Mitschrieb zur Vorlesung: Numerische Mathematik II

Prof. Dr. Wieners

Vorlesung Wintersemester 2005/2006

Letzte Aktualisierung und Verbesserung: 18. Februar 2006

Mitschrieb der Vorlesung Numerische Mathematik II
von Herrn Prof. Dr. Wieners im Wintersemester 2005/2006
von Marco Schreck.

Inhaltsverzeichnis

Kapitel 1

Anfangswertaufgaben

1.1 Existenz- und Eindeutigkeitstheorie

Gute numerische Approximationen sind nur zu erwarten, wenn das kontinuierliche Problem zwei Bedingungen erfüllt, nämlich erstens eine eindeutige Lösung besitzt und zweitens wenn diese eindeutige Lösung nicht zu empfindlich von den Daten abhängt. (Das heißt, unser Problem muss „gut konditioniert“ sein.) Eine Differentialgleichung heißt sachgemäß gestellt, wenn die Lösung eindeutig ist und stetig von den Daten abhängt.

Beispiel:

Wir betrachten die Differentialgleichung $\dot{u} = \sqrt{1-u^2}$. (Dies bedeutet $\dot{u}(t) = \sqrt{1-(u(t))^2}$.) $u(t) = \sin(t+c)$ mit $c \in \mathbb{R}$ oder auch $u \equiv \pm 1$ löst diese Gleichung.

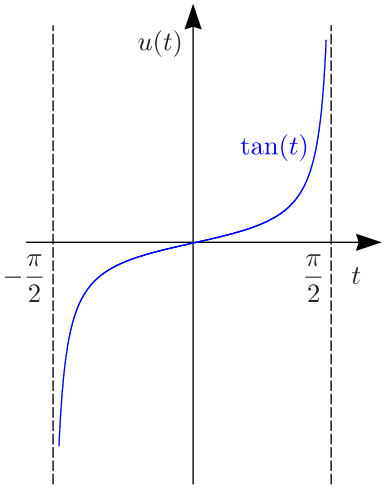
i.) $u(0) = 0$: $u(t) = \sin(t)$

ii.) $u(0) = 1$: Hier stellt sich heraus, dass die Lösung entweder von der Form $u(t) = \sin(t + \pi/2)$ oder $u \equiv 1$ ist. Damit ist die Lösung nicht eindeutig.

iii.) $u(0) = 2$: Dies ist für keine Lösung, die uns zur Verfügung stellen, zu erfüllen.

Beispiel:

Nun betrachten wir $\dot{u} = 1 + u^2$ mit $u(0) = 0$. Die Lösung ist gegeben durch $u(t) = \tan(t)$.



Die Lösung existiert nur im Intervall $[0, \pi/2)$.

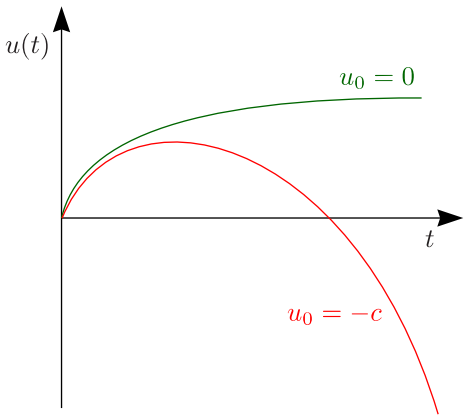
Beispiel:

Nun betrachten wir:

$$\dot{u} = 10\left(u - \frac{c^2}{1+t^2}\right) + \frac{2t}{(1+t^2)^2} \text{ mit } u(0) = u_0$$

Hieraus ergibt sich die Lösung:

$$u(t) = u_0 \exp(10t) + \frac{t^2}{1+t^2}$$



Die Differentialgleichung ist also instabil, denn kleine Änderungen von u_0 ergeben große Änderungen in u!

Definition 1:

Sei $t_0 \in \mathbb{R}$ und eine Simulationszeit $T > 0$. $G \subseteq \mathbb{R}^m$ sei ein Gebiet (offen und zusammenhängend). Zu einem Anfangswert $u_0 \in G$ und $f \in C([t_0, t_0+T] \times G, \mathbb{R}^m)$ suchen wir eine Lösung $u \in C^1([t_0, t_0+T], G)$ der Anfangswertaufgabe (AWA) $\dot{u}(t) = f(t, u(t))$ mit $t \in (t_0, t_0+T)$ und $u(t_0) = u_0$.

Bezeichnungen:

Einfache Betragsstriche sollen für die euklidische Norm $|z| = \sqrt{z^\intercal z}$ mit $z^\intercal z = z \cdot z$ mit $z \in \mathbb{R}^m$ stehen. Die Norm in Funktionenräumen wollen wir mit Doppelstrichen bezeichen:

$$\|u\|_\infty = \max_{t \in [t_0, t_0+T]} |u(t)|$$

$C([t_0, t_0+T], \mathbb{R}^n)$ ist ein BANACHraum. $C^1([t_0, t_0+T], \mathbb{R}^n)$ ist ein BANACHraum bezüglich der Norm

$$\|u\| = \max\{\|u\|_\infty, \|\dot{u}\|_\infty\}$$

Lemma 2:

Für $u \in C([t_0, t_0+T], G)$ ist äquivalent:

a.) Es ist $u \in C^1([t_0, t_0+T], G)$ und u löst die Anfangswertaufgabe.

b.) u löst die Integralgleichung:

$$u(t) = u_0 + \int_{t_0}^{t} f(s, u(s))\,\mathrm{d}s \;\forall\, t \in [t_0, t_0+T]$$

Beweis „⇒“:

$$\int_{t_0}^{t} f(s, u(s))\,\mathrm{d}s = \int_{t_0}^{t} \dot{u}(s)\,\mathrm{d}s = u(s)|_{t_0}^{t} = u(t) - u_0 \qquad \square$$

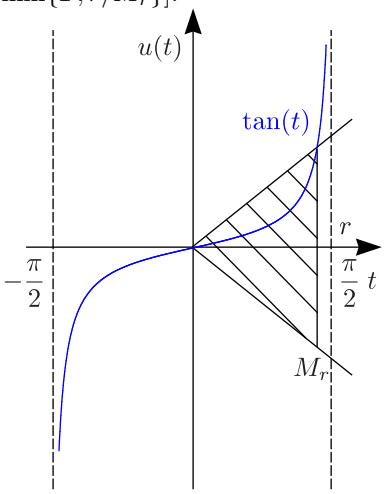
Beweis „⇐“:

$$v(t) = u_0 + \int_{t_0}^{t} f(s, u(s))\,\mathrm{d}s \Rightarrow \dot{v} = f(s, u(s)),\; v(t_0) = u_0$$

Wegen $u \equiv v$ gilt $\dot{u} = \dot{v}$ und damit (1.1) $\square$

Lemma 3:

Zu $r > 0$ mit $B_r(u_0) := \{z \in \mathbb{R}^m : |z - u_0| \leq r\} \subset G$ setze $M_r := \max\{|f(t,z) : (t,z) \in [t_0, t_0 + T] \times B_r(u_0)\}$. Dann gilt für jede Lösung u von (1.1) die a-priori-Schranke $|u(t) - u_0| \leq (t - t_0)M_r$ mit $t \in [t_0, t_0 + \min\{T, r/M_r\}]$.



Beweis:

Wir definieren $d(t) = |u(t) - u_0|$ stetig mit $d(t_0) = 0$. Da dies stetig ist, existiert ein $\bar{t} \in (t_0, t_0 + T]$ mit $d(t) < r$ für $t \in [t_0, \bar{t})$ und im ersten Fall $\bar{t} = t_0 + T$ und im zweiten Fall $d(\bar{t}) = r$. Aus Stetigkeitsgründen können wir daraus schließen, dass im Bereich $[t_0, \bar{t}]$ gilt: $u(t) \in B_r(u_0)$.

$$d(t) = \left| u_0 + \int_{t_0}^{t} f(s, u(s))\,\mathrm{d}s - u_0 \right| \leq \int_{t_0}^{t} |f(s, u(s))|\,\mathrm{d}s \leq (t - t_0)M_r \text{ für } t \in [t_0, \bar{t}]$$

Im zweiten Fall ist $r = d(\bar{t}) \leq (\bar{t} - t_0)M_r$ und daraus folgt $\bar{t} \geq t_0 + r/M_r$. □
Im folgenden wählen wir zu $r > 0$ immer M_r und $T \leq r/M_r$.

Lemma 3:

Zu $N \in \mathbb{N}$ ist der EULERsche Polygonzug $u^N \in C([t_0, t_0 + T], G)$ mit folgender rekursiven Definition, nämlich $u^N(t) = u^N(t^N_{n-1}) + (t - t^N_{n-1})f(t^N_{n-1}, u^N(t^N_{n-1}))$ mit $t \in [t^N_{n-1}, t^N_n]$, $t^N_n = t_0 + n\tau_N$ und $\tau_N = T/N$ wohldefiniert und $|u^N(t) - u_0| \leq (t - t_0)M_r$.

Beweis:

Es ist induktiv zu zeigen, dass $|u^N(t) - u_0| \leq (t - t_0)M_r$ für $[t_0, t^N_n]$. Für $n = 1$ gilt:

$|u^N(t) - u_0| \leq (t - t_0)|f(t_0, u_0)| \leq (t - t_0)M_r$

Dann machen wir den Induktionsschluss $n - 1 \mapsto n$:

$|u^N(t_{n-1}) - u_0| \leq (t^N_{n-1} - t_0)M_r \leq TM_r \leq r \Rightarrow u^N(t_{n-1}) \in G$

$|u^N(t) - u_0| \leq |u^N(t) - u^N(t^N_{n-1})| + |u^N(t) - u_0| \leq (t - t^N_{n-1})|f(t_{n-1}, u^N(t_{n-1}))| + (t^N_{n-1} - t_0)M_r \leq (t - t_0)M_r$□

Definition 2:

$v \in C([t_0, t_0 + T], \mathbb{R}^m)$ ist LIPSCHITZ-stetig, wenn $L > 0$ existiert mit $|v(s) - v(t)| \leq L|s - t|$ wobei s, $t \in [t_0, t_0 + T]$. Die LIPSCHITZ-stetigen Funktionen bilden einen BANACHraum $C^{0,1}([t_0, t_0 + T], \mathbb{R}^m)$ mit der Norm $\|v\| = \max\left\{ \|v\|_\infty, \sup_{t_0 \leq s < t \leq t_0 + T} \frac{|v(s) - v(t)|}{|s - t|} \right\}$.

1.1.1 Folgerung aus dem Satz von Arzela-Ascoli

Der Satz von ARZELA-ASCOLI besagt, dass jede beschränkte Folge $\{v^N : N \in \mathbb{N}\}$ in $C^{0,1}([t_0, t_0 + T], \mathbb{R}^m)$ eine konvergente Teilfolge $\{v^{N_k} : k \in \mathbb{N}$ in $C([t_0, t_0 + T], \mathbb{R}^m)$ besitzt.

1.1.2 Satz von Peano

Die Folge $\{u^N : N \in \mathbb{N}\}$ aus dem Lemma 3 besitzt eine konvergente Teilfolge $\{u^{N_k} : k \in \mathbb{N}\}$, die gegen eine Lösung $u \in C^1([t_0, t_0+T], G)$ der Anfangswertaufgabe konvergiert.

Beweis:

Der erste Schritt ist, zu zeigen, dass u^N in $C^{0,1}$ beschränkt ist. Mit der Dreiecksungleichung ergibt sich:

$\|u\|_\infty \leq \max\limits_{t\in[t_0,t_0+T]}(|u_0| + |u(t) - u_0|) \leq |u_0| + TM_r$

Für $t_0 \leq s < t \leq t_0 + T$ folgt:

$$u^N(t) - u^N(s) = \sum_{s-\tau_N \leq t_{n-1}^N \leq t} \int_{\max\{s,t_{n-1}^N\}}^{\min\{t,t_n^N\}} f(t_{n-1}^N, u^N(t_{n-1}^N))\,\mathrm{d}s$$

Hieraus können wir dann folgern:

$|u^N(t) - u^N(s)| \leq |t - s|M_r$

Damit ist dies gleichmäßig Lipschitz-stetig mit $L = M_r$. Somit ist $\{u^N\}$ in dem Raum $C^{0,1}$ beschränkt. Aus dem Satz von Arzela-Ascoli folgt, dass eine Teilfolge $\{u^{N_k} : k \in \mathbb{N}\}$ existiert und $u \in C([t_0, t_0+T], \mathbb{R}^m)$ mit $\lim_{k\mapsto\infty} \|u^{N_k} - u\|_\infty = 0$. Jetzt wollen wir im zweiten Schritt zeigen, dass u unser Anfangswertproblem löst. Wir definieren eine kompakte Menge $K_r = [t_0, t_0 + T] \times B_r(u_0)$. In einer kompakten Menge ist f gleichmäßig stetig. Das heißt, dass zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert mit der Eigenschaft $|f(t,z) - f(s,y)| \leq \varepsilon$ für $|z - y| \leq \delta$ und $|t - s| \leq \delta$. Wähle ein $k_0 > 0$ mit $\tau^{N_{k_0}} \leq \delta$ und $\tau^{N_k} M_r < \delta$ und $\|u^{N_k} - u\|_\infty \leq \min\{\varepsilon, \delta\}$ für $k \geq k_0$.

$$\left| u^{N_k}(t) - u_0 - \int_{t_0}^{t} f(s, u^{N_k}(s))\,\mathrm{d}s \right| = \left| \sum_{t_{n-1}^N < t} \int_{t_{n-1}^N}^{\min\{t_{n,t}^N\}} \left[f(t_{n-1}^{N_k}, u^{N_k}(t_{n-1}^{N_k})) - f(s, u^{N_k}(s)) \right] \mathrm{d}s \right| \leq (t - t_0)\varepsilon$$

$$\left| u(t) - u_0 - \int_{t_0}^{t} f(s, u(s)\,\mathrm{d}s \right| \leq (t - t_0)\varepsilon + \left| u(t) - u^{N_k}(t) - \int_{t_0}^{t} \big(f(s, u(s)) - f(s, u^{N_k}(s)) \big)\,\mathrm{d}s \right| \leq$$

$$\leq (t - t_0)\varepsilon + \varepsilon + (t - t_0)\varepsilon \square$$

Bemerkung:

1.) Der Satz von Peano garantiert keine Eindeutigkeit.

Beispiel: Die Lösung der Differentialgleichung $\dot{u} = 3u^{\frac{2}{3}}$ mit $u(0) = 0$ ist $u(t) = \max\{0, (x - a)^3\}$, wobei $a \geq 0$.

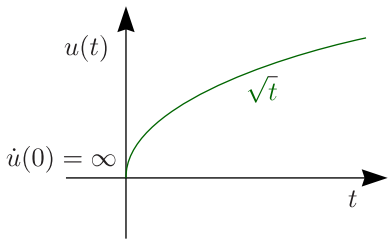
2.) Die Konvergenz ist beliebig langsam.

Definition:

$f \in C([t_0, t_0 + T] \times G, \mathbb{R}^m)$ ist der zweiten Komponente Liptschitz-stetig (erfüllt eine L-Bedingung) in G, wenn $L > 0$ existiert mit $f(t - y) - f(t, z) \leq L|y - z|$, wobei $t \in [t_0, t_0 + T]$ und y, $z \in G$.

Beispiel:

a.) $f(u) = u^\alpha$ mit $0 \in G$ ist Lipschitz-stetig genau dann, wenn $\alpha \geq 1$.



b.) $f(u) = 1 + u^2$

Für $G = [0, 1]$ ist $L = 2$ und für $G = \mathbb{R}$ existiert kein $L < \infty$.

Lemma 4:

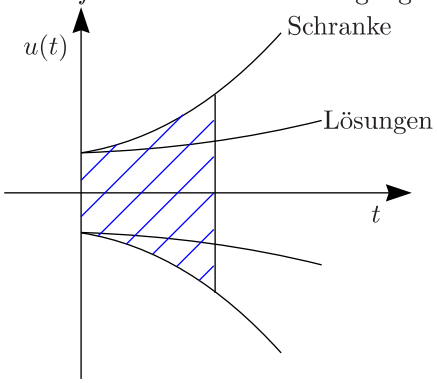
Sei $f \in C^1([t_0, t_0+T] \times \overline{G}, \mathbb{R}^m)$, $\overline{G} \subset \mathbb{R}^m$ beschränkt. Dann erfüllt f eine LIPSCHITZ-Bedingung in G.

Satz 2:

Seien u, $v \in C^1([t_0, t_0+T], G)$ Lösungen von $\dot{u}(t) = f(t, u(t))$ bzw. $\dot{v}(t) = f(t, v(t))$ für $t \in [t_0, t_0+T]$. Wenn f eine LIPSCHITZ-Bedingung in G erfüllt, dann gilt $|u(t) - v(t)| \leq \exp(L(t-t_0))|u(t_0) - v(t_0)|$.

Folgerung:

Wenn f eine LIPSCHITZ-Bedingung erfüllt, ist die Anfangswertaufgabe (1.1) eindeutig lösbar.



Beispiel: Das Lorenz-System im $\mathbb{R}^3$

Wir betrachten:

$$\dot{u} = f(u) \text{ mit } f\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} -10u_1 + 10u_2 \\ 28u_1 - u_2 - u_1u_2 \\ u_1u_2 - \frac{8}{3}u_3 \end{pmatrix} \text{ mit } u(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

ist ein Beispiel für ein „chaotisches“ dynamisches System. Für große Zahlen ist die Lösung auch für nahe beieinander liegende Anfangswerte sehr verschieden.

a.) Es existiert ein $r > 0$ mit $u(t) \leq B_r(u_0)$ für alle $t > 0$.

b.) f erfüllt eine LIPSCHITZ-Bedingung in $B_r(u_0)$. Hieraus folgt, dass die Lösung für alle Zeiten eindeutig bestimmt, aber nicht berechenbar ist.

1.1.3 Gronwall-Lemma

Seien w, a, b: $[t_0, t_0+T] \mapsto \mathbb{R}_{\geq 0} := \{x \in \mathbb{R}, x \geq 0\}$ stückweise stetige Funktionen. Sei b nicht fallend, das heißt, $b(t) \geq b(s)$ für $t \geq s$ und es gelte die Integralungleichung $w(t) \leq \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s + b(t)$ für $t \in [t_0, t_0+T]$. Dann gilt:

$$w(t) \leq b(t)\exp(A(t)) \text{ mit } A(t) = \int_{t_0}^{t} a(s)\,\mathrm{d}s$$

Beweis:

Wir definieren:

$$\varphi(t) = \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s \text{ und } \psi(t) = w(t) - \varphi(t) \Rightarrow \psi(t) \leq b(t)$$

Hieraus ergibt sich $\dot{\varphi}(t) = a(t)w(t)$ und

$$a(t)\psi(t) = a(t)\left(w(t) - \varphi(t)\right) = \dot{\varphi}(t) - a(t)\varphi(t)$$

φ erfüllt somit die Anfangswertaufgabe $\dot{\varphi}(t) = a(t)(\varphi(t) - \psi(t))$ mit $\varphi(t_0) = 0$. Diese Anfangswertaufgabe hat die eindeutige Lösung

$$\varphi(t) = \exp(A(t)) \int_{t_0}^{t} a(s)\psi(s)\exp(-A(s))\,\mathrm{d}s$$

$$\left[\dot{\varphi} = \dot{A}\varphi + \exp(A) + \psi\exp(-A) = a\varphi - a\psi\right]$$

Aus $a(s) \geq 0$, $\psi(s) \leq b(s) \leq b(t)$ für $s \leq t$ folgt:

$$g(t) = \exp(A(t)) \int_{t_0}^{t} a(s)b(t)\exp(-A(s))\,\mathrm{d}x = b(t)\exp(A(t))\left[-\exp(-A(s))\right]_{t_0}^{t} = b(t)\left[\exp(A(t)) - 1\right]$$

Nach Voraussetzung ist $w(t) \leq g(t) + b(t) = b(t)\exp(A(t))$. □

Beweis von Satz 2:

Wir setzen $w(t) = 1/2|u(t) - v(t)|^2$, wobei $|z|^2 = z^\intercal z$.

$$\begin{aligned}\dot{w}(t) &= (u(t) - v(t))^\intercal(\dot{u}(t) - \dot{v}(t)) = (u(t) - v(t))^\intercal(f(t, u(t)) - f(t, v(t))) \leq \\ &\leq |u(t) - v(t)| \underbrace{|f(t, u(t)) - f(t, v(t))|}_{\leq L|u(t)-v(t)} \leq L|u(t) - v(t)|^2 = 2Lw(t)\end{aligned}$$

Setze $a(t) \equiv 2L$, $b(t) \equiv w(t_0)$. Dann gilt $A(t) = (t - t_0) \cdot 2L$. Hieraus folgt mit dem Gronwall-Lemma:

$$w(t) \leq w(t_0)\exp(2L(t - t_0))$$

$$|u(t) - v(t)| = \sqrt{2w(t)} \leq \sqrt{2w(t_0)} \cdot \sqrt{\exp(2L(t - t_0))} = |u(t_0) - v(t_0)\exp(L(t - t_0))$$ □

Lemma 5:

Der Integraloperator F: $C([t_0, t_0 + T], G) \mapsto C([t_0, t_0 + T], G)$, $v \mapsto F(v)(t) := u_0 + \int_{t_0}^{t} f(s, v(s))\,\mathrm{d}$ ist wohldefiniert.

Beweis:

$$|F(v)(t) - u_0| \leq (t - t_0)M_r \leq TM_r \leq r$$

Man kann nun die Aussage 1.2 analog so formulieren: u löst die Anfangswertaufgabe $\Leftrightarrow$ $u = F(u)$ Fixpunkt □

1.1.4 Satz von Picard-Lindelöff

Die Lipschitz-Bedingung (1.8) sei erfüllt. Dann gilt für $T > 0$ mit $LT < 1$, dass die sogenannte „Picard-Folge“ $v^0 = u_0$, $v^k = F(v^{k-1})$ für $k = 1, 2, 3, \ldots$ gegen eine Lösung u von (1.1) konvergiert.

Beweis:

$$\begin{aligned}F(v) - F(w)\|_\infty &\leq \max_{t\in[t_0,t_0+T]} \left| \int_{t_0}^{t} [f(s, v(s)) - f(s, w(s))]\,\mathrm{d}s \right| \leq \max_{t\in[t_0,t_0+T]} \int_{t_0}^{t} |f(s, v(s)) - f(s, w(s))|\,\mathrm{d}s \leq \\ &\leq \max_{t\in[t_0,t_0+T]} \int_{t_0}^{t} L|v(s) - w(s)|\,\mathrm{d}s \leq \underbrace{TL}_{<1} \|v - w\|_\infty\end{aligned}$$

Damit ist F kontrahierend und wir können den Banachschen Fixpunktsatz anwenden. Die Folge konvergiert nach dem Fixpunktsatz gegen einen Fixpunkt $u \in C([t_0, t_0 + T], \mathbb{R}^m)$. Hieraus folgt mit (1.2), dass u die Anfangswertaufgabe (1.1) löst. □

Bemerkung:

Das ganze ist fortsetzbar bis zur a-priori-Schranke $T = r/M_r$.

1.2 Explizite Einschrittverfahren

Definition (1.15):

Zur Funktion $f \in C([t_0, t_0 + T] \times G, \mathbb{R}^m)$ definieren wir den Fluss Φ: $D \subset [t_0, t_0 + T] \times \mathbb{R}_{\geq 0} \times G \mapsto \mathbb{R}^m$ durch $\Phi(t, \tau, v_0) = v(t + \tau)$, wobei $v \in C^1([t_0, t + \tau], G)$ Lösung der Anfangswertaufgabe $\dot{v}(s) = f(s, v(s))$ für $s \in (t, t + \tau)$ mit $v(t) = v_0$ ist. D umfasst alle Werte, für die diese Anfangswertaufgabe eindeutig lösbar ist.

Definition (1.16):

Ein explizites Einschritt-Verfahren wird durch eine Verfahrensfunktion Ψ: $[t_0, t_0+T]\times\mathbb{R}_{\geq 0}\times G \mapsto \mathbb{R}^m$ definiert. Zu $u_0 \in G$ setze $u_n = u_{n-1} + \tau_n \Psi(t_{n-1}, \tau_n, u_{n-1})$. Dabei ist $\tau_n = \tau_n - \tau_{n-1}$ Schrittweite auf dem Gitter $\Delta = \{t_0, \ldots, t_N\}$, u^N: $\Delta \mapsto G$: $u^N(t_n) = u_n$. $\Phi_\tau(t, \tau, z) = z + \tau\Psi(t, \tau, z)$ ist der sogenannte „diskrete Fluss“.

Beispiele:

a.) Explizite EULERverfahren: $\Psi(t, \tau, z) = f(t, z)$

b.) Verfahren von Heum: $k_1 = f(t, z)$ und $k_2 = f(t + \tau, z + \tau k_1)$
Dann ist $\Psi(t, \tau, z) = 1/2k_1 + 1/2k_2$.

c.) Klassisches RUNGE-KUTTA-Verfahren: $k_1 = f(t, z)$, $k_2 = f(t+\tau/2, z+1/2\tau k_1)$, $k_3 = f(t+\tau/2, z+1/2\tau k_2)$, $k_4 = f(t + \tau, z + \tau k_3)$

$$\Psi(t, \tau, z) = \frac{1}{6}k_1 + \frac{2}{6}k_2 + \frac{2}{6}k_3 + \frac{1}{6}k_4$$

Bemerkung:

Im Spezialfall $f(t, z) = g(t)$ gilt:

$$u(t_n) - u(t_{n-1}) = \int_{t_{n-1}}^{t_n} \dot{g}(t) \approx \begin{cases} \tau_n g(t_{n-1}) & \\ \tau_n/2\,(g(t_{n-1}) + g(t_n)) & \text{Trapezregel} \\ \tau_n/6\,(g(t_{n-1}) + 4g((t_{n-1} + t_n)/2) + g(t_n)) & \text{Simpsonregel} \end{cases}$$

Definition (1.17):

Durch $e_n = u(t_n) - u_n$ ist der „globale Fehler“ gegeben.

$$g_n = \frac{1}{\tau_n}(u(t_n) - u(t_{n-1})) - \Psi(t_n, \tau_n, u(t_{n-1}))$$

g_n bezeichnet man als „**lokalen Diskretisierungsfehler**“.

Bemerkung:

a.) Der globale Fehler e_n hängt von $[t_0, \ldots, t_n]$ ab!

b.) Der „lokale Diskretisierungsfehler“ g_n misst in $[t_{n-1}, t_n]$ den Unterschied zwischen der Tangente an die Lösung und der Approximation durch die Verfahrensfunktion.

