{-1}\right\| &\leq K\cdot e^{-\beta(s-t)} && \text{für } s> t, \quad s,t\in[t_0,\infty)\,.\end{aligned}$$

Falls $\alpha,\beta>0$, so spricht man von *exponentieller Dichotomie*.

Beispiel. Für $m=3$ betrachte die DGL

$$x'(t) = \begin{pmatrix} -2 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 3\end{pmatrix} x(t)\,, \qquad t\in[0,\infty)\,.$$

Dann ist

$$X(t) = \begin{pmatrix} e^{-2t} & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & e^{3t}\end{pmatrix} \quad \text{und} \quad X(s)^{-1} = \begin{pmatrix} e^{2s} & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & e^{-3s}\end{pmatrix}.$$

Wir versuchen

$$P = \begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{pmatrix}, \qquad I-P = \begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 1\end{pmatrix}.$$

Dann ist

$$X(t)\,PX(s)^{-1} = \begin{pmatrix} e^{-2(t-s)} & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{pmatrix} \quad \text{und} \quad X(t)\,(I-P)\,X(s)^{-1} = \begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & e^{-3(s-t)}\end{pmatrix},$$

also

$$\begin{aligned}\left\|X(t)\,PX(s)^{-1}\right\| &= e^{-2(t-s)} && \text{für } s\leq t,\\ \left\|X(t)\,(I-P)\,X(s)^{-1}\right\| &= e^{-3(s-t)} && \text{für } s\geq t.\end{aligned}$$

D.h. die DGL ist exponentiell dichotomisch mit $K=1$, $\alpha=2$ und $\beta=3$.

Versuchen wir $P = \begin{pmatrix} 1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}$, so ergibt sich ebenfalls $K=1$, $\alpha=2$, $\beta=3$.

Bemerkung. Der Projektor P teilt die Lösungen in nicht wachsende und nicht fallende Lösungen auf. Konstante Lösungen können zu beiden Unterräumen hinzugenommen werden (in Abhängigkeit von P), wie das obige Beispiel zeigt.

Die allgemeine Lösung der DGL $x'(t) = A(t)\,x(t)$ ist

$$x(t) = X(t)\,x_0 = X(t)\,Px_0 + X(t)\,(I-P)\,x_0.$$

Definiere nun

$$\begin{aligned} S_-(t) &:= \{X(t)\,Px_0\,|x_0 \in \mathbb{R}^m\} = \mathrm{im}X(t)\,P, \\ S_+(t) &:= \{X(t)\,(I-P)\,x_0\,|x_0 \in \mathbb{R}^m\} = \mathrm{im}X(t)\,(I-P). \end{aligned}$$

Bemerkung. Die Mengen $S_-(t)$ und $S_+(t)$ bilden eine direkte Zerlegung des $\mathbb{R}^m$:
Sei $z \in S_-(t) \cap S_+(t)$ für ein $t \geq t_0$. Dann ist

$$z = X(t)\,Pw_1, \qquad z = X(t)\,(I-P)\,w_2$$

für Punkte $w_1, w_2 \in \mathbb{R}^m$. Da $X(t)$ regulär ist, folgt

$$Pw_1 = (I-P)\,w_2.$$

Multiplikation mit dem Projektor P $(P^2 = P)$ liefert dann

$$Pw_1 = P^2w_1 = Pw_2 - P^2w_2 = Pw_2 - Pw_2 = 0.$$

Also ist $z = X(t)\,Pw_1 = 0$, d.h.

$$S_-(t) \oplus S_+(t) = \mathbb{R}^m.$$

Bemerkung. $S_-(t)$ enthält die nicht wachsenden Lösungen, $S_+(t)$ die nicht fallenden Lösungen:
Sei $x(t) \in S_-(t)$ mit $x(t) \neq 0$, d.h. $Px_0 \neq 0$. Dann gilt für alle $s \leq t$

$$\frac{|x(t)|}{|x(s)|} = \frac{|X(t)\,Px_0|}{|X(s)\,Px_0|} = \frac{\left|X(t)\,PX(s)^{-1}\,X(s)\,Px_0\right|}{|X(s)\,Px_0|} \leq \left\|X(t)\,PX(s)^{-1}\right\| \leq Ke^{-\alpha(t-s)},$$

d.h. die Lösung x ist nicht wachsend, denn für alle $t \geq s$ ist

$$|x(t)| \leq Ke^{-\alpha(t-s)}\,|x(s)|.$$

Sei nun $x(t) \in S_+(t)$ mit $x(t) \neq 0$, d.h. $(I-P)\,x_0 \neq 0$. Dann gilt für alle $s > t$

$$\frac{|x(t)|}{|x(s)|} \leq \left\|X(t)\,(I-P)\,X(s)^{-1}\right\| \leq Ke^{-\beta(s-t)},$$

d.h. die Lösung x ist nicht fallend, denn für alle $s > t$ ist

$$|x(s)| > \frac{1}{K}e^{\beta(s-t)}\,|x(t)|.$$

Nicht jede DGL $x'(t) = A(t)\,x(t)$, $A \in C([t_0,\infty), L(\mathbb{R}^m))$ ist dichotomisch:

Beispiel. Sei $m = 2$, $\varepsilon > 0$ und für $t \in [-1,\infty)$ betrachte die DGL

$$\begin{aligned} x_1'(t) &= -\frac{1}{\varepsilon}2tx_1(t), \\ x_2'(t) &= x_2(t). \end{aligned}$$

Dann ist

$$\tilde{X}(t) = \begin{pmatrix} e^{-\frac{t^2}{\varepsilon}} & 0 \\ 0 & e^t \end{pmatrix} \qquad \text{und} \quad \tilde{X}(s)^{-1} = \begin{pmatrix} e^{\frac{s^2}{\varepsilon}} & 0 \\ 0 & e^{-s} \end{pmatrix}.$$

eine Fundamentalmatrix mit $\tilde{X}(0) = I$. Sei

$$X(t) = \tilde{X}(t) R$$

mit der regulären Matrix $R = \left(\tilde{X}(-1)\right)^{-1}$, so daß $X(-1) = \tilde{X}(-1) R = I$.

Wir versuchen $P = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$: Dann ist

$$\tilde{X}(t) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tilde{X}(s)^{-1} = \begin{pmatrix} e^{-\frac{1}{\varepsilon}\left(t^2-s^2\right)} & 0 \\ 0 & 0 \end{pmatrix}$$

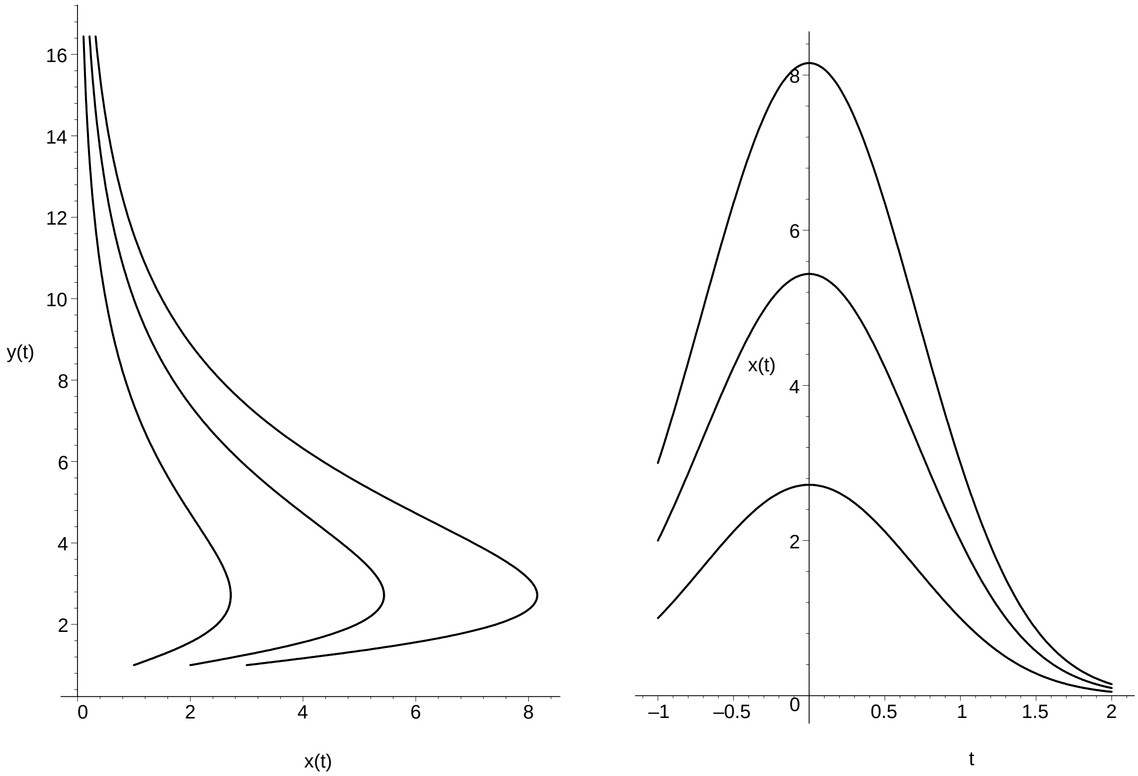
und

$$\left\| \tilde{X}(t) P \tilde{X}(s)^{-1} \right\| = e^{-\frac{1}{\varepsilon}\left(t^2-s^2\right)} = e^{-\frac{1}{\varepsilon}(t+s)(t-s)}.$$

Hier ist keine Bedingung

$$e^{-\frac{1}{\varepsilon}(t+s)(t-s)} \overset{?}{\leq} K \cdot e^{-\alpha(t-s)}$$

für $-1 \leq s \leq t < \infty$ erreichbar.



Lösungen der DGL für $\varepsilon = 1$ und Startwerte $(x(-1), y(-1)) \in \{(1,1), (2,1), (3,1)\}$.

Satz 4.6. *M sei regulär, $M = B_0 X(t_0) + B_T X(T)$.*

1. *Die DGL $x'(t) = A(t) x(t)$ sei dichotomisch. Für die Randbedingungen (4.2) gelte:*

$$B_0 X(t_0)(I - P) = 0, \qquad B_T X(T) P = 0.$$

Dann ist $\kappa_2 \leq K$ und

$$\|\mathcal{G}(t,s)\| \leq K \cdot \begin{cases} e^{-\alpha(t-s)} & s \leq t, \\ e^{-\beta(s-t)} & s > t. \end{cases}$$

2. *Es gelte* $r_0 + r_T = m$, $r_0 = \operatorname{Rang}(B_0)$, $r_T = \operatorname{Rang}(B_T)$, *und*
$$\|\mathcal{G}(t,s)\| \leq K \cdot \begin{cases} e^{-\alpha(t-s)} & s \leq t, \\ e^{-\beta(s-t)} & s > t, \end{cases}$$
für alle $s,t \in [t_0, T]$ *mit Konstanten* $K > 0$, $\alpha, \beta \geq 0$.
Dann ist die DGL $x'(t) = A(t)\,x(t)$ *dichotomisch mit den Konstanten* K, α, β.

