Numerik gewöhnlicher Differentialgleichungen

gehalten von

Werner Römisch

Vorlesung
im Wintersemester 2012/2013

Humboldt-Universität Berlin
Institut für Mathematik

Inhalt: Seite

1 Theorie gewöhnlicher Differentialgleichungen

Problem: (1) $x'(t) = f(x(t), t), \quad \forall t \in I, x(t_0) = x_0,$

wobei $I \subseteq \mathbb{R}$ ein Intervall, $t_0 \in I, x_0 \in \mathbb{R}^m$ und $f : \mathbb{R}^m \times I \to \mathbb{R}^m$.

Satz 1.1 *(Cauchy-Peano)*
Es seien I ein Intervall, $t_0 \in I$, $x_0 \in \mathbb{R}^m$, $r > 0$ und $f : \mathbb{R}^m \times I \to \mathbb{R}^m$ sei stetig auf $B(x_0, r) \times I$, wobei $B(x_0, r) = \{x \in \mathbb{R}^m : \|x - x_0\| \leq r\}$.
Dann existieren ein Intervall $I_0 \subseteq I$ mit $t_0 \in I_0$ und eine differenzierbare Funktion $x : I_0 \to \mathbb{R}^m$ mit

$$x'(t) = f(x(t), t), \ \forall t \in I_0, \ x(t_0) = x_0.$$

Beweisidee:
O.B.d.A. sei t_0 nicht rechter Randpunkt von I. Wir wählen $\eta > 0$ so, dass $M(\eta, r)\eta \leq r$, wobei $M(\eta, r) = \{\|f(y, t)\| : (y, t) \in B(x_0, r) \times [t_0, t_0 + \eta]\}$, und wir setzen $I_0 = [t_0, t_0 + \eta]$.
Wir betrachten die Menge $X = \{x \in C(I_0, \mathbb{R}^m) : \|x(t) - x_0\| \leq r, \ \forall t \in I_0\}$ im Banachraum $C(I_0, \mathbb{R}^m)$ mit der Norm $\|\cdot\|_\infty$ und den Operator

$$(Tx)(t) := x_0 + \int_{t_0}^{t} f(x(s), s)ds, \ t \in I_0,$$

von X in $C(I_0, \mathbb{R}^m)$. Man zeigt nun, dass X beschränkt, konvex und abgeschlossen ist, $T(X) \subseteq X$ gilt und dass T stetig ist und beschränkte Mengen in relativ kompakte abbildet (man sagt "T ist vollstetig"). Letzteres zeigt man mit dem Satz von Arzela-Ascoli. Dann folgt die Aussage aus dem Fixpunktsatz von Schauder. □

Frage: Ist unter den Voraussetzungen von Satz 1.1 die Lösung von (1) eindeutig bestimmt?

Beispiel 1.2
Wir betrachten die Anfangswertaufgabe $x'(t) = x(t)^{\frac{1}{3}}, \forall t \in [0, +\infty[, x(0) = 0.$
D.h. $f(y, t) = y^{\frac{1}{3}}, \forall (y, t) \in \mathbb{R} \times [0, +\infty[\rightsquigarrow$ *f ist stetig,* $\rightsquigarrow \exists$ *"lokale" Lösung.*
Lösung: (Verifikation durch Nachrechnen)

$$x(t) := \left(\frac{2}{3}t\right)^{\frac{3}{2}}, \forall t \in [0, +\infty[.$$

Es existieren aber unendlich viele Lösungen:

$$x_c(t) := \begin{cases} (\frac{2}{3}(t-c)^{\frac{3}{2}}, & \forall t \in [c, +\infty[\\ 0, & \forall t \in [0, c[\end{cases} \quad \forall c \geq 0.$$

Ursache: Wir benötigen stärkere Stetigkeitseigenschaften von f, um eindeutige Lösungen zu erhalten!

Satz 1.3 *(Picard-Lindelöf)*
Es sei $I \subseteq \mathbb{R}$ ein Intervall, es seien $t_0 \in I, x_0 \in \mathbb{R}^m, r > 0$, und $f : \mathbb{R}^m \times I \to \mathbb{R}^m$ sei stetig auf $B(x_0, r) \times I$. Ferner existiere eine Konstante $L > 0$ mit

$$\|f(y,t) - f(z,t)\| \leq L\|y - z\|, \forall (y,t), (z,t) \in B(x_0, r) \times I.$$

Dann existieren ein Intervall $I_0 \subseteq I$ mit $t_0 \in I_0$ und eine eindeutig bestimmte differenzierbare Funktion $x : I_0 \to \mathbb{R}^m$, so dass

$$x'(t) = f(x(t), t),\ \forall t \in I_0,\ x(t_0) = x_0.$$

Beweisidee:
Man wählt $\eta > 0$ wie im Beweis von Satz 1.1.
Die Menge X verstehen wir als vollständigen metrischen Raum mit der durch $\|\cdot\|_\infty$ definierten Metrik d. Zusätzlich betrachten wir die Metriken

$$d_k(x, \tilde{x}) = \max_{t \in I_0} \exp(-k(t - t_0))\|x(t) - \tilde{x}(t)\| \quad (k > 0)$$

auf X. Da für diese gilt

$$\exp(-k\eta)d(x, \tilde{x}) \leq d_k(x, \tilde{x}) \leq d(x, \tilde{x}) \quad (\forall x, \tilde{x} \in X)$$

ist auch (X, d_k) ein vollständiger metrischer Raum. Der Operator $T : X \to X$ ist wieder gegeben durch

$$(Tx)(t) := x_0 + \int_{t_0}^{t} f(x(s), s)ds,\ t \in I_0.$$

Er ist wohldefiniert und kontraktiv als Abbildung von X in X bzgl. d_k mit $k > L$. Letzteres folgt aus der Abschätzung

$$\begin{aligned}\exp(-k(t - t_0))\|(Tx)(t) - (T\tilde{x})(t)\| &\leq L \exp(-k(t - t_0)) \int_{t_0}^{t} \exp(k(s - t_0))ds\, d_k(x, \tilde{x}) \\ &\leq \frac{L}{k} d_k(x, \tilde{x}) \quad (\forall t \in I_0).\end{aligned}$$

Die Aussage folgt dann aus dem Banachschen Fixpunktsatz. □

Da wir bisher nur Lösungen auf kleinen Intervallen, die t_0 enthalten, erhalten haben, also "lokale" Lösungen, besteht unser nächstes Ziel in der Fortsetzung lokaler Lösungen.

Definition 1.4 **a)** *Ein Paar $(\tilde{I}, \tilde{x})$ heißt lokale Lösung von (1), falls $\tilde{I} \subseteq I$ ein Intervall mit $t_0 \in \tilde{I}$, und $\tilde{x}$ Lösung von $x'(t) = f(x(t), t)$, $\forall t \in \tilde{I}, x(t_0) = x_0$, ist.*

b) *Eine lokale Lösung $(\hat{I}, \hat{x})$ heißt Fortsetzung von $(\tilde{I}, \tilde{x})$, falls $\tilde{I} \subseteq \hat{I}$ und $\tilde{x}(t) = \hat{x}(t), \forall t \in \tilde{I}$. Man nennt die Fortsetzung strikt, falls $\tilde{I} \subset \hat{I}$.*

c) *Eine lokale Lösung von $(\tilde{I}, \tilde{x})$ heißt maximale Lösung von (1), falls keine lokale Lösung von (1) existiert, die strikte Fortsetzung von $(\tilde{I}, \tilde{x})$ ist.*

d) *$(\tilde{I}, \tilde{x})$ heißt globale Lösung von (1), falls $(\tilde{I}, \tilde{x})$ eine lokale Lösung von (1) ist und $\tilde{I} = I$ gilt.*

Beispiel 1.5
Wir betrachten eine skalare Differentialgleichung mit getrennten Variablen der Form

$$x'(t) = \frac{c(t)}{g(x(t))} \quad (t \in I, t_0 \in I, x(t_0) = x_0)$$

mit stetigen Funktionen $c : I \to \mathbb{R}$ und $g : D(g) \to \mathbb{R} \setminus \{0\}$. Diese kann mit der Kettenregel äquivalent umgeschrieben werden als

$$g(x(t))x'(t) = \frac{d}{dt} G(x(t)) = c(t) \quad (t \in I),$$

wobei G eine Stammfunktion von g ist. In aufintegrierter Form hat letztere Gleichung die Gestalt

$$G(x(t)) = G(x_0) + \int_{t_0}^{t} c(s)ds \quad (t \in I).$$

Mit der inversen Funktion G^{-1} von G, falls sie existiert, ergibt sich als Lösung

$$x(t) = G^{-1}\Big(G(x_0) + \int_{t_0}^{t} c(s)ds\Big) \quad (t \in I)$$

Diese Prozedur wenden wir auf die folgenden Beispiele an.

a) $x'(t) = x(t)^2, \forall t \in \mathbb{R}, x(1) = -1$.
Es gilt $c(t) = 1$, $g(y) = y^{-2}$, $D(g) = \mathbb{R} \setminus \{0\}$, $x_0 = -1$, $t_0 = 1$, $I = [0, +\infty)$. Dann ist $G(y) = -y^{-1}$ eine Stammfunktion von g und $G(x_0) = 1$. Also entsteht als Lösung

$$x(t) = G^{-1}(1 + t - t_0) = -t^{-1} \quad (t \in (0, +\infty)).$$

Diese ist maximale Lösung der Anfangswertaufgabe und kann nicht zu einer globalen Lösung fortgesetzt werden.

b) $x'(t) = -2tx(t)^2, \forall t \in \mathbb{R}, x(1) = \frac{1}{2}$.
Man erhält mit der gleichen Prozedur (Übung)

$$x(t) = \frac{1}{1 + t^2} \quad (t \in \mathbb{R})$$

Diese Funktion x ist globale Lösung der Anfangswertaufgabe.

Lemma 1.6
Existiert eine lokale Lösung von (1), so existiert auch eine maximale Lösung von (1), die diese lokale Lösung fortsetzt.

Beweis:
O.B.d.A. sei im folgenden t_0 linker Randpunkt von I. Alle anderen Fälle lassen sich analog behandeln. Es sei $(\tilde{I}, \tilde{x})$ eine lokale Lösung von (1).
$\mathcal{M}_0 :=$ Menge aller lokalen Lösungen von (1), die $(\tilde{I}, \tilde{x})$ fortsetzen.
Nach Definition und Voraussetzung gilt $\mathcal{M}_0 \neq \emptyset$.
Sei $\tau_0 := \sup\{t : \exists([t_0, t], x) \in \mathcal{M}_0\}$. Wir unterscheiden nun die beiden Fälle $\tau_0 < +\infty$ und $\tau_0 = +\infty$. Im ersten Fall setzen wir wie folgt fort:
Wir wählen $([t_0, t_1), x_{(1)}) \in \mathcal{M}_0$, so dass $\tau_0 \geq t_1 \geq \max\{t_0, \tau_0 - 1\}$.
$\mathcal{M}_1 :=$ Menge aller lokalen Lösungen von (1), die $([t_0, t_1], x_{(1)})$ fortsetzen.
Es sei $\tau_1 := \sup\{t : \exists([t_0, t], x) \in \mathcal{M}_1\} \leq \tau_0$.
Wir wählen $([t_0, t_2), x_{(2)}) \in \mathcal{M}_1$, so dass $\tau_1 \geq t_2 \geq \max\{t_0, \tau_1 - \frac{1}{2}\}$, usw.
Allgemein definieren wir für jedes i:
$\mathcal{M}_i :=$ Menge aller lokalen Lösungen von (1), die $([t_0, t_i], x_{(i)})$ fortsetzen.
$\tau_i := \sup\{t : \exists([t_0, t], x) \in \mathcal{M}_i\}$ und $\tau_i \geq t_{i+1} \geq \max\{t_0, \tau_i - \frac{i}{i+1}\}$.
Nach Konstruktion gilt: $t_{i+1} \geq t_i, \tau_{i+1} \leq \tau_i, t_{i+1} \leq \tau_{i+1}$.
$\rightsquigarrow t_i \leq t_{i+1} \leq \tau_{i+1} \leq \tau_i, \quad \forall i \in \mathbb{N}$.
$\rightsquigarrow (t_i)_{i\in\mathbb{N}}$ und $(\tau_i)_{i\in\mathbb{N}}$ sind als monotone beschränkte Folgen beide konvergent gegen denselben Grenzwert $\tau \in I$.

$$\rightsquigarrow \lim_{i\to\infty} t_i = \tau = \lim_{i\to\infty} \tau_i.$$

Wir definieren eine Funktion $\hat{x} : [t_0, \tau[\to \mathbb{R}^m$, $\hat{x}(t) := x_{(i)}(t), \quad \forall t \in [t_0, t_i]$.
$\hat{x}$ ist damit wohl-definiert, da $x_{(i+1)}$ Fortsetzung von $x_{(i)}$ ist, $\forall i \in \mathbb{N}$.
Außerdem ist $([t_0, \tau[, \hat{x})$ lokale Lösung , da, $\forall i \in \mathbb{N}, ([t_0, t_i], x_{(i)})$ lokale Lösung ist.
Ferner ist $([t_0, \tau[, \hat{x})$ eine Fortsetzung von $(\tilde{I}, \tilde{x})$.
Mit dem so konstruierten τ bzw. im Fall $\tau := \tau_0 = +\infty$ unterscheiden wir nun die folgenden zwei Fälle:

(i) $([t_0, \tau[, \hat{x})$ ist maximale Lösung,

(ii) es existiert eine lokale Lösung $([t_0, \bar{t}], \bar{x})$ von (1), die eine strikte Fortsetzung von $([t_0, \tau[, \hat{x})$ ist, d. h. $\tau \leq \bar{t}$.
$\rightsquigarrow$ nach Konstruktion müßte dann gelten: $([t_0, \bar{t}], \bar{x}) \in \mathcal{M}_i, \forall i \in \mathbb{N}$,
$\rightsquigarrow \tau_i \geq \bar{t}, \forall i \in \mathbb{N}$.
$\rightsquigarrow \tau \leq \bar{t} \leq \tau_i, \forall i \in \mathbb{N}$, und $\tau_i \to \tau$.
$\rightsquigarrow \bar{t} = \tau$.
$\rightsquigarrow ([t_0, \tau], \bar{x})$ ist maximale Lösung von (1). □

Satz 1.7
Es sei $I \subseteq \mathbb{R}$ ein Intervall, $t_0 \in I$ sei linker Randpunkt von $I, x_0 \in \mathbb{R}^m$, und die Funktion $f : \mathbb{R}^m \times I \to \mathbb{R}^m$ sei stetig.
Es existiert eine maximale Lösung $(\tilde{I}, \tilde{x})$ von (1), die entweder eine globale Lösung von (1) ist oder $\tilde{I}$ hat die Form $[t_0, t_1[$ mit $t_1 \in I$, und $\tilde{x}$ ist nicht beschränkt auf $\tilde{I}$ (d.h. $\tilde{x}$ "explodiert").

Beweis:
Nach dem Satz 1.1 existiert eine lokale Lösung von (1) und nach Lemma 1.6 folglich auch eine maximale Lösung $(\tilde{I}, \tilde{x})$ von (1), die die lokale Lösung fortsetzt.
Wir unterscheiden vier Fälle:

i) $\tilde{I} = I$, d.h. $(\tilde{I}, \tilde{x})$ ist globale Lösung.

ii) $\tilde{I} = [t_0, t_1] \subseteq I, t_1 \in \text{ int}(I)$. Wir setzen $x_1 := \tilde{x}(t_1)$ und betrachten die Anfangswertaufgabe

$$x'(t) = f(x(t), t), \forall t \in [t_1, +\infty[\cap I, x(t_1) = x_1 \qquad (*)$$

Nach Satz 1.1 existieren ein $\eta_1 > 0$ und eine differenzierbare Funktion y auf $([t_1, t_1+\eta_1])$, die Lösung von $(*)$ ist. Wir definieren eine Funktion $\hat{x} : [t_0, t_1+\eta_1] \to \mathbb{R}^m$ durch

$$\hat{x}(t) := \begin{cases} \tilde{x}(t), \forall t \in [t_0, t[\\ y(t), \text{ sonst} \end{cases}$$

$\rightsquigarrow \hat{x}$ ist differenzierbar in $[t_0, t_1-\eta_1]$ und lokale Lösung von (1), die $(\tilde{I}, \tilde{x})$ fortsetzt.
$\rightsquigarrow$ Widerspruch!

iii) $\tilde{I} = [t_0, t_1[$, und $\tilde{x}$ ist nicht beschränkt auf $\tilde{I}$.

iv) $\tilde{I} = [t_0, t_1[$, und $\tilde{x}$ ist beschränkt auf $\tilde{I}$.
Wir definieren $\hat{x}(t) = \tilde{x}(t), \forall t \in \tilde{I}$.
Ziel: $\hat{x}$ auf das Intervall $[t_0, t_1]$ fortsetzen und $\hat{x}$ ist lokale Lösung von (1). $\rightsquigarrow$ Widerspruch zu $(\tilde{I}, \tilde{x})$ maximale Lösung.

Wir betrachten die Integrale $\int_{t_0}^{t} f(\tilde{x}(s), s)ds, \forall t \in \tilde{I}$.

Wir zeigen: $\exists \lim_{t \to t_1} \int_{t_0}^{t} f(\tilde{x}(s), s)ds$.

Sei (τ_n) eine monoton wachselnde Folge in $\tilde{I}$ mit $\tau_n \to t_1$. Dann gilt für $m > n$:

$$\| \int_{t_0}^{\tau_n} f(\tilde{x}(s), s)ds - \int_{t_0}^{\tau_m} f(\tilde{x}(s), s)ds \| \leq \int_{\tau_n}^{\tau_m} \|f(\tilde{x}(s), s)\|ds \leq C(\tau_m - \tau_n),$$

wobei in C die Beschränktheit von $\tilde{x}$ und die Stetigkeit von f "eingeht".
$\rightsquigarrow \exists \lim_{t \to t_1} \int_{t_0}^{t} f(\tilde{x}(s), s)ds,$
$\rightsquigarrow \hat{x}(t_1) := x_0 + \lim_{t \to t_1} \int_{t_0}^{t} f(\underbrace{\tilde{x}(s)}_{=\hat{x}(s)}, s)ds = \lim_{t \to t_1} \tilde{x}(t) = x_0 + \int_{t_0}^{t_1} f(\hat{x}(s), s)ds.$
$\rightsquigarrow ([t_0, t_1], \hat{x})$ ist lokale Lösung von (1) $\rightsquigarrow$ Widerspruch! □

Frage: Welche Bedingungen an die DGL (d.h. also an die Funktion f) verhindern eine Explosion von maximalen Lösungen?

Lemma 1.8 *(Gronwall)*
Es seien w, g stetige Funktionen auf $[a, b], c > 0$ und es gelte die "Integralungleichung"

$$0 \leq w(t) \leq g(t) + c \int_a^t w(s)ds, \quad t \in [a, b].$$

Dann gilt: $w(t) \leq \max_{t \in [a,b]} |g(t)| \exp(c(t-a)), \forall t \in [a, b].$

Beweis: Wir betrachten die Funktion $u(t) := \max_{t\in[a,b]} |g(t)| + c\int_a^t w(s)ds, \forall t \in [a,b]$.
u ist stetig differenzierbar und es gilt $w(t) \le u(t), u'(t) = cw(t), \forall t \in [a,b]$.
Die Funktion $v(t) := u(t)\exp(-c(t-a))$, $\forall t \in [a,b]$, ist monoton fallend, da
$v'(t) = (u'(t) - cu(t))\exp(-c(t-a)) \le 0, \forall t \in [a,b]$, gilt.
$\rightsquigarrow v(t) \le v(a) = u(a) = \max_{t\in[a,b]} |g(t)|$, $\forall t \in [a,b]$.
$\rightsquigarrow w(t) \le u(t) \le \max_{t\in[a,b]} |g(t)|\exp(c(t-a)), \forall t \in [a,b]$. □

Satz 1.9 *(Existenz globaler Lösungen)*
Es sei I ein kompaktes Intervall und t_0 linker Randpunkt von I, $f : \mathbb{R}^m \times I \to \mathbb{R}^m$ sei stetig, und es existiere ein Skalarprodukt $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^m$ mit zugehöriger Norm $\|\cdot\|$ und eine Konstante L, so dass

$$\langle f(y,t), y\rangle \le L(1+\|y\|^2), \forall (y,t) \in \mathbb{R}^m \times I \text{ (”einseitiges lineares Wachstum”).}$$

Dann existiert (mindestens) eine globale Lösung von (1).
Insbesondere gilt dies, falls f ”lineares Wachstum” besitzt, d.h.

$$\|f(y,t)\| \le \ell(1+\|y\|),\ \forall (y,t) \in \mathbb{R}^m \times I,$$

gilt mit einer Konstanten $\ell > 0$ und einer beliebigen Norm in $\mathbb{R}^m$.

Beweis:
Wir wenden Satz 1.7 an und zeigen, daß jede lokale Lösung von (1) beschränkt ist. Satz 1.7 besagt dann die Existenz einer maximalen Lösung, die globale Lösung sein muss.
Es sei also $(\tilde{I}, x)$ eine lokale Lösung von (1).
Wir betrachten die Funktion $w : I \to \mathbb{R}$, $w(t) := \|x(t)\|^2, \quad \forall t \in \tilde{I}$.
Diese ist differenzierbar und es gilt: $w'(t) = 2\langle x'(t), x(t)\rangle$, $\forall t \in \tilde{I}$.

$$\begin{aligned}
&\rightsquigarrow \quad \tfrac{d}{dt}\|x(t)\|^2 = 2\langle f(x(t),t), x(t)\rangle \le 2L(1+\|x(t)\|^2), \quad \forall t \in \tilde{I}.\\
&\rightsquigarrow \quad \|x(t)\|^2 - \|x(t_0)\|^2 \le 2L\int_{t_0}^t (1+\|x(s))\|^2)ds, \quad \forall t \in \tilde{I}.\\
&\rightsquigarrow \quad \underbrace{\|x(t)\|^2}_{=:w(t)} \le \underbrace{\left[\|x(t_0)\|^2 + c\int_{t_0}^t ds\right]}_{=:g(t)} + c\int_{t_0}^t \|x(s)\|^2 ds, \quad \forall t \in \tilde{I}.
\end{aligned}$$

wobei $c := 2|L|$. Aus dem Lemma von Gronwall folgt dann

$$w(t) = \|x(t)\|^2 \le \sup_{t\in I} |g(t)|\exp(c(t-t_0)), \quad \forall t \in \tilde{I},$$

d. h. die beliebig gewählte lokale Lösung ist beschränkt! Der Nachsatz folgt im Fall, daß die Normen identisch sind, aus der Ungleichung

$$\langle f(y,t), y\rangle \le \|f(y,t)\|\|y\| \le \ell(1+\|y\|)\|y\| \le 2\ell(1+\|y\|^2), \forall y \in \mathbb{R}^m.$$

Sind die Normen verschieden, gehen zusätzlich die Normäquivalenz-Konstanten ein. □

Beispiel 1.10
Gleichung der Populationsdynamik: $x'(t) = ax(t) - bx(t)^2, t \in [t_0, +\infty), x(t_0) = x_0$.
Die rechte Seite f *der Differentialgleichung hat die Form*

$$f(y,t) := \begin{cases} ay - by^2 & , y \geq 0 \quad (a > b > 0) \\ 0 & , sonst \end{cases}$$

$\rightsquigarrow f : \mathbb{R} \times I \to \mathbb{R}$ *ist stetig und es gilt wegen* $\langle r, \tilde{r} \rangle := r\tilde{r}, \forall r, \tilde{r} \in \mathbb{R}$:

$$\rightsquigarrow f(y,t)y = \begin{cases} ay^2 - by^3 & , \quad 0 \leq ay^2 \leq a(1+y^2),\ y \geq 0 \\ 0 & , \quad sonst \end{cases}$$

Satz 1.9 $\rightsquigarrow$ *die DGL besitzt auf jedem kompakten Intervall eine Lösung.*

Nächstes Ziel: Bedingungen für die Eindeutigkeit von Lösungen!

Satz 1.11 *(Existenz und Eindeutigkeit globaler Lösungen)*
Sei I *ein abgeschlossenes Intervall,* $t_0 \in I$ *linker Randpunkt von* I, $x_0 \in \mathbb{R}^m$, $f : \mathbb{R}^m \times I \to \mathbb{R}^m$ *sei stetig und es existiere ein Skalarprodukt* $\langle \cdot, \cdot \rangle$ *auf* $\mathbb{R}^m$ *mit Norm* $\| \cdot \|$ *und eine Konstante* γ *so, dass*

$$(*) \qquad \langle f(y,t) - f(\tilde{y},t), y - \tilde{y} \rangle \leq \gamma \|y - \tilde{y}\|^2,\ \forall y, \tilde{y} \in \mathbb{R}^m,\ \forall t \in I.$$

Dann besitzt (1) eine globale Lösung, und jede lokale Lösung von (1) ist Einschränkung dieser globalen Lösung.
Überdies gilt

$$\|x(t) - \tilde{x}(t)\| \leq \exp(\gamma(t - t_0))\|x(t_0) - \tilde{x}(t_0)\| \quad \forall t \in I,$$

für zwei Lösungen x *und* $\tilde{x}$ *auf dem Intervall* I *mit zugehörigen Anfangswerten* $x(t_0)$ *bzw.* $\tilde{x}(t_0)$.
Insbesondere ist $(*)$ *erfüllt mit* $\gamma = L$, *falls ein* $L > 0$ *existiert mit*

$$\|f(y,t) - f(\tilde{y},t)\| \leq L\|y - \tilde{y}\|, \quad \forall y, \tilde{y} \in \mathbb{R}^m \forall t \in I.$$

Beweis: Wir nehmen zunächst an, dass I auch beschränkt ist.

1. Existenz (mit Hilfe von Satz 1.9)
   Wir zeigen, dass die Wachstumsbedingung von Satz 1.9 erfüllt ist.
   $\tilde{y} = 0 \rightsquigarrow \langle f(y,t) - f(0,t), y \rangle \leq L\|y\|^2$.

   $$\begin{aligned} \rightsquigarrow \langle f(y,t), y \rangle &\leq \langle f(0,t), y \rangle + L\|y\|^2 \leq \|f(0,t)\|\|y\| + L\|y\|^2 \\ &\leq (\max_{t \in I} \|f(0,t)\| + L)(1 + \|y\|^2), \forall y \in \mathbb{R}^m, \forall t \in I. \end{aligned}$$

   $\rightsquigarrow$ Satz 1.9 liefert die Existenz einer globalen Lösung auf I.

2. Es seien $(\tilde{I}, \tilde{x}), (I, x)$ eine lokale bzw. globale Lösung von (1).
Wir zeigen: $\tilde{x}(t) = x(t), \quad \forall t \in \tilde{I}$.
Wir betrachten die Funktion $\varphi(t) := \exp(-2\gamma(t-t_0))\|x(t) - \tilde{x}(t)\|^2, \forall t \in \tilde{I}$.

$$\begin{aligned} \rightsquigarrow \varphi(t_0) &= 0, \text{ da } x(t_0) = \tilde{x}(t_0) = x_0, \\ \rightsquigarrow \varphi'(t) &= \|x(t) - \tilde{x}(t)\|^2(-2\gamma)\exp(-2\gamma(t-t_0)) \\ &\quad +2\exp(-2\gamma(t-t_0))\langle x'(t) - \tilde{x}'(t), x(t) - \tilde{x}(t)\rangle \\ &= 2\exp(-2\gamma(t-t_0))[\langle x'(t) - \tilde{x}'(t), x(t) - \tilde{x}(t)\rangle - \gamma\|x(t) - \tilde{x}(t)\|^2] \\ &\leq 0, \forall t \in \tilde{I} \quad (\text{wegen}(^*)) \end{aligned}$$

$\rightsquigarrow \varphi$ ist monoton fallend, $\rightsquigarrow 0 \leq \varphi(t) \leq \varphi(t_0) = \|x_0 - x_0\|^2 = 0$,
$\rightsquigarrow \varphi(t) = 0, \forall t \in \tilde{I} \rightsquigarrow \tilde{x}(t) = x(t), \quad \forall t \in \tilde{I}$.

3. Sind nun x und $\tilde{x}$ zwei Lösungen von (1) mit verschiedenen Anfangswerten $x(t_0)$ und $\tilde{x}(t_0)$. Mit demselben Ansatz für φ wie in 2. aber auf I erhalten wir, dass φ monoton fallend auf I ist und folglich

$$\varphi(t) := \exp(-2\gamma(t-t_0))\|x(t) - \tilde{x}(t)\|^2 \leq \varphi(t_0) = \|x(t_0) - \tilde{x}(t_0)\|^2, \quad \forall t \in I,$$

oder

$$\|x(t) - \tilde{x}(t)\| \leq \exp(\gamma(t-t_0))\|x(t_0) - \tilde{x}(t_0)\|, \quad \forall t \in I.$$

4. Für alle $t \in I$ gilt:

$$\langle f(y,t) - f(\tilde{y},t), y - \tilde{y}\rangle \leq \|f(y,t) - f(\tilde{y},t)\|\|y - \tilde{y}\| \leq L\|y - \tilde{y}\|^2 \ \forall y, \tilde{y} \in \mathbb{R}^m.$$

Offenbar sind die Beweise in 2.–4. auch richtig, wenn I unbeschränkt ist. Ist nun x_i eine globale Lösung von (1) auf dem Intervall $[t_0, i]$ für jedes $i \in \mathbb{N}$ mit $t_0 \leq i$, so definieren wir im Fall $I = [t_0, +\infty)$ eine Funktion

$$\hat{x} : I \to \mathbb{R}^m, \quad \hat{x}(t) := x_i(t), \ t \in [t_0, i].$$

Diese ist Lösung von (1) auf I und jede lokale Lösung ist Einschränkung von $\hat{x}$ auf ein kleineres Intervall. Damit ist alles bewiesen. □

Beispiel 1.12
DGL der Populationsdynamik (Forsetzung von Beispiel 1.10):

$$x(t) = \frac{ax_0}{bx_0 + (a - bx_0)\exp(-a(t-t_0))} \quad \textit{ist Lösung auf } [t_0, +\infty).$$

$$f(y,t) := \begin{cases} ay - by^2 & , \ y \in [0, \frac{a}{b}], (a > b > 0) \\ 0 & , \ \textit{sonst} \end{cases}.$$

Nachprüfen der Voraussetzungen von Satz 1.11:

$$(f(y,t) - f(\tilde{y},t))(y - \tilde{y}) = \begin{cases} (a(y-\tilde{y}) - b(y^2 - \tilde{y}^2))(y - \tilde{y}) \leq a(y - \tilde{y})^2, y, \tilde{y} \in [0, \frac{a}{b}] \\ (ay - by^2)(y - \tilde{y}) \leq a(y - \tilde{y})^2, y \in [0, \frac{a}{b}], \tilde{y} < 0 \\ (by^2 - ay)(\tilde{y} - y) \leq a(\tilde{y} - y)^2, y \in [0, \frac{a}{b}], \tilde{y} > \frac{a}{b} \\ 0 \leq a(\tilde{y} - y)^2, y, \tilde{y} \notin [0, \frac{a}{b}] \end{cases}$$

Die Abschätzung in Satz 1.11 ist also erfüllt. Folglich stimmen lokale und eingeschränkte globale Lösungen überein.

Definition 1.13
Die DGL (1) mit $I = [t_0, +\infty)$ heißt kontraktiv (schwach kontraktiv), falls ein Skalarprodukt $\langle \cdot, \cdot \rangle$ auf $\mathbb{R}^m$, eine zugehörige Norm $\| \cdot \|$ und eine Konstante $\gamma < 0$ existieren, so dass

$$\langle f(x,t) - f(\tilde{x},t), x - \tilde{x} \rangle \leq \gamma \|x - \tilde{x}\|^2 \quad \forall x, \tilde{x} \in \mathbb{R}^m, \forall t \in [t_0, +\infty).$$

$$(\langle f(x,t) - f(\tilde{x},t), x - \tilde{x} \rangle \leq 0\,)$$

Folgerung 1.14 *Ist die DGL (1) mit $I = [t_0, +\infty)$ kontraktiv, so gilt für zwei Lösungen x und $\tilde{x}$ mit den Anfangswerten $x(t_0)$ bzw. $\tilde{x}(t_0)$ stets*

$$\lim_{t \to +\infty} \|x(t) - \tilde{x}(t)\| = 0$$

in jeder beliebigen Norm $\| \cdot \|$ auf dem $\mathbb{R}^m$.
Ist (1) schwach kontraktiv, so existiert eine Norm $\| \cdot \|$ auf $\mathbb{R}^m$, so dass

$$\|x(t) - \tilde{x}(t)\| \leq \|x(t_0) - \tilde{x}(t_0)\| \quad \forall t \in I.$$

Beweis: folgt sofort aus Satz 1.11 und im ersten Teil aus der Äquivalenz aller Normen auf $\mathbb{R}^m$. □

Ab jetzt beschäftigen wir uns mit linearen Differentialgleichungen der Form

$$(2) \quad x'(t) = A(t)x(t) + b(t), \quad \forall t \in I,\, x(t_0) = x_0,$$

wobei die Matrixfunktion $A(\cdot)$ und die Funktion $b(\cdot)$ auf I stetig sind.

Satz 1.15 *Sei I ein abgeschlossenes Intervall, $t_0 \in I$ linker Randpunkt von I, $x_0 \in \mathbb{R}^m$. Es seien $b : I \to \mathbb{R}^m$ und $A : I \to \mathbb{R}^{m \times m}$ stetig. Dann existiert genau eine globale Lösung x von (2). Diese hat die Form*

$$x(t) := X(t) \left[x_0 + \int_{t_0}^{t} X^{-1}(s) b(s) ds \right], \forall t \in I,$$

wobei die Matrixfunktion $X : I \to \mathbb{R}^{m \times m}$ die Spalten $x_i(\cdot)$, $i = 1, \ldots, m$, besitzt und es gilt $x_i'(t) = A(t)x_i(t)$, $t \in I$, $x_i(t_0) = e_i$, mit dem i-ten kanonischen Einheitsvektor e_i, $i = 1, \ldots, m$.

Beweis: Die Funktion $f : \mathbb{R}^m \times I \to \mathbb{R}^m$, $f(y,t) := A(t)y + b(t)$ erfüllt die Voraussetzungen von Satz 1.11. Folglich existiert genau eine globale Lösung von (2). Die Matrixfunktion $X(\cdot) : I \to \mathbb{R}^{m \times m}$ ist wohldefiniert nach Satz 1.11 und es gilt

$$\frac{d}{dt} X(t) = A(t)X(t),\ \forall t \in I,\ X(t_0) = E \ (E \text{ ist die Einheitsmatrix im } \mathbb{R}^{m \times m}).$$

Nach Konstruktion ist die Funktion $x_y : I \to \mathbb{R}^m$, $x_y(t) = X(t)y$, $t \in I$, die einzige Lösung des linearen Anfangswertproblems

$$x'(t) = A(t)x(t),\ \forall t \in I,\ x(t_0) = y \in \mathbb{R}^m\,.$$

Für jedes $y \neq 0$ besitzt x_y keine Nullstelle in I. Würde es nämlich ein $\bar{t} \in I$ geben mit $x_y(\bar{t}) = 0$, so löst die Nullfunktion die homogene lineare DGL

$$x'(t) = A(t)x(t),\ t \in I,\ x(\bar{t}) = 0,$$

was ein Widerspruch zur Einzigkeit von x_y wäre. Deshalb gilt für alle $y \in \mathbb{R}^m$, $y \neq 0$, und alle $t \in I$, dass

$$x_y(t) = X(t)y = \sum_{i=1}^{m} y_i x_i(t) \neq 0,$$

d.h. die Spalten von $X(t)$ sind für jedes $t \in I$ linear unabhängig und folglich existiert die Inverse $X^{-1}(t)$ für jedes $t \in I$. Damit ist die Funktion $x(\cdot)$ in der Aussage wohldefiniert. Durch Nachrechnen sieht man, dass $x(\cdot)$ die Aufgabe (2) löst. □

Schließlich kümmern wir uns noch um das asymptotische Verhalten von Lösungen von (2) im Fall einer *konstanten Matrixfunktion* $A(t) = A \in \mathbb{R}^{m\times m}$, $\forall t \in [t_0, +\infty)$. Dazu verwenden wir Satz 1.11 und die dortige Bedingung (*).

Satz 1.16 *Sei $A \in \mathbb{R}^{m\times m}, \lambda_1, \cdots, \lambda_p \in \mathbb{C}$ seien die Eigenwerte von A.*

a) Gilt $\max\limits_{i=1,\cdots,p} Re\,[\lambda_i] < 0$, so existiert ein Skalarprodukt $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^m$ mit zugehöriger Norm $\|\cdot\|$, so daß

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

mit einer Konstanten $\gamma < 0$.

b) Gilt $Re\,[\lambda_i] \leq 0, i = 1, \cdots, p$, und gehören zu rein imaginären Eigenwerten ausschließlich Jordan-Kästchen der Dimension 1, so existiert ein Sklarprodukt $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^m$, so daß

$$\langle Ax, x\rangle \leq 0, \qquad \forall x \in \mathbb{R}^m.$$

Vorbemerkung 1.17

a) Sei $A \in \mathbb{C}^{m\times m}$ hermitisch, d.h. $A = A^* := \bar{A}^T$.
Dann existieren reelle Eigenwerte $\lambda_1, \cdots, \lambda_m$ und Eigenvektoren $v_1, \cdots, v_m \in \mathbb{C}^m$, so daß $\sum\limits_{j=1}^{m}\langle x, v_j\rangle v_j$ und $\langle v_i, v_j\rangle = \delta_{ij}, i \neq j$. Dann gilt

$$\langle Ax, x\rangle = \left\langle \sum_{j=1}^{m}\langle x, v_j\rangle\lambda_j v_j, \sum_{i=1}^{m}\langle x, v_i\rangle v_i \right\rangle = \sum_{j=1}^{m} |\langle x, v_j\rangle|^2\lambda_j \leq \overbrace{\max_{j=1,\cdots,m}\lambda_j}^{=\lambda\max(A)} \|x\|^2.$$

b) Sei $A \in \mathbb{C}^{m\times m}$ beliebig. Für bel. $x \in \mathbb{C}^m$ gilt:

$$\begin{aligned}&\langle Ax, x\rangle = \langle x, A^*x\rangle = \overline{\langle A^*x, x\rangle}\\ &\rightsquigarrow \langle Ax, x\rangle + \langle A^*x, x\rangle = 2\ Re\ \langle Ax, x\rangle\\ &\rightsquigarrow\ Re\ \langle Ax, x\rangle = \langle \tfrac{1}{2}(A + A^*)x, x\rangle \leq \lambda_{\max}(\tfrac{1}{2}(A + A^*))\|x\|^2,\end{aligned}$$

da $\frac{1}{2}(A + A^*)$ hermitisch ist.

c) Es sei $C \in \mathbb{C}^{m\times m}$ invertierbar. Wir betrachten die Funktion

$$s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, s(x,y) = \ Re\ \langle Cx, Cy\rangle, \quad \forall x, y \in \mathbb{R}^m.$$

Beh. s ist auf $\mathbb{R}^m$ ein Skalarprodukt.
Bew.

(i) $s(x,x) = \ Re\ \langle Cx, Cx\rangle = \|Cx\|^2 \geq 0, = 0$ *gdw.* $x = 0$

(ii)

$$\begin{aligned} s(x,y) &= Re\ \langle C^*Cx, y\rangle = \ Re\ \langle x, C^*Cy\rangle \\ &= Re\ \langle C^*Cy, x\rangle = \ Re\ \langle Cy, Cx\rangle = s(y,x) \end{aligned}$$

(iii)

$$\begin{aligned} s(\alpha x + \beta z, y) &= \ Re\ \langle C^*C(\alpha x + \beta z), y\rangle = \ Re\ \langle \alpha C^*Cx + \beta C^*Cz, y\rangle \\ &= \alpha\ Re\ \langle C^*Cx, y\rangle + \beta\ Re\ \langle C^*Cz, y\rangle \\ &= \alpha s(x,y) + \beta s(z,y), \quad \forall \alpha, \beta \in \mathbb{R}, \forall x, y, z \in \mathbb{R}^m. \end{aligned}$$

□

Beweis von Satz 1.16
Es sei $\varepsilon > 0$ beliebig gewählt. $D_\varepsilon = \mathrm{diag}(\varepsilon, \varepsilon^2, \cdots, \varepsilon^m)$, sei J Jordansche Normalform von A, d. h. $J = T^{-1}AT$ mit einer invertierbaren Matrix $T \in \mathbb{C}^{m\times m}$.
Wir definieren

$$J_\varepsilon := D_\varepsilon^{-1} J D_\varepsilon = D_\varepsilon^{-1}T^{-1}ATD_\varepsilon = (\overbrace{TD_\varepsilon}^{=:C_\varepsilon})^{-1}A(TD_\varepsilon) = C_\varepsilon^{-1}AC_\varepsilon$$

J_ε hat die Form $\begin{pmatrix} \lambda_1 & \delta & 0 & \cdots & 0 \\ 0 & \ddots & \delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \ddots & \delta \\ 0 & 0 & \cdots & 0 & \lambda_p \end{pmatrix}, \quad \delta \in \{0, \varepsilon\}.$.