Beachte:

$$g_n = \frac{1}{\tau_n}\left(\phi(t_{n-1}, \tau, u(t_{n-1})) - z\right) - \frac{1}{\tau}\left(\phi_\tau(t, \tau, z) - z\right)$$

Beispiel:

a.) Explizites EULERverfahren:

$$g_n = \frac{1}{\tau_n}(u(t_n) - u(t_{n-1})) - f(t_{n-1}, u(t_{n-1})) = \frac{1}{\tau_n}\int_{t_{n-1}}^{t_n} (\dot{u}(t_n) - f(t_{n-1}, u(t_{n-1}))) \, \mathrm{d}t =$$

$$= \frac{1}{\tau_n}\int_{t_{n-1}}^{t_n} (\dot{u}(t_n) - \dot{u}(t_{n-1})) \, \mathrm{d}t = \frac{1}{\tau_n}\int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{t} \ddot{u}(s) \, \mathrm{d}s \, \mathrm{d}t$$

Wir können dann den lokalen Fehler wie folgt abschätzen:

$$|g_n| \leq \frac{1}{\tau_n} \max_{s\in[t_{n-1},t_n]} |\ddot{u}(s)| \int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{t} \mathrm{d}s\mathrm{d}t = \frac{\tau_n}{2}\|\ddot{u}\|_\infty$$

Definition (1.18):

a.) Ein Einschrittverfahren heißt „konsistent“, wenn $\lim_{\tau_n \mapsto} g_n(t, \tau, z) \mapsto 0$. Es heißt „konsistent von der Ordnung p“, wenn $|g_n| = O(\tau_n^p)$.

b.) Es heißt „konvergent“, wenn $\lim_{\tau\mapsto 0} \max_{n=1,\ldots,N_\tau} e_n = 0$ ($n = 0, \ldots, N$ und $\tau = \max \tau_n$ mit $n = 1, \ldots, N$). Es heißt „konvergent von der Ordnung p“, wenn $|e_n| = O(\tau^p)$.

Definition (1.19):

Für die Verfahrensfunktion gelte $|\Psi(t, \tau, z) - \Psi(t, \tau, y)| \leq \Lambda|z - y|$ für $t \in [t_0, t_0 + T]$, $\tau \leq \tau_0$ und $z, y \in G$. Dann gilt:

$$|u(t_n) - u_n| \leq \exp(\Lambda(t_n - t_0))|u(t_0) - u_0| + \max_{j=1,\ldots,n} |g_j| \frac{1}{\Lambda} (\exp(\Lambda(t_n - t_0)) - 1)$$

Bemerkung:

a.) Für das explizite EULERverfahren gilt $\Psi = f$ und damit $\Lambda = L$.

Folgerung:

Verfahren der Konsistenzordnung p sind konvergent von der Ordnung p.

Diskretes Gronwall-Lemma (1.21):

Seien $\delta_n > 0$ und η_n, $z_n \geq 0$ für $n = 0, \ldots, N$ mit $z_n \leq (1 + \delta_n)z_{n-1} + \eta_n$ mit $n = 1, \ldots, N$. Dann gilt:

$$z_n \leq z_0 \exp(\Delta_n) + \max_{j=1,\ldots,n} \frac{\eta_j}{\delta_j}(\exp(\Delta_n) - 1) \text{ mit } \Delta_n = \sum_{j=1}^{n} \delta_j$$

Beweis:

Setze $M_n = \max_{j=1,\ldots,n} \eta_j/\delta_j$. Wir führen den Beweis mit vollständiger Induktion. Der Induktionsanfang für $\delta_1 > 0$ und z_0, $\eta_0 \geq 0$ lautet:

$$z_1 \leq (1 + \delta_1)z_0 + \eta_1 = (1 + \delta_1)z_0 + \frac{\eta_l}{\delta_l}\delta_l \leq \exp(\delta_l)z_0 + M_1(\exp(\delta_1) - 1) = \exp(\Delta_n)z_0 + M_1(\exp(\Delta_1) - 1)$$

Induktionsschritt $n - 1 \mapsto n$:

$$z_n \leq (1 + \delta_n)z_{n-1} + \eta_n \leq (1 + \delta_n)\left(z_0 \exp(\Delta_{n-1}) + M_{n-1}(\exp(\Delta_{n-1}) - 1) + \frac{\eta_l}{\delta_l}\delta_l\right) \leq$$

$$\leq z_0 \exp(\Delta_{n-1}) \exp(\delta_n) + M_n \left[(1 + \delta_n) (\exp(\Delta_{n-1}) - 1) + \delta_n\right] =$$

$$= \exp(\Delta_{n-1} + \delta_n) + M_n \left[(1 + \delta_n) \exp(\Delta_{n-1}) - 1\right] \leq$$

$$\leq z_0 \exp(\Delta_n) + M_n (\exp(\Delta_n) - 1)$$

□

Beweis von (1.19):

$$\begin{aligned} e_n &= u(t_n) - u_n = e_{n-1} + u(t_n) - u(t_{n-1}) - (u_n - u_{n-1}) = \\ &= e_{n-1} + u(t_n) - u(t_{n-1}) - \tau_n \Psi(t_{n-1}, \tau_n, u_{n-1}) + \tau_n \Psi(t_{n-1}, \tau_n, u(t_{n-1})) - \tau_n \Psi(t_{n-1}, \tau_n, u(t_{n-1})) = \\ &= e_{n-1} + \tau_n g_n - \tau_n \left(\Psi(t_{n-1}, \tau_n, u_{n-1}) - \Psi(t_{n-1}, \tau_n, u(t_{n-1}))\right) \end{aligned}$$

Hieraus ergibt sich weiter:

$$|e_n| \leq |e_{n-1}| + \tau_n |g_n| + \tau_n \Lambda |u_{n-1} - u(t_{n-1})| = (1 + \Lambda\tau_n)|e_{n-1}| + \tau_n |g_n|$$

Wähle $\delta_n = \tau_n \Lambda$, $z_n = |e_n|$ und $\eta_n = |g_n|\tau_n$. Wir können nun das Gronwall-Lemma (1.21) anwenden und erhalten:

$$|e_n| \leq |e_0| \exp(\Lambda(t_n - t_0)) + \max_{j=1,\ldots,n} \frac{|g_j|}{\Lambda} \left[\exp(\Lambda(t_n - t_0)) - 1\right] \qquad \square$$

Wiederholung:

Für $\psi(t, \tau, z) = f(t, z)$ für das explizite Eulerverfahren:

$$\begin{aligned} u(t_n) - u_n &= u(t_{n-1} - u_{n-1} + \frac{\tau_n}{\tau_n} \left(u(t_n) - u(t_{n-1})\right) - u_n + u_{n-1} = \\ &= e_{n-1} + \tau_n g_n + f(t_{n-1}, u(t_{n-1})) - f(t_{n-1}, u_{n-1}) = \\ &= e_{n-1} + \tau_n g_n + \tau_n \left(f(t_{n-1}, u(t_{n-1})) - f(t_{n-1}, u_{n-1})\right) \end{aligned}$$

Hieraus folgt weiter:

$$|e_n| \leq |e_{n-1}| + \tau_n |g_n| + \tau_n L |e_{n-1}| = (1 + \tau_n L)|e_{n-1} + \tau_n |g_n|$$

Vollständige Induktion für $|e_0| = 0$:

$$|e_n| \leq \max_{j=1,\ldots,n} |g_j| \left(\frac{\exp(L(t_n - t_0)) - 1}{L}\right)$$

Für $n = 1$ gilt $|e_1| \leq |g_1|\tau_1 + O(\tau_1^2)$.

Definition 1.21:

Ein allgemeines explizites Runge-Kutta-Verfahren mit s Stufen wird durch Stützstellen c_i mit $i = 1, \ldots, s$ durch Gewichte b_j mit $j = 1, \ldots, s$ und Koeffizienten a_{ij} mit $i, j = 1, \ldots, s$ definiert. (Das Verfahren ist explizit.) Dann setze

$$k_1 = f(t, z),\ k_2 = f(t + c_2\tau, z + \tau a_{21} k_1),\ \ldots,\ k_s = f\left(t + c_s\tau, z + \tau \sum_{j=1}^{s-1} a_{ij} k_j\right) \text{ mit } a_{ij} = 0 \text{ für } i \leq j$$

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^\intercal \end{array}$$ (Butcher-Schema)

und

$$\psi(t, \tau, z) = \sum_{j=1}^{s} k_j b_j$$

Dabei sind Stützstellen und Gewichte einer Qudratur in $[0, 1]$ zugleich

$$\int_0^1 f(t)\,\mathrm{d}t \approx \sum_{j=1}^{3} b_j f(c_j)$$

Beispiele:

a.) Explizites EULERverfahren: $s = 1$, Konsistenzordnung $p = 1$

0	
	1

b.) Verfahren von HEUM: $s = 2$, Konsistenzordnung $p = 2$

0		
1	1	
	1/2	1/2

c.) Klassische RUNGE-KUTTA-Verfahren: $s = 4$, Konsistenzordnung $p = 4$

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/6	1/6	1/6

Lemma 1.22:

a.) Ein explizites RUNGE-KUTTA-Verfahren ist genau dann konsistent für alle $f \in C([t_0, t_0 + T] \times \mathbb{R}^m, \mathbb{R}^m)$, wenn sie Summe der Gewichte, also $\sum_{j=1}^{s} b_j = 1$.

b.) Wenn ein explizites RUNGE-KUTTA-Verfahren für alle $f \in C^\infty([t_0, t_0+T]\times\mathbb{R}^n, \mathbb{R}^n)$ die Konsistenzordnung p hat, dann gilt $p \leq s$.

Beweis:

a.) Aussage ①:

$$\lim_{\tau_n \mapsto 0} g_n = \lim_{\tau_n \mapsto 0} \left(\left[\frac{1}{\tau_n}(u(t_n) - u(t_{n-1})) \right] - \psi(t_{n-1}, \tau_n, u(t_{n-1})) \right) = \dot{u}(t_n) - \sum_{j=1}^{k} b_j k_j =$$

$$= \dot{u}(t_n) - \sum_{j=1}^{s} b_j f(t_n, u(t_n)) = \dot{u}(t_n) \left(1 - \sum_j b_j \right)$$

□

b.) Aussage ②:

Wähle $f(z) = z$ $(m = 1)$ und $u(0) = 1$. Aus $\dot{u} = u$ folgt $u(t) = \exp(t)$. Hieraus ergibt sich:

$$g_1 = \underbrace{\frac{1}{\tau}(u(\tau) - u(0))}_{=1+1/2\tau+1/6\tau^2+\ldots} - \underbrace{\psi(0, \tau, 1)}_{\in \mathbb{P}_{s-1}}$$

Es gilt nun $k_1 = 1 \in \mathbb{P}_0$, $k_2 = 1 + \tau a_{21} \in \mathbb{P}_1$, $k_3 = 1 + \tau(a_{31} + a_{32}(1 + \tau a_{21})) \in \mathbb{P}_2(\tau)$. Hieraus ergibt sich:

$$g_1 = \frac{1}{(s+1)!}\tau^s + \ldots$$

Dies ist optimal! □

Bemerkung:

Für die maximale Konsistenzordnung von expliziten RUNGE-KUTTA-Verfahren gilt:

p	1	2	3	4	5	6	7	8	10
s	1	2	3	4	6	7	9	11	18

1.3 Autonomisierung

Ein System $\dot{u} = f(t, u(t))$ mit $u(t_0) = u_0$ ist äquivalent zu $\dot{y}(t) = \hat{f}(y(t))$ mit $y(t_0) = y_0$, wobei

$$y(t) = \begin{pmatrix} u(t) \\ t \end{pmatrix} \in \mathbb{R}^{n+1},\ \hat{f}(y) = \hat{f}\begin{pmatrix} z \\ t \end{pmatrix} = \begin{pmatrix} f(t,z) \\ 1 \end{pmatrix} \text{ und } y_0(t_0) = \begin{pmatrix} u_0 \\ t_0 \end{pmatrix} \in \mathbb{R}^{n+1}$$

Eine Verfahrensfunktion ist „invariant gegen Autonomisierung“, wenn

$$\begin{pmatrix} \psi(t,\tau,z) \\ 1 \end{pmatrix} = \hat{\psi}\left(t, \tau, \begin{pmatrix} z \\ t \end{pmatrix}\right)$$

gilt.

Lemma:

Ein explizites Runge-Kutta-Verfahren ist genau dann invariant gegen Autonomisierung, wenn es konsistent ist und wenn $c_i = \sum_{j=1}^{s} a_{ij}$ mit $i = 1, \ldots, s$ gilt.

Beweis:

$$k_i = f\left(t + \tau c_i, z + \tau \sum_{j=1}^{i-1} a_{ij} b_j\right)$$

$$\begin{pmatrix} \hat{k}_i \\ \theta_i \end{pmatrix} = \hat{f}\left[\begin{pmatrix} z \\ t \end{pmatrix} + \tau \sum_j a_{ij} \begin{pmatrix} \hat{k}_j \\ \theta_j \end{pmatrix}\right] = \begin{pmatrix} f\left(t + \tau \sum_j a_{ij}, z + \tau \sum_j a_{ij} \hat{k}_j\right) \\ 1 \end{pmatrix}$$

Hieraus folgt $\theta_i = 1$.

$$\begin{pmatrix} \sum_j b_j k_j \\ 1 \end{pmatrix} = \sum b_j \begin{pmatrix} \hat{k}_j \\ 1 \end{pmatrix} \Rightarrow \sum_j b_j = 1 \text{ und } k_j = \hat{k}_j \text{ mit } f \text{ beliebig}$$

Hieraus ergibt sich

$$\sum_j a_{ij} = c_i \qquad \square$$

Satz 1.24:

Ein Runge-Kutta-Verfahren ist genau dann konsistent von der Ordnung

- $p = 1$: $\sum_j b_j h = 1$
- $p = 2$: $\sum_j b_j c_j = \frac{1}{2}$
- $p = 3$: $\sum_j b_j c_j^2 = \frac{1}{3}$, $\sum_{i,j} b_i a_{ij} c_j = \frac{1}{6}$
- $p = 4$: $\sum_j b_j c_j^2 = \frac{1}{3}$, $\sum_{i,j} b_i c_i a_{ij} c_j = \frac{1}{8}$, $\sum_{i,j} b_i a_{ij} c_j^2 = \frac{1}{12}$, $\sum_{i,j,k} b_i a_{ij} a_{jk} c_k = \frac{1}{24}$

Beispiel:

Für die Simpsonregel gilt $b_1 = 1/6$, $b_2 = b_3 = 2/6$, $b_4 = 1/6$, $c_1 = 0$, $c_2 = c_3 = 1/2$ und $c_4 = 1$. Durch Einsetzen findet man heraus, dass das Simpsonverfahren konsistent von der Ordnung $p = 1$ und $p = 2$ ist.

$$c_2 = a_{21} = \frac{1}{2}$$

$$b_3 a_{32} c_2 + b_4(a_{42} c_2 + a_{43} c_3) = \frac{1}{6} \text{ und } b_3 c_3 a_{32} c_2 + b_4 c_4(a_{42} c_2 + a_4 3 c_3) = \frac{1}{8} \Rightarrow a_{22} = \frac{1}{2} \text{ und } a_{42} c_2 + a_{43} c_3 = \frac{1}{2}$$

$$b_4 a_{43} a_{32} c_2 = \frac{1}{24} \Rightarrow a_{43} = 1$$

1.3.1 Realisierung

Für ein derartiges Verfahren benötigen wir einen Startwert u_0 und eine Zeit t_0.

a.) Wähle die Schrittweite $\tau > 0$. (Ohne Kenntnis des Problems und der Größenordnung ist es nicht einfach, eine sinnvolle Schrittweite zu wählen.) Setzte $t = t_0$ und $u = u_0$.

b.) $k_1 = f(t, u)$, $k_2 = f(t + \tau/2, u + \tau/2k_1)$, $k_3 = f(\tau/2, u + \tau/2k_2)$ und $k_4 = f(t + \tau, u + \tau k_3)$

c.) $u := u + \dfrac{\tau}{6}(k_1 + 2k_2 + 2k_3 + k_4)$ und $t := t + \tau$. Gehe dann zu Schritt b.) zurück.

Für $e_0 = u(t_0) - u_0 = 0$ gilt die Abschätzung $|g_n| \leq C\tau^p$. Für dieses Beispiel gilt $p = 4$. Dies gilt für $\tau \leq \tau_0$ (maximal erlaubte Schrittweite). Hieraus ergibt sich:

$$|u(t_n) - u_n| \leq C\tau^p \frac{\exp(\Lambda(t_n - t_0)) - 1}{\Lambda}$$

Beweis von (1.24) für $p = 2$:

Wir nehmen ohne Einschränkung an, dass die Differentialgleichung autonom ist. Wir machen eine Taylor-Entwicklung:

$$u(t + \tau) + u(t) + \dot{u}\tau + \frac{1}{2}\ddot{u}\tau^2 + \frac{1}{6}\ddot{u}(t)\tau^3 + O(\tau^4)$$

Offensichtlich gilt $\dot{u}(t) = f(u(t)) = f$. Hieraus ergibt sich:

$$\ddot{u}(t) = \frac{\mathrm{d}}{\mathrm{d}t} f(u(t)) = Df(u(t)) \cdot \dot{u}(t) = f' \cdot f$$

$$\dddot{u}(t) = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 f(u(t)) = \frac{\mathrm{d}}{\mathrm{d}t}\left(Df(u(t)) \cdot \dot{u}(t)\right) = \left(\frac{\mathrm{d}}{\mathrm{d}t} Df(u(t))\right)\dot{u}(t) + Df(u(t))\frac{\mathrm{d}}{\mathrm{d}t}\dot{u}(t) = f''ff + f'f'f$$

Wir wollen nun auch die Verfahrensfunktion entwickeln:

$$k_i = f\left(z + \tau \sum_{j=1}^{i-1} a_{ij} k_j\right) = f(z) + \tau Df(z) \sum_{j=1}^{i-1} a_{ij} b_j + O(\tau^2) = f(z) + \tau Df(z) \sum_{j=1}^{i-1} a_{ik} f(z) + O(\tau^2) \text{ mit } \sum_{j=1}^{i-1} a_{ij} \equiv c_i$$

Hieraus ergibt sich also:

$$\begin{aligned}
g &= \frac{1}{\tau}\left[u(t + \tau) - u(t)\right] - \psi(t, \tau, u(t)) = \\
&= \dot{u}(t) + \frac{\tau}{2}\ddot{u}(t) + O(\tau^2) - \sum_j b_j k_j = \\
&= f(u(t)) + \frac{\tau}{2} Df(u(t)) f(u(t)) - \sum_j b_j f(u(t)) - \tau Df(u(t)) c_j f(u(t)) + O(\tau^2) = \\
&= f(u(t)) \underbrace{\left(1 - \sum_j b_j\right)}_{\stackrel{!}{=}0} + \tau Df(u(t)) f(u(t)) \underbrace{\left(\frac{1}{2} - \sum_j b_j c_j\right)}_{\stackrel{!}{=}0} + O(\tau^2)
\end{aligned}$$

□

Lemma (1.25):

Wenn f eine L-Bedingung erfüllt, dann ist die Λ-Bedingung in (1.19) für explizite Runge-Kutta-Verfahren erfüllt.

Beweis:

Setze

$$k_i = f\left(z + \tau \sum_j a_{ij} k_j\right) \text{ und } \hat{k}_i = f\left(\hat{z} + \tau \sum_j a_{ij} \hat{k}_j\right)$$

Hieraus ergibt sich:

$$|\psi(t,\tau,z) - \psi(t,\tau,\hat{z})| \leq \left|\sum_i b_i (k_i - \hat{k}_i)\right| \leq \sum_i |b_i| \max |k_i - \hat{k}_i|$$

$$|k_1 - \hat{k}_1| = |f(z) - f(\hat{z})| \leq L|z - \hat{z}| \equiv L_1 |z - \hat{z}|$$

Unter Verwendung der Dreiecksungleichung und von $|k_1 - \hat{k}_1|$ ergibt sich:

$$|k_2 - \hat{k}_2| = \left|f(z + \tau a_{21} k_1) - f(\hat{z} + \tau a_{21} \hat{k}_1)\right| \leq L|z + \tau a_{21} k_1 - \hat{z} - \tau a_{21} \hat{k}_1| \leq L(1 + \tau |a_{21} L)|z - \hat{z}| \equiv L_2 |z - \hat{z}|$$

Damit ergibt sich dann:

$$|k_s - \hat{k}_s| = \left|f\left(z + \sum_j a_{ij} b_j\right) - f\left(\hat{z} + \sum_j a_{ij} b_j\right)\right| \leq L\left(1 + \tau \sum_j |a_{ij}| L_j\right) |z - \hat{z}| \equiv L_s |z - \hat{z}|$$

Wähle im letzten Schritt $\Lambda = L_s$, womit das Lemma bewiesen ist. □

Frage: Lässt sich die Schrittweite τ_n zu gegebenem $\varepsilon > 0$ (Genauigkeit) variabel in der Art steuern, dass folgendes $|g_n| < \varepsilon$? Dies bedeutet:

$$|u(t_n) - u_n| \leq \varepsilon \frac{\exp(\Lambda(t - t_0)) - 1}{\Lambda}$$

Die Idee ist, zwei Verfahrensfunktionen ψ, $\hat{\psi}$ der Konsistenzordnung $p > \hat{p}$ zu verwenden.

$$|g_n - \hat{g}_n| = |\psi(t_{n-1}, \tau_n, u(t_{n-1})) - \hat{\psi}(t_{n-1}, \tau_n, u(t_{n-1}))|$$

Dann folgt aus der „Saturationsannahme“

$$\frac{|g_n|}{|\hat{g}_n|} = \frac{O(\tau^p)}{O(\tau^{\hat{p}})} = O(\tau^{p-\hat{p}}) \leq \theta < 1 \text{ für } \tau < \tau_0$$

mit der Dreiecksungleichung

$$|\hat{g}_n| \leq |g_n - \hat{g}_n| + |g_n| \leq |g_n - \hat{g}_n| + \theta |\hat{g}_n| \Rightarrow |\hat{g}_n| \leq \frac{1}{1 - \theta} |g_n - \hat{g}_n|$$

und außerdem

$$|g_n - \hat{g}_n| \leq |g_n| + |\hat{g}_n| \leq (1 + \theta)|\hat{g}_n| \Rightarrow \frac{1}{1 + \theta} |g_n - \hat{g}_n| \leq |\hat{g}_n|$$

Das heißt, wir haben den lokalen Diskretisierungsfehler also sowohl nach oben als auch nach unten abgeschätzt. Also gilt

$$\eta_n := \psi(t_{n-1}, \tau_n, u_{n-1}) - \hat{\psi}(t_{n-1}, \tau_{n-1}, u_{n-1}) \approx g_n - \hat{g}_n \approx \hat{g}_n$$

Dies ist ein „Schätzer für den lokalen Diskretisierungsfehler“. Der tatsächliche Fehler ist kleiner oder gleich einer Konstante multipliziert mit dem „Fehlerschätzer“. (Im Ferhlerschätzer vergleicht man den lokalen Diskretisierungsfehler einer Verfahrensfunktion mit dem lokalen Diskretisierungsfehler einer anderen Verfahrensfunktion.)

Bemerkung:

Der Fehlerschätzer ist „asymptotisch exakt“, denn $\eta_n \mapsto \hat{g}_n$ für $\tau \mapsto 0$ ($\theta \mapsto 0$ für $\tau \mapsto 0$).

1.3.2 Schrittweitenvorhersage

Sei $|\eta_n| \approx |\hat{g}_n| = C\tau_n^{\hat{p}} + O(\tau_n^{\hat{p}+1}) \approx C\tau_n^{\hat{p}} \not\approx \varepsilon$. Wir wählen τ_{n+1} mit der Eigenschaft $|\eta_{n+1}| \approx C\tau_{n+1}^{\hat{p}} \approx \varepsilon$. Wegen $C \approx |\eta_n|/\tau_n^{\hat{p}}$ gilt dann:

$$\tau_{n+1} = \sqrt[\hat{p}]{\frac{\varepsilon}{c}} = \tau_n \sqrt[\hat{p}]{\frac{\varepsilon}{|\eta_n|}}$$

Definition (1.27):

c	$\mathcal{A}$
	$b^\intercal$
	$\hat{b}^\intercal$

Eingebettete RUNGE-KUTTA-Verfahren

$$k_i = f\left(t + c_i\tau, z + \tau\sum_j a_{ij}k_j\right) \text{ mit } \psi(t,\tau,z) = \sum_i b_i k_i \text{ (Ordnung } p) \text{ und } \hat{\psi}(t,\tau,z) = \sum_j \hat{b}_i k_i \text{ (Ordnung } \hat{p})$$

Betrachten wir als Beispiel die SIMPSONregel:

0					
1/2	1/2				
1/2	0	1/2			
1	0	0	1		
1	1/6	2/6	2/6	1/6	
	1/6	2/6	3/6	1/6	0
	1/6	2/6	2/6	0	1/6

$$\psi(t,\tau,z) - \hat{\psi}(t,\tau,z) = \frac{1}{6}(k_4 - k_5)$$

1.3.3 Adaptives eingebettetes Runge-Kutta-Verfahren

Wir gehen aus von den Startwerten u_0 und t_0.

a.) Wähle $\varepsilon > 0$, $\tau > 0$ und $\vartheta \in (0,1)$. Setze $u = u_0$, $t = t_0$, $n = 1$ und $k_1 = f(t,u)$.

b.) Berechne

$$k_2 = f\left(t + \frac{\tau}{2}, u + \frac{\tau}{2}k_1\right),\ k_3 = f\left(t + \frac{\tau}{2}, u + \frac{\tau}{2}k_2\right)$$

$$k_4 = f\left(t + \tau, u + \tau k_3\right),\ k_5 = f\left(t + \tau, u + \frac{\tau}{6}(k_1 + 2k_2 + 2k_3 + k_4)\right)$$

Der Fehlerschätzer lautet $\eta = 1/6(k_4 - k_5)$.

c.) Falls $|\psi| < \varepsilon$ wird definiert:

$$t := t + \tau \text{ und } u := u + \frac{\tau}{6}(k_1 + 2k_2 + 2k_3 + k_4),\ k_1 = k_5 \text{ und } n := n + 1$$

d.) $\tau := \tau\sqrt[3]{\dfrac{\varepsilon}{|\eta|}}$ und gehe zu Schritt b.)

$\eta(t,\tau,z) = g(t,\tau,z) - \hat{g}(t,\tau,z) = \hat{\psi}(t,\tau,z) - \psi(t,\tau,z)$ mit $\hat{p} < p$

Speziell für das RUNGE-KUTTA-Verfahren gilt:

$$\eta(t,\tau,z) = \sum_j \left(\hat{b}_j - b_j\right) k_j$$

$$\lim_{\tau\mapsto 0}\max_{n=1,\ldots,N_\tau} |\eta_n + \hat{g}_n| =$$

$$= \lim_{\tau\mapsto 0}\max_{n=1,\ldots,N_\tau} |\eta(t_{n-1},\tau_n,u_{n-1}) - \eta(t_{n-1},\tau,u(t_{n-1})) + \eta(t_{n-1},\tau_n,u(t_{n-1})) + \hat{g}(t_{n-1},\tau_n,u(t_{n-1}))| \leq$$

$$\leq \max_{n=1,\ldots,N_\tau} \underbrace{(\Lambda + \hat{\Lambda})|u_{n-1} - u(t_{n-1})|}_{O(\tau^p)} + \underbrace{|g_n|}_{O(\tau^p)} \Rightarrow \frac{|\eta_n + \hat{g}_n|}{|\hat{g}_n|} = O(\tau^{p-\hat{p}}) \mapsto 0$$

Hieraus ergibt sich $|\eta_n| \mapsto |\hat{g}_n|$; der Fehlerschätzer η_n ist also asymptotisch exakt.