Beweis. 1. Sei $s \leq t$:
$$\begin{aligned} \mathcal{G}(t,s) &= X(t)\,M^{-1}B_0X(t_0)\,X(s)^{-1} \\ &= X(t)\,M^{-1}B_0X(t_0)\,PX(s)^{-1}, \qquad \text{da } B_0X(t_0)(I-P) = 0 \\ &= X(t)\,M^{-1}(M - B_TX(T))\,PX(s)^{-1} \\ &= X(t)\,PX(s)^{-1}, \qquad \text{da } B_TX(T)\,P = 0. \end{aligned}$$
Also folgt
$$\|\mathcal{G}(t,s)\| \leq K \cdot e^{-\alpha(t-s)}.$$
Sei $s > t$:
$$\begin{aligned} \mathcal{G}(t,s) &= -X(t)\,M^{-1}B_TX(T)\,X(s)^{-1} \\ &= -X(t)\,M^{-1}B_TX(T)(I-P)\,X(s)^{-1}, \qquad \text{da } B_TX(T)\,P = 0 \\ &= -X(t)\,M^{-1}(M - B_0X(t_0))(I-P)\,X(s)^{-1} \\ &= -X(t)(I-P)\,X(s)^{-1}, \qquad \text{da } B_0X(t_0)(I-P) = 0. \end{aligned}$$
Also folgt
$$\|\mathcal{G}(t,s)\| \leq K \cdot e^{-\beta(s-t)}.$$

2. Wir verwenden $\Phi(t) = X(t)\,M^{-1}$ mit $B_0\Phi(t_0) + B_T\Phi(T) = I$. Dann ist
$$\mathcal{G}(t,s) = \begin{cases} \Phi(t)\,B_0\Phi(t_0)\,\Phi(s)^{-1} & s \leq t, \\ -\Phi(t)\,B_T\Phi(T)\,\Phi(s)^{-1} & s > t. \end{cases}$$
Wir zeigen, daß $P := B_0\Phi(t_0)$ der gesuchte Projektor ist.
Sei $E \in L(\mathbb{R}^m)$ orthogonal $\left(E^T = E^{-1}\right)$, so daß
$$EB_T = \begin{pmatrix} * \\ 0 \end{pmatrix} \begin{matrix} \}r_T \\ \end{matrix}.$$
Wegen $B_0\Phi(t_0) + B_T\Phi(T) = I$ ist dann
$$EB_0\Phi(t_0)\,E^T + EB_T\Phi(T)\,E^T = I.$$
Daraus folgt
$$EB_0\Phi(t_0)\,E^T = I - EB_T\Phi(T)\,E^T = I - \begin{pmatrix} * & * \\ 0 & 0 \end{pmatrix} \begin{matrix} \}r_T \\ \}r_0 \end{matrix} =: \begin{pmatrix} H_{11} & H_{12} \\ 0 & I \end{pmatrix}.$$
Wegen $\operatorname{Rang}(B_0) = r_0$ ist auch $\operatorname{Rang}\left(EB_0\Phi(t_0)\,E^T\right) = r_0$, also muß $H_{11} = 0$ sein. Dann ist
$$EB_0\Phi(t_0)\,E^T = \begin{pmatrix} 0 & H_{12} \\ 0 & I \end{pmatrix}$$
ein Projektor $\left(\begin{pmatrix} 0 & H_{12} \\ 0 & I \end{pmatrix}^2 = \begin{pmatrix} 0 & H_{12} \\ 0 & I \end{pmatrix}\right)$, und dann auch
$$B_0\Phi(t_0) = E^T \begin{pmatrix} 0 & H_{12} \\ 0 & I \end{pmatrix} E.$$
□

4.3 Nichtlineare Randwertaufgaben

Wir betrachten nun Gleichungen der Form

$$\boxed{x'(t) = f(x(t), t), \qquad t \in [t_0, T]} \tag{4.7}$$

mit $f : \mathbb{R}^m \times I_f \to \mathbb{R}^m$ stetig, $[t_0, T] \subseteq I_f \subseteq \mathbb{R}$, f_x stetig, und Randbedingungen

$$\boxed{r(x(t_0), x(T)) = 0} \tag{4.8}$$

mit $r = (r_1, r_2) : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ stetig differenzierbar.

Dazu betrachten wir die Anfangswertaufgabe

$$\boxed{x'(t) = f(x(t), t), \quad t \in [t_0, T], \qquad x(t_0) = x_0} \tag{4.9}$$

Sei

$$\boxed{S(x_0) := r(x_0, x(T, x_0)) \qquad x_0 \in \mathbb{R}^m}$$

wobei $x(\cdot, x_0)$ eine Lösung der AWA (4.9) bezeichne.

$$\boxed{S(x_0) = 0}$$

ist eine nichtlineare Gleichung bzgl. x_0.

Für eine Lösung $x_* \in C^1([t_0, T], \mathbb{R}^m)$ von (4.7), (4.8) gilt $S(x_*(t_0)) = 0$.

Wir linearisieren die Randwertaufgabe (4.7), (4.8) in x_* (Quasilinearisierung längs $x_*(t)$, $t \in [t_0, T]$), indem wir folgende *homogene* Randwertaufgabe definieren:

$$\boxed{\begin{aligned} z'(t) &= A_*(t) z(t), \\ 0 &= B_{*0} z(t_0) + B_{*T} z(T), \end{aligned}} \tag{4.10}$$

mit

$$\begin{aligned} A_*(t) &:= f_x(x_*(t), t), \qquad t \in [t_0, T] \\ B_{*0} &:= r_1'(x_*(t_0), x_*(T)) \\ B_{*T} &:= r_2'(x_*(t_0), x_*(T)) \end{aligned}$$

Die Fundamentalmatrix $X_*(t)$ zur DGL $z'(t) = A_*(t) z(t)$ ist Lösung von

$$X_*'(t) = A_*(t) X_*(t), \quad t \in [t_0, T], \qquad X_*(t_0) = I,$$

und es sei

$$M_* := B_{*0} + B_{*T} X_*(T).$$

Satz 4.7. *Sei $x_* \in C^1([t_0, T], \mathbb{R}^m)$ eine Lösung der RWA (4.7), (4.8), und sei die lineare homogene RWA (4.10) nur trivial lösbar. Dann gilt:*

1. *x_* ist isolierte Lösung der RWA.*
2. *Für hinreichend kleines $\delta > 0$ sind die gestörten RWAn*

$$\begin{aligned} x'(t) &= f(x(t), t) + q(t), \qquad t \in [t_0, T] && (4.11) \\ r(x(t_0), x(T)) &= d, && (4.12) \end{aligned}$$

$\|q\|_\infty + |d| \le \delta$, *lokal eindeutig lösbar. Die Lösung x der RWA hängt stetig von (q, d) ab.*

3. *Alle AWA (4.9) mit* $x_0 \in \bar{B}(x_*(t_0), \delta)$, $\delta > 0$ *hinreichend klein, besitzen Lösungen, die (mindestens) auf* $[t_0, T]$ *definiert sind.* $x(t, x_0)$ *ist stetig differenzierbar bzgl.* x_0.
4. *Die Abbildung* $S(x_0) := r(x_0, x(T, x_0))$ *ist wohldefiniert auf* $\bar{B}(x_*(t_0), \delta)$ *und dort stetig differenzierbar. Es gilt* $S'(x_*(t_0)) = M_*$.

Beweis. 1. Wir definieren die Abbildung

$$\mathcal{F} : C^1([t_0, T], \mathbb{R}^m) \to C([t_0, T], \mathbb{R}^m) \times \mathbb{R}^m,$$

$$\mathcal{F}(x) := (x'(\cdot) - f(x(\cdot), \cdot),\ r(x(t_0), x(T))), \qquad x \in C^1([t_0, T], \mathbb{R}^m).$$

Es ist

$$\mathcal{F}(x) = 0 \quad \Leftrightarrow \quad x \text{ erfüllt die RWA (4.7), (4.8)}$$

für $x \in B_{C^1}(x_*, \delta)$,

$$B_{C^1}(x_*, \delta) := \left\{ x \in C^1([t_0, T], \mathbb{R}^m) \middle| \|x - x_*\|_\infty + \|x' - x_*'\|_\infty \leq \delta \right\}.$$

Wir zeigen, daß $\mathcal{F}$ Frechet-differenzierbar ist und $\mathcal{F}'(x_*)$ regulär. Daraus folgt dann die Behauptung.

Proposition. *$\mathcal{F}$ ist Frechet-differenzierbar und für $z \in C^1([t_0, T], \mathbb{R}^m)$ ist*

$$\mathcal{F}'(x_*) z = (z' - A_* z, B_{*0} z(t_0) + B_{*T} z(T))$$

($\mathcal{F}'(x_*)$ ist eine lineare beschränkte Abbildung von $C^1([t_0, T], \mathbb{R}^m)$ in $C([t_0, T], \mathbb{R}^m) \times \mathbb{R}^m$).

Beweis. Sei $x \in C^1([t_0, T], \mathbb{R}^m)$. Wir berechnen

$$\lim_{\tau \to 0} \frac{1}{\tau} (\mathcal{F}(x + \tau z) - \mathcal{F}(x)), \qquad z \in C^1([t_0, T], \mathbb{R}^m).$$

Für die erste Komponente von $\frac{1}{\tau}(\mathcal{F}(x + \tau z) - \mathcal{F}(x))$ ergibt sich

$$\begin{aligned} & \frac{1}{\tau} \{x'(t) + \tau z'(t) - f(x(t) + \tau z(t), t) - x'(t) + f(x(t), t)\} \\ = \quad & z'(t) - \frac{1}{\tau} \{f(x(t) + \tau z(t), t) - f(x(t), t)\} \\ \xrightarrow{\tau \to 0} \quad & z'(t) - f_x(x(t), t) z(t), \qquad t \in [t_0, T]. \end{aligned}$$

Für die zweite Komponente von $\frac{1}{\tau}(\mathcal{F}(x + \tau z) - \mathcal{F}(x))$ ergibt sich

$$\begin{aligned} & \frac{1}{\tau} \{r(x(t_0) + \tau z(t_0), x(T) + \tau z(T)) - r(x(t_0), x(T))\} \\ = \quad & \frac{1}{\tau} \left\{ r(x(t_0), x(T)) + r'(x(t_0), x(T)) \begin{pmatrix} \tau z(t_0) \\ \tau z(T) \end{pmatrix} + o\left(\begin{pmatrix} \tau z(t_0) \\ \tau z(T) \end{pmatrix} \right) - r(x(t_0), x(T)) \right\} \\ = \quad & \frac{1}{\tau} \{r_1'(x(t_0), x(T)) \tau z(t_0) + r_2'(x(t_0), x(T)) \tau z(T) + o(\tau)\} \\ \xrightarrow{\tau \to 0} \quad & r_1'(x(t_0), x(T)) z(t_0) + r_2'(x(t_0), x(T)) z(T). \end{aligned}$$

Es ist also

$$\mathcal{F}'(x) z = (z' - f_x(x(\cdot), \cdot) z, r_1'(x(t_0), x(T)) z(t_0) + r_2'(x(t_0), x(T)) z(T)).$$

eine lineare beschränkte Abbildung von $C^1([t_0, T], \mathbb{R}^m)$ in $C([t_0, T], \mathbb{R}^m) \times \mathbb{R}^m$, und es ist

$$\mathcal{F}'(x_*) z = \left(z' - \underbrace{f_x(x_*(t), t)}_{A_*} z, \underbrace{r_1'(x_*(t_0), x_*(T))}_{B_{*0}} z(t_0) + \underbrace{r_2'(x_*(t_0), x_*(T))}_{B_{*T}} z(T) \right) \quad \square$$

Die Gleichung

$$\mathcal{F}'(x_*)\, z = 0$$

ist die Randwertaufgabe (4.10). Nach Voraussetzung ist diese nur trivial lösbar, d.h. $\mathcal{F}'(x_*)\, z = 0 \Leftrightarrow z = 0$ ist. Also ist $\mathcal{F}'(x_*)$ injektiv. Mit der Lösbarkeitsbedingung (Abschnitt 4.1.1, Seite 60) folgt dann die Regularität von M_* und damit die Bijektivität von $\mathcal{F}'(x_*)$.