Wir betrachten das Skalarprodukt auf $\mathbb{R}^m$ (nach 1.17 c)

$$\langle x, y\rangle_\varepsilon := \mathrm{Re}\langle C_\varepsilon^{-1}x, C_\varepsilon^{-1}y\rangle, \quad \forall x, y \in \mathbb{R}^m.$$

Dann gilt:

$$\begin{aligned} \langle Ax, x\rangle_\varepsilon &= \mathrm{Re}\langle C_\varepsilon^{-1}Ax, C_\varepsilon^{-1}x\rangle = \mathrm{Re}\langle C_\varepsilon^{-1}AC_\varepsilon C_\varepsilon^{-1}x, C_\varepsilon^{-1}x\rangle \\ &= \mathrm{Re}\langle J_\varepsilon C_\varepsilon^{-1}x, C_\varepsilon^{-1}x\rangle - \langle \tfrac{1}{2}(J_\varepsilon + J_\varepsilon^*)C_\varepsilon^{-1}x, C_\varepsilon^{-1}x\rangle \\ \text{(nach 1.17b)} \ &\leq \lambda_{\max}(\tfrac{1}{2}(J_\varepsilon + J_\varepsilon^*)) \underbrace{\|C_\varepsilon^{-1}x\|^2}_{=\|x\|_\varepsilon^2} \end{aligned}$$

Ferner gilt:

$$\begin{aligned} \tfrac{1}{2}(J_\varepsilon + J_\varepsilon^*) &= \begin{pmatrix} \mathrm{Re}\lambda_1 & \frac{\delta}{2} & 0 \\ \frac{\delta}{2} & \ddots & \frac{\delta}{2} \\ 0 & \frac{\delta}{2} & \mathrm{Re}\ \lambda_p \end{pmatrix} \qquad \text{wobei } \delta \in \{0, \varepsilon\} \\ &\xrightarrow[\varepsilon\to 0]{} \begin{pmatrix} \mathrm{Re}\lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathrm{Re}\ \lambda_p \end{pmatrix} \end{aligned}$$

Daraus folgt:

$$\lambda_{\max}(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*) \overrightarrow{\varepsilon \to 0} \max_{i=1,\cdots,p} \text{Re } \lambda_i$$

a) Es gelte nun $\max_{i=1,\cdots,p} \text{Re } [\lambda_i] < 0$.
Wir wählen $\gamma < 0$, so dass $\max_{i=1,\cdots,p} \text{Re } [\lambda_i] < \gamma < 0$
$\rightsquigarrow \exists \varepsilon > 0 : \lambda_{\max}(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)) \leq \gamma < 0$.
Dann gilt: $\langle Ax, x\rangle_\varepsilon \leq \gamma \|x\|_\varepsilon^2 \quad \forall x \in \mathbb{R}^m$, und
$\rightsquigarrow \langle \cdot, \cdot \rangle_\varepsilon : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ ist Skalarprodukt auf $\mathbb{R}^m$.

b) Es gelte $\max_{i=1,\cdots,p} \text{Re}[\lambda_i] \leq 0$.
Seien $\lambda_1, \ldots, \lambda_r$ die Eigenwerte von A in $i\mathbb{R}$ (d.h. rein imaginär) und $\lambda_{r+1}, \ldots, \lambda_p$ die restlichen Eigenwerte in $\mathbb{C}$. Das bedeutet, dass die Jordan-Kästchen zu λ_i, $i = 1, \ldots, r$, Dimension 1 besitzen und $\max_{i=r+1,\ldots,p} \text{Re}[\lambda_i] < 0$ gilt. Also folgt

$$\frac{1}{2}(J_\varepsilon + J_\varepsilon^*) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \text{Re}(\lambda_{r+1}) & \frac{\delta}{2} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \frac{\delta}{2} & \text{Re}(\lambda_{r+2}) & \frac{\delta}{2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{\delta}{2} & \text{Re}(\lambda_{r+2}) & \frac{\delta}{2} \\ 0 & 0 & 0 & \cdots & 0 & 0 & \frac{\delta}{2} & \text{Re}(\lambda_p) \end{pmatrix}.$$

für $\delta \in \{0, \varepsilon\}$. Das charakteristische Polynom $\varphi(\mu)$ von $\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)$ besitzt die Form

$$\varphi(\mu) = \det\Big(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*) - \mu I\Big) = (-\mu)^r \varphi_\varepsilon(\mu),$$

wobei $\varphi_\varepsilon(\mu)$ das charakteristische Polynom der rechten unteren $(m-r)$-dimensionalen Teilmatrix von $\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)$ ist. Für die rechte untere Teilmatrix läßt sich Teil a) des Beweises anwenden. Deshalb existiert ein ε, so dass $\lambda_{\max}(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*)) \leq 0$ und damit $\langle Ax, x\rangle_\varepsilon \leq 0$, $\forall x \in \mathbb{R}^m$ gilt. □

Folgerung 1.18
Sei $A \in \mathbb{R}^{m\times m}, \lambda_1, \cdots, \lambda_p \in \mathbb{C}$ seien die Eigenwerte von A. Die DGL $x'(t) = Ax(t)$ ist kontraktiv, falls $\max_{i=1,\cdots,p} Re\ [\lambda_i] < 0$.
Die DGL ist schwach kontraktiv, falls $\max_{i=1,\cdots,p} Re\ [\lambda_i] \leq 0$ und alle Jordan-Kästchen zu rein imaginären Eigenwerten Dimension 1 besitzen.
Für die einzige Lösung x mit Anfangswert $x(t_0) = x_0$ gilt dann entweder

$$\lim_{t\to+\infty} \|x(t)\| = 0$$

für jede Norm $\|\cdot\|$ in $\mathbb{R}^m$ oder es existiert eine Norm $\|\cdot\|$ in $\mathbb{R}^m$, so dass

$$\|x(t)\| \leq \|x_0\|,\ \forall t \in [t_0, +\infty).$$

Beweis: Die Aussage folgt aus Satz 1.11, Folgerung 1.14 und Satz 1.16. □

2 Diskretisierung von Operatorgleichungen

Seien $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ lineare normierte Räume. Sei $A : X \to Y$ ein i. a. (nichtlinearer) Operator und 0 das Nullelement in Y. Wir betrachten die Operatorgleichung

$$(1) \qquad Ax = 0\,.$$

Beispiel 2.1
Wir wählen $X = C([t_0, T], \mathbb{R}^m)$ *oder* $X = C^1([t_0, T], \mathbb{R}^m)$ *und* $Y = \mathbb{R}^m \times C([t_0, T], \mathbb{R}^m)$ *sowie*

$$Ax(\cdot) := \begin{pmatrix} r(x(t_0), x(T)) \\ x(\cdot) - x(t_0) - \int_{t_0}^{\cdot} f(x(s), s)ds \end{pmatrix} \text{ oder } Ax(\cdot) := \begin{pmatrix} r(x(t_0), x(T)) \\ x'(\cdot) - f(x(\cdot), \cdot) \end{pmatrix},$$

wobei $f : \mathbb{R}^m \times [t_0, T] \to \mathbb{R}^m$ *und* $r : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ *gegebene stetige Funktionen sind. Im Fall* $r(x, y) = x - x_0$ *für gegebenes* $x_0 \in \mathbb{R}^m$ *entspricht die Gleichung* $Ax = 0$ *einem* Anfangswertproblem *für eine gewöhnliche Differentialgleichung und im allgemeinen Fall einem* Randwertproblem.

Diskretisierung: $\qquad (1n) \qquad A_n x_n = 0 \quad (n \in \mathbb{N}),$

wobei $0 \in Y_n$ das Nullelement, $A_n : X_n \to Y_n$, und $(X_n, \|\cdot\|_{X_n})$, $(Y_n, \|\cdot\|_{Y_n})$ lineare normierte Räume sind.

Ziel: Bedingungen finden, so dass eine Folge von Lösungen von $A_n x_n = 0$ gegen Lösung von $Ax = 0$ konvergiert.

Definition 2.2
Sei $(X, \|\cdot\|)$ *linearer normierter Raum.*
$(X, X_n, r_n)_{n\in\mathbb{N}}$ *heißt diskrete Approximation von* X*, falls*

(i) $(X_n, \|\cdot\|_{X_n})$ *linearer normierter Raum,* $\forall n \in \mathbb{N}$*, ist und*

(ii) $r_n : X \to X_n$ *linearer Operator,* $\forall n \in \mathbb{N}$*, ist, für die gilt*

$$\lim_{n\to\infty} \|r_n x\|_{X_n} = \|x\|_X$$

Die r_n *heißen Restriktionsoperatoren (von* X *in* X_n*).*

Definition 2.3
Sei $(X, X_n, r_n)_{n\in\mathbb{N}}$ *diskrete Approximation von* X*.*
Eine Folge $(x_n)_{n\in\mathbb{N}}$*, mit* $x_n \in X_n$*,* $\forall n \in \mathbb{N}$*, heißt diskret konvergent gegen* $x \in X$*, falls*

$$\lim_{n\to\infty} \|x_n - r_n x\|_{X_n} = 0. \qquad \text{Bez.: } x_n \xrightarrow{d} x\,.$$

Lemma 2.4 *Sei* $(X, X_n, r_n)_{n\in\mathbb{N}}$ *eine diskrete Approximation von* X*. Dann gilt:*

a) $r_n x \xrightarrow{d} x$

b) Der Grenzwert einer diskret konvergenten Folge ist eindeutig bestimmt.

Beweis:
Teil a) ist trivial, da $\|r_n x - r_n x\|_{X_n} = 0$.
b) Annahme: Es existiert eine Folge (x_n), für die $x_n \xrightarrow{d} x$ und $x_n \xrightarrow{d} \tilde{x}$ mit $x \neq \tilde{x}$ gilt. Aus der Dreiecksungleichung für $\|\cdot\|_{X_n}$ folgt

$$\|r_n x - r_n \tilde{x}\|_{X_n} \leq \|r_n x - x_n\|_{X_n} + \|x_n - r_n \tilde{x}\|_{X_n}$$

Deshalb folgt

$$\lim_{n\to\infty} \|r_n x - r_n \tilde{x}\|_{X_n} = 0 = \|x - \tilde{x}\|_X \quad \text{(vgl. Def. 2.2).}$$

□

Beispiel 2.5

a) Sei $(X, \|\cdot\|)$ ein linearer normierter Raum, $X_n \subseteq X$ lineare, endlichdimensionale Teilräume, $r_n : X \to X_n$ lineare Abbildung mit $\lim_{n\to\infty} \|r_n x - x\| = 0$, $\forall x \in X$. Dann ist $(X, X_n, r_n)_{n\in\mathbb{N}}$ eine diskrete Approximation von X.
Bew.: Ist (x_n) Folge in X mit $x_n \xrightarrow{d} x$, d.h., $\lim_{n\to\infty} \|x_n - x\| = 0$. Es gilt

$$\|x_n - r_n x\| \leq \|x_n - x\| + \|x - r_n x\|$$

d.h. $\lim_{n\to\infty} \|x_n - r_n x\| = 0$.

b) Wir konstruieren eine diskrete Approximation von $X = C([a,b], \mathbb{R}^m)$. Dazu betrachten wir eine Folge von Unterteilungen $t_0^{(n)} = a < t_1^{(n)} < \cdots < t_{k_n}^{(n)} = b$, $\forall n \in \mathbb{N}$, von $[a,b]$ mit $h_n := \max_{i=1,\ldots,k_n} |t_i^{(n)} - t_{i-1}^{(n)}| \longrightarrow 0$ für $n \to \infty$. Wir definieren

$$\begin{aligned} X_n &:= \mathbb{R}^{(k_n+1)m} \quad \textit{mit} \quad \|x_n\|_{X_n} := \max_{i=0,\ldots,k_n} \|x_{ni}\|, \\ r_n &: X \to X_n \quad \textit{mit} \quad r_n x := (x(t_0^{(n)}), \ldots, x(t_{k_n}^{(n)})) \quad (\forall n \in \mathbb{N}). \end{aligned}$$

Dann ist r_n linear und es gilt $\|r_n x\| = \max_{i=0,\ldots,k_n} \|x(t_i^{(n)}\|$ für jedes $n \in \mathbb{N}$. Außerdem gilt $\|r_n x\|_{X_n} \leq \|x\|_X = \max_{t\in[a,b]} \|x(t)\|$ und damit zunächst

$$\limsup_{n\to\infty} \|r_n x\|_{X_n} \leq \|x\|_X .$$

Ist $t_* \in [a,b]$ so gewählt, dass $\|x(t_*)\| = \|x\|_X$, so existiert für jedes $n \in \mathbb{N}$ ein $i(n) \in \{0, \ldots, k_n\}$ mit $t_* \in [t_{i-1}^{(n)}, t_i^{(n)}]$. Es folgt $0 \leq t_{i(n)}^{(n)} - t_* \leq h_n$ und damit die Konvergenz von $(t_{i(n)}^{(n)})$ gegen t_*. Daraus schlussfolgern wir

$$\liminf_{n\to\infty} \|r_n x\|_{X_n} \geq \liminf \|x(t_{i(n)}^{(n)})\| = \|x(t_*)\| = \|x\|_X .$$

also gilt $\lim_{n\to\infty} \|r_n x\|_{X_n} = \|x\|_X$. □

c) Wir betrachten $X = L_p([a,b],\mathbb{R}^m)$ *mit der Norm* $\|x\|_X = (\int_a^b |x(t)|^p dt)^{\frac{1}{p}}$ *für* $1 \le p < +\infty$. *Wir definieren eine diskrete Approximation durch*

$$X_n := \mathbb{R}^{mk_n} \quad \text{mit der Norm} \quad \|x_n\|_{X_n} := \Big(\sum_{i=1}^{k_n} h_{n,i}\|x_{ni}\|^p\Big)^{\frac{1}{p}},$$

$$r_n \;:\; X \to X_n, \quad r_n x = \Big(\frac{1}{h_{n,1}}\int_{t_0^n}^{t_1^{(n)}} x(t)dt, \ldots, \frac{1}{h_{n,k_n}}\int_{t_{k_n-1}^n}^{t_{k_n}^{(n)}} x(t)dt\Big).$$

Hierbei ist $h_{n,i} = t_i^{(n)} - t_{i-1}^{(n)}$, $i = 1,\ldots,k_n$, *und* $t_0^{(n)} = a$ *bzw.* $t_{k_n}^{(n)} = b$.
Es lassen sich wieder alle Eigenschaften einer diskreten Approximation nachweisen (vgl. Buch von Vainikko, Kap.2; Übung).

Definition 2.6
Seien $(X, X_n, r_n)_{n\in\mathbb{N}}$ *und* $(Y, Y_n, \tilde{r}_n)_{n\in\mathbb{N}}$ *diskrete Approximationen.*
Seien $A : X \to Y$ *und* $A_n : X_n \to Y_n$, $n \in \mathbb{N}$.

(i) A *und* $(A_n)_{n\in\mathbb{N}}$ *heißen konsistent in* $x \in X$, *falls*

$$\lim_{n\to\infty} \|A_n r_n x - \tilde{r}_n A x\|_{Y_n} = 0$$

(konsistent, falls dies für jedes $x \in X$ *gilt).*

(ii) diskret konvergent in $x \in X$, *falls für jede gegen* $x \in X$ *diskret konvergente Folge* $(x_n)_{n\in\mathbb{N}}$ *gilt*

$$\lim_{n\to\infty} \|A_n x_n - \tilde{r}_n A x\|_{Y_n} = 0.$$

(diskret konvergent, falls dies für alle $x \in X$*) gilt)*

(iii) Die Folge (A_n) *heißt invers stabil, falls eine Konstante* $S > 0$ *und ein* $n_0 \in \mathbb{N}$ *existieren, so dass für alle* $n \ge n_0$ *und alle* $x_n, \tilde{x}_n \in X_n$ *gilt*

$$\|x_n - \tilde{x}_n\|_{X_n} \le S\|A_n x_n - A_n \tilde{x}_n\|_{Y_n}.$$

Bemerkung 2.7
Inverse Stabilität von (A_n) *gilt, falls für alle* $n \ge n_0$ *die Operatoren* A_n *injektiv und* A_n^{-1} *gleichmäßig Lipschitzstetig auf dem Wertebereich* $R(A_n)$ *von* A_n *ist.*
Eine Folge linearer Operatoren A_n *ist folglich invers stabil, falls* $\|A_n^{-1}\|$ *gleichmäßig beschränkt ist !*

Satz 2.8 *(Konvergenzsatz)*
Es seien $(X, X_n, r_n)_{n\in\mathbb{N}}$ *und* $(Y, Y_n, \tilde{r}_n)_{n\in\mathbb{N}}$ *diskrete Approximationen und* $A : X \to Y$ *sowie* $A_n : X_n \to Y_n$, $n \in \mathbb{N}$, *Operatoren.*
Die Folge (A_n) *sei invers stabil und* A *und* (A_n) *seien konsistent in einer Lösung* x^* *von* $Ax = 0$.
Sind $(x_n^*)_{n\ge n_1}$ *Lösungen von* $A_n x_n = 0$, *so gilt*

$$\lim_{n\to\infty} \|x_n^* - r_n x^*\|_{X_n} = 0, \;\; \text{d.h. } x_n^* \xrightarrow{d} x^*$$

und die Fehlerabschätzung für große n:

$$\|x_n^* - r_n x^*\|_{X_n} \leq S\|A_n r_n x^*\|_{Y_n}$$

Beweis: Seien $S > 0$ und $n_0 \in \mathbb{N}$ wie in Def. 2.6 (iii) gewählt. Dann gilt für $n \geq \max\{n_0, n_1\}$:

$$\|x_n^* - r_n x^*\|_{X_n} \leq S\|A_n x_n^* - A_n r_n x^*\|_{Y_n} = S\|A_n r_n x^*\|_{Y_n} = S\|A_n r_n x^* - \tilde{r}_n A x^*\|_{Y_n}$$

Die rechte Seite konvergiert gegen 0 wegen der vorausgesetzten Konsistenz. □

Bemerkung 2.9
Untersuchung der Konvergenzgeschwindigkeit von $\|A_n r_n x^*\|_{Y_n}$ ist wichtig!
Z.B. unter gewissen Voraussetzungen an die Lösung x^*.
Inverse Stabilität ist eine starke Bedingung, läßt sich aber in einer Reihe von Anwendungen nachweisen.

Definition 2.10
Eine Folge $(x_n)_{n\in\mathbb{N}}, x_n \in X_n, \forall n \in \mathbb{N}$, heißt diskret kompakt, falls für jede Teilmenge $\mathbb{N}' \subseteq \mathbb{N}$ ein $\mathbb{N}'' \subseteq \mathbb{N}'$ und ein $x \in X$ existieren, so daß

$$\lim_{\substack{n\to\infty \\ n\in\mathbb{N}''}} \|x_n - r_n x\|_{X_n} = 0.$$

Definition 2.11
Eine Folge von $(A_n)_{n\in\mathbb{N}}$ von Operatoren $A_n : X_n \to Y_n, n \in \mathbb{N}$, konvergiert regulär gegen $A : X \to Y$, falls

(i) $(A_n)_{n\in\mathbb{N}}$ diskret gegen A konvergiert.

(ii) Aus $\|x_n\|_{X_n} \leq$ const und der diskreten Kompaktheit von $(A_n x_n)_{n\in\mathbb{N}}$ folgt, daß $(x_n)_{n\in\mathbb{N}}$ diskret kompakt ist.

Bemerkung 2.12
Die reguläre Konvergenz ist insbesondere für lineare, beschränkte Operatoren von Bedeutung. Dafür brauchen wir weitere Begriffe. Es seien X und Y lineare normierte Räume und
$A : X \to Y$ linear und beschtänkt, d.h. $A \in L(X, Y)$.
$N(A) := \{x \in X : Ax = 0\}$ Nullraum von A,
$R(A) := \{Ax : x \in X\}$ Wertebereich von A.
Beides sind lineare Teilräume von X bzw. Y.

A heißt fredholmsch, falls $R(A)$ abgeschlossen ist und $\dim N(A)$ sowie codim $R(A)$ endlich sind. Man definiert seinen Index

$$ind\ (A) := \dim N(A) - codim\ R(A).$$

(Es gilt codim $R(A) = n$, falls n linear unabhängige Elemente $y_1, \ldots, y_n$ in $Y \setminus R(A)$ existieren, so dass $Y = span\{R(A), y_1, \ldots, y_n\}$.)
Sind X und Y endlichdimensional, so ist jeder lineare Operator A fredholmsch mit $ind\ (A) = \dim X - \dim Y$.

Satz 2.13
Es seien $A \in L(X,Y)$ und $A_n \in L(X_n, Y_n), \forall n \in \mathbb{N}$. Dann sind die folgenden beiden Aussagen äquivalent:
(i) (A_n) *konvergiert diskret gegen* A.
(ii) $\|A_n\| \leq const (\forall n \in \mathbb{N})$ *und* $\lim\limits_{n\to\infty} \|A_n r_n x - \tilde{r}_n A x\|_{Y_n} = 0, \forall x \in X$.

Beweis:

(i) $\Rightarrow$ (ii): Annahme: $\exists \mathbb{N}' \subseteq \mathbb{N}: \quad \|A_n\| \xrightarrow[n\in\mathbb{N}']{} +\infty$
$\rightsquigarrow \exists x'_n \in X_n : \|x'_n\|_{X_n} = 1$ und $\|A_n x'_n\|_{Y_n} \xrightarrow[n\in\mathbb{N}']{} +\infty$
$(n \in \mathbb{N}')$
$\rightsquigarrow (x_n), x_n := \frac{x'_n}{\|Ax'_n\|_{Y_n}} \xrightarrow{d} 0$, aber
$\|A_n x_n\|_{y_n} = 1, \forall n \in \mathbb{N}' \not\rightsquigarrow A_n x_n \xrightarrow{d} A0 = 0$
$\rightsquigarrow (A_n)$ konvergiert nicht diskret gegen A. $\rightsquigarrow$ Widerspruch.

(ii) $\Rightarrow$ (i): Sei (x_n) diskret gegen $x \in X$ konvergent.
Zu zeigen: $\|A_n x_n - \tilde{r}_n A x\|_{Y_n} \to 0$.

$$\begin{aligned} \|A_n x_n - \tilde{r}_n A x\|_{Y_n} &\leq \|A_n x_n - A_n r_n x\|_{Y_n} + \|A_n r_n x - \tilde{r}_n A x\|_{Y_n} \\ &\leq \|A_n\| \|x_n - r_n x\|_{X_n} + \|A_n r_n x - \tilde{r}_n A x\|_{Y_n} \qquad \square \\ &\leq C \quad \overrightarrow{n \to \infty}\ 0 \qquad \overrightarrow{n \to \infty}\ 0 \end{aligned}$$

Satz 2.14
Für $A \in L(X,Y), A_n \in L(X_n, Y_n)$ gelte: $N(A) = \{0\}$, d.h. A ist injektiv, (A_n) konvergiert regulär gegen A und die A_n sind fredholmsch mit ind $(A_n) = 0, \forall n \geq n_0$.
Dann ist A surjektiv und es gilt $A^{-1} \in L(Y,X)$. Für genügend große n sind auch die A_n bijektiv und es gilt $\|A_n^{-1}\| \leq const.$, d.h. (A_n) ist invers stabil.

Beweis:
Wir zeigen zunächst: $\exists \gamma > 0 \quad \exists n_0 \in \mathbb{N}:$
$\|A_n x_n\|_{Y_n} \geq \gamma \|x_n\|_{X_n}, \forall x_n \in X_n, \forall n \geq n_0$.
Annahme: $\exists$ Folge $(x_n)_{n\in\mathbb{N}'}$ mit $\|x_n\|_{X_n} = 1$ und $\|A_n x_n\|_{Y_n} \xrightarrow[n\in\mathbb{N}']{} 0$
Weil (A_n) regulär gegen A konvergiert, ist $(x_n)_{n\in\mathbb{N}'}$ diskret kompakt
$\rightsquigarrow \exists \mathbb{N}'' \subseteq \mathbb{N}' : x_n \xrightarrow[n\in\mathbb{N}'']{d} x$ und $\|x\|_X = 1$
$\rightsquigarrow A_n x_n \xrightarrow[n\in\mathbb{N}'']{d} Ax \neq 0$ (wegen $N(A) = \{0\}$) $\rightsquigarrow$ Widerspruch.
$\rightsquigarrow A_n$ ist injektiv und es gilt $A_n^{-1} : R(A_n) \to X_n$ sowie $\|A_n^{-1}\| \leq \frac{1}{\gamma}$.
Wegen $N(A_n) = \{0\}$ und ind $(A_n) = 0$ folgt aber codim $R(A_n) = 0$
$\rightsquigarrow R(A_n) = Y_n \rightsquigarrow A_n^{-1} \in L(Y_n, X_n)$.
Betrachtet man für ein beliebig gewähltes $x \in X$ die Folge $(r_n x)$, so gilt $r_n x \xrightarrow{d} x$ und $A_n r_n x \xrightarrow{d} Ax$. Deshalb kann man in der Ungleichung

$$\|A_n r_n x\|_{Y_n} \geq \gamma \|r_n x\|_{X_n}$$

zum Grenzwert $n \to \infty$ übergehen und erhält

$$\|Ax\|_Y \geq \gamma \|x\|_X \quad \text{d.h.} \quad \|A^{-1} y\|_X \leq \frac{1}{\gamma} \|y\|_Y \quad (\forall y \in R(A)).$$

Noch zu zeigen ist: $R(A) = Y$. Sei $y \in Y$. Dann gilt $\tilde{r}_n y \xrightarrow{d} y$.
Für $x_n := A_n^{-1} r_n y$ gilt $\|x_n\|_{X_n} \leq \frac{1}{\gamma} \|r_n y\|_{Y_n}, \forall n \geq n_0$
$\rightsquigarrow (\|x_n\|)$ ist beschränkt und $A_n x_n = r_n y \xrightarrow{d} y$.
Da (A_n) regulär gegen A konvergiert, ist die Folge (x_n) diskret kompakt
$\rightsquigarrow \exists \mathbb{N}' \subseteq \mathbb{N} : x_n \overrightarrow{n \in \mathbb{N}'}$, $x \in X$
$\rightsquigarrow A_n x_n = r_n y \overrightarrow{n \to \infty} Ax \rightsquigarrow Ax = y$ (Lemma 2.4)
$\rightsquigarrow y \in R(A)$ und damit $R(A) = Y$, d.h. A ist surjektiv. □

Wir betrachten jetzt lineare Operatorgleichungen

$$(2) \quad Ax = y, \text{ wobei } A \in L(X, Y), y \in Y$$

und ihre Diskretisierungen

$$(2n) \quad A_n x_n = y_n, \text{ wobei } A_n \in L(X_n, Y_n), y_n \in Y_n, n \in \mathbb{N}$$

Satz 2.15 *(Konvergenzsatz für Diskretisierungen linearer Gleichungen)*
Es seien die folgenden Bedingungen erfüllt:
(i) $A \in L(X, Y), N(A) = \{0\}, y \in Y$.
(ii) $A_n \in L(X_n, Y_n) (n \in \mathbb{N})$ *sind fredholmsch mit Index* 0,
(iii) (A_n) *konvergiert gegen* A *regulär,*
(iv) $y_n \xrightarrow{d} y$.
Dann hat die Gleichung (2) genau eine Lösung $x^* \in X$.
Überdies existiert ein $n_0 \in \mathbb{N}$, *so daß die Gleichungen* $(2n)$ *für* $n \geq n_0$ *eine eindeutig bestimmte Lösung* x_n^* *besitzen und es gilt*

$$x_n^* \xrightarrow{d} x^*$$

sowie

$$C_1 \|A_n r_n x^* - y_n\|_{Y_n} \leq \|x_n^* - r_n x^*\|_{X_n} \leq C_2 \|A_n r_n x * - y_n\|_{Y_n}$$

für alle $n \geq n_0$ *mit gewissen Konstanten* $C_1, C_2 > 0$.

Beweis:
Nach Satz 2.13 existieren $A^{-1} \in L(Y, X)$ und $A_n^{-1} \in L(Y_n, X_n)$ und es gilt $\|A_n^{-1}\| \leq C_2 = \text{ const. }, \forall n \geq n_0$.
Damit sind die Gleichungen (2) und (2n), $n \geq n_0$, eindeutig lösbar mit Lösungen $x^* \in X$ bzw. $x_n^* \in X_n$.
Nach Satz 2.12 gilt aber auch $\|A_n\| \leq \frac{1}{c_1} = \text{const.}$ $(n \in \mathbb{N})$. Wegen $A_n(x_n^* - r_n x^*) = y_n - A_n r_n x^*$ folgt

$$\begin{aligned} \|y_n - A_n r_n x^*\|_{Y_n} &\leq \|A_n\| \|x_n^* - r_n x^*\|_{X_n} \\ &\leq \tfrac{1}{C_1} \|x_n^* - r_n x^*\|_{X_n} \end{aligned}$$

und $\|x_n^* - r_n x^*\|_{X_n} \leq \|A_n^{-1}(y_n - A_n r_n x^*)\|_{X_n} \leq C_2 \|y_n - A_n r_n x^*\|_{Y_n}$. Da überdies $\|A_n r_n x^* - \tilde{r}_n A x_*\|_{Y_n} = \|A_n r_n x^* - \tilde{r}_n y\|_{Y_n} \to 0$ wegen der diskreten Konvergenz von (A_n) gegen A, folgt

$$\|A_n r_n x^* - y_n\|_{Y_n} \leq \underbrace{\|A_n r_n x^* - \tilde{r}_n y\|_{Y_n}}_{n \to \infty \; 0} + \underbrace{\|\tilde{r}_n y - y_n\|_{Y_n}}_{(iv) \; 0}$$

□

Bemerkung 2.16
Da nach Satz 2.14 die (A_n) invers stabil sind, folgt die diskrete Konvergenz $x_n^* \xrightarrow{d} x^*$ auch aus Satz 2.8. Neu ist hier, daß man eine Fehlerabschätzung für

$$\|r_n x^* - x_n^*\|_{X_n}$$

hat, die beidseitig ist und zeigt, daß die gewünschte Konvergenz von $(\|r_n x^* - x_n^*\|_{X_n})$ genau so schnell erfolgt wie die von $(\|A_n r_n x^* - y_n\|_{Y_n})$ gegen 0. Wählt man $y_n = \tilde{r}_n y$, so gilt

$$\|A_n r_n x^* - y_n\|_{Y_n} = \|A_n r_n x^* - \tilde{r}_n A x^*\|_{Y_n}$$

d. h. es tritt wieder der "Konsistenzterm" in x^* auf!

Frage: Wie läßt sich ein solches Resultat auf nichtlineare Operatorgleichungen (1) und (1n) erweitern?
Der Einfachheit halber nehmen wir an, daß auch die nichtlinearen Operatoren A bzw. A_n auf X bzw. X_n definiert sind.

Definition 2.17
Ein Operator $A : X \to Y$ heißt Frechét-differenzierbar in $x_0 \in X$, falls ein Operator $L_{x_0} \in L(X, Y)$ existiert, so daß für alle x aus einer Umgebung von x_0 $Ax = Ax_0 + L_{x_0}(x - x_0) + \omega(x, x_0)$ gilt, wobei $\frac{\|\omega(x,x_0)\|_y}{\|x - x_0\|} \to 0$ für $x \to x_0$.
Bezeichnung: $A'(x_0) := L_{x_0}$

Für die Frechét Ableitung gelten die Eigenschaften:

(i) Sind A_1 und A_2 Frechét-differenzierbar in x_0, so auch $A_1 + A_2$.

(ii) Ist $B \in L(X, Y)$, so ist B Frechét-differenzierbar in jedem $x_0 \in X$ und es gilt $B'(x_0) = B$.

(iii) Sind A_1 und A_2 Frechét-differenzierbar in x_0, so auch ihre Komposition $A_2 \circ A_1$. Dabei gilt $(A_2 \circ A_1)'(x_0) = A_2'(A_1(x_0)) A_1'(x_0)$.

Wir benötigen im folgenden eine erweiterte Version des Mittelwertsatzes für die Frechét-Ableitung.

Lemma 2.18 *(verallgemeinerter Mittelwertsatz)*
Es sei $A : X \to Y$ Frechét-differenzierbar in jedem Element einer Kugel $B(x_0, r) \subseteq X$ mit Mittelpunkt $x_0 \in X$ und Radius $r > 0$. Ist $B \in L(X, Y)$, so gilt $\forall\, x, \tilde{x} \in B(x_0, r)$:

$$\|A(\tilde{x}) - A(x) - B(\tilde{x} - x)\|_Y \leq \sup_{t \in [0,1]} \|A'(x + t(\tilde{x} - x)) - B\| \|\tilde{x} - x\|_X$$

Beweis: Es seien $x, \tilde{x} \in B(x_0, r)$ beliebig.
Wir betrachten die folgende Funktion $\varphi : [0,1] \to Y$

$$\varphi(t) := A(x + t(\tilde{x} - x)) - B(x + t(\tilde{x} - x)), \forall t \in [0,1].$$

Die Abbildung ist wohldefiniert wegen der Konvexität von $B(x_0, r)$ und es folgt aus der Kettenregel

$$\varphi'(t) = A'(x + t(\tilde{x} - x))(\tilde{x} - x) - B(\tilde{x} - x), \forall t \in [0,1].$$

Dann gilt der folgende Mittelwertsatz für $\varphi(\cdot)$:

$$\|\varphi(1) - \varphi(0)\|_Y \leq \sup_{t \in [0,1]} \|\varphi'(t)\|_Y$$

Bew.: Wir setzen $M := \sup_{t \in [0,1]} \|\varphi'(t)\|$ und nehmen an, daß $M < +\infty$ (ansonsten ist die Aussage trivial!). Sei $\varepsilon > 0$ beliebig gewählt.
Wir zeigen: $\|\varphi(1) - \varphi(0)\| \leq M + \varepsilon$.
Dazu definieren wir die folgende Menge:
$I := \{t \in [0,1] : \|\varphi(s) - \varphi(0)\| \leq (M + \varepsilon)s, \quad \forall s \in [0,t]\}$
Beh.: I ist ein abgeschlossenes Intervall mit $0 \in I$.
Bew.: $0 \in I$ ist klar nach Definition. Sei $t \in I$ bel. und $s \in [0, t[$.
$\rightsquigarrow \|\varphi(\tau) - \varphi(0)\| \leq (M + \varepsilon)\tau \quad \forall \tau \in [0, s]$, da $s < t$
$\rightsquigarrow s \in I \rightsquigarrow I$ ist ein Intervall der Gestalt $[0, \gamma[$ oder $[0, \gamma]$.
sei $s \in [0, \gamma] \rightsquigarrow \|\varphi(s) - \varphi(0)\| \leq (M + \varepsilon)s$
$s \to \gamma : \|\varphi(\gamma) - \varphi(0)\| \leq (M + \varepsilon)\gamma \rightsquigarrow \gamma \in I$
$\rightsquigarrow I$ ist auch abgeschlossen (da φ stetig ist).
Folglich hat I die Gestalt $I = [0, \gamma]$ mit $\gamma \in [0,1]$.
Annahme: $\gamma < 1 \rightsquigarrow \exists \delta > 0 : \gamma + \delta < 1$.
Nach Voraussetzung kann δ noch so klein gewählt werden, daß

$$\left\| \frac{\varphi(\gamma + h) - \varphi(\gamma)}{h} - \varphi'(\gamma) \right\| \leq \varepsilon, \quad \text{falls } h \in]0, \delta].$$

$\rightsquigarrow$ für $h \in]0, \delta]$ gilt:

$$\|\varphi(\gamma + h) - \varphi(\gamma)\| \leq \|\varphi'(\gamma)\|h + \varepsilon h \leq (M + \varepsilon)h$$

$$\begin{aligned} \rightsquigarrow \|\varphi(\gamma + h) - \varphi(0)\| &\leq \|\varphi(\gamma + h) - \varphi(\gamma)\| + \|\varphi(\gamma) - \varphi(0)\| \\ &\leq (M + \varepsilon)h + (M + \varepsilon)\gamma = (M + \varepsilon)(h + \gamma). \end{aligned}$$

$\rightsquigarrow \gamma + \delta \in I \rightsquigarrow$ Widerspruch!
$\rightsquigarrow I = [0,1]$ und $\|\varphi(1) - \varphi(0)\| \leq M + \varepsilon$.
Da $\varepsilon > 0$ beliebig gewählt war, ist der Mittelwertsatz für φ gezeigt.
Daraus folgt dann

$$\begin{aligned} \|\varphi(1) - \varphi(0)\| &= \|A(\tilde{x}) - A(x) - B(\tilde{x} - x)\| \\ &\leq \sup_{t \in [0,1]} \|A'(x + t(\tilde{x} - x)) - B\| \|x - \tilde{x}\|_X. \end{aligned}$$

□

Lemma 2.19 *Es sei X_n ein Banachraum mit Norm $\|\cdot\|$.*
Der Operator $A_n : X_n \to Y_n$ sei Frechét-differenzierbar in der Kugel
$B(x_n^0, \delta_0) := \{x_n \in X_n : \|x_n - x_n^0\| \leq \delta_0\}$.
Es existiere $[A'(x_n^0)]^{-1} \in L(Y_n, X_n)$, wobei für gewisse $\tau, \kappa > 0$ und $\varrho \in [0,1)$:

$$\|A_n'(x_n^0)\| \leq \tau, \|[A_n'(x_n^0)]^{-1}\| \leq \kappa$$

$$\sup_{x_n \in B(x_n^0, \delta_0)} \|A_n'(x_n) - A_n'(x_n^0)\| \leq \frac{\varrho}{\kappa}$$

$$\|A_n(x_n^0)\| \leq \frac{\delta_0(1-\varrho)}{\kappa}$$

gelte. Dann hat die Gleichung $A_n x_n = 0$ eine eindeutig bestimmte Lösung x_n^* in $B(x_n^0, \delta_0)$ und es gilt

$$\frac{\alpha_n}{1+\varrho} \leq \|x_n^* - x_n^0\| \leq \frac{\alpha_n}{1-\varrho}$$

wobei $\alpha_n := \|[A_n'(x_n^0)]^{-1}(A_n x_n^0)\|$ und $\frac{1}{\tau}\|A_n x_n^0\| \leq \alpha_n \leq \kappa \|A_n x_n^0\|$.

Beweis:
Die Gleichung $A_n x_n = 0$ ist äquivalent zur Fixpunkt-Gleichung

$$x_n = \hat{A}_n x_n$$

wobei $\hat{A}_n : X_n \to X_n, \hat{A}_n x_n := x_n - [A_n'(x_n^0)]^{-1} A_n x_n$.

Wir zeigen: $\hat{A}_n(B(x_n^0, \delta_0) \subseteq B(x_n^0, \delta_0)$ und $\hat{A}_n$ ist kontraktiv.

(a) Sei $x_n \in B(x_n^0, \delta_0)$ beliebig gewählt.

$$\begin{aligned}
\|\hat{A}_n x_n - x_n^0\| &= \|x_n - x_n^0 - [A_n'(x_n^0)]^{-1} A_n x_n\| \\
&\leq \|[A_n'(x_n^0)]^{-1}\| \|A_n'(x_n^0)(x_n - x_n^0) - A_n x_n\| \\
&\leq \kappa(\| - A_n'(x_n^0)(x_n - x_n^0) + A_n x_n - A_n x_n^0\| + \|A_n x_n^0\|) \\
&\overset{2.18}{\leq} \kappa\Big(\sup_{t\in[0,1]} \|A_n'(x_n^0 + t(x_n - x_n^0)) - A_n'(x_n^0)\| \|x_n - x_n^0\| + \|A_n x_n^0\|\Big) \\
&\leq \kappa\left(\tfrac{\varrho}{\kappa}\delta_0 + \tfrac{1-\varrho}{\kappa}\delta_0\right) = \delta_0
\end{aligned}$$

(b) Seien $x_n, \tilde{x}_n \in B(x_n^0, \delta_0)$:

$$\begin{aligned}
\|\hat{A}_n x_n - \hat{A}_n \tilde{x}_n\| &= \|x_n - \tilde{x}_n - [A_n'(x_n^0)]^{-1}(A_n x_n - A_n \tilde{x}_n)\| \\
&\leq \|[A_n'(x_n^0)]^{-1}\| \|A'(x_n^0)(x_n - \tilde{x}_n) - (A_n x_n - A_n \tilde{x}_n)\| \\
&\overset{2.18}{\leq} \kappa \tfrac{\varrho}{\kappa} \|\tilde{x}_n - x_n\| = \varrho \|\tilde{x}_n - x_n\|.
\end{aligned}$$

Die Kugel $B(x_n^0, \delta_0)$ ist abgeschlossen im Banachraum X_n und deshalb versehen mit der durch die Norm definierten Metrik ein vollständiger metrischer Raum.
Die Aussage über die eindeutige Existenz von x_n^* folgt aus dem Fixpunktsatz von

Banach. Die Abschätzungen für α_n sind klar wegen der Definition der Konstanten τ und κ. Es bleibt, die Abschätzung für $\|x_n^* - x_n^0\|$ zu beweisen. Es gilt

$$\begin{aligned} \|\hat{A}_n x_n^0 - \hat{A}_n x_n^*\| &= \|x_n^0 - x_n^* - [A_n'(x_n^0)]^{-1}(A_n x_n^0 - \underbrace{A_n x_n^*}_{=0})\| \\ &\leq \varrho \|x_n^0 - x_n^*\| \end{aligned}$$

$$\rightsquigarrow \left| \|x_n^* - x_n^0\| - \|[A'(x_n^0)]^{-1}(A_n x_n^0)\| \right| \leq \varrho \|x_n^* - x_n^0\|$$

$$\rightsquigarrow \frac{\alpha_n}{1+\varrho} \leq \|x_n^* - x_n^0\| \leq \frac{\alpha_n}{1-\varrho}$$

□

Satz 2.20 *(Konvergenzsatz für nichtlineare differenzierbare Operatorgleichungen) Es seien die folgenden Voraussetzungen erfüllt:*

(i) *Die Gleichung (1) $Ax = 0$ besitzt eine Lösung $x^* \in X$.*

(ii) *Der Operator $A : X \to Y$ ist in x^* Frechét-differenzierbar.*

(iii) $N(A'(x^*)) = \{0\}$.