1.4 Fehlerschätzen durch Extrapolation

Vergleiche $\hat{u}_n = u_{n-1} + \tau_n \psi(t_{n-1}, \tau_n, u_{n-1})$ mit demselben Verfahren bei halber Schrittweite $\widetilde{u}_{n-\frac{1}{2}} = u_{n-1} + \frac{\tau_n}{2}\psi(t_{n-1}, \tau_n/2, u_{n-1})$.

$$\widetilde{u}_n = \widetilde{u}_{n-\frac{1}{2}} + \frac{\tau_n}{2}\psi\left(t_{n-1} + \frac{\tau_n}{2}, \frac{\tau_n}{2}, \widetilde{u}_{n-\frac{1}{2}}\right) = u_{n-1} + \tau_n \widetilde{\psi}(t_{n-1}, \tau_n, u_{n-1}) \text{ mit}$$
$$\widetilde{\psi}(t, \tau, z) = \frac{1}{2}\psi\left(t + \frac{\tau}{2}, \frac{\tau}{2}, z + \frac{\tau}{2}\psi\left(t, \frac{\tau}{2}, z\right)\right)$$

Wir setzen folgendes voraus:

$$g(t_{n-1}, \tau_n, u_{n-1}) = C\tau_n^p + O(\tau_n^{p+1}) \text{ und } \widetilde{g}(t_{n-1}, \tau_n, u_{n-1}) = C\left(\frac{\tau_n}{2}\right)^p + O(\tau_n^{p+1})$$

Hieraus ergibt sich

$$u_n - \widetilde{u}_n = \tau_n(g(t_{n-1}, \tau_n, u_{n-1})) - \widetilde{g}(t_{n-1}, \tau_n, u_{n-1}) = \tau_n\left[C\left(1 - \frac{1}{2^p}\right)\tau_n^p + O(\tau_n^{p+1})\right]$$

und weiter für den Fehlerschätzer:

$$\eta_n := \frac{1}{2^p - 1}\frac{1}{\tau_n}(u_n - \widetilde{u}) = C\left(\frac{\tau_n}{2}\right)^p + O(\tau_n^{p+1}) = \widetilde{g}_n + O(\tau_n^{p+1}) = C\left(\frac{\tau_n}{2}\right)^p + O(\tau_n^{p+1}) = \widetilde{g}_n + O(\tau_n^{p+1})$$

Wir haben also einen asymptotisch exakten Fehlerschätzer!

1.5 Adaptiver Algorithmus zu einem Verfahren ψ der Ordnung p

a.) Schritt 0: Wähle $\varepsilon > 0$, $\underline{\tau} \leq \overline{\tau}$, $M > 0$ und $\vartheta \in (0, 1)$. Wähle außerdem Startwerte $u = u_0$, $t = t_0$ und $\tau > 0$.

b.) Schritt 1: $\hat{u} = u + \tau\psi(t, \tau, u)$ (Lösungsvorschlag) und halbierte Schrittweite $\widetilde{u} = u + \tau/2\psi(t, \tau/2, u)$ (Lösungsvorschlag)

$$\widetilde{u} := \widetilde{u} + \frac{\tau}{2}\psi\left(t + \frac{\tau}{2}, \tau, \widetilde{u}\right) \text{ und } \eta = \frac{1}{2^p - 1}\frac{1}{\tau}(\hat{u} - \widetilde{u})$$

c.) Schritt 2: Falls $|\eta| < \varepsilon$ und $t = T$, breche das Verfahren ab. Falls andererseits $t + \tau > T$, setze $\tau = T - t$ und $t = t + \tau$.

d.) Schritt 3: Anpassung der Schrittweite an den Fehlerschätzer:

Falls $t \neq T$ ist, setze $\tau := \vartheta\tau \sqrt[p]{\frac{\varepsilon}{|\eta|}}$. Falls $\tau > \overline{\tau}$ setze $\tau := \overline{\tau}$.

e.) Schritt 4: Falls $|u| > M$ breche das Verfahren ab; ebenso, falls $\tau < \underline{\tau}$, also die Schrittweite unter die vorgegebene kleinste zugelassene Schrittweite sinkt.

Sei a: $\mathbb{R} \mapsto \mathbb{R}^m$ glatt mit $a(\tau) = a(0) + a_p\tau^p + a_{p+1}\tau^{p+1} + \ldots + O(\tau^{p+2})$ (TAYLORentwicklung). $a(\tau)$ sei berechenbar für $\tau > 0$, aber nicht $a(0)$. Extrapoliere dann $a(0)$ aus den Werten $a(\tau_0)$, $a(\tau_1)$, …, $a(\tau_k)$ für $\tau_0 > \tau_1 > \ldots > \tau_k > 0$. Bestimmte ein Polynom $P \in \mathbb{P}_k$ mit $P(\tau_j) = a(\tau_j)$. Hieraus ergibt sich $P(0) = a(0) + O(\tau^{p+1})$.

Beispiel:

Sei $k = 1$, $\tau_0 = \tau$ und $\tau_1 = \tau/2$.

$$a(\tau) = a(0) + a_p\tau^p + O(\tau^{p+1}) \text{ und } a\left(\frac{\tau}{2}\right) = a(0) + a_p\left(\frac{\tau}{2}\right)^p + O(\tau^{p+1})$$

Damit ergibt sich das Interpolationspolynom:

$$P(\tau) = \frac{2^p + a\left(\frac{\tau}{2}\right) - a(\tau)}{2^p - 1} + 2\frac{a(\tau) - a\left(\frac{\tau}{2}\right)}{\tau^p - (2^p - 1)}\tau^p \Rightarrow P(0) = \frac{1}{2^p - 1}\left(2^p a\left(\frac{\tau}{2}\right) - a(\tau)\right) = a(0) + O(\tau^{p+1})$$

1.5.1 Anwendung: Numerisches Differenzieren

$$a(\tau) = \frac{1}{\tau}(f(t+\tau) - f(t))$$

Wir betrachten $f \in C^3$. Mit einer TAYLORentwicklung ergibt sich dann:

$$a(\tau) = \frac{1}{\tau}\left[f(t) + \tau f'(t) + \frac{1}{2}\tau^2 f''(t) + O(t^3) - f(t)\right] = f'(t) + \frac{1}{2}\tau f''(t) + O(\tau^2)$$

Für $k = 1$ und $p = 1$ ergibt sich in Approximation 2.Ordnung:

$$f'(t) = 2a\left(\frac{\tau}{2}\right) - a(\tau) + O(\tau^2) = \frac{1}{\tau}\left[4f\left(t + \frac{\tau}{2}\right) - 3f(t) - f(t+\tau)\right] + O(\tau^2)$$

1.5.2 Anwendung: Symmetrischer Differenzenquotient

Man bezeichnet

$$a(\tau) = \frac{1}{2\tau}[f(t+\tau) - f(t-\tau)]$$

als symmetrischen Differenzenquotienten. Wir nehmen für unsere folgenden Überlegungen an, dass $f \in C^5$. Dann ergibt sich TAYLORentwicklung bis zur fünften Ordnung:

$$\begin{aligned} 2\tau a(\tau) = &\left(f(t) + f'(t)\tau + \frac{1}{2}f''(t)\tau^2 + \frac{1}{6}f'''(t)\tau^3 + \frac{1}{24}f^{(4)}\tau^4 + O(\tau^2)\right) - \\ &+ \left(f(t) - f'(t)\tau + \frac{1}{2}f''(t)\tau^2 - \frac{1}{6}f'''(t)\tau^3 + \frac{1}{24}f^{(4)}(t)\tau^4 + O(\tau^5)\right) \end{aligned}$$

Damit erhalten wir:

$$a(\tau) = f'(t) + \frac{1}{6}f''(t)\tau^2 + O(\tau^4)$$

$$a(0) = \frac{1}{3\tau}\left[2f\left(t + \frac{\tau}{2}\right) + 2f\left(t - \frac{\tau}{2}\right) - \frac{1}{2}f(t+\tau) - \frac{1}{2}f(t-\tau)\right] + O(\tau^4) = f'(t) + O(\tau^4)$$

Damit kann die Ableitung viel genauer berechnet werden als mit dem gewöhnlichen Differenzenquotienten (wegen Rundungsfehlern). Die Voraussetzung $f \in C^5$ setzt jedoch die Anwendung bei partiellen Differentialgleichungen sehr ein, da dort oft $f \notin C^5$ ist.

Satz (1.28):

Sei $f \in C^k([t_0, t_0+T] \times G, \mathbb{R}^m)$ und $u \in C^1([t_0, t_0+T], G)$ eine Lösung von $\dot{u}(t) = f(t, u(t))$. Dann gilt $u \in C^{k+1}([t_0, t_0+T], G)$.

Beweis:

Wir können ganz banal die zweite Ableitung hinschreiben:

$$\ddot{u}(t) = \frac{\mathrm{d}}{\mathrm{d}t}f(t, u(t)) = D_1 f(t, u(t)) + D_2 f(t, u(t))\dot{u}(t) =: F_2(t, u(t), \dot{u}(t))$$

Das ist stetig, damit ist $\ddot{u}$ stetig, also gilt $u \in C^2$ und $F_2 \in C^{k-1}$. Kommen wir nun zur dritten Ableitung:

$$\dddot{u}(t) = \frac{\mathrm{d}}{\mathrm{d}t}F_2(t, u(t), \dot{u}(t)) = D_1 F_2(t, u(t), \dot{u}(t)) + D_2 F_2(t, u(t), \dot{u}(t))\dot{u}(t) + D_3 F_2(t, u(t), \dot{u}(t)) =: F_3(t, u(t), \dot{u}(t), \ddot{u}(t))$$

Dies ist auch stetig, also ist $u \in C^3$ und $F_3 \in C^{k-2}$. Der Beweis ergibt sich dann durch vollständige Induktion. □

Satz (1.29):

Sei f hinreichend glatt und u Lösung einer Anfangswertaufgabe in $[t_0, t_0+T]$. Des weiten sei ψ ein Verfahren der Ordnung p. Dann existieren glatte Funktionen c_0, c_1, c_2, ... mit $c_j(t_0) = 0$ und $u^\tau(t) = u(t) + c_0(t)\tau^p + c_1(t)\tau^{p+1} + \ldots + O(\tau^{p+l})$ für alle k und $t \in t_0 + \mathbb{N}\tau$ $(t \leq T)$. Dabei sei $u^\tau(t_0 + n\tau) = u^\tau(t_0 + (n-1)\tau) + \tau\psi(t_0 + (n-1)\tau, \tau, u^\tau(t_0 + (n-1)\tau))$.

Beweis:

Wir wollen nur den Spezialfall $\psi(t,\tau,z) = f(t,z)$, $p=1$, $k=1$ und $n=1$ (skalare Anwendung, explizites EULERverfahren) beweisen. Sei c_0 Lösung von $\dot{c}_0(t) = D_2 f(t,u(t))c_0(t) - 1/2\ddot{u}(t)$, $c_0(t)=0$. (Dies ist eine lineare homogene Differentialgleichung. Aus der Existenztheorie wissen wir, dass solche Gleichungen eine Lösung besitzen.) Es gilt für festes τ und $t_n = t_0 + n\tau$ und $u_n = u^\tau(t_n)$ mittels TAYLORentwicklung:

$$u(t_{n+1}) = u(t_n) + \dot{u}(t_n)\tau + \frac{1}{2}\ddot{u}(t_n)\tau^2 + \frac{1}{6}\dddot{u}(\xi_n)\tau^3 \text{ mit } \xi_n \in (t_n, t_{n+1})$$

$$u(t_{n+1}) = u(t_n) + f(t_n,u(t_n))\tau + \frac{1}{2}\ddot{u}(t_n)\tau^2 + \frac{1}{6}\dddot{u}(\xi_n)\tau^3$$

Für den Fehler $e_n = u_n - u(t_n)$ ergibt sich wieder mit TAYLORentwicklung:

$$\begin{aligned} e_{n+1} &= e_n - u(t_{n+1}) + u(t_n) + \tau f(t_n,u_n) = e_n + \tau(f(t_n,u_n)) - f(t,u(t_n)) - \frac{1}{2}\ddot{u}(t_n)\tau^2 - \frac{1}{6}\tau^3\dddot{u}(\xi_n) = \\ &= e_n + \tau D_2 f(t_n,u(t_n))e_n + \frac{1}{2}\tau D_2^2 f(t_n,u(t_n))e_n^2 - \frac{1}{2}\tau^2\ddot{u}(t_n) - \frac{1}{6}\tau^3\dddot{u}(\xi_n) \end{aligned}$$

Dann gilt für $\overline{e}_n = 1/\tau e_n$:

$$\overline{e}_{n+1} = \overline{e}_n + \tau(D_2 f(t_n,u(t_n))\overline{e}_n - \frac{1}{2}\ddot{u}(t)) + \tau^2 r_n \text{ mit } r_n = \frac{1}{2}D_2^2 f(t_n,u(t_n))\overline{e}_n^2 - \frac{1}{6}\dddot{u}(\xi_n)$$

Für $c_{n+1} = c_n + \tau h(t_n,c_n)$ mit $h(t,c) = D_2 f(t,u(t))c - 1/2\ddot{u}(t)$ folgt, dass $c_0(t_n) - c_n = O(\tau)$.

$$e_n = \tau\overline{e}_n = \tau\left(c_n + \sum_n \tau^2 r_n\right) = \tau c_0(t_n) + \tau\underbrace{(c_n - c_0(t_n))}_{O(\tau)} + \underbrace{\tau\sum_n \tau^2 r_n}_{\leq \tau^2 T|r_n|} = \tau c_0(t_n) + O(\tau^2) \qquad \square$$

Bemerkung:

Die Entwicklung konvergiert für festes k und $\tau \mapsto 0$, aber im allgemeinen nicht für festes t und τ und $k \mapsto \infty$. In der Anwendung lässt sich die Extrapolation mit dem NEVILLE-Schema auswerten. Sei $\phi_\tau(t,\tau,z)$ ein diskreter Fluss. Definiere $\phi_\tau(t,\tau,z,k) = y$ rekursiv durch $y=z$, für $j=1,\ldots,k$: $y = \phi_\tau(t+(j-1)\tau,\tau/k,y)$. Dann gilt für $p=1$: $k=2^j$.

$U_{00} = \phi_\tau(t,\tau,z)$, $U_{10} = \phi_\tau(t,\tau,z,2)$, $U_{20} = \phi_\tau(t,\tau,z,4)$, $\ldots$, $U_{k0} = \phi_\tau(t,\tau,z,2^k)$

$U_{11} = 2U_{10} - U_{00}$, $U_{21} = 2U_{20} - U_{10}$ und $U_{23} = \frac{1}{3}(4U_{21} - U_{11})$

Allgemein gilt:

$$U_{kj} = \frac{1}{2^j - 1}\left(2^j U_{k,j-1} - U_{k-1,j-1}\right) \text{ mit Fehlerschätzer } \eta_{kj} = \frac{1}{2^j-1}(U_{kj} - U_{k+1,j})\frac{1}{\tau}$$

Wir können hier sowohl die Schrittweite τ als auch die Ordnung k steuern.

1.5.3 Mittelpunktregel

Ein Verfahren heißt „reversibel“, wenn für den diskreten Fluss $\phi_\tau(t+\tau,-\tau,\phi_\tau(t,\tau,z)) = z$ gilt.

Beispiel:

Dies gilt für die Trapezregel (implizit):

$$u_n = u_{n-1} + \tau_n f\left(t + \frac{\tau_n}{2}, \frac{1}{2}u_n + \frac{1}{2}u_{n-1}\right) \Rightarrow u_{n-1} = u_n - \tau_n f\left((t+\tau_n) - \frac{\tau_n}{2}, \frac{1}{2}u_{n-1} + \frac{1}{2}u_n\right)$$

Dies ist eine sehr schöne Eigenschaft, um Energierhaltung in mechanischen Systemen zu gewährleisten.

Bemerkung:

1.) Es gibt keine reversiblen expliziten RUNGE-KUTTA-Verfahren.

2.) Für reversible Verfahren gilt:

$$u^\tau(t) = u(t) + \hat{c}_0(t)\tau^{2q} + \hat{c}_1(t)\tau^{2q+2} + \ldots + O(\tau^{2(q+k)})$$

3.) Explizite Mittelpunktregel $u_1 = u_0 + \tau f(t_0, u_0)$, $u_{n+1} = u_{n-1} + 2\tau f(t_n, u_n)$ für $n = 1, 2, \ldots$
Diese ist reversibel und insbesondere gilt 2.)

4.) Oszillationen in der Fehlerentwicklung können verringert werden:

$$\widetilde{u}_n = \frac{1}{4}(u_{n-1} + 2u_n + u_{n+1}) = u_n + O(\tau^2)$$

Auch für diese Entwicklung gilt immer noch 2.)

Kapitel 2

Lineare Mehrschrittverfahren

Definition:

Sei $\Delta = \{t_n = t_0 + \tau n \text{ mit } n = 0, \ldots, N\}$ ein äquidistantes Gitter in $[t_0, t_0+T]$ und fester Schrittweite $\tau = T/N$. Seien u_0, u_1, …, u_{k-1} Näherungen für die Anfangswertaufgabe $\dot{u} = f(t,u)$ und $u(t_0) = u_0$ an den Stellen t_0, t_1, …, t_{k-1}. Ein „lineares Mehrschrittverfahren“ definiert Werte u_k, …, u_N rekursiv durch folgende Gleichung:

$$\sum_{i=0}^{k} \alpha_{k-1} u_{n-i} = \tau \sum_{i=0}^{k} \beta_{k-1} f_{n-i} \text{ mit } n = k, k+1, \ldots, N \text{ und } f_n = f(t_n, u_n)$$

O.B.d.A. sei $\alpha_k = 1$. Es ist durch α_0, …, α_{k-1}, β_0, …, β_k bekannt; für $\beta_k = 0$ nämlich explizit, sonst implizit.

Beispiel:

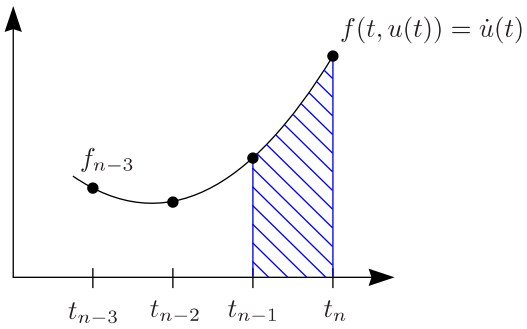
a.) Adams-Bashforth (explizit)

Approximiere die Funktionswerte $f(t,u(t))$ im Intervall $[t_{n-1}, t_n]$ durch ein Polynom $P \in \mathbb{P}_{k-1}$ mit $P(t_{n-i}) = f_{n-i}$ für $i = 1, \ldots, k$. Gilt

$$u(t_n) = u(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(t, u(t))\,\mathrm{d}t$$

approximiere

$$u_n = u_{n-1} + \int_{t_{n-1}}^{t_n} P(t)\,\mathrm{d}t = u_{n-1} + \tau \sum_{i=1}^{k} \beta_{k-i} f_{n-i} \text{ wobei } \alpha_k = 1,\ \alpha_{k-1} = -1 \text{ und } \alpha_{k-i} = 0\,(i > 1)$$



Gilt

$$P(t) = \sum_{i=1}^{k} L_i(t) f_{n-i} \text{ mit } L_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{t - t_{n-j}}{t_{n-j} - t_{n-i}} \text{ wobei } L_i(t_{n-j}) = 0$$

$$\int_{t_{n-1}}^{t_n} L_i(t)\,\mathrm{d}t = \tau \int_0^1 L_i(t_{n-1} + s\tau)\,\mathrm{d}s = \tau \int_0^1 \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{t_{n-1} + s\tau - t_{n-j}}{t_{n-j} - t_{n-i}}\,\mathrm{d}s = \tau \int_0^1 \prod_{\substack{j=1 \\ i \neq j}}^{k} \frac{s - (1-j)}{j - i}\,\mathrm{d}s = \tau \beta_{k-i}$$

Betrachten wir nun Spezialfälle:

i.) $k = 1$: $u_n = u_{n-1} + \tau f_{n-1}$ (explizites EULERverfahren)

ii.) $k = 4$: $u_n = u_{n-1} + \frac{\tau}{24}(55f_{n-1} + 59f_{n-2} + 37f_{n-3} - 9f_{n-4})$

b.) ADAMS-MOULTON (implizit)

Approximiere $f(t, u(t))$ in $[t_{n-1}, t_n]$ durch ein Polynom $P \in \mathbb{P}_k$ mit der Eigenschaft $P(t_{n-i})f_{n-i}$ ($i = 0, \ldots, k$). Hieraus folgt:

$$u_n = u_{n-1} + \tau \sum_{i=1}^{k} \hat{\beta}_{k-i} f_{k-i} \text{ mit } \hat{\alpha}_k = 1,\ \hat{\alpha}_{k-1} = -1 \text{ und } \hat{\alpha}_{k-i} = 0 \text{ für } i > 1$$

Die $\hat{beta}_{k-i}$ sind folgendermaßen definiert:

$$\hat{\beta}_{k-i} = \int_0^1 \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{s-1+j}{j-i}\, \mathrm{d}s$$

Schauen wir uns wieder einige Spezialälle an:

i.) $k = 0$: $u_n = u_{n-1} + \tau f_n$ (implizites EULERverfahren)

ii.) $k = 3$: $u_n = u_{n-1} + \frac{\tau}{24}(9f_n + 19f_{n-1} - 5f_{n-2} + f_{n-3})$

Dieses Verfahren ist aufgrund der größeren Unterschiede zwischen den Koeffizienten weniger auslöschend als das Verfahren ii.) unter Punkt a.)

c.) NYSTRÖM (explizit)

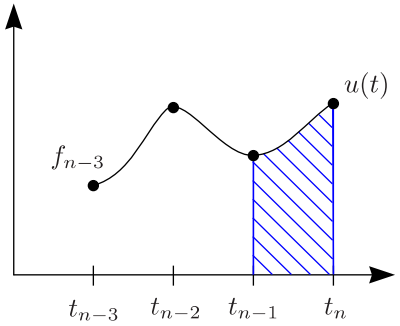
Approximiere $f(t, u(t))$ in $[t_{n-1}, t_n]$ durch $P \in \mathbb{P}_{k-1}$ mit $P(t_{n-i}) = f_{n-i}$ für $i = 1, \ldots, k$. Dies führt zu:

$$u_n = u_{n-2} + \tau \sum_{i=1}^{k} \widetilde{\beta}_{k-i} f_{n-i} \text{ mit } \widetilde{\beta}_{k-i} = \int_0^2 \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{s-2+j}{j-i}\, \mathrm{d}s$$

i.) $k = 1$: $u_n = u_{n-1} + 2\tau f_{n-1}$ (explizite Mittelpunktregel)

d.) BDF (backward difference formula, implizit)

Diese Verfahren werden oft beim Lösen von partiellen Differentialgleichungen eingesetzt.



Approximiere u in $[t_{n-k}, t_n]$ durch $P \in \mathbb{P}_k$ mit $P(t_{n-i}) = u_{n-i}$ für $i = 0, \ldots, k$. Setze $\dot{P}(t) = f_n$; das heißt:

$$\sum_{i=0}^{k} \alpha_{k-i} u_{n-i} = \tau f_n \text{ mit } \tau\alpha_{k-i} = \frac{\mathrm{d}}{\mathrm{d}t} L_i(t_n) \text{ mit } L_i(t) = \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{t - t_{n-j}}{t_{n-i} - t_{n-j}}$$

Für $k = 1$ ergibt sich das implizite EULERverfahren, nämlich $u_n - u_{n-1} = \tau f_n$.

Lemma 2.2:

Sei für f eine L-Bedingung erfüllt und sei $\tau L|\beta_k| < 1$. Dann konvergiert für jeden Startwert $u_n^{(0)} \in G$ die Iteration

$$u_n^{(j)} = -\sum_{i=1}^{k} \alpha_{k-i} u_{n-i} + \tau \sum_{i=1}^{k} \beta_{k-i} f_{n-i} + \tau\beta_k f(t_n, u_n^{(j-1)}) \text{ für } j = 1, 2, \ldots \text{ und } u_n^{(j)} = F(u_n^{(j-1)})$$

gegen die Lösung des Mehrschrittverfahrens.