$\mathcal{F}'(x_*)$ ist also eine beschränkte bijektive Abbildung in Banach-Räumen. Dann ist $\mathcal{F}'(x_*)$ ein Homöomorphismus.

2. Zu $x \in C^1([t_0,T], \mathbb{R}^m)$, $p = (q,d) \in C([t_0,T], \mathbb{R}^m) \times \mathbb{R}^m$ sei nun

$$\mathcal{H}(x,p) := \mathcal{F}(x) - p.$$

Die Gleichung

$$\mathcal{H}(x,p) = 0$$

entspricht der gestörten RWA (4.11), (4.12). Es ist

$$\mathcal{H}(x_*,0) = \mathcal{F}(x_*) = 0,$$

$\mathcal{H}$ ist Frechet-differenzierbar, und es ist

$$\mathcal{H}_x(x_*,0) = \mathcal{F}'(x_*) \quad \text{bijektiv.}$$

Nach dem Satz über implizite Funktionen existiert dann eine eindeutig bestimmte Funktion

$$\varphi : \bar{B}_{C\times\mathbb{R}^m}(0,\delta) \to \bar{B}_{C^1}(x_*,\rho), \qquad \rho \le \delta,$$

mit der Eigenschaft

$$\mathcal{H}(\varphi(p), p) = 0 \qquad \text{für alle } p \in \bar{B}_{C\times\mathbb{R}^m}(0,\delta).$$

Also hat die Gleichung $\mathcal{F}(x) = (q,d)$ für $q \in C([t_0,T], \mathbb{R}^m)$, $d \in \mathbb{R}^m$, $\|q\|_\infty + |d| \le \delta$, genau eine Lösung $\varphi((q,d))$ auf $\bar{B}_{C^1}(x_*,\rho)$.

3. Wie untersuchen die „künstliche" RWA (4.7) mit der Randbedingung $x(t_0) = x_*(t_0)$:

$$\begin{aligned} x'(t) &= f(x(t),t), \quad t \in [t_0,T] \\ \breve{r}(x(t_0), x(T)) &= 0 \\ \breve{r}(u,v) &:= u - x_*(t_0), \qquad u,v \in \mathbb{R}^m. \end{aligned}$$

Diese RWA hat die Lösung x_*. Es ist

$$\breve{r}_1'(u,v) \equiv I \quad \text{und} \quad \breve{r}_2'(u,v) \equiv 0,$$

also $\breve{B}_{*0} = I$, $\breve{B}_{*T} = 0$, $\breve{M}_* = I$, $\breve{A}_* = A_*$.

Nach 2. sind dann die gestörten „künstlichen" RWAn

$$\begin{aligned} x'(t) &= f(x(t),t) + q(t), \qquad t \in [t_0,T], \\ \breve{r}(x(t_0), x(T)) &= d, \qquad (\Leftrightarrow x(t_0) = x_*(t_0) + d) \end{aligned}$$

lokal eindeutig lösbar. Die Lösung $x = \breve{\varphi}(q,d)$ ist stetig bzgl. (q,d). Speziell ist für $q = 0$ die Lösung $x = \breve{\varphi}(0,d)$ stetig bzgl. d. D.h. die AWA (4.9)

$$\begin{aligned} x'(t) &= f(x(t),t), \qquad t \in [t_0,T], \\ x(t_0) &= x_*(t_0) + d, \end{aligned}$$

ist für $|d| \le \delta$ lokal eindeutig lösbar, die Lösung $x = \breve{\varphi}(0,d) \in C^1([t_0,T], \mathbb{R}^m)$ ist stetig bzgl. d.

4. Nach 3. ist $S(x_0) = r(x_0, x(T, x_0))$ wohldefiniert für $x_0 \in \bar{B}(x_*(t_0), \delta)$ und dort stetig differenzierbar.

Es bleibt $S'(x_*(t_0)) = M_*$ zu zeigen.

Es ist

$$\frac{\partial}{\partial x_0} S(x_0) = r_1'(x_0, x(T, x_0)) + r_2'(x_0, x(T, x_0)) \cdot \frac{\partial}{\partial x_0} x(T, x_0).$$

Wegen $x'(t, x_0) = f(x(t, x_0), t)$, $t \in [t_0, T]$ und $x(t_0, x_0) = x_0$ folgt nun

$$\frac{d}{dt}\frac{\partial}{\partial x_0} x(t, x_0) = \frac{\partial}{\partial x_0} x'(t, x_0) = \frac{\partial}{\partial x_0} f(x(t, x_0), t) = f_x(x(t, x_0), t) \frac{\partial}{\partial x_0} x(t, x_0), \quad t \in [t_0, T],$$

$$\frac{\partial}{\partial x_0} x(t_0, x_0) = I.$$

Für $x_0 = x_*(t_0)$, $x(t, x_0) = x_*(t)$, $X(t, x_0) := \frac{\partial}{\partial x_0} x(t, x_0)$ also

$$\begin{aligned} X'(t, x_*(t_0)) &= \underbrace{f_x(x_*(t), t)}_{A_*(t)} X(t, x_*(t_0)), \\ X(t, x_*(t_0)) &= I. \end{aligned}$$

Diese AWA hat die Lösung $X_*(t)$, es ist also

$$\frac{\partial}{\partial x_0} x(t, x_*(t_0)) = X(t, x_*(t_0)) = X_*(t),$$

und damit

$$\begin{aligned} S'(x_*(t_0)) &= r_1'(x_*(t_0), x_*(T)) + r_2'(x_*(t_0), x_*(T)) X_*(T) \\ &= B_{*0} + B_{*T} X_*(T) \\ &= M_*. \end{aligned}$$

□

4.4 Periodische Lösungen

Wir betrachten nun Differentialgleichungen

$$x'(t) = f(x(t), t), \qquad t \in \mathbb{R}, \quad I_f = \mathbb{R},$$

mit

$$\boxed{f(x, t) = f(x, t + T), \qquad x \in \mathbb{R}^m, \quad t \in \mathbb{R}.}$$

Definition 4.8 (T-periodisch). Eine Lösung $x_* \in C^1(\mathbb{R}, \mathbb{R}^m)$ der DGL $x'(t) = f(x(t), t)$ heißt *T-periodisch*, falls

$$\boxed{x_*(t) = x_*(t + T) \qquad t \in \mathbb{R}.}$$

Falls die RWA

$$\boxed{\begin{aligned} x'(t) &= f(x(t), t), \qquad t \in [0, T] \\ x(0) - x(T) &= 0 \end{aligned}} \tag{4.13}$$

die Lösung $x_* \in C^1([0, T], \mathbb{R}^m)$ hat, so kann x_* zu einer T-periodischen Lösung fortgesetzt werden.

Falls die Lösbarkeitsmatrix

$$\boxed{M_* = I - X_*(T)}$$

regulär ist, können wir dann Satz 4.7 anwenden?

Beispiel. Für $m = 1$, $T = 2\pi$ und $\alpha \neq 0$ betrachte die Differentialgleichung

$$x'(t) = \alpha x(t) + \sin(t)$$

(lineare DGL mit erzwungener Schwingung). Die Fundamentalmatrix ist

$$X(t) = e^{\alpha t},$$

die Lösbarkeitsmatrix ist

$$M = 1 - e^{\alpha 2\pi} \neq 0.$$

Die allgemeine Lösung der DGL ist

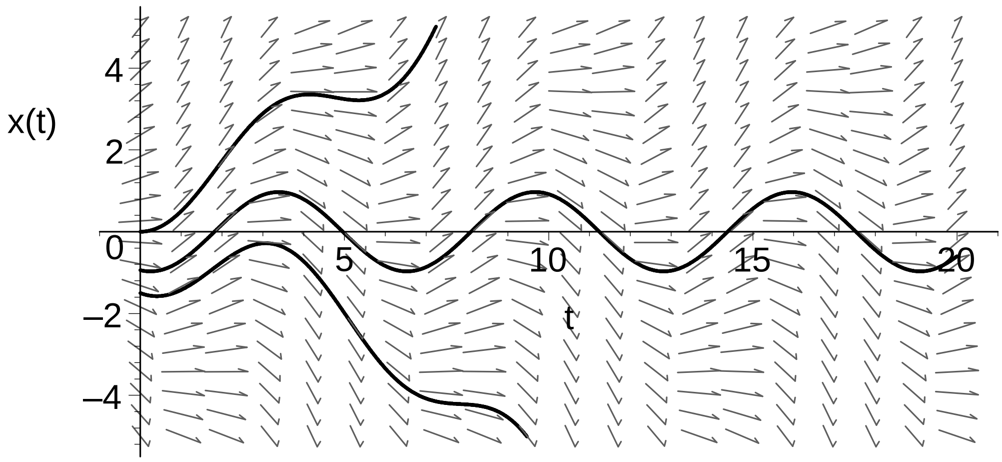
$$x(t) = e^{\alpha t} x_0 + \int_0^t e^{\alpha(t-\tau)} \sin \tau d\tau.$$

Aus der Periodizitätsbedingung $x_*(2\pi) = x_0$ erhält man dann

$$x_0 = -\frac{1}{1+\alpha^2}$$

und die periodische Lösung

$$x_*(t) = -\frac{1}{1+\alpha^2}(\alpha \sin t + \cos t).$$



Lösungen der DGL $x'(t) = \frac{1}{4}x(t) + \sin(t)$ für Startwerte $x_0 \in \left\{-\frac{3}{2}, -\frac{16}{17}, 0\right\}$.

Beispiel. Wir betrachten nochmals Beispiel 4.1 (Seite 59): Sei $m = 2$ und die DGL

$$x'(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x(t)$$

gegeben. Die Fundamentalmatrix ist

$$X(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix},$$

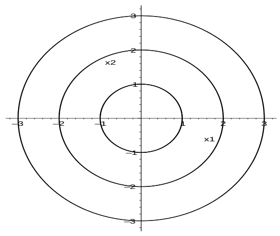
die Lösbarkeitsmatrix ist

$$M = I - X(T) = \begin{pmatrix} 1-\cos T & -\sin T \\ \sin T & 1-\cos T \end{pmatrix}.$$

Falls $T = 2\pi$ ist, so ist $M = 0$. Alle Lösungen,

$$x(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} x_0, \qquad x_0 \in \mathbb{R}^m,$$

sind periodisch.