(iv) *Die Operatoren $A_n : X_n \to Y_n$ sind in den Kugeln $B(r_n x^*, \delta)$ Frechét-differenzierbar, wobei $\delta > 0$ nicht von n abhängt.*

(v) *Für jedes $\varepsilon > 0$ existiert ein $\delta(\varepsilon) \in (0, \delta]$, so daß für alle $n \in \mathbb{N}$ gilt:*

$$\|x_n - r_n x^*\|_{X_n} \leq \delta(\varepsilon) \Rightarrow \|A_n'(x_n) - A_n'(r_n x^*)\| \leq \varepsilon$$

(vi) *Die Operatoren $A_n'(r_n x^*) \in L(X_n, Y_n)$ sind fredholmsch mit $ind\,(A_n'(r_n x^*)) = 0$ und die X_n sind Banachräume für jedes $n \in \mathbb{N}$.*

(vii) *Die Folge $(A_n'(r_n x^*))$ konvergiert regulär gegen $A'(x^*)$*

(viii) $A_n r_n x^* \xrightarrow{d} A x^*$.

Dann existieren ein $n_0 \in \mathbb{N}$ und ein $\delta_0 \in (0, \delta]$, so daß die Gleichungen

$$A_n x_n = 0$$

in $B(r_n x^*, \delta_0)$ eine eindeutig bestimmte Lösung x_n^* besitzen. Es gilt:

$$x_n^* \xrightarrow{d} x^*$$

und es gilt die Fehlerabschätzung

$$C_1 \|A_n r_n x^*\|_{Y_n} \leq \|x_n^* - r_n x^*\|_{X_n} \leq C_2 \|A_n r_n x^*\|_{Y_n}$$

mit geeigneten Konstanten $C_1, C_2 > 0$.

Beweis:
Aus den Sätzen 2.13 und 2.14 und den Bedingungen (iii),(vi) und (vii) folgt:

$$\|A_n'(r_n x^*)\| \leq \tau, \quad \|[A_n'(r_n x^*)]^{-1}\| \leq \kappa, \quad \forall n \geq n_0$$

mit Konstanten τ und κ, die unabhängig von n sind. Mit $x_n^0 := r_n x^*$ wenden wir jetzt für jedes n Lemma 2.19 an. Dabei sei $\varrho \in (0,1)$ vorgegeben.
Aus (v) folgt, daß ein $\delta_0 \in (0, \delta]$ exisitert, so daß für jedes $n \in \mathbb{N}$ die Bedingung

$$\sup_{x_n \in B(x_n^0, \delta_0)} \|A_n'(x_n) - A_n'(r_n x^*)\| \leq \frac{\varrho}{\kappa}$$

erfüllt ist. Wegen (viii) läßt sich für hinreichend großes $n_0 \in \mathbb{N}$ auch zeigen

$$\|A_n r_n x^*\|_{Y_n} \leq \frac{\delta_0(1-\varrho)}{\kappa}, \quad \forall n \geq n_0.$$

Dann folgt aus Lemma 2.19 für $n \geq n_0$ die eindeutige Existenz einer Lösung x_n^* von $A_n x_n = 0$ in $B(r_n x^*, \delta_0)$. Überdies gilt

$$\frac{\|A_n r_n x^*\|_{Y_n}}{\tau(1+\varrho)} \leq \|x_n^* - r_n x^*\|_{X_n} \leq \frac{\kappa}{1-\varrho} \|A_n r_n x^*\|_{Y_n}.$$

Die Konvergenz $x_n^* \xrightarrow{d} x^*$ folgt aus $A_n r_n x_n^* \xrightarrow{d} A x^* = 0$. □

Bemerkung 2.21
Unter den Voraussetzungen von Satz 2.20 gilt nach Satz 2.14, dass $A'(x^*)$ surjektiv (d.h. bijektiv) ist und $(A'(r_n x^*))$ invers stabil.

Lemma 2.22 *(Isoliertheit von Lösungen)*
Seien $x^* \in X$ eine Lösung von (1), $A : X \to Y$ Frechét-differenzierbar in x^* und es gelte $[A'(x^*)]^{-1} \in L(X,Y)$.
Dann existiert eine $\delta > 0$, so dass die Gleichung (1) keine weitere Lösung in $B(x^*, \delta)$ besitzt, d. h. die Lösung x^* ist isoliert!

Beweis: Annahme: $\exists$ Folge (x_n^*) von Lösungen von (1) mit $x_n^* \to x^*$.
Wegen (ii) gilt dann

$$\underbrace{A x_n^*}_{=0} = \underbrace{A x^*}_{=0} + A'(x^*)(x_n^* - x^*) + \omega(x_n^*, x^*).$$

Daraus folgt $A'(x^*)(x_n^* - x^*) = -\omega(x_n^*, x^*)$.
Da $A'(x^*)$ bijektiv und $[A'(x^*)]^{-1}$ linear und beschränkt ist, folgt

$$\frac{x_n^* - x^*}{\|x_n^* - x^*\|_X} = [A'(x^*)]^{-1} \left(\frac{-\omega(x_n^*, x^*)}{\|x_n^* - x^*\|_X} \right),$$

und damit

$$1 = \frac{\|x_n^* - x^*\|_X}{\|x_n^* - x^*\|_X} \leq \|[A'(x^*)]^{-1}\| \frac{\|\omega(x_n^* -, x^*)\|_X}{\|x_n^* - x^*\|_X}$$

Da die rechte Seite gegen 0 konvergiert, ist dies ein Widerspruch. □

Beispiel 2.23 *(Quadraturformelmethoden für Hammersteinsche Integralgleichungen) Wir betrachten $X = Y = C([0,1])$ mit der Norm $\|x\|_X = \max_{t\in[0,1]} |x(t)|$ sowie den Operator $A : X \to X$ definiert durch*

$$(Ax)(t) := x(t) - \int_0^1 f(t,s,x(s))ds \quad (t \in [0,1], x \in X),$$

wobei die Funktion $f : [0,1] \times [0,1] \times \mathbb{R} \to \mathbb{R}$ stetig ist und die partielle Ableitung $\frac{\partial f}{\partial x}$ existiert und ebenfalls stetig auf $[0,1] \times [0,1] \times \mathbb{R}$ ist.
Dann ist A auf X Frechét-differenzierbar und es gilt

$$(A'(x^*)x)(t) = x(t) - \int_0^1 \frac{\partial f}{\partial x}(t,s,x^*(s))x(s)ds \quad (t \in [0,1], x, x^* \in X).$$

Zur Einführung der Quadraturformelmethoden betrachten wir Gitter $G_n = \{t_0^{(n)} = 0 < t_1^{(n)} < \cdots < t_{k(n)}^{(n)} = 1\}$ in $[0,1]$ und Quadraturformeln

$$\int_0^1 x(s)ds \approx \sum_{j=0}^{k(n)} w_j^{(n)} x(t_j^{(n)})$$

für jedes $n \in \mathbb{N}$, die konvergent für jede Funktion $x \in X$ sind, d.h.

$$\varphi_n(x) = \left| \int_0^1 x(s)ds - \sum_{j=0}^{k(n)} w_j^{(n)} x(t_j^{(n)}) \right| \to 0 \quad (\textit{für } n \to \infty \ \textit{ und für alle } x \in X).$$

Dann betrachten wir die Räume $X_n = \mathbb{R}^{m(k(n)+1)}$ mit $\|x_n\|_{X_n} = \max_{i=0,1,\ldots,k(n)} |x_{ni}|$ für $x_n = (x_{n0}, x_{n1}, \ldots, x_{n,k(n)}) \in X_n$ und definieren die Operatoren $A_n : X_n \to X_n$ durch

$$[A_n x_n]_i = x_{ni} - \sum_{j=0}^{k(n)} w_j^{(n)} f(t_i^{(n)}, t_j^{(n)}, x_{nj}) \quad (i = 0, 1, \ldots, k(n))$$

für jedes $n \in \mathbb{N}$. Die Restriktionsoperatoren r_n seien wie in Beispiel 2.5b) definiert. Dann sind auch die Operatoren A_n Frechét-differenzierbar auf X_n und die Bedingung (v) von 2.20 läßt sich nachweisen. Bedingung (vi) ist automatisch erfüllt.
Ist ferner $x^* \in X$ eine Lösung von (1) und besitzt die homogene lineare (Fredholmsche) Integralgleichung

$$x(t) - \int_0^1 \frac{\partial f}{\partial x}(t,s,x^*(s))x(s)ds = 0 \quad (t \in [0,1])$$

nur die triviale Lösung $x = 0$, so ist (iii) von 2.20 erfüllt und es bleiben (vii) und (viii) nachzuweisen. Diese resultieren schließlich aus der Konvergenz der Quadraturformeln. Satz 2.20 liefert dann

$$\|x_n^* - r_n x^*\| = O(\varphi_n(v^*)) \quad (\textit{wobei } v^*(s) = \|f(\cdot, s, x^*(s))\|_X).$$

Literatur:

G. Vainikko: Funktionalanalysis der Diskretisierungsmethoden, Teubner, Leipzig 1976.

H. J. Reinhardt: Analysis of approximation methods for differential an integral equations, Springer, New York 1985.

3 Numerische Behandlung von Anfangswertproblemen für gewöhnliche Differentialgleichungen

Wir betrachten die Anfangswertaufgabe

$$(1) \qquad x'(t) = f(x(t), t),\ t \in [t_0, T],\ x(t_0) = x_0\,,$$

wobei $f : \mathbb{R}^m \times [t_0, T] \to \mathbb{R}^m$ stetig ist, $x_0 \in \mathbb{R}^m$, und vorausgesetzt wird, dass (1) eine eindeutig bestimmte (globale) Lösung x_* besitzt.

3.1 Integrationsverfahren für Anfangswertprobleme: Grundprinzipien und Beispiele

Grundidee: Zu einem Gitter $G = \{t_0 < t_1 < \cdots < t_N = T\}$ in $[t_0, T]$ ist ein Prinzip zur sukzessiven Bestimmung von Näherungen x_ℓ von $x_*(t_\ell)$, $\ell = 1, \ldots, N$ (ein sog. *Integrationsverfahren*) gesucht.

Ansätze für Integrationsverfahren:

(i) Ersetzung der Ableitung $x'(t_\ell)$ in den Gitterpunkten t_ℓ, $\ell = 1, \ldots, N$ durch Differenzenquotienten. Beispiel:

$$\frac{1}{h_\ell}(x(t_\ell) - x(t_{\ell-1})) \approx x'(t_{\ell-1}) = f(x(t_{\ell-1}), t_{\ell-1}),\quad h_\ell = t_\ell - t_{\ell-1},\ \ell = 1, \ldots, N.$$

Diese Idee führt zum *expliziten Euler-Verfahren* (bzw. Euler vorwärts)

$$x_\ell = x_{\ell-1} + h_\ell f(x_{\ell-1}, t_{\ell-1}),\ \ell = 1, \ldots, N.$$

(ii) Ersetzung der Integrale in der aufintegrierten Gleichung (1) durch Formeln zur numerischen Integration. Beispiel:

$$\int_{t_{\ell-1}}^{t_\ell} f(x(s), s)ds \approx \frac{h_\ell}{2}(f(x(t_\ell), t_\ell) + f(x(t_{\ell-1}), t_{\ell-1})) \quad \text{(Trapezregel)}.$$

Diese Idee führt zur sog. *Trapezregel*

$$x_\ell = x_{\ell-1} + \frac{h_\ell}{2}(f(x_\ell, t_\ell) + f(x_{\ell-1}, t_{\ell-1})),\ \ell = 1, \ldots, N.$$

Beispiel 3.1 *(Integrationsverfahren)*

(a) *Lineare Mehrschrittverfahren:*

$$\sum_{j=0}^{k} a_{\ell j} x_{\ell-j} = h_\ell \sum_{j=0}^{k} b_{\ell j} f(x_{\ell-j}, t_{\ell-j}),\ \ell = 1, \ldots, N,$$

wobei $k \in \mathbb{N}$, $a_{\ell 0} = 1$, $a_{\ell j} = 0$, $j = \ell + 1, \ldots, k$, heißt k-schrittiges lineares Mehrschrittverfahren.

Spezialfall: k-schrittige interpolative Mehrschrittverfahren
Ersetzung des Integranden in der aufintegrierten Gleichung (1) durch ein Interpolationspolynom in den Gitterpunkten $t_{\ell-k}, \ldots, t_\ell$, d.h.

$$\int_{t_{\ell-1}}^{t_\ell} f(x(s), s)ds \approx \sum_{j=0}^{k} \int_{t_{\ell-1}}^{t_\ell} \prod_{\substack{i=0 \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} dt f(x(t_{\ell-j}), t_{\ell-j}) \quad (\ell = k, \ldots, N).$$

Dies führt zu den interpolativen linearen Mehrschrittverfahren vom Adams-Typ

$$x_\ell = x_{\ell-1} + h_\ell \sum_{j=0}^{k} \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} \prod_{\substack{i=0 \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} dt f(x_{\ell-j}, t_{\ell-j}) \quad (\ell = k, \ldots, N).$$

(b) *Runge-Kutta-Verfahren (RKV):*
Idee: Ersetzung des Integrals in der aufintegrierten Gleichung (1) durch interne (neue) Stützstellen in $[t_{\ell-1}, t_\ell]$ und näherungsweise Berechnung der Werte von $x(\cdot)$ an diesen Stützstellen aus einem nichtlinearen Gleichungssystem.
*Allgemeine Gestalt eines p-*stufigen Runge-Kutta-Verfahrens*:*

$$\begin{aligned} x_\ell &= x_{\ell-1} + h_\ell \sum_{j=1}^{p} \gamma_j f(\bar{x}_\ell^j, t_{\ell-1} + \alpha_j h_\ell)\,,\ \ell = 1, \ldots, N, \\ \bar{x}_\ell^i &= x_{\ell-1} + h_\ell \sum_{j=1}^{p} \beta_{ij} f(\bar{x}_\ell^j, t_{\ell-1} + \alpha_j h_\ell)\,,\ i = 1, \ldots, p, \end{aligned}$$

wobei $\gamma = (\gamma_1, \ldots, \gamma_p)$ und $B = (\beta_{ij})_{i,j=1,\ldots,p}$ die Runge-Kutta-Parameter sind und oft $\alpha_i = \sum_{j=1}^{p} \beta_{ij}$, $i = 1, \ldots, p$, gesetzt wird.
Spezialfall: RKV ist explizit*, falls $\beta_{ij} = 0$, $\forall i \geq j$, sonst* implizit*.*
Im Unterschied zu (a) geht f nichtlinear in die Verfahrensvorschrift ein.

Definition 3.2 *(allgemeine Verfahrensklasse)*

$$\textit{(IV)} \qquad \sum_{j=0}^{k} a_{\ell j} x_{\ell-j} = h_\ell \varphi_\ell(x_\ell, \ldots, x_{\ell-k})\,,\ \ell = 1, \ldots, N \quad (\textit{realisiert auf } G),$$

wobei $a_{\ell 0} = 1$, $a_{\ell j} = 0$, $j = \ell + 1, \ldots, k$, und $\varphi_\ell : \mathbb{R}^{m(k+1)} \to \mathbb{R}^m$, $\ell = 1, \ldots, N$, heißen k-schrittige Integrationsverfahren*.*
Für $k = 1$ heißt (IV) Einschrittverfahren*, sonst* Mehrschrittverfahren*.*
(IV) heißt explizit*, falls φ_ℓ für jedes ℓ nicht von x_ℓ abhängt, sonst* implizit*.*
(IV) heißt linear*, falls $\varphi_\ell(x_\ell, \ldots, x_{\ell-k}) = \sum_{j=0}^{k} b_{\ell j} f(x_{\ell-j}, t_{\ell-j})$ mit gewissen reellen Parametern $b_{\ell j}$ gilt, sonst* nichtlinear*.*

Die Verfahrensklasse (IV) umfasst die Beispiele 3.1(a) und 3.1(b).

3.2 Konsistenz, Stabilität und Konvergenz von Integrationsverfahren

Unser Ziel ist die Anwendung der Konzepte und Ergebnisse aus Kapitel 2. Dazu betrachten wir die linearen normierten Räume $X = C^1([t_0, T], \mathbb{R}^m)$ mit der Norm $\|x\|_X = \max_{t\in[t_0,T]} \|x(t)\|$ (und einer Norm $\|\cdot\|$ auf $\mathbb{R}^m$) und $Y = \mathbb{R}^m \times C([t_0, T], \mathbb{R}^m)$ mit $\|y\|_Y = \|a\| + \max_{t\in[t_0,T]} \|x(t)\|$ für $y = (a, x) \in Y$. Der Operator $A : X \to Y$ sei definiert durch

$$(Ax)(\cdot) = (x(t_0) - x_0, x'(\cdot) - f(x(\cdot), \cdot)).$$

Damit hat das Anfangswertproblem (1) als Operatorgleichung die Form

$$Ax = 0\,.$$

Zu jedem Gitter $G = \{t_0 < t_1 < \cdots < t_N = T\}$ in $[t_0, T]$ mit den Schrittweiten $h_\ell = t_\ell - t_{\ell-1}$, $\ell = 1, \ldots, N$, $N = N(G)$, und der maximalen Schrittweite $h(G) = \max_{\ell=1,\ldots,N} h_\ell$ betrachten wir die endlichdimensionalen linearen normierten Räume $X_G = \mathbb{R}^{(N+1)m}$ mit der Norm

$$\|y_G\|_G = \max_{\ell=0,\ldots,N} \|y_\ell\| \quad \text{für jedes } y_G = (y_0, y_1, \ldots, y_N) \in X_G,$$

sowie $Y_G = \mathbb{R}^{(N+1)m}$ mit der Norm

$$\|y_G\|_G = \|y_0\| + \max_{\ell=1,\ldots,N} \|y_\ell\| \quad \text{für jedes } y_G = (y_0, y_1, \ldots, y_N) \in Y_G,$$

und die Restriktionsoperatoren $r_G : X \to X_G$ definiert durch

$$r_G x = (x(t_0), x(t_1), \ldots, x(t_N)) \quad \text{für alle } x \in X,$$

bzw. $\tilde{r}_G : Y \to Y_G$ gegeben durch

$$\tilde{r}_G y = (a, x(t_1), \ldots, x(t_N)) \quad \text{für alle } y = (a, x) \in Y.$$

Analog zu Beispiel 2.5 gilt, dass $(X, X_{G_n}, r_{G_n})_{n\in\mathbb{N}}$ bzw. $(Y, Y_{G_n}, \tilde{r}_{G_n})_{n\in\mathbb{N}}$ diskrete Approximationen von X bzw. Y sind, falls für die Folge (G_n) von Gittern in $[t_0, T]$ die Folge $(h(G_n))$ gegen Null konvergiert.
Schließlich definieren wir 'diskretisierte' Operatoren $A_G : X_G \to Y_G$ durch

$$\begin{aligned} [A_G y_G]_0 &= y_0 - x_0\,, \\ [A_G y_G]_\ell &= \frac{1}{h_\ell} \sum_{j=0}^{k} a_{\ell,j} y_{\ell-j} - \varphi_\ell(y_\ell, \ldots, y_{\ell-k}) \quad (\ell = 1, \ldots, N) \end{aligned}$$

wobei $y_G = (y_0, y_1, \ldots, y_N)$ und $[\cdot]_\ell$ die ℓ-te Komponente bedeutet.

Definition 3.3 *Es sei $\mathcal{G}$ eine Klasse von Gittern in I und (1) besitze eine einzige Lösung x^*.*

(a) *Das Integrationsverfahren (IV) heißt* konsistent *mit (1) (auf $\mathcal{G}$), falls für jede Folge (G_n) von Gittern in $\mathcal{G}$ mit $h(G_n) \to 0$, A und (A_{G_n}) konsistent in x^* sind.*

(b) *Das Integrationsverfahren (IV) heißt* konvergent *(auf $\mathcal{G}$), falls für jede Lösung x_G^* von (IV) gilt*

$$\lim_{h(G)\to 0} \|x_G^* - r_G x^*\|_G = 0\,.$$

(c) *Das Integrationsverfahren ist* stabil auf $\mathcal{G}$*, falls eine Konstante $S = S(\mathcal{G})$ existiert, so dass für jedes $G \in \mathcal{G}$ und alle x_G und $\tilde{x}_G$ in X_G gilt*

$$\|x_G - \tilde{x}_G\|_G \le S\|A_G x_G - A_G \tilde{x}_G\|_G.$$

Bemerkung 3.4
Konsistenz mit (1) auf $\mathcal{G}$ bedeutet

$$\begin{aligned}\|A_G r_G x_*\|_G &= \max_{\ell=1,\dots,N} \|[A_G r_G x_*]_\ell\| \\ &= \max_{\ell=1,\dots,N} \Big\| \frac{1}{h_\ell} \sum_{j=0}^{k} a_{\ell,j} x_*(t_{\ell-j}) - \varphi_\ell(x_*(t_\ell), \dots, x_*(t_{\ell-k})) \Big\| \to 0\,,\end{aligned}$$

falls $h(G) \to 0$ mit $G \in \mathcal{G}$.

Satz 3.5 *(Konvergenzsatz)*
Es sei $\mathcal{G}$ eine Klasse von Gittern in $[t_0, T]$ und das Integrationsverfahren (IV) sei stabil auf $\mathcal{G}$ und konsistent mit (1) auf $\mathcal{G}$.
Dann existiert eine Konstante $S = S(\mathcal{G})$, so dass für jedes Gitter $G \in \mathcal{G}$, jede Lösung $x_G^* = (x_0^*, \dots, x_{N(G)}^*)$ von (IV) und die einzige Lösung x_* von (1) gilt

$$\begin{aligned}\|x_G^* - r_G x_*\|_G = \max_{\ell=0,\dots,N(G)} \|x_\ell^* - x_*(t_\ell)\| &\le S \max_{\ell=1,\dots,N(G)} \|[A_G r_G x_*]_\ell\| \\ \lim_{h(G)\to 0} \max_{\ell=0,\dots,N(G)} \|x_\ell^* - x_*(t_\ell)\| &= 0\,.\end{aligned}$$

Beweis: Wie im Beweis von Satz 2.5 setzen wir $x_G = x_G^*$ und $\tilde{x}_G = r_G x_*$ in der Stabilitätsungleichung in Definition 3.3(c) und erhalten:

$$\|x_G^* - r_G x_*\| = \max_{\ell=0,\dots,N(G)} \|x_\ell^* - x_*(t_\ell)\| \le S\|A_G x_G^* - A_G r_G x_*\|_G = \|A_G r_G x_*\|_G$$

Wegen der vorausgesetzten Konsistenz gilt $\lim_{h(G)\to 0} \|A_G r_G x_*\|_G = 0$. □

Definition 3.6 *Es sei x_* die einzige Lösung von (1). Dann sagt man, das Integrationsverfahren (IV) besitzt*

(a) Konsistenzordnung $s \in \mathbb{N}$ *(auf $\mathcal{G}$), falls $\|A_G r_G x_*\|_G = O(h(G)^s)$ für alle $G \in \mathcal{G}$ gilt,*

(b) Konvergenzordnung $s \in \mathbb{N}$ *(auf $\mathcal{G}$), falls $\|x_G^* - r_G x_*\|_G = O(h(G)^s)$ für alle $G \in \mathcal{G}$ gilt.*

Der Term $\tau_\ell(G) = \|[A_G r_G x_*]_\ell\|$ *heißt* lokaler Diskretisierungsfehler *von (IV) im Schritt* ℓ *(auf* $G \in \mathcal{G}$*).*

Folgerung 3.7 *Das Integrationsverfahren (IV) sei stabil auf einer Klasse* $\mathcal{G}$ *von Gittern in* $[t_0, T]$*. Ist (IV) konsistent mit (1) auf* $\mathcal{G}$ *(mit Konsistenzordnung* $s \in \mathbb{N}$*, so ist (IV) auch konvergent auf* $\mathcal{G}$ *(mit Konvergenzordnung* $s \in \mathbb{N}$*).*

Beweis: Die Aussage folgt sofort aus Satz 3.5. □

3.3 Einschrittverfahren

Wir betrachten die folgenden *Einschrittverfahren* zur Lösung von (1)

$$\text{(ESV)} \qquad x_\ell = x_{\ell-1} + h_\ell \Phi(x_{\ell-1}, t_{\ell-1}; h_\ell)\,,\ \ell = 1, \ldots, N,$$

wobei $\Phi : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ mit einem gewissen $H > 0$ und $h_\ell = t_\ell - t_{\ell-1}$, $\ell = 1, \ldots, N = N(G)$, $G = \{t_0 < t_1 < \cdots < t_N\} \in \mathcal{G}$ derart, dass $h(G) \le H$.
(Ist $H \ge T - t_0$, so ist die Bedingung $h(G) \le H$ stets erfüllt.)

Wir untersuchen jetzt Bedingungen an Φ, so dass (ESV) stabil und konsistent (mit Konsistenzordnung $s \in \mathbb{N}$) auf $\mathcal{G}$ ist.

Satz 3.8 *(Konsistenz)*
Es sei $\Phi : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ *stetig.*

(a) Das Einschrittverfahren (ESV) ist konsistent mit (1) auf $\mathcal{G}$*, falls* $\Phi(x, t; 0) = f(x, t)$*,* $\forall (x, t) \in \mathbb{R}^m \times [t_0, T]$*.*

(b) Es sei $f \in C^s(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$ *und die partiellen Ableitungen* $\frac{\partial^j \Phi}{\partial h^j} : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$*,* $j = 1, \ldots, s \in \mathbb{N}$*, mögen existieren und stetig sein.*
Dann besitzt (ESV) die Konsistenzordnung s*, falls die folgenden Bedingungen erfüllt sind:*

$$\frac{\partial^j \Phi}{\partial h^j}(x, t; 0) = \frac{1}{j+1} f^{(j)}(x, t)\,, \quad \forall j = 0, \ldots, s-1, \forall (x, t) \in \mathbb{R}^m \times [t_0, T].$$

Hierbei ist $f^{(0)}(x, t) = f(x, t)$ *und* $f^{(j)}(x, t) = \frac{\partial f^{(j-1)}}{\partial x}(x, t) f^{(j-1)}(x, t) + \frac{\partial f^{(j-1)}}{\partial t}(x, t)$*,* $\forall j = 1, \ldots, s, \forall (x, t) \in \mathbb{R}^m \times [t_0, T]$*.*

Beweis:
Vorbemerkung: Falls die entsprechenden Ableitungen existieren und stetig sind, folgt aus (1) mit der Kettenregel für alle $t \in [t_0, T]$

$$\begin{aligned} x_*''(t) &= \frac{\partial f}{\partial x}(x_*(t), t) x_*'(t) + \frac{\partial f}{\partial t}(x_*(t), t) \\ &= \frac{\partial f}{\partial x}(x_*(t), t) f(x_*(t), t) + \frac{\partial f}{\partial t}(x_*(t), t) = f^{(1)}(x_*(t), t) \end{aligned}$$

bzw. allgemein

$$x_*^{(j+1)}(t) = f^{(j)}(x_*(t), t) \quad (j = 1, \ldots, s).$$

(a) Gemäß Definition 3.3(a) betrachten wir für ein bel. Gitter $G = \{t_0 < \cdots < t_N\}$ und bel. $\ell \in \{1, \ldots, N\}$

$$\begin{aligned}
\|[A_G r_G x_*]_\ell\| &= \Big\| \frac{1}{h_\ell}(x_*(t_\ell) - x_*(t_{\ell-1})) - \Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell) \Big\| \\
&= \Big\| \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} x_*'(t)dt - x_*'(t_{\ell-1}) + f(x_*(t_{\ell-1}), t_{\ell-1}) - \Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell) \Big\| \\
&\le \Big\| \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} (x_*'(t) - x_*'(t_{\ell-1}))dt \Big\| + \|f(x_*(t_{\ell-1}), t_{\ell-1}) - \Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell)\| \\
&\le \max_{t \in [t_{\ell-1}, t_\ell]} \|x_*'(t) - x_*'(t_{\ell-1})\| + \|\Phi(x_*(t_{\ell-1}), t_{\ell-1}; 0) - \Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell)\| \\
&\le \max_{|t-s| \le h} \|x_*'(t) - x_*'(s)\| + \max_{\substack{t \in [t_0, T] \\ \eta \in [0,h]}} \|\Phi(x_*(t), t; 0) - \Phi(x_*(t), t; \eta)\|
\end{aligned}$$

Die Aussage von (a) resultiert nun aus der gleichmäßigen Stetigkeit von x_*' auf $[t_0, T]$ sowie von Φ auf $\{(x_*(t), t, \eta) : t \in [t_0, T], \eta \in [0, h]\}$ für $h = h(G)$ gegen 0.

(b) Nach Voraussetzung ist die Lösung x_* $(s+1)$-mal stetig differenzierbar. Nach dem Satz von Taylor existiert deshalb für jedes $\ell \in \{1, \ldots, N\}$ und jede Komponente $i \in \{1, \ldots, m\}$ (von x_*) ein $\xi_{\ell i} \in (t_{\ell-1}, t_\ell)$, so dass

$$x_*(t_\ell) = \sum_{j=0}^{s} \frac{x_*^{(j)}(t_{\ell-1})}{j!} h_\ell^j + \frac{x_*^{(s+1)}(\xi_\ell)}{(s+1)!} h_\ell^{s+1} \,,$$

wobei wir den letzten Term hier und im folgenden so verstehen, dass in jeder Komponente von $x_*^{(s+1)}$ das Argument $\xi_{\ell i}$ auftritt. Dies wird sich später wegen der Verwendung der max-Norm im $\mathbb{R}^m$ als unproblematisch erweisen. Dann folgt:

$$\begin{aligned}
\frac{1}{h_\ell}(x_*(t_\ell) - x_*(t_{ell-1})) &= \sum_{j=1}^{s} \frac{f^{(j-1)}(x_*(t_{\ell-1}), t_{\ell-1})}{j!} h_\ell^{j-1} + \frac{x_*^{(s+1)}(\xi_\ell)}{(s+1)!} h_\ell^s \\
&= \sum_{j=0}^{s-1} \frac{1}{j!} \frac{\partial^j \Phi}{\partial h^j}(x_*(t_{\ell-1}), t_{\ell-1}; 0) h_\ell^j + \frac{x_*^{(s+1)}(\xi_\ell)}{(s+1)!} h_\ell^s
\end{aligned}$$

Nach Voraussetzung existieren nach dem Satz von Taylor auch $\eta_{\ell i} \in (0, h_\ell)$ für alle $\ell = 1, \ldots, N$ und alle Komponenten $i \in \{1, \ldots, m\}$ von Φ, so dass

$$\Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell) = \sum_{j=0}^{s-1} \frac{1}{j!} \frac{\partial^j \Phi}{\partial h^j}(x_*(t_{\ell-1}), t_{\ell-1}; 0) h_\ell^j + \frac{1}{s!} \frac{\partial^s \Phi}{\partial h^s}(x_*(t_{\ell-1}), t_{\ell-1}; \eta_\ell) h_\ell^s \,,$$

wobei wieder die obige Vereinbarung für den letzten Term verwendet wird. Insgesamt folgt daraus die Abschätzung

$$\begin{aligned}
\max_{\ell=1,\ldots,N} \tau_\ell(G) &= \max_{\ell=1,\ldots,N} \|\frac{1}{h_\ell}(x_*(t_\ell) - x_*(t_{\ell-1})) - \Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell)\| \\
&\le \max_{\ell=1,\ldots,N} \Big\| \frac{x_*^{(s+1)}(\xi_\ell)}{(s+1)!} - \frac{1}{s!} \frac{\partial^s \Phi}{\partial h^s}(x_*(t_{\ell-1}), t_{\ell-1}; \eta_\ell) \Big\| h_\ell^s \\
&= \Big(\max_{t \in [t_0, T]} \Big\| \frac{x_*^{(s+1)}(t)}{(s+1)!} \Big\| + \max_{\substack{t \in [t_0, T] \\ \eta \in [0,h]}} \Big\| \frac{1}{s!} \frac{\partial^s \Phi}{\partial h^s}(x_*(t), t; \eta) \Big\| \Big) h^s = O(h^s)
\end{aligned}$$

mit $h = h(G)$. Beim Übergang von der vorletzten zur letzten Zeile wurde dabei komponentenweise argumentiert und die Definition der max-Norm im $\mathbb{R}^m$ verwendet. $\square$

Beispiel 3.9 *(Taylor-Verfahren)*
Es sei $f \in C^s(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$ für ein $s \in \mathbb{N}$. Dann heißt das Verfahren

$$x_\ell = x_{\ell-1} + h_\ell \sum_{j=0}^{s-1} \frac{h_\ell^j}{(j+1)!} f^{(j)}(x_{\ell-1}, t_{\ell-1}) \,,\ \ell = 1, \ldots, N,$$

Taylor-Verfahren der Ordnung s. *Es ist ein Einschrittverfahren und besitzt die Konsistenzordnung s auf jeder Klasse von Gittern.*
Bew.: Wir wenden Satz 3.8 an und überprüfen die dortige Voraussetzung and die Funktion $\Phi : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ $(H > 0)$

$$\Phi(x, t; h) = \sum_{j=0}^{s-1} \frac{h^j}{(j+1)!} f^{(j)}(x, t) \ \forall (x, t, h) \in \mathbb{R}^m \times [t_0, T] \times [0, H].$$

Φ ist stetig und die Differenzierbarkeitseigenschaften von Φ bsgl. h sind erfüllt. Es gilt

$$\frac{\partial^i \Phi}{dh^i}(x, t; h) = \sum_{j=i}^{s-1} \frac{1}{j+1} \frac{h^{j-i}}{(j-i)!} f^{(j)}(x, t) \,,\ i = 0, \ldots, i.$$

Wir folgern $\frac{\partial^i \Phi}{dh^i}(x, t; 0) = \frac{1}{i+1} f^{(i)}(x, t)$, $\forall i = 0, \ldots, s-1$, $\forall (x, t) \in \mathbb{R}^m \times [t_0, T]$. $\square$

Durch die aktuellen Möglichkeiten des algorithmischen Differenzierens *haben Taylor-Verfahren deutlich an Potential gewonnen.*
A. Griewank: Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation*, SIAM, Philadelphia 2000. (Second Edition with co-author A. Walther, 2008).*

Satz 3.10 *(Stabilität)*
Es sei $\Phi : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ stetig und es existiere ein $L > 0$, so dass

$$\|\Phi(x, t; h) - \Phi(\tilde{x}, t; h)\| \le L \|x - \tilde{x}\| \,,\ \forall x, \tilde{x} \in \mathbb{R}^m, \forall t \in [t_0, T], \forall h \in [0, H].$$

Dann ist das Einschrittverfahren (ESV) stabil auf jeder Klasse $\mathcal{G}$ von Gittern in $[t_0, T]$ mit $h(G) \le H$, $\forall G \in \mathcal{G}$.

Beweis: Es seien $x_0, \tilde{x}_0 \in \mathbb{R}^m$ und ein Gitter $G = \{t_0 < \cdots < t_N\}$ in $\mathcal{G}$ beliebig gewählt. Wir betrachten den Operator $A_G : X_G \to X_G$

$$[A_G x_G]_\ell = \frac{1}{h_\ell}(x_\ell - x_{\ell-1}) - \Phi(x_{\ell-1}, t_{\ell-1}; h_\ell)$$

für jedes $\ell = 1, \ldots, N$ und $x_G = (x_0, x_1, \ldots, x_N) \in X_G$. Wir setzen

$$\varepsilon_\ell := [A_G x_G]_\ell - [A_G \tilde{x}_G]_\ell \quad (\ell = 1, \ldots, N).$$

Dann ergibt sich für alle $\ell = 1, \ldots, N$:

$$\begin{aligned}
\|x_\ell - \tilde{x}_\ell\| &= \|x_{\ell-1} - \tilde{x}_{\ell-1} + h_\ell(\Phi(x_{\ell-1}, t_{\ell-1}; h_\ell) - \Phi(\tilde{x}_{\ell-1}, t_{\ell-1}; h_\ell)) + h_\ell \varepsilon_\ell\| \\
&\leq \|x_{\ell-1} - \tilde{x}_{\ell-1}\| + h_\ell L \|x_{\ell-1} - \tilde{x}_{\ell-1}\| + h_\ell \|\varepsilon_\ell\| \\
&\leq \exp(h_\ell L)\|x_{\ell-1} - \tilde{x}_{\ell-1}\| + h_\ell \|\varepsilon_\ell\| \\
&\leq \exp(h_\ell L)[\exp(h_{\ell-1} L)\|x_{\ell-2} - \tilde{x}_{\ell-2}\| + h_{\ell-1}\|\varepsilon_{\ell-1}\|]h_\ell \|\varepsilon_\ell\| \\
&\leq \exp((t_\ell - t_{\ell-2})L)\|x_{\ell-2} - \tilde{x}_{\ell-2}\| + \sum_{j=0}^{1} \exp((t_\ell - t_{\ell-j})L) h_{\ell-j}\|\varepsilon_{\ell-j}\| \\
&\leq \exp((t_\ell - t_0)L)\|x_0 - \tilde{x}_0\| + \sum_{j=0}^{\ell-1} \exp((t_\ell - t_{\ell-j})L) h_{\ell-j}\|\varepsilon_{\ell-j}\| \\
&\leq \exp((T - t_0)L)(\|x_0 - \tilde{x}_0\| + \sum_{j=0}^{\ell-1} h_{\ell-j}\|\varepsilon_{\ell-j}\|) \\
&\leq \exp((T - t_0)L)(\|x_0 - \tilde{x}_0\| + (T - t_0) \max_{j=1,\ldots,\ell-1} \|\varepsilon_{\ell-j}\|)
\end{aligned}$$

Insgesamt erhält man

$$\max_{\ell=0,\ldots,N} \|x_\ell - \tilde{x}_\ell\| \leq \exp((T - t_0)L) \max\{1, T - t_0\} \max_{\ell=0,\ldots,N} \|[A_G x_G]_\ell - [A_G \tilde{x}_G]_\ell\|$$

und damit die Stabilitätskonstante $S = \exp((T - t_0)L) \max\{1, T - t_0\}$. □

Satz 3.11 *(Konvergenz)*
Es seien die Voraussetzungen für Φ von Satz 3.10 erfüllt. Dann ist das Einschrittverfahren (ESV) konvergent (mit Ordnung s), falls es konsistent (mit Ordnung s) auf $\mathcal{G}$ mit mit $h(G) \leq H$, $\forall G \in \mathcal{G}$ ist.

Beweis: Nach Satz 3.10 gilt mit Satz 3.5 für Einschrittverfahren

$$\max_{\ell=1,\ldots,N} \|x_\ell - x_*(t_\ell)\| \leq S \max_{\ell=1,\ldots,N} \tau_\ell(G),$$

wobei $\tau_\ell(G) = \|\frac{1}{h_\ell}(x_*(t_\ell) - x_*(t_{\ell-1})) - \Phi(x_*(t_{\ell-1}), t_{\ell-1}; h_\ell)\|$, $\ell = 1, \ldots, N$. Ist das Einschrittverfahren konsistent (mit Ordnung s), so konvergiert $\max_{\ell=1,\ldots,N} \tau_\ell(G)$ gegen 0, falls $h(G) \to 0$ ($\max_{\ell=1,\ldots,N} \tau_\ell(G) = O(h(G)^s)$). □

Bemerkung 3.12 *Gemäß Satz 3.10 ist die Stabilitätskonstante S für (ESV) insbesondere dann groß, wenn die Lipschitzkonstante L von Φ groß bzw. das Intervall $[t_0, T]$ lang ist ! Ein ähnliches Resultat werden wir in Kapitel 3.5 auch für lineare Mehrschrittverfahren erhalten.*
Ist (ESV) stabil und sind die Stetigkeits- und Differenzierbarkeitsforderungen an f und Φ erfüllt, so entscheiden die nichtlinearen Gleichungen

$$\frac{\partial^j \Phi}{\partial h^j}(x, t; 0) = \frac{1}{j+1} f^{(j)}(x, t)\,, \quad \forall j = 0, \ldots, s-1, \forall (x, t) \in \mathbb{R}^m \times [t_0, T]$$

über die Konvergenzordnung s.

Beispiel 3.9 (*Fortsetzung*)
Die Taylor-Verfahren der Ordnung s sind nach Satz 3.10 stabil, falls ein $L > 0$ existiert, so dass

$$\|f^{(j)}(x,t) - f^{(j)}(\tilde{x},t)\| \leq L\|x - \tilde{x}\| \quad (\forall x, \tilde{x} \in \mathbb{R}^m, \forall j = 0, \ldots, s-1).$$

Diese Lipschitzbedingung folgt i.a. nicht aus der Differenzierbarkeitsforderung $f \in C^s(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$.

3.4 Runge-Kutta Verfahren

Wir betrachten die folgende Funktion $\Phi : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ mit einem noch festzulegenden $H > 0$

$$\begin{aligned} \Phi(x,t;h) &:= \sum_{j=1}^{p} \gamma_j K_j(x,t;h)\,, \quad \text{wobei} \\ K_i(x,t;h) &= f(x + h\sum_{j=1}^{P} \beta_{ij} K_j(x,t;h), t + \alpha_i h)\,,\ i = 1, \ldots, p. \end{aligned}$$

und $p \in \mathbb{N}$ die Anzahl der Stufen, $B = (\beta_{ij}) \in \mathbb{R}^{p\times p}$ die Runge-Kutta Matrix, γ, $\alpha \in \mathbb{R}^p$ die restlichen Runge-Kutta Parameter sind. Dann stellt das Einschrittverfahren

$$x_\ell = x_{\ell-1} + h_\ell \Phi(x_{\ell-1}, t_{\ell-1}; h_\ell),\ \ell = 1, \ldots, N,$$

das in Beispiel 3.1 angegebene Runge-Kutta Verfahren in umformulierter Form dar. Wir beantworten zunächst die Frage, wann Runge-Kutta Verfahren stabil sind.