Beweis:

$$|u_n^{(j+1)} - u_n^{(j)}| = \tau|\beta_k| \left| f(t_n, u_n^{(j)}) - f(t_n, u_n^{(j-1)}) \right| \leq \tau|\beta_k| L |u_n^{(j)} - u_n^{(j-1)}|$$

Dies konvergiert nach dem BANACHschen Fixpunktsatz. □

Definition 2.3:

Zu $u \in C^1([t_n, t_n + T], G)$ definieren wir den „lokalen Diskretisierungsfehler“

$$g_n = \frac{1}{\tau} \sum_{i=0}^{k} \alpha_{k-i} u(t_{n-i}) - \sum_{i=0}^{k} \beta_{k-i} \dot{u}(t_{n-i})$$

Lemma 2.4:

Sei u analytisch , das heißt wir können u als Potenzreihe entwickeln:

$$u(t - s\tau) = \sum_{j=0}^{\infty} \frac{(-s\tau)^j}{j!} \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^j u(t)$$

Dann lässt sich der lokale Diskretisierungsfehler schreiben als:

$$g_n = \frac{1}{\tau} \sum_{j=0}^{\infty} (-1)^j C_j \tau^j \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^j u(t_n) \text{ mit } C_0 = \sum_{i=1}^{k} \alpha_i \text{ und } C_j = \frac{1}{j!} \sum_{i=0}^{j} i^j \alpha_i - \frac{1}{(j-1)!} \sum_{i=0}^{k} i^{j-1} \beta_i \text{ für } i > 0$$

Beweis:

Wenn man davon ausgeht, dass u als obige Potenzreihe darstellbar ist, lässt sich $\dot{u}$ folgendermaßen darstellen:

$$\dot{u}(t - s\tau) = \sum_{j=0}^{\infty} \frac{(-s\tau)^j}{j!} \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^{j+1} u(t)$$

Damit ergibt sich durch Einsetzen:

$$g_n = \sum_{j=0}^{\infty} \left[\frac{1}{\tau} \sum_{i=0}^{k} \alpha_{k-i} \tau^j (-1)^j \frac{i^j}{j!} \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^j u(t_n) - \sum_{i=0}^{k} \beta_{k-i} \tau^j (-1)^j \frac{1}{j!} i^j \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^{j+1} u(t_n) \right]$$

Durch Koeffizientenvergleich folgt die Behauptung. □

2.0.4 Alternative

Setze $f_j = f(t_j, u_j) = f[t_j]$. Dann gilt folgender Satz, den wir früher schon bewiesen haben:

$$f[t_i, \ldots, t_j] = \frac{1}{t_j - t_i} \left(f[t_{i+1}, \ldots, t_j] - f[t_i, \ldots, t_{j-1}] \right)$$

Den Spezialfall, dass die Stützstellen äquidistant zueinander liegen, verwenden wir $\vec{\nabla}^0 f_j = f_j$ und $\vec{\nabla}^i f_j = \vec{\nabla}^{i-1} f_j - \vec{\nabla}^{i-1} f_{j-1}$. Auf einem äquidistanten Gitter $t_n = t_0 + n\tau$ gilt $\vec{\nabla}^k f_j = k! \tau^k f[\tau_{j-k}, \ldots, t_j]$.

a.) ADAMS-BASHFORTH:

Approximiere $f(t, u(t))$ durch

$$P(t) = f_{n-1} + f[t_{n-2}, t_{n-1}](t - t_{n-1}) + \ldots + f[t_{n-k}, \ldots, t_{n-1}](t - t_{n-1}) \ldots (t - t_{n-k+1})$$

Wir setzen nun t als Zwischenstelle zwischen t_{n-1} und t_n ein:

$$P(t_{n-1} + s\tau) = \nabla^0 f_{n-1} + s\nabla^a f_{n-1} + \frac{s(s+1)}{1 \cdot 2} \nabla^2 f_{n-1} + \ldots + \frac{s(s+1)\ldots}{(k-1)!} \nabla^{k-1} f_{n-1} =$$

$$= \sum_{i=0}^{k-1} (-1)^i \binom{-s}{k} \nabla^i f_{n-1} \text{ mit } \binom{s}{n} = 1 \text{ und } \binom{s}{i} = \prod_{j=1}^{i} \frac{s-j}{j}$$

Dies setzen wir in unsere Formel ein:

$$u_n = u_{n-1} + \tau \int_0^1 P(t_{n-1} + s\tau)\,\mathrm{d}s = u_{n-1} + \tau \sum_i \gamma_i \nabla^i f_{n-1} \text{ mit } \gamma_i = (-1)^i \int_0^1 \binom{-s}{i}\,\mathrm{d}s$$

b.) Adams-Moulton:

$$u_n = u_{n-1} + \tau \sum_{i=0}^{k-1} \widetilde{\gamma}_i \nabla^i f_n \text{ mit } \widetilde{\gamma}_i = (-1)^i \int_0^1 \binom{1-s}{i}\,\mathrm{d}s$$

c.) BDF:

Approximiere u in $[t_{n-k}, t_n]$ durch

$$P(t_n + s\tau) = \sum_{i=0}^{k} (-1)^i \binom{-s}{i} \nabla^i u_n$$

Mit

$$\frac{\mathrm{d}}{\mathrm{d}s}(-1)^i \binom{-s}{i}\bigg|_{s=0} = \begin{cases} 0 & \text{für } i = 0 \\ \frac{1}{i} & \text{für } i > 0 \end{cases} \quad \text{und } \dot{P}(t_n) = \tau f_n$$

erhalten wir:

$$\sum_{i=1}^{k} \frac{1}{i} \nabla^i u_n = \tau f_n$$

Für den Spezialfall $k = 2$ ergibt sich dann:

$$\nabla^1 u_n + \frac{1}{2}\nabla^2 u_n = u_n - u_{n-1} + \frac{1}{2}(u_n - 2u_{n-1} + u_{n-2}) = \frac{3}{2}u_n - 2u_{n-1} + \frac{1}{2}u_{n-2} = \tau f_n$$

Definition 2.5:

Ein Mehrschrittverfahren ist konsistent von der Ordnung p, wenn $C_0 = C_1 = \ldots = C_p = 0$ und $C_{p+1} \neq 0$. C_{p+1} bezeichnet man auch als **Fehlerkonstante**.

Lemma 2.6:

Ein Mehrschrittverfahren ist konsistent von der Ordnung p, wenn der lokale Diskretisierungsfehler für Polynome vom Grad p verschwindet.

Beweis:

Durch Taylor-Entwicklung an der Stelle t_{n-k} erhalten wir:

$$u(t) = \sum_{j=0}^{p} \frac{1}{j!} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^j u(t_{n-k})(t - t_{n-k})^j + R(t) \text{ mit } R(t) = \frac{1}{p!} \int_{t_{n-k}}^{t} (t-s)^p \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{p+1} u(s)\,\mathrm{d}s$$

Damit können wir den lokalen Diskretisierungsfehler auswerten:

$$|g_n| = \left|0 + \frac{1}{\tau}\sum_{i=0}^{k} R(t_{n-k})\alpha_{k-i} - \sum_{i=0}^{k} \beta_{k-i} \frac{\mathrm{d}}{\mathrm{d}t} R(t_{n-i})\right| \leq$$

$$\leq \left\| \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{p+1} u \right\|_\infty \left[\frac{1}{\tau} \sum_{i=0}^{k} |\alpha_{k-i}| \int_{t_{n-k}}^{t_n} (t_n - s)^p\,\mathrm{d}s + \sum_{i=1}^{k} |\beta_{k-i}| (t_n - t_{n-k})^p \right] \leq C\tau^p$$

□

Konsistenz alleine genügt jedoch nicht! Warum ist das so? Betrachten wir dazu folgendes Beispiel:

$$u_n = 4u_{n-1} + 5u_{n-2} + \tau(4f_{n-1} + 2f_{n-2})$$

$\alpha_2 = 1$, $\alpha_1 = 4$, $\alpha_0 = -5$ und $\beta_2 = 0$, $\beta_1 = 4$, $\beta_0 = 2$

Damit ergibt sich:

$$C_0 = \sum_i \alpha_i = 1 + 4 - 5 = 0$$

$$C_1 = \sum_i i \cdot \alpha_i - \sum_i \beta_i = 0 \cdot (-5) - 1 \cdot 4 + 2 \cdot 1 - (2 + 4) = 0$$

$$C_2 = \frac{1}{2} \sum_i i^2 \cdot \alpha_i - \sum_i i \cdot \beta_i = \frac{1}{2}(1 \cdot 4 + 4 \cdot 1) - (0 \cdot 2 + 1 \cdot 4) = 0$$

$$C_3 = \frac{1}{6} \sum_i i^3 \cdot \alpha_i - \sum_i i^2 \cdot \beta_i = \frac{1}{6}(1 \cdot 4 + 8 \cdot 1) - \frac{1}{2}(1 \cdot 4) = 0$$

$$C_4 = \frac{1}{24} \sum_i i^4 \cdot \alpha_i - \frac{1}{6} \sum_i i^3 \cdot \beta_i = \frac{1}{24}(1 \cdot 4 + 16 \cdot 1) - \frac{1}{6}(1 \cdot 4) \neq 0$$

Damit gilt $p = 3$. Wähle nun $f(t, u) = -u$. Setzen wir dies in das Schema ein, so erhalten wir die Vorschrift $u_n = -4u_{n-1} + 5u_{n-2} - \tau(4u_{n-1} + 2u_{n-2})$. Dies kann man entsprechend als Differenzengleichung schreiben:

$$u_n + (4 + 4\tau)u_{n-1} + (-5 + 2\tau)u_{n-2} = 0$$

Für diese Differenzengleichung machen wir den Ansatz $u_n = z^n$. Damit ergibt sich die quadratische Gleichung $z^2 + (4 + 4\tau)z + (-5 + 2\tau) = 0$. Die Lösungen dieser Gleichung lauten $z_{1/2} = -2 + 2\tau \pm 3\sqrt{1 + \alpha(\tau)}$, womit weiterhin folgt:

$$u_n = c_1 z_1^n + c_2 z_2^n \Rightarrow \lim_{\tau \mapsto 0} |u_n| = |c_2| \cdot |-5|^n \mapsto \infty \text{ für } n \mapsto \infty$$

Definition 2.7:

Zu einem Mehrschrittverfahren definieren wir das „charakteristische Polynom“ durch:

$$\chi(z) = \sum_{i=0}^{k} \alpha_i z^i$$

Definition 2.8:

Ein Mehrschrittverfahren heißt **stabil** (0-stabil), wenn für alle Nullstellen λ_i von χ gilt, dass $|\lambda_i| \leq 1$ ist. Für $|\lambda_i| = 1$ soll λ_i einfach sein. (Das heißt: $\chi'(\lambda_i) \neq 0$, wenn $|\lambda_i| = 1$ und $\chi(\lambda_i) = 0$)

Satz 2.9:

Wenn ein Mehrschrittverfahren nicht stabil ist, dass ist die diskrete Lösung für $\tau \mapsto 0$ und für fast alle Anfangswerte unbeschränkt.

Beweis:

Für $\tau \mapsto 0$ konvergiert u_n gegen eine Lösung von der linearen Differenzengleichung $\sum_i \alpha_{k-i} z_{n-i} = 0$. Die Behauptung folgt dann aus (2.10). □

Satz 2.10:

Sei

$$\chi(\lambda) = \sum_{i=0}^{k} \alpha_i \lambda^i = \prod_{\nu=1}^{r} (\lambda - \lambda_\nu)^{m_\nu} \text{ für } \lambda_\nu \neq \lambda_\mu \text{ für } \nu \neq \mu \text{ und } \sum_{\nu=1}^{r} m_\nu = k$$

Dann hat jede Lösung $(z_n)_{n=0,1,2,\ldots}$ der linearen Differenzengleichung

$$\sum_{i=0}^{k} \alpha_{k-i} z_{n-i} \text{ für } n = k,\ k+1, \ldots$$

die Form:

$$z_n = \sum_{\nu=1}^{r} \sum_{j=0}^{m_\nu - 1} c_{\nu,j} \frac{n!}{(n-j)!} \lambda_\nu^n$$

Existiert eine mehrfache Nullstelle λ_i des charakteristischen Polynoms mit $|\lambda_i| = 1$, so wächst z_n proportional zu $n!$, ist also nicht mehr beschränkt. Die Koeffizienten $c_{\nu,j}$ sind eindeutig durch z_0, z_1, …, z_{k-1} bestimmt.

Beweis:

Die Lösungen bilden einen Vektorraum der Dimension k. Also genügt es zu zeigen, dass

$$z_n = \frac{n!}{(n-j)!} \lambda_\nu^n \text{ für } 0 \leq j \leq m_\nu$$

Lösung ist. Betrachten wir zuerst den Fall $j = 0$:

$$\chi(\lambda_\nu) = 0 \Rightarrow \sum_{i=0}^{k} \alpha_{k-i} \lambda_\nu^{n-i} = \lambda_\nu^{n-k} \chi(\lambda_\nu) = 0$$

Für den Fall $j = 1$ ergibt sich:

$$\begin{aligned} \sum_{i=0}^{k} \alpha_{k-i}(n-i)\lambda_\nu^{n-i} &= \lambda_\nu^{n-k} \sum_{i=0}^{k} \alpha_{k-i}(n-k)\lambda_\nu^{k-i} + \lambda_\nu^{n-k+1} \sum_{i=0}^{k} \alpha_{k-i}(k-i)\lambda_\nu^{k-i-1} = \\ &= \lambda_\nu^{n-k}(n-k)\chi(\lambda_\nu) + \lambda_\nu^{n-k+1}\chi'(\lambda_\nu) = 0 \end{aligned}$$

Dies folgt daraus, dass wir eine doppelte Nullstelle haben. Auf diese Weise kann dies für $j > 2$ fortgesetzt werden. □

Beispiel:

a.) $\chi(z) = z^k - z^{k-1} = z^{k-1}(z-1)$ (Dies gilt auch für Verfahren b.)

c.) $\chi(z) = z^k - z^{k-2} = z^{k-2}(z-1)(z+1)$

d.) $k = 2$: $\chi(\lambda) = \frac{3}{2}\lambda^2 - 2\lambda + \frac{1}{2} = \frac{1}{2}(3\lambda - 1)(\lambda - 1)$

Warnung: BDF ist für hohe Ordnungen nicht stabil (da es Nullstellen λ_i mit $|\lambda_i| > 1$ gibt)!

Satz 2.11:

Sei $u \in C^1([t_0, t_0 + T], G)$ Lösung der Anfangswertaufgabe (1.1) und sei für f eine passende L-Bedingung erfüllt und sei $\tau L|\beta_k| < 1$. Dann gilt für ein stabiles Mehrschrittverfahren: Es existieren Konstanten K, $L^* > 0$ mit folgenden Eigenschaften:

$$|u_n - u(t_n)| \leq K \left[\max_{j=0,\ldots,k-1} |u_i - u(t_i)| \exp(L^*(t_n - t_0)) + \max_{j=k,\ldots,n} |g_n| \left(\frac{\exp(L^*(t_n - t_0) - 1)}{L^*} \right) \right]$$

Lemma 2.12:

Sei $A \in \mathbb{R}^{k,k}$ mit dem Spektralradius $\varrho(A) = \varrho = \max_{\mu \in \sigma(A)} |\mu|$. Wenn für jeden Eigenwert $\lambda \in \sigma(A)$ mit $|\lambda| = \varrho$ die algebraische Vielfachheit genau der geometrischen Vielfachheit entspricht, dann existiert eine symmetrische und positiv definite Matrix $S \in \mathbb{R}^{k,k}$ mit der Eigenschaft $|Az|_S \leq \varrho |z|_S$ mit $z \in \mathbb{R}^k$, wobei $|z|_S = \sqrt{z^\intercal S z}$, $|A|_S = \sup_{|z|_S = 1} |Az|_S$, $|z| = \sqrt{z^\intercal z}$ und $|B| = \sup_{|z|=1} |Bz|$.

Beweis:

Zu einer Matrix A existiert eine Matrix $Q \in \mathbb{C}^{k,k}$ mit der Eigenschaft $Q^{-1}AQ = \mathcal{J}$, wobei $\mathcal{J}$ die JORDANsche Normalform ist:

$$\mathcal{J} = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_r \end{pmatrix} \text{ mit den Blöcken } J_i = \begin{pmatrix} \lambda_j & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_j & 1 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & \lambda_j & 1 & \ddots & \ddots & 0 \\ 0 & \ddots & 0 & \ddots & \ddots & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \lambda_j & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_j \end{pmatrix}$$

Zu $\varepsilon > 0$ setze $E = \operatorname{diag}(\varepsilon^i)$ wie folgt (Standard-Skalierungstrick):

$$E = \begin{pmatrix} \varepsilon & & & 0 \\ & \varepsilon^2 & & \\ & & \varepsilon^2 & \\ 0 & & & \ddots \end{pmatrix}$$

Hieraus ergibt sich:

$$E^{-1}\mathcal{J}E = \begin{pmatrix} J_1^\varepsilon & & 0 \\ & \ddots & \\ 0 & & J_r^\varepsilon \end{pmatrix} \text{ mit } J_j^\varepsilon = \begin{pmatrix} \lambda_j & \varepsilon & 0 & 0 & 0 \\ 0 & \ddots & \varepsilon & \ddots & \ddots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & \lambda_j & \varepsilon \\ 0 & 0 & 0 & 0 & \lambda_j \end{pmatrix} \equiv \lambda_j I + \varepsilon N_j \text{ mit } N_j = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & \ddots & 1 & \ddots & \ddots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Damit ergibt sich:

$$|E^{-1}\mathcal{J}E| = \max|J_j^\varepsilon| \le \max\left\{ \max_{|\lambda_j|=\varrho} |\lambda_j|, \max_{|\lambda_j|<\varrho} |\lambda_j| + \varepsilon|N_j| \right\}$$

Wähle $\varepsilon > 0$ so klein, dass $|\lambda_j| + \varepsilon|N_j| \le \varrho$ für alle $|\lambda_j| < \varrho$. Wähle $S = Q^{-\intercal}E^{-\intercal}E^{-1}Q^{-1}$ per Konstruktion symmetrisch und positiv semidefinit. So können wir $|Az|_S$ ausrechnen:

$$\begin{aligned} |Az|_S &= \sqrt{z^\intercal A^\intercal Q^{-\intercal}E^{-\intercal}E^{-1}Q^{-1}Az} = |E^{-1}Q^{-1}Az| = |E^{-1}Q^{-1}AQEE^{-1}Q^{-1}z| = |E^{-1}\mathcal{J}EE^{-1}Q^{-1}z| \le \\ &\le |E^{-1}\mathcal{J}E||E^{-1}Q^{-1}z| \le \varrho|z|_S \end{aligned}$$

□

Beweis von 2.11:

Sei $m = 1$. Aus $\alpha_k = 1$ folgt:

$$\begin{aligned} e_n &= u(t_n) - u_n = \\ &= \tau g_n - \sum_{i=1}^{k} \alpha_{k-i} u(t_{n-i}) + \tau \sum_{i=0}^{k} \beta_{k-i} f(t_{n-i}, u(t_{n-i})) + \sum_{i=1}^{k} \alpha_{k-i} u_{n-i} - \tau \sum_{i=0}^{k} \beta_{k-i} f(t_{n-i}, u_{n-i}) = \\ &= \tau g_n - \sum_{i=1}^{k} \alpha_{k-i} e_{n-i} + \tau \sum_{i=0}^{k} \beta_{k-i} q_{n-i} e_{n-i} \text{ mit } q_n = \begin{cases} \dfrac{f(t_n, u(t_n)) - f(t_n, u_n)}{e_n} & \text{für } e_n \neq 0 \\ 0 & \text{für } e_n = 0 \end{cases} \end{aligned}$$

Hieran können wir gut erkennen, dass q_n abschätzbar ist:

$$|q_n| \le \frac{|f(t_n, u(t_n)) - f(t_n, u_n)|}{|u(t_n) - u_n|} \le L$$

Damit folgt also:

$$(1-\tau\beta_n q_n)e_n = \tau g_n - \sum_{i=1}^{k} \alpha_{k-i} e_{n-i} + \tau \sum_{i=1}^{k} \beta q_{n-i} e_{n-i} = -\sum_{i=1}^{k} \alpha_{k-i} e_{n-i} + \tau \cdot r_n \text{ mit } r_n = g_n + \sum_{i=1}^{k} \beta_{k-i} q_{n-i} e_{n-i}$$

Bilde

$$\underline{e}_n = \begin{pmatrix} k_n \\ k_{n-1} \\ \vdots \\ k_{n-k+1} \end{pmatrix} \in \mathbb{R}^k$$

womit wir den obigen Ausdruck folgendermaßen schreiben können:

$$D_n \underline{e}_n = A\underline{e}_{n-1} + \tau \underline{r}_n \text{ mit } \underline{r}_n = \begin{pmatrix} r_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \underline{g}_n = \begin{pmatrix} g_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} \text{ und } D_n = \operatorname{diag}(1 + \tau\beta_k q_n, 1, \ldots, 1)$$

Mit diesen Definitionen ist es möglich, die Matrix A zu schreiben in der Form:

$$A = \begin{pmatrix} -\alpha_{k-1} & -\alpha_{k-2} & \ldots & \ldots & -\alpha_0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Das charakteristische Polynom lautet damit $\chi(z) = \sum_{i=1}^{k} \alpha_i z^i$, was man am Beispiel

$$A = \begin{pmatrix} -\alpha_2 & -\alpha_1 & -\alpha_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

nachvollziehen kann. Wegen Lemma 2.12 existiert eine symmetrische, positiv definite Matrix S mit $|Az|_S \leq |z|_S$, woraus $|A|_S \leq 1$ folgt. Normäquivalenz:

$c|z| \leq |z|_S \leq C|z|$ mit $C > 1 > c > 0$

Schauen wir uns nun die Norm folgender Matrix an:

$|D_n - I| = |\tau\beta_n q_n| \leq \tau|\beta_k|L < 1$

Damit können wir die Neumannsche Reihe anwenden und $D_n = I - (I - D_n)$ ist invertierbar.

$$|D_n^{-1}| \leq \left| \sum_{j\geq 0} (I - D_n)^j \right| \leq \sum_{j\geq 0} (\tau|\beta_k|L)^j = \frac{1}{1-\tau|\beta_k|L} \Rightarrow |D_n^{-1} - I| \leq \frac{\tau|\beta_k|L}{1-\tau|\beta_k|L}$$

Hieraus ergibt sich dann folgende Abschätzung:

$$|D_n^{-1}|_S \leq |I|_S + |D_n^{-1} - I|_S = 1 + \sup_{z\neq 0} \frac{|D_n^{-1}z - z|_S}{|z|_S} \leq 1 + \frac{C}{c}|D_n^{-1} - I| \leq 1 + \tau L_0 \text{ mit } L_0 = \frac{|\beta_k|L}{1-\tau|\beta_k|L}\frac{C}{c}$$

$$|r_n|_S \leq |g_n|_S + |r_n - g_n|_S \leq |g_n|_S + C|r_n - q_n| \leq |g_n|_S + C\sum_{i=1}^{k} |\beta_{k-i}||q_{n-i}||\underline{e}_{n-i}| \leq$$

$$\leq |g_n|_S + CL\left(\sum_{i=1}^{k} |\beta_{k-i}|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{k} |\underline{e}_{n-i}|^2\right)^{\frac{1}{2}} \leq |g_n|_S + L_1|\underline{e}_{n-i}|_S \text{ mit } L_1 = \frac{C}{c}L\sqrt{\sum_{i=1}^{k} |\beta_{k-i}|^2}$$

Somit können wir weiterrechnen:

$$|\underline{e}_n|_S = |D_n^{-1}(A\underline{e}_{n-i} + \tau\underline{r}_n)|_S \leq |D_n^{-1}|_S \left(|A|_S|\underline{e}_{n-i}|_S + \tau|\underline{r}_n|_S\right) \leq$$
$$\leq (1+\tau L_0)\left(|\underline{e}_{n-1}|_S + \tau(|g_n| + L_1|\underline{e}_{n-i})|\right) \leq (1+\tau L^*)(|\underline{e}_{n-i}|_S + \tau|g_n|_S) \text{ mit } L^* = L_0 + L_1 + \tau L_0 L_1$$

Jetzt ist es uns möglich, das diskrete Gronwall-Lemma anzuwenden:

$$|\underline{e}_n|_S \leq |\underline{e}_k|_S \exp(L^*(t_n - t_{n-k})) + \max_{j=k,\ldots,n} |g_j|_S \frac{\exp(L^*(t_n - t_k)) - 1}{L^*}$$

$$|u_n - u(t_n)| \leq |\underline{e}_n| \leq \frac{1}{c}|\underline{e}_n|_S \leq \frac{C}{c}\left[|\underline{e}_k| \exp(L^*(t_n - t_{n-k})) + \max_{j=k,\ldots,n} |g_j| \frac{\exp(L^*(t_n - t_k)) - 1}{L^*}\right]$$

$$|\underline{e}_n| \leq \sqrt{k} \max_{j=0,\ldots,k-1} |u_j - u(t_j)| \text{ mit } K = \frac{C}{c}\sqrt{k}$$ □

Aus dem Beweis sollte man sich folgende Tricks merken:

1.) Wähle geeignete Normen! Man kann mit ihnen viel einfacher rechnen als mit Spektralreihen.

2.) Nutze die Neumannsche Reihe aus. Das ist ein sehr gutes Hilfsmittel, um inverse Matrizen abschätzen.

3.) Verwenden des Gronwall-Lemmas

2.0.5 Stabilität

Wir schauen uns die kontinuierlichen Lösungen u und v des Problems $\dot{u} = f(t, u(t))$ bzw. $\dot{v} = f(t, v(t))$ an. Das kontinuierliche Gronwall-Lemma sagt uns dann, dass $|u(t) - v(t)| \leq |u(t_0) - v(t_0)| \exp(L(t - t_0))$ gilt. Einschrittverfahren:

$u_n = u_{n-1} + \tau\psi(t_{n-1}, \tau_n, u_{n-1})$ und $v_n = v_{n-1} + \tau\psi(t_{n-1}, \tau_n, v_{n-1})$

Auch hier gilt analog $|u_n - v_n| \leq |u_0 - v_0| \exp(\Lambda(t - t_0))$, wobei $\Lambda \mapsto L$ für $\tau \mapsto 0$. (Die Kondition wird also kodiert durch die Lipschitz-Konstante.) Bei den Mehrschrittverfahren macht es Sinn, die Lipschitz-Bedingungen der Vektoren $\underline{u}_n$ bzw. $\underline{v}_n$ anzugeben:

$$|\underline{u}_n - \underline{v}_n|_S \leq |\underline{u}_{k-1} - \underline{v}_{k-1}| \exp(L^*(t - t_0)) \text{ wobei } \underline{u}_{k-1} = \begin{pmatrix} u_0 \\ \vdots \\ u_{k-1} \end{pmatrix} \in \mathbb{R}^k$$

Hierbei gilt $L^* \mapsto \infty$ für $\tau L|\beta_k| \mapsto 1$.

Folgerung 2.13:

Wenn das Mehrschrittverfahren konsistent von der Ordnung p ist und wenn $u_1, \ldots, u_{k-1}$ mit einem Einschrittverfahren der Ordnung $p-1$ berechnet werden, dann ist $|u(t_n) - u_n| = O(\tau^p)$.

Beweis:

Für die ersten Schritte eines Einschrittverfahrens haben wir folgende Abschätzung:

$$|u(t_i) - u_i| \leq \max_{j=1,\ldots,i} |g_j| \frac{\exp(\Lambda(t_i - t_0)) - 1}{\Lambda} \leq O(\tau^{p-1})\left(i\tau + \frac{1}{2}\Lambda(i\tau)^2 + \ldots\right) \leq$$
$$\leq O(\tau^{p-1})\left(k\tau + O(\tau^2)\right) = O(\tau^p) \text{ für } i < k$$

□

Beispiel:

$u_1 = u_0 + \tau f_0$

$u_2 = u_0 + 2\tau f_1$

$u_3 = -\alpha_2 u_2 - \alpha_1 u_1 - \alpha_0 u_0 + \tau \sum \beta f$

Satz 2.14:

Für stabile Mehrschrittverfahren der Ordnung p gilt:

$$p \begin{cases} k+2 & \text{für } k \text{ gerade} \\ k+1 & \text{für } k \text{ ungerade} \\ k & \text{für } \beta_k = 0 \text{ explizit} \end{cases}$$

Beispiel:

Wir machen folgenden Ansatz:

$u_n + \alpha_1 u_{n-1} + \alpha_0 u_{n-2} = \tau\left(\beta_2 f_n + \beta_1 f_{n-1} + \beta_0 f_{n-2}\right)$ wobei $\alpha_2 = 1$, $k = 2$

Wir lesen $C_0 = 1+\alpha_1+\alpha_0 = 0$ ab. Setzen wir dann $\alpha = \alpha_0$ und $\alpha_1 = -1-\alpha$, so ergibt sich das charakteristische Polynom:

$\chi(z) = z^2 - (1+\alpha)z + \alpha = (z-1)(z-\alpha)$

Das Verfahren ist also stabil für $-1 \leq \alpha < 1$.