2π-periodische Lösungen der DGL

Im weiteren untersuchen wir autonome Differentialgleichungen:

$$\boxed{x'(t) = f(x(t)), \qquad f \in C^1(\mathbb{R}^m, \mathbb{R}^m).}$$

Sei x_* eine Lösung. Dann ist

$$x_*'(t) = f(x_*(t))$$

stetig differenzierbar bzgl. t, d.h. $x_* \in C^2(\mathbb{R}, \mathbb{R}^m)$,

$$x_*''(t) = f_x(x_*(t)) \cdot x_*'(t) = A_*(t)\, x_*'(t).$$

Für

$$z := x_*'$$

folgt dann

$$z_*'(t) = A_*(t)\, z_*(t).$$

Ist x_* eine T-periodische Lösung, so ist

$$\begin{aligned} x_*(0) &= x_*(T), \\ x_*'(0) &= x_*'(T), \end{aligned}$$

also $z_*(0) = z_*(T)$. Dann ist z_* eine Lösung der linearisierten Randwertaufgabe (vgl. (4.10), S. 68)

$$\boxed{\begin{aligned} z_*'(t) &= A_*(t)\, z_*(t), \\ z_*(0) - z_*(T) &= 0. \end{aligned}} \tag{4.14}$$

Falls x_* eine nicht-konstante Lösung ist, so ist $x_*' \not\equiv 0$, d.h. z_* ist eine nichttriviale Lösung von (4.14). Dann läßt sich der Satz 4.7 so nicht anwenden.

Der Vektor $z_*(0) = x_*'(0) \neq 0$ ist Eigenvektor von $X_*(T)$ zum Eigenwert 1, denn aus

$$z_*(t) = X_*(t)\, z_*(0)$$

und der Periodizität $z_*(0) = z_*(T)$ folgt

$$z_*(T) = X_*(T)\, z_*(0) = X_*(T)\, z_*(T).$$

Dann ist die Matrix

$$M_* = I - X_*(T)$$

singulär ($M_* z_*(0) = 0$, $z_*(0) \neq 0$), d.h. es können mehrere periodische Lösungen existieren.

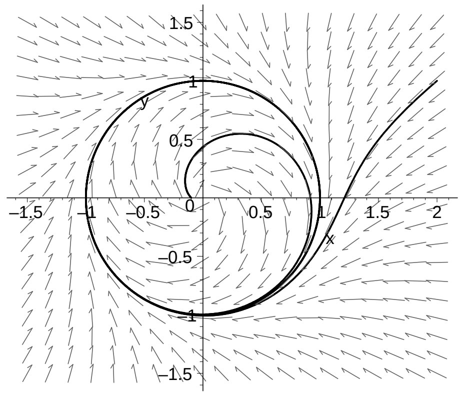
Definition 4.9 (periodisch, Grenzzykel, Attraktor). Eine Lösung einer autonomen DGL heißt *periodisch*, falls es ein $T > 0$ gibt, mit dem sie T-periodisch ist ($x_*(t) = x_*(t+T)$, $t \in \mathbb{R}$).

Eine nicht-konstante periodische Lösung (bzw. deren Orbit) heißt *Grenzzykel*, wenn in einer Umgebung dieses Orbits keine weitere periodische Lösung enthalten ist.

Ein stabiler Grenzzykel heißt *Attraktor*.

Beispiel (Grenzzykel). Für $m = 2$ betrachte die DGL

$$\begin{aligned} x_1' &= x_2 + x_1\left(1 - x_1^2 - x_2^2\right) \\ x_2' &= -x_1 + x_2\left(1 - x_1^2 - x_2^2\right) \end{aligned}$$



Lösungen der DGL für Startwerte $x_0 \in \left\{(1,0), (2,1), \left(-\frac{1}{10}, 0\right)\right\}$.

Die Funktion

$$x_*(t) = (\sin t, \cos t)$$

ist eine 2π-periodische Lösung. Weiterhin ist

$$\begin{aligned} \frac{d}{dt}|x(t)|_2^2 &= 2\langle x'(t), x(t)\rangle_2 \\ &= 2\left\langle \begin{pmatrix} x_2 + x_1\left(1 - x_1^2 - x_2^2\right) \\ -x_1 + x_2\left(1 - x_1^2 - x_2^2\right) \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\rangle \\ &= 2\left(x_1^2\left(1 - x_1^2 - x_2^2\right) + x_2^2\left(1 - x_1^2 - x_2^2\right)\right) \\ &= 2|x(t)|_2^2\left(1 - |x(t)|_2^2\right). \end{aligned}$$

D.h. für Lösungen $x(t)$ mit $|x(t)|_2 < 1$ ist $\frac{d}{dt}|x(t)|_2^2 > 0$, d.h. $|x(t)|$ wächst. Ist $|x(t)|_2 > 1$, so ist $\frac{d}{dt}|x(t)|_2^2 < 0$, d.h. $|x(t)|$ fällt.

4.4.1 Modifikation der Aufgabenstellung

Falls die Periode T im Voraus nicht bekannt ist, so haben wir folgende Aufgabe zu betrachten:

$$\begin{aligned} x'(t) &= f(x(t)), \qquad t \in [0, T], \\ x(0) &= x(T), \qquad T \text{ unbekannt.} \end{aligned}$$

Dazu transformieren wir $[0, T] \to [0, 1]$ mittels $s = \frac{t}{T}$ und definieren

$$\bar{x}(s) := x(sT) = x(t).$$

Aus

$$\frac{d}{ds}\bar{x}(s) = \frac{d}{ds}x(sT) = T \cdot \frac{dx}{dt}(sT) = T \cdot f(x(st)) = T \cdot f(\bar{x}(s))$$

erhalten wir dann die Aufgabe

$$\begin{aligned} \frac{d}{ds}\bar{x}(s) &= T \cdot f(\bar{x}(s)), \\ \bar{x}(0) &= \bar{x}(1). \end{aligned}$$

Wir fügen nun die Funktion

$$\bar{x}_{m+1}(s) := T$$

mit $\frac{d}{ds}\bar{x}_{m+1}(s) = 0$ als neue unbekannte Funktion hinzu, so daß wir die Aufgabe

$$\begin{aligned} \frac{d}{ds}\bar{x}(s) &= \bar{x}_{m+1}(s) \cdot f(\bar{x}(s)), \\ \frac{d}{ds}\bar{x}_{m+1}(s) &= 0, \\ \bar{x}(0) - \bar{x}(1) &= 0 \end{aligned}$$

erhalten. Da nun eine Randbedingung fehlt, fügen wir noch die zusätzliche Anfangsbedingung

$$\bar{x}_{i_*}(0) = \alpha$$

mit $\alpha \neq 0$ für ein $i_* \in \{1, \ldots, m\}$ hinzu („*Phasennormalisierung*“).

Kapitel 5

Schießverfahren

Wir betrachten Randwertaufgaben der Form

$$\boxed{\begin{array}{rcll} x'(t) & = & f(x(t), t), & t \in [t_0, T] \\ r(x(t_0), x(T)) & = & 0 & \end{array}} \tag{5.1}$$

Sei $x_* \in C^1([t_0, T], \mathbb{R}^m)$ eine Lösung und sei M_* regulär.

Nach Satz 4.7 (Seite 68) ist die „*Schießabbildung*“ S,

$$\boxed{S(x_0) := r(x_0, x(T, x_0))}$$

in einer Umgebung von $x_*(t_0)$ definiert und stetig differenzierbar. Es gilt $S'(x_*(t_0)) = M_*$.

Um das Randwertproblem (5.1) zu lösen, müssen wir x_0 so bestimmen, daß $(x_0, x(T, x_0))$ die Randbedingung erfüllt, d.h. wir müssen die Gleichung

$$\boxed{S(x_0) = 0}$$

lösen. Für jeden Wert x_0 kann man $S(x_0)$ berechnen. Man hat dazu den Wert $x(T, x_0)$ zu bestimmen. Eine Berechnung von $S(x_0)$ läuft also auf die Lösung eines Anfangswertproblems hinaus. Dies kann man z.B. mit den Methoden aus Kapitel 2 realisieren.

Zur Berechnung der Nullstelle von $S(\cdot)$ kann man im Prinzip beliebige Methoden verwenden, die eine Nullstelle einer Funktion bestimmen. Kennt man z.B. Werte $x_0^{(1)}$ und $x_0^{(2)}$ mit $S\left(x_0^{(1)}\right) < 0$ und $S\left(x_0^{(2)}\right) > 0$, so kann man eine Nullstelle durch ein einfaches Bisektionsverfahren berechnen. Man sollte aber beachten, daß eine Auswertung der Funktion S immer eine Lösung einer Anfangswertaufgabe bedeutet. Da S eine stetig differenzierbare Funktion ist, kann man auch das Newton-Verfahren (oder Quasi-Newton-Verfahren) zur Bestimmung der Nullstelle benutzen. Da dieses Verfahren i.a. eine höhere Konvergenzrate besitzt, sollte man erwarten, daß dann auch weniger Funktionsauswertungen benötigt werden, um ein Ergebnis mit ähnlicher Genauigkeit wie beim Bisektionsverfahren zu erhalten.

Man könnte nun meinen, daß das Problem des Lösens der Randwertaufgabe durch Bestimmung des Startwertes x_0 gelöst ist, da man den Wert $x(t, x_0)$ der Lösung an jeder anderen Stelle t durch Behandlung des Anfangswertproblems $x'(t) = f(x(t), t)$, $x(t_0) = x_0$, bestimmen kann. In der Praxis treten dabei aber häufig erhebliche Ungenauigkeiten auf, wenn die Lösung $x(t) = x(t, x_0)$ sehr empfindlich von x_0 abhängt. Eine approximative Berechnung von x_0 reicht dann häufig nicht aus, wie das nächste Beispiel zeigt:

Beispiel ([SB90, §7.3.4]). Wir betrachten die Differentialgleichung

$$x'(t) = \begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix} x(t), \qquad t \in [0,10].$$

Die Eigenwerte der Matrix $\begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix}$ sind $\lambda_1 = -10$ und $\lambda_2 = 11$. Die Lösung $x(t,z)$ der Differentialgleichung, die der Anfangsbedingung $x(0,z) = z$ genügt, hat dann die Form

$$x(t,z) = \frac{11z_1 - z_2}{21} e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + \frac{10z_1 + z_2}{21} e^{11t} \begin{pmatrix} 1 \\ 10 \end{pmatrix}.$$

Wir wollen nun die Lösung $x_*(t)$ bestimmen, die den Randwertbedingungen

$$x_{*1}(0) = 1, \qquad x_{*1}(10) = 1$$

genügt. Die exakte Lösung ist

$$z_* := x_*(0) = \begin{pmatrix} 1 \\ -10 + 21\frac{1-e^{-100}}{e^{110}-e^{-100}} \end{pmatrix}.$$

Können wir z_* aber nur mit z.B: 10-stelliger Genauigkeit berechnen, so erhalten wir den Näherungswert

$$\bar{z} = \begin{pmatrix} 1 \\ -10 + 10^{-9} \end{pmatrix}.$$

Die exakte Lösung $x(t,\bar{z})$ zum Anfangswert $\bar{z}$ liefert nun aber

$$\begin{aligned} x_1(10,\bar{z}) &= \frac{21 - 10^{-9}}{21} e^{-100} + \frac{10^{-9}}{21} e^{110} \\ &\approx \frac{10^{-9}}{21} e^{110} \\ &\approx 2.8 \cdot 10^{37} \neq 1 \end{aligned}$$

Das Beispiel zeigt, daß die Berechnung des Startwertes x_0 selbst mit hoher Genauigkeit nicht garantiert, daß sich die Werte $x(t,x_0)$ mit ähnlicher Genauigkeit bestimmen lassen. Für Systeme, die die einseitige Lipschitzbedingung (1.3) erfüllen, haben wir in Satz 1.6 (Seite 8) die Abschätzung

$$|x(t,x_0) - x(t,\bar{x}_0)| \leq e^{\gamma(t-t_0)} |x_0 - \bar{x}_0|$$

erhalten, d.h. der Einfluß fehlerhafter Anfangsdaten kann exponentiell mit t wachsen.