Satz 3.13
Es sei $f : \mathbb{R}^m \times [t_0, T] \to \mathbb{R}^m$ stetig und es existiere eine Konstante $L > 0$, so dass

$$\|f(x,t) - f(\tilde{x},t)\| \leq L\|x - \tilde{x}\|\,,\ \forall x, \tilde{x} \in \mathbb{R}^m, \forall t \in [t_0, T].$$

Für beliebig vorgegebene $p \in \mathbb{N}$, $B \in \mathbb{R}^{p\times p}$, $\gamma, \alpha \in \mathbb{R}^p$, existiert ein $H > 0$, so dass die oben definierte Verfahrensfunktion $\Phi : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ eines p-stufigen Runge-Kutta Verfahrens wohldefiniert und stetig ist. Überdies erfüllt Φ die Lipschitzbedingung von Satz 3.10 und das Runge-Kutta Verfahren ist stabil.

Beweis: Die Runge-Kutta Parameter p, B, γ, α seien gegeben und wir wählen $H > 0$ so, dass

$$LH \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\beta_{ij}| =: \kappa < 1\,.$$

Es sei nun $(x,t;h) \in \mathbb{R}^m \times [t_0, T] \times [0, H]$ beliebig gewählt. Wir setzen $w = (x,t;h)$ und definieren die Abbildung $\mathcal{K}_w : \mathbb{R}^{mp} \to \mathbb{R}^{mp}$ durch

$$[\mathcal{K}_w(K)]_i := f\Big(x + h\sum_{j=1}^{p} \beta_{ij} K_j, t + h\alpha_i\Big)\,,\ i = 1, \ldots, p.$$

Für beliebige $K, \tilde{K} \in \mathbb{R}^{mp}$ gilt bzgl. der Norm $\|\cdot\|_\infty$ auf $\mathbb{R}^{mp}$

$$\begin{aligned}\|\mathcal{K}_w(K) - \mathcal{K}_w(\tilde{K})\|_\infty &= \max_{i=1,\dots,p} \|[\mathcal{K}_w(K)]_1, \dots, [\mathcal{K}_w(K)]_p\| \\ &\le L\,h \max_{i=1,\dots,p} \sum_{j=1}^{p} |\beta_{ij}| \|K_j - \tilde{K}_j\| \\ &\le \kappa \|K - \tilde{K}\|_\infty.\end{aligned}$$

Also existiert für jedes $w = (x, t; h) \in \mathbb{R}^m \times [t_0, T] \times [0, H]$ genau ein $K(w) \in \mathbb{R}^{mp}$, das Fixpunkt von $\mathcal{K}_w$ auf $\mathbb{R}^{mp}$ ist. Damit ist die Verfahrensfunktion $\Phi(w) = \Phi(x, t; h) = \sum_{i=1}^{p} \gamma_i K_i(x, t; h)$ wohldefiniert.
Ist überdies $\tilde{w} = (\tilde{x}, \tilde{t}, \tilde{h}) \in \mathbb{R}^m \times [t_0, T] \times [0, H]$ und $K(\tilde{w})$ der zugehörige Fixpunkt von $\mathcal{K}_{\tilde{w}}$, so gilt

$$\begin{aligned}\|K(w) - K(\tilde{w})\|_\infty &= \|\mathcal{K}_w(K(w)) - \mathcal{K}_{\tilde{w}}(K(\tilde{w}))\|_\infty \\ &\le \|\mathcal{K}_w(K(w)) - \mathcal{K}_{\tilde{w}}(K(w))\|_\infty + \|\mathcal{K}_{\tilde{w}}(K(w)) - \mathcal{K}_{\tilde{w}}(K(\tilde{w}))\|_\infty \\ &\le \|\mathcal{K}_w(K(w)) - \mathcal{K}_{\tilde{w}}(K(w))\|_\infty + \kappa \|K(w) - K(\tilde{w})\|_\infty\end{aligned}$$

und folglich

$$\|K(w) - K(\tilde{w})\|_\infty \le \frac{1}{1-\kappa} \|\mathcal{K}_w(K(w)) - \mathcal{K}_{\tilde{w}}(K(w))\|_\infty.$$

Insgesamt erhält man daraus

$$\begin{aligned}\|\Phi(x, t; h) - \Phi(\tilde{x}, \tilde{t}; \tilde{h})\| &\le \sum_{j=1}^{p} |\gamma_j| \|K_j(w) - K_j(\tilde{w})\| \le \sum_{j=1}^{p} |\gamma_j| \|K(w) - K(\tilde{w}\|_\infty \\ &\le \sum_{j=1}^{p} |\gamma_j| \frac{1}{1-\kappa} \|\mathcal{K}_w(K(w)) - \mathcal{K}_{\tilde{w}}(K(w))\|_\infty \\ &= \sum_{j=1}^{p} |\gamma_j| \frac{1}{1-\kappa} \max_{i=1,\dots,p} \Big\| f\Big(x + h \sum_{j=1}^{p} \beta_{ij} K_j(w), t + h\alpha_i\Big) \\ &\qquad - f\Big(\tilde{x} + \tilde{h} \sum_{j=1}^{p} \beta_{ij} K_j(w), \tilde{t} + \tilde{h}\alpha_i\Big) \Big\|.\end{aligned}$$

Da f stetig ist, ist deshalb auch Φ stetig. Überdies ergibt sich

$$\|\Phi(x, t; h) - \Phi(\tilde{x}, t; h)\| \le \frac{L}{1-\kappa} \sum_{j=1}^{p} |\gamma_j| \|x - \tilde{x}\| \;\; (\forall x, \tilde{x} \in \mathbb{R}^m, (t, h) \in [t_0, T] \times [0, H])$$

und deshalb aus Satz 3.10 die Stabilität des Runge-Kutta Verfahrens.
Der Beweisweg über den Banachschen Fixpunktsatz ist überflüssig, falls das Runge-Kutta Verfahren explizit ist. Dann folgt die Stetigkeit von K_i, $i = 1, \dots, p$, rekursiv aus der Identität $K_i(x, t; h) = f(x + h \sum_{j=1}^{i-1} \beta_{ij} K_j(x, t; h), t + h\alpha_i)$ für $i = 1, \dots, p$. $\square$

Satz 3.14
Es sei $f : \mathbb{R}^m \times [t_0, T] \to \mathbb{R}^m$ *stetig und es existiere eine Konstante* $L > 0$*, so dass*

$$\|f(x,t) - f(\tilde{x},t)\| \leq L\|x - \tilde{x}\|\,, \forall x, \tilde{x} \in \mathbb{R}^m, \forall t \in [t_0, T].$$

Es seien $p \in \mathbb{N}$*,* $B \in \mathbb{R}^{p\times p}$ *und* $\gamma, \alpha \in \mathbb{R}^p$ *die Parameter eines Runge-Kutta Verfahrens. Das Runge-Kutta Verfahren ist konvergent , falls* $\gamma^\top e = 1$ *mit* $e = (1, 1, \ldots, 1)^\top \in \mathbb{R}^p$*. Ist zusätzlich* $f \in C^s(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$ *für ein* $s \in \{1, 2, 3, 4\}$*, so ist das Runge-Kutta Verfahren konvergent mit Ordnung* s*, falls die folgenden Gleichungen erfüllt sind:*
$s = 1$*:* $\gamma^\top e = 1$ *.*
$s = 2$*:* $\gamma^\top e = 1$*,* $\gamma^\top Be = \frac{1}{2}$*,* $Be = Ae$*, wobei* $A := diag\,(\alpha_1, \ldots, \alpha_p)$*.*
$s = 3$*:* $\gamma^\top e = 1$*,* $\gamma^\top Be = \frac{1}{2}$*,* $Be = Ae$*,* $\gamma^\top A^2 e = \frac{1}{3}$*,* $\gamma^\top BAe = \frac{1}{6}$*.*
$s = 4$*:* $\gamma^\top e = 1$*,* $\gamma^\top Be = \frac{1}{2}$*,* $Be = Ae$*,* $\gamma^\top A^2 e = \frac{1}{3}$*,* $\gamma^\top BAe = \frac{1}{6}$*,* $\gamma^\top A^3 e = \frac{1}{4}$*,*
$\gamma^\top BA^2 e = \frac{1}{12}$*,* $\gamma^\top B^2 Ae = \frac{1}{24}$*,* $\gamma^\top ABAe = \frac{1}{8}$*.*

Beweis: Aus dem Beweis von Satz 3.13 entnehmen wir die Gleichung

$$K_j(x,t;0) = f(x,t) \quad (\forall j = 1, \ldots, p, \forall (x,t) \in \mathbb{R}^m \times [t_0, T]).$$

Deshalb folgt

$$\Phi(x,t;0) = \sum_{j=1}^{p} \gamma_j K_j(x,t;0) = \sum_{j=1}^{p} \gamma_j f(x,t) \quad (\forall (x,t) \in \mathbb{R}^m \times [t_0, T]).$$

Nach Satz 3.8 ist das Runge-Kutta Verfahren konsistent und damit konvergent, falls $\sum_{j=1}^{p} \gamma_j = \gamma^\top e = 1$.
Es sei nun $f \in C^s(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$ für $s \in \{1, 2, 3, 4\}$. Nach Satz 3.13 genügt die Funktion $K : \mathbb{R}^m \times [t_0, T] \times [0, H] \to \mathbb{R}^m$ dem nichtlinearen Gleichungssystem

$$K_i(x,t;h) = f\Big(x + h\sum_{j=1}^{p} \beta_{ij} K_j(x,t;h), t + h\alpha_i\Big) \quad (i = 1, \ldots, p).$$

Da die Funktion

$$(h, K) \to f\Big(x + h\sum_{j=1}^{p} \beta_{ij} K_j, t + h\alpha_i\Big)$$

für festes (x,t) s-mal stetig differenzierbar ist, ist auch K (und damit Φ) bzgl. h s-mal stetig differenzierbar.
Nach Satz 3.8 ist die Konsistenzordnung gerade $s = 1$, falls $f \in C^1(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$ und deshalb die partielle Ableitung

$$\frac{\partial \Phi}{\partial h} \text{ auf } \mathbb{R}^m \times [t_0, T] \times [0, H] \text{ existiert und stetig ist und } \Phi(x,t;0) = f(x,t)$$

für alle $(x,t) \in \mathbb{R}^m \times [t_0, T]$ erfüllt ist. Letzteres haben wir bereits gezeigt.
Für die Konsistenzordnung $s = 2$ muss nach Satz 3.8 zusätzlich

$$\frac{\partial^2 \Phi}{\partial h^2} \text{ auf } \mathbb{R}^m \times [t_0, T] \times [0, H] \text{ existieren und stetig sein sowie}$$

$$\frac{\partial\Phi}{\partial h}(x,t;0) = \frac{1}{2}f^{(1)}(x,t) = \frac{1}{2}\Big(\frac{\partial f}{\partial x}(x,t)f(x,t) + \frac{\partial f}{\partial t}(x,t)\Big) \text{ gelten.}$$

Ersteres gilt wegen $f \in C^2(\mathbb{R}^m \times [t_0,T],\mathbb{R}^m)$. Zum Nachprüfen der zweiten Bedingung berechnen wir für $i \in \{1,\dots,p\}$:

$$\begin{aligned}
\frac{\partial K_i}{\partial h}(x,t;h) &= \frac{\partial f}{\partial x}\Big(x + h\sum_{j=1}^{p}\beta_{ij}K_j(x,t;h), t + h\alpha_i\Big)\Big(\sum_{j=1}^{p}\beta_{ij}K_j(x,t;h) \\
&\quad + h\sum_{j=1}^{p}\beta_{ij}\frac{\partial K_j}{\partial h}(x,t;h)\Big) + \frac{\partial f}{\partial t}\Big(x + h\sum_{j=1}^{p}\beta_{ij}K_j(x,t;h), t + h\alpha_i\Big)\alpha_i \\
\frac{\partial K_i}{\partial h}(x,t;0) &= \frac{\partial f}{\partial x}(x,t)\sum_{j=1}^{p}\beta_{ij}f(x,t) + \frac{\partial f}{\partial t}(x,t)\alpha_i
\end{aligned}$$

Daraus folgt

$$\begin{aligned}
\frac{\partial\Phi}{\partial h}(x,t;0) &= \sum_{i=1}^{p}\gamma_i\sum_{j=1}^{p}\beta_{ij}\frac{\partial f}{\partial x}(x,t)f(x,t) + \sum_{i=1}^{p}\gamma_i\alpha_i\frac{\partial f}{\partial t}(x,t) \\
&= \gamma^\top Be\frac{\partial f}{\partial x}(x,t)f(x,t) + \gamma^\top Ae\frac{\partial f}{\partial t}(x,t) \\
&= \frac{1}{2}\Big(\frac{\partial f}{\partial x}(x,t)f(x,t) + \frac{\partial f}{\partial t}(x,t)\Big) = \frac{1}{2}f^{(1)}(x,t)
\end{aligned}$$

für alle $(x,t) \in \mathbb{R}^m \times [t_0,T]$. Für $s \in \{3,4\}$ sei auf die Kapitel 5.5 und 5.6 im Buch von Crouzeix-Mignot (Masson, Paris, 1984) verwiesen. □

Beispiel 3.15

(a) Wir betrachten die Verfahrensklasse der linearen Einschrittverfahren

$$x_\ell = x_{\ell-1} + h_\ell(b_1 f(x_{\ell-1},t_{\ell-1}) + b_2 f(x_\ell,t_\ell)),\ \ell = 1,\dots,N.$$

Dies sind Runge-Kutta Verfahren mit $p = 2$, $\gamma = (b_1,b_2)$, $\alpha = (0,1)$ *und*

$K_1(x,t;h) = f(x,t)$, *d.h.* $\beta_{11} = \beta_{12} = 0$ *und*
$K_2(x,t;h) = f(x + h(b_1K_1(x,t;h) + b_2K_2(x,t;h), t+h))$, *d.h.* $\beta_{21} = b_1, \beta_{22} = b_2$.

Dies impliziert $K_2(x_{\ell-1},t_{\ell-1};h) = f(t_\ell,t_\ell)$, $\ell = 1,\dots,N$.
Erfüllt f *die Voraussetzungen aus Satz 3.14, so ist das Verfahren konvergent, falls* $b_1 + b_2 = \gamma^\top e = 1$. *Überdies gilt* $Be = \alpha = Ae$.
Konvergenzordnung $s = 1$ *liegt vor, falls außerdem* $f \in C^1(\mathbb{R}^m \times [t_0,T],\mathbb{R}^m)$.
Konvergenzordnung $s = 2$ *liegt vor, falls überdies* $\gamma^\top Ae = \frac{1}{2}$, *d.h.* $b_2 = \frac{1}{2}$ *und folglich auch* $b_1 = \frac{1}{2}$ *gilt.*
Also hat die Trapezregel Konvergenzordnung $s = 2$, *falls* $f \in C^2(\mathbb{R}^m \times [t_0,T],\mathbb{R}^m)$.
Das explizite und das implizite Euler-Verfahren haben Konvergenzordnung $s = 1$, *falls* $f \in C^1(\mathbb{R}^m \times [t_0,T],\mathbb{R}^m)$.

(b) Welche Konvergenzordnung kann ein einstufiges Runge-Kutta Verfahren erreichen ? Wir setzen also $p = 1$, $\gamma = 1$, $B = (\beta)$, $\beta = \alpha$ *und fordern* $\gamma\alpha = \alpha = \frac{1}{2}$. *Dann wird Konvergenzordnung* $s = 2$ *von folgendem Verfahren erreicht:*

$$\begin{aligned} x_\ell &= x_{\ell-1} + h_\ell K(x_{\ell-1}, t_{\ell-1}; h_\ell) \\ K(x,t;h) &= f\Big(x + \frac{h}{2}K(x,t;h), t + \frac{h}{2}\Big)\,,\ \forall (x,t;h) \in \mathbb{R}^m \times [t_0, T] \times [0, H]. \end{aligned}$$

$\rightsquigarrow \frac{h_\ell}{2}K(x_{\ell-1}, t_{\ell-1}; h_\ell) = \frac{1}{2}(x_\ell - x_{\ell-1})$
$\rightsquigarrow K(x_{\ell-1}, t_{\ell-1}; h_\ell) = f(\frac{1}{2}(x_{\ell-1} + x_\ell), \frac{1}{2}(t_{\ell-1} + t_\ell))$

$\rightsquigarrow$ implizite Mittelpunktregel*:*

$$x_\ell = x_{\ell-1} + h_\ell f\Big(\frac{1}{2}(x_{\ell-1} + x_\ell), \frac{1}{2}(t_{\ell-1} + t_\ell)\Big)\,,\ \ell = 1, \ldots, N.$$

(c) Welche Konvergenzordnung kann ein zweistufiges explizites Runge-Kutta Verfahren erreichen ? Wir wissen $p = 2$, $\gamma_1 + \gamma_2 = 1$ *und* $B = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix}$.
Konsistenzbedingungen für $s = 2$*:* $Be = \alpha \rightsquigarrow \alpha_1 = 0$ *und* $\alpha_2 = \beta_{21}$ *sowie* $\gamma^\top Ae = \gamma_2\alpha_2 = \frac{1}{2}$.
Zusätzliche Bedingungen für $s = 3$*:* $\gamma^\top A^2 e = \gamma_2\alpha_2^2 = \frac{1}{3}$ *und* $\gamma^\top BAe = 0 \neq \frac{1}{6}$.
Es wird also höchstens die Konvergenzordnung $s = 2$ *erreicht und zwar von der einparametrischen Verfahrensfamilie*

$$x_\ell = x_{\ell-1} + h_\ell(\gamma f(x_{\ell-1}, t_{\ell-1}) + (1-\gamma) f(x_{\ell-1} + \alpha h_\ell f(x_{\ell-1}, t_{\ell-1}), t_{\ell-1} + \alpha h_\ell))$$

für $\ell = 1, \ldots, N$, *wobei* $(1-\gamma)\alpha = \frac{1}{2}$.
Dazu gehört insbesondere das verbesserte Euler-Verfahren *mit* $\gamma = 0$ *und* $\alpha = \frac{1}{2}$*:*

$$x_\ell = x_{\ell-1} + h_\ell f\Big(x_{\ell-1} + \frac{h_\ell}{2} f(x_{\ell-1}, t_{\ell-1}), t_{\ell-1} + \frac{h_\ell}{2}\Big)\,,\ \ell = 1, \ldots, N.$$

(d) "Klassisches" Runge-Kutta Verfahren mit $p = 4$ *und Konvergenzordnung* $s = 4$*:*

$$\gamma = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{6} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \textit{und} \quad \alpha = Be\,.$$

Rechenbeispiel: *(Vergleich der obigen Verfahren)*
Wir betrachten die Differentialgleichung $x'(t) = x(t)$, $t \in [0, 1]$, $x(0) = x_0 = 1$, *d.h.* $x_*(t) = \exp(t)$. *Die Differentialgleichung wird numerisch integriert mit konstanter*

Schrittweite $h = 0.1$ *und*

Euler vorwärts	$x_\ell = (1+h)x_{\ell-1} = \prod_{j=1}^{\ell}(1+h)$	$x_{10} = 2.5337$
Trapezregel	$x_\ell = \frac{1+0.5h}{1-0.5h}x_{\ell-1} = \prod_{j=1}^{\ell}\frac{1+0.5h}{1-0.5h}$	$x_{10} = 2.72055$
verbesserter Euler	$x_\ell = (1+h+\frac{1}{2}h^2)x_{\ell-1} = \prod_{j=1}^{\ell}(1+h+\frac{1}{2}h^2)$	$x_{10} = 2.71408$
klassisches RKV	$x_\ell = \prod_{j=1}^{\ell}(1+h+\frac{1}{2}h^2+\frac{1}{6}h^3+\frac{1}{24}h^4)$	$x_{10} = 2.718279$

Die exakte Lösung ist $x_*(1) = e = 2.7182818$. *Man erkennt den Wert der höheren Konsistenzordnung.*

Definition 3.16
Vereinfachende Konsistenzbedingungen von Butcher *für RKV:*

Bedingung	$B(s)$	$\gamma^\top A^{k-1}e$	$= \frac{1}{k}$	$(k = 1, \dots s)$,
Bedingung	$C(l)$	$BA^{k-1}e$	$= \frac{1}{k}A^k e$	$(k = 1, \dots, l)$,
Bedingung	$D(q)$	$\gamma^\top A^{k-1}B$	$= \frac{1}{k}\gamma^\top(I - A^k)$	$(k = 1, \dots, q)$.

Lemma 3.17
Für ein p-stufiges RK-Verfahren gelte $\gamma_j \neq 0, j = 1, \dots, p$, *und die* $\alpha_j, j = 1, \dots, p$, *seien paarweise verschieden. Dann gilt:*
(i) $B(p+n), C(p) \Rightarrow D(n)$
(ii) $B(p+l), D(p) \Rightarrow C(l)$

Beweis:
Wir beweisen (i); (ii) wird analog bewiesen. $D(n)$ ist gleichbedeutend mit

$$0 = r_k := \gamma^\top A^{k-1}B - \frac{1}{k}\gamma^\top(I - A^k) \quad (k = 1, \dots, n).$$

Dazu betrachten wir die nach Voraussetzung invertierbare Vandermonde-Matrix

$$A_p := \begin{pmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_p \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{p-1} & \alpha_2^{p-1} & \cdots & \alpha_p^{p-1} \end{pmatrix}$$

und beweisen, dass

$$A_p r_k = 0 \quad (k = 1, \dots, n).$$

Seien $k \in \{1,\dots,n\}$, $\nu \in \{1,\dots,p\}$. Dann gilt

$$\begin{aligned}
[A_p r_k]_\nu &= \sum_{j=1}^{p}\sum_{i=1}^{p} \gamma_i \beta_{ij} \alpha_i^{k-1} \alpha_j^{\nu-1} - \sum_{j=1}^{p} \frac{1}{k}\gamma_j (1-\alpha_j^k)\alpha_j^{\nu-1} \\
&= \sum_{i=1}^{p} \gamma_i \alpha_i^{k-1} \underbrace{\sum_{j=1}^{p} \beta_{ij}\alpha_j^{\nu-1}}_{\substack{\frac{1}{\nu}\alpha_i^\nu \\ C(p)}} - \frac{1}{k}\underbrace{\sum_{j=1}^{p}\gamma_j \alpha_j^{k+\nu-1}}_{\substack{=\frac{1}{\nu} \\ B(p+n)}} + \frac{1}{k}\underbrace{\sum_{j=1}^{p}\gamma_j\alpha_j^{k+\nu-1}}_{\substack{=\frac{1}{k+\nu} \\ B(p+n)}} \\
&= \frac{1}{\nu}\underbrace{\sum_{i=1}^{p}\gamma_i \alpha_i^{\nu+k-1}}_{\substack{\frac{1}{k+\nu} \\ B(p+n)}} - \frac{1}{k}\frac{1}{\nu} + \frac{1}{k}\frac{1}{k+\nu} = \frac{1}{\nu}\frac{1}{k+\nu} - \frac{1}{k}\frac{1}{\nu} + \frac{1}{k}\frac{1}{k+\nu} \\
&= \frac{k-(k+\nu)+\nu}{\nu k(k+\nu)} = 0\,.
\end{aligned}$$

Wir erhalten also $A_p r_k = 0$ und damit $r_k = 0$, $k = 1,\dots,n$. □

Folgerung 3.18
Unter den Voraussetzungen von Lemma 3.17 gilt:
(i) $B(2p), C(p) \Rightarrow D(p)$
(ii) $B(2p-1), C(p) \Rightarrow D(p-1)$

Satz 3.19 *(Butcher 1964)*
Es seien die vereinfachenden Konsistenzbedingungen $B(s), C(l)$ und $D(q)$ mit $s \le l+q+1$,$s \le 2l+2$ für ein p-stufiges Runge-Kutta-Verfahren erfüllt. Dann besitzt es Konsistenzordnung s, falls $f \in C^s(\mathbb{R}^m \times [t_0,T], \mathbb{R}^m)$.

(ohne Beweis)

Bemerkung 3.20
Die vereinfachenden Konsistenzbedingungen lassen keine höhere Konsistenzordnung für p-stufige Runge-Kutta-Verfahren als $s = 2p$ zu!
Nach Satz 3.19 würde $s = 2p+1$ die Bedingung $B(2p+1)$ erfordern, d. h. $\gamma^\top A^{k-1} = \frac{1}{k}$, $k = 1,\dots,2p+1$.
Die Bedingung $B(n)$ ist äquivalent dazu, dass die Quadraturformel

$$(*) \qquad \int_{t_{\ell-1}}^{t_\ell} f(t)dt \approx h_\ell \sum_{i=1}^{p} \gamma_i f(t_{\ell-1} + \alpha_i h_\ell)$$

exakt für alle Polynome vom Grad $\le n-1$ ist.
Bew.: Sei f das Polynom $f(t) = (t - t_{\ell-1})^{k-1}$ mit $k \le n$. Dann gilt

$$\begin{aligned}
\int_{t_{\ell-1}}^{t_\ell} f(t)dt &= \frac{1}{k}h_\ell^k = h_\ell \sum_{i=1}^{p}\gamma_i \alpha_i^{k-1} h_\ell^{k-1} \\
&= h_\ell \sum_{i=1}^{p}\gamma_i f(t_{\ell-1}+\alpha_i h_\ell). \qquad \square
\end{aligned}$$

Also wäre $B(2p+1)$ äquivalent dazu, dass die Quadraturformel exakt für alle Polynome $\leq 2p$ ist, d.h. auch für das Polynom

$$f(t) = \prod_{j=1}^{p} (t - (t_{\ell-1} + \alpha_j h_\ell))^2.$$

Dafür gilt aber

$$h_\ell \sum_{i=1}^{p} \gamma_i f(t_{\ell-1} + \alpha_i h_\ell) = 0 < \int_{t_{\ell-1}}^{t_\ell} f(t)dt.$$

Deshalb ist $s = 2p$ die höchstmögliche Konsistenzordnung.

Wir zeigen als nächstes, dass die Konsistenzordnung $s = 2p$ durch geeignete Wahl von $\gamma, \alpha \in \mathbb{R}^p$ erreicht werden kann.

Satz 3.21
Es sei $B(p)$ erfüllt für γ und α in $\mathbb{R}^p$. Für $t_{\ell j} := t_{\ell-1} + \alpha_j h_\ell$, $j = 1, \ldots, p$, sei

$$\omega(t) = \prod_{j=1}^{p} (t - t_{\ell j})$$

das zugehörige Stützstellenpolynom.
Dann ist die Quadraturformel (*) genau für alle Polynome vom Grad $2p-1$ gdw.

$$\int_{t_{\ell-1}}^{t_\ell} \omega(t) t^j dt = 0\,,\ \forall j = 0, \ldots, p-1.$$

Beweis:

($\Rightarrow$) (*) sei genau für alle Polynome vom Grad $2p-1$. Sei $j \in \{0, \ldots, p-1\}$. Da das Polynom $t \to \omega(t)t^j$ einen Grad $\leq 2p-1$ hat, ist (*) dafür genau, d.h. es gilt

$$\int_{t_{\ell-1}}^{t_\ell} \omega(t) t^j = h_\ell \sum_{i=1}^{p} \gamma_i \omega(t_{\ell i}) t_{\ell i}^j = 0\,.$$

($\Leftarrow$) Es gelte

$$\int_{t_{\ell-1}}^{t_\ell} \omega(t) t^j dt = 0\,,\ j = 0, \ldots, p-1.$$

Zu zeigen: $\int_{t_{\ell-1}}^{t_\ell} f(t)dt = h_\ell \sum_{i=1}^{p} \gamma_i f(t_{\ell i})$ für jedes Polynom f vom Grad $\leq 2p-1$.
Wir schreiben ein solches Polynom f in der Form

$$f = \omega q + r \quad \text{(Polynom-Division von } f \text{ durch } \omega \text{ mit Rest)}$$

Dann gilt: Der Grad von q und r ist $\leq p-1$.

$$\begin{aligned} \rightsquigarrow \qquad f(t_{\ell j}) &= r(t_{\ell j}), \quad j=1,\dots,p, \text{ und} \\ \int\limits_{t_{\ell-1}}^{t_\ell} f(t)dt &= \underbrace{\sum_{j=0}^{s-1} a_j \int\limits_{t_{\ell-1}}^{t_\ell} \omega(t)t^j dt}_{=0} + \int\limits_{t_{\ell-1}}^{t_\ell} r(t)dt \end{aligned}$$

Nach Bemerkung 3.20 folgt:

$$\int\limits_{t_{\ell-1}}^{t_\ell} f(t)dt = \int\limits_{t_{\ell-1}}^{t_\ell} r(t)dt = h_\ell \sum_{i=1}^{p} \gamma_i r(t_{\ell i}) = h_\ell \sum_{i=1}^{p} \gamma_i f(t_{\ell i})$$

□

Bemerkung 3.22
Wählt man $\alpha = (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^p$ so, daß die Bedingung in Satz 3.21 erfüllt ist, ergibt sich:

$$\begin{aligned} \int_{t_{\ell-1}}^{t_\ell} \omega(t)t^j dt &= \int_{t_{\ell-1}}^{t_\ell} \prod_{i=1}^{p} (t-(t_{\ell-1}+\alpha_j h_\ell))t^j dt \quad (\textit{Subst. } t = t_{\ell-1}+\tau h_\ell) \\ &= h_\ell \int\limits_0^1 \prod_{i=1}^{p} (\tau h_\ell - \alpha_i h_\ell)(t_{\ell-1}+\tau h_\ell)^j dt \\ &= h_\ell^{p+1} \int\limits_0^1 \prod_{i=1}^{p} (\tau-\alpha_i)(t_{\ell-1}+\tau h_\ell)^j d\tau = 0 \quad (\forall j = 0,\dots,p-1). \end{aligned}$$

Daraus folgt nach Satz 3.21:

$$\begin{aligned} \int_{t_{\ell-1}}^{t_\ell} \omega(t)t^j dt = 0 \ (\forall j=0,\dots,p-1) &\Leftrightarrow \int\limits_0^1 \prod_{i=1}^{p} (\tau-\alpha_i)\tau^j d\tau = 0 \ (\forall j=0,\dots,p-1) \\ &\Leftrightarrow B(2p) \end{aligned}$$

Satz 3.23
Es sei P ein Polynom vom Grad p und es gelte

$$\int\limits_0^1 P(t)t^j dt = 0, \quad \forall j = 0,\dots,p-1.$$

Dann hat P genau p paarweise verschiedene Nullstellen in $(0,1)$.

Beweis:
Es seien $t_j \in \mathbb{C}, j = 1,\dots,n \leq p$ die Nullstellen von P mit den algebraischen Vielfachheiten $\nu_j, \jmath = 1,\dots,n$, d.h.

$$P(t) = c\prod_{j=1}^{n} (t-t_j)^{\nu_j} \text{ mit } c \neq 0, \sum_{j=1}^{n} \nu_j = p.$$

Es seien $M = \{\tilde{t}_1, \ldots, \tilde{t}_m\} = \{t_j \in (0,1) : \nu_j \text{ ist ungerade }\} \rightsquigarrow 0 \leq m \leq n \leq p$

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

Dann besitzen die Nullstellen des Polynoms $P \cdot r$ in $(0,1)$ gerade Vielfachheit. Also wechselt $P \cdot r$ das Vorzeichen in $(0,1)$ <u>nicht</u>, d. h.

$$\int_0^1 P(t)r(t)dt \neq 0.$$

Da aber nach Voraussetzung $\int_0^1 P(t)r(t)dt = 0$ gilt, falls r ein Polynom vom Grad $\leq p-1$ ist, muß r vom Grad p sein. $\rightsquigarrow m = n = p$ und M enthält p Elemente.
$\rightsquigarrow t_1, \ldots, t_p \in (0,1)$ und alle Nullstellen sind einfach. □

Satz 3.24
Ist die Quadraturformel

$$\int_{t_{\ell-1}}^{t_\ell} f(t)dt \approx h_\ell \sum_{i=1}^{p} \gamma_i f(t_{\ell-1} + \alpha_i h_\ell)$$

genau für alle Polynome f vom Grad $\leq 2p-2$, so gilt $\gamma_i > 0$, $\forall i = 1, \ldots, p$.

<u>Beweis:</u>
Wir betrachten die bei der Lagrange'schen Form des Interpolationspolynoms auftretenden Polynome $(p-1)$-ten Grades

$$L_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{p} \frac{t - t_{\ell j}}{t_{\ell i} - t_{\ell j}}, \quad i = 1, \ldots, p, \; t_{\ell j} := t_{\ell-1} + \alpha_j h_\ell, \; j = 1, \ldots, p.$$

Dann gilt:

$$0 < \int_{t_{\ell-1}}^{t_\ell} L_i^2(t)dt = h_\ell \sum_{j=1}^{p} \gamma_j \underbrace{L_i^2(t_{\ell j})}_{=\delta_{ij}} = h_\ell \gamma_i, \quad i = 1 \ldots, p.$$

□

Die Konsistenzbedingungen aus Definition 3.16 lassen sich auch wie folgt ausdrücken:

$$\begin{array}{lcl} B(p) & \Leftrightarrow & A_p \gamma = e_p \\ C(p) & \Leftrightarrow & B A_p^\top = C_p \\ D(p) & \Leftrightarrow & A_p \text{diag } (\gamma_1, \ldots, \gamma_p) B = (N_p - C_p) \text{ diag } (\gamma_1, \ldots, \gamma_p) \end{array}$$

wobei A_p wie im Beweis von Lemma 3.17 definiert ist und

$$C_p = \begin{pmatrix} \alpha_1 & \frac{\alpha_1^2}{2} & \dots & \frac{\alpha_1^p}{p} \\ \vdots & \vdots & & \\ \alpha_p & \frac{\alpha_p^2}{2} & \dots & \frac{\alpha_p^p}{p} \end{pmatrix} \in \mathbb{R}^{p\times p} \quad e_p = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{p} \end{pmatrix} \in \mathbb{R}^p$$

$$N_p = \begin{pmatrix} 1 & \frac{1}{2} & \dots & \frac{1}{p} \\ \vdots & \vdots & & \\ 1 & \frac{1}{2} & \dots & \frac{1}{p} \end{pmatrix} \in \mathbb{R}^{p\times p}.$$

Ist also A_p invertierbar, so gilt $\gamma = A_p^{-1}e_p$ und $B = C_p(A_p^{-1})^\top$, d.h. die Runge-Kutta Parameter γ und B lassen sich aus α bestimmen.

Satz 3.25 *(Gaußsche implizite Runge-Kutte-Verfahren)*
Es sei P ein Polynom vom Grad p, das die Bedingung

$$(*) \qquad \int_0^1 P(t)t^j dt = 0, \quad j = 0, \dots, p-1.$$

erfüllt und $\alpha_1, \dots, \alpha_p$ seien seine Nullstellen. Wir betrachten die zugehörige Vandermonde-Matrix A_p (vgl. Bew. von 3.17) und wir definieren die Runge-Kutta Parameter

$$\gamma = A_p^{-1}e_p, \quad B = C_p(A_p^{-1})^\top.$$

Das zugehörige p-stufige implizite Runge-Kutta-Verfahren besitzt Konsistenzordnung $s = 2p$, falls $f \in C^{2p}(\mathbb{R}^m \times [t_0, T], \mathbb{R}^m)$.

Beweis:
Ein solches Polynom P vom Grad p, das die Bedingung (*) erfüllt, kann stets konstruiert werden. Macht man nämlich den Ansatz

$$P(t) = t^p + \sum_{i=1}^{p} a_i t^{i-1},$$

so genügen die Koeffizienten a_i, $i = 1, \dots, p$, wegen (*) dem linearen Gleichungssystem

$$\sum_{i=1}^{p} \frac{1}{i+j}\, a_i = -\frac{1}{p+j+1} \quad (j = 0, \dots, p-1),$$

dessen quadratische Matrix gerade die Hilbert-Matrix der Ordnung p (vgl. Vorlesung Wiss. Rechnen II) und deshalb nichtsingulär ist. Die Nullstellen $\alpha_1, \dots, \alpha_p$ von P sind nach Satz 3.23 paarweise verschieden und gehören zu $(0, 1)$.
Dann ist A_p invertierbar und nach Voraussetzung sind $B(p)$ und $C(p)$ erfüllt. Aus Satz 3.21 und Bemerkung 3.22 folgt, dass sogar $B(2p)$ gültig ist. Gemeinsam mit Lemma 3.17 resultieren also $B(2p)$, $C(p)$ und $D(p)$. Schließlich folgt aus Satz 3.19, dass das Runge-Kutta Verfahren die Konsistenzordnung $s = 2p$ besitzt. □

Beispiel 3.26 *(spezielle Gaußsche implizite Runge-Kutta-Verfahren)*

(a) $p = 1$*: Die Bedingungen* $B(2p) = B(2)$ *und* $C(p) = C(1)$ *bedeuten*

$$\gamma\alpha^{k-1} = \frac{1}{k},\ k = 1, 2,\ \textit{und}\ \beta = \alpha$$

$\rightsquigarrow \gamma = 1,\ \alpha = \frac{1}{2},\ \beta = \frac{1}{2}$.
$\rightsquigarrow$ *implizite Mittelpunktregel mit Konsistenzordnung* $s = 2p = 2$.

(b) $p = 2$*: Das Polynom* $P(t) = t^2 + a_2 t + a_1$ *berechnet sich aus dem linearen Gleichungssystem*

$$\frac{1}{j+1}\,a_1 + \frac{1}{j+2}\,a_2 = -\frac{1}{j+3} \quad (j = 0, 1)$$

$\rightsquigarrow P(t) = t^2 - t + \frac{1}{6}$ *mit den Nullstellen* $\alpha_{1/2} = \frac{1}{2}\left(1 \mp \frac{1}{3}\sqrt{3}\right)$.

$$\gamma = A_2^{-1} e_2 = \frac{1}{\alpha_2 - \alpha_1} \begin{pmatrix} \alpha_2 & -1 \\ -\alpha_1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

$$\begin{aligned} B \;&=\; C_2 (A_2^{-1})^\top = \frac{1}{\alpha_2 - \alpha_1} \begin{pmatrix} \alpha_1 & \frac{\alpha_1^2}{2} \\ \alpha_2 & \frac{\alpha_2^2}{2} \end{pmatrix} \begin{pmatrix} \alpha_2 & -\alpha_1 \\ -1 & 1 \end{pmatrix} \\ &=\; \frac{1}{\alpha_2 - \alpha_1} \begin{pmatrix} \alpha_1\alpha_2 - \frac{\alpha_1^2}{2} & -\frac{\alpha_1^2}{2} \\ \frac{\alpha_2^2}{2} & -\alpha_1\alpha_2 + \frac{\alpha_2^2}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \end{pmatrix} \end{aligned}$$

Die Konsistenzordnung dieses impliziten Runge-Kutta Verfahrens ist $s = 2p = 4$.

Bemerkung 3.27
Radau-Methoden: Fixiere $\alpha_1 = 0$ *oder* $\alpha_p = 1$ *und wähle die restlichen Komponenten von* α *wie vorher* $\rightsquigarrow B(2p-1)$.
Beispiel: Radau IIa: $\alpha_p = 1,\ B = C_p (A_p^{-1})^\top,\ \gamma = A_p^{-1} e_p$
$\rightsquigarrow B(2p-1), C(p) \rightsquigarrow D(p-1)$ *(Folgerung 3.23)*
$\rightsquigarrow$ *Konsistenzordnung:* $s = 2p - 1$
($p = 1$ *: implizites Euler-Verfahren)*

Lobatto-Methoden: Fixiere $\alpha_1 = 0$ *und* $\alpha_p = 1$ *und wähle die restlichen Komponenten von* α *wie vorher.*
$\rightsquigarrow$ *Konsistenzordnung:* $s = 2p - 2$
(verschiedene Varianten zur Wahl von B*: erfülle* $C(p)$*,* $D(p)$ *und* $B(2p-2)$*)*

Bemerkung 3.28
Trotz vieler guter Eigenschaften (hohe Konsistenzordnung, günstiges asymptotisches Verhalten vgl. Kap. 3.6) impliziter Runge-Kutta-Verfahren ist ihre Anwendung limitiert durch den hohen Aufwand der Lösung des nichtlinearen Gleichungssystems

$$K_i = f\Big(x + h_\ell \sum_{j=1}^{p} \beta_{ij} K_j, t + h_\ell \alpha_i\Big) \quad (i = 1, \ldots, p).$$

der Dimension $p \cdot m$.
Für semiimplizite (bzw. diagonalimplizite) Runge-Kutta-Verfahren mit $\beta_{ij} = 0$, falls $i < j$, (zusätzlich $\beta_{ii} = \beta$, $i = 1, \ldots, p$) zerfällt das mp-dimensionale nichtlineare Gleichungssystem in p $m-$dimensionale Gleichungssysteme. Gauß-, Radau- und Lobatto-Verfahren sind i.a. nicht semiimplizit.

Literatur:
J.C. Butcher: The Numerical Analysis of Ordinary Differential Equations, Wiley, Chichester, 1987.