$C_1 = \alpha_1 + 2 - (\beta_0 + \beta_1 + \beta_2)$ und $C_2 = \frac{1}{2}(\alpha_1 + 4) - (\beta_1 + 2\beta_2)$

$C_3 = \frac{1}{6}(\alpha_1 + 8) - \frac{1}{2}(\beta_1 + 4\beta_2)$ und $C_4 = \frac{1}{24}(\alpha_1 + 16) - \frac{1}{6}(\beta_1 + 8\beta_2)$

Aus $C_1 = C_2 = C_3 = 0$ ergibt sich dann:

$\beta_0 = -\frac{1}{12}(1+5\alpha)$, $\beta_1 = \frac{2}{3}(1-\alpha)$, $\beta_2 = \frac{1}{12}(5+\alpha)$ und $C_4 = \frac{1}{24}(1+\alpha)$

Für $\alpha = -5$ ist der Verfahren explizit der Ordnung $p = 3$. Damit ist aber das Stabilitätskriterium nicht erfüllt, womit das Verfahren instabil ist. Für $\alpha = -1$ gilt p_4 und das Verfahren ist optimal!

$u_n = u_{n-2} + \frac{1}{3}\tau(f_n + 4f_{n-1} + f_{n-2})$

Dies ist gerade die SIMPSON-Regel.

2.0.6 Prediktor-Korrektor-Verfahren

- Prediktor mit explizitem Mehrschrittverfahren:

$$u_n^0 = -\sum_{i=1}^{k} \alpha_{k-i} u_{n-i} + \tau \sum_{i=1}^{k} \beta_{k-i} f_{n-i}$$

- Korrektor mit implizitem Mehrschrittverfahren:

$$u_n^j = -\sum_{i=1}^{k} \widehat{\alpha}_{k-i} u_{n-i} + \tau \widehat{\beta}_k f(t_n, u_n^{j-i}) + \tau \sum_{i=1}^{k} \widehat{\beta}_{k-i} u_{k-i}$$

Wir machen die Korrektur nur so lange bis zur Konvergenz. Konvergenzkriterium ist hier $|u_n^j - u_n^{j-1}| \leq c\tau^p$.

Lemma 2.15:

Sei p^{p} die Fehlerordnung des Prediktors und sei p^{c} die Ordnung des Korrektors. Dann ist $p \equiv \min\{p^{\mathrm{p}} + j, p^{\mathrm{c}}\}$ die Ordnung des Prediktor-Korrektur-Verfahrens.

Beispiel:

Betrachten wir:

$u_n^0 = u_{n-1} + \frac{\tau}{24}\left(55f_{n-1} - 59f_{n-2} - 37f_{n-3} - 9f_{n-4}\right)$ und $u_n^j = u_{n-1} + \frac{\tau}{24}\left(9f(t_n, u_n^{j-1}) + 19f_{n-1} - 5f_{n-2} + f_{n-3}\right)$

Wir schätzen den Fehler ab unter der Annahme, dass $|u(t_{n-i} - u_{n-i}|$ klein ist für $i = 1, \ldots, k$.

$$g_n = \frac{1}{\tau}\left[u(t_n) - u_n\right] \approx C_{p+1}\tau^p \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{p+1} u(t_{n-k}) + O(\tau^{p+1})$$

Für den Prediktor wählen wir das Adams-Bachforth-Verfahren (mit $k = 4$, $p = 4$) und den den Korrektor das Adams-Moulton-Verfahren (mit $k = 3$, $p = 4$). Man kann dann berechnen, dass folgendes gilt:

$$C_5^{\mathrm{p}} = \frac{251}{720} \text{ und } C_5^{\mathrm{c}} = -\frac{19}{720}$$

Damit erhalten wir

$$u(t_n) - u_n^0 = C_5^{\mathrm{p}}\tau^5 \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^5 u(t_{n-k}) + O(\tau^6)$$

$$u(t_n) - u_n^j = C_5^{\mathrm{c}}\tau^5 \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^5 u(t_{n-k}) + O(\tau^6)$$

womit sich ergibt:

$$u_n^j - u_n^0 = (C_5^{\mathrm{c}} - C_5^{\mathrm{p}})\tau^5 \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^5 u(t_{n-k}) + O(\tau^6)$$

$$g_n^{\mathrm{c}} = \frac{1}{\tau}(u(t_n) - u_n^j) + O(\tau^5) = C_5^{\mathrm{c}}\tau^4 \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^5 u(t_{n-k}) + O(\tau^5) = \frac{C_5^{\mathrm{c}}}{C_5^{\mathrm{c}} + C_5^{\mathrm{p}}}\frac{1}{\tau}(u_n^j - u_n^0) + O(\tau^5)$$

Wähle nun den **Fehlerschätzer**:

$$\eta_n = \frac{1}{\tau}\frac{C_5^{\mathrm{c}}}{C_5^{\mathrm{p}} - C_5^{\mathrm{c}}}(u_n^j - u_n^0)$$

2.0.7 Schrittweitensteuerung

1.) Wähle $\varepsilon > 0$.

2.) Speichere immer $i = 0, \ldots, K$ Werte ($K = 2k$)

3.) Für $|\eta_n| > \varepsilon$ setze $\tau := 1/2\tau$. Berechne Zwischenwerte mit Interpolation der Ordnung p.

4.) Für $|\eta_n| < 2^{-p}\varepsilon$. Wenn genügend Werte gespeichert sind, rechne weiter mit f_n, f_{n-2}, f_{n-4} und $\tau := 2\tau$.

Kapitel 3

Steife Differentialgleichungen

Beispiel:

Das wichtigste Beispiel ist die lineare Differentialgleichung $\dot{u}(t) = Au(t)$, wobei $A \in \mathbb{R}^{m,m}$ und $u(0) = u_0$ (autonom). Das lokale Verhalten eines jeden dynamischen Systems ist durch eine Gleichung dieser Form (mit A als JACOBI-Matrix von f) gegeben. Die Lösung können wir direkt angeben, nämlich:

$$u(t) = \exp(tA)u_0 \text{ mit } \exp(B) = \sum_{k\geq 0} \frac{1}{k!} B^k$$

Da die Reihenentwicklung der Exponentialfunktion absolut konvergent ist, funktioniert sie auch für Matrizen. Matrizen bilden einen nichtkommutativen Ring, also gilt:

$$\exp(A+B) = \exp(A)\exp(B) \Leftrightarrow AB = BA$$

Matrizen, die miteinander vertauschen, lassen sich simultan diagonalisieren. Es gilt außerdem, falls $\mathrm{Re}(\lambda) < 0$ ist für alle $\lambda \in \sigma(A)$, dass $\lim_{t\mapsto\infty} u(t) = 0$. Genauer gilt folgende Abschätzung:

$$|u(t)| \leq C_\varrho \exp(-\varrho t) \text{ für alle } \mathrm{Re}(\sigma(A)) < -\varrho < 0$$

Falls $\mathrm{Re}(\lambda) \leq 0$ für alle $\lambda \in \sigma(A)$ und zusätzlich für alle $\mathrm{Re}(\lambda) = 0$ gilt, dass die algebraische Vielfachheit mit der geometrischen Vielfachheit von λ übereinstimmt, dann gilt $|u(t)| \leq |u_0|$ für alle $t \geq 0$.

Beispiel:

Wir betrachten die Matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ mit } \sigma(A) = \{0\}$$

Die Lösung der Differentialgleichung

$$\dot{u}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} u(t) \text{ mit } u(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

ist gegeben durch:

$$u(t) = \begin{pmatrix} t \\ 1 \end{pmatrix} \text{ mit } \dot{u}(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Ist die Bedingung, dass algebraische Vielfachheit und geometrische Vielfachheit übereinstimmen, also nicht erfüllt, so ist die Lösung nicht stabil. (Die obige Lösung wächst mit t, geht also gegen ∞ für $t \mapsto \infty$.)

Beispiel:

Für das explizite RUNGE-KUTTA-Verfahren (für $f(t,z) = Az$) gilt:

$$k_1 = Az,\ k_2 = A(z + \tau a_{21} k_1) = (A + \tau a_{21} A^2)z$$

$$z + \tau\psi(t,\tau,z) = z + \tau \sum_j b_j \cdot k_j = P(\tau A)z \text{ mit } P \in \mathbb{P}_s,\ P \notin \mathbb{R}$$

Also gilt:

$$u_n = P(\tau A)u_{n-1} = P(\tau A)^n u_0$$

Wegen $P(\tau A) \mapsto \infty$ für $\tau \mapsto \infty$ muss τ beschränkt sein!

3.1 Implizite Runge-Kutta-Verfahren

Definition 3.1:

Ein RUNGE-KUTTA-Verfahren der Stufe s ist durch das BUTCHER-Schema

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & \\ & b^\intercal \end{array} \text{ mit } c,\, b \in \mathbb{R}^s,\, \mathcal{A} \in \mathbb{R}^{s,s}$$

und durch

$$k_i = f\left(t + c_i\tau, z + \tau \sum_{j=1}^{s} a_{ij} \cdot k_j\right),\ \psi(t,r,z) = \sum_{j=1}^{s} b_j \cdot k_j \text{ für } i = 1, \ldots, s$$

definiert.

Spezialfälle:

a.) $a_{ij} = 0$ für $j \geq i$ (explizite Verfahren)

b.) $a_{jj} = b_j$ (Diese Eigenschaft bezeichnet man als „steif“.)

c.) $a_{ij} = 0$ für $j > i$ (diagonal-implizite Verfahren)
Diese Verfahren spielen eine Rolle bei partiellen Differentialgleichungen. k_i in diesem Falle ist Lösung von $g_i(v) = 0$ mit

$$g_i(v) := v - f\left(t + c_i\tau, z + \tau \sum_{j=1}^{i-1} a_{ij}k_j + \tau a_{ii} v\right)$$

d.) Im allgemeinen löse $G(k) = 0$ mit

$$G(k) = \begin{pmatrix} k_1 - f\left(t + c_1\tau, z + \tau \sum_j a_{ij} b_j\right) \\ \vdots \\ k_s - f\left(t + c_s\tau, z + \tau \sum_j a_{sj} b_j\right) \end{pmatrix} \in \mathbb{R}^{s-n}$$

Die Idee ist, die Differentialgleichung in eine Integralgleichung umzuschreiben:

$$u(t_n) = u(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(t, u(t))\,\mathrm{d}t$$

Diese wird durch

$$u_n = u_{n-1} + \tau \sum_{i=1}^{s} b_i \underbrace{f(t_{n,i}, u_{n,i})}_{:=k_j} \text{ mit } t_{n,i} = t_{n-1} + c_i\tau\, u_{n,i} = u_{n-1} + \tau \sum_{j=1}^{s} a_{ij}k_j$$

approximiert.
Die Konsistenzordnung von RUNGE-KUTTA-Verfahren können wir durch Vergleich mit der „TAYLOR-Methode“ bestimmen.

Lemma 3.2:

Für $m = 1$ ist die „TAYLOR-Methode“

$$u_n = u_{n-1} + \tau \sum_{j=1}^{p} \frac{\tau^{j-1}}{j!} f^{(j-1)}(t_{n-1}, u_{n-1})$$

konsistent von der Ordnung p. Dabei ist $f^{(0)}(t,z) = f(t,z)$ und weiter

$f^{(1)}(t,z) = D_1 f(t,z) + D_2 f(t,z) f(t,z)$

$f^{(2)}(t,z) = D_1^2 f(t,z) + D_2^2 f(t,z)(f(t,z))^2 + D_2 f(z,t) f(z,t) f^{(1)}$ usw.

Vorsicht: Im allgemeinen kann für $m > 1$ die Ordnung von RUNGE-KUTTA-Verfahren geringer sein als für $m = 1$.

Beispiel:

Bestimme das BUTCHER-Schema

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^\intercal \end{array} \text{ mit } \sum_j b_j k_j = \sum_{j=1}^{p} \frac{\tau^{j-1}}{j!} f^{(j-1)}(t,z)$$

Dazu machen wir Gebrauch von der verallgemeinerten Mittelpunktregel ($s = 1$):

$$u_n = u_{n-1} + \tau b \underbrace{f\left(t_{n-1} + c\tau, u_{n-1} + \tau a k\right)}_{=k}$$

$$bk = bf(t + c\tau, z + \tau a k) = b\left[f(t,z) + D_1 f(t,z)\tau c + D_2 f(t,z)\tau a f(t,z) + O(\tau^2)\right] \stackrel{!}{=} f(t,z) + \frac{\tau}{2} f^{(1)}(t,z) + O(\tau^2)$$

Durch Koeffizientenvergleich ergibt sich $b = 1$ und $a = c = 1/2$.

$$u_n = u_{n-1} + \tau \underbrace{f\left(t_{n-1} + \frac{\tau}{2}, u_{n-1} + \frac{\tau}{2} k\right)}_{=k = 1/\tau(u_n - u_{n-1})} = u_{n-1} + \tau f\left(\frac{t_n + t_{n-1}}{2}, \frac{1}{2}(u_n + u_{n-1})\right)$$

Definition 3.3:

Zu Stützstellen $0 \le c_1 < c_2 < \ldots < c_s \le 1$ definieren wir ein „Kollokationsverfahren“ durch: Wähle ein Polynom $P \in \mathbb{P}_s$

$$P(t_{n-1}) = u_{n-1} \text{ mit } \frac{\mathrm{d}}{\mathrm{d}t} P(t_{n,i}) = f(t_{n,i}, P(t_{n,i}))$$

und setze $u_n = P(t_n)$ (falls die Interpolation lösbar ist).

Lemma:

Das Kollokationsverfahren ist ein RUNGE-KUTTA-Verfahren mit

$$b_j = \int_0^1 L_j(t)\,\mathrm{d}t \text{ mit } L_j(t) = \prod_{\substack{j=1 \\ j \ne i}}^{s} \frac{t - c_i}{c_j - c_i} \text{ und } a_{ij} = \int_0^{c_i} L_j(t)\,\mathrm{d}t$$

Beweis:

$$k_i = \frac{\mathrm{d}}{\mathrm{d}t} P(t_{n,i}) \Rightarrow u_n = u_{n-1} + \int_{t_{n-1}}^{t_n} + \int_{t_{n-1}}^{t_n} \frac{\mathrm{d}}{\mathrm{d}t} P(t)\,\mathrm{d}t = u_{n-1} + \tau \sum_{j=1}^{s} b_j k_j \text{ wobei } u_n = P(t_n),\ u_{n-1} = P(t_{n-1})$$

$$P(t_{n,i}) = P(t_{n-1}) + \int_{t_{n-1}}^{t_{n,i}} \frac{\mathrm{d}}{\mathrm{d}t} P(t)\,\mathrm{d}t = u_{n-1} + \int_0^{c_i} \dot{P}(t_{n-1} + \xi\tau)\tau\,\mathrm{d}\xi = u_{n-1} + \tau \sum_{j=1}^{s} b_j \int_0^{c_i} L_j(t)\,\mathrm{d}t$$

Dies gilt, da $\dot{P}(t_{n-1} + s\tau) = \sum_{j=1}^{s} L_j(s) b_j$ ist. □

3.2 Störungsrechnung

3.2.1 Lineare nicht-autonome Anfangswertaufgaben

Wir betrachten $\dot{u}(t) = Au(t)$ in $\mathbb{R}^m$ (mit $A \in C(\mathbb{R}, \mathbb{R}^{m,m})$). Dafür existiert eine Lösung über ein Fundamentalsystem $\dot{U}(t) = A(t)U(t)$ und $U(t_0) = I$ mit $U \in C^1(\mathbb{R}, \mathbb{R}^{m,m})$. Für die sogenannte „WRONSKI-Determinante“ $w(t) = \det(U(t))$ gilt folgende Differentialgleichung:

$\dot{w}(t) = \mathrm{Sp}(A(t))w(t)$ mit $w(t_0) = 1$

Da $w(t_0) > 0$, ist das Fundamentalsystem $U(t)$ regulär für alle $t \in \mathbb{R}$.

3.2.2 Lineare nicht-homogene Anfangswertaufgabe

Wir betrachten $\dot{u}(t) = A(t)u(t) + b(t)$ mit $u(t_0) = u_0$. Die Lösung kann bekanntlich mit der „Variation der Konstanten“ gefunden werden.

$$u(t) = U(t)u_0 + \int_{t_0}^{t} U(t)^{-1}U(s)b(s)\,\mathrm{d}s$$

3.2.3 Parameterabhängige Anfangswertaufgabe

Wir betrachten:

$\frac{\partial}{\partial t}u(t,\lambda) = f(t, u(t,\lambda), \lambda)$ mit $u(t_0, \lambda) = u_0$

Die Lösung hängt stetig differenzierbar von den Parametern λ ab und es gilt für $v_\lambda(t) = \partial/\partial\lambda u(t,\lambda)$:

$$\begin{aligned}\dot{v}_\lambda(t) &= \frac{\partial}{\partial t}\frac{\partial}{\partial\lambda}u(t,\lambda) = \frac{\partial}{\partial\lambda}\frac{\partial}{\partial t}u(t,\lambda) = \frac{\partial}{\partial\lambda}f(t,u(t,\lambda),\lambda) = D_2f(t,u(t,\lambda),\lambda)\frac{\partial}{\partial\lambda}u(t,\lambda) + D_3(t,u(t,\lambda),\lambda) = \\ &= A_\lambda(t)v_\lambda(t) + b_\lambda(t) \text{ mit } v_\lambda(t_0) = 0\end{aligned}$$

$A_\lambda(t) = D_2f(t,u(t,\lambda),\lambda)$ ist die JACOBI-Matrix und $b_\lambda = D_3f(t,u(t,\lambda),\lambda)$. Die Parameterabhängigkeit führt als zu einer Differentialgleichung wie wir sie zuvor behandelt haben. Also existiert ein $V(t,\lambda)$ mit der Eigenschaft

$$\frac{\partial}{\partial t}V(t,\lambda) = A(t)V(t,\lambda) \Rightarrow v_\lambda(t) = \int_{t_0}^{t} V(t,\lambda)^{-1}V(s,\lambda)b(s)\,\mathrm{d}s$$

Satz 3.6:

Seien f und eine Störung $\delta f \in C^1([t_0, t_0+T]\times G, \mathbb{R}^m)$ mit $G \subset \mathbb{R}^m$ offen, $u_0 \in G$ gegeben. Dann existiert $\widetilde{T} > 0$ und eine Matrixfunktion $M \in C(\Delta, \mathbb{R}^{m,m})$ mit dem Dreieck $\Delta = \{(t,s) : t_0 \in t \le t_0 + \widetilde{T}, t_0 \le s \le t\}$. Die Lösung der Anfangswertaufgabe $\dot{u}(t) = f(t,u(t))$ und der gestörten Anfangswertaufgabe $\dot{\widetilde{u}}(t) = f(t,\widetilde{u}(t)) + \delta f(t,\widetilde{u}(t))$ unterscheiden sich durch die Störung (Variation) $\delta u(t) = \widetilde{u}(t) - u(t)$. Für die Störung gilt:

$$\delta u(t) = \int_{t_0}^{t} M(t,s)\delta f(s,\widetilde{u}(s))\,\mathrm{d}s$$

Beweis:

Wir müssen ein passendes parameterabhängiges Problem definieren. Sei $u(t,\lambda)$ Lösung vom parameterabhängigen Problem

$\frac{\partial}{\partial t}u(t,\lambda) = f(t,u(t,\lambda)) + \lambda\delta f(t,\widetilde{u}(t))$

Hieraus ergibt sich:

$$\delta u(t) = \widetilde{u}(t) - u(t) = u(t,1) - u(t,0) = \int_0^1 \frac{\partial}{\partial\lambda}u(t,\lambda)\,\mathrm{d}\lambda = \int_0^1\int_{t_0}^{t} V(t,\lambda)^{-1}V(s,\lambda)\delta f(s,\widetilde{u}(s))\,\mathrm{d}s\,\mathrm{d}\lambda =$$

$$= \int_{t_0}^{t} M(t,s)\delta f(s,\widetilde{u}(s))\,\mathrm{d}s \text{ mit } M(t,s) = \int_0^1 V(t,\lambda)^{-1}V(s,\lambda)\,\mathrm{d}\lambda$$

□

Ergänzung zu 3.1:

Wir haben den allgemeinen Fall definiert wie folgt:

$$G(k) = \left(k_i - f\left(t + c_i\tau, z + \tau \sum_{j=1}^{s} a_{ij} k_j \right) \right) \in \mathbb{R}^{m-s}$$

Wenn die c_i, α_{ij}, usw. angegeben sind, so können wir dies schreiben als $k - T(k) = 0$. Die Lösung ist ein Fixpunkt von T. Wir wollen den BANACHschen Fixpunktsatz anwenden. Infolgedessen müssen wir berechnen:

$$T\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} f\left(z + a_{11}k_1 + a_{12}k_2\right) \\ f\left(z + a_{21}k_1 + a_{22}k_2\right) \end{pmatrix}, \left| T\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} - T\begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \right| \leq L \left| \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} k_1 - l_1 \\ k_2 - l_2 \end{pmatrix} \right|$$

$$|T(k) - T(l)| \leq L \left| \left(z + \tau \sum_j a_{ij} k_j - z - \tau \sum_j a_{ij} k_j \right)_i \right| = L\tau \left| \left(\sum_j a_{ij}(k_j - l_j) \right)_i \right| \leq L\tau |\mathcal{A}| |k - l|$$

Das ganze ist kontrahierend für $L\tau|\mathcal{A}| < 1$.

3.2.4 Skalierung

Wenn die Quadratur

$$\int_0^1 h(t) = \sum_j b_j h(c_j) + R(h)$$

exakt ist für Polynome vom Grad $p-1$, dann gilt für $h \in C^p([t_{n-1}, t_n])$:

$$\int_{t_{n-1}}^{t_n} h(t)\,\mathrm{d}t = \tau \sum_{j=1}^{s} h(t_{n,j}) b_j + R(h) \text{ mit } |R(h)| \leq C\tau_n^{p+1} \max_{t \in [t_1, \ldots, t_n]} \left| \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^p h(t) \right|$$

Satz 3.5:

Wenn die Quadratur (c_i, b_i) exakt für Polynome vom Grad $p-1$ ist, dann hat das zugehörige Kollokationsverfahren die Ordnung p. Für die GAUSS-Quadratur gilt $p = 2s$.

Beweis:

Wir schauen uns den lokalen Diskretisierungsfehler an:

$$g_n = \frac{1}{\tau_n}\left(u(t_n) - u(t_{n-1})\right) - \psi\left(t_{n-1}, \tau_n, u(t_{n-1})\right) = \frac{1}{\tau_n}\left(u(t_n) - P(t_n)\right)$$

P ist ein Polynom vom Grad s mit $P(t_{n-1}) = u(t_{n-1})$ und $\dot{P}(t_{n,i}) = f(t_{n,i}, P(t_{n,i}))$. Es gilt $\dot{u}(t) = f(t, u(t))$ und $\dot{P}(t) = f(t, P(t)) + (\dot{P}(t) - f(t, P(t)))$, wobei $\dot{P}(t) - f(t, P(t)) = \delta f$.

$$u(t_n) - P(t_n) = \int_{t_{n-1}}^{t_n} M(t_n, t)\delta f(t)\,\mathrm{d}t == \tau_n \sum_{j=1}^{s} b_j M(t_n, t_{n,j}) \underbrace{\delta f(t_{n,j})}_{\substack{=0 \text{ nach} \\ \text{Konstruktion}}} + R$$

Es ist $|u(t_n) - P(t_n)| \leq \widetilde{C}\tau_n^{p+1}$, wenn $\widetilde{C}$ unabhängig von τ ist. Dies müssen wir jetzt noch zeigen. Behauptung:

$$\max_{t_{n-1} \leq t \leq t_n} \left| \left(\frac{\mathrm{d}}{\mathrm{d}t} \right)^k (u - P)(t) \right| \leq C\tau_n^{s+1-k} \text{ für } \tau_n < \tau^* \text{ und } k = 0, \ldots, s$$

Wir verwenden nun, was wir bereits wissen:

$$\dot{u}(t_{n-1} + \theta\tau_n) = \sum_{j=1}^{s} \dot{u}(t_{n,i}) L_i(\theta) + \tau^s r(\theta)\omega(\theta) \text{ mit } r(\theta) = u[t_{n-1} + \theta\tau_n, t_{n,1}, \ldots, t_{n,s}],\ \omega(\theta) = \prod_i (\theta - c_i)$$

Des weiteren gilt:

$$\max_{0\le\theta\le1}\left|\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^k r(\theta)\right| \le \tau_n^k \frac{1}{k!}\max_{t_{n-1}\le t\le t_n}\left|\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{s+1+\lambda} u(t)\right|$$

$$u(t_{n-1}+\theta\tau_n) - P(t_{n-1}+\theta\tau_n) = \tau_n \int_0^\theta \left[\dot{u}(t_{n-1}+\eta\tau_n) - \dot{P}(t_{n-1}+\eta\tau_n)\right]\,\mathrm{d}\eta =$$

$$= \tau_n \int_0^\theta \sum_{j=1}^{s} \left[f(t_{n,j}, u(t_{n,j})) - f(t_{n,j}, P(t_{n,j}))\right] L_j(\eta)\,\mathrm{d}\eta + \tau^{s+1}\int_0^\theta r(\eta)u(\eta)\,\mathrm{d}\eta$$

$$\varepsilon(\tau_n) := \max|u(t)-P(t)| \le \tau_n L\varepsilon(\tau_n)\Lambda + \tau^{s+1}C \text{ mit } \Lambda = \max_{u_0}\int_0^\theta |L_j(\eta)|\,\mathrm{d}\eta$$

Der Fall $k > 0$ funktioniert analog. □

3.3 Lineare Stabilitätsanalyse

3.3.1 Beispiel: Wärmeleitungsgleichung

Wir betrachten parabolische Anfangsrandwertaufgaben. Bestimme u: $[0,\infty]\times[a,b] \mapsto \mathbb{R}$ mit

$$\frac{\partial}{\partial t}u(t,x) = K\frac{\partial^2}{\partial x^2}u(t,x) \text{ mit } K > 0$$

K ist der sogenannte Wärmeleitkoeffizient. Wir müssen Anfangswerte zum Zeitpunkt Null vorgeben, also $u(0,x) = g(x)$ und außerdem Randwerte zu beliebigen Zeiten, nämlich $u(t,a) = c_a$, $u(t,b) = c_b$, wobei $x \in [a,b]$. An diesem Modell lassen sich allerhand Dinge untersuchen. Wir wollen das Verhalten anhand einer vorgegebenen Ortsdiskretisierung betrachten. Wähle ein äquidistantes Gitter $\Delta = \{x_i = a + \mathrm{i}h : i = 1,\ldots,M\}$ mit $h = (b-a)/M$ und approximiere $\partial^2/\partial x^2 u(x,t)$ durch:

$$u(t,x\pm h) = u(t,x) \pm \frac{\partial}{\partial x}u(t,x)h + \frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x)h^2 \pm \frac{1}{6}\frac{\partial^3}{\partial x^3}u(t,x)h^3 + O(h^4)$$

Hieraus ergibt sich dann:

$$u(t,x-h) - 2u(t,x) + u(t,x+h) = \frac{\partial^2}{\partial x^2}u(t,x)h^2 + O(h^4)$$

Damit wird $u(t,x_i)$ durch die lineare Anfangswertaufgabe

$$\dot{u}_i = Kh^{-2}(u_{i-1} - 2u_i + u_{i+1}) \text{ mit } u_i(0) = g(x_i) \text{ und } u_0 = c_a,\ u_M = c_b \qquad (i = 1,\ldots,m = M-1)$$

approximiert. Damit haben wir die Differentialgleichung

$$\dot{u} = Au + c \text{ in } \mathbb{R}^m \text{ mit } A = -\frac{K}{h^2}A_0 \text{ wobei } A_0 = \begin{pmatrix} 2 & -1 & & \\ -1 & & & \\ & & & \end{pmatrix} \text{ und } c = \frac{K}{h^2}\begin{pmatrix} c_a \\ 0 \\ \vdots \\ 0 \\ c_b \end{pmatrix}$$

Sei $a = 0$ und $b = 1$.

i.) A_0 hat die Eigenwerte $\lambda_i = 4\sin\left(\frac{\mathrm{i}\pi h}{2}\right)^2$ für $i = 1,\ldots,n$. Wir erkennen:

$$\lim_{h\mapsto0}\frac{\lambda_i}{h^2} = \pi^2 \text{ und } \lim_{h\mapsto0}\lambda_m = \lim_{h\mapsto0} 4\sin^2\left(\frac{m\pi}{2(m+1)}\right) = 4\sin\left(\frac{\pi}{2}\right) = 4$$

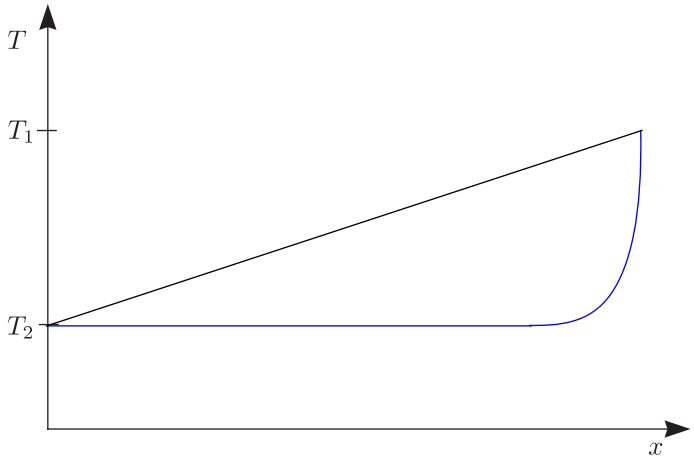
Wir können damit über das Spektrum der Matrix A folgende Aussage treffen:

$$\sigma(A) = \sigma\left(-\frac{K}{h^2}A_0\right) < 0$$

Es gilt $\lim_{t\mapsto\infty}\exp(tA) = 0$.