Diese Abschätzung zeigt aber auch, daß der Einfluß ungenauer Anfangsdaten durch Verkleinerung des Intervalls $[t_0,t]$ beliebig klein gemacht werden kann. Dies führt zur Idee der Mehrzielmethode:

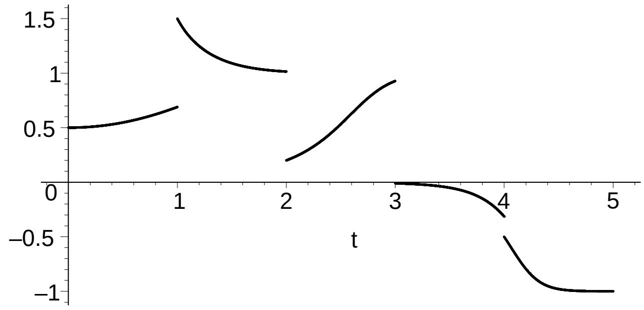
5.1 Mehrzielmethode

Wir zerlegen $[t_0,T]$ in Teilintervalle, $t_0 < t_1 < \ldots < t_N = T$.

Bei der Mehrzielmethode werden die Werte $x(t_j)$ der exakten Lösung $x(t)$ eines Randwertproblems 5.1 an den Stellen $t_0,\ldots,t_n$ gleichzeitig iterativ berechnet. Wir betrachten dazu die Anfangswertaufgaben

$$\boxed{\begin{aligned} x'(t) &= f(x(t),t) \qquad t \in [t_{j-1},t_j], \\ x(t_{j-1}) &= z_{j-1}, \end{aligned} \qquad j = 1,\ldots,N.}$$

und bezeichnen deren Lösungen mit $x(t;z_{j-1},t_{j-1})$, $t \in [t_{j-1},t_j]$.



Die Aufgabe besteht nun darin, die Vektoren $z_0, \ldots, z_{N-1}$ so zu bestimmen, daß die aus den $x(t; z_{j-1}, t_{j-1})$ stückweise zusammengesetze Funktion $x(t)$,

$$\begin{aligned} x(t) &:= x(t, z_{j-1}, t_{j-1}) \qquad \text{für } t \in [t_{j-1}, t_j), \quad j = 1, \ldots, N, \\ x(t_N) &:= x(t_N, z_{N-1}, t_{N-1}), \end{aligned}$$

stetig ist, also eine Lösung der Differentialgleichung $x'(t) = f(x(t), t)$, $t \in [t_0, T]$, darstellt und darüber hinaus die Randbedingung $r(x(t_0), x(T)) = 0$ erfüllt.

Mit den Bedingungen (Stetigkeitsbedingungen)

$$\boxed{x(t_j, z_{j-1}, t_{j-1}) = z_j, \qquad j = 1, \ldots, N-1}$$

werden die Segmente zu einer Lösung der DGL auf $[t_0, T]$ zusammengeklebt.

Zusätzlich benötigen wir die Randbedingung

$$\boxed{r(z_0, x(t_N, z_{N-1}, t_{N-1})) = 0.}$$

Dies ergibt ein nichtlineares Gleichungssystem im $\mathbb{R}^{mN}$ bzgl. $z_0, \ldots, z_{N-1} \in \mathbb{R}^m$:

$$\boxed{S(z) = 0}$$

mit Schießabbildung

$$\boxed{S(z) := \begin{pmatrix} x(t_1, z_0, t_0) - z_1 \\ \vdots \\ x(t_{N-1}, z_{N-2}, t_{N-2}) - z_{N-1} \\ r(z_0, x(t_N, z_{N-1}, t_{N-1})) \end{pmatrix}, \qquad z = \begin{pmatrix} z_0 \\ \vdots \\ z_{N-1} \end{pmatrix}.}$$

Sei x_* Lösung der RWA (5.1) und sei M_* regulär. Sei

$$z_* := \begin{pmatrix} z_{*0} \\ \vdots \\ z_{*N-1} \end{pmatrix} := \begin{pmatrix} x_*(t_0) \\ \vdots \\ x_*(t_{N-1}) \end{pmatrix}.$$

Nach Satz 4.7 ist $S(z)$ dann wohldefiniert für $z_j \in B(z_{*j}, \delta)$, $j = 0, \ldots, N-1$, und stetig differenzierbar.

Um das Newton-Verfahren anwenden zu können, ist die Regularität von $S'(z_*)$ wünschenswert. Definiere dazu

$$\begin{aligned} X_j(t) &:= \frac{\partial}{\partial z_j} x(t; z_j, t_j), \qquad j = 0, \ldots, N-1, \\ X_{*j}(t) &:= \left. \frac{\partial}{\partial z_j} x(t; z_j, t_j) \right|_{z_j = z_{*j}}, \qquad j = 0, \ldots, N-1. \end{aligned}$$

Wie wir im Beweis von Satz 4.7, Teil 4, gesehen haben, ist

$$X_j'(t) = f_x(x(t,z_j,t_j),t)\,X_j(t), \quad X_j(t_j) = I, \qquad j = 0,\dots,N-1,$$

und

$$X_{*j}'(t) = A_*(t)\,X_{*j}(t), \quad X_{*j}(t_j) = I, \qquad j = 0,\dots,N-1.$$

Dann ist $S'(z) =$

$$\begin{pmatrix} \frac{\partial}{\partial z_0}(x(t_1;z_0,t_0)-z_1) & \frac{\partial}{\partial z_1}(x(t_1;z_0,t_0)-z_1) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\partial}{\partial z_{N-2}}(\cdots) & \frac{\partial}{\partial z_{N-1}}(x(t_{N-1};z_{N-2},t_{N-2})-z_{N-1}) \\ \frac{\partial}{\partial z_0}r(z_0,x(t_N,z_{N-1},t_{N-1})) & 0 & \cdots & 0 & \frac{\partial}{\partial z_{N-1}}r(z_0,x(t_N,z_{N-1},t_{N-1})) \end{pmatrix}$$

$$= \begin{pmatrix} X_0(t_1) & -I & 0 & \cdots & 0 \\ 0 & X_1(t_2) & -I & 0 & \vdots \\ \vdots & \ddots & \ddots & & 0 \\ 0 & & \ddots & X_{N-2}(t_{N-1}) & -I \\ r_1'(z_0,x(t_N,z_{N-1},t_{N-1})) & 0 & \cdots & 0 & r_2'(z_0,x(t_N,z_{N-1},t_{N-1}))\,X_{N-1}(t_N) \end{pmatrix}$$

und speziell für $z = z_*$:

$$\begin{aligned} S(z_*) &= 0, \\ S'(z_*) &= \begin{pmatrix} X_{*0}(t_1) & -I & 0 & \cdots & 0 \\ 0 & X_{*1}(t_2) & -I & 0 & \vdots \\ \vdots & \ddots & \ddots & & 0 \\ 0 & & \ddots & X_{*N-2}(t_{N-1}) & -I \\ B_{*0} & 0 & \cdots & 0 & B_{*T}X_{*N-1}(t_N) \end{pmatrix}. \end{aligned}$$

Sei $y = (y_0,\dots,y_{N-1})^T \in \mathbb{R}^{mN}$ mit

$$S'(z_*)\,y = 0.$$

Für die Regularität von $S'(z_*)$ müssen wir $y = 0$ zeigen. Aus $S'(z_*)\,y = 0$ folgt

$$\begin{aligned} X_{*0}(t_1)\,y_0 &= y_1, \\ &\vdots \\ X_{*N-3}(t_{N-2})\,y_{N-3} &= y_{N-2}, \\ X_{*N-2}(t_{N-1})\,y_{N-2} &= y_{N-1}, \\ B_{*0}y_0 + B_{*T}X_{*N-1}(t_N)\,y_{N-1} &= 0, \end{aligned}$$

und die letzte Zeile wird zu

$$B_{*0}y_0 + B_{*T}X_{*N-1}(t_N)\cdots X_{*0}(t_1)\,y_0 = 0. \tag{5.2}$$

Es galt: $X_*'(t) = A_*(t)\,X_*(t)$, $t \in [t_0,T]$, $X_*(t_0) = I$. Nach der Eindeutigkeit der Lösung von Anfangswertaufgaben folgt

$$X_*(t) \equiv X_{*0}(t),$$

also speziell für $t = t_1$

$$X_{*0}(t_1) = X_*(t_1).$$

$X_{*1}(t)\, X_{*0}(t_1)$ ist Fundamentalmatrix und

$$\underbrace{X_{*1}(t_1)}_{I} X_{*0}(t_1) = X_{*0}(t_1) = X_*(t_1).$$

Dann folgt wegen der Eindeutigkeit der Lösung von AWAn, daß

$$X_{*1}(t)\, X_{*0}(t_1) \equiv X_*(t),$$

also speziell für $t = t_2$

$$X_{*1}(t_2)\, X_{*0}(t_1) = X_*(t_2).$$

Setzen wir dies fort, so erhalten wir

$$X_{*N-1}(t)\, X_{*N-2}(t_{N-1}) \cdots X_{*0}(t_1) \equiv X_*(t)$$

und speziell für $t = t_N$

$$X_{*N-1}(t_N) \cdots X_{*0}(t_1) = X_*(t_N).$$

Einsetzen in (5.2) ergibt

$$0 = B_{*0} y_0 + B_{*T} X_*(t_N)\, y_0 = M_* y_0.$$

Aus der Regularität von M_* folgt dann $y_0 = 0$ und zurückeinsetzen ergibt dann $y_1 = 0$, $y_2 = 0$, $\ldots$, $y_{N-1} = 0$, also $y = 0$. D.h. $S'(z_*)$ ist regulär.

Satz 5.1. *Sei $x_* \in C^1([t_0, T], \mathbb{R}^m)$ Lösung der Randwertaufgabe (5.1), M_* sei regulär. Es gelte $\|B_{*0}\| \leq 1$ und $\|B_{*T}\| \leq 1$. Seien κ_1 und κ_2 die Konditionszahlen der in x_* linearisierten RWA.*

Dann ist die Abbildung S der Mehrzielmethode zur Zerlegung $t_0 < t_1 < \ldots < t_N = T$ auf einer Umgebung von $z_* = (x_*(t_0), \ldots, x_*(t_{N-1}))^T \in \mathbb{R}^{mN}$ definiert und dort stetig differenzierbar. Es gilt $S(z_*) = 0$, $S'(z_*)$ ist regulär und

$$\boxed{\operatorname{cond}(S'(z_*)) \leq (\kappa_1 + (N-1)\,\kappa_2) \cdot \left(1 + \max_{j=1,\ldots,N} \|X_{*j-1}(t_j)\|\right)}$$

bzgl. der Block-Zeilensummen-Norm auf $L\left(\mathbb{R}^{mN}\right)$.