3.5 Lineare Mehrschrittverfahren

Wir betrachten ein k-schrittiges lineares Mehrschrittverfahren auf einem Gitter $G = \{t_0 < t_1 < \cdots, t_N = T\}$ der Form

$$\sum_{j=1}^{k} a_{\ell j} x_{\ell - j} = h_\ell \sum_{j=0}^{k} b_{\ell j} f(x_{\ell - j}, t_{\ell - j}) \quad (\ell = 1, \ldots, N),$$

wobei $k \in \mathbb{N}$, $a_{\ell j}$, $b_{\ell j} \in \mathbb{R}$, $a_{\ell 0} = 1$, $a_{\ell j} = 0$, $j = \ell + 1, \ldots, k$. Im Unterschied zu Runge-Kutta Verfahren hängen die Koeffizienten $a_{\ell j}$, $b_{\ell j}$, $j = 1, \ldots, k$, $\ell = 1, \ldots, N$, vom Gitter G ab.

Definition 3.29 *(Klassen von Gittern)*
Die Menge $\mathcal{G}_0 = \{G_h : \frac{(T-t_0)}{h} \in \mathbb{N}\}$ heißt Klasse der äquidistanten Gitter in $[t_0, T]$. Eine Menge $\mathcal{G}$ heißt Klasse von zulässigen *Gittern in $[t_0, T]$, falls Konstanten $C_i > 0$, $i = 1, 2, 3$, existieren, so dass für alle Gitter $G = \{t_0 < t_1 < \cdots, t_N = T\} \in \mathcal{G}$ gilt:*

$$C_1 \leq \frac{h_\ell}{h_{\ell - 1}} \leq C_2 \quad (\forall \ell = 2, \ldots, N), \quad h(G) N(G) \leq C_3 \,,$$

wobei $h_\ell := t_\ell - t_{\ell - 1}$, $\forall \ell = 1, \ldots, N$, $h(G) := \max_{\ell = 1, \ldots, N} h_\ell$.
Bezeichnung*: $\mathcal{G} = \mathcal{G}(C_1, C_2, C_3)$.*

Satz 3.30 *(Konsistenz)*
Ein k-schrittiges lineares Mehrschrittverfahren ist konsistent auf einer Klasse $\mathcal{G}$ von zulässigen Gittern, falls eine Konstante $K > 0$ existiert, so dass

$$(*) \quad \sum_{j=1}^{k} |a_{\ell j}| \leq K \;\; \textit{und} \;\; \sum_{j=0}^{k} |b_{\ell j}| \leq K \quad (\forall \ell = 1, \ldots, N)$$

gilt und die Bedingungen

$$(**) \quad \sum_{j=0}^{k} a_{\ell j} = 0 \;\; \textit{und} \;\; \sum_{j=0}^{k} [a_{\ell j}(t_{\ell - j} - t_\ell) - h_\ell b_{\ell j}] = 0 \quad (\forall \ell = 1, \ldots, N)$$

für alle Gitter $G = \{t_0 < t_1 < \cdots, < t_N = T\} \in \mathcal{G}$ erfüllt sind.
Es besitzt die Konsistenzordnung $s \leq 2k$, falls zusätzlich die linearen Gleichungen

$$(***) \quad \sum_{j=0}^{k} [a_{\ell j}(t_{\ell-j} - t_\ell) - h_\ell b_{\ell j} i](t_{\ell-j} - t_\ell)^{i-1} = 0 \quad (\forall i = 1, \ldots, s,\ \ell = 1, \ldots, N)$$

gültig sind und $f \in C^s([t_0, T] \times \mathbb{R}^m, \mathbb{R}^m)$ gilt. Die maximale Konsistenzordnung für k-schrittige implizite (explizite) lineare Mehrschrittverfahren ist $s = 2k$ ($s = 2k - 1$).

Beweis: Wir betrachten den lokalen Diskretisierungsfehler $\|\tau_\ell(G)\|$ auf $G \in \mathcal{G}$ für festes $\ell \in \{1, \ldots, N\}$ und beginnen mit

$$\begin{aligned} \tau_\ell(G) &= \frac{1}{h_\ell} \sum_{j=0}^{k} a_{\ell j} x_*(t_{\ell-j}) - \sum_{j=0}^{k} b_{\ell j} f(x_*(t_{\ell-j}), t_{\ell-j}) \\ &= \frac{1}{h_\ell} \sum_{j=0}^{k} [a_{\ell j} x_*(t_{\ell-j}) - h_\ell b_{\ell j} x'_*(t_{\ell-j})] \end{aligned}$$

Nach dem Mittelwertsatz existiert für jedes $j \in \{1, \ldots, k\}$ und jede Komponente $i \in \{1, \ldots, m\}$ ein $\xi_{ji} \in (t_{\ell-j}, t_\ell)$, so dass

$$x_*(t_{\ell-j}) = x_*(t_\ell) + x'_*(t_\ell)(t_{\ell-j} - t_\ell) + (x'_*(\xi_j) - x'_*(t_\ell))(t_{\ell-j} - t_\ell)\,,$$

wobei analog zum Beweis von Satz 3.8 in jeder Komponente i von $x'_*(\xi_j)$ das Argument ξ_{ji} auftritt. Dann gilt

$$\begin{aligned} \tau_\ell(G) &= \frac{1}{h_\ell} \sum_{j=0}^{k} a_{\ell j} x_*(t_\ell) + \frac{1}{h_\ell} \sum_{j=0}^{k} [a_{\ell j}(t_{\ell-j} - t_\ell) - h_\ell b_{\ell j}] x'_*(t_\ell) \\ &\quad + \frac{1}{h_\ell} \sum_{j=1}^{k} a_{\ell j}(t_{\ell-j} - t_\ell)(x'_*(\xi_j) - x'_*(t_\ell)) + \sum_{j=0}^{k} b_{\ell j}(x'_*(t_\ell) - x'_*(t_{\ell-j})) \end{aligned}$$

Sind $(*)$ und $(**)$ erfüllt, so folgt daraus

$$\|\tau_\ell(G)\| \leq K\Big(\Big|\frac{t_{\ell-k} - t_\ell}{h_\ell}\Big| + 1\Big) \max_{|t-s| \leq kh} \|x'_*(t) - x'_*(s)\|$$

und damit die Konsistenz wegen der gleichmäßigen Stetigkeit von x'_* auf $[t_0, T]$ und weil $|\frac{t_{\ell-k} - t_\ell}{h_\ell}|$ für zulässige Gitter gleichmäßig beschränkt ist.
Für $r \in \mathbb{N}$, $r \leq 2k$, gelte jetzt $f \in C^r([t_0, T] \times \mathbb{R}^m, \mathbb{R}^m)$. Dann ist x_* $(r+1)$-mal stetig differenzierbar und für jedes $j \in \{1, \ldots, k\}$ und jede Komponente $l \in \{1, \ldots, m\}$ von x_* existieren $\theta_{jl} \in (0, 1)$ und $\hat{\theta}_{jl}$, so dass

$$\begin{aligned} x_*(t_{\ell-j}) &= \sum_{i=0}^{r} \frac{x_*^{(i)}(t_\ell)}{i!}(t_{\ell-j} - t_\ell)^i + \frac{x_*^{(r+1)}(t_\ell + \theta_j(t_{\ell-j} - t_\ell))}{(r+1)!}(t_{\ell-j} - t_\ell)^{r+1} \\ x'_*(t_{\ell-j}) &= \sum_{i=1}^{r} \frac{x_*^{(i)}(t_\ell)}{(i-1)!}(t_{\ell-j} - t_\ell)^{i-1} + \frac{x_*^{(r+1)}(t_\ell + \hat{\theta}_j(t_{\ell-j} - t_\ell))}{r!}(t_{\ell-j} - t_\ell)^r. \end{aligned}$$

Hierbei steht im Argument der jeweiligen l-ten Komponente θ_{jl} (bzw. $\hat{\theta}_{jl}$) anstelle von θ_j (bzw. $\hat{\theta}_j$). Insgesamt ergibt sich für den lokalen Diskretisierungsfehler

$$\begin{aligned}\tau_\ell(G) = &\frac{1}{h_\ell}\Big(\sum_{j=0}^{k} a_{\ell j}\Big)x_*(t_\ell) + \frac{1}{h_\ell}\sum_{i=1}^{r}\sum_{j=0}^{k}\Big[a_{\ell j}\frac{(t_{\ell-j}-t_\ell)^i}{i!} - h_\ell b_{\ell j}\frac{(t_{\ell-j}-t_\ell)^{i-1}}{(i-1)!}\Big]x_*^{(i)}(t_\ell)\\ &+\frac{1}{h_\ell}\sum_{j=0}^{k} a_{\ell j}\frac{(t_{\ell-j}-t_\ell)^{r+1}}{(r+1)!}\,x_*^{(r+1)}(t_\ell+\theta_j(t_{\ell-j}-t_\ell))\\ &+\sum_{j=0}^{k} b_{\ell j}\frac{(t_{\ell-j}-t_\ell)^{r}}{r!}\,x_*^{(r+1)}(t_\ell+\hat{\theta}_j(t_{\ell-j}-t_\ell))\\ \|\tau_\ell(G)\| \le\ & K\frac{|t_{\ell-k}-t_\ell|^r}{(r+1)!}\Big[\Big(\Big|\frac{t_{\ell-k}-t_\ell}{h_\ell}\Big|+r+1\Big)\max_{t\in[t_0,T]}\|x_*^{(r+1)}(t)\| = O(h(G)^r)\,,\end{aligned}$$

da wieder $|\frac{t_{\ell-j}-t_\ell}{h_\ell}|$ für zulässige Gitter gleichmäßig beschränkt ist.
Die linearen Gleichungen (**) und (***) enthalten die $s = 2k+1$ ($s = 2k$) Unbekannten $a_{\ell j}$, $j = 1,\dots,k$, $b_{\ell j}$, $j = 0(1),\dots,k$, im impliziten (expliziten) Fall. Deshalb wären sie unlösbar bei mehr als s Gleichungen. □

Bemerkung 3.31 *Im Fall von äquidistanten Gittern $G \in \mathcal{G}_0$ hängen die Konsistenzbedingungen (**) und (***) nicht mehr vom Gitter ab und nehmen die Form an*

$$\sum_{j=0}^{k} a_{\ell j} = 0 \quad \textit{und} \quad \sum_{j=0}^{k}(j^{\,i}a_{\ell j} - b_{\ell j} i j^{\,i-1}) = 0 \quad (i = 1,\dots,s).$$

Deshalb sind die Koeffizienten $a_{\ell j}$ und $b_{\ell j}$ unabhängig von ℓ, d.h.

$$a_j = a_{\ell j} \quad \textit{und} \quad b_j = b_{\ell j} \quad (j = 0,\dots,k).$$

Der Beweis von Satz 3.30 zeigt auch, dass Konsistenzordnung s bedeutet, dass das lineare Mehrschrittverfahren die Differentialgleichung exakt integriert, falls die Lösung x_* ein Polynom vom Grad s ist. Dann verschwinden die Restglieder und die Koeffizienten erfüllen die Gleichungen (**) und (***).

Beispiel 3.32 *(Adams-Verfahren)*
Ausgangspunkt: aufintegrierte Differentialgleichung in Gitter G

$$(*) \qquad x(t_\ell) = x(t_{\ell-1}) + \int_{t_{\ell-1}}^{t_\ell} f(x(t),t)dt,\ \ell = 1,\dots,N.$$

Der Integrand $f(x(\cdot),\cdot)$ wird in den Stützstellen
(i) $t_{\ell-k},\dots,t_\ell$ (implizites Verfahren)
(ii) $t_{\ell-k},\dots,t_{\ell-1}$ (explizites Verfahren)
interpoliert. Das Interpolationspolynom hat die Form

$$P(t) = \sum_{j=a}^{k}\prod_{\substack{i=a\\ i\neq j}}^{k}\frac{t-t_{\ell-i}}{t_{\ell-j}-t_{\ell-i}} f(x(t_{\ell-j}),t_{\ell-j}) \quad (a = 0 \textit{ für } (i),\ a = 1 \textit{ für } (ii)).$$

Gemeinsam mit (*) entsteht das Verfahren

$$\begin{aligned} x_\ell &= x_{\ell-1} + \sum_{j=a}^{k} \int_{t_{\ell-1}}^{t_\ell} \prod_{\substack{i=a \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} dt\, f(x_{\ell-j}, t_{\ell-j}) \\ &= x_{\ell-1} + h_\ell \sum_{j=a}^{k} b_{\ell j} f(x_{\ell-j}, t_{\ell-j}), \quad \textit{wobei} \\ b_{\ell j} &= \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} \prod_{\substack{i=a \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} dt \quad (\textit{Subst.: } t = t_{\ell-1} + \tau h_\ell) \\ &= \int_0^1 \prod_{\substack{i=a \\ i\neq j}}^{k} \frac{\tau h_\ell + t_{\ell-1} - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} d\tau \quad (j = a, \ldots, k) \\ &= \begin{cases} \int_0^1 \prod_{i=0}^{j-1} \frac{(1-\tau)h_\ell + h_\ell + \cdots + h_{\ell-i+1}}{h_{\ell-i} + \cdots + h_{\ell-j+1}} \prod_{i=j+1}^{k} \frac{(\tau-1)h_\ell + h_\ell + \cdots + h_{\ell-i+1}}{h_{\ell-j} + \cdots + h_{\ell-k+1}} d\tau & , \ a = 0 \\ \int_0^1 \prod_{i=1}^{j-1} \frac{-\tau h_\ell + h_{\ell-1} + \cdots + h_{\ell-j+2}}{h_{\ell-i} + \cdots + h_{\ell-j+1}} \prod_{i=j+1}^{k} \frac{\tau h_\ell + h_{\ell-1} + \cdots + h_{\ell-j+2}}{h_{\ell-j} + \cdots + h_{\ell-k+1}} d\tau & , \ a = 1 \end{cases} \end{aligned}$$

Eine Division von Zähler und Nenner unter den Produktzeichen durch $h_{\ell-1}$ für $\ell = 2, \ldots, N$ zeigt, dass die Koeffizienten Funktionen der k Argumente $\kappa_\ell, \ldots, \kappa_{\ell-k+1}$ sind, wobei $\kappa_\ell = \frac{h_\ell}{h_{\ell-1}}$.
Bezeichnung:
$a = 0$: Adams-Moulton-Verfahren
$a = 1$: Adams-Bashforth-Verfahren.
Beispiele:

(1) $a = k = 0$: $x_\ell = x_{\ell-1} + h_\ell b_{\ell 0} f(x_\ell, t_\ell)$, *wobei* $b_{\ell 0} = \int_0^1 d\tau = 1$ (implizites Euler-Verfahren).

(2) $k = 1$, $a = 0$: $x_\ell = x_{\ell-1} + h_\ell(b_{\ell 0} f(x_\ell, t_\ell) + b_{\ell 1} f(x_{\ell-1}, t_{\ell-1}))$, *wobei*

$$\begin{aligned} b_{\ell 0} &= \int_0^1 \frac{\tau h_\ell}{t_\ell - t_{\ell-1}} d\tau = \int_0^1 \tau d\tau = \frac{1}{2} \\ b_{\ell 1} &= \int_0^1 \frac{\tau h_\ell + t_{\ell-1} - t_\ell}{t_{\ell-1} - t_\ell} = \int_0^1 (-\tau + 1) d\tau = \Big[\tau - \frac{\tau^2}{2}\Big]_0^1 = \frac{1}{2} \end{aligned}$$

(Trapezregel)

(3) $k = 2$, $a = 1$: $x_\ell = x_{\ell-1} + h_\ell(b_{\ell 1} f(x_{\ell-1}, t_{\ell-1}) + b_{\ell 2} f(x_{\ell-2}, t_{\ell-2}))$, *wobei*

$$\begin{aligned} b_{\ell 1} &= \int_0^1 \frac{\tau h_\ell + h_{\ell-1}}{h_{\ell-1}} d\tau = \Big[\frac{\tau^2}{2}\frac{h_\ell}{h_{\ell-1}} + \tau\Big]_0^1 = \frac{h_\ell}{2h_{\ell-1}} + 1 \\ b_{\ell 2} &= \int_0^1 \frac{\tau h_\ell}{-h_{\ell-1}} d\tau = \Big[-\frac{\tau^2}{2}\frac{h_\ell}{h_{\ell-1}}\Big]_0^1 = -\frac{h_\ell}{2h_{\ell-1}} \end{aligned}$$

(zweischrittiges Adams-Bashforth-Verfahren)

Für die effiziente Implementierung der Adams-Verfahren ist es notwendig, eine Möglichkeit zu finden, die Koeffizienten $b_{\ell j}$, $j = a, \ldots, k$, schnell zu berechnen. Dies wurde von Krogh mittels eines geigneten Zusammenspiels von Lagrange- und Newton-Form des Interpolationspolynoms realisiert (vgl. Chapt. III.5, Hairer-Nørsett-Wanner 1993).

Satz 3.33 *Für die k-schrittigen Adams-Verfahren sind die Summen $\sum_{j=a}^{k} |b_{\ell j}|$ auf jeder Klasse zulässiger Gitter gleichmäßig beschränkt. Sie besitzen die Konsistenzordnung $s = k + 1 - a$.*

Beweis: Untersucht man die Konsistenz der Adams-Verfahren, so bedeutet ihr Verfahrensansatz, dass die Ableitung der Lösung x_* durch ein Polynom der Ordnung $k - a$ interpoliert wird. Ist also x_* selbst ein Polynom der Ordnung $k + 1 - a$, so wird x'_* exakt dadurch dargestellt und die Differentialgleichung exakt integriert. Deshalb ist die Konsistenzordnung $s = k + 1 - a$ (vgl. Bem. 3.31).
Die als letzte angegebene Darstellung für die $b_{\ell j}$, $j = a, \ldots, k$, zeigt bei Division durch $h_{\ell-1}$, dass der Integrand eine rationale Funktion von $(\kappa_{\ell-k+1}, \ldots, \kappa_\ell)$ ist. Diese ist aber für zulässige Gitter gleichmäßig beschränkt. □

Definition 3.34
Man sagt, ein Polynom p mit $p(1) = 0$ erfüllt die (starke) Wurzelbedingung nach Dahlquist*, falls alle seine Nullstellen im Einheitskreis $\{\lambda \in \mathbb{C} : |\lambda| \leq 1\}$ liegen und alle Nullstellen auf seinem Rand einfach sind (und lediglich $\lambda = 1$ auf seinem Rand liegt).*

Lemma 3.35
Es sei B eine reelle $k \times k$-Matrix, deren charakteristisches Polynom die Wurzelbedingung nach Dahlquist erfüllt. Dann existiert eine Norm $\|\cdot\|_*$ auf $\mathbb{C}^k$, so dass

$$\|B\|_* = \max_{\substack{x \in \mathbb{C}^k \\ \|x\|_* \leq 1}} \|Bx\|_* \leq 1 \,.$$

Analog existiert für die Tensorprodukt-Matrix $B \otimes I$ vom Typ $km \times km$ mit der $m \times m$ Einheitsmatrix I eine Norm $\|\cdot\|_*$ auf $\mathbb{C}^{km}$, so dass

$$\|B \otimes I\|_* \leq 1 \,.$$

Dabei ist $B \otimes I$ die Matrix mit den $m \times m$ Blöcken $b_{ij} I$, $i, j = 1, \ldots, k$.

Beweis: Es seien $\lambda_1, \ldots, \lambda_r$, $r \leq k$, die Eigenwerte von B, die auf dem Rand des Einheitskreises von $\mathbb{C}$ liegen. Sei $J \in \mathbb{C}^{k \times k}$ die Jordan-Normalform zu B, d.h. es existiert eine invertierbare Matrix $T \in \mathbb{C}^{k \times k}$, so dass

$$J = T^{-1} B T = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \lambda_r & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_{r+1} & * & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_{k-1} & * \\ 0 & 0 & 0 & \cdots & 0 & 0 & \lambda_k \end{pmatrix}$$

wobei λ_i, $i = 1, \ldots, k$, die Eigenwerte von B sind und $*$ für 0 oder 1 steht. Wir wählen nun $\varepsilon > 0$ so, dass $\max_{j=r+1,\ldots,k} |\lambda_j| + \varepsilon \leq 1$. Es bezeichne D_ε die Diagonalmatrix

$$D_\varepsilon = \operatorname{diag}(1, \varepsilon, \ldots, \varepsilon^{k-1})$$

und wir betrachten die Matrix

$$J_\varepsilon := D_\varepsilon^{-1} J D_\varepsilon = (TD_\varepsilon)^{-1} B (TD_\varepsilon).$$

Dann hat J_ε die Form

$$J_\varepsilon = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \lambda_r & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_{r+1} & *_\varepsilon & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_{k-1} & *_\varepsilon \\ 0 & 0 & 0 & \cdots & 0 & 0 & \lambda_k \end{pmatrix}$$

wobei $*_\varepsilon$ für 0 oder ε steht.
Wir definieren nun die Norm $\|\cdot\|_*$ auf $\mathbb{C}^k$ durch

$$\|x\|_* := \|(TD_\varepsilon)^{-1} x\|_1 \quad (\forall x \in \mathbb{C}^k),$$

wobei $\|y\|_1 := \sum_{j=1}^k |y_j|$ für jedes $y = (y_1, \ldots, y_k) \in \mathbb{C}^k$. Dann ergibt sich

$$\begin{aligned} \|Bx\|_* &= \|(TD_\varepsilon)^{-1} B (TD_\varepsilon)(TD_\varepsilon)^{-1} x\|_1 = \|J_\varepsilon (TD_\varepsilon)^{-1} x\|_1 \leq \|J_\varepsilon\|_1 \|x\|_* \\ \rightsquigarrow \|B\|_* &\leq \|J_\varepsilon\|_1 \leq \max\{|\lambda_1|, |\ldots, |\lambda_r|, |\lambda_j| + \varepsilon : j = r+1, \ldots, k\} \leq 1. \end{aligned}$$

Die Matrix $B \otimes I$ hat die Jordan-Normalform $J \otimes I$ und die Norm $\|\cdot\|_*$ auf $\mathbb{C}^{km}$ wird definiert durch

$$\|x\|_* := \|((TD_\varepsilon)^{-1} \otimes I) x\|_1 \quad (\forall x \in \mathbb{C}^{km}).$$

Man erhält analog $\|B \otimes I\|_* \leq 1$. □

Satz 3.36 *(Stabilität)*
Es existieren reelle Funktionen a_j, $j = 1, \ldots, k$, und b_j, $j = 0, \ldots, k$, die auf einer Umgebung U von $e = (1, \ldots, 1) \in \mathbb{R}^k$ definiert und stetig in e sind, so dass für jedes Gitter G die Darstellungen

$$a_{\ell j} = a_j(\kappa_\ell, \ldots, \kappa_{\ell-k+1}),\ b_{\ell j} = b_j(\kappa_\ell, \ldots, \kappa_{\ell-k+1}),\ \kappa_\ell = \frac{h_\ell}{h_{\ell-1}},\ \ell = 2, \ldots, N,$$

gültig sind und $\sum_{j=0}^k a_{\ell j} = 0$ gilt. Das charakteristische Polynom

$$p(\lambda) = \lambda^k + \sum_{j=1}^k a_j(e) \lambda^{k-j}$$

erfülle die starke Wurzelbedingung nach Dahlquist und es existiere eine Konstante $L > 0$, so dass

$$\|f(x,t) - f(\tilde{x},t)\| \leq L\|x - \tilde{x}\| \quad (\forall x, \tilde{x} \in \mathbb{R}^m, \forall t \in [t_0, T]).$$

Die Startphase des k-schrittigen linearen Mehrschrittverfahrens

$$(*) \quad \sum_{j=0}^{k} a_{\ell j} x_{\ell-j} = h_\ell \sum_{j=1}^{k} b_{\ell j} f(x_{\ell-j}, t_{\ell-j}),\ \ell = k, \ldots, N,$$

zur Berechnung von $x_1, \ldots, x_{k-1}$ werde mit einem stabilen Verfahren durchgeführt. Dann existieren Konstanten $H > 0$ und $C_1, C_2 > 0$, so dass das lineare Mehrschrittverfahren (*) stabil auf $\mathcal{G} = \mathcal{G}(C_1, C_2, C_3)$ mit $h(G) \leq H$ und beliebigem $C_3 > 0$ ist. Dabei werden die Konstanten $C_1, C_2 > 0$ so gewählt, dass alle Polynome

$$p(\lambda) = \lambda^k + \sum_{j=1}^{k} a_j(\kappa_1, \ldots, \kappa_k)\lambda^{k-j}$$

mit $C_1 \leq \kappa_j \leq C_2$, $j = 1, \ldots, k$, die Wurzelbedingung nach Dahlquist erfüllen und $\sum_{j=0}^{k} |b_{\ell j}|$ gleichmäßig beschränkt ist auf $\mathcal{G}$.

Beweis: Das Polynom p erfüllt nach Voraussetzung die starke Wurzelbedingung nach Dahlquist. Wegen der stetigen Abhängigkeit der Nullstellen eines Polynoms von dessen Koeffizienten existieren Konstanten $C_1 \leq 1$ und $C_2 \geq 1$ so, dass die charakteristischen Polynome

$$p_\ell(\lambda) = \lambda^k + \sum_{j=1}^{k} a_{\ell j}\lambda^{k-j}$$

mit $a_{\ell j} = a_j(\kappa_\ell, \ldots, \kappa_{\ell-k+1})$, $j = 1, \ldots, k$, die Wurzelbedingung nach Dahlquist erfüllen, falls

$$C_1 \leq \kappa_{\ell-i} = \frac{h_{\ell-i}}{h_{\ell-i-1}} \leq C_2 \quad (i = 0, \ldots, k-1).$$

Die $k \times k$-Matrix

$$A_\ell = \begin{pmatrix} -a_{\ell 1} & -a_{\ell 2} & \cdots & \cdots & -a_{\ell k} \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

besitzt das charakteristische Polynom p_ℓ. Nach Lemma 3.35 existiert eine Norm $\|\cdot\|_*$ auf $\mathbb{C}^{km}$ mit $\|A_\ell \otimes I\|_* \leq 1$ (mit der $m \times m$ Einheitsmatrix I) für alle $\ell = k, \ldots, N$. Das k schrittige lineare Mehrschrittverfahren kann mit Hilfe des Operators A_G angewandt auf $x_G = (x_0, x_1, \ldots, x_N) \in \mathbb{R}^{m(N+1)}$ formuliert werden

$$(*) \quad [A_G x_G]_\ell = \frac{1}{h_\ell} \sum_{j=0}^{k} a_{\ell j} x_{\ell-j} - \sum_{j=0}^{k} b_{\ell j} f(x_{\ell-j}, t_{\ell-j}) \quad (\ell = k, \ldots, N).$$

Die letztere Identität läßt sich mit den Bezeichnungen $X_\ell = (x_\ell, \dots, x_{\ell-k+1}) \in \mathbb{R}^{km}$, $d_\ell = [A_G x_G]_\ell$ und $\Phi_\ell(X_\ell, X_{\ell-1}) = \sum_{j=0}^{k} b_{\ell j} f(x_{\ell-j}, t_{\ell-j})$ in Form eines "gestörten" Einschrittverfahrens

$$X_\ell = (A_\ell \otimes I)X_{\ell-1} + h_\ell(e_1 \otimes I)\Phi_\ell(X_\ell, X_{\ell-1}) + h_\ell(e_1 \otimes I)d_\ell \quad (\ell = k, \dots, N)$$

schreiben, wobei e_1 der erste kanonische Einheitsvektor im $\mathbb{R}^k$ ist. Diese entsteht aus (*) indem die $k-1$ Identitäten $x_{\ell-j} = x_{\ell-j}$, $j = 1, \dots, k-1$, hinzugefügt werden. Ist nun $\tilde{x}_G = (\tilde{x}_0, \dots, \tilde{x}_N)$ ein weiters Element aus $\mathbb{R}^{m(N+1)}$ und $\tilde{X}_\ell := (\tilde{x}_\ell, \dots, \tilde{x}_{\ell-k+1})$ sowie $\tilde{d}_\ell = [A_G \tilde{x}_G]_\ell$, so entsteht als zweite Gleichung

$$\tilde{X}_\ell = (A_\ell \otimes I)\tilde{X}_{\ell-1} + h_\ell(e_1 \otimes I)\Phi_\ell(\tilde{X}_\ell, \tilde{X}_{\ell-1}) + h_\ell(e_1 \otimes I)\tilde{d}_\ell \quad (\ell = k, \dots, N).$$

Für die Differenz $Y_\ell := X_\ell - \tilde{X}_\ell$ folgt dann mit $D_\ell := d_\ell - \tilde{d}_\ell$ und geeigneten Normäquivalenzkonstanten K_* und $\bar{K}_*$:

$$\begin{aligned}
Y_\ell &= (A_\ell \otimes I)Y_{\ell-1} + h_\ell(e_1 \otimes I)(\Phi_\ell(X_\ell, X_{\ell-1}) - \Phi_\ell(\tilde{X}_\ell, \tilde{X}_{\ell-1})) + h_\ell(e_1 \otimes I)D_\ell \\
\|Y_\ell\|_* &\le \|Y_{\ell-1}\|_* + K_* h_\ell(\|\Phi_\ell(X_\ell, X_{\ell-1}) - \Phi_\ell(\tilde{X}_\ell, \tilde{X}_{\ell-1})\| + \|D_\ell\|) \\
&\le \|Y_{\ell-1}\|_* + K_* L h_\ell \sum_{j=0}^{k} |b_{\ell j}| \|x_{\ell-j} - \tilde{x}_{\ell-j}\| + K_* h_\ell \|D_\ell\| \\
&\le \|Y_{\ell-1}\|_* + K_* K L h_\ell(\|Y_\ell\| + \|Y_{\ell-1}\|) + K_* h_\ell \|D_\ell\| \\
&\le \|Y_{\ell-1}\|_* + K_* K \bar{K}_* L h(\|Y_\ell\|_* + \|Y_{\ell-1}\|_*) + K_* h \|D_\ell\|
\end{aligned}$$

Dabei ist h die maximale Schrittweite des Gitters und K so, dass $\sum_{j=0}^{k} |b_{\ell j}| \le K$. Letzteres ist wegen der vorausgesetzten Stetigkeit der b_j, $j = a, \dots, k$, in e möglich. Es sei nun $C := K_* K \bar{K}_* L$ und $H > 0$ so gewählt, dass $CH < 1$. Dann gilt für $h \le H$ die Abschätzung

$$\|Y_\ell\|_* \le \frac{1 + Ch}{1 - Ch}\|Y_{\ell-1}\|_* + \frac{K_* h}{1 - Ch}\|d_\ell - \tilde{d}_\ell\| \quad (\ell = k, \dots, N).$$

Daraus folgt rekursiv

$$\begin{aligned}
\|Y_\ell\|_* &\le \left(\frac{1 + Ch}{1 - Ch}\right)^{N-k} \left[\|Y_{k-1}\|_* + \frac{K_* h(N-k)}{1 - Ch} \max_{\ell=k,\dots,N} \|d_\ell - \tilde{d}_\ell\|\right] \\
&\le \exp\left(\frac{2CC_3}{1 - CH}\right) \left[\|Y_{k-1}\|_* + \frac{K_* C_3}{1 - CH} \max_{\ell=k,\dots,N} \|d_\ell - \tilde{d}_\ell\|\right] \ (\ell = k. \dots, N).
\end{aligned}$$

Hierbei wurde die folgende Ungleichung verwendet:

$$\begin{aligned}
\left(\frac{1 + Ch}{1 - Ch}\right)^{N-k} &= \left(1 + \frac{2hC}{1 - hC}\right)^{N-k} = \left[\left(1 + \frac{2hC}{1 - hC}\right)^{\frac{1-Ch}{2Ch}}\right]^{\frac{2Ch}{1-Ch}(N-k)} \\
&\le \exp\left(\frac{2Ch(N-k)}{1 - Ch}\right) \le \exp\left(\frac{2CC_3}{1 - CH}\right)
\end{aligned}$$

Folglich existiert eine Konstante $\hat{S} > 0$, so dass

$$\|Y_\ell\|_* \le \hat{S}(\|Y_{k-1}\|_* + \max_{\ell=k,\dots,N} \|d_\ell - \tilde{d}_\ell\|) \quad (\ell = k, \dots, N)$$

und bei Übergang zur Maximum-Norm entsteht mit einer durch den Normübergang modifizierten Konstante $S > 0$ die endgültige Abschätzung

$$\max_{\ell=k,\dots,N} \|x_\ell - \tilde{x}_\ell\| \le S \left(\max_{\ell=0,\dots,k-1} \|x_\ell - \tilde{x}_\ell\| + \max_{\ell=k,\dots,N} \|[A_G x_G]_\ell - [A_G \tilde{x}_G]_\ell\| \right)$$

Da nach Voraussetzung auch die Startphase des linearen Mehrschrittverfahrens stabil ist, folgt die Aussage. □

Beispiel *(Fortsetzung von 3.32)*
Für die k-schrittigen Adams-Verfahren hat das charakteristische Polynom die Gestalt

$$p(\lambda) = \lambda^k - \lambda^{k-1} = \lambda^{k-1}(\lambda - 1)\,.$$

Es hat also die einfache Nullstelle $\lambda = 1$ und die $(k-1)$-fache Nullstelle $\lambda = 0$. Also sind die Adams-Verfahren nach Satz 3.36 stabil und damit konvergent mit Ordnung $s = k + 1 - a$ (Satz 3.33) auf jeder Klasse zulässiger Gitter.

Bemerkung 3.37 *Der Beweis von Satz 3.36 zeigt auch, dass ein lineares Mehrschrittverfahren auf der Klasse der äquidistanten Gitter $\mathcal{G}_0$ stabil ist, falls sein charakteristisches Polynom die Wurzelbedingung nach Dahlquist erfüllt.*

Satz 3.38 *(erste Dahlquist-Barriere)*
Für die Konsistenzordnung s eines auf $\mathcal{G}_0$ konsistenten und stabilen k-schrittigen linearen Mehrschrittverfahrens

$$\sum_{j=0}^{k} a_j x_{\ell-j} = h \sum_{j=0}^{k} b_j f(x_{\ell-j}, t_{\ell-j}), \quad (\ell = k, \dots, N)$$

sind die folgenden Bedingungen gültig:

(a) $s \le k + 2$, *falls k gerade ist;* $s = k + 2$ *impliziert* $a_j = -a_{k-j}$, $b_j = b_{k-j}$, $j = 0, \dots, k$.

(b) $s \le k + 1$, *falls k ungerade ist,*

(c) $s \le k$, *falls* $b_0 \le 0$ *(insbesondere, falls das Verfahren explizit ist).*

Beweis: vgl. Hairer-Nørsett-Wanner 1993, Chapt. III.3.
Beweisidee: Man betrachtet die beiden Polynome

$$\rho(z) = \sum_{j=0}^{k} a_j z^{k-j} \quad \text{und} \quad \sigma(z) = \sum_{j=0}^{k} b_j z^{k-j}.$$

Nach Voraussetzung erfüllt ρ die Wurzelbedingung nach Dahlquist und es gilt $\rho(1) = 0$. Man führt die Transformation

$$z = \frac{\zeta + 1}{\zeta - 1} \quad \text{bzw.} \quad \zeta = \frac{z + 1}{z - 1}$$

durch. Dabei wird $\{z \in \mathbb{C} : |z| \leq 1\}$ auf $\mathbb{C}_- = \{\zeta \in \mathbb{C} : \mathrm{Re}(\zeta) \leq 0\}$, 1 auf ∞, -1 auf 0 und $\{z \in \mathbb{C} : |z| < 1\}$ auf $\{\zeta \in \mathbb{C} : \mathrm{Re}(\zeta) < 0\}$ abgebildet. Wir betrachten die transformierten Polynome

$$R(\zeta) = \left(\frac{\zeta - 1}{2}\right)^k \rho\left(\frac{\zeta + 1}{\zeta - 1}\right) \quad \text{und} \quad S(\zeta) = \left(\frac{\zeta - 1}{2}\right)^k \sigma\left(\frac{\zeta + 1}{\zeta - 1}\right),$$

wobei die Nullstellen von R in $\mathbb{C}_-$ liegen und auf der imaginären Achse nur einfache Nullstellen von R sind. Daraus schlussfolgert man, dass R vom Grad $k-1$ ist und alle Koeffizienten von R dasselbe Vorzeichen besitzen.
Die Bedingungen für Konsistenzordnung s sind nach Bemerkung 3.31

$$\sum_{j=0}^{k} \left(a_j j^i - i j^{i-1} b_j\right) = 0 \quad (i = 0, \ldots, s).$$

Diese sind äquivalent zu (Hairer-Nørsett-Wanner 1993, Chapt. III.2)

$$\frac{\rho(z)}{\log(z)} - \sigma(z) = O((1-z)^s) \ \text{ für } \ z \to 1\,.$$

und deshalb äquivalent zu

$$R(\zeta)\left(\log\frac{\zeta+1}{\zeta-1}\right)^{-1} - S(\zeta) = O(\zeta^{k-s}) \ \text{ für } \ \zeta \to \infty\,.$$

Man betrachtet die Laurent Reihe von $\log\frac{\zeta+1}{\zeta-1}$ und zeigt, dass die letztere äquivalente Bedingung in der Tat gültig ist. □

Beispiel 3.39 *(rückwärtige Differentiationsformeln, BDF)*
Ausgangspunkt: $x'(t_\ell) = f(x(t_\ell), t_\ell)$, $\ell = 1, \ldots, N$.

Die Funktion $x(\cdot)$ wird in den Stützstellen $t_{\ell-k}, \ldots, t_\ell$ mit den Werten $x_{\ell-k}, \ldots, x_\ell$ interpoliert durch das Polynom

$$P_k(t) = \sum_{j=0}^{k} \prod_{\substack{i=0 \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} x_{\ell-j}.$$

Ansatz: $P_k'(t_\ell) = f(x_\ell, t_\ell)$, $\ell = k, \ldots, N$, *wobei*

$$\begin{aligned} P_k'(t) &= \sum_{j=0}^{k} \frac{\mathrm{d}}{\mathrm{dt}} \left[\prod_{\substack{i=0 \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}}\right] x_{\ell-j} \\ &= \sum_{j=0}^{k} \left[\sum_{\substack{r=0 \\ r\neq j}}^{k} \frac{1}{t_{\ell-j} - t_{\ell-r}} \prod_{\substack{i=0 \\ i\neq r \\ i\neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}}\right] x_{\ell-j} \\ \rightsquigarrow P_k'(t_\ell) &= \left[\sum_{r=1}^{k} \frac{1}{t_\ell - t_{\ell-r}}\right] x_\ell + \sum_{j=1}^{k} \left[\frac{1}{t_{\ell-j} - t_\ell} \prod_{\substack{i=1 \\ i\neq j}}^{k} \frac{t_\ell - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}}\right] x_{\ell-j} \end{aligned}$$

Rückwärtige Differentiationsformeln der Ordnung k:

$$x_\ell + \sum_{j=1}^{k} \underbrace{\left[\left[\sum_{r=1}^{k} \frac{1}{t_\ell - t_{\ell-r}}\right]^{-1} \frac{1}{t_{\ell-j} - t_\ell} \prod_{\substack{i=1 \\ i\neq j}}^{k} \frac{t_\ell - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}}\right]}_{a_{\ell j}} x_{\ell-j} = h_\ell \underbrace{\left[\sum_{r=1}^{k} \frac{h_\ell}{t_\ell - t_{\ell-r}}\right]^{-1}}_{b_{\ell 0}} f(x_\ell, t_\ell)$$

Wegen $\sum_{r=1}^{k} \frac{h_\ell}{t_\ell - t_{\ell-r}} \geq \frac{h_\ell}{t_\ell - t_{\ell-1}} = 1$ *gilt immer* $0 \leq b_{\ell 0} \leq 1$.

Beispiele:

$k = 1$: $x_\ell - x_{\ell-1} = h_\ell f(x_\ell, t_\ell), \quad \ell = 1, \ldots, N$ *(impliziter Euler bzw. BDF1).*

$k = 2$: $x_\ell + a_{\ell 1} x_{\ell-1} + a_{\ell 2} x_{\ell-2} = h_\ell b_{\ell 0} f(x_\ell, t_\ell), \ \ell = 2, \ldots, N$ *(BDF2), wobei*

$$\begin{aligned} b_{\ell 0} &= \left(1 + \frac{h_\ell}{h_\ell + h_{\ell-1}}\right)^{-1} = \frac{h_\ell + h_{\ell-1}}{2h_\ell + h_{\ell-1}} = \frac{1+\kappa_\ell}{1+2\kappa_\ell} \\ a_{\ell 1} &= -\frac{h_\ell(h_\ell + h_{\ell-1})}{2h_\ell + h_{\ell-1}} \frac{h_\ell + h_{\ell-1}}{h_\ell h_{\ell-1}} = -\frac{(h_\ell + h_{\ell-1})^2}{(2h_\ell + h_{\ell-1})h_{\ell-1}} = -\frac{(1+\kappa_\ell)^2}{1+2\kappa_\ell} \\ a_{\ell 2} &= \frac{h_\ell(h_\ell + h_{\ell-1})}{2h_\ell + h_{\ell-1}} \frac{h_\ell}{(h_\ell + h_{\ell-1})h_{\ell-1}} = \frac{h_\ell^2}{(2h_\ell + h_{\ell-1})h_{\ell-1}} = \frac{\kappa_\ell^2}{1+2\kappa_\ell} \\ &\quad \textit{mit} \ \ \kappa_\ell := \frac{h_\ell}{h_{\ell-1}} \,. \end{aligned}$$

Satz 3.40
Für $1 \leq k \leq 6$ *erfüllen die BDF-Verfahren die starke Wurzelbedingung nach Dahlquiststst auf* $\mathcal{G}_0$ *und besitzen die Konsistenzordnung* $s = k$. *Für* $k > 6$ *sind sie nicht mehr stabil auf* $\mathcal{G}_0$. *Das zweischrittige BDF-Verfahren ist stabil auf jeder Klasse zulässiger Gitter mit* $0 \leq \frac{h_\ell}{h_{\ell-1}} \leq C_2 < 1 + \sqrt{2}$, $\ell = 2, \ldots, N$.