2.) Lineare Verbindung zwischen den beiden Anfangswerten a und b:

$$\widehat{u}_i = \frac{b - x_i}{b - a} c_a + \frac{x_i - a}{b - a} c_b$$



Es gilt $\lim_{t \mapsto \infty} u(t) = \widehat{u}$. Zum Beweis betrachten wir $w = u - \widehat{u}$. Hieraus folgt $\dot{w} = \dot{u} = Au + c = Au - A\hat{u} = Aw$ und damit:

$$\lim_{t \mapsto \infty} |w(t)| = \lim_{t \mapsto \infty} |\exp(tA)||w(0)| = 0$$

Für das explizite Cubeverfahren gilt:

$$w^n = (I + \tau A) w^{n-1} \ldots = (I + \tau A)^n w^0$$

Wählen wir speziell w^0 als Eigenvektor zu λ_m, so ergibt sich:

$$w^n = \left(1 - \frac{k}{h^2} \tau \lambda_m\right)^n w^0 \text{ also } \lim_{n \mapsto \infty} |w^n| = 0$$

$$\left|1 - \frac{K}{h^2} \tau \lambda_m\right| < 1 \overset{\lambda_m > 0}{\Leftrightarrow} \tau \frac{K}{h^2} \lambda_m < 2 \Leftrightarrow \tau < \frac{h^2}{K} \frac{2}{\lambda_m}$$

Für $\lambda_m \mapsto 4$ für $m \mapsto \infty$ ist das explizite EULERverfahren nur stabil für $\tau \leq h^2/2K$. Das Problem ist, dass explizite Verfahren sehr kleine Zeitschritte erzwingen, um numerisch stabile Lösungen zu erzielen!

Definition 3.7:

Die lineare Anfangswertaufgabe $\dot{u} = Au$ mit $u(0) = u_0$ heißt „stabil“, wenn für alle Anfangswerte u_0 gilt: $|u(t)| \leq u_0$. Sie heißt „asymptotisch stabil“, wenn für alle u_0 gilt: $\lim_{t \mapsto 0} |u(t)| = 0$.

Satz 3.8:

Für lineare Anfangswertaufgaben hat ein RUNGE-KUTTA-Verfahren die Form $u_n = R(\tau_n A) u_{n-1}$ mit $R(z) = P(z)/Q(z)$ (rationales Polynom) mit $P, Q \in \mathbb{P}_s$. Zusatz: Es gilt $R(z) \in \mathbb{P}_s$, wenn das RUNGE-KUTTA-Verfahren explizit ist. Außerdem gilt für ein RUNGE-KUTTA-Verfahren der Ordnung p:

$$R(z) = \sum_{j=0}^{r} \frac{1}{j!} z^j + O(z^{p+1})$$

(Man bezeichnet diese Entwicklung als „PALE-Approximation“ der Exponentialfunktion.)

Beweis:

Wir betrachten $\dot{u} = \lambda u$ mit $u(0) = 1$. Der erste Schritt ist gegeben durch:

$$u_1 = u_0 + \tau \sum_{i=1}^{s} b_i k_i \text{ mit } k_i = \lambda \left(u_0 + \tau \sum_{i=1}^{s} a_{ij} k_j\right)$$

Setze $z = \lambda\tau$ mit

$$d = \frac{1}{\lambda}\begin{pmatrix} k_1 \\ \vdots \\ k_s \end{pmatrix} \in \mathbb{R}^s \text{ und } e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^s$$

Damit ergibt sich der nächste Schritt:

$$u_1 = 1 + zb^\intercal d \text{ mit } d = e + z\mathcal{A}d \Rightarrow d = (I - z\mathcal{A})^{-1}e \qquad (*)$$

$$u_1 = 1 + zb^\intercal(I - z\mathcal{A})^{-1}e \stackrel{!}{=} R(z)u_0$$

Dies können wir durch Anwendung der CRAMERschen Regel zeigen. Diese ergibt für das Gleichungssystem $(*)$, also $(I - z\mathcal{A})d = e$:

$$d_i = \frac{P_i(z)}{Q(z)} \text{ mit } P_i \in \mathbb{P}_{s-1} \text{ und } Q(z) = \det(I - z\mathcal{A}) \in \mathbb{P}_s$$

Auf diese Weise ergibt sich:

$$R(z)u_0 = 1 + z\sum_i b_i \frac{P_i(z)}{Q(z)} = \frac{Q(z) + \sum_i b_i z P_i(z)}{Q(z)} = \frac{P(z)}{Q(z)}$$

Ist das RUNGE-KUTTA-Verfahren explizit, so resultiert

$$(I - z\mathcal{A})^{-1} = \sum_{j\geq 0}(z\mathcal{A})^j = \sum_{j=0}^{s-1}(z\mathcal{A})^j \in \mathbb{P}_{s-1}$$

da $\mathcal{A}$ nilpotent ist. Die letzte Aussage des Satzes zeigen wir durch folgende Abschätzung:

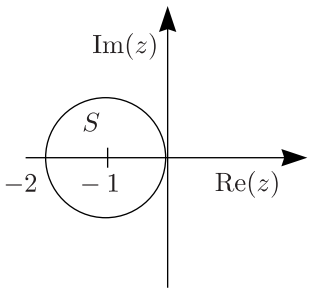
$$|u(\tau) - R(z)u_0| = |\exp(\tau\lambda) - R(z)| = O(\tau^{p+1}) \qquad \square$$

Definition 3.9:

Ein RUNGE-KUTTA-Verfahren heißt „A-stabil“, wenn die linke komplexe Halbebene $\mathbb{C}_- = \{z \in \mathbb{C} : \mathrm{Re}(z) \leq 0\}$ im Stabilitätsgebiet $S = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ enthalten ist. (Das Stabilitätsgebiet gilt an, welche τ zulässig sind.)

Beispiel:

a.) Explizites EULERverfahren: $u_n = (I + \tau A)u_{n-1} \Rightarrow R(z) = 1 + z$

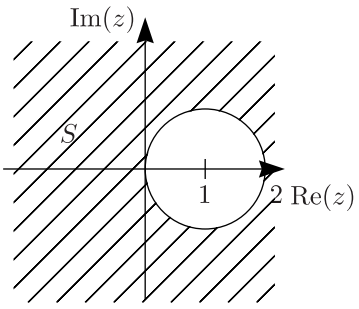


Das Stabilitätsgebiet ist nur ein winziger Bruchteil der linken Halbebene. Damit ist das Verfahren nicht A-stabil.

b.) Implizites EULERverfahren: $u_n = u_{n-1} + \tau A u_n \Rightarrow u_n = (I - \tau A)^{-1}u_{n-1}$

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

$$\left|\frac{1}{1-z}\right|^2 = \frac{1}{(1-z)(1-\overline{z})} = \frac{1}{1 - 2\mathrm{Re}(z) + |\overline{z}|^2} \leq 1 \text{ für alle } z \text{ mit } \mathrm{Re}(z) \leq 0$$



Damit ist $\mathbb{C}_-$ im Stabilitätsgebiet S enthalten.

c.) Allgemeines explizites RUNGE-KUTTA-Verfahren:

Hier ist $R(z)$ ein Polynom. Es gilt immer $\lim_{z\mapsto\infty}|R(z)| = \infty$. Grundsätzlich gilt also $C_- \not\subset S$; das Verfahren ist nicht A-stabil.

Bemerkung:

Sei $R(z) = P(z)/Q(z)$ eine rationale Approximation der Exponentialfunktion der Ordnung p. Dann gilt $p \leq \operatorname{grad} P + \operatorname{grad} Q$.

Beweis:

Wir nehmen an, dass P und Q existieren mit $\operatorname{grad} P \leq k$ und $\operatorname{grad} Q \leq j$, wobei $k + j < p$. Hieraus folgt:

$$\frac{P(z)}{Q(z)} - \exp(z) = O(z^{k+j+2}) \Rightarrow P(z) - Q(z)\exp(z) = O(z^{k+j+2}) \text{ und } \exp(-z)P(z) - Q(z) = O(z^{k+j+2})\,(z \mapsto 0)$$

Dies gilt, da wir davon ausgehen können, dass $P(z)$ und $Q(z)$ für $z \mapsto 0$ in der Umgebung von 1 liegt. Wir führen eine Induktion über k durch und zeigen, dass $P \equiv Q \equiv 0$ ist (was zu einem Widerspruch führt). Sei zunächst $k = 0$. Leiten wir die zweite der obigen Gleichungen $j + 1$ mail nach z ab:

$(-1)^{j+1}\exp(-z)P = Q(z)$ für $z \mapsto 0$

Für $z \mapsto 0$ folgt $P = 0$ und damit $Q \equiv 0$. Führen wir nun den Induktionsschritt $k - 1 \mapsto k$:

$P'(z) + (Q(z) + Q'(z))\exp(z) = O(z^{k+j+1})$

Damit ist $\operatorname{grad} P'(z) \leq k - 1$ und $\operatorname{grad}(Q(z) + Q'(z)) \leq j$, woraus nach Induktionsvoraussetzung $P' \equiv 0$ und damit $Q \equiv 0$ resultiert. □

Satz 3.10:

Für A-stabile RUNGE-KUTTA-Verfahren gilt: Wenn die lineare Anfangswertaufgabe stabil ist, dann ist auch die numerische Lösung stabil mit $|u_n| \leq |u_0|$ für alle Schrittweiten $\tau > 0$.

Beweis:

Betrachten wir den Fall, dass A diagonalisierbar ist:

1.) A und $R(\tau A)$ haben die gleichen Eigenvektoren. Weiterhin ist $\lambda \in \sigma(A)$ genau dann. wenn $R(\tau\lambda) \in \sigma(R(\tau A))$.

2.) $\{z \in \mathbb{C} : \operatorname{Re}(z) < 0\} \subset \operatorname{int}(\mathbb{C}_-) \subset \operatorname{int}(S)$

Hieraus folgt $|R(\tau\lambda)| < 1$ für $\lambda < 0$.

3.) Aus $|R(\tau\lambda)| = 1$ folgt $\operatorname{Re}(\lambda) = 0$ und damit, dass die algebraische Vielfachheit gleich der geometrischen Vielfachheit ist.

4.) $|R(\tau A)^n u_0| \leq \max\limits_{\mu\in\sigma(R(\tau A))} |\mu^n||u_0| \leq |u_0|$ □

Beispiel: Implizite Trapezregel

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

$k_1 = \lambda u_0$, $k_2 = \lambda\left(u_0 + \frac{\tau}{2}k_1 + \frac{\tau}{2}k_2\right) \Rightarrow k_2 = \frac{\lambda\left(1+\frac{\tau}{2}\lambda\right)}{1-\frac{\tau}{2}\lambda}u_0$

Damit folgt weiter:

$$u_1 = u_0 + \frac{\tau}{2}(k_1 + k_2) = u_0 + \frac{1}{2}\tau\lambda u_0 + \frac{1}{2}\tau\frac{\lambda\left(1+\frac{\tau}{2}\lambda\right)}{1-\frac{\tau}{2}\lambda}u_0$$

Setzen wir $z = \tau\lambda$, so gilt:

$$u_1 = R(z)u_0 \text{ mit } R(z) = 1 + \frac{1}{2}z + \frac{1}{2}z\cdot\frac{1+\frac{2}{z}}{1-\frac{2}{z}} = 1 + \frac{z}{1-\frac{1}{2}z} = \frac{1+\frac{2}{z}}{1-\frac{2}{z}} = \frac{2+z}{2-z}$$

$$|R(z)|^2 = R(z)R(\overline{z}) = \frac{2+z}{2-z}\cdot\frac{2+\overline{z}}{2-\overline{z}} = \frac{4+4\mathrm{Re}(z)+|z|^2}{4-4\mathrm{Re}(z)+|z|^2} \leq 1 \text{ für } \mathrm{Re}(z) \leq 0$$

Also ist die implizite Trapezregel A-stabil.

Beispiel: Implizite Mittelpunktsregel

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

$k_1 = \lambda\left(u_0 + \frac{\tau}{2}k_1\right) \Rightarrow k_1 = \frac{1}{1-\frac{\lambda\tau}{2}}u_0 \Rightarrow u_1 = u_0 + \tau k_1$

Damit erhalten wir $R(z)$:

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

Die rationale Funktion für die implizite Mittelpunktsregel stimmt also mit der rationalen Funktion der impliziten Trapezregel überein.

Bemerkung:

Wegen $R(z) = 1 + z + O(z^2)$ für konsistente Verfahren gilt immer $R'(0) = 1$. Also gilt $|R(z)| > 1$ für $z \in (0, \varepsilon)$ und $|R(z)| < 1$ für $z \in (-\varepsilon, 0)$. Damit ist $0 \in \partial S$. Wir definieren zu A eine „charakteristische Schrittweite“ $\tau_c = \sup\{\tau > 0 : |R(\widetilde{\tau}A)| \leq 1 \text{ für alle Schrittweiten } \widetilde{\tau} \in (0, \tau)\}$ (abhängig von A).

Beispiel:

a.) Wärmeleitungsgleichung:

A ist diagonalisierbar und $\sigma(A) \subset \mathbb{R}_-$. Damit existiert zu jedem Verfahren eine charakteristische Schrittweite $\tau_c > 0$.

b.) Wir betrachten:

$$\dot{u} = Au \text{ mit } u(0) = \begin{pmatrix}1\\0\end{pmatrix} \text{ und } A = \begin{pmatrix}0 & 1\\-1 & 0\end{pmatrix}$$

Da A eine Drehmatrix ist, wissen wir $\sigma(A) = \{\pm\mathrm{i}\}$. Die Anfangswertaufgabe ist stabil, aber nicht asymptotisch stabil. Für das explizite EULERverfahren ist die rationale Funktion $R(z) = 1 + z$. Dies ist in der komplexen Ebene eine Einheitskugel um -1. Damit gilt $\tau_c = 0$. Das explizite EULERverfahren ist für ein endliches Intervall $[0, T]$ konvergent mit $O(\tau)$. In $[0, \infty)$ ist u_n unbeschränkt für alle τ. Das explizite EULERverfahren ist für keine Schrittweite stabil!

Es funktioniert jedoch mit dem klassischen RUNGE-KUTTA-Verfahren.

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$$

Man berechnet $\tau_c = 2\sqrt{2} \approx 2,828$. u_n bleibt also stabil für $\tau < \tau_c$.

Definition 2.11:

Ein A-stabiles RUNGE-KUTTA-Verfahren heißt „L-stabil“, wenn $R(\infty) = \lim_{|z|\mapsto\infty} R(z) = 0$ ist. Folgerung: Für asymptotisch stabile Anfangswertaufgaben gilt dann $R(\tau A) \mapsto 0$ für $\tau \mapsto 0$. Das heißt. der asymptotische Zustand lässt sich in einem Zeitschritt berechnen.

Beispiel:

a.) Für das implizite EULERverfahren ist $R(z) = 1/(1-z)$. Damit gilt $R(\infty) = 0$.

b.) Für $R(z) = P(z)/Q(z)$ mit $\mathrm{Grad}(Q) > \mathrm{Grad}(P)$ gilt $R(\infty) = 0$.

c.) Implizite Mittelpunktsregel:

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

Satz 3.12:

Sei $\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^\intercal \end{array}$ ein A-stabiles RUNGE-KUTTA-Verfahren. Sei zusätzlich $\mathcal{A}$ invertierbar und $a_{kj} = b_j$. Dann ist es L-stabil.

Beweis:

Wir gehen aus von der rationalen Funktion:

$$R(z) = 1 + zb^\intercal(I - z\mathcal{A})^{-1}e = 1 + b^\intercal\left(\frac{1}{z}I - \mathcal{A}\right)^{-1}e \Rightarrow R(\infty) = \lim_{|z|\mapsto\infty} 1 + b^\intercal\left(\frac{1}{z}I - \mathcal{A}\right)^{-1}e = 1 - b^\intercal A^{-1}e$$

Setzen wir

$$b^\intercal = (e^s)^\intercal\mathcal{A} \text{ mit } e^s = \begin{pmatrix}0\\ \vdots \\ 1\end{pmatrix} \Leftrightarrow \mathcal{A}^\intercal e^s = b$$

oben ein, so ergibt sich:

$$R(\infty) = 1 - (e^s)^\intercal A^{-1}Ae = 1 - 1 = 0$$ □

3.3.2 Radau-Verfahren

Dies ist ein Kollokationsverfahren mit der Eigenschaft, dass $c_s = 1$ ist. $s = 3$ und $p = 5$:

$(4-\sqrt{6})/10$			
$(4+\sqrt{6})/10$			
1	$(16-\sqrt{6})/36$	$(16+\sqrt{6})/36$	$1/9$
	$(16-\sqrt{6})/36$	$(16+\sqrt{6})/36$	$1/9$

Dieses Verfahren ist L-stabil.

3.4 Beispiele und Anwendungen

A.) Künstlich konstruiertes Beispiel:

Wir betrachten folgende Differentialgleichung:

$$\dot{u} = 10\left(u - \frac{t^2}{1+t^2}\right) + \frac{2t}{(1+t^2)^2} \text{ mit } u(0) = 0$$

Deren Lösung ist gegeben durch $u(t) = t^2/(1+t^2)$.

$|u(t) - \widetilde{u}(t)| \leq |u(s) - \widetilde{u}(s)|\exp(L(t+s))$

Man hat keine Change, auf lange Zeiten zu simulieren. Die Lösung läuft aufgrund von Rundungsfehlern für große Zeiten davon. (Dies ist ein prinzipielles Problem. Man sagt auch, dass die Aufgabe schlecht konditioniert ist.)

B.) HAMILTON-Mechanik:

Wir betrachten ein System von Massepunkten $x_i \in \mathbb{R}^3$. Die Beschleunigung dieser Punkte ist von folgender Form:

$$\ddot{x}_j = \sum_{i=1}^{N} m_i \frac{x_j - x_i}{|x_j - x_i|^3}$$

Die Differentialgleichung ist nicht steif. Einfache Verfahren respektieren die Energieerhaltung nicht (wie das explizite EULERverfahren). Beispielsweise ist die explizite Mittelpunktsregel energieerhaltend. Geschlossene Bahnen existieren in der Nähe der numerisch berechneten Lösungen.

C.) Chemische Reaktion: Oregonator ($HBrO_2$, Br^-, Ce(IV))

$$\dot{u}_1 = 77,24\left[u_2 + u_1\left(1 - 8,375 \cdot 10^{-6} u_1 - u_2\right)\right]$$

$$\dot{u}_2 = \frac{1}{77,27}\left[u_3 - (1+u_1)u_2\right]$$

$$\dot{u}_3 = 0,161(u_1 - u_3)$$

Die Funktionen u_1, u_2 und u_3 beschreiben die Konzentrationen der verschiedenen Stoffe (Spezies). Typisch ist, dass die einzige Nichtlinearität quadratisch ist. Das Problem hier sind die verschiedenen Größenordnungen der Skalierungen. Die chemischen Reaktionen laufen in unterschiedlichen Zeitskalen ab (Millisekunden bis Minuten). Die Reaktion verläuft zyklisch. Hier verwendet man grundsätzlich explizite Verfahren (eingebettete RUNGE-KUTTA-Verfahren).

D.) Chemischer Reaktor:

$$\dot{u} = a + u^2 v - (b+1)u + \epsilon z''$$

$$\dot{v} = bu - u^2 v + \epsilon v''$$

In Abhängigkeit von ϵ sind sehr kleine Schrittweiten zu wählen. Hat das Verfahren zu große Schrittweiten, so ergeben sich künstliche Oszillationen. Dann muss man die Schrittweite verringern.

Ist das Problem ungekoppelt, so wird die Strömung mit der NAVIER-STOKES-Gleichung berechnet. Die Strömung transportiert die verschiedenen Spezies, die an den Reaktionen teilnehmen. Wird die Strömung selbst durch die chemische Reaktion beeinflusst, so ist das Problem gekoppelt und damit sehr aufwändig zu lösen.

E.) Künstlich konstruiertes Problem:

Reagieren wenige Spezies miteinander, so ist das RADAU-Verfahren geschickt:

$A \xrightarrow{0,04} B$ (langsam), $B + B \xrightarrow{3,10^7} C + B$ (sehr schnell), $B + C \xrightarrow{10^4} A + C$

$\dot{u}_1 = -0,04u_1 + 10^4 u_2 u_3$, $\dot{u}_2 = 0,04u_1 - 10^4 u_2 u_3 - 3 \cdot 10^7 u_2^2$, $\dot{u}_3 = 3 \cdot 10^7 u_2^2$

Bei 300 Spezies wird man sich jedoch überlegen, ob man das RADAU-Verfahren verwenden will! Steifheit des Problems ist durch die schnelle Reaktion eingebaut.

F.) Schwingungen eines elastischen Stabes:

Lösungen müssen nicht immer glatt sein! Große Schwingungen können von kleinen Schwingungen überlagert werden. Die Frage ist dann nur, ob uns die Lösung interessiert, welche die kleinen Schwingungen mit berücksichtigt (Gitarre, Bauwerke (Mikroschwingung)). Entweder handelt es sich um L-stabile Verfahren, die alles gut abbilden, oder Verfahren, die über numerische Diffusion verfügen und die nichtglatten Lösungen herausfiltern. (Dominante Moden werden verstärkt und im Resultat unerwünschte Oszillationen, die im System vorhanden sind, weggedämpft.)

G.) Elektrischer Schaltkreis:

$\dot{u}_1 = u_2$, $\dot{u}_2 = (1 - u_1^2)u_2 - u_1$

Simulation mit VAN DER POL-Gleichung

Gleichungen kann man heute genauer Lösungen, als die Anfangsdaten bekannt sind (Masse der Erde, Reaktionsdaten bei chemischen Reaktionen.)

3.5 B-Stabilität

Eine Anfangswertaufgabe $\dot{u} = f(t, u(t))$, $u(t_0) = u_0$ mit $t \in [t_0, t_0 + T]$ und $f \in C([t_0, t_0 + T] \times G, \mathbb{R}^m)$ heißt dissipativ, wenn f bezüglich einem geeigneten Skalarprodukt $(\bullet, \bullet)$ in $\mathbb{R}^m$ negativ monoton ist, das heißt, $(f(t, z) - f(t, y), z - y) \leq 0$ für $t \in [t_0, t_0 + T]$ und z, $y \in G$.

Bemerkung:

Wenn f eine einseitige Lipschitz-Bedingung $(f(t, z) - f(t, y), z - y) \leq L|z - y|^2$ erfüllt, dann gilt für $\dot{u} = f(t, u)$ und $\dot{v} = f(t, v)$: $|u(t) - v(t) \leq \exp(L(t - t_0))|u(t_0) - v(t_0)|$. Für $L = 0$ ist f monoton und die Anfangswertaufgabe ist dissipativ. Für $L < 0$ ist f streng monoton, so ist die Anfangswertaufgabe strikt dissipativ. (Zum Beweis benötigt man das Gronwall-Lemma. Er wird in den Übungen durchgeführt.)

Beispiele:

a.) $f(z) = Az$

Ist A normal und negativ semidefinit, so ist $-f$ monoton.

b.) Wir betrachten $f = -\operatorname{grad} F$, wobei F ein konvexes Potential ist. „Konvex“ bedeutet, dass $F((1 - \lambda)z + \lambda y) < (1 - \lambda)F(z) + \lambda F(y)$ für $\lambda \in [0, 1]$ und z, $y \in G$. Wenn $F \in C^2(G)$ ist, dann ist die Hessematrix D^2F positiv semidefinit und es gilt mit $-Df = D^2F$:

$$-(f(z) - f(y), z - y) = \left(\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\lambda} DF((1 - \lambda)y + \lambda z)\, \mathrm{d}\lambda, z - y \right) =$$

$$= \int_0^1 \left(D^2F((1 - \lambda)y + \lambda z)(z - y), z - y \right) \mathrm{d}\lambda \geq 0$$

Damit ist $-f$ monoton.