Beweis. Die Wohldefiniertheit und stetige Differenzierbarkeit von S in einer Umgebung von z_* folgen aus Satz 4.7. Die Regularität von $S'(z_*)$ haben wir gerade gezeigt. Es bleibt die Abschätzung der Kondition zu zeigen. Die Kondition ist definiert als

$$\operatorname{cond}(S'(z_*)) = \|S'(z_*)\| \left\| S'(z_*)^{-1} \right\|.$$

Es ist

$$S'(z_*)^{-1} = \begin{pmatrix} \mathcal{G}_*(t_0, t_1) & \cdots & \mathcal{G}_*(t_0, t_{N-1}) & X_*(t_0)\, M_*^{-1} \\ \vdots & & \vdots & \vdots \\ \mathcal{G}_*(t_{N-1}, t_1) & \cdots & \mathcal{G}_*(t_{N-1}, t_{N-1}) & X_*(t_{N-1})\, M_*^{-1} \end{pmatrix}$$

mit der Greenschen Funktion (vgl. (4.5), Seite 62)

$$\mathcal{G}_*(t,s) = \begin{cases} X_*(t)\, M_*^{-1} B_{*0} X_*(t_0)\, X_*(s)^{-1} & s \leq t, \\ -X_*(t)\, M_*^{-1} B_{*T} X_*(T)\, X_*(s)^{-1} & s > t, \end{cases}$$

denn

$$\begin{aligned}
\left(S'(z_*)\,S'(z_*)^{-1}\right)_{11} &= -X_{*0}(t_1)\,\mathcal{G}_*(t_0,t_1) + \mathcal{G}_*(t_1,t_1) \\
&= \underbrace{X_{*0}(t_1)\,X_*(t_0)}_{X_*(t_1)}\,M_*^{-1}B_{*T}X_*(T)\,X_*(t_1)^{-1} + X_*(t_1)\,M_*^{-1}B_{*0}X_*(t_0)\,X_*(t_1)^{-1} \\
&= X_*(t_1)\,M_*^{-1}\left(B_{*T}X_*(T) + B_{*0}X_*(t_0)\right)X_*(t_1)^{-1} \\
&= I \\
\left(S'(z_*)\,S'(z_*)^{-1}\right)_{12} &= -X_{*0}(t_1)\,\mathcal{G}_*(t_0,t_2) + \mathcal{G}_*(t_1,t_2) \\
&= X_{*0}(t_1)\,X_*(t_0)\,M_*^{-1}B_{*T}X_*(T)\,X_*(t_2)^{-1} - X_*(t_1)\,M_*^{-1}B_{*T}X_*(T)\,X_*(t_2)^{-1} \\
&= 0
\end{aligned}$$

usw. Nach Definition von κ_1 und κ_2 ((4.6), Seite 62) ist

$$\begin{aligned}
\left\|X_*(t)\,M_*^{-1}\right\| &\leq \kappa_1 \qquad t\in[t_0,T] \\
\left\|\mathcal{G}_*(t,s)\right\| &\leq \kappa_2 \qquad t,s\in[t_0,T]
\end{aligned}$$

Damit ist also

$$\left\|S'(z_*)^{-1}\right\| = \max_{i=0,\ldots,N-1}\left(\sum_{j=1}^{N-1}\left\|\mathcal{G}_*(t_i,t_j)\right\| + \left\|X_*(t_i)\,M_*^{-1}\right\|\right) \leq (N-1)\,\kappa_2 + \kappa_1$$

und

$$\begin{aligned}
\left\|S'(z_*)\right\| &= \max\left\{\max_{i=0,\ldots,N-2}\left(1+\left\|X_{*i}(t_{i+1})\right\|\right), \underbrace{\left\|B_{*0}\right\|}_{\leq 1} + \underbrace{\left\|B_{*T}X_{*N-1}(t_N)\right\|}_{\leq\left\|X_{*N-1}(t_N)\right\|}\right\} \\
&\leq \max_{i=0,\ldots,N-1}\left(1+\left\|X_{*i}(t_{i+1})\right\|\right).
\end{aligned}$$

□

Bemerkung. Bei skalierten Randbedingungen gilt $\kappa_1 \leq 2\kappa_2$ (Lemma 4.4, Seite 63). Dann ist

$$\kappa_1 + (N-1)\,\kappa_2 \leq (N+1)\,\kappa_2.$$

Bemerkung. Zur Abschätzung von $\left\|X_{*j-1}(t_j)\right\|$ definiere

$$\begin{aligned}
\mu_j &:= \max_{t\in[t_j,t_{j+1}]}\left\|A_*(t)\right\| \qquad j=0,\ldots,N-1, \\
\mu &:= \max_{t\in[t_0,T]}\left\|A_*(t)\right\| = \max\left(\mu_0,\ldots,\mu_{N-1}\right),
\end{aligned}$$

d.h. $\mu_j \leq \mu$. Wegen

$$\begin{aligned}
X'_{*j-1}(t) &= A_*(t)\,X_{*j-1}(t) \qquad t\in[t_{j-1},t_j] \\
X_{*j-1}(t_{j-1}) &= I
\end{aligned}$$

ist dann

$$\left\|X_{*j-1}(t_j)\right\| \leq e^{\mu_{j-1}(t_j-t_{j-1})}.$$

Im Spezialfall $h=\frac{T-t_0}{N}$, $t_j=t_0+jh$, $\mu_j\leq\mu$, folgt dann

$$\operatorname{cond}\left(S'(z_*)\right) \leq \left(\kappa_1+(N-1)\,\kappa_2\right)\left(1+e^{\mu h}\right) \leq \left(\kappa_1+\frac{T-t_0}{h}\kappa_2\right)\left(1+e^{\mu h}\right).$$

Im Idealfall möchte man $\left\|X_{*j}(t_{j+1})\right\|\leq\kappa$, $j=0,\ldots,N-1$, mit $\kappa\approx 10$ oder $\kappa\approx 100$ erreichen.

Beispiel (AMR, S. 151). Sei $m = 2$ und wir betrachten das Randwertproblem

$$\begin{aligned} x'(t) &= \begin{pmatrix} 1 - 19\cos(2t) & 1 + 19\sin(2t) \\ -1 + 19\sin(2t) & 1 + 19\cos(2t) \end{pmatrix} x(t) + q(t) \qquad t \in [0, \pi], \\ x(0) + x(\pi) &= \begin{pmatrix} 1 + e^{\pi} \\ 1 + e^{\pi} \end{pmatrix}, \end{aligned}$$

mit „passender" Funktion q. Dann ist

$$B_0 = I, \quad B_T = I \qquad \kappa_1, \kappa_2 \approx 2.$$

Die exakte Lösung ist

$$x_*(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}$$

und die Fundamentalmatrix ist

$$X(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} e^{-18t} & 0 \\ 0 & e^{20t} \end{pmatrix} \quad \text{mit } X(0) = I,\ X(\pi) = I.$$

Dann ist

$$M = I + \begin{pmatrix} e^{-18\pi} & 0 \\ 0 & e^{20\pi} \end{pmatrix} = \begin{pmatrix} 1 + e^{-18\pi} & 0 \\ 0 & 1 + e^{20\pi} \end{pmatrix},$$

so daß

$$\operatorname{cond}(M) \approx e^{20\pi} \approx 1.9 \cdot 10^{27}.$$

Für die Konditionszahlen von $S'(z_*)$ erhält man

N	$\left\Vert S'(z_*)^{-1} \right\Vert$	$\operatorname{cond}(S'(z_*))$
1	1	$1.9 \cdot 10^{27}$
3	1	$1.2 \cdot 10^{9}$
9	1	$1.1 \cdot 10^{3}$
15	1	$6.9 \cdot 10^{1}$
24	1.1	$1.6 \cdot 10^{1}$

Kapitel 6

Kollokationsverfahren

Wir stellen die Randwertaufgabe

$$\boxed{\begin{array}{rcll} x'(t) & = & f(x(t), t), & t \in [t_0, T], \\ r(x(t_0), x(T)) & = & 0, & \end{array}} \tag{6.1}$$

wieder in der Form

$$\boxed{\mathcal{F}x = 0}$$

für einen Operator

$$\mathcal{F} : X \to Y$$

mit

$$X := C^1([t_0, T], \mathbb{R}^m), \qquad Y := V \times \mathbb{R}^m, \quad V := C([t_0, T], \mathbb{R}^m)$$

und

$$\mathcal{F}x := \begin{pmatrix} x'(\cdot) - f(x(\cdot), \cdot) \\ r(x(t_0), x(T)) \end{pmatrix}$$

dar. Die Idee der *Kollokationsverfahren* ist es nun, die Räume X und Y durch endlichdimensionale Räume $X_{(n)}$ und $Y_{(n)}$ zu ersetzen (Diskretisierung, Finitisierung), und für einen entsprechenden Operator

$$\mathcal{F}_{(n)} : X_{(n)} \to Y_{(n)}$$

die Gleichung

$$\boxed{\mathcal{F}_{(n)}u = 0}$$

zu lösen.

Dazu wählen wir zunächst Zerlegungen

$$\pi_{(n)} : \quad t_0 < t_1^{(n)} < \ldots < t_{N^{(n)}}^{(n)} = T, \qquad n \in \mathbb{N}$$

von $[0, T]$ in $N^{(n)}$ Teilintervalle. Sei

$$h_i^{(n)} := t_i^{(n)} - t_{i-1}^{(n)}, \quad i = 1, \ldots, N^{(n)}, \qquad h^{(n)} := \max_{j=1,\ldots,N^{(n)}} h_j^{(n)},$$

und

$$t_{j-\frac{1}{2}}^{(n)} := t_{j-1}^{(n)} + \frac{1}{2}h_j^{(n)}.$$

Sei von nun an eine Zerlegung $\pi_{(n)}$ fixiert. Wir verzichten im folgenden (weitestgehend) auf den Index „n".

Als Diskretisierung wählen wir nun den Raum $X_{(n)}$ der *kubischen Splinefunktionen* bzgl. $\pi_{(n)}$:

$$\boxed{X_{(n)} := \left\{ u \in X \,\middle|\, u \text{ ist auf jedem Teilintervall von } \pi_{(n)} \text{ ein Polynom mit Grad} \leq 3 \right\}}$$

$X_{(n)} \subset C^1\left([t_0, T], \mathbb{R}^n\right)$ ist ein endlich-dimensionaler Teilraum.

Durch die Vorgaben

$$u(t_i) = u_i, \quad u'(t_i) = u_i', \qquad i = 0, \ldots, N,$$

ist ein $u \in X_{(n)}$ eindeutig bestimmt. Es hat auf dem Teilintervall $[t_{i-1}, t_i]$ folgende Darstellung:

$$\boxed{\begin{array}{rcl} u(t) & = & u_{i-1} + (t - t_{i-1}) \frac{u_i - u_{i-1}}{h_i} + (t - t_{i-1})(t - t_i) v_i + (t - t_{i-1})^2 (t - t_i) w_i \\ v_i & = & \frac{1}{h_i}\left(\frac{u_i - u_{i-1}}{h_i} - u_{i-1}'\right) \\ w_i & = & \frac{1}{h_i^2}\left(u_{i-1}' - 2\frac{u_i - u_{i-1}}{h_i} + u_i'\right) \end{array}} \tag{6.2}$$

Die Darstellung für v_i und w_i ergibt sich dabei aus den Bedingungen $u'(t_i) = u_i'$, $i = 1, \ldots, N$:

Aus der Bedingung $u_{i-1}' = u'(t_{i-1})$ folgt

$$u_{i-1}' = \frac{u_i - u_{i-1}}{h_i} - h_i v_i, \quad \text{d.h. } v_i = \frac{1}{h_i}\left(\frac{u_i - u_{i-1}}{h_i} - u_{i-1}'\right).$$

Aus der Bedingung $u_i' = u'(t_i)$ folgt

$$\begin{array}{rcl} u_i' & = & \frac{u_i - u_{i-1}}{h_i} + h_i v_i + h_i^2 w_i \\ \text{d.h. } w_i & = & \frac{1}{h_i^2}\left(u_i' - \frac{u_i - u_{i-1}}{h_i} - \frac{u_i - u_{i-1}}{h_i} + u_{i-1}'\right) = \frac{1}{h_i^2}\left(u_i' - 2\frac{u_i - u_{i-1}}{h_i} + u_{i-1}'\right). \end{array}$$

Zu einer Funktion $x \in X$ ist eine Funktion $u \in X_{(n)}$ durch $u_i := x(t_i)$, $u_i' := x'(t_i)$, $i = 0, \ldots, N$, eindeutig bestimmt. Wir definieren daher die Abbildung

$$\boxed{R_{(n)} : X \to X_{(n)},}$$

indem wir einer Funktion $x \in X$ die Funktion $R_{(n)} x := u$ zuweisen, für die

$$\begin{array}{rcll} u(t_i) & = & x(t_i), & i = 0, \ldots, N, \\ u'(t_i) & = & x'(t_i), & i = 0, \ldots, N, \end{array}$$

gilt. Die Restriktionsabbildung $R_{(n)}$ ist linear, beschränkt und ein Projektor ($R_{(n)}^2 = R_{(n)}$).