Beweis: Wir beweisen nur den zweiten Teil der Aussage und verweisen für den ersten Teil auf die Literatur. Das charakteristische Polynom des zweischrittigen BDF-Verfahrens hat in Schritt ℓ die Form

$$p_\ell(\lambda) = \lambda^2 - \frac{(1+\kappa_\ell)^2}{1+2\kappa_\ell}\lambda + \frac{\kappa_\ell^2}{1+2\kappa_\ell} \,.$$

Wegen $1 - \frac{(1+\kappa_\ell)^2}{1+2\kappa_\ell} + \frac{\kappa_\ell^2}{1+2\kappa_\ell} = 0$ ist $\lambda_1 = 1$ eine Nullstelle von p_ℓ. Damit gilt

$$p_\ell(\lambda) = (\lambda - 1)\left(\lambda - \frac{\kappa_\ell^2}{1+2\kappa_\ell}\right) \,.$$

Die Wurzelbedingung nach Dahlquist ist deshalb erfüllt, falls

$$\left|\frac{\kappa_\ell^2}{1+2\kappa_\ell}\right| \leq 1 \ \Leftrightarrow \ \kappa_\ell^2 \leq 2\kappa_\ell + 1 \ \Leftrightarrow \ \kappa_\ell \leq 1 + \sqrt{2} \,.$$

Die Aussage folgt deshalb aus Satz 3.36. □

Bemerkung 3.41 *Stabilitätsresultate für die BDF-Verfahren auf variablen Gittern wurden von März 80 und Grigorieff 83 bewiesen. Die folgende Charakterisierung von Klassen zulässiger Gitter, auf denen die BDF der Ordnung k stabil sind, stammt von Grigorieff 83:*

k	2	3	4	5
C_1	0	0.836	0.979	0.997
C_2	$1+\sqrt{2}$	1.127	1.019	1.003

Bei den BDF-Verfahren der Ordnung $k > 3$ ist es also sehr schwer Schrittweiten zügig zu verkleinern bzw. zu vergrößern. Deshalb werden sie oft nur bis $k = 3$ implementiert und es findet vor allem die BDF2 Anwendung (z.B. beim Design elektronischer Bauelemente).

Bemerkung 3.42 *(Realisierung impliziter Verfahren)*
Für implizite Verfahren ist in jedem Schritt ℓ das nichtlineare Gleichungssystem

$$x_\ell = h_\ell b_{\ell 0} f(x_\ell, t_\ell) - \sum_{j=1}^{k} [a_{\ell j} x_{\ell-j} - h_\ell b_{\ell j} f(x_{\ell-j}, t_{\ell-j})]$$

zur Berechnung von x_ℓ zu lösen. Ist die Lipschitzkonstante L von f (bzgl. x) nicht zu groß, so kann dies mit der Iteration nach dem Banachschen Fixpunktsatz erfolgen. Die Kontraktionskonstante ist gerade $b_{\ell 0} h_\ell L$ und ist deshalb kleiner 1, wenn h_ℓ nicht zu groß ist. Dabei kann als Startwert $x_{\ell-1}$ oder ein besserer Startwert mit einem expliziten Verfahren berechnet werden. Es entstehen dann sog. Prädiktor-Korrektor-Verfahren *(z.B. im Fall der k-schrittigen expliziten und impliziten Adams-Verfahren).*
Sind die Lipschitz-Konstanten aber groß, so würde die Bedingung $b_{\ell 0} h_\ell L < 1$ eine zu große Einschränkung an die Schrittweite darstellen und das nichtlineare Gleichungssystem muss mit dem Newton-Verfahren mit der Jacobi-Matrix

$$I - h_\ell b_{\ell 0} \frac{\partial f}{\partial x}(x_\ell, t_\ell)$$

iterativ gelöst werden (vgl. Kapitel 3.7). Als Bedingung an die Schrittweite h_ℓ ist dann nur noch die Konvergenz des Newton-Verfahrens zu fordern.

Bemerkung 3.43 *(Schrittweiten-Steuerung)*
Variable Gitter werden i.a. angestrebt, um möglichst wenige Schritte (d.h. ein kleines $N(G)$) bei hinreichender Fehlerkontrolle zu machen, um den Gesamtaufwand des Verfahrens gering zu halten. Zur Fehlerkontrolle werden z.B. (i) der lokale Diskretisierungsfehler $\tau_\ell(G)$ oder (ii) der Fehler $x_\ell - \bar{x}_\ell$ zweier Näherungen von $x_*(t_\ell)$ verwendet.
(i) Ausgangspunkt ist eine Darstellung von $\tau_\ell(G)$ der Form

$$\tau_\ell(G) = C_s x_*^{(s+1)}(t_\ell) h_\ell^s + O(h^{s+1}) \,.$$

Man versucht dann den Term $C_\ell = C_s x_*^{(s+1)}(t_\ell)$ zu schätzen, in dem man $\tau_\ell(G)$ für zwei verschiedene Schrittweiten h_ℓ und $\tilde{h}_\ell$ berechnet und den Term $O(h^{s+1})$ vernachlässigt. Danach wird h_ℓ aus einer Bedingung der Form $\|C_\ell\| h_\ell^s \leq$ tol berechnet.

(ii) Man berechnet x_ℓ mit Schrittweite h_ℓ und $\bar{x}_\ell$ mit einer kleineren Schrittweite αh_ℓ mit $\alpha < 1$ oder mit einem Verfahren höherer Ordnung (z.B. bei den Adams-Verfahren) und versucht asymptotische bzw. andere Darstellungen für $x_\ell - \bar{x}_\ell$ zu finden, die es erlauben, mit einer vorgegebenen Toleranz tol *für $\|x_\ell - \bar{x}_\ell\|$ eine akzeptable Schrittweite h_ℓ zu bestimmen.*
In beiden Fällen versucht man drastische *Änderungen der Schrittweite zu vermeiden und die aktuelle Schrittweite h_ℓ so zu bestimmen, dass $h_{\ell+1} = h_\ell$ wahrscheinlich ist.*

Literatur:
E. Hairer, S.P. Nørsett, G. Wanner: Solving Ordinary Differential Equations I, Springer, Berlin 1993.

3.6 Asymptotisches Verhalten von Integrationsverfahren

Wir betrachten die Differentialgleichung

$$(*) \quad x'(t) = f(x(t), t), \quad t \in [t_0, +\infty),$$

und setzen voraus, daß sie (schwach) kontraktiv ist, d. h. es existiert ein Skalarprodukt $\langle \cdot, \cdot \rangle$ auf $\mathbb{R}^m$ und eine Konstante $\gamma \leq 0$, so dass $\forall t \in [t_0, +\infty)$:

$$\langle f(x,t) - f(\tilde{x}, t), x - \tilde{x} \rangle \leq \gamma \|x - \tilde{x}\|^2 \quad \text{(vgl. Definition 1.13)}$$

Nach Satz 1.11 gilt für zwei Lösungen:

$$\|x(t) - \tilde{x}(t)\| \leq \exp(\gamma(t - t_0)) \|x(t_0) - \tilde{x}(t_0)\|$$

Dies motiviert die folgende Definition für ein Integrationsverfahren:

Definition 3.44
Ein Integrationsverfahren heißt B-stabil (bzgl. $\| \cdot \| = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$), falls es bei Anwendung auf eine (bzgl. $\langle \cdot, \cdot \rangle$) schwach kontraktive Differentialgleichung $(*)$ mit Anfangswerten x_0, $\tilde{x}_0$ und beliebigen Gittern $G = \{t_0 < t_1 < t_2 < \ldots\}$ Folgen $(x_\ell)_{\ell \in \mathbb{N}_0}$ und $(\tilde{x}_\ell)_{\ell \in \mathbb{N}_0}$ erzeugt, für die gilt:

$$\|x_\ell - \tilde{x}_\ell\| \leq \|x_{\ell-1} - \tilde{x}_{\ell-1}\|, \quad \forall \ell \in \mathbb{N}.$$

Beispiel 3.45
Wir betrachten $x'(t) = -kx(t), \quad t \in [t_0, +\infty)$ $(k > 0)$

a) explizites Euler-Verfahren: $x_\ell = x_{\ell-1} + h_\ell f(x_{\ell-1}, t_{\ell-1})$

$$\begin{aligned}
&\rightsquigarrow \quad x_\ell = x_{\ell-1} + h_\ell(-kx_{\ell-1}) = (1 - h_\ell k) x_{\ell-1} \\
&\qquad \bar{x}_\ell = (1 - h_\ell k) \bar{x}_{\ell-1} \\
&\rightsquigarrow \quad |x_\ell - \bar{x}_\ell| \leq |1 - h_\ell k| \; |x_{\ell-1} - \bar{x}_{\ell-1}| \\
&\rightsquigarrow \quad \textit{das Verfahren ist B-stabil, falls} |1 - h_\ell k| \leq 1 \textit{ oder } h_\ell \leq \tfrac{2}{k} \\
&\rightsquigarrow \quad \textit{Bedingung an die Schrittweiten, d. h.} \\
&\qquad \textit{die Abschätzung gilt nicht für beliebige Gitter.}
\end{aligned}$$

b) impliziertes Euler-Verfahren: $x_\ell = x_{\ell-1} + h_\ell f(x_\ell, t_\ell)$

$$\rightsquigarrow \quad (1+h_\ell k)x_\ell = x_{\ell-1} \rightsquigarrow \quad \begin{array}{l} x_\ell = \frac{1}{1+h_\ell k} x_{\ell-1} \\ \bar{x}_\ell = \frac{1}{1+h_\ell k}\bar{x}_{\ell-1} \end{array}$$

$$\rightsquigarrow \quad |x_\ell - \bar{x}_\ell| \le \tfrac{1}{1+h_\ell k}|x_{\ell-1} - \bar{x}_{\ell-1}| \le |x_{\ell-1} - \bar{x}_{\ell-1}|, \forall \ell \in \mathbb{N}$$

$\rightsquigarrow$ *evtl. B-stabil*

Definition 3.46
Ein p-stufiges Runge-Kutta-Verfahren heißt algebraisch stabil, falls $\gamma_i \ge 0$, $i = 1, \dots, p$, *und die (symmetrische) Matrix* $M = (m_{ij}) \in \mathbb{R}^{p\times p}$, $m_{ij} := \beta_{ij}\gamma_i + \beta_{ji}\gamma_j - \gamma_i\gamma_j$, $i, j = 1, \dots, p$, *positiv semidefinit ist, d.h.*

$$\langle Mx, x\rangle \ge 0, \quad \forall x \in \mathbb{R}^p.$$

(Es gilt $M = B^T\ diag\ (\gamma_1, \dots, \gamma_p) + diag\ (\gamma_1, \dots, \gamma_p)B - \gamma\gamma^T$*).*

Satz 3.47 *(Crouzeix 79, Burrage-Butcher 79)*
Die Differentialgleichung (*) sei schwach kontraktiv (bzgl. $\langle\cdot,\cdot\rangle$*). Dann ist jedes algebraisch stabile Runge-Kutta-Verfahren B-stabil (bzgl.* $\|\cdot\| = \langle\cdot,\cdot\rangle^{\frac{1}{2}}$*).*

Beweis: Wir betrachten ein p-stufiges Runge-Kutta-Verfahren

$$\begin{aligned} x_\ell &= x_{\ell-1} + h_\ell \sum_{j=1}^p \gamma_j F_{\ell j}, \quad F_{\ell j} := f(\bar{x}_\ell^j, t_{\ell-1} + \alpha_j h_\ell), \\ (+) \quad \bar{x}_\ell^i &= x_{\ell-1} + h_\ell \sum_{j=1}^p \beta_{ij} F_{\ell j}, \quad i = 1, \dots, p \end{aligned}$$

Ist die Differentialgleichung schwach kontraktiv, so folgt mit $\tilde{F}_{\ell j} = f(\tilde{\bar{x}}_\ell^j, t_{\ell-1} + \alpha_j h_\ell)$

$$\langle F_{\ell i} - \tilde{F}_{\ell i}, \bar{x}_\ell^i - \tilde{\bar{x}}_\ell^i\rangle \le 0, \quad i = 1, \dots, p, \ell \in \mathbb{N}_0.$$

Zu zeigen: $\|x_\ell - \tilde{x}_\ell\|^2 \le \|x_{\ell-1} - \tilde{x}_{\ell-1}\|^2, \quad \forall \ell \in \mathbb{N}.$

$$\begin{aligned} x_\ell - \tilde{x}_\ell &= x_{\ell-1} - \tilde{x}_{\ell-1} + h_\ell \sum_{j=1}^p \gamma_j (F_{\ell j} - \tilde{F}_{\ell j}) \\ \rightsquigarrow \|x_\ell - \tilde{x}_\ell\|^2 &= \|x_{\ell-1} - \tilde{x}_{\ell-1}\|^2 + k_\ell, \quad \text{wobei} \\ k_\ell &:= 2h_\ell \sum_{j=1}^p \gamma_j \langle F_{\ell j} - \tilde{F}_{\ell j}, x_{\ell-1} - \tilde{x}_{\ell-1}\rangle + h_\ell^2 \sum_{i,j=1}^p \gamma_i\gamma_j \langle F_{\ell j} - \tilde{F}_{\ell j}, F_{\ell i} - \tilde{F}_{\ell i}\rangle \end{aligned}$$

Zu zeigen: $k_\ell \le 0$

Aus (+) folgt: $x_{\ell-1} - \tilde{x}_{\ell-1} = \bar{x}_\ell^j - \tilde{\bar{x}}_\ell^j - h_\ell \sum_{i=1}^p \beta_{ji} U_i$ mit $U_i := F_{\ell i} - \tilde{F}_{\ell i}$ und

$$2h_\ell \sum_{j=1}^p \gamma_j \langle F_{\ell j} - \tilde{F}_{\ell j}, x_{\ell-1} - \tilde{x}_{\ell-1}\rangle = 2h_\ell \sum_{j=1}^p \gamma_j \underbrace{\langle F_{\ell j} - \tilde{F}_{\ell j}, \bar{x}_\ell^j - \tilde{\bar{x}}_\ell^j\rangle}_{\le 0} - 2h_\ell^2 \sum_{i,j=1}^p \gamma_j\beta_{ji}\langle U_j, U_i\rangle.$$

$$\rightsquigarrow k_\ell \le -h_\ell^2 \sum_{i,j=1}^{p} (-\gamma_i\gamma_j + 2\gamma_j\beta_{ji})\langle U_j, U_i\rangle$$

Es gilt: $\sum\limits_{i,j=1}^{p} 2\gamma_j\beta_{ji} = \sum\limits_{i,j=1}^{p} (\gamma_i\beta_{ij} + \gamma_j\beta_{ji})$

$$\rightsquigarrow k_\ell \le -h_\ell^2 \sum_{i,j=1}^{p} m_{ij}\langle U_j, U_i\rangle, \quad U_j := F_{\ell j} - \tilde{F}_{\ell j} \quad (j = 1, \ldots, p)$$

Zu zeigen: $\sum\limits_{i,j=1}^{p} m_{ij}\langle U_i, U_j\rangle \ge 0, \quad \forall U_i, i = 1, \ldots, p.$
Es sei $M = R^T R$ die Cholesky-Zerlegung der positiv semidefiniten Matrix M mit einer oberen Dreiecksmatrix $R = (r_{ij})$. Dann folgt

$$\begin{aligned}\sum_{i,j=1}^{p} m_{ij}\langle U_i, U_j\rangle &= \sum_{i,j=1}^{p}\sum_{k=1}^{p} r_{ki}r_{kj}\langle U_i, U_j\rangle \\ &= \sum_{k=1}^{p}\left\langle \sum_{i=1}^{p} r_{ki}U_i, \sum_{j=1}^{p} r_{kj}U_j\right\rangle = \sum_{k=1}^{p}\left\| \sum_{i=1}^{p} r_{ki}U_i\right\|^2 \ge 0\,.\end{aligned}$$

□

Satz 3.48
Die impliziten Runge-Kutta-Verfahren vom Typ Gauß und Radau IIa sind algebraisch stabil und damit B-stabil.

Beweis:
Nach Satz 3.24 sind für beide Verfahrenstypen die γ_i, $i = 1, \ldots, p$ stets positiv. Wir betrachten die Matrix

$$\begin{aligned} M &= (m_{ij}), \quad m_{ij} = \gamma_i\beta_{ij} + \gamma_j\beta_{ji} - \gamma_i\gamma_j, \quad i, j = 1, \ldots, p, \text{ und} \\ Q &= A_p M A_p^\top, \quad A_p = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_p \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_1^{p-1} & \alpha_2^{p-1} & \cdots & \alpha_p^{p-1} \end{pmatrix}. \end{aligned}$$

Da die Knoten $\alpha_1, \ldots, \alpha_p$ in beiden Fällen sämtlich verschieden sind, gilt:
M ist positiv semidefinit gdw. Q positiv semidefinit.
Für $Q = (q_{\ell k})_{\ell,k=1,\ldots,p}$ und $\ell, k = 1, \ldots, p$ gilt:

$$q_{\ell k} = \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_j^{\ell-1} m_{ji}\alpha_i^{k-1} = \sum_{i=1}^{p}\sum_{j=1}^{p} (\gamma_i\beta_{ij} + \gamma_j\beta_{ji} - \gamma_i\gamma_j)\alpha_j^{\ell-1}\alpha_i^{k-1}\,.$$

1. Gauß-Verfahren: $B(2p)$ und $C(p)$ sind erfüllt, d.h.

$$\begin{aligned} \sum_{j=1}^{p} \gamma_j\alpha_j^{k-1} &= \tfrac{1}{k}, \quad k = 1, \ldots, 2p \\ \sum_{j=1}^{p} \beta_{ij}\alpha_j^{k-1} &= \tfrac{1}{k}\alpha_i^k, \quad k = 1, \ldots, p, \quad i = 1, \ldots, p \end{aligned}$$

$$\rightsquigarrow \sum_{i,j=1}^{p} \gamma_i\gamma_j\alpha_j^{\ell-1}\alpha_i^{k-1} = \tfrac{1}{k\ell}, \quad k,\ell = 1,\dots,p$$

$$\begin{aligned}\sum_{i,j=1}^{p} \gamma_i\beta_{ij}\alpha_j^{\ell-1}\alpha_i^{k-1} &= \sum_{i=1}^{p} \gamma_i\alpha_i^{k-1}\left(\sum_{j=1}^{p}\beta_{ij}\alpha_j^{\ell-1}\right)\\ &= \sum_{i=1}^{p}\gamma_i\alpha_i^{k-1}\cdot\tfrac{1}{\ell}\alpha_i^{\ell}\\ &= \tfrac{1}{\ell}\sum_{i=1}^{p}\gamma_i\alpha_i^{k+\ell-1} = \tfrac{1}{\ell}\tfrac{1}{k+\ell}\end{aligned}$$

Analog: $\displaystyle\sum_{i,j=1}^{p}\gamma_j\beta_{ji}\alpha_j^{\ell-1}\alpha_i^{k-1} = \frac{1}{k}\frac{1}{k+\ell}$

$$\rightsquigarrow q_{k\ell} = \frac{1}{\ell}\frac{1}{k+\ell} + \frac{1}{k}\frac{1}{k+\ell} - \frac{1}{k\ell} = \frac{k+\ell-(k+\ell)}{k\cdot\ell\cdot(k+\ell)} = 0 \ (k,\ell = 1,\dots,p)$$

und $Q = 0$ ist positiv semidefinit.

2. Radau IIa: $B(2p-1)$ und $C(p)$ sind erfüllt.
   Analog wie oben folgt $q_{\ell k} = 0$ für $\ell + k < 2p$.
   Also gilt: $Q = \begin{pmatrix} 0 & \cdots & 0 & 0\\ \vdots & & & \vdots\\ 0 & \cdots & 0 & 0\\ 0 & \cdots & 0 & q_{pp}\end{pmatrix}$

   $\rightsquigarrow Q$ ist positiv semidefinit gdw. $q_{pp} \geq 0$.
   Es gilt

$$\begin{aligned}q_{pp} &= \sum_{i,j=1}^{p} m_{ij}\alpha_j^{p-1}\alpha_i^{p-1}\\ &= 2\left(\sum_{i=1}^{p}\gamma_i\alpha_i^{p-1}\right)\underbrace{\left(\sum_{j=1}^{p}\beta_{ij}\alpha_i^{p-1}\right)}_{=\frac{1}{p}\alpha_i^p} - \underbrace{\left(\sum_{i=1}^{p}\gamma_i\alpha_i^{p-1}\right)}_{=\frac{1}{p}}\underbrace{\left(\sum_{j=1}^{p}\gamma_j\alpha_j^{p-1}\right)}_{=\frac{1}{p}}\\ &= \tfrac{2}{p}\left(\sum_{i=1}^{p}\gamma_i\alpha_i^{2p-1} - \tfrac{1}{2p}\right)\end{aligned}$$

   Wir setzen $t_{\ell i} = t_{\ell-1} + \alpha_i h_\ell$, $i = 1,\dots,p$ mit $t_{\ell p} = t_\ell$ wegen $\alpha_p = 1$, betrachten das Polynom

$$P(t) = \prod_{i=1}^{p-1}(t - t_{\ell i})^2(t - t_\ell)$$

   vom Grad $2p-1$ und schreiben es in der Form

$$P(t) = (t - t_{\ell-1})^{2p-1} + P_*(t - t_{\ell-1}) \text{ mit Grad } P_* \leq 2p-2.$$

$\rightsquigarrow 0 = P(t_{\ell i}) = (\alpha_i h_\ell)^{2p-1} + P_*(\alpha_i h_\ell)$. Wegen der Wahl der $\alpha_1, \ldots, \alpha_{p-1}$ gilt:

$$\begin{aligned}
\int_{t_{\ell 1}}^{t_\ell} P_*(t - t_{\ell-1})dt &= h_\ell \sum_{i=1}^{p} \gamma_i P_*(t_{\ell i} - t_{\ell-1}) \\
&= h_\ell \sum_{i=1}^{p} \gamma_i P_*(\alpha_i h_\ell) \\
&= -h_\ell \sum_{i=1}^{p} \gamma_i (\alpha_i h_\ell)^{2p-1} \\
\rightsquigarrow \int_{t_{\ell-1}}^{t_\ell} P(t)dt &= \int_{t_{\ell-1}}^{t_\ell} (t - t_{\ell-1})^{2p-1} dt + \int_{t_{\ell-1}}^{t_\ell} P_*(t - t_{\ell-1})dt \\
&= \tfrac{1}{2p} h_\ell^{2p} - h_\ell \sum_{i=1}^{p} \gamma_i (\alpha_i h_\ell)^{2p-1} \\
&= h_\ell^{2p} \left(\tfrac{1}{2p} - \sum_{i=1}^{p} \gamma_i \alpha_i^{2p-1} \right) = -\tfrac{p}{2} q_{pp} h_\ell^{2p}
\end{aligned}$$

Wegen $\int_{t_{\ell-1}}^{t_\ell} P(t)dt < 0$, folgt $q_{pp} > 0$. □

Beispiel 3.49 *(B-stabile Runge-Kutta-Verfahren)*
a) implizite Mittelpunktregel (Gauß, $p = 1$)
b) impliziter Euler (Radau IIa, $p = 1$)
c) Lineare Einschrittverfahren:

$$x_\ell = x_{\ell-1} + h_\ell((1-b)f(x_{\ell-1}, t_{ell-1}) + bf(x_\ell, t_{ell})),\ \ell \in \mathbb{N}.$$

Es gilt (siehe Beispiel 3.15a)) $\gamma_1 = 1 - b$, $\gamma_2 = b$ und

$$B = \begin{pmatrix} 0 & 0 \\ 1-b & b \end{pmatrix}.$$

$\rightsquigarrow$ *γ_1 und γ_2 sind nichtnegativ gdw. $b \in [0, 1]$.*
$\rightsquigarrow$ *$M = \begin{pmatrix} -(1-b)^2 & 0 \\ 0 & b^2 \end{pmatrix}$ wegen $m_{ij} = \gamma_i\beta_{ij} + \gamma_i\beta_{ji} - \gamma_i\gamma_j$, $i, j = 1, 2$.*
$\rightsquigarrow$ *M ist positiv semidefinit gdw. $b = 1$ (Euler implizit).*
D.h. die Trapezregel ist nicht algebraisch stabil.

Eine Aussage zur B-Stabilität variabler linearer Mehrschrittverfahren läßt sich nicht herleiten. Es existiert allerdings eine Erweiterung des Konzeptes der B-Stabilität, dass Anwendungen auf gewisse lineare Mehrschrittverfahren auf äquidistanten Gittern erlaubt.
Wir betrachten dazu ein k-schrittiges lineares Mehrschrittverfahren auf äquidistanten Gittern

$$(*) \qquad \sum_{j=0}^{k} a_j x_{\ell-j} = h \sum_{j=0}^{k} b_j f(x_{\ell-j}, t_{\ell-j}),\ \ell = k, k+1, \ldots,$$

mit den beiden Polynomen

$$\rho(\lambda) = \sum_{j=0}^{k} a_j \lambda^{k-j} \quad \text{und} \quad \sum_{j=0}^{k} b_j \lambda^{k-j},$$

wobei hier eine andere Normierung verwendet wird, nämlich $\sigma(1) = 1$ anstatt $a_0 = 1$.

Definition 3.50 *Das lineare Mehrschrittverfahren (auf $\mathcal{G}_0$)*

$$\sum_{j=0}^{k} a_j x_{\ell-j} = hf\Big(\sum_{j=0}^{k} b_j x_{\ell-j}, \sum_{j=0}^{k} b_j t_{\ell-j}\Big)$$

heißt das zu $(*)$ assoziierte one-leg Verfahren.

Beispiel 3.51

(a) Das zur Trapezregel assoziierte one-leg Verfahren ist die implizite Mittelpunktregel.

(b) Die k-schrittige BDF und das assoziierte one-leg Verfahren sind identisch.

Definition 3.52 *Ein k-schrittiges one-leg Verfahren heißt G-stabil, falls eine symmetrische, positiv definite Matrix $G = (g_{ij}) \in \mathbb{R}^{k\times k}$ existiert, so dass für zwei vom Verfahren erzeugten numerischen Lösungen $(x_\ell)_{\ell\in\mathbb{N}}$ und $(\tilde{x}_{ell})_{\ell\in\mathbb{N}}$*

$$\big\|X_\ell - \tilde{X}_\ell\big\|_G \le \big\|X_{\ell-1} - \tilde{X}_{\ell-1}\big\|_G \, \ell = k, k = 1, \ldots$$

gilt mit $X_\ell = (x_\ell, x_{\ell-1}, \ldots, x_{\ell-k+1})$, $\tilde{X}_\ell = (\tilde{x}_\ell, \tilde{x}_{\ell-1}, \ldots, \tilde{x}_{\ell-k+1})$ in $\mathbb{R}^{mk}$ und

$$\langle X, Y\rangle_G = \sum_{i,j=1}^{k} g_{ij}\langle X_i, Y_i\rangle \quad (\forall X = (X_1, \ldots, X_k),\, Y = (Y_1, \ldots, Y_k) \in \mathbb{R}^{mk})$$

sowie $\|X\|_G = \langle X, X\rangle_G^{\frac{1}{2}}$.

Satz 3.53 *Die Differentialgleichung sei schwach kontraktiv bzgl. $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^m$. Falls eine symmetrische, positiv definite Matrix $G = (g_{ij}) \in \mathbb{R}^{k\times k}$ und reelle Zahlen α_i, $i = 0, \ldots, k$, existieren, so dass für Polynome ρ und σ vom Grad k gilt*

$$(G) \quad \frac{1}{2}(\rho(\lambda)\sigma(\omega) + \rho(\omega)\sigma(\lambda)) = (\lambda\omega - 1)\sum_{i,j=1}^{k} g_{ij}\lambda^{k-i}\omega^{k-j} + \sum_{i,j=0}^{k} \alpha_i\alpha_j\lambda^{k-i}\omega^{k-j},$$

so ist das mit ρ und σ gebildete k-schrittige one-leg Verfahren G-stabil (dabei ist die Konstante $\frac{1}{2}$ durch eine beliebige positive Konstante ersetzbar).

Beweis: Wir betrachten das mit den Polynomen ρ und σ (mit den Bezeichnungen wie früher) gebildete one-leg Verfahren und erhalten wegen der schwachen Kontraktivität

$$\begin{aligned}
0 \;\ge\; & \frac{h}{2}\Big\langle f\Big(\sum_{i=0}^{k} b_i x_{\ell-i}, \sum_{i=0}^{k} b_i t_{\ell-i}\Big) - f\Big(\sum_{i=0}^{k} b_i \tilde{x}_{\ell-i}, \sum_{i=0}^{k} b_i t_{\ell-i}\Big), \sum_{j=0}^{k} b_j (x_{\ell-j} - \tilde{x}_{\ell-j})\Big\rangle \\
= \;& \frac{1}{2}\Big\langle \sum_{i=0}^{k} a_i (x_{\ell-i} - \tilde{x}_{\ell-i}), \sum_{j=0}^{k} b_j (x_{\ell-j} - \tilde{x}_{\ell-j})\Big\rangle \\
= \;& \frac{1}{2}\sum_{i,j=0}^{k} a_i b_j \big\langle x_{\ell-i} - \tilde{x}_{\ell-i}, x_{\ell-j} - \tilde{x}_{\ell-j}\big\rangle
\end{aligned}$$

Wir ersetzen nun formal $\langle x_{\ell-i} - \tilde{x}_{\ell-i}, x_{\ell-j} - \tilde{x}_{\ell-j} \rangle$ durch $\lambda^{k-i}\omega^{k-j}$ und erhalten aus (G)

$$\begin{aligned}
0 \;\geq\; & \frac{1}{2}(\rho(\lambda)\sigma(\omega) + \rho(\omega)\sigma(\lambda)) \\
= \; & \sum_{i,j=1}^{k} g_{ij}\big(\lambda^{k-i+1}\omega^{k-j+1} - \lambda^{k-i}\omega^{k-j}\big) + \sum_{i,j=0}^{k} \alpha_i\alpha_j\lambda^{k-i}\omega^{k-j} \\
= \; & \sum_{i,j=1}^{k} g_{ij}\big(\langle x_{\ell-i+1} - \tilde{x}_{\ell-i+1}, x_{\ell-j+1} - \tilde{x}_{\ell-j+1}\rangle - \langle x_{\ell-i} - \tilde{x}_{\ell-i}, x_{\ell-j} - \tilde{x}_{\ell-j}\rangle\big) \\
& \qquad + \sum_{i,j=0}^{k} \alpha_i\alpha_j\langle x_{\ell-i} - \tilde{x}_{\ell-i}, x_{\ell-j} - \tilde{x}_{\ell-j}\rangle \\
= \; & \|X_\ell - \tilde{X}_\ell\|_G^2 - \|X_{\ell-1} - \tilde{X}_{\ell-1}\|_G^2 + \Big\| \sum_{i=0}^{k} \alpha_i(x_{\ell-i} - \tilde{x}_{\ell-i})\Big\|^2 \\
\geq \; & \|X_\ell - \tilde{X}_\ell\|_G^2 - \|X_{\ell-1} - \tilde{X}_{\ell-1}\|_G^2
\end{aligned}$$

wobei wir beim vorletzten Gleichheitszeichen die formale Ersetzung rückgängig gemacht haben. Also gilt

$$\|X_\ell - \tilde{X}_\ell\|_G \leq \|X_{\ell-1} - \tilde{X}_{\ell-1}\|_G.$$

und die Aussage ist bewiesen. □

Beispiel 3.54 *Wir betrachten die zweischrittige BDF auf einem Gitter in $\mathcal{G}_0$ mit Schrittweite h*

$$\frac{3}{2}x_\ell - 2x_{\ell-1} + \frac{1}{2}x_{\ell-2} = hf(x_\ell, t_\ell),\ \ell = 2, 3, \ldots.$$

Dann ist $\rho(\lambda) = \frac{3}{2}\lambda^2 - 2\lambda + \frac{1}{2}$ und $\sigma(\lambda) = \lambda^2$. Satz 3.53 legt nun nahe g_{ij}, $i, j = 1, 2$, und α_i, $i = 0, 1, 2$, so zu bestimmen, dass die Gleichung (G) gültig ist. Durch Ausmultiplikation in der linken und rechten Seite von (G) entstehen die Gleichungen

$$\begin{aligned}
\frac{3}{2} \;&=\; g_{11} + \alpha_0^2, \quad -2 = 2g_{21} + 2\alpha_1\alpha_0, \quad \frac{1}{2} = 2\alpha_0\alpha_2, \\
0 \;&=\; -g_{22} + \alpha_2^2, \quad 0 = -2g_{21} + 2\alpha_2\alpha_1, \quad 0 = g_{22} - g_{11} + \alpha_1^2.
\end{aligned}$$

Durch Addition aller 6 Gleichungen ergibt sich $(\alpha_0+\alpha_1+\alpha_2)^2 = 0$, d.h. $\alpha_0+\alpha_1+\alpha_2 = 0$. Durch Addition der zweiten und fünften Gleichung entsteht $-1 = \alpha_1(\alpha_0+\alpha_2)$ und damit

$$(\alpha_0 + \alpha_2)^2 = 1 \quad \textit{und} \quad \alpha_0\alpha_2 = \frac{1}{4}.$$

Folglich erhält man $\alpha_0 = \alpha_2 = \frac{1}{2}$, $\alpha_1 = -1$, und die positiv definite Matrix

$$G = \frac{1}{4}\begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}.$$

Um weitere Schlussfolgerungen ableiten zu können, spezialisieren wir die DGL zu einem linearen Differentialgleichungssystem mit konstanter Matrix

$$(**) \quad x'(t) = Ax(t), t \in [t_0, +\infty), \text{ mit } A \in \mathbb{R}^{m\times m}$$

und wissen nach Folgerung 1.18, dass $(**)$ schwach kontraktiv ist, falls für die Eigenwerte λ_i, $i = 1, \ldots, p$ von A $\max\limits_{i=1,\cdots,p} \mathrm{Re}(\lambda_i) \leq 0$ gilt und alle Jordan-Kästchen zu rein imaginären Eigenwerten Dimension 1 besitzen.

Lemma 3.55
Für die Eigenwerte λ_i, $i = 1, \ldots, p \leq m$ von $A \in \mathbb{R}^{m\times m}$ gelte $\max\limits_{i=1,\cdots,p} \mathrm{Re}(\lambda_i) \leq 0$ und alle Jordan-Kästchen zu rein imaginären Eigenwerten besitzen Dimension 1.
Es existiert eine reguläre Matrix $T \in \mathbb{C}^{m\times m}$, so daß jede Komponente von $z(t) = Tx(t)$ beschränkt ist auf $[t_0, +\infty)$ und

$$(***) \quad z'_{j_i}(t) = \lambda_i z_{j_i}(t) \quad (t \in [t_0, +\infty), i = 1, \ldots, p)$$

für gewisse Komponenten $j_i \in \{1, \ldots, m\}$, $i = 1, \ldots, p$, gilt.

Beweis: Es existiert eine reguläre Matrix $T \in \mathbb{C}^{m\times m}$, so daß die Jordan-Normalform J von A mit den Jordan-Kästchen J_j, $j = 1, \ldots, r$, $r \geq p$, die Form hat

$$J = TAT^{-1} = \begin{pmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_r \end{pmatrix}.$$

Die Jordan-Kästchen der Dimension $\nu_j > 1$ haben die Form

$$J_j = \begin{pmatrix} \lambda_{i_j} & 1 & 0 & \cdots & 0 \\ 0 & \lambda_{i_j} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{i_j} & 1 \\ 0 & 0 & \cdots & 0 & \lambda_{i_j} \end{pmatrix} \quad \text{mit} \quad \mathrm{Re}(\lambda_{i_j}) < 0\,.$$

Bei Anwendung der Transformation auf die Differentialgleichung (**) entsteht

$$\begin{aligned} z'(t) &= Tx'(t) = TAx(t) = TAT^{-1}z(t) = Jz(t) \\ z'_{j_i}(t) &= \lambda_i z_{j_i}(t) \quad (i = 1, \ldots, p), \text{ d.h. } z_{j_i}(t) = \exp(\lambda_i(t - t_0))z_{j_i}(t_0) \end{aligned}$$

wobei die letzten p Gleichungen jeweils den letzten Zeilen der Jordan-Kästchen entsprechen. Die restlichen Gleichungen des Systems $z'(t) = Jz(t)$ führen zu Lösungen der Gestalt

$$z_{j_i-\nu}(t) = \left(\frac{t}{t_0}\right)^{\nu} \exp(\lambda_i(t - t_0))z_{j_i-\nu}(t_0) \quad (\nu = 0, \ldots, \nu_i - 1),$$

mit $\mathrm{Re}(\lambda_i) < 0$. Deswegen konvergieren diese Lösungen gegen 0 für $t \to +\infty$ und sind beschränkt. □

Nach Lemma 3.55 genügt es, nur die skalare Testgleichung

$$(+) \quad x'(t) = \lambda x(t), \quad (t \in [t_0, +\infty),\, x(t_0) = x_0,\, \lambda \in \mathbb{C})$$

zu betrachten.

Als nächstes charakterisieren wir die Integrationsverfahren (realisiert auf $\mathcal{G}_0$), für die die nachfolgende Theorie anwendbar ist.

Definition 3.56
Ein Integrationsverfahren heißt rational, falls es bei Anwendung auf die Testgleichung (+) mit konstanter Schrittweite $h > 0$ die Gestalt

$$\sum_{j=0}^{k} \eta_j(h\lambda) x_{\ell-j} = 0 \quad \textit{oder} \quad x_\ell + \sum_{j=1}^{k} \frac{\eta_j(h\lambda)}{\eta_0(h\lambda)} x_{\ell-j} = 0 \quad (\ell = k, k+1, \ldots)$$

mit $k \in \mathbb{N}$ und Polynomen $\eta_j(\cdot)$, $j = 0, \ldots, k$, annimmt, wobei wenigstens eins der Polynome nicht Grad 0 besitzt.

Beispiel 3.57

a) Lineare Mehrschrittverfahren (der Ordnung $k \in \mathbb{N}$ auf $\mathcal{G}_0$):

$$\sum_{j=0}^{k} a_j x_{\ell-j} = h \sum_{j=0}^{k} b_j f(x_{\ell-j}, t_{\ell-j}), \quad \ell = k, k+1, \ldots$$

$$\rightsquigarrow \sum_{j=0}^{k} (a_j - h\lambda b_j) x_{\ell-j} = 0, \quad \ell = k, k+1, \ldots$$
$$\rightsquigarrow \eta_j(z) := a_j - b_j z,\ j = 0, \ldots k,\ a_0 = 1.$$

b) Runge-Kutta-Verfahren (p-stufig auf $\mathcal{G}_0$):

$$\begin{aligned}
x_\ell &= x_{\ell-1} + h \sum_{j=1}^{p} \gamma_j f(\bar{x}_\ell^j, t_{\ell-1} + \alpha_j h) \\
\bar{x}_\ell^i &= x_{\ell-1} + h \sum_{j=1}^{p} \beta_{ij} f(\bar{x}_\ell^j, t_{\ell-1} + \alpha_j h),\ i = 1, \ldots, p \\
\rightsquigarrow x_\ell &= x_{\ell-1} + h\lambda \sum_{j=1}^{p} \gamma_j \bar{x}_\ell^j = x_{\ell-1} + h\lambda \langle \gamma, \bar{x}_\ell \rangle \\
\bar{x}_\ell^i &= x_{\ell-1} + h\lambda \sum_{j=1}^{p} \beta_{ij} \bar{x}_\ell^j \quad \textit{oder} \quad (I - h\lambda B)\bar{x}_\ell = e\, x_{\ell-1} \\
\rightsquigarrow \bar{x}_\ell &= (I - h\lambda B)^{-1} e\, x_{\ell-1} \\
\rightsquigarrow x_\ell &= (1 + h\lambda \langle \gamma, (I - h\lambda B)^{-1} e \rangle) x_{\ell-1}.
\end{aligned}$$

$$\rightsquigarrow\ k=1 \quad \textit{und} \quad -\frac{\eta_1(z)}{\eta_0(z)} = 1 + z\langle \gamma, (I - zB)^{-1} e\rangle.$$

Aus der Cramerschen Regel folgt die alternative Darstellung

$$-\frac{\eta_1(z)}{\eta_0(z)} = \frac{\det(I - zB + ze\gamma^\top)}{\det(I - zB)}.$$

Dabei ist I die $p \times p$ Einheitsmatrix und $e = (1, \ldots, 1)^\top \in \mathbb{R}^p$.

Satz 3.58

Für die Gleichungen $\sum_{j=0}^{k} \eta_j(z) x_{\ell-j} = 0$, $\ell = k, k+1, \ldots$, bei gegebenen $x_0, \ldots, x_{k-1}$ gilt:

$\{x_\ell : \ell \in \mathbb{N}\}$ ist beschränkt in $\mathbb{C}$, falls das Polynom $\sigma(z, \mu) = \sum_{j=0}^{k} \eta_j(z) \mu^{k-j}$ die Wurzelbedingung nach Dahlquist erfüllt.