Anwendung: Newtonsche Mechanik

Wir betrachten $x^n \in \mathbb{R}^3$ für $n = 1, \ldots, N$. Das Gravitationspotential ist gegeben durch:

$$F(x) = G \sum_{n<m} \frac{M_n M_m}{|x^n - x^m|}$$

Definition 3.14:

Ein Einschritt-Verfahren heißt „B-stabil“, wenn für eine dissipative Anfangswertaufgabe gilt: $|u_n - v_n| \leq |u_{n-1} - v_{n-1}|$, wobei $u_n = u_{n-1} + \tau_n \psi(t_{n-1}, \tau_n, u_{n-1})$ und $v_n = v_{n-1} + \tau_n \psi(t_{n-1}, \tau_n, v_{n-1})$.

Lemma 3.15:

B-stabile Runge-Kutta-Verfahren sind A-stabil.

Beweis:

Wähle $z = \alpha + \mathrm{i}\beta \in \mathbb{C}_n$ (das heißt $\alpha = \operatorname{Re}(z) \leq 0$) und zeige $|R(z)| \leq 1$. Identifiziere $\mathbb{R}^2$ und $\mathbb{C}$ und definiere $A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$. Zur Anfangswertaufgabe $\dot{u} = Au = f(u)$ betrachte $u_1 = R(\tau A)u_0$ und $v_1 = R(\tau A)v_0$.

$(f(z) - f(y))^\intercal (z - y) = (z - y)^\intercal A^\intercal (z - y) = \alpha |z - y|^2 \leq 0$ da $\alpha \leq 0$

Also ist die Anfangswertaufgabe dissipativ. Ist ein Runge-Kutta-Verfahren B-stabil, so ergibt sich $|u_1 - v_1| = |R(\tau A)(u_0 - v_0)| \leq |u_0 - v_0|$ mit beliebigem $u_0 - v_0$. Aus $|R(\tau A)| \leq 1$ folgt dann $|R(\tau z)| \leq 1$. □

Definition 3.16:

Ein Runge-Kutta-Verfahren $\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^\intercal \end{array}$ heißt algebraisch stabil, wenn

a.) $M := \operatorname{diag}(b_i)\mathcal{A} + A^\intercal \operatorname{diag}(b_i) - bb^\intercal$ positiv semidefinit ist und

b.) die Gewichte $b_i \geq 0$ sind.

Satz 3.17:

Algebraisch stabile RUNGE-KUTTA-Verfahren sind B-stabil.

Beweis:

Ohne Einschränkung sei die Anfangswertaufgabe autonom.

$$u_{n,i} = u_{n-1} + \tau \sum_{j=1}^{s} a_{ij} k_j \text{ mit } k_i = f(u_{n,i}) \qquad u_n = u_{n-1} + \tau \sum_{j=1}^{s} b_j k_j$$

$$v_{n,i} = v_{n-1} + \tau \sum_{j=1}^{s} a_{ij} l_j \text{ mit } l_i = f(v_{n,i}) \qquad v_n = v_{n-1} + \tau \sum_{j=1}^{s} b_j l_j$$

Sei $U_0 = u_{n-1} - v_{n-1}$, $U_i = u_{n,i} - v_{n,i}$ und $K_{n,i} = \tau(k_i - l_i)$. Dass die Anfangswertaufgabe dissipativ ist, bedeutet $(f(u_{n,i}) - f(v_{n,i}), u_{n,i} - v_{n,i}) \leq 0$, also mit den obigen Bezeichnungen $(K_i, U_i) \leq 0$.

$$\begin{aligned}
|u_n - v_n|^2 &= \Big|U_0 + \sum_{i=1}^{s} K_i\Big|^2 = |U_0|^2 + 2\sum_{i=1}^{s}(K_i, U_0) + \sum_{\substack{i=1\\j=1}}^{s} k_i k_j (K_i, K_j) = \\
&= |U_0|^2 + 2\sum_{i=1}^{s} b_i \underbrace{(K_i, U_i)}_{\leq 0} + 2\sum_{i=1}^{s} b_i (U_0 - U_i, K_i) + \sum_{\substack{i=1\\j=1}}^{s} k_i k_j (K_i, K_j) \leq \\
&\underset{b.)}{\leq} |U_0|^2 - 2\sum_{i=1}^{s} b_i \sum_{j=1}^{s} a_{ij}(K_j, K_i) + \sum_{\substack{i=1\\j=1}}^{s} k_i k_j (K_i, K_j) = |U_0|^2 - \sum_{\substack{i=1\\j=1}}^{s} \underbrace{(b_i a_{ij} - b_j a_{ji} - b_i b_j)}_{M} (K_i, K_j) \leq \\
&\underset{a.)}{\leq} |U_0|^2 = |u_{n-1} - v_{n-1}|^2
\end{aligned}$$

Dies folgt, da die Matrix M positiv semidefinit ist. □

Bemerkung:

Da M positiv semidefinit ist, gilt:

$$\sum_{i,j} m_{ij}(K_i, K_j) = \sum_{l} \sum_{i,j} m_{ij} K_{il} K_{jl} = K_{il}^{\mathsf{T}} M K_{jl} \geq 0$$

Beispiel:

a.) ALEXANDER-Schema:
$$\begin{array}{c|cc} \alpha & \alpha & 0 \\ 1 & 1-\alpha & \alpha \\ \hline & 1-\alpha & \alpha \end{array} \text{ mit } \alpha = 1 - 1/\sqrt{2} \ (p = 2)$$

Man kann nachrechnen, dass das Verfahren algebraisch stabil ist. Daraus folgt dann B-, A- und L-Stabilität.

b.) RADAU-Verfahren ($s = 2$):
$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 1/4 & 1/4 \end{array} \ (p = 3)$$

Auch dieses Verfahren ist algebraisch stabil und damit auch B-, A-, L-stabil.

Satz 3.18:

Die Kollokationsverfahren zur GAUSS- und RADAU-Quadratur sind B-stabil.

Bemerkung:

Die implizite Trapezregel $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$, aber nicht B-stabil. Betrachte folgende dissipative Anfangswertaufgabe:

$$\dot{u} = \begin{cases} |u|^3 & \text{für} \quad u \leq 0 \\ -u^2 & \text{für} \quad u \geq 0 \end{cases}$$

$u(0) = -2$, $\tau = \frac{36}{7}$, $u_1 = \frac{5}{2}(v_0 = v_1 = 0)$

3.6 Reversibilität und Energieerhaltung

Ein dynamisches System ist **reversibel**, das heißt $\phi(t+\tau, -\tau, \phi(t, \tau, z)) = z$. Wir wollen dies nun speziell nicht über den diskreten Fluss definieren, sondern wie folgt.

Definition 3.19:

Ein RUNGE-KUTTA-Verfahren ist „reversibel“, wenn $R(z)R(-z) = 1$ ist für alle $z \in \mathbb{C}$ und $R(z) \neq \infty$.

Beispiel:

Betrachten wir als einfachstes GAUSS-Verfahren die explizite Mittelpunktsregel mit dem Schema $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$.

$$k_1 = f\left(t_{n-1} + \frac{\tau_n}{2}, u_{n-1} + \frac{\tau_n}{2}k_1\right) = f\left(t_{n-1} + \frac{\tau_n}{2}, \frac{1}{2}u_{n-1} + \frac{1}{2}u_n\right)$$

Aus $u_n = u_{n-1} + \tau k_1$ ergibt sich $\tau k_1 = u_n - u_{n-1}$. Dies haben wir oben eingesetzt. Damit ergibt sich:

$$u_n = u_{n-1} + \tau f\left(\frac{1}{2}(t_{n-1} + t_n), \frac{1}{2}(u_{n-1} + u_n)\right) \text{ und } u_{n-1} = u_n - \tau f\left(\frac{1}{2}(t_n + t_{n-1}), \frac{1}{2}(u_n + u_{n-1})\right)$$

Damit ist der Fluss, wenn wir vorwärts oder rückwärts laufen, derselbe. Mit

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

folgt $R(z)R(-z) = 1$, also ist das Verfahren reversibel.

Lemma 3.20:

Sei $S = \mathbb{C}_-$. Dann ist das RUNGE-KUTTA-Verfahren reversibel.

Beweis:

Wir nehmen an, dass ein $x \in \mathbb{R}$ existiert, so dass $|R(\mathrm{i}x)| < 1$ ist. Daraus ergibt sich, dass ein $z \in \mathbb{C} \setminus \mathbb{C}_-$ mit $|R(z)| < 1$ existiert. Dies stellt ein Widerspruch dar.

$$1 = |R(\mathrm{i}x)|^2 = R(\mathrm{i}x)\overline{R(\mathrm{i}x)} = R(\mathrm{i}x)R(\overline{\mathrm{i}x}) = R(\mathrm{i}x)R(-\mathrm{i}x)$$

Es ist $R(z)R(-z) = 1$ für $z \in \mathrm{i}\mathbb{R}$. Da R meromorph ist, folgt $R(z)R(-z) = 1$ für alle $z \in \mathbb{C}$. □

Bemerkung:

Ein RUNGE-KUTTA-Verfahren ist A-stabil (bzw. R hat keinen Pol in $\mathbb{C}_-$) und R ist rerversibel. Dann ergibt sich daraus $S = \mathbb{C}_-$.

Satz 3.21:

Für $A \in \mathbb{R}^{m \times m}$ ist äquivalent:

a.) Schiefsymmetrie: $A^\intercal = -A$

b.) $\exp(tA)$ ist orthogonal, also $\exp(tA)^\intercal \exp(tA) = I$.

c.) Für alle Anfangswerte u_0 und $\dot{u} = Au$ mit $u(0) = u_0$ gilt $|u(t)| = |u_0|$ für $t \in \mathbb{R}$.

Der Beweis wird in den Übungen durchgeführt.

Bemerkung:

Dann ist die Differentialgleichung dissipativ, also gilt:

$(Az - Ay)^\intercal(z-y) = (z-y)^\intercal A^\intercal(z-y) = (z-y)^\intercal A(z-y) = (z-y)^\intercal(-A)^\intercal(z-y)$

Hieraus ergibt sich $(Az - Ay)^\intercal(z-y) = 0$.

Satz 3.22:

Sei $A = -A^\intercal$ (schiefsymmetrisch) und sei R eine rationale Funktion zu einem RUNGE-KUTTA-Verfahren mit $S = \mathbb{C}_-$. Dann ist $R(\tau A)$ orthogonal und für $u_n = R(\tau A)u_{n-1}$ gilt $|u_n| = |u_{n-1}|$.

Beweis:

Die Matrix $\mathrm{i}A$ ist hermitesch, weil $(\mathrm{i}A)^H = \overline{\mathrm{i}A}^\intercal = -\mathrm{i}A^\intercal = \mathrm{i}A$ gilt. Damit ist das Spektrum $\sigma(\mathrm{i}A)$ reell und $\mathrm{i}A$ ist außerdem diagonalisierbar. $\sigma(A)$ ist dann rein imaginär und diagonalisierbar. R hat keine Polstellen auf $\mathrm{i}\mathbb{R}$ (wohldefiniert auf allen Eigenwerten), womit sich ergibt:

$R(\tau A)^\intercal = R(\tau A^\intercal) == R(-\tau A) \Rightarrow R(\tau A)R(-\tau A) = R(\tau A)R(\tau A)^\intercal = I$ □

Bemerkung:

Aus der Tatsache, dass R reversibel ist, folgt $M = \mathrm{diag}(b)\mathcal{A} + \mathcal{A}^\intercal\mathrm{diag}(b) - bb^\intercal = 0$.

Satz 3.23:

Das GAUSS-Verfahren ist reversibel.

Beweis:

Sei $P \in \mathbb{P}_s$.

$u_{n-1} = P(t_{n-1})$, $\dot{P}(t_{n-1} + c_i\tau_n) = f(P(t_{n-1} + c_i\tau_n))$, $u_n = P(t_n)$

Eine Eigenschaft der GAUSS-Quadratur ist $c_i = 1 - c_{s+1-i}$. Wir definieren nun ein $Q \in \mathbb{P}_s$: $Q(t_n - \theta\tau_n) = P(t_{n-1} + \theta\tau_n)$.

$u_n = Q(t_n)$, $\dot{Q}(t_n - c_i\tau_n) = f(Q(t_n - c_i\tau_n))$, $u_{n-1} = Q(t_{n-1})$

3.6.1 Übergang zu nichtlinearen Problemen

Definition 3.24:

Eine Funktion $\mathcal{E}$: $G \mapsto \mathbb{R}$ heißt „erstes Integral“ zur Differentialgleichung $\dot{u} = f(u)$, falls $\mathcal{E}(u(t)) \equiv$ const. ist für alle t.

Beispiel:

Wir betrachten $\mathcal{E}(z) = z^\intercal z$, $f(z) = Az$ mit $A^\intercal = -A$. Dann gilt $\mathcal{E}(u(t)) =$ const..

Lemma 3.25:

$\mathcal{E} \in C^1(G)$ ist genau dann erstes Integral einer autonomen Differentialgleichung $\dot{u} = f(u)$, wenn $D\mathcal{E}(z)f(z) = 0$ (JACOBI-Matrix von $\mathcal{E}$) ist.

Beweis:

1.) „$\Rightarrow$“: Aus $\mathcal{E}(u(t)) =$ const. ergibt sich durch Berücksichtigung der Kettenregel:

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(u(t)) = D\mathcal{E}(u(t))\dot{u}(t) = D\mathcal{E}(u(t))f(u(t))$$

2.) „$\Leftarrow$“:

$$\mathcal{E}(u(t)) = \mathcal{E}(u(t_0)) + \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{E}(u(s))\,\mathrm{d}s = \mathcal{E}(u(t_0))$$

Satz 3.36:

Die Differentialgleichung $\dot{u} = f(u)$ besitze ein quadratisches erstes Integral $\mathcal{E}(z) = z^\intercal Ez + d^\intercal z + e$. Dann ist $\mathcal{E}(u_n) = \mathcal{E}(u_{n-1})$ für das GAUSS-Verfahren.

Beweis:

Sei $P \in \mathbb{P}_s$ mit $P(t_{n-1}) = u_{n-1}$, $\dot{P}(t_{n,i}) = f(P(t_{n,i}))$, $u_n = P(t_n)$ mit $t_{n,i} = t_n + c_i\tau_n$. Außerdem definieren wir $Q(t) = \mathcal{E}(P(t)) \in \mathbb{P}_{2s}$.

$$\mathcal{E}(u_n) - \mathcal{E}(u_{n-1}) = \mathcal{E}(P(t_n)) - \mathcal{E}(P(t_{n-1})) = \int_{t_{n-1}}^{t_n} \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(P(t))\,\mathrm{d}t \overset{\dot{Q}\in\mathbb{P}_{2s-1}}{=} \tau\sum_{i=1}^{s} b_i\dot{Q}(t_{n,i}) =$$

$$= \tau\sum_{i=1}^{s} b_i D\mathcal{E}(P(t_{n,i}))\dot{P}(t_{n,u}) = \tau\sum_{i=1}^{s} b_i D\mathcal{E}(P(t_{n,i}))f(P(t_{n,i})) = 0$$

Dass dies verschwindet folgt mit $D\mathcal{E}(z)f(z) = 0$ aus Lemma 3.25. □

3.6.2 Anwendung: Hamilton-Systeme

$q \in \mathbb{R}^m$ seien die Koordinaten und $p \in \mathbb{R}^m$ die Impulse. Eine HAMILTONfunktion ist eine Funktion $H\colon \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}$ mit

$$\dot{q} = \frac{\partial}{\partial p}H(p,q) \text{ und } \dot{p} = -\frac{\partial}{\partial q}H(p,q)$$

$$\dot{u} = JDH(u) \text{ mit } J = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix} \text{ und } u = \begin{pmatrix} q \\ p \end{pmatrix} \in \mathbb{R}^{2m}$$

$\mathcal{E}(z) = H(z)$ ist erstes Integral, da

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(u(t)) = DH(u(t))^\intercal JDH(u(t)) = 0$$

Dies ergibt sich aus der Tatsache, dass die Matrix J schiefsymmetrisch ist.

Beispiel:

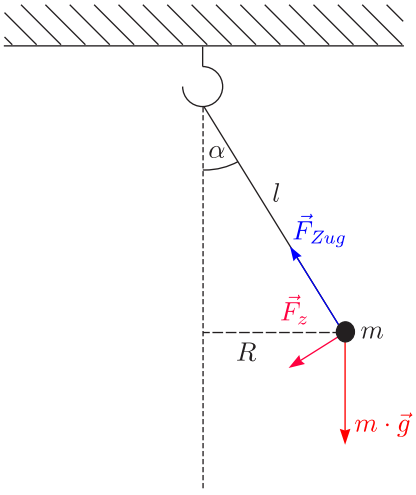
a.) Mathematisches Pendel: $H(p,q) = \frac{1}{2}p^2 - \cos(q)$

b.) Vielteilchensystem/Mehrkörpersystem: $H(p,q) = \frac{1}{2}\sum_i M_i|p_i|^2 - G\sum_{i<j}\frac{M_iM_j}{|q_i - q_j|}$

c.) $G(q) = 0$ (Beschreibung der festgehaltenen Positionen)

3.7 DAE-Systeme

3.7.1 Beispiel: Mathematisches Pendel



Wir suchen Funktionen $(x, y)^\intercal$ mit folgenden Eigenschaften:

a.) Energieerhaltung:

$$\frac{m}{2}(\dot{x}^2 + \dot{y}^2) + mgy = \text{const.}$$

b.) Nebenbedingung: $x^2 + y^2)l^2$

Man kann das ganze nun beschreiben als ein HAMILTON-System mit Nebenbedingungen. Dazu führen wir verallgemeinerte Koordinaten ein, nämlich $q_1 = x$, $q_2 = y$ und $p_1 = m\dot{x}$, $p_2 = m\dot{y}$.

$H(p, q) = \frac{1}{2m}|p|^2 - mgq_2$ mit $|q|^2 = l^2$

Das zugehörige LAGRANGE-Funktional lautet dann unter Berücksichtigung der Nebenbedingung mittels des LAGRANGE-Parameters λ:

$\mathcal{L}(q, \dot{q}, \lambda) = H(m\dot{q}, q) + \lambda(|q|^2 - l^2)$

Unter Verwendung der Optimierungstheorie erhalten wir die EULER-LAGRANGE-Gleichung:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{q}}\mathcal{L}(q, \dot{q}, \lambda)\right) = \frac{\partial}{\partial q}\mathcal{L}(q, \dot{q}, \lambda)$$

Wenden wir dies auf das mathematische Pendel an, so ergibt sich:

$$\boxed{m\ddot{q} = -\begin{pmatrix} 0 \\ mg \end{pmatrix} - 2\lambda q \text{ und } 0 = |q|^2 - l^2}$$

Wir haben jetzt also eine Differentialgleichung mit einer zusätzlichen algebraischen Gleichung als Nebenbedingung. Solche Gleichungen wollen wir nun untersuchen.

Definition 8.1:

a.) Sei $F \in C(I \times \mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^m)$ mit $I = [t_0, t_0 + T]$. Dann heißt $F(t, u(t), \dot{u}(t)) = 0$ mit $t \in I$ und der zusätzlichen Bedingung $u(t_0) = t_0$ implizit gestelltes Anfangswertaufgabe.

b.) $F(t, u, \dot{u})$ heißt differentiell-algebraisches System (DAE), wenn die JACOBI-Matrix für festes t und $u(t)$, also $D_3F(t, u(t), \dot{u}(t))$, nicht regulär ist.

c.) Der (differentielle) Index ist die kleinste Zahl k, so dass $\dot{u}$ durch das System $F(t, u, \dot{u}) = 0$, $(\mathrm{d}/\mathrm{d}j)^j F(t, u, \dot{u}) = 0$ für $j = 1, \ldots, k$ eindeutig in Abhängigkeit von u bestimmt ist.

Beispiel:

a.) Sei $F(t_0, u_0, v_0) = 0$ und $D_3F(t_0, u_0, v_0)$ regulär. Dann existiert $f\colon U \subset I \times \mathbb{R}^m \mapsto \mathbb{R}^m$ mit der Eigenschaft $F(t, z, f(t, z)) = 0$. Dann können wir die Differentialgleichung in der Form $\dot{u}(t) = f(t, u(t))$ mit $u(t_0) = u_0$ darstellen. (Alleine durch f ist $\dot{u}$ automatisch bestimmt, also ist der (differentielle) Index gleich 0.)

b.) Ein wichtiger Fall sind die linear-impliziten Anfangswertaufgaben. Man geht dabei davon aus, dass $F(t, u, v) = f(t, u) + M(t, u)v$ gilt. Dann sieht unsere Differentialgleichung folgendermaßen aus: $M(t, u)\dot{u} = f(t, u(t))$.

c.) Sei $M = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Dann ist $\dot{u} = f(t, u, v)$ und $0 = g(t, u, v)$. Da v nicht bestimmt ist, müssen wir $g(t, u, v)$ nach t ableiten:

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t, u, v) = D_1g(t, u, v) + D_2g(t, u, v)\dot{u} + D_3g(t, u, v)\dot{v} = 0$$

Falls $D_3g(t, u, v)$ regulär ist, können wir mit der inversen Matrix durchmultiplizieren und damit $\dot{v}$ darstellen. Es gilt also:

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} f(t, u, v) \\ -D_3g(t, u, v)^{-1}\left[D_1g(t, u, v) + D_2g(t, u, v)f(t, u, v)\right] \end{pmatrix}$$

Durch einfaches Ableiten haben wir es also geschafft, eine Anfangswertaufgabe der gewöhnlichen Form zu erhalten. Das Problem in der Praxis ist jedoch, dass man die Ableitungen benötigt. Die erhaltene Index-0-Gleichung ist viel komplizierter als die vorherige Index-1-Gleichung, von der wir ausgegangen sind.

d.) Betrachten wir $\dot{u} = f(t,u,v)$ und $0 = g(t,u)$. Dann reicht die erste Ableitung nicht aus, wir müssen zweimal ableiten:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 g(t,u) = D_1^2 g(t,u) + 2D_1 D_2 g(t,u) f(t,u,v) + D_2^2 g(t,u) f(t,u,v)^2 + \\ + D_2 g(t,u)\left[D_1 f(t,u,v) + D_2 f(t,u,v)\dot{u} + D_3 f(t,u,v)\dot{v}\right]$$

Dies ist nach $\dot{v}$ auflösbar, falls $D_2 g(t,u) D_3 f(t,u,v) := G$ regulär ist.

$$\begin{pmatrix}\dot{u}\\ \dot{v}\end{pmatrix} = \begin{pmatrix} f(t,u,v) \\ G(t,u,v)^{-1} h(t,u,v)\end{pmatrix}$$

Somit ist unser System vom Index 2.

Wir betrachten Index-1-Systeme $\dot{u} = f(t,u,v)$, $0 = g(t,u,v)$ mit den Anfangswerten $u(t_0) = u_0$, $v(t_0) = v_0$, welche kompatibel mit der Nebenbedingung sind, also $g(t_0,u_0,v_0) = 0$ ist, und für die $G \in C(I \times \mathbb{R}^m, \mathbb{R}^m)$ mit $g(t,z,G(t,z)) = 0$ existiert.

Satz 3.2:

Wenn f, g stetig differenzierbar und $D_3 g(t_0,u_0,v_0)$ regulär ist, dann existiert eine Simulationszeit $T > 0$, so dass die DAE in $[t_0, t_0+T]$ eine eindeutige Lösung besitzt.

Beweis:

Sei $h(t,z) = f(t,z,G(t,z))$ und $\dot{u} = h(t,u)$. h erfüllt eine L-Bedingung. Damit existiert ein $T > 0$ mit eindeutiger Lösung in $[t_0, t_0+T]$ und wir können (u,v) als $(u,v) = (u, G(t,u))$ definieren. Dies löst die DAE. □

Satz 3.3:

Sei $\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^\intercal \end{array}$ ein RUNGE-KUTTA-Schema der Ordnung p mit $a_{s_i} = b_i$ (steif genau) und sei

$$\begin{pmatrix}u_n\\ v_n\end{pmatrix} = \begin{pmatrix}u_{n-1}\\ v_{n-1}\end{pmatrix} + \tau \sum_{i=1}^{s} b_i \begin{pmatrix}k_i\\ l_i\end{pmatrix} \text{ mit } \begin{pmatrix}k_i\\ 0\end{pmatrix} = \begin{pmatrix} f(t_{n,i},u_{n,i},v_{n,i}) \\ g(t_{n,i},u_{n,i},v_{n,i})\end{pmatrix} \text{ wobei}$$

$$t_{n,i} = t_{n-1} + c_i\tau_n,\ u_{n,i} = u_{n-1} + \tau\sum_{j=1}^{s} a_{ij}k_j,\ v_{n,i} = v_{n-1} + \tau\sum_{j=1}^{s} a_{ij}l_j$$

Dann gilt:

$$\left|\begin{pmatrix}u_n\\ v_n\end{pmatrix} - \begin{pmatrix}u(t_n)\\ v(t_n)\end{pmatrix}\right| = O(\tau^p)$$

Beweis:

Es gilt offensichtlich $G(t_{n,i},u_{n,i}) = v_{n,i}$. Daraus folgt $k_i = h(t_{n,i},u_{n,i})$, also berechnet sich u_n aus u_{n-1} mit dem RUNGE-KUTTA-Verfahren zu $\dot{u} = h(t,u)$. Somit erfüllt u die Gleichung $|u_n - u(t_n)| = O(\tau^p)$. „Steif genau“ bedeutet nun $v_n = v_{n,s} = G(t_{n,s},u_{n,s}) = G(t_n,u_n)$, woraus sich ergibt:

$$|v_n - v(t_n)| = |G(t_n,u_n) - G(t_n,u(t_n))| \le L|u_n - u(t_n)| = O(\tau^p)$$ □

Satz 3.4:

Wenn zusätzlich zu Satz 3.3 $\mathcal{A}$ invertierbar ist, dann konvergiert die RUNGE-KUTTA-Lösung zu einem regularisierten Problem, nämlich $\dot{u}^\varepsilon = f(t,u^\varepsilon,v^\varepsilon)$, $\varepsilon\dot{v}^\varepsilon = g(t,u^\varepsilon,v^\varepsilon)$ mit $u^\varepsilon(t_0) = u_0$, $v^\varepsilon(t_0) = v_0$ und außerdem $g(t_0,u_0,v_0) = 0$ für $\varepsilon \mapsto 0$ gegen die Lösung von 3.3.