Zur Lösung der Randwertaufgabe (6.1) war eine Funktion $x_* \in X$ mit $\mathcal{F} x_* = 0$ gesucht. Bzgl. der Diskretisierung $X_{(n)}$ suchen wir nun eine Funktion $u \in X_{(n)}$ mit

$$\boxed{\begin{array}{rcll} u'(t_i) & = & f(u(t_i), t_i) & i = 0, \ldots, N, \\ u'\left(t_{i-\frac{1}{2}}\right) & = & f\left(u\left(t_{i-\frac{1}{2}}\right), t_{i-\frac{1}{2}}\right) & i = 1, \ldots, N, \\ r(u(t_0), u(t_N)) & = & 0, & \end{array}} \tag{6.3}$$

d.h. wir suchen eine Funktion $u \in X_{(n)} \subset X$, die die Randbedingung $r\left(u\left(t_0\right), u\left(t_N\right)\right) = 0$ und die Differentialgleichung $u'(t) = f\left(u(t), t\right)$ in den Diskretisierungspunkten $t_0, t_{\frac{1}{2}}, t_1, \ldots, t_N$ erfüllt.

Aus (6.2) erhalten wir

$$\begin{aligned}
u'\left(t_{i-\frac{1}{2}}\right) &= \frac{u_i - u_{i-1}}{h_i} + 0 + \left(-2\frac{h_i^2}{4} + \frac{h_i^2}{4}\right) w_i \\
&= \frac{u_i - u_{i-1}}{h_i} - \frac{h_i^2}{4} \cdot \frac{1}{h_i^2}\left(u'_{i-1} - 2\frac{u_i - u_{i-1}}{h_i} + u'_i\right) \\
&= \frac{3}{2}\left(\frac{u_i - u_{i-1}}{h_i}\right) - \frac{1}{4}\left(u'_i + u'_{i-1}\right), \qquad (6.4) \\
u\left(t_{i-\frac{1}{2}}\right) &= \frac{1}{2}\left(u_{i-1} + u_i\right) - \frac{1}{8} h_i \left(u'_i - u'_{i-1}\right). \qquad (6.5)
\end{aligned}$$

Damit erhalten wir aus (6.3) das Gleichungssystem

$$\begin{aligned}
u'_i &= f\left(u_i, t_i\right) & i &= 0, \ldots, N, \\
\frac{3}{2}\frac{u_i - u_{i-1}}{h_i} - \frac{1}{4}\left(u'_i + u'_{i-1}\right) &= f\left(\frac{1}{2}\left(u_{i-1} + u_i\right) - \frac{1}{8} h_i \left(u'_i - u'_{i-1}\right), t_{i-\frac{1}{2}}\right) & i &= 1, \ldots, N, \\
r\left(u_0, u_N\right) &= 0,
\end{aligned}$$

aus $(2N+2)\,m$ Gleichungen in den Unbekannten u_i, u'_i, $i = 0, \ldots, N$.

Ersetzen wir nun noch $u'_i = f\left(u_i, t_i\right)$, so erhalten wir das Gleichungssystem

$$\boxed{\begin{aligned}
\frac{u_i - u_{i-1}}{h_i} &= \frac{2}{3} f\left(\frac{1}{2}\left(u_{i-1} + u_i\right) - \frac{1}{8} h_i \left(f\left(u_i, t_i\right) - f\left(u_{i-1}, t_{i-1}\right)\right), t_{i-\frac{1}{2}}\right) \\
&\quad + \frac{1}{6} f\left(u_{i-1}, t_{i-1}\right) + \frac{1}{6} f\left(u_i, t_i\right) \qquad i = 1, \ldots, N, \\
r\left(u_0, u_N\right) &= 0,
\end{aligned}} \qquad (6.6)$$

aus $(N+1)\,m$ Gleichungen in den Unbekannten $u_0, \ldots, u_N$.

Um den Operator $\mathcal{F}_{(n)}$ definieren zu können, benötigen wir noch eine passende Diskretisierung $V_{(n)}$ von V. Für $u \in X_{(n)}$ ist offensichtlich $u' \in V_{(n)}$ für

$$\boxed{V_n := \left\{ v \in V \,\middle|\, v \text{ ist auf jedem Teilintervall von } \pi_{(n)} \text{ ein Polynom mit Grad} \leq 2 \right\}.}$$

Ein Element $v \in V_{(n)}$ ist eindeutig bestimmt durch die Werte

$$\begin{aligned}
v\left(t_i\right) &= v_i, & i &= 0, \ldots, N, \\
v\left(t_{i-\frac{1}{2}}\right) &= v_{i-\frac{1}{2}}, & i &= 1, \ldots, N.
\end{aligned}$$

Für $t \in \left[t_{i-1}, t_i\right]$ besitzt v die Darstellung

$$\boxed{v(t) = v_{i-1} + \left(t - t_{i-1}\right)\frac{v_{i-\frac{1}{2}} - v_{i-1}}{\frac{1}{2}h_i} - \left(t - t_{i-1}\right)\left(t - t_{i-\frac{1}{2}}\right)\left(\frac{v_{i-\frac{1}{2}} - v_{i-1}}{\frac{1}{2}h_i^2} - \frac{v_i - v_{i-\frac{1}{2}}}{\frac{1}{2}h_i^2}\right).} \qquad (6.7)$$

Wir definieren daher die Restriktionsabbildung

$$\boxed{R_{(n)}^V : V \to V_{(n)}}$$

indem wir einer Funktion $v \in V$ die Funktion $R^V_{(n)} v$ zuordnen, die durch die Bedingungen

$$\begin{aligned} \left(R^V_{(n)} v\right)(t_i) &= v(t_i), \qquad i = 0, \ldots, N, \\ \left(R^V_{(n)} v\right)\left(t_{i-\frac{1}{2}}\right) &= v\left(t_{i-\frac{1}{2}}\right), \qquad i = 1, \ldots, N, \end{aligned}$$

eindeutig bestimmt ist. $R^V_{(n)}$ ist ebenfalls ein linearer, beschränkter Projektor.

Für $v \in V_{(n)}$ gilt: $v(t) = 0 \Leftrightarrow v_i = 0,\ i = 0, \ldots, N$ und $v_{i-\frac{1}{2}} = 0,\ i = 1, \ldots, N$.

Zu $Y = V \times \mathbb{R}^m$ definieren wir weiterhin

$$\boxed{Y_{(n)} := V_{(n)} \times \mathbb{R}^m.}$$

Der Operator $\mathcal{F} : X \to Y$ repräsentierte die RWA (6.1) ($\mathcal{F}x = 0$). Der Operator

$$\boxed{\mathcal{F}_{(n)} : X_{(n)} \to Y_{(n)}}$$

soll nun die Diskretisierung (6.3) repräsentieren. Zu $u \in X_{(n)}$ definiere deshalb

$$\begin{pmatrix} v \\ \gamma \end{pmatrix} := \mathcal{F}_{(n)} u$$

vermöge

$$\boxed{\begin{aligned} v(t_i) &:= u'(t_i) - f(u(t_i), t_i), \qquad i = 0, \ldots, N, \\ v\left(t_{i-\frac{1}{2}}\right) &:= u'\left(t_{i-\frac{1}{2}}\right) - f\left(u\left(t_{i-\frac{1}{2}}\right), t_{i-\frac{1}{2}}\right), \qquad i = 1, \ldots, N, \\ \gamma &:= r(u(t_0), u(t_N)). \end{aligned}} \tag{6.8}$$

Die Gleichung

$$\boxed{\mathcal{F}_{(n)} u = 0}$$

entspricht dann (6.3). (6.6), zusammen mit $u_i' = f(u_i, t_i)$, $i = 0, \ldots, N$, liefert ein Gleichungssystem zur Bestimmung von $u \in X_{(n)}$.

Es ergibt sich folgende schematische Darstellung:

$$\boxed{\begin{array}{ccccc} X & \xrightarrow{\mathcal{F}} & Y & \qquad & \mathcal{F}x = 0 \\ {\scriptstyle R_{(n)}}\big\downarrow & & \big\downarrow{\scriptstyle R^Y_{(n)}} & & \\ X_{(n)} & \xrightarrow{\mathcal{F}_{(n)}} & Y_{(n)} & & \mathcal{F}_{(n)} u = 0 \end{array}}$$

Definition 6.1 (lokaler Diskretisierungsfehler). Sei die Randwertaufgabe (6.1) lösbar, $x_* \in X$ sei Lösung.

$$\boxed{\tau_{(n)} := \mathcal{F}_{(n)} R_{(n)} x_*}$$

heißt *lokaler Diskretisierungsfehler* (Defekt der exakten Lösungswerte in der Näherungsgleichung).