Beweisidee:
Seien $\mu_j(z), j = 1, \ldots, r$, die Nullstellen des Polynoms $\sigma(z, \cdot)$ der Ordnung k mit den Vielfachheiten n_j, $j = 1, \ldots, r$, d.h. $\sum_{j=1}^{r} n_j = k$. Dann hat die allgemeine Lösung der Gleichungen bei gegebenen $x_0, \ldots, x_{k-1}$ die Form

$$x_\ell = \sum_{j=1}^{r} p_j(\ell)(\mu_j(z))^\ell \quad (\ell = k, k+1, \ldots),$$

wobei die p_j Polynome vom Grad $n_j - 1$ sind für $j = 1, \ldots, r$. Deshalb gilt

$$|x_\ell| \le \sum_{j=1}^{r} |p_j(\ell)||\mu_j(z)|^\ell \quad (\ell = k, k+1, \ldots).$$

Ist also die Wurzelbedingung nach Dahlquist erfüllt, so sind die $|x_\ell|$, $\ell \in \mathbb{N}$, gleichmäßig beschränkt. □

Definition 3.59

Es sei ein rationales Integrationsverfahren mit den Polynomen $\eta_j(\cdot), j = 0, \ldots, k$ gegeben und wir betrachten das Polynom $\sigma(z, \mu) := \sum_{j=0}^{k} \eta_j(z) \mu^{k-j}$.

Die Menge $H_A := \{z \in \mathbb{C} : \sigma(z, \cdot)$ erfüllt die Wurzelbedingung nach Dahlquist$\}$ heißt Bereich absoluter Stabilität des Verfahrens.
Ein solches Verfahren heißt
A-stabil, falls $\mathbb{C}_- := \{z \in \mathbb{C} : \mathrm{Re}(z) \le 0\} \subseteq H_A$,
$A(\alpha)$-stabil, $0 < \alpha < \frac{\pi}{2}$, falls $S_\alpha := \{z \in \mathbb{C} : |\arg(-z)| < \alpha, z \ne 0\} \subseteq H_A$.
A(0)-stabil, falls ein $\alpha \in (0, \frac{\pi}{2})$ existiert, so dass $S_\alpha \subseteq H_A$.

Folgerung 3.60

Für ein rationales Integrationsverfahren mit Schrittweite $h > 0$ in Anwendung auf $x'(t) = Ax(t)$ mit einer Matrix $A \in \mathbb{R}^{m \times m}$, deren Eigenwerte in $\mathbb{C}_-$ liegen und für rein imaginäre Eigenwerte nur Jordan-Kästchen der Dimension 1 auftreten, gilt:
Die Näherungslösung $\|x_\ell\|$ ist gleichmäßig beschränkt wie die Lösung der DGL, falls $h\lambda \in H_A$ für jeden Eigenwert λ von A gilt. Ist das Integrationsverfahren A-stabil, so gilt die Beschränktheit der Näherungslösung für alle Schrittweiten $h > 0$.

Beweis:
Die erste Aussage folgt aus Lemma 3.55 und Satz 3.58. Mit $\lambda \in \mathbb{C}_-$ gilt auch $h\lambda \in \mathbb{C}_- \subseteq H_A$. □

Die Beschränkungen an die Schrittweite h bleiben gering, wenn H_A einen möglichst großen Teil von $\mathbb{C}_-$ enthält. Sie werden allerdings wesentlich, wenn H_A beschränkt ist.

Beispiel 3.61
Lineare Einschrittverfahren: $x_\ell = x_{\ell-1} + h((1-b)f(x_{\ell-1}, t_{\ell-1}) + bf(x_\ell, t_\ell))$ *(*$b \in [0,1]$*).*
Bei Anwendung auf $x'(t) = \lambda x(t), t \in [t_0, +\infty)$ *entsteht*

$$\eta_0(h\lambda)x_\ell + \eta_1(h\lambda)x_{\ell-1} = 0, \quad \eta_0(z) := 1 - bz, \quad \eta_j(z) := -1 - (1-b)z \ (z \in \mathbb{C}).$$

Einzige Nullstelle von $\sigma(z, \cdot)$*:* $\mu_1(z) = \frac{1+(1-b)z}{1-bz}$
Bereich absoluter Stabilität:

$$H_A = \left\{ z \in \mathbb{C} : \left| \frac{1+(1-b)z}{1-bz} \right| \leq 1 \right\}$$

$$\begin{aligned} \left| \frac{1+(1-b)(x+iy)}{1-b(x+iy)} \right|^2 \leq 1 \quad &\Leftrightarrow \quad \frac{(x+(1-bx))^2 + (1-b)^2y^2}{(1-bx)^2 + b^2y^2} \leq 1 \\ &\Leftrightarrow \quad x^2 + 2x(1-bx) + (1-2b)y^2 \leq 0 \\ &\Leftrightarrow \quad (x^2+y^2)(1-2b) \leq -2x \end{aligned}$$

$\rightsquigarrow$ *Ungleichung richtig für jedes* $\mathrm{Re}(z) = x \leq 0$*, falls* $1 - 2b \leq 0$.
$\rightsquigarrow$ *lineare Einschrittverfahren sind A-stabil gdw.* $b \in [\frac{1}{2}, 1]$.
Bereich absoluter Stabilität:

$$\begin{aligned} &\textit{expliziter Euler:} && H_A = \{z \in \mathbb{C} : |1+z| \leq 1\} \\ &\textit{impliziter Euler:} && H_A = \left\{z \in \mathbb{C} : \left|\tfrac{1}{1-z}\right| \leq 1\right\} = \{z \in \mathbb{C} : 1 \leq |1-z|\} \\ &\textit{Trapezregel:} && H_A = \left\{ z \in \mathbb{C} : \left| \tfrac{1+\frac{1}{2}z}{1-\frac{1}{2}z} \right| \leq 1 \right\} = \mathbb{C}_- \end{aligned}$$

Lemma 3.62 *Ist* p *ein Polynom vom Grad* n *über* $\mathbb{C}$*, so existieren Konstanten* $K > 0$ *und* $r_0 > 0$*, so dass*

$$|p(z)| \geq Kr^n \quad \forall z \in \mathbb{C}, |z| = r \geq r_0.$$

Beweis: Wir schreiben p in der Form

$$p(z) = z^n \Big(a_0 + \sum_{j=1}^{n} \frac{a_j}{z^j} \Big).$$

Daraus folgt für alle $z \in \mathbb{C}$, $|z| = r$:

$$|p(z)| \geq r^n \Big| |a_0| - \Big| \sum_{j=1}^{n} \frac{a_j}{z^j} \Big| \Big|$$

Wir wählen nun $r_0 > 0$ so, dass für $z \in \mathbb{C}$, $|z| = r \geq r_0$ gilt

$$\Big|\sum_{j=1}^{n} \frac{a_j}{z^j}\Big| \leq \sum_{j=1}^{n} \frac{|a_j|}{r^j} \leq \sum_{j=1}^{n} \frac{|a_j|}{r_0^j} \leq \frac{|a_0|}{2}.$$

Deshalb folgt für $z \in \mathbb{C}$, $|z| = r \geq r_0$: $|p(z)| \geq \frac{|a_0|}{2} r^n$. □

Satz 3.63
Für kein explizites (d.h. $\eta_0(z) = 1$) rationales Integrationsverfahren ist H_A unbeschränkt.

Beweis:
Annahme: Es existiert ein explizites rationales Integrationsverfahren mit unbeschränktem H_A.
Nach dem Satz von Vieta gilt für das Polynom $\sigma(z, \mu) = \mu^k + \sum_{j=1}^{k} \eta_j(z)\mu^{k-j}$:

$$\eta_j(z) = (-1)^j \sum_{1 \leq i, < \cdots < i_j \leq k} \mu_i(z) \cdots \mu_{i_j}(z)$$

(insbesondere: $\eta_1(z) = -\sum_{j=1}^{k} \mu_j(z)$, $\eta_k(z) = (-1)^k \prod_{j=1}^{k} \mu_j(z)$)
Daraus folgt:

$$|\eta_j(z)| \leq \binom{k}{j} (\max_{j=1,\ldots,k} |\mu_j(z)|)^k$$

$$\rightsquigarrow \quad |\eta_j(z)| \leq \binom{k}{j}, \ \forall z \in H_A \quad (j = 1, \ldots, k).$$

Nach Definition existiert $j_0 \in \{1, \ldots, k\}$, so dass η_{j_0} Grad $s \geq 1$ besitzt. Deshalb existieren positive Konstanten K und r_0 mit

$$\binom{k}{j_0} \geq |\eta_{j_0}(z)| \geq K r^s,$$

falls $z \in H_A$, $|z| = r \geq r_0$, gilt. Dies ist ein Widerspruch zur Annahme der Unbeschränktheit von H_A. □

Satz 3.64
B-stabile Runge-Kutta-Verfahren sind auch A-stabil.

Beweis:
Wir betrachten die sog. Stabilitätsfunktion $R(z) := 1 + z\langle \gamma, [I - zB]^{-1} e \rangle$ und wissen (vgl. Bsp. 3.57b)), daß Runge-Kutta-Verfahren bei Anwendung auf $x'(t) = \lambda x(t)$ die Form $x_\ell = R(h\lambda) x_{\ell-1}$ annehmen. B-Stabilität impliziert nun, daß aus $\mathrm{Re}(\lambda) \leq 0$ folgt

$$|x_\ell| \leq |x_{\ell-1}|$$

bei beliebig gewähltem Anfangswert x_0.
$\rightsquigarrow |R(h\lambda)| \leq 1$ falls $\mathrm{Re}(\lambda) \leq 0$.
Da $R(z)$ gerade die Nullstelle des charakteristischen Polynoms $\sigma(z, \mu) = \mu - R(z)$ ist, folgt $\mathbb{C}_- \subseteq H_A$. □

Satz 3.65 *(Dahlquist)*
Falls die Polynome ρ und σ keinen gemeinsamen Teiler besitzen, ist das auf ρ bzw. σ beruhende lineare Mehrschrittverfahren A-stabil gdw. das assoziierte one-leg Verfahren G-stabil ist.

Beweis: Hairer-Wanner II, Satz V.6.7.

Beispiel 3.66 *(Rückwärtige Differentiationsformeln, BDF)*

$$\sum_{j=0}^{k} a_j x_{\ell-1} = h b_0 f(x_\ell, t_\ell) \quad (k \leq 6)$$

Diese sind stabil für $1 \leq k \leq 6$ und konsistent mit Ordnung $s = k$.
Wie sehen die Bereiche absoluter Stabilität aus?
$k = 1$: Euler implizit, A-stabil.
$k = 2$: $x_\ell - \frac{4}{3}x_{\ell-1} + \frac{1}{3}x_{\ell-2} = h\frac{2}{3}f(x_\ell, t_\ell)$.
Dieses Verfahren ist G-stabil nach Beispiel 3.54 und deshalb A-stabil nach Satz 3.65.
$3 \leq k \leq 6$: Die Verfahren sind $A(\alpha)$-stabil, wobei

k	3	4	5	6
α	$86,03°$	$73,35°$	$51,84°$	$17,84°$

Satz 3.67 *(Zweite Dahlquist-Barriere)*
Jedes konsistente A-stabile lineare Mehrschrittverfahren besitzt eine Konsistenzordnung $s \leq 2$, Die Trapezregel ist unter diesen linearen Mehrschittverfahren mit Konsistenzordnung $s = 2$ dasjenige mit der besten Fehlerkonstanten ($C = -\frac{1}{12}$).

Beweis: siehe Kapitel V.1 in Hairer-Wanner II.

Bemerkung 3.68
Es bieten sich 2 Auswege für das Erreichen höherer Konsistenzordnung und das Auftreten unbeschränkter Bereiche H_A an:
(i) implizite Runge-Kutta-Verfahren z.B. vom Typ Gauß und Radau IIa.
(ii) Zu jedem $\alpha \in (0, \frac{\pi}{2})$ existieren $A(\alpha)$-stabile lineare Mehrschrittverfahren beliebiger Konsistenzordnung (jedoch mit i.a. großen Fehlerkonstanten je größer α)! Diese werden aber nicht empfohlen. Möglich sind allerdings die BDF-Verfahren für $s = k \leq 4$ (vgl. Beispiel 3.66).

Bisher war nur unser Ziel, beschränkte bzw. abklingende Lösungen mit den Integrationsverfahren korrekt "nachzubilden". Wie steht es aber, falls auch anwachsende Lösungskomponenten existieren ?

Definition 3.69 *(Griepentrog)*
Ein rationales Integrationsverfahren heißt
a) asymtotisch exakt, falls ein $r > 0$ existiert, so daß $H_A \cap B(0,r) = \mathbb{C}_- \cap B(0,r)$.
b) global asymptotisch exakt, falls $H_A = \mathbb{C}_-$.

Bemerkung 3.70
Asymptotische Exaktheit bedeutet, daß aus

$$|z| \le r \quad \begin{array}{l} \mathrm{Re}(z) < 0 \text{ folgt } |\mu_j(z)| < 1 \text{ für alle } j = 1, \ldots, p \\ \mathrm{Re}(z) > 0 \text{ folgt } |\mu_j(z)| > 1 \text{ für mindestens ein } j. \end{array}$$

Bei hinreichend kleiner Schrittweite erfolgt die richtige Nachbildung des asymptotischen Verhaltens. Bei globaler asymptotischer Exaktheit ist diese richtige Nachbildung unabhängig von der Wahl der Schrittweite.

Satz 3.71 *(Jeltsch, Griepentrog)*
Ein streng stabiles $(\mu_1(0) = 1,\ |\mu_j(0)| < 1, j = 2, \ldots, k)$ *rationales Integrationsverfahren ist global asymptotisch exakt gdw.* $k = 1$, $\eta_1(\cdot)$ *besitzt nur Nullstellen* z *mit* $\mathrm{Re}(z) < 0$ *und es gilt* $\eta_1(z) = -\eta_0(-z)$, $\forall z \in \mathbb{C}$.

Beweis: Griepentrog: Wiss. Zeitschrift HUB, Math.-Nat. Reihe 19 (1970).

Folgerung 3.72
Das einzige streng stabile, global asymptotisch exakte lineare Mehrschrittverfahren ist die Trapezregel.

Beweis: Für lineare Einschrittverfahren gilt: $\eta_0(z) = 1 - bz$ und $\eta_1(z) = -1 - (1-b)z$. Deshalb gilt $\eta_1(z) = -\eta_0(-z)$ gdw. $b = 1 - b$, d.h. $b = \frac{1}{2}$. □

Definition 3.73
Eine rationale Funktion $\frac{\sum_{i=0}^{k} a_i z^i}{\sum_{i=0}^{j} b_i z^i} =: R_{jk}(z)$ *(mit* $b_0 = 1$*) heißt Padé-Approximation an* e^z *vom Index* (j,k)*, wenn gilt*

$$\frac{d^r}{dz^r} R_{jk}(z)\Big|_{z=0} = \frac{d^r}{dz^r} e^z\Big|_{z=0}, r = 0, \ldots, j+k, \quad \text{oder } R_{jk}(z) = e^z + O(z^{j+k+1}) \text{ (für } z \to 0).$$

Aussage: Padé-Approximationen von e^z existieren für jeden Index (j,k) ($j+k+1$ Bedingungen für $j+k+1$ unbekannte Koeffizienten). Es gilt:

$$a_s = \sum_{i=0}^{s} b_{s-i} \quad (s = 0, 1, \ldots).$$

Satz 3.74
Die Funktion $R(z) = 1 + z\langle \gamma, (I - zB)^{-1} e \rangle$ *eines p-stufigen Runge-Kutta-Verfahrens vom Typ Gauß (bzw. Radau IIa) ist die Padé-Approximation an* e^z *vom Index* (p,p) *(bzw.* $(p, p-1)$*).*
Die p-stufigen Runge-Kutta-Verfahren vom Typ Gauß sind B-stabil (insbesondere A-stabil) und global asymptotisch exakt.

Beweis: vgl. Kapitel 6.22 in Strehmel-Weiner 1995.

3.7 Integration steifer Differentialgleichungen

Definition 3.75
Die Anfangswertaufgabe $x'(t) = f(x(t), t)$, $t \in [t_0, T]$, $x(t_0) = x_0$ mit Lösung x_* heißt steif, falls für die Eigenwerte $\lambda_i(t), i = 1, \ldots, m$, der Jacobi-Matrix $\frac{\partial f}{\partial x}(x_*(t), t)$, $t \in [t_0, T]$, gilt:
(i) $\mathrm{Re}(\lambda_i(t)) < 0, \quad i = 1, \ldots, m, t \in [t_0, T]$
(ii) $\sigma = \frac{\max\limits_i |\mathrm{Re}(\lambda_i(t))|}{\min\limits_i |\mathrm{Re}(\lambda_i(t))|} \gg 1, \quad t \in [t_0, T]$

(dabei meint man Steifheitsfaktoren $\sigma \geq 10^3$)

Für steife Aufgaben ist die Lösung stabil, aber $L_f(T - t_0)$ "groß". Je kleiner $\min\limits_i |\mathrm{Re}(\lambda_i(t))|$ desto weiter muss man integrieren, um den stationären Zustand zu erreichen, je größer $|\mathrm{Re}(\lambda_i(t))|$ desto größer wird die Lipschitzkonstante L_f von f; einseitige Lipschitzkonstanten sind davon unbeeinflusst.
Konsequenz:
Verwendung von Integrationsverfahren, die einen unbeschränkten Bereich H_A besitzen (am besten A-stabil oder $A(\alpha)$-stabil mit großem α), B-stabil oder G-stabil sind.

Bemerkung 3.76 *(Anwendungen steifer Differentialgleichungen)*

- *Gleichungen der chemischen Kinetik*
- *Gleichungen der Schaltkreissimulation*
- *Semidiskretisierung (Ortsdiskretisierung) parabolischer partieller Differentialgleichungen*

Beispiel 3.77 *(Integrationsverfahren für steife Differentialgleichungen)*
a) impliziter Euler, Trapezregel, implizite Mittelpunktregel
b) BDF 1-3
c) Runge-Kutta-Verfahren vom Typ Gauß oder Radau IIa.

Treten Eigenwerte der Jacobi-Matrix mit großem $|\mathrm{Re}(\lambda_i(t))|$ auf, so erwartet man vom Integrationsverfahren, dass diese Anteile schnell abklingen. Das führt zu folgendem Konzept.

Definition 3.78
Ein rationales Integrationsverfahren heißt L-stabil (mit Dämpfungsordnung $\varepsilon > 0$), falls

$$\lim_{\mathrm{Re}(z) \to -\infty} |\mu_j(z)| = 0, \; j = 1, \ldots, k \; (\max_{j=1,\ldots,k} |\mu_j(z)| = 0(z^{-\varepsilon}) \text{ für } \mathrm{Re}(z) \to -\infty)$$

Satz 3.79 *(Jeltsch)*
Ein rationales Integrationsverfahren mit charakteristischem Polynom
$\sigma(z, \mu) = \sum\limits_{j=0}^{k} \eta_j(z) \mu^{k-j}$ *besitzt die Dämpfungsordnung*

$$\varepsilon = \min \left\{ \frac{q_0 - q_j}{j} : q_j \geq 0, \quad j = 1, \ldots, k \right\}$$

wobei $q_j := \begin{cases} -1 & , \ \eta_j(z) \equiv 0 \\ \text{Grad } \eta_j & , \ \text{sonst} \end{cases}$ $(j = 0, \ldots, k)$.
Speziell gilt $\varepsilon > 0$, *falls Grad* $(\eta_0) >$ *Grad* (η_j), $j = 1, \ldots, k$, *und* $\varepsilon \in [0, \frac{q_0}{k}]$,
falls Grad $(\eta_0) \geq$ *Grad* (η_j), $j = 1, \ldots, k$, $\eta_k(z) \neq 0$.

Beispiel 3.80

a) *Lineare Einschrittverfahren:* $\mu_1(z) = \frac{1+(1-b)z}{1-bz} \to 0$ *für* $\mathrm{Re}(z) \to -\infty \Leftrightarrow b = 1$
D. h. nur Euler implizit ist L-stabil.

b) *Lineare Mehrschrittverfahren sind L-stabil, falls*

$$\max_{j=1,\ldots,k} \left| \frac{a_j - b_j z}{1 - b_0 z} \right| \longrightarrow_{\mathrm{Re}(z)\to-\infty} 0 \quad (\textit{Satz von Vieta !}).$$

Dies gilt, falls $b_j = 0$, $j = 1, \ldots, k$, $b_0 \neq 0$. *Deshalb sind die stabilen k-schrittigen BDF* $(1 \leq k \leq 6)$ *L-stabil mit Dämpfungsordnung*

$$\varepsilon = \min \left\{ \frac{1}{j} : j = 1, \ldots, k \right\} = \frac{1}{k} \quad (\textit{nach Satz 3.79}).$$

c) *Runge-Kutta-Verfahren:*
$\mu_1(z) = -\frac{\eta_1(z)}{\eta_0(z)} = 1 + z\langle \gamma, (E - zB)^{-1} e \rangle$
$\rightsquigarrow |\mu_1(z)| \to 0$ *für* $\mathrm{Re}(z) \to -\infty$, *falls Grad* $\eta_0 >$ *Grad* η_1.
Dabei ist die Dämpfungsordnung gerade $\varepsilon :=$ *Grad* $\eta_0 -$ *Grad* η_1.
Die p-stufigen Runge-Kutta-Verfahren vom Typ Radau IIa sind L-stabil mit Dämpfungsordnung $\varepsilon = 1$ *(nach Satz 3.79).*

Bemerkung 3.81 *Da bei steifen DGln insbesondere die Lipschitzkonstanten von f sehr groß sein können, sind die Konstanten bei Konsistenz- bzw. Konvergenzordnungsaussagen enorm groß. Dies macht sich praktisch durch eine scheinbare "Ordnungsreduktion" bemerkbar. Deshalb wurde eine Theorie für die Herleitung von Konvergenzresultaten entwickelt, die nur mit einseitigen Lipschitz-Bedingungen an f auskommt (sog. B-Konvergenz bei RKV). Danach sind p-stufige RKV vom Typ Gauß und Radau IIa nur mit* $O(h^p)$ *konvergent (statt* $O(h^{2p})$*). Es existieren auch spezielle Existenzaussagen für Lösungen impliziter Verfahrensgleichungen auf der Basis einseitiger Lipschitzbedingungen und mit nur moderaten Einschränkungen an Schrittweiten* h_ℓ *(vgl. Hairer-Wanner II, Kap. IV.14, IV.15 und V.6).*

Bemerkung 3.82 *(Implementierung von Verfahren für steife DGLn.)*

- *Wichtig sind Steifheitstests (z. B. Versagen expliziter Verfahren oder Schätzung von Lipschitzkonstanten)*
- *Wesentlich ist ein Gesamtkonzept zur Schrittweitensteuerung und Lösung der nichtlinearen Verfahrensgleichungen bei impliziten Integrationsverfahren.*

- *Es existiert nicht das Verfahren für steife Differentialgleichungen; wichtig ist die Analyse der praktischen Aufgabenstellung, der Dimension m und des asymptotischen Verhaltens von Lösungen. Ist m groß, so sind wahrscheinlich implizite Runge-Kutta-Verfahren ungeeignet $\rightsquigarrow$ Verfahren mit kleinerer Konsistenzordnung!*

Literatur:
E. Hairer, G. Wanner: Solving Ordinary Differential Equations II, Springer, 1991.
K. Strehmel, R. Weiner: Numerik gewöhnlicher Differentialgleichungen, Teubner, 1995.

4 Numerische Methoden für Randwertaufgaben gewöhnlicher Differentialgleichungen

Wir betrachten Aufgabenstellungen der Form

$$x'(t) = f(x(t), t), t \in [t_0, T], r(x(t_0), x(T)) = 0\,,$$

wobei $f : \mathbb{R}^m \times [t_0, T] \to \mathbb{R}^m$, $r : \mathbb{R}^m \times R^m \to \mathbb{R}^m$.

4.1 Korrekt formulierte lineare Randwertaufgaben (RWAn) und Greensche Funktion

Wir betrachten lineare Differentialgleichungen

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

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

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

mit $B_0, B_T \in \mathbb{R}^{m\times m}$ und $d \in \mathbb{R}^m$.
Spezialfälle von Randbedingungen sind:

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

Dabei ist E die $m \times m$ Einheitsmatrix.
Das folgende Beispiel zeigt, daß die (eindeutige) Lösbarkeit der Randwertaufgaben von den Randbedingungen abhängt.

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

$$\begin{array}{rcl} x_1'(t) &=& x_2(t) \quad \textit{oder} \quad x'' + x = 0 \\ x_2'(t) &=& -x_1(t) \quad \textit{bzw.} \qquad x' = Ax \quad \textit{mit} \quad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \end{array}$$

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:

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

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

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

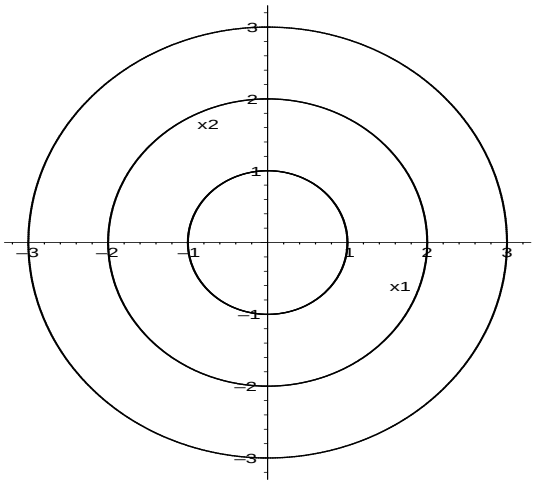
Ist nun $T \neq \pi k$ für alle $k \in \mathbb{Z}$, so erhalten wir $c_2 = \frac{\beta}{\sin T}$, also die eindeutige Lösung

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

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

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

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 $x_1(2\pi) = 0$)

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

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

besitzt eine Lösung $x(\cdot\,, x_0) \in C^1([t_0, T], \mathbb{R}^m)$. Diese hat nach Satz 1.15 die Form

$$x(t, x_0) = X(t)x_0 + x_q(t) = X(t)\left(x_0 + \int_{t_0}^{t} X^{-1}(s)q(s)ds\right) \quad t \in [t_0, T],$$

mit einer Fundamentalmatrix $X \in C^1([t_0, T], \mathbb{R}^{m\times m})$, die das homogene Differentialgleichungssystem löst.
Durch Einsetzen der Lösungen in die Randbedingungen (4.2) erhält man

$$\begin{aligned} B_0X(t_0)x_0 + B_T(X(T)x_0 + x_q(T)) &= d, \quad \text{oder} \\ (B_0X(t_0) + B_TX(T))x_0 &= d - B_Tx_q(T). \end{aligned}$$

Satz 4.2
Die lineare Randwertaufgabe ist eindeutig lösbar für alle $d \in \mathbb{R}^m$, $q(\cdot)$, falls die Matrix $M := B_0 \underbrace{X(t_0)}_{=I} + B_1x(T)$ regulär ist. Wenn die sog. Lösbarkeitsmatrix

$$\boxed{M := B_0X(t_0) + B_TX(T)}$$

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

$$x_0 = M^{-1}\left(d - B_Tx_q\left(T\right)\right).$$

Für die Regularität von

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

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

Beispiel: (Fortsetzung Bsp. 4.1)
Wir betrachten nochmals die Randwertaufgabe aus Beispiel 4.1:

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

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

Die Fundamentalmatrix ist dann

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

so dass man folgende Lösbarkeitsmatrix erhält

$$\begin{aligned} M = B_0X\left(0\right) + B_TX\left(T\right) &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \cos T & \sin T \\ -\sin T & \cos T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \cos T & \sin T \end{pmatrix}. \end{aligned}$$

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

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

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

$$\begin{aligned} \text{und mit } I &= M^{-1}B_0 X(t_0) + M^{-1}B_T X(T) \text{ erhalten wir weiter} \\ &= X(t)\, M^{-1}d + \int_{t_0}^{t} X(t)\, M^{-1}B_0 X(t_0)\, X(s)^{-1}\, q(s)\, ds \\ &\quad - \int_{t}^{T} X(t)\, M^{-1}B_T X(T)X(s)^{-1}q(s)ds. \end{aligned}$$

Satz 4.3 *Ist M regulär, so hat die eindeutig bestimmte Lösung der Randwertaufgabe die Form*

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

mit der Greenschen Funktion

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

Definition 4.4
Die Zahlen

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

heißen Konditionszahlen *der Randwertaufgabe.*

Folgerung 4.5 *Es sei M regulär. 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*

$$\boxed{\|x\|_\infty \leq \kappa_1 \|d\| + \kappa_2 (T - t_0) \|q\|_\infty .}$$

D.h. die Randwertaufgabe ist korrekt gestellt in Abhängigkeit von den Daten (d, q).

4.2 Nichtlineare Randwertaufgaben

Wir betrachten nun Gleichungen der Form

$$\boxed{x'(t) = f(x(t), t), \quad t \in [t_0, T]} \tag{4.7}$$

mit $f : \mathbb{R}^m \times I_f \to \mathbb{R}^m$ stetig differenzierbar, $[t_0, T] \subseteq I_f \subseteq \mathbb{R}$, und den 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

$$\boxed{x'(t) = f(x(t), t), \quad t \in [t_0, T], \quad x(t_0) = x_0} \tag{4.9}$$

Sei

$$\boxed{S(x_0) := r(x_0, x(T, x_0)) \quad x_0 \in \mathbb{R}^m}$$

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 lineare *homogene* Randwertaufgabe betrachten:

$$\boxed{\begin{aligned} z'(t) &= A_*(t)z(t), \\ 0 &= B_{*0}z(t_0) + B_{*T}z(T), \end{aligned}} \tag{4.10}$$

mit

$$\begin{aligned} A_*(t) &:= \frac{\partial f}{\partial x}(x_*(t), t), \quad t \in [t_0, T], \\ B_{*0} &:= r_1'(x_*(t_0), x_*(T)), \quad B_{*T} := r_2'(x_*(t_0), x_*(T)). \end{aligned}$$

Hierbei sind r_1' und r_2' die partielle Frechét-Ableitungen von r nach der ersten bzw. zweiten Variable.

Die Fundamentalmatrix $X_*(t)$ zur DGL $z'(t) = A_*(t)\, z(t)$ ist Lösung von

$$X_*'(t) = A_*(t)X_*(t), \quad t \in [t_0, T], \quad X_*(t_0) = I,$$

und es sei

$$M_* := B_{*0} + B_{*T}X_*(T).$$

Wir formulieren die RWA (4.7) und (4.8) als Operatorgleichung im (reellen) Banachraum $C^1([t_0, T], \mathbb{R}^m)$ mit der Norm $\|x\| := \|x\|_\infty + \|x'\|_\infty$.

Lemma 4.6 *Die Abbildung* $F : C^1([t_0,T],\mathbb{R}^m) \to C([t_0,T],\mathbb{R}^m) \times \mathbb{R}^m$ *definiert durch*

$$F(x) := (x'(\cdot) - f(x(\cdot),\cdot), r(x(t_0),x(T))$$

ist Frechét-differenzierbar und für $x, z \in C^1([t_0,T],\mathbb{R}^m)$ *gilt*

$$\begin{aligned} F'(x)z &= (z'(\cdot) - A(\cdot)z(\cdot), B_0 z(t_0) + B_T z(T)) \\ A(t) &= \frac{\partial f}{\partial x}(x(t),t),\ t \in [t_0,T], \quad B_0 := r_1'(x(t_0),x(T)), \quad B_T := r_2'(x(t_0),x(T)), \end{aligned}$$

und $F'(x)$ *ist linear und beschränkt von* $C^1([t_0,T],\mathbb{R}^m)$ *in* $C([t_0,T],\mathbb{R}^m) \times \mathbb{R}^m$.
Für jedes $x \in C^1([t_0,T],\mathbb{R}^m)$ *und* $\varepsilon > 0$ *existiert ein* $\delta > 0$ *mit*

$$\tilde{x} \in C^1([t_0,T],\mathbb{R}^m)\,,\ \|x - \tilde{x}\| \le \delta \Longrightarrow \|F'(x) - F'(\tilde{x})\| \le \varepsilon\,.$$

Beweis: Seien $x, z \in C^1([t_0,T],\mathbb{R}^m)$. Dann gilt

$$\begin{aligned} \|\omega(x,z)\| &= \|F(x+z) - F(x) - F'(x)z\| \\ &= \max_{t\in[t_0,T]} \Big\| f(x(t),t) - f(x(t)+z(t),t) + \frac{\partial f}{\partial x}(x(t),t)z(t) \Big\| \\ &\quad + \|r(x(t_0)+z(t_0), x(T)+z(T)) - r(x(t_0),x(T)) \\ &\qquad - r_1'(x(t_0),x(T))z(t_0) - r_2'(x(t_0),x(T))z(T)\| \\ &\le \sup_{t\in[t_0,T]} \Big\| \int_0^1 \Big(\frac{\partial f}{\partial x}(x(t),t)z(t) - \frac{\partial f}{\partial x}(x(t)+\tau z(t),t)z(t) \Big) d\tau \Big\| \\ &\quad + \Big\| \int_0^1 (r_1'(x(t_0)+\tau z(t_0), x(T)+\tau z(T))z(t_0) - r_1'(x(t_0),x(T))z(t_0)) d\tau \Big\| \\ &\quad + \Big\| \int_0^1 (r_2'(x(t_0)+\tau z(t_0), x(T)+\tau z(T))z(T) - r_1'(x(t_0),x(T))z(T)) d\tau \Big\| \\ &\le \Big(\sup_{t\in[t_0,T]} \int_0^1 \Big\| \frac{\partial f}{\partial x}(x(t),t) - \frac{\partial f}{\partial x}(x(t)+\tau z(t),t) \Big\| d\tau \\ &\quad + \int_0^1 \Big\| r_1'(x(t_0)+\tau z(t_0), x(T)+\tau z(T)) - r_1'(x(t_0),x(T)) \Big\| d\tau \\ &\quad + \int_0^1 \Big\| r_2'(x(t_0)+\tau z(t_0), x(T)+\tau z(T)) - r_2'(x(t_0),x(T)) \Big\| d\tau \Big) \|z\|_\infty \end{aligned}$$

nach dem Mittelwertsatz im $\mathbb{R}^m$ (vgl. Heuser, Lehrbuch der Analysis, Teil 2, Kap. 167). Wegen der Stetigkeit der partiellen Ableitungen $\frac{\partial f}{\partial x}$, r_1' und r_2' existiert für jedes vorgegebene $\varepsilon > 0$ eine $\delta > 0$, so dass der Ausdruck in der Klammer kleiner als ε ist, falls $\|z\|_\infty \le \delta$, d.h.

$$\|\omega(x,z)\| \le \varepsilon \|z\|_\infty \le \varepsilon \|z\|, \quad \text{falls} \quad \|z\| \le \delta\,.$$

D.h. $\|\omega(x,z)\| \to 0$ falls $\|z\| \to 0$.
Offenbar ist $F'(x)$ eine lineare Abbildung von $C^1([t_0,T],\mathbb{R}^m)$ in $C([t_0,T],\mathbb{R}^m) \times \mathbb{R}^m$. Sie ist aber auch beschränkt wegen

$$\begin{aligned} \|F'(x)z\| &= \Big\| z' - \frac{\partial f}{\partial x}(x(\cdot),\cdot)z \Big\|_\infty + \|r_1'(x(t_0),x(T))z(t_0) + r_2'(x(t_0),x(T))z(T)\| \\ &\le \|z'\|_\infty + \max_{t\in[t_0,T]} \Big\| \frac{\partial f}{\partial x}(x(t),t) \Big\| \|z\|_\infty + \max_{i=1,2} \|r_i'(x(t_0),x(T))\| \|z\|_\infty \\ &\le K(\|z'\|_\infty + \|z\|_\infty) \end{aligned}$$

mit einer geeigneten Konstanten $K > 0$. Deshalb ist $F'(x)$ die Frechét-Ableitung von F in x. Außerdem gilt für $\tilde{x}, z \in C^1([t_0, T], \mathbb{R}^m)$ mit $\|z\| \leq 1$, dass

$$\begin{aligned}\|F'(x)z - F'(\tilde{x})z\| \leq & \max_{t\in[t_0,T]} \left\| \frac{\partial f}{\partial x}(x(t),t) - \frac{\partial f}{\partial x}(\tilde{x}(t),t) \right\| \\ & +2 \max_{i=1,2} \|r_i'(x(t_0),x(T)) - r_i'(\tilde{x}(t_0),\tilde{x}(T))\| \\ \leq & \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon\,, \quad \text{falls} \quad \|x - \tilde{x}\|_\infty \leq \delta\,,\end{aligned}$$

wobei $\delta \in (0, 1]$ aus der gleichmäßigen Stetigkeit von $\frac{\partial f}{\partial x}$ auf $B(0, \|x\|_\infty + 1) \times [t_0, T]$ und von r_i', $i = 1, 2$, auf $B(0, \|x\|_\infty + 1) \times B(0, \|x\|_\infty + 1)$ zu vorgegebenem $\frac{\varepsilon}{3}$ gewählt ist. □

Satz 4.7
Sei $x_* \in C^1([t_0, T], \mathbb{R}^m)$ eine Lösung der RWA (4.7), (4.8), und es sei die lineare homogene RWA (4.10) nur trivial lösbar. Dann gilt:

(a) x_* ist isolierte Lösung der RWA.

(b) Für hinreichend kleines $\delta > 0$ sind die gestörten RWAn

$$x'(t) = f(x(t), t) + q(t), \ \ t \in [t_0, T], \tag{4.11}$$
$$r(x(t_0), x(T)) = d, \tag{4.12}$$

eindeutig lösbar, falls $\|q\|_\infty + \|d\| \leq \delta$.
Die Lösung x der RWA hängt stetig von (q, d) ab.

(c) Alle AWA (4.9) mit $x_0 \in B(x_*(t_0), \delta)$, $\delta > 0$ hinreichend klein, besitzen Lösungen, die auf $[t_0, T]$ definiert sind. $x(t, x_0)$ ist stetig differenzierbar bzgl. x_0.

(d) Die Abbildung $S(x_0) := r(x_0, x(T, x_0))$ ist wohldefiniert auf $B(x_*(t_0), \delta)$ und dort stetig differenzierbar. Es gilt $S'(x_*(t_0)) = M_*$.

Beweis:

(a) Mit der Abbildung F aus Lemma 4.6 haben wir die Äquivalenz

$$F(x) = 0 \quad \Leftrightarrow \quad x \text{ erfüllt die RWA (4.7), (4.8)}$$

Nach Lemma 4.6 ist F stetig Frechét-differenzierbar. Die Gleichung

$$F'(x_*)z = 0$$

ist die Randwertaufgabe (4.10). Nach Voraussetzung ist diese nur trivial lösbar, d.h. $F'(x_*)z = 0 \Leftrightarrow z = 0$. Also ist $F'(x_*)$ injektiv. Mit der Lösbarkeitsbedingung (Satz 4.2) folgt dann die Regularität von M_* und damit die Bijektivität von $F'(x_*)$. Die Aussage folgt schließlich aus Lemma 2.22.

(b) 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

$$H(x,p) := F(x) - p.$$

Die Gleichung $H(x,p)=0$ entspricht der gestörten RWA (4.11), (4.12). Es ist

$$H(x_*,0) = F(x_*) = 0,$$

H ist Frechét-differenzierbar bzgl. x (nach Lemma 4.6), und es ist

$$H'_x(x_*,0) = F'(x_*) \quad \text{bijektiv.}$$

Nach dem Satz über implizite Funktionen im Banachraum $C^1 \times (C \times \mathbb{R}^m)$ existiert dann eine eindeutig bestimmte Funktion

$$\varphi : B_{C\times\mathbb{R}^m}(0,\delta) \to B_{C^1}(x_*,\rho), \quad \rho \le \delta,$$

mit der Eigenschaft

$$H(\varphi(p),p) = 0 \quad \text{für alle } p \in B(0,\delta).$$

Also hat die Gleichung $F(x)=(q,d)$ für $q \in C([t_0,T],\mathbb{R}^m)$, $d \in \mathbb{R}^m$, $\|q\|_\infty + \|d\| \le \delta$, genau eine Lösung $\varphi(q,d)$ in $B_{C^1}(x_*,\rho)$. Überdies ist φ stetig differenzierbar in $B(0,\delta)$.

(c) Wir untersuchen die ,,künstliche" RWA (4.7) mit der Randbedingung $x(t_0) = x_*(t_0)$:

$$\begin{aligned} x'(t) &= f(x(t),t), \quad t \in [t_0,T] \\ \hat{r}(x(t_0),x(T)) &= 0 \\ \hat{r}(u,v) &:= u - x_*(t_0), \quad u,v \in \mathbb{R}^m. \end{aligned}$$

Diese RWA hat die Lösung x_*. Es ist

$$\hat{r}'_1(u,v) \equiv E \quad \text{und} \quad \hat{r}'_2(u,v) \equiv 0,$$

also $\hat{B}_{*0} = E$, $\hat{B}_{*T} = 0$, $\hat{M}_* = E$, $\hat{A}_* = A_*$.

Nach (b) sind dann die gestörten ,,künstlichen" RWAn

$$\begin{aligned} x'(t) &= f(x(t),t) + q(t), \quad t \in [t_0,T], \\ \hat{r}(x(t_0),x(T)) &= d, \quad (\Leftrightarrow x(t_0) = x_*(t_0) + d) \end{aligned}$$

lokal eindeutig lösbar. Die Lösung $x = \hat{\varphi}(q,d)$ ist stetig differenzierbar bzgl. (q,d). Speziell ist für $q=0$ die Lösung $x = \hat{\varphi}(0,d)$ stetig differenzierbar bzgl. d. D.h. die AWA (4.9)

$$\begin{aligned} x'(t) &= f(x(t),t), \quad t \in [t_0,T], \\ x(t_0) &= x_*(t_0) + d, \end{aligned}$$

ist für $\|d\| \le \delta$ lokal eindeutig lösbar, die Lösung $x = \hat{\varphi}(0,d) \in C^1([t_0,T],\mathbb{R}^m)$ ist stetig differenzierbar bzgl. d.