Wir untersuchen $\tau_{(n)}$ genauer. Sei dazu

$$u_* := R_{(n)} x_*.$$

Nach Definition von $R_{(n)}$ ist

$$\begin{aligned} u_*(t_i) &= x_*(t_i), \qquad i = 0, \ldots, N, \\ u_*'(t_i) &= x_*'(t_i), \qquad i = 0, \ldots, N. \end{aligned}$$

Sei

$$\begin{pmatrix} v_* \\ 0 \end{pmatrix} := \tau_n = \mathcal{F}_{(n)} R_{(n)} x_* = \mathcal{F}_{(n)} u_*.$$

Aus (6.8) erhalten wir nun

$$v_*(t_i) = 0, \qquad i = 0, \ldots, N,$$

und für $i = 1, \ldots, N$ folgt mit (6.4), (6.5)

$$\begin{aligned}
v_*\left(t_{i-\frac{1}{2}}\right) &= u_*'\left(t_{i-\frac{1}{2}}\right) - f\left(u_*\left(t_{i-\frac{1}{2}}\right), t_{i-\frac{1}{2}}\right) \\
&= \frac{3}{2}\frac{x_*(t_i) - x_*(t_{i-1})}{h_i} - \frac{1}{4}\left(x_*'(t_i) + x_*'(t_{i-1})\right) \\
&\quad -f\left(\frac{1}{2}\left(x_*(t_{i-1}) + x_*(t_i)\right) - \frac{1}{8}h_i\left(x_*'(t_i) - x_*'(t_{i-1})\right), t_{i-\frac{1}{2}}\right) \\
&= \frac{3}{2}\left(x_*'\left(t_{i-\frac{1}{2}}\right) + O(h_i)\right) - \frac{1}{4}\left(x_*'\left(t_{i-\frac{1}{2}}\right) + O(h_i) + x_*'\left(t_{i-\frac{1}{2}}\right) + O(h_i)\right) \\
&\quad -f\left(x_*\left(t_{i-\frac{1}{2}}\right) + O\left(h_i^2\right), t_{i-\frac{1}{2}}\right) \\
&= x_*'\left(t_{i-\frac{1}{2}}\right) - f\left(x_*\left(t_{i-\frac{1}{2}}\right) + O\left(h_i^2\right), t_{i-\frac{1}{2}}\right) + O(h_i) \\
&= f\left(x_*\left(t_{i-\frac{1}{2}}\right), t_{i-\frac{1}{2}}\right) - f\left(x_*\left(t_{i-\frac{1}{2}}\right) + O\left(h_i^2\right), t_{i-\frac{1}{2}}\right) + O(h_i) \\
&= O(h_i)
\end{aligned}$$

Aus (6.7) folgt dann (wegen $v_*(t_i) = 0$, $i = 0, \ldots, N$) für $t \in [t_{i-1}, t_i]$

$$\begin{aligned}
v_*(t) &= (t - t_{i-1})\frac{v_*\left(t_{i-\frac{1}{2}}\right)}{\frac{1}{2}h_i} - (t - t_{i-1})\left(t - t_{i-\frac{1}{2}}\right)\left(\frac{v_*\left(t_{i-\frac{1}{2}}\right)}{\frac{1}{2}h_i^2} + \frac{v_*\left(t_{i-\frac{1}{2}}\right)}{\frac{1}{2}h_i^2}\right) \\
&= 2\frac{(t - t_{i-1})}{h_i}\left(1 - \left(t - t_{i-\frac{1}{2}}\right)\frac{2}{h_i}\right) v_*\left(t_{i-\frac{1}{2}}\right),
\end{aligned}$$

also

$$|v_*(t)| \le 2(1+1)\left|v_*\left(t_{i-\frac{1}{2}}\right)\right| = O(h_i),$$

d.h.

$$\max_{t \in [t_0, T]} |v_*(t)| \le O\left(h^{(n)}\right),$$

so daß

$$\boxed{\tau_{(n)} \xrightarrow{h^{(n)} \xrightarrow{n \to \infty} 0} 0.}$$

Definition 6.2 (stabil). Die Diskretisierung des Problems $\mathcal{F}x = 0$ mit der Lösung x_* durch die Problemfolge $\mathcal{F}_{(n)}u = 0$, $n \in \mathbb{N}$, heißt *stabil*, falls es Konstanten $S > 0$, $\rho > 0$ gibt mit:

$$\boxed{\|\bar{u} - \bar{\bar{u}}\|_X \le S\left\|\mathcal{F}_{(n)}\bar{u} - \mathcal{F}_{(n)}\bar{\bar{u}}\right\|_Y \qquad \text{für alle } \bar{u}, \bar{\bar{u}} \in \bar{B}\left(R_{(n)}x_*, \rho\right) \cap X_{(n)}} \tag{6.9}$$

Bemerkung. Für eine Lösung x_* von (6.1) mit M_* regulär ist $\mathcal{F}x_* = 0$ und $\mathcal{F}'(x_*)$ ein Homöomorphismus (Satz 4.7).

Seien $\bar{x}, \bar{\bar{x}} \in \bar{B}(x_*, \rho_0)$ für ein $\rho_0 > 0$ hinreichend klein. Dann ist

$$\mathcal{F}\bar{x} - \mathcal{F}\bar{\bar{x}} = \underbrace{\int_0^1 \underbrace{\mathcal{F}'\left(s\bar{x} + (1-s)\bar{\bar{x}}\right)}_{\approx \mathcal{F}'(x_*)} ds}_{=:\mathcal{M}(\bar{x}, \bar{\bar{x}}) \approx \mathcal{F}'(x_*)} \cdot (\bar{x} - \bar{\bar{x}})$$

Nach dem Banach-Lemma (Störungslemma) ist $\mathcal{M}$ stetig invertierbar und $\left\|\mathcal{M}^{-1}(\bar{x},\bar{\bar{x}})\right\| \leq S_0$ für alle $\bar{x},\bar{\bar{x}} \in \bar{B}(x_*,\rho_0)$, also

$$\begin{aligned} \bar{x}-\bar{\bar{x}} &= \mathcal{M}(\bar{x},\bar{\bar{x}})^{-1}(\mathcal{F}\bar{x}-\mathcal{F}\bar{\bar{x}}) \\ \|\bar{x}-\bar{\bar{x}}\|_X &\leq S_0 \cdot \|\mathcal{F}\bar{x}-\mathcal{F}\bar{\bar{x}}\|_Y \qquad \bar{x},\bar{\bar{x}} \in \bar{B}(x_*,\rho_0). \end{aligned}$$

Sei die Diskretisierung stabil.

Sei $x_{(n)*} \in X_{(n)}$ eine Lösung der Gleichung $\mathcal{F}_{(n)}u = 0$,

$$\mathcal{F}_{(n)}x_{(n)*} = 0.$$

Sei zudem $x_{(n)*} \in \bar{B}\left(R_{(n)}x_*,\rho\right)$. Dann folgt aus (6.9)

$$\boxed{\left\|R_{(n)}x_* - x_{(n)*}\right\|_X \leq S\left\|\tau_{(n)}\right\|_Y}$$

als Abschätzung für den *globalen Fehler*.

Treten beim Lösung von $\mathcal{F}_{(n)}u = 0$ Fehler der Größe $\delta_{(n)}$ auf,

$$\mathcal{F}_{(n)}x_{(n)*} = \delta_{(n)},$$

so erhält man

$$\left\|R_{(n)}x_* - x_{(n)*}\right\|_X \leq S\left\|\tau_{(n)} - \delta_{(n)}\right\|_Y.$$

Satz 6.3. *Die RWA (6.1) habe eine Lösung x_*, M_* sei regulär.*

Die durch (6.3) erzeugte Diskretisierung ist stabil. Die Gleichungen $\mathcal{F}_{(n)}u = 0$ besitzen bei hinreichend feinen Zerlegungen $\pi_{(n)}$ je genau eine Lösung $x_{(n)*} \in \bar{B}\left(R_{(n)}x_*,\rho\right)$. $\mathcal{F}'_{(n)}\left(x_{(n)*}\right)$ ist regulär.

Beweis. Für den Speziallfall von Anfangswertaufgaben führe Exponentialabschätzungen durch wie im Abschnitt 2.3.

Der Fall der Randwertaufgabe wird auf den Fall der Anfangswertaufgabe zurückgeführt. □

6.1 Zusammenhang zu Runge-Kutta-Verfahren

Zur Berechnung des $u \in X_{(n)}$ mit $F_{(n)}u = 0$ war die Lösung des Gleichungssystems (6.6) nötig:

$$\begin{aligned} \frac{u_i - u_{i-1}}{h_i} &= \frac{1}{6}f(u_{i-1},t_{i-1}) + \frac{2}{3}f\left(\mathrm{Arg},t_{i-\frac{1}{2}}\right) + \frac{1}{6}f(u_i,t_i) \qquad i=1,\ldots,N, \\ \mathrm{Arg} &:= \frac{1}{2}(u_{i-1}+u_i) - \frac{1}{8}h_i\,(f(u_i,t_i) - f(u_{i-1},t_{i-1})), \\ r(u_0,u_N) &= 0. \end{aligned} \tag{6.6'}$$

Nun ist

$$\begin{aligned} \mathrm{Arg} &= u_{i-1} + \frac{1}{2}(u_i - u_{i-1}) - \frac{1}{8}h_i f(u_i,t_i) + \frac{1}{8}h_i f(u_{i-1},t_{i-1}) \\ &\overset{(6.6')}{=} u_{i-1} + \frac{1}{2}h_i\left\{\frac{1}{6}f(u_{i-1},t_{i-1}) + \frac{2}{3}f\left(\mathrm{Arg},t_{i-\frac{1}{2}}\right) + \frac{1}{6}f(u_i,t_i)\right\} - \frac{1}{8}h_i f(u_i,t_i) + \frac{1}{8}h_i f(u_{i-1},t_{i-1}) \\ &= u_{i-1} + \frac{5}{24}h_i f(u_{i-1},t_{i-1}) + \frac{1}{3}h_i f\left(\mathrm{Arg},t_{i-\frac{1}{2}}\right) - \frac{1}{24}h_i f(u_i,t_i) \end{aligned}$$

Bezeichnen wir nun

$$\begin{aligned} U'_{i1} &= f\left(u_{i-1}, t_{i-1}\right) \\ U'_{i2} &= f\left(u_{i-1} + \frac{5}{24} h_i U'_{i1} + \frac{1}{3} h_i U'_{i2} - \frac{1}{24} h_i U'_{i3}\right) = f\left(\mathrm{Arg}, t_{i-\frac{1}{2}}\right) \\ U'_{i3} &= f\left(u_i, t_i\right), \end{aligned}$$

so erhalten wir aus (6.6'):

$$u_i = u_{i-1} + h_i \left(\frac{1}{6} U'_{i1} + \frac{2}{3} U'_{i2} + \frac{1}{6} U'_{i3}\right).$$

Dies entspricht gerade dem Lobatto IIIA-Verfahren für $s = 3$ ($p = 2s - 1$, A-stabil, nicht L-stabil):

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

6.2 Praktische Verwendung von Kollokation

Um die Kollokationsverfahren anwenden zu können, müssen wir das Gleichungssystem (6.6) lösen. Linearisiert man die Gleichungen, so erhält man ein System dessen linke Seite die Form

$$\begin{bmatrix} -I + h_1 F_{11} & I + h_1 F_{12} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -I + h_{N-1} F_{N-1,N-1} & I + h_{N-1} F_{N-1,N} \\ r'_1 & 0 & \cdots & \cdots & r'_2 \end{bmatrix} =$$

$$\begin{bmatrix} -I + h_1 F_{11} & I & 0 & \cdots & 0 \\ 0 & \left(-I + h_2 F_{22}\right)\left(I + h_1 F_{12}\right)^{-1} & I & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \left(-I + h_{N-1} F_{N-1,N-1}\right)\left(\cdots\right)^{-1} & I \\ r'_1 & 0 & \cdots & \cdots & r'_2\left(\cdots\right)^{-1} \end{bmatrix}$$

$$\cdot \begin{bmatrix} I & & & \\ & \left(I + h_1 F_{12}\right) & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

hat. Für diese nichlinearen Gleichungen in $\mathbb{R}^{(N+1)m}$ mit zyklischer Struktur, N groß, ist ein numerisch stabiler linearer Löser zu verwenden!

Literaturverzeichnis

[SW95] K. Strehmel, R. Weiner, Numerik gewöhnlicher Differentialgleichungen, Teubner Studienbücher Mathematik, Stuttgart 1995

[AP98] Uri M. Ascher, Linda R. Petzold, Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations, siam, 1998

[DB94] Deuflhard, Bornemann, Numerische Mathematik II - Integration gewöhnlicher Differentialgleichungen, de Gruyter Lehrbuch, New York 1994

[SH96] A.M. Stuart, A.R. Humphries, Dynamical Systems and Numerical Analysis, Cambridge University Press 1996

[SB90] Stoer, Bulirsch, Numerische Mathematik 2, Springer 1990

Index