(d) Nach (c) ist $S(x_0) = r(x_0, x(T, x_0))$ wohldefiniert für $x_0 \in B(x_*(t_0), \delta)$ und dort stetig differenzierbar. Es bleibt $S'(x_*(t_0)) = M_*$ zu zeigen. Es gilt

$$S'(x_0) = r_1'(x_0, x(T, x_0)) + r_2'(x_0, x(T, x_0)) \cdot \frac{\partial}{\partial x_0} x(T, x_0).$$

und die Berechnung erfordert die Lösung des linearen Matrizen-Anfangswertproblems

$$\frac{d}{dt}\frac{\partial}{\partial x_0} x(t, x_0) = \frac{\partial f}{\partial x}(x(t, x_0), t) \frac{\partial}{\partial x_0} x(t, x_0), \quad t \in [t_0, T], \quad \frac{\partial}{\partial x_0} x(t_0, x_0) = I.$$

Mit $x_0 = x_*(t_0)$, $x(t, x_0) = x_*(t)$, $X(t, x_0) := \frac{\partial}{\partial x_0} x(t, x_0)$ folgt

$$X'(t, x_*(t_0)) = \underbrace{\frac{\partial f}{\partial x}(x_*(t), t)}_{A_*(t)} X(t, x_*(t_0)), \quad X(t, x_*(t_0)) = I\,.$$

Diese AWA hat die Lösung $X_*(t) = X(t, x_*(t_0))$. Folglich gilt

$$\begin{aligned} S'(x_*(t_0)) &= r_1'(x_*(t_0), x_*(T)) + r_2'(x_*(t_0), x_*(T)) X_*(T) \\ &= B_{*0} + B_{*T} X_*(T) = M_*\,. \end{aligned}$$

□

4.3 Schießverfahren

Wir betrachten Randwertaufgaben der Form

$$\boxed{\begin{aligned} x'(t) &= f(x(t), t), \quad t \in [t_0, T] \\ r(x(t_0), x(T)) &= 0 \end{aligned}} \tag{4.13}$$

Sei $x_* \in C^1([t_0, T], \mathbb{R}^m)$ eine Lösung und sei M_* regulär.
Nach Satz 4.7 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 (4.13) zu lösen, müssen wir x_0 so bestimmen, dass das Paar $(x_0, x(T, x_0))$ die Randbedingung erfüllt, d.h. wir müssen die Gleichung

$$\boxed{S(x_0) = 0}$$

lösen.

Folgerung 4.8
Es sei $x_* \in C^1([t_0, T], \mathbb{R}^m)$ eine Lösung der RWA (4.13) und M_* sei regulär. Dann gilt für das nichtlineare Gleichungssystem

$$S(x_0) = 0\,,$$

dass eine Lösung $x_{*0} = x_*(t_0)$ existiert mit $S'(x_{*0}) = M_*$ regulär. Folglich können Newton-ähnliche Verfahren der Form

$$x_0^{n+1} := x_0^{(n)} - A_n^{-1} S(x_0^{(n)}), \quad n = 0, 1, 2, \ldots,$$

mit $\|x_0^{(0)} - x_{*0}\| \leq \delta$ und hinreichend kleinem $\delta > 0$ angewendet werden.

Es sei angemerkt, dass die Matrix M_* sehr schlecht konditioniert und $\delta > 0$ außerordentlich klein sein kann.
Zur Berechnung von $S(x_0)$ muss $x\,(T, x_0)$ bestimmt und folglich ein Anfangswertproblem gelöst werden. Dies wird praktisch mit für die Anfangswertaufgabe geeigneten Integrationsverfahren (aus Kap. 3) realisiert. Da die Berechnung von Funktions- und Ableitungswerten also aufwendig ist, sollten Methoden mit sehr guten Konvergenzeigenschaften verwendet werden. Dies spricht für Newton-ähnliche Verfahren mit überlinearer Konvergenz. Da Ableitungen einen besonderen Aufwand darstellen (m Anfangswertprobleme !), erscheinen Quasi-Newton-Verfahren als besonders geeignet.

Häufig hängen aber die Lösungen $x(\cdot, x_0)$ sehr empfindlich von x_0 ab, wie das folgende Beispiel verdeutlicht.

Beispiel 4.9 *(Stoer-Bulirsch)*
Wir betrachten die Randwertaufgabe

$$x'(t) = \begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix} x(t), \quad t \in [0, 10],\ x_1(0) = 1,\ x_1(10) = 1,$$

mit dem charakteristischen Polynom $p(\lambda) = \lambda(\lambda - 1) - 110$ und den Eigenwerten $\lambda_1 = -10$ und $\lambda_2 = 11$.
Die Lösung $x(t, z)$ der Differentialgleichung, die der Anfangsbedingung $x(0, z) = z$ genügt, hat dann die Form

$$x(t, z) = \frac{11z_1 - z_2}{21} e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + \frac{10z_1 + z_2}{21} e^{11t} \begin{pmatrix} 1 \\ 10 \end{pmatrix}.$$

Die Lösung $x_*(\cdot)$ zu den Randbedingungen besitzt den Anfangswert

$$z_* := x_*(0) = \begin{pmatrix} 1 \\ -10 + 21 \frac{1 - e^{-100}}{e^{110} - e^{-100}} \end{pmatrix}.$$

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 von Satz 1.11 erfüllen, gilt die Abschätzung

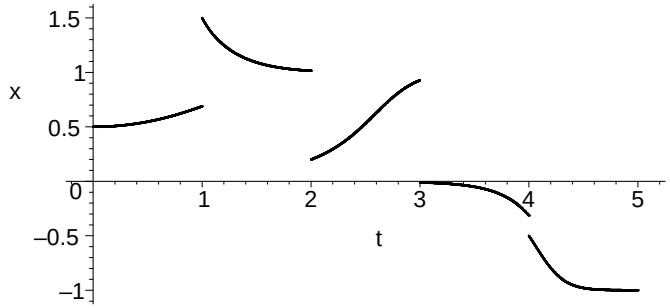
$$\|x(t, x_0) - x(t, \bar{x}_0)\| \leq \exp(\gamma(t - t_0))\|x_0 - \bar{x}_0\|,$$

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.

Wir zerlegen $[t_0, T]$ in Teilintervalle $t_0 < t_1 < \ldots < t_N = T$. Die Werte $x(t_j)$ der exakten Lösung $x(t)$ der Randwertaufgabe 4.13 an den Stellen $t_0, \ldots, t_n$ sollen nun gleichzeitig iterativ berechnet werden.. Wir betrachten dazu die Anfangswertaufgaben

$$\boxed{\begin{aligned} x'(t) &= f(x(t), t) \quad t \in [t_{j-1}, t_j], \\ x(t_{j-1}) &= z_{j-1}, \end{aligned} \qquad (j = 1, \ldots, N).}$$

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 zusammengesetzte Funktion $x(t)$,

$$\begin{aligned} x(t) &:= x(t, z_{j-1}, t_{j-1}), \quad t \in [t_{j-1}, t_j), \quad j = 1, \ldots, N, \\ x(t_N) &:= x(t_N, z_{N-1}, t_{N-1}), \end{aligned}$$

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 Stetigkeitsbedingungen

$$\boxed{x\left(t_j, z_{j-1}, t_{j-1}\right) = z_j, \quad j = 1, \ldots, N-1}$$

werden die Segmente zu einer Funktion auf $[t_0, T]$ zusammengesetzt. Diese Funktion ist automatisch stetig differenzierbar und eine Lösung wegen der Erfülltheit der Gleichungen.
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 der Schießabbildung

$$\boxed{S(z) := \begin{pmatrix} x(t_1, z_0, t_0) - z_1 \\ \vdots \\ x(t_{N-1}, z_{N-2}, t_{N-2}) - z_{N-1} \\ r(z_0, x(t_N, z_{N-1}, t_{N-1})) \end{pmatrix}, \quad z = \begin{pmatrix} z_0 \\ \vdots \\ z_{N-1} \end{pmatrix}.}$$

Sei x_* Lösung der RWA (4.13) 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 Newton-ähnliche Verfahren anwenden zu können, ist die Regularität von $S'(z_*)$ wünschenswert. Definiere dazu

$$\begin{aligned} X_j(t) &:= \frac{\partial}{\partial z_j} x(t; z_j, t_j), \quad j = 0, \ldots, N-1, \\ X_{*j}(t) &:= \frac{\partial}{\partial z_j} x(t; z_j, t_j)\Big|_{z_j = z_{*j}}, \quad j = 0, \ldots, N-1. \end{aligned}$$

Wie wir im Beweis von Satz 4.7(d) gesehen haben, ist

$$X_j'(t) = \frac{\partial f}{\partial x}(x(t, z_j, t_j), t) X_j(t), \quad X_j(t_j) = I, \quad j = 0, \ldots, N-1,$$

und

$$X_{*j}'(t) = A_*(t) X_{*j}(t), \quad X_{*j}(t_j) = I, \quad j = 0, \ldots, N-1.$$

Dann ist $S'(z) =$

$$\begin{pmatrix} \frac{\partial}{\partial z_0}(x(t_1, z_0, t_0) - z_1) & \frac{\partial}{\partial z_1}(x(t_1, z_0, t_0) - z_1) & 0 & \cdots & 0 \\ 0 & \frac{\partial}{\partial z_1}(x(t_2, z_1, t_1) - z_2) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{\partial}{\partial z_{N-2}}(\cdots) & \frac{\partial}{\partial z_{N-1}}(x(t_{N-1}, z_{N-2}, t_{N-2}) - z_{N-1}) \\ \frac{\partial}{\partial z_0} r(z_0, x(t_N, z_{N-1}, t_{N-1})) & 0 & \cdots & 0 & \frac{\partial}{\partial z_{N-1}} r(z_0, x(t_N, z_{N-1}, t_{N-1})) \end{pmatrix}$$

$$= \begin{pmatrix} X_0(t_1) & -I & 0 & \cdots & 0 \\ 0 & X_1(t_2) & -I & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \vdots & \vdots & X_{N-2}(t_{N-1}) & -I \\ r_1'(z_0, x(t_N, z_{N-1}, t_{N-1})) & 0 & \cdots & 0 & r_2'(z_0, x(t_N, z_{N-1}, t_{N-1})) X_{N-1}(t_N) \end{pmatrix}$$

und speziell für $z = z_*$:

$$
\begin{aligned}
S(z_*) &= 0\,, \\
S'(z_*) &= \begin{pmatrix}
X_{*0}(t_1) & -I & 0 & \cdots & 0 \\
0 & X_{*1}(t_2) & -I & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & X_{*N-2}(t_{N-1}) & -I \\
B_{*0} & 0 & \cdots & 0 & B_{*T}X_{*N-1}(t_N)
\end{pmatrix}.
\end{aligned}
$$

Sei $y = (y_0, \ldots, y_{N-1})^T \in \mathbb{R}^{mN}$ mit

$$S'(z_*)y = 0\,.$$

Für die Regularität von $S'(z_*)$ müssen wir $y = 0$ zeigen. Aus $S'(z_*)y = 0$ folgt

$$
\begin{aligned}
X_{*0}(t_1)y_0 &= y_1, \\
&\vdots \\
X_{*N-2}(t_{N-1})y_{N-2} &= y_{N-1}, \\
B_{*0}y_0 + B_{*T}X_{*N-1}(t_N)y_{N-1} &= 0\,,
\end{aligned}
$$

und die letzte Zeile wird zu

$$B_{*0}y_0 + B_{*T}X_{*N-1}(t_N)\cdots X_{*0}(t_1)y_0 = 0\,. \tag{4.14}$$

Es galt:
$X'_*(t) = A_*(t)X_*(t)$, $t \in [t_0, T]$, $X_*(t_0) = E$. 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

$$\underbrace{X_{*1}(t_1)}_{=E}X_{*0}(t_1) = X_{*0}(t_1) = X_*(t_1)\,.$$

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 (4.14) 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 $y_1 = 0$, $y_2 = 0$, ..., $y_{N-1} = 0$, also $y = 0$. D.h. $S'(z_*)$ ist regulär.

Beispiel 4.10
Sei $m = 2$ und wir betrachten das lineare Randwertproblem

$$\begin{aligned} x'(t) &= \begin{pmatrix} 1 - 19\cos(2t) & 19\sin(2t) \\ 19\sin(2t) & 1 + 19\cos(2t) \end{pmatrix} x(t) + q(t)\,,\ t \in [0, \pi], \\ x(0) + x(\pi) &= \begin{pmatrix} 1 + e^{\pi} \\ 1 + e^{\pi} \end{pmatrix}, \end{aligned}$$

mit ,,passender" Funktion q. Dann ist

$$B_0 = E, \quad B_T = E, \quad \kappa_1, \kappa_2 \approx 2\,.$$

$$q(t) = \begin{pmatrix} 19e^t(\sin(2t) - \cos(2t)) \\ 19e^t(\sin(2t) + \cos(2t)) \end{pmatrix}$$

$$\begin{aligned} \varphi(\lambda) = \det(A(t) - \lambda I) &= \begin{pmatrix} 1 - 19\cos(2t) - \lambda & 19\sin 2t \\ 19\sin 2t & 1 + 19\cos(2t) - \lambda \end{pmatrix} \\ &= 1 - 19^2(\cos^2(2t) + \sin^2(2t)) - 2\lambda + \lambda^2 \\ &= -360 - 2\lambda + \lambda^2 \\ \lambda_{1|2} &= 1 \pm \sqrt{1 + 360} = 1 \pm 19 \end{aligned}$$

Die Lösung des Randwertproblems ist

$$x_*(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}$$

und die Fundamentalmatrix ist

$$X(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} e^{-18t} & 0 \\ 0 & e^{20t} \end{pmatrix} \quad \textit{mit} \quad X(0) = E\,.$$

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

$$cond_1(M) = \|M\|_1 \|M^{-1}\|_1 \approx e^{20\pi} \approx 1.9 \cdot 10^{27}.$$

Für die Konditionszahlen von $S'(z_*) = M$ erhält man

N	$\left\|S'(z_*)^{-1}\right\|$	$cond(S'(z_*))$
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	$1.6 \cdot 10^{1}$

Nachdem sich am Beispiel verbesserte Konditionszahlen für $S'(z_*)$ für wachsendes N gezeigt hatten, stellt sich die Frage nach einem allgemeinen Resultat für diese Konditionszahlen.

Satz 4.11 *Sei $x_* \in C^1([t_0,T],\mathbb{R}^m)$ Lösung der Randwertaufgabe (4.13), M_* sei regulär. Es gelte $\|B_{*0}\| \le 1$ und $\|B_{*T}\| \le 1$. 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

$$\boxed{\operatorname{cond}(S'(z_*)) \le (\kappa_1 + (N-1)\kappa_2) \cdot \left(1 + \max_{j=1,\ldots,N} \|X_{*j-1}(t_j)\|\right)}$$

bzgl. der Block-Zeilensummen-Norm auf $\mathbb{R}^{mN\times mN}$.

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))

$$\mathcal{G}_*(t,s) = \begin{cases} X_*(t)M_*^{-1}B_{*0}X_*(t_0)X_*(s)^{-1} & s \le t, \\ -X_*(t)M_*^{-1}B_{*T}X_*(T)X_*(s)^{-1} & s > t, \end{cases}$$

denn es gilt blockweise

$$\begin{aligned}
(S'(z_*)S'(z_*)^{-1})_{11} &= -X_{*0}(t_1)\mathcal{G}_*(t_0,t_1) + \mathcal{G}_*(t_1,t_1) \\
&= \underbrace{X_{*0}(t_1)X_*(t_0)}_{X_*(t_1)} M_*^{-1}B_{*T}X_*(T)X_*(t_1)^{-1} + X_*(t_1)M_*^{-1}B_{*0}X_*(t_0)X_*(t_1)^{-1} \\
&= X_*(t_1)M_*^{-1}(B_{*T}X_*(T) + B_{*0}X_*(t_0))X_*(t_1)^{-1} \\
&= I \\
(S'(z_*)S'(z_*)^{-1})_{12} &= -X_{*0}(t_1)\mathcal{G}_*(t_0,t_2) + \mathcal{G}_*(t_1,t_2) \\
&= X_{*0}(t_1)X_*(t_0)M_*^{-1}B_{*T}X_*(T)X_*(t_2)^{-1} - X_*(t_1)M_*^{-1}B_{*T}X_*(T)X_*(t_2)^{-1} \\
&= 0
\end{aligned}$$

usw. Nach Definition von κ_1 und κ_2 (vgl. (4.6)) ist

$$\begin{aligned} \left\|X_*(t)M_*^{-1}\right\| &\le \kappa_1 \quad t \in [t_0, T] \\ \|\mathcal{G}_*(t,s)\| &\le \kappa_2 \quad t, s \in [t_0, T] \end{aligned}$$

Damit ist also

$$\left\|S'(z_*)^{-1}\right\| = \max_{i=0,\ldots,N-1} \Big(\sum_{j=1}^{N-1} \|\mathcal{G}_*(t_i,t_j)\| + \left\|X_*(t_i)M_*^{-1}\right\| \Big) \le (N-1)\kappa_2 + \kappa_1$$

und

$$\begin{aligned} \|S'(z_*)\| &= \max \Big\{ \max_{i=0,\ldots,N-2} (1 + \|X_{*i}(t_{i+1})\|) \,, \underbrace{\|B_{*0}\|}_{\le 1} + \underbrace{\|B_{*T}X_{*N-1}(t_N)\|}_{\le \|X_{*N-1}(t_N)\|} \Big\} \\ &\le \max_{i=0,\ldots,N-1} \Big(1 + \|X_{*i}(t_{i+1})\| \Big). \end{aligned}$$

□

Obwohl sich durch den Term $(N-1)\kappa_2$ eine gewisse Vergrößerung der Konditionszahl mit wachsendem N ergibt, besteht der entscheidende Gewinn in der Ersetzung von $\|X_*(T)\|$ durch $\max_{i=0,\ldots,N-1} \|X_{*i}(t_{i+1})\|$.

Bemerkung 4.12 *Zur Abschätzung von $\|X_{*j-1}(t_j)\|$ definiere*

$$\begin{aligned} \mu_j &:= \max_{t\in[t_j,t_{j+1}]} \|A_*(t)\| \quad (j = 0, \ldots, N-1), \\ \mu &:= \max_{t\in[t_0,T]} \|A_*(t)\| = \max\{\mu_0, \ldots, \mu_{N-1}\}, \end{aligned}$$

d.h. $\mu_j \le \mu$. Wegen

$$X'_{*j-1}(t) = A_*(t)X_{*j-1}(t), \quad t \in [t_{j-1}, t_j], \quad X_{*j-1}(t_{j-1}) = I,$$

ist dann

$$\|X_{*j-1}(t_j)\| \le e^{\mu_{j-1}(t_j - t_{j-1})}.$$

Im Spezialfall $h = \frac{T-t_0}{N}$, $t_j = t_0 + jh$, $\mu_j \le \mu$, folgt dann

$$cond(S'(z_*)) \le (\kappa_1 + (N-1)\kappa_2)(1 + e^{\mu h}) \le (\kappa_1 + \frac{T-t_0}{h}\kappa_2)(1 + e^{\mu h}).$$

Im Idealfall möchte man $\|X_{*j}(t_{j+1})\| \le \kappa$, $j = 0, \ldots, N-1$, mit $\kappa \approx 10$ oder $\kappa \approx 100$ erreichen.

4.4 Kollokationsverfahren

Wir betrachten wieder die nichtlineare Randwertaufgabe (4.7), (4.8) und nehmen an, dass sie eine Lösung x^* besitzt, und der Einfachheit halber, dass $m = 1$ gilt, d.h. die Differentialgleichung skalar ist. Am Schluss des Kapitels diskutieren wir kurz die Erweiterung der Ergebnisse auf $m > 1$. Wir betrachten die Zerlegungen

$$Z_n : t_0 < t_1^{(n)} < \cdots < t_{N_n}^{(n)} = T \quad (n \in \mathbb{N})$$

des Intervalls $[t_0, T]$ mit $h_n := \max\limits_{i=1,\ldots,N_n} h_j^{(n)}$, $h_j^{(n)} := t_j^{(n)} - t_{j-1}^{(n)}$, $i = 1, \ldots, N_n$, und $t_{j-\frac{1}{2}}^{(n)} := t_{j-1}^{(n)} + \frac{1}{2} h_j^{(n)}$ (Mittelpunkt von $[t_{j-1}^{(n)}, t_j^{(n)}]$). Nachfolgend werden wir den Index n an den Gitterpunkten und an N meist weglassen.
Wir betrachten wieder den Banachraum $X := C^1([t_0, T])$ mit der Norm $\|x\| := \|x\|_\infty + \|x'\|_\infty$ wie vor Lemma 4.6 sowie für $n \in \mathbb{N}$ die Teilräume

$$X_n := \{u \in C^1([t_0, T]) : u \text{ ist ein kubischer Spline auf } Z_n, \text{ d.h. } u \text{ ist auf jedem Teilintervall von } Z_n \text{ ein Polynom vom Grad } \leq 3\}.$$

Jedes $u \in X_n$ ist eindeutig bestimmt, falls

$$u_i := u(t_i),\ u'(t_i) := u_i', \quad i = 0, 1, \ldots, N,$$

vorgegeben sind.
Ansatz:

$$u(t) = u_{i-1} + (t - t_{i-1}) \frac{u_i - u_{i-1}}{h_i} + (t - t_{i-1})(t - t_i) v_i + (t - t_{i-1})^2 (t - t_i) w_i \quad (t \in [t_{i-1}, t_i])$$

Die Koeffizienten v_i, w_i werden aus den Bedingungen

$$u'(t_i) = u_i', \quad i = 0, 1, \ldots, N,$$

bestimmt. Es gilt nach Ansatz

$$\begin{aligned}
u'(t) &= \tfrac{u_i - u_{i-1}}{h_i} + (t - t_i) v_i + (t - t_{i-1}) v_i + 2(t - t_{i-1})(t - t_i) w_i + (t - t_{i-1})^2 w_i \\
\rightsquigarrow u_{i-1}' &= \tfrac{u_i - u_{i-1}}{h_i} - h_i v_i \quad \text{oder} \quad v_i = \tfrac{1}{h_i} \left(\tfrac{u_i - u_{i-1}}{h_i} - u_{i-1}' \right) \\
u_i' &= \tfrac{u_i - u_{i-1}}{h_i} + h_i v_i + h_i^2 w_i \\
\rightsquigarrow w_i &= \tfrac{1}{h_i^2} \left(u_i' - 2 \tfrac{u_i - u_{i-1}}{h_i} + u_{i-1}' \right)
\end{aligned}$$

Wir definieren nun den Restriktionsoperator $r_n; X \to X_n$ durch $r_n x(t_i) = x(t_i)$ und $(r_n x)'(t_i) = x'(t_i)$ für $i = 0, \ldots, N$. Dann ist r_n linear und ein Projektor.

Wir betrachten weiterhin eine Diskretisierung für $Y = C([t_0, T]) \times \mathbb{R}$ mit der Norm $\|(v, a)\| = \|v\|_\infty + |a|$, nämlich die Teilräume $Y_n := V_n \times \mathbb{R}$, wobei

$$V_n := \{v \in C([t_0, T]) : v \text{ ist stückweise quadratisch auf } Z_n\} \quad (n \in \mathbb{N}).$$

$v \in V_n$ ist eindeutig bestimmt durch $v(t_i) = v_i$, $i = 0, \ldots, N$, und $v(t_{i-\frac{1}{2}}) = v_{i-\frac{1}{2}}$, $i = 1, \ldots, N$. Für $t \in [t_{i-1}, t_i]$ besitzt $v(\cdot)$ die Darstellung

$$(*) \quad v(t) = v_{i-1} + (t - t_{i-1}) \frac{v_{i-\frac{1}{2}} - v_{i-1}}{\frac{1}{2} h_i} + (t - t_{i-1})(t - t_{i-\frac{1}{2}}) \frac{v_i - 2v_{i-\frac{1}{2}} + v_{i-1}}{\frac{1}{2} h_i^2}$$

Restriktionsoperator: $\tilde{r}_n : Y \to Y_n$

$$\tilde{r}_n(v, a) = (v_n, a), \text{ wobei } \begin{array}{ll} v_n(t_i) = v(t_i), & i = 0, \ldots, N, \\ v_n(t_{i-\frac{1}{2}}) = v(t_{i-\frac{1}{2}}), & i = 1, \ldots, N. \end{array}$$

$\tilde{r}_n$ ist ebenfalls ein linearer Operator.

Lemma 4.13
Falls $h_n \to 0$, so sind (X, X_n, r_n) und $(Y, Y_n, \tilde{r}_n)$ diskrete Approximationen und es gilt

$$\lim_{n\to\infty} \|r_n x - x\| = 0 \quad (\forall x \in X) \text{ und } \lim_{n\to\infty} \|\tilde{r}_n(v, a) - (v, a)\| = 0 \quad (\forall (v, a) \in Y).$$

Die diskrete Konvergenz entspricht der Konvergenz in X bzw. Y.

Beweis: Die Funktionen $r_n x$ stellen interpolierende kubische Splines mit Hermite-Randbedingungen dar. Deshalb konvergieren sie und ihre Ableitungen gleichmäßig gegen x bzw. deren Ableitung x', falls sie in Gitterpunkten $t_j^{(n)}$, $j = 0, \ldots, N_n$, $n \in \mathbb{N}$, interpolieren mit $h_n \to 0$ (vgl. Hämmerlin-Hoffmann, Kap. 6.5.4). Die erste Komponente von $\tilde{r}_n(v, a)$ stellt einen interpolierenden quadratischen Spline dar, der gleichmäßig gegen v konvergiert (vgl. Powell: Approximation theory, Cambridge Univ. Press 1991). □

Der Operator $F : X \longrightarrow Y$ wird wie in Lemma 4.6 definiert durch

$$F(x) := (x'(\cdot) - f(x(\cdot), \cdot),\ r(x(t_0), x(T))) .$$

Diskretisierung des Operators:

$$F_n : X_n \longrightarrow Y_n , \quad F_n u = (v, r(u(t_0), u(t_N))),$$

wobei $v \in V_n$ die Eigenschaft hat

$$\begin{array}{rll} v(t_i) = & u'(t_i) - f(u(t_i), t_i) , & i = 0, \ldots, N, \\ v(t_{i-\frac{1}{2}}) = & u'(t_{i-\frac{1}{2}}) - f(u(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) , & i = 1, \ldots, N. \end{array}$$

Diskretisierte Operatorgleichung:

$$F_n x_n = 0 \tag{4.15}$$

bedeutet $x_n \in X_n$ mit $r(x_n(t_0), x_n(t_N)) = 0$ und x_n genügt der Differentialgleichung in den Diskretisierungspunkten

$$t_0, t_{\frac{1}{2}}, t_1, t_{\frac{3}{2}}, t_2, \ldots, t_{N-1}, t_{N-\frac{1}{2}}, t_N \quad \text{(Kollokation)}.$$

Die Gleichung (4.15) ist deshalb äquivalent zu $(2N+2)$ Gleichungen für die Unbekannten $x_n(t_i)$, $x_n'(t_i)$, $i = 0, \ldots, N$, wodurch $x_n \in X_n$ eindeutig bestimmt ist.

Nach Lemma 4.6 ist F Frechét-differenzierbar auf X und $F'(x^*)$ ist injektiv (d.h. $N(F'(x^*)) = \{0\}$), falls die "Lösbarkeitsmatrix" $M_* \neq 0$ ist (vgl. Satz 4.7).

Lemma 4.14 *Die Operatoren $F_n : X_n \longrightarrow Y_n$ sind Frechét-differenzierbar auf X_n und die Operatoren $F_n'(r_n x^*)$ sind fredholmsch mit ind$(F_n'(r_n x^*)) = 0$. Überdies gilt: $\forall \varepsilon > 0$ $\exists \delta > 0$ so dass $\forall n \in \mathbb{N}$*

$$\|x_n - r_n x^*\| \leq \delta \Rightarrow \|F_n'(x_n) - F_n'(r_n x^*)\| \leq \varepsilon\,.$$

Beweis: Nach Definition des Operators F_n gilt analog zu Lemma 4.6 für dessen Frechét-Ableitung in $u \in X_n$

$$F_n'(u)z = (v, r_1'(u(t_0), u(T))z(t_0) + r_2'(u(t_0), u(T))z(T)) \quad (\forall z \in X_n),$$

wobei $v \in V_n$ die Funktion mit der Eigenschaft

$$v(t_j) = z'(t_j) - \frac{\partial f}{\partial x}(u(t_j), t_j)z(t_j) \quad (j = 0, \frac{1}{2}, 1, \ldots, N - \frac{1}{2}, N).$$

ist. Wegen der Gestalt von v (siehe (*)) gilt

$$\|v\|_\infty \leq 9 \max\{|v(t_j)|, |v(t_{i-\frac{1}{2}}| : j = 0, 1, \ldots, N, i = 1, \ldots, N\}$$

Deshalb ist $F_n'(u)$ linear und auch beschränkt wegen

$$\|F_n'(u)z\|_\infty \leq 9\|z'\|_\infty + \Big(9 \max_{t\in[t_0,T]} \Big|\frac{\partial f}{\partial x}(u(t), t)\Big| + 2 \max_{i=1,2} |r_i'(u(t_0), u(T))|\Big)\|z\|_\infty.$$

Die Ableitungen $F_n'(u)$ fredholmsch mit ind $(F_n'(u)) = \dim X_n - \dim Y_n = 0$. Ferner gilt für die beiden Komponenten von $F_n'(u)z - F_n'(\tilde{u})z$ für $\|z\|_\infty \leq 1$:

$$\begin{aligned} \|[F_n'(u)z - F_n'(\tilde{u})z]_1\|_\infty &\leq \max_{t\in[t_0,T]} \Big|\frac{\partial f}{\partial x}(u(t), t) - \frac{\partial f}{\partial x}(\tilde{u}(t), t)\Big| \\ |[F_n'(u)z - F_n'(\tilde{u})z]_2| &\leq 2 \max_{i=1,2} |r_i'(u(t_0), u(T)) - r_i'(\tilde{u}(t_0), \tilde{u}(T))| \end{aligned}$$

Wegen der gleichmäßigen Stetigkeit von $\frac{\partial f}{\partial x}$ und r_i', $i = 1, 2$, auf kompakten Mengen ist deshalb für gegebenes u und kleines $\|u - \tilde{u}\|_\infty$ auch $\|F_n'(u) - F_n'(\tilde{u})\|$ klein. □

Lemma 4.15 *Es gelte $h_n \to 0$.*
Die Folge $(F_n'(r_n x^*))$ konvergiert regulär gegen $F'(x^*)$, d.h. $F_n'(r_n x^*) \xrightarrow{d} F'(x^*)$ und $\|x_n\| \leq$ const und $(F_n'(r_n x^*)x_n)$ diskret konvergent $\Rightarrow (x_n)$ ist diskret kompakt.

Beweis: Für die Operatoren $F_n'(r_n x^*)$ gilt für jedes $z \in X_n$ mit $\|z\| \leq 1$

$$\begin{aligned} F_n'(r_n x^*)z &= (v_n, B_{*0}z(t_0) + B_{*T}z(T)) \\ \|F_n'(r_n x^*)z\| &\leq \|v_n\|_\infty + |B_{*0}| + |B_{*T}|, \end{aligned}$$

wobei v_n der quadratische Spline in V_n ist mit

$$\begin{aligned} v_n(t_i) &= z'(t_i) - A_*(t_i)z(t_i) =, \ i = 0, 1, \ldots, N, \\ v_n(t_{i-\frac{1}{2}}) &= z'(t_{i-\frac{1}{2}}) - A_*(t_{i-\frac{1}{2}})z(t_{i-\frac{1}{2}}), \ i = 1, \ldots, N, \end{aligned}$$

und $u_n = r_n x^*$ der kubische Spline mit $u_n(t_j) = x^*(t_j)$ und $u_n'(t_j) = x'^*(t_j)$ für alle $j = 0, \ldots, N$, ist. Da u_n in X konvergiert und $A_*(\cdot)$ stetig ist, existiert eine Konstante

$c > 0$, so dass $\max\{|v(t_j)|, |v(t_{i-\frac{1}{2}}| : j = 0, 1, \ldots, N, i = 1, \ldots, N\} \leq c$. Wegen (*) gilt dann für $\|v_n\|_\infty$, dass

$$\|v_n\|_\infty \leq 9c\,.$$

Folglich ist die Norm $\|F_n'(r_n x^*)\|$ gleichmäßig beschränkt und es genügt nach Satz 2.13 für $F_n'(r_n x^*) \xrightarrow{d} F'(x^*)$ zu zeigen, dass $\|F_n'(r_n x^*) r_n z - \tilde{r}_n F'(x^*) z\| \to 0$ falls $n \to \infty$. Dies ist nicht schwierig zu zeigen (Übung).
Ist schließlich $x_n \in X_n$ mit $\|x_n\| \leq C = \text{const}$ für alle $n \in \mathbb{N}$ und (in Y) konvergenter Folge $(F_n'(r_n x^*) x_n)$, so gilt $\|x_n\|_\infty \leq C$ und $|x_n(t) - x_n(s)| \leq |t - s|$ für alle $t, s \in [t_0, T]$ und $n \in \mathbb{N}$. Dann ist (x_n) zunächst relativ kompakt in $C([t_0, T])$ mit $\|\cdot\|_\infty$ nach dem Satz von Arzela-Ascoli. Die erste Komponente von $F_n'(r_n x^*) x_n$ ist $v_n \in V_n$ mit

$$v_n(t_j) = x_n'(t_j) - \frac{\partial f}{\partial x}(r_n x^*(t_j), t_j) x_n(t_j) \quad (j = 0, \frac{1}{2}, 1, \ldots, N - \frac{1}{2}, N).$$

Da die Folge $(\frac{\partial f}{\partial x}(r_n x^*(t_j), t_j) x_n(t_j))$ für jedes j eine konvergente Teilfolge besitzt und (v_n) gleichmäßig konvergiert, muss auch die Folge $(x_n'(t_j))$ für jedes j eine konvergente Teilfolge besitzen. Daraus folgt schließlich, dass auch die Folge (x_n') eine Teilfolge besitzt, die gleichmäßig konvergiert. □

Satz 4.16 *(Konvergenzsatz für Kollokationsmethoden)*
Es seien f und r stetig differenzierbar, x^* sei eine Lösung der nichtlinearen RWA und M_* sei regulär. Es gelte $h_n \to 0$.
Dann existieren ein $n_0 \in \mathbb{N}$ und ein $\delta_0 > 0$, so daß die Gleichungen

$$F_n x_n = 0 \quad (n \geq n_0)$$

eine lokal eindeutige Lösung x_n^* besitzen und

$$x_n^* \longrightarrow x^* \quad (\text{in } X)$$

sowie

$$C_1 \|F_n r_n x^*\| \leq \|x_n^* - r_n x^*\| \leq C_2 \|F_n r_n x^*\| = O(h_n)$$

Beweis: Unser Ziel ist die Anwendung von Satz 2.20. Wegen der Lemmata 4.14 und 4.15 sind neben (i)–(iv) auch die Voraussetzungen $(v), (vi)$ und (vii) von Satz 2.20 erfüllt. Es bleibt $(viii)$ zu zeigen, nämlich

$$F_n r_n x^* \xrightarrow{d} F x^* = 0\,.$$

Dann würde die Aussage aus Satz 2.20 folgen.
Wir betrachten zunächst $u_n := r_n x^*$. Nach Definition ist u_n der kubische Spline, für den $u_n(t_j) = x^*(t_j)$ und $u_n'(t_j) = x'^*(t_j)$, $j = 0, \ldots, N$, gilt.

$$\begin{aligned} F_n r_n x^* = F_n u_n &= (v_n, r(u_n(t_0), u_n(t_N))) = (v_n, r(x^*(t_0), x^*(t_N))) \\ &= (v_n, 0)\,, \end{aligned}$$

wobei $v_n \in V_n$ der quadratische Spline mit

$$\begin{aligned} v_n(t_i) &= u_n'(t_i) - f(u_n(t_i), t_i) = x'^*(t_i) - f(x^*(t_i), t_i) = 0\,,\ i = 0, 1, \ldots, N, \\ v_n(t_{i-\frac{1}{2}}) &= u_n'(t_{i-\frac{1}{2}}) - f(u_n(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}),\ i = 1, \ldots, N. \end{aligned}$$

Zu zeigen ist folglich

$$v_n \to 0 \ \text{ d.h. } \max_{t\in[t_0,T]} |v_n(t)| \to 0 \ \text{ für } \ n \to \infty\,.$$

Wir zeigen zunächst, daß gilt: $v_n(t_{i-\frac{1}{2}}) \to 0$

$$\begin{aligned} v_n(t_{i-\frac{1}{2}}) &= u_n'(t_{i-\frac{1}{2}}) - f(u_n(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) \\ &= \frac{u_n(t_i) - u_n(t_{i-1})}{h_i} - \frac{1}{2}h_i v_i + \frac{1}{2}h_i v_i - 2\frac{h_i^2}{4}w_i + \frac{h_i^2}{4}w_i - f(u_n(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) \\ &= \frac{u_n(t_i) - u_n(t_{i-1})}{h_i} - \frac{h_i^2}{4}w_i - f(u_n(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) \end{aligned}$$

wegen der Darstellung für Elemente aus X_n bzw. deren Ableitungen. Es ergibt sich

$$\begin{aligned} v_n(t_{i-\frac{1}{2}}) &= \frac{u_n(t_i) - u_n(t_{i-1})}{h_i} - \frac{h_i^2}{4}\frac{1}{h_i^2}\Big(u_n'(t_{i-1}) - 2\frac{u_n(t_i) - u_n(t_{i-1})}{h_i} + u_n'(t_i)\Big) \\ &\quad -f(u_n(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) \\ &= \frac{3}{2}\left(\frac{u_n(t_i) - u_n(t_{i-1})}{h_i}\right) - \frac{1}{4}(u_n'(t_i) + u_n'(t_{i-1})) - f(u_n(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) \\ &= \frac{3}{2}\frac{x^*(t_i) - x^*(t_{i-1})}{h_i} - \frac{1}{4}(x'^*(t_i) + x'^*(t_{i-1})) \\ &\quad -f\Big(\underbrace{\frac{1}{2}(x_*(t_{i-1}) + x_*(t_i)) - \frac{1}{8}h_i(x_*'(t_i) - x_*'(t_{i-1}))}_{=u_n(t_{i-\frac{1}{2}})}, t_{i-\frac{1}{2}}\Big) \end{aligned}$$

und deshalb

$$\begin{aligned} \rightsquigarrow v_n(t_{i-\frac{1}{2}}) &= \frac{3}{2}(x_*'(t_{i-\frac{1}{2}}) + O(h_i)) - \frac{1}{4}(x_*'(t_{i-\frac{1}{2}}) + x_*'(t_{i-\frac{1}{2}}) + O(h_i)) \\ &\quad -f(x_*(t_{i-\frac{1}{2}}) + O(h_i^2), t_{i-\frac{1}{2}}) + O(h_i) \\ &= x_*'(t_{i-\frac{1}{2}}) - f(x_*(t_{i-\frac{1}{2}}) + O(h_i^2), t_{i-\frac{1}{2}}) + O(h_i) \\ &= f(x_*(t_{i-\frac{1}{2}}), t_{i-\frac{1}{2}}) - f(x_*(t_{i-\frac{1}{2}}) + O(h_i^2), t_{i-\frac{1}{2}}) + O(h_i) \\ &= O(h_i) \end{aligned}$$

Wegen der Darstellung von $v_n \in V_n$ und $v_n(t_i) = 0$, $i = 0, \ldots, N$, gilt

$$\begin{aligned} v_n(t) &= (t - t_{i-1})\frac{v_n(t_{i-\frac{1}{2}})}{\frac{1}{2}h_i} - (t - t_{i-1})(t - t_{i-\frac{1}{2}})\frac{2v_n(t_{i-\frac{1}{2}})}{\frac{1}{2}h_i^2} \\ &= 2\frac{t - t_{i-1}}{h_i}\left(1 - (t - t_{i-\frac{1}{2}})\frac{1}{\frac{1}{2}h_i}\right)v_n(t_{i-\frac{1}{2}}), t \in [t_{i-1}, t_i] \\ \rightsquigarrow |v_n(t)| &\le 2 \cdot (1+1)|v_n(t_{i-\frac{1}{2}})| = O(h_i),\ t \in [t_{i-1}, t_i] \\ \rightsquigarrow \|v_n\|_\infty &= \max_{t\in[t_0,T]} |v_n(t)| = O(h_n) \\ \rightsquigarrow \|F_n r_n x^*\| &= \|v_n\|_\infty = O(h_n) \end{aligned}$$

und die Aussage folgt aus Satz 2.20. □

Abschließend sei angemerkt, dass die Konvergenzaussagen für Kollokationsverfahren auf den Fall $m > 1$ erweitert werden können, indem in den Räumen X_n und Y_n die interpolierenden kubischen bzw. quadratischen Splines für jede Komponente von x bzw. y extra definiert werden.

Literatur:
U. Ascher, R. Matthej, R. Russell: Numerical solution of boundary value problems for ordinary differential equations, Prentice-Hall, 1988.

5 ODE Software

NAG (Numerical Algorithms Group) Library (Fortran, C) `www.nag.co.uk`

- Explizite Runge-Kutta-Verfahren
- variable order, variable step Adams-method
- BDF
- Collocation code
- multiple shooting code

ODEPACK (Collection of ODE Solvers) `www.netlib.org/odepack/opkd-sum`
COLSYS (collocation code) `www.netlib.org/ode/colsys`