Beweis:

$$\varepsilon v_{n,i}^{\varepsilon} = \varepsilon v_{n-1}^{\varepsilon} + \tau \sum_{j=1}^{s} a_{ij} g(t_{n,j}, u_{n,j}, v_{n,j})$$

Hieraus ergibt sich

$$\tau g(t_{n,j}, u_{n,j}^{\varepsilon}, v_{n,j}^{\varepsilon}) = \varepsilon \sum_{i=1}^{s} \overline{a}_{ij} (v_{n,i}^{\varepsilon} - v_{n-1}^{\varepsilon}) \text{ mit } \mathcal{A}^{-1} = (\overline{a}_{ij})$$

und weiter $g(t_{n,j}, u_{n,j}^{\varepsilon}, v_{n,j}^{\varepsilon}) \mapsto 0$ für $\varepsilon \mapsto 0$. Aus

$$\varepsilon v_{n}^{\varepsilon} = \varepsilon v_{n,s}^{\varepsilon} = \varepsilon v_{n-1} + \tau \sum_{i=1}^{s} b_i g(t_{n,i}, u_{n,i}^{\varepsilon}, v_{n,i}^{\varepsilon})$$

folgt:

$$v_n^{\varepsilon} = \left(1 - \sum_{j=1}^{s} b_j \sum_{i=1}^{s} \overline{a}_{ij}\right) v_{n-1}^{\varepsilon} + \sum_{i,j=1}^{s} b_i \overline{a}_{ij} v_{n,j}^{\varepsilon} = \underbrace{R(\infty)}_{=0} v_{n-1}^{\varepsilon} + v_{n,s}^{\varepsilon} = v_{n,s}^{\varepsilon} \qquad \square$$

Beispiel:

a.) Allgemeine mechanische Systeme mit Nebenbedingungen

$$\dot{q} = v \text{ mit } v = \frac{\partial}{\partial p} H(p,q) = M(q)^{-1} p \text{ da } H(p,q) = \frac{1}{2} p M(q)^{-1} p + U(q)$$

Hierbei ist $M(q)$ positiv definit. $0 = g(q)$ sei die Nebenbedingung.

$$M(q)\dot{v} = f(q,u) - Dg(q)\lambda \text{ mit } f(q,v) = -\frac{\partial}{\partial q} H(p,q)$$

Diese Gleichung ist vom Index 3, wenn $Dg(q)M(q)Dg(q)^{\intercal}$ invertierbar ist. Aus $g(q) = 0$ folgt

$$\frac{\mathrm{d}}{\mathrm{d}t} g(q) = Dg(q)\dot{q} \Rightarrow \begin{pmatrix} \dot{q} \\ \dot{v} \\ 0 \end{pmatrix} = \begin{pmatrix} v \\ M(q)^{-1}[f(q,u) - Dg(q)^{\intercal}\lambda] \\ Dg(q)v \end{pmatrix}$$

Dieses System ist jetzt nur noch vom Index 2. Aus $0 = Dg(q)v$ folgt:

$$0 = v^{\intercal} D^2 g(q) v + Dg(q)\dot{v} \Rightarrow \begin{pmatrix} M(q) & Dg(q)^{\intercal} \\ Dg(q) & 0 \end{pmatrix} \begin{pmatrix} \dot{v} \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q,v) \\ -v^{\intercal} Dg^2(q) v \end{pmatrix}$$

Das System ist jetzt nur noch vom Index 1.

b.) VAN-DER-POL-Gleichung (einfaches Modell für einen elektrischen Schaltkreis)
Wir betrachten:

$$\dot{x} = y \text{ und } \dot{y} = \mu(1 - x^2)y - x$$

$$f(x,y) = \begin{pmatrix} y \\ \mu(1-x^2)y - x \end{pmatrix} \Rightarrow Df(x,y) = \begin{pmatrix} 0 & 1 \\ -\mu \cdot 2xy - 1 & \mu(1-x^2) \end{pmatrix}$$

Bemerkung:

1.) Für große x und μ hat Df Eigenwerte mit großem negativen Realteil. Damit ist die Anfangswertaufgabe steif.

2.) Die Anfangswertaufgabe hat eine periodische Lösung; sie wird mit μ größer.

Wir machen eine Analyse als singulär gestörtes System. Wir führen eine Reskalierung der Form $\widetilde{t} = 1/\mu t$ durch und führen entsprechend reskalierte Größen $\widetilde{x}(\widetilde{t}) = x(t)$ und $\widetilde{y}(\widetilde{t}) = \mu y(t)$. Hieraus folgt:

$$\frac{\mathrm{d}}{\mathrm{d}\widetilde{t}}\widetilde{x}(\widetilde{t}) = \mu\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \mu y(t) = \widetilde{y}(\widetilde{t})$$

$$\frac{\mathrm{d}}{\mathrm{d}\widetilde{t}}\widetilde{y} = \mu^2\frac{\mathrm{d}}{\mathrm{d}t}y(t) = \mu^2[\mu(1-x(t)^2)y(t) - x(t)] = \mu^2\left[(1-\widetilde{x}(\widetilde{t})^2)\widetilde{y} - \widetilde{x}\right]$$

Setzen wir $\varepsilon = 1/\mu^2$ und $u = \widetilde{x}$, so folgt:

$$\ddot{u} = \frac{1}{\varepsilon}\left((1-u^2)\dot{u} - u\right) \Rightarrow \varepsilon\ddot{u} + (u^2-1)\dot{u} + u = 0$$

An dieser Stelle müssen wir noch einen kleinen Trick machen:

$$v = -\int_0^t u(s)\,\mathrm{d}s + C \Rightarrow \dot{v} = -u \Rightarrow \dot{v} = \varepsilon\ddot{u} + (u^2-1)\dot{u}$$

Durch Integration folgt weiter:

$$v = \varepsilon\dot{u} + \left(\frac{u^3}{3} - u\right) \text{ (Wahl von c)}$$

Unser System sieht damit folgendermaßen aus:

$$\boxed{\dot{v} = -u \text{ und } \varepsilon\dot{u} = v - \left(\frac{u^3}{3} - u\right)}$$

Was sagt uns diese neue Struktur? Machen wir den Grenzübergang $\varepsilon \mapsto 0$, so folgt:

$$\dot{v} = -u \text{ und } 0 = g(v,u) := v - \left(\frac{u^3}{3} - u\right)$$

Es gilt $D_2 g(v,u) = u^2 - 1 \neq 0$ für $u \neq \pm 1$. Durch Ableiten folgt weiter:

$$0 = \dot{v} - (u^2-1)\dot{u} = -u - (u^2-1)\dot{u} \Rightarrow \dot{u} = \frac{u}{1-u^2}$$

Die kritischen Stellen liegen bei $u = \pm 1$, $v = \mp 2/3$. Das DAE-System generiert einen Fluss auf dieser eindimensionalen Mannigfaltigkeit im zweidimensionalen Raum. Dies ist ein für Schaltkreise typisches periodisches Verhalten.

3.8 Lösungsverfahren für lineare implizite DAEs

$$M(u)\dot{u} = f(t,u) \text{ mit } M(u) = \begin{cases} I & \text{ODE} \\ \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} & \text{steif} \\ \text{singulär} & \text{DAE} \end{cases}$$

3.8.1 Diagonal-implizites Runge-Kutta-Verfahren

$$\begin{array}{c|cc} c_1 & a_{11} & \\ c_2 & a_{21} & a_{22} \\ & & \ddots \\ & & & a_{ss} \end{array}$$

Für $i = 1, \ldots, s$ löse $G(k_i) = 0$ mit

$$G_i(k_i) = \widetilde{G}_i\left(u_{n-1} + \tau_n \sum_{j=1}^{i} a_{ij}k_j\right) \text{ mit } \widetilde{G}(z) = M(z) - f(t_{n,i}, z)$$

Bemerkung:

Falls $M(z) = M$ eine positiv semidefinite Matrix und $-f$ [strikt] monoton ist, so folgt $DG(z) = M - \tau_n a_{ii} Df(z)$ und für $a_{ii} > 0$ ist $DG(z)$ positiv [semi]definit.

3.8.2 Mehrschrittverfahren

Wir lösen

$$M \sum_{i=0}^{k} \alpha_{k-i} u_{n-i} = \tau \sum_{i=1}^{k} \beta_{k-i} f(u_{n-i})$$

Analog zu oben gilt für unendlich-stabile BDF-Verfahren: $|u_n - u(t_n)| = O(\tau^p)$.

3.8.3 Rosenbrock-Verfahren (linear implizite Verfahren)

Beobachtung:

$$k_1 = f\left(u_{n-1} + \tau \sum_{j=1}^{i} a_{ij} k_j\right) \simeq f\left(u_{n-1} + \tau \sum_{j=1}^{i-1} a_{ij} k_j\right) + \tau a_{ii} Df(u_{n-1}) k_i$$

Zu b, c, $d \in \mathbb{R}^s$ und $\mathcal{A}$, $\mathcal{E} \in \mathbb{R}^{s,s}$:

$$u_n = u_{n-1} + \tau \sum_i b_i k_i$$

$$Mk_i = f\left(t_{n-1} + c_i \tau_n, u_{n-1} + \tau \sum_{j=1}^{i-1} a_{ij} k_j\right) + d_i \tau D_1 f(t_{n-1}, u_{n-1}) + \tau D_2 f(t_{n-1}, u_{n-1}) \sum_{j=1}^{i} e_{ij} k_j$$

Bemerkung:

A-Stabilität überträgt sich, B-Stabilität gilt jedoch nicht!

3.9 Randwertaufgaben

3.9.1 Lineare Randwertaufgaben

Sehr viele Überlegungen zu Randwertaufgaben haben sich entwickelt an einem ganz speziellen System 2.Ordnung.

Definition 4.1:

In einem Intervall $I = [a, b]$ zu $p \in C^1(I)$, q, r, $f \in C(I)$ und α_i, β_i und $\gamma_i \in \mathbb{R}$ ist die „Sturm-Liouville-Randwertaufgabe“ durch

$$-\frac{\mathrm{d}}{\mathrm{d}t}(p(t)\dot{u}(t)) + q(t)\dot{u}(t) + r(t)u(t) = f(t) \text{ für } t \in (a, b)$$

$\alpha_0 u(a) + \alpha_1 \dot{u}(a) + \beta_0 u(b) + \beta_1 \dot{u}(b) = \gamma_0$ und $\alpha_2 u(a) + \alpha_3 \dot{u}(a) + \beta_2 u(b) + \beta_3 \dot{u}(b) = \gamma_1$

bestimmt.

Beispiele:

Durch Wahl der Koeffizienten in den Randbedingungen können wir verschiedene Systeme beschreiben. Es gibt viele Funktionen in der Physik, die definiert sind durch Lösung der Sturm-Liouville-Gleichung.

a.) Freie Schwingung: $\ddot{u} + \omega^2 u = 0$ (mit $\omega > 0$)

Die Rückstellkraft (und damit die Beschleunigung) ist proportional zur Auslenkung. Die Lösung der Gleichung lautet:

$u(t) = c_0 \sin(\omega t) + c_1 \cos(\omega t)$

Betrachte für diese Schwingung das Intervall $[a, b] = [0, \pi/\omega]$. Dazu können wir nun verschiedene Fälle untersuchen:

1.) Periodische Randbedingungen: $u(a) = u(b)$, $\dot{u}(a) = \dot{u}(b)$ (durch entsprechende Wahl der Koeffizienten α_i, β_i und γ_i)
Dann ist $\dot{u} \equiv 0$ die eindeutige Lösung.

2.) $u(a) = 0$, $u(b) = 0$: Hier gibt es unendlich viele Lösungen.

3.) $u(a) = 0$, $u(b) = 1$: Es gibt keine Lösung.

Anfangswertaufgaben besitzen immer eine eindeutige Lösung, Randwertaufgaben jedoch nicht!

3.9.2 Schreibweise als System

Sei $v = \dot{u}$ und $p \neq 0$.

$$-p\dot{v} - \dot{p}v + qv + ru = f \Leftrightarrow \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \frac{v}{p} & \frac{q \cdot \dot{p}}{p} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{f}{p} \end{pmatrix}$$

Die Randwertgleichungen können dann in folgender Form geschrieben werden:

$$\begin{pmatrix} \alpha_0 & \alpha_1 \\ \alpha_2 & \alpha_3 \end{pmatrix} \begin{pmatrix} u(a) \\ v(a) \end{pmatrix} + \begin{pmatrix} \beta_0 & \beta_1 \\ \beta_2 & \beta_3 \end{pmatrix} \begin{pmatrix} u(b) \\ v(b) \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix}$$

Definition 4.2:

Zu $I = [a, b]$, $A \in C(I, \mathbb{R}^{m \times m})$, $b \in C(I, R^m)$ und B_a, $B_b \in \mathbb{R}^{m \times m}$, $g \in \mathbb{R}^m$ ist die allgemeine inhomogene lineare Randwertaufgabe durch

$\dot{u}(t) = A(t)u(t) + b(t)$ mit $t \in I$

$B_a u(a) + B_b u(b) = g$

bestimmt.

Beispiel:

Sei $B_a = I$ und $B_b = 0$. Dann ist dies eine gewöhnliche Anfangswertaufgabe.

Satz 4.3:

Sei $u_0 \in C^1(I)$ Lösung der inhomogenen Anfangswertaufgabe $u_0(t) = A(t)u_0(t) + b(t)$ mit $u_0(a) = 0$. Sei außerdem $u_i \in C^1(I)$ Lösung der homogenen Anfangswertaufgabe $u_i(t) = A(t)u_i(t)$mit $u_i(a) = e^i$. (e^i ist Einheitsvektor im $\mathbb{R}^m$, $i = 1, \ldots, m$). Dann ist die Matrix $U(t) = (u_1(t), \ldots, u_n(t))$ ein Fundamentalsystem und (4.2) hat die Form $u(t) = u_0(t) + \sum_{i=1}^m y_i u_i(t)$ mit $y \in \mathbb{R}^m$ als Lösung von $(B_a + B_b U(b))y = g - B_b u_0(b)$. Also gilt (Fredholmsche Alternative):

a.) $Q := (B_a + B_b U(b))$ ist regulär. Dann ist (4.2) immer eindeutig lösbar.

b.) Q ist singular, also existieren entweder mehrfache oder keine Lösungen (abhängig von $g - B_b u_0(b)$)

Der Fall B ist numerisch nicht einfach handzuhaben.

Beweis:

Wir setzen die angegebene Lösung ein und erhalten:

$$g = B_a \left(u_0(a) + \sum_{i=1}^m y_i u_i(a) \right) + B_b \left(u_0(b) + \sum_{i=1}^m y_i u_i(b) \right) = B_a y + B_b \left(u_0(b) + U(b)y \right)$$
$$= (B_a + B_b U(b))y + B_b u_0(b)$$

Damit ist alles gezeigt. □

Satz 4.4:

Sei u_0^τ diskrete Lösung der Anfangswertaufgabe $\dot{u}_0 = Au_0 + b$, $u(a) = 0$ und u_i^τ diskrete Lösung der Anfangswertaufgabe $\dot{u}_i = Au_i$, $u(a) = e^i$ (mit $i = 1, \ldots, n$). Sei $y^\tau \in \mathbb{R}^m$ Lösung von $Q^\tau y^\tau = g - B_b u_0^\tau(b)$ mit $Q^\tau = B_a + B_b U^\tau(b)$ mit $U^\tau = (u_1^\tau, \ldots, u_m^\tau)$. Setze $u^\tau = u_0^\tau + \sum_{i=1}^m y_i^\tau u_i^\tau$. Für u_i^τ gelte $|u_i(t_n) - u_i^\tau(t_n)| = O(\tau^p)$ für $i = 0, \ldots, m$ und $t_m \leq \Delta$, $b \in \Delta$. Dann gilt: Wenn Q regulär ist, dann existiert ein τ_0, so dass Q^τ für $\tau < \tau_0$ regulär und für die Lösung der Randwertaufgabe gilt $|u(t_n) - u^\tau(t_n)| = O(\tau^p)$.

Beweis:

Als erstes schauen wir uns folgenden Fehler an:

$|Q - Q^\tau| \leq |B_b||u(b) - u^\tau(b)| \leq |B_b| \max |u_i(b) - u_i^\tau(b)| = O(\tau^p)$

Zu einer Konstanten $C < 1$ existiert ein τ_0 mit $|Q-Q^\tau| \leq C/|Q^{-1}|$ für alle $\tau < \tau_0$. Dann folgt $|Q^{-1}(Q-Q^\tau)| < 1$. Unter dieser Voraussetzung ist $Q^{-1}Q^\tau = I - Q^{-1}(Q - Q^\tau)$ regulär und die NEUMANNsche Reihe besagt:

$$(Q^{-1}Q^\tau)^{-1} = \sum_{k\geq 0} [Q^{-1}(Q - Q^\tau)]^k$$

Hieraus ergibt sich weiter:

$$|(Q^\tau)^{-1}| \leq \frac{|Q^{-1}|}{1 - |Q^{-1}||Q - Q^\tau|} \leq \frac{|Q^{-1}|}{1 - C} \Rightarrow |(Q^\tau)^{-1} - Q^{-1}| = |(Q^\tau)^{-1}(Q^\tau - Q)Q^{-1}| \leq \frac{|Q^{-1}|^2}{1 - C}|Q - Q^\tau| = O(\tau^p)$$

$$\begin{aligned} y - y^\tau &= Q^{-1}(g - B_b u_0(b)) - (Q^\tau)^{-1}(g - B_b u_0^\tau(b)) = (Q^{-1} - (Q^\tau)^{-1})(g - B_b u_0^\tau(b)) - Q^{-1}B_b(u_0(b) - u_0^\tau(b)) = \\ &= O(\tau^p) + O(\tau^p) = O(\tau^p) \end{aligned}$$

Also gilt:

$$|u(t_n) - u^\tau(t_n)| \leq |u_0(t_n) + u_0^\tau(t_n)| + \sum_i \underbrace{|y_i u_i(t_n) - y_i^\tau u_i^\tau(t_n)|}_{=(y_i - y_i^\tau)u_i(t_n) + y_i^\tau(u_i(t_m n) - u_i^\tau(t_n))} = O(\tau^p) \qquad \square$$

Problem:

Es gilt $|u(t_n) - u^\tau(t_n)| \leq C\tau^p \exp(L(t_n - a))$. Für $(t_n - a)L \gg 1$ unbrauchbar!

Beispiel:

Wir betrachten die STURM-LIOUVILLE-Gleichung $\ddot{u} - \dot{u} - 110u = 0$ für $I = [0, 10]$. Die Lösung $u(t)$ ist gegeben durch:

$u(t) = c_0 \exp(-10t) + c_1 \exp(11t)$

$u_0 \equiv 0$, $u_1(t) = \frac{11}{21}\exp(-10t) + \frac{10}{21}\exp(11t)$ mit $u_1(0) = 1$, $\dot{u}_1(0) = 0$

$u_2(t) = -\frac{1}{21}\exp(-10t) + \frac{1}{21}\exp(11t)$ mit $u_2(0) = 0$, $\dot{u}_2(0) = 1$

Randwerte: $u(0) = u(10) = 0$

$$u = u_1 + c_2 u_2 \text{ mit } c_2 = -10 + \frac{21 + \exp(-100)}{\exp(110) - \exp(-100)} \approx -10 + 3,5 \cdot 10^{-47}$$

Wenn wir die Steigung um $\varepsilon = 3,5 \cdot 10^{-47}$ ändern, dann erhalten wir eine völlig andere Lösung. Damit sehen wir, dass das Problem extrem konditionsabhängig ist.

3.10 Übergang zu nichtlinearen Problemen

3.10.1 Beispiel: Nichtlineare Zweipunktrandwertaufgabe 2.Ordnung

Wir betrachten $\ddot{u} = f(t, u, \dot{u})$ mit $u(a), u_a$, $u(b) = u_b$ und $t \in [a, b] = I$, $f \in C(I \times \mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^m)$. Schießverfahren: Zu $v \in \mathbb{R}$ berechne die Lösung der Anfangswertaufgabe von $\ddot{u}^v = f(t, u^v, \dot{u}^v)$ mit $u(a) = u_a$ und $\dot{u}(a) = v$, $t \in I$. Setze $F(v) = u^v(b) - u_b$. Aus $F(v) = 0$ folgt, dass $u = u^v$ die Randwertaufgabe löst. Wir nehmen an, dass F differenzierbar ist und F' sich durch den Differenzenquotienten $F'(v) = 1/\delta[F(v + \delta) - F(v)]$ (mit geeignetem δ) approximieren lässt. Daraus machen wir einen Algorithmus:

1.) Gegeben sei ein Startvektor v, ein Parameter δ und eine Genauigkeit $\varepsilon > 0$. Berechne u^v.

2.) Falls $|u^v(b) - u_b| < \varepsilon$ ist, beende die Rechnung.

3.) Berechne $u^{v+\varepsilon}$ und $J = 1/\delta(u^{v+\delta}(b) - u^b(b))$.

4.) Berechne daraus $v := v - 1/J(u^v(b) - u_b)$ $(\approx v - F(v)/F'(v))$.

Definition 4.5:

Sei $f \in C(I \times \mathbb{R}^m, \mathbb{R}^m)$ mit $r \in C(\mathbb{R}^m \times \mathbb{R}^m)$ gegeben. Dann lautet die allgemeine Randwertaufgabe: Bestimme $u \in C^1(I, \mathbb{R}^m)$ mit $\dot{u} = f(t, u)$ mit $t \in [a, b] = I$ und den Randwerten, die durch $0 = r(u(a), u(b)$ gegeben sind.

Beispiel:

Wir betrachten $f(t, u) = Au + b$ mit $r(z, y) = B_a z + B_b y - g$.
Schießverfahren: Bestimme einen Anfangsvektor $v \in \mathbb{R}^m$, so dass die Lösung der Anfangswertaufgabe $\dot{u}^v = f(t, u^v)$, $u^v(a) = v$ die Randwerte $r(v, u^v(b)) = 0$ erfüllt. Damit das „NEWTON-Verfahren“ sinnvoll ist, muss die Lösung u^v der Anfangswertaufgabe differenzierbar von v abhängen.

Satz 4.7:

Sei $f \in C^1(I \times \mathbb{R}^m, \mathbb{R}^m)$. Dann ist u^v nach v differenzierbar mit $J = D_v u^v \in C^1(I, \mathbb{R}^{m \times m})$. J erfüllt die lineare Matrix-Anfangswertaufgabe $\dot{J} = D_2 f(t, u^v(t))J$ mit $J(a) = I$. Es gilt:

$$J_{ij}(t) = \lim_{\delta \mapsto 0} \left(u_i^{v+\delta e^j}(t) - u_i^v(t) \right)$$

Beweis:

Setze $w = v + \delta e^j$, wobei e^j der j-te Einheitsvektor ist.

$$\frac{1}{\delta}\left[u^w(t) - u^v(t)\right] = \frac{1}{\delta}\left[u^w(a) - u^v(a)\right] + \int_a^t \frac{1}{\delta}(\dot{u}^w(s) - \dot{u}^v(s))\,\mathrm{d}s =$$

$$= \frac{1}{\delta}(w - v) + \int_0^t \frac{1}{\delta}(f(s, u^w(s)) - f(s, u^v(s)))\,\mathrm{d}s = e^j + \int_a^t \frac{1}{\delta} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\lambda} f(s, u^v(s) + \lambda(u^w(s) - u^v(s)))\,\mathrm{d}\lambda\,\mathrm{d}s =$$

$$= e^j + \int_a^t \frac{1}{\delta} \int_0^1 D_2 f(s, u^v(s) + \lambda(u^w(s) - u^v(s)))(u^w(s) - u^v(s))\,\mathrm{d}\lambda\,\mathrm{d}s$$

Das GRONWALL-Lemma hat uns gezeigt, dass $\lim_{\delta \mapsto 0} |u^w - u^v| = 0$ ist.

$$\dot{J} = D_2 f(s, u^v(s))J \Rightarrow J(t)e^j = e^j + \int_a^t \dot{J}(s)e^j\,\mathrm{d}s = e^j + \int_a^t D_2(s, u^v(s))J(s)e^j\,\mathrm{d}s$$

Durch Vergleich ergibt sich dann:

$$\lim_{\delta \mapsto 0} \frac{1}{\delta}(u^w(s) - u^v(s)) = J(s)e^j \qquad \square$$

Der Algorithmus sieht nun wie folgt aus:

1.) Wähle Startvektor $v \in \mathbb{R}^m$.

2.) Berechne die Anfangswertaufgabe für u^v mit $\dot{u}^v = f(t, u^v)$ mit $u^v(a) = v$. Berechne $F(v) := r(v, u^v(b))$. Ist $|F(v)|$ klein genug, breche ab.

3.) Berechne eine Approximation ΔF von $DF(v) = D_1 r(v, u^v(b)) + D_2 r(v, u^v(b)) D_v u^v(b)$ mit $J(b) = D_v u^v(b)$ spaltenweise für $\delta > 0$ und für jede Spalte e^j setze $\Delta F(v)e^j = 1/\delta(F(v + \delta e^j) - F(v))$.

4.) Berechne $v := v - (\Delta F(v))^{-1} F(v)$ und gehe zu Schritt (2) zurück.

Bemerkung:

Seien f und r hinreichend glatt und u eine isolierte (lokal eindeutige Lösung) der Randwertaufgabe. Dann gilt:

a.) $F \in C^2(U, \mathbb{R}^m)$ in einer Umgebung U von $u(a)$.

b.) $F(u(a)) = r(u(a), u(b)) = 0$

c.) $DF(u(a))$ ist regulär.

Daraus folgt, dass das Schießverfahren konvergiert für Startwerte nahe bei $u(a)$.

3.11 Differenzenverfahren

Definition:

Zu $Lu(t) := \dot{u}(t) - A(t)u(t) = f(t)$ mit $u \in C^1(I = [a,b], \mathbb{R}^m)$ mit der Randbedingung $B_a u(a) + B_b u(b) = r$ betrachte eine Gitterfunktion u^h auf $\Delta = \{t_n = a + nh : n = 0, \ldots, N\}$ und $h = (b-a)/N$ als Lösung der „Differenzengleichung“ $(L_h u^h)_n := \sum_{j=0}^{N} C_{n_j}(h) u_j^h = E_n(hf)$ mit $R_h u^h := B_a u^h(a) + B_b u^h(b) = r$ für $n = 1$, …, N.

$$\mathcal{A}_h u^h = F^h \text{ mit } \mathcal{A}_h = \begin{pmatrix} B_a & 0 & \ldots & 0 & B_b \\ C_{10}(h) & & & & C_{1N}(h) \\ \vdots & & & & \vdots \\ C_{N0}(h) & & & & C_{NN}(h) \end{pmatrix} \in \mathbb{R}^{m(N+1),m(N+1)} \text{ und } F^h = \begin{pmatrix} r \\ F_1(h,f) \\ \vdots \\ F_N(h,f) \end{pmatrix}$$

3.11.1 Beispiel: Box-Schema

Das bekannteste Schema dieser Form ist das sogenannte „Box-Schema“.

$$(L_h u^j)_n = \frac{1}{h}(u_n^n - u_{n-1}^h) - A_{n-\frac{1}{2}} \left(\frac{1}{2} u_n + \frac{1}{2} u_{n-1} \right), \; F_n(h,f) = f(t_{n+\frac{1}{2}})$$

$t_{n-\frac{1}{2}} = \frac{1}{2}(t_{n-1} + t_n)$, $A_{n-\frac{1}{2}} = A(t_{n-\frac{1}{2}})$

Im Spezialfall $A = \text{const.}$ und $f = \text{const.}$ ist dies das Adams-Moulton-Verfahren.

$$\begin{pmatrix} B_a & 0 & \ldots & 0 & B_b \\ -I + \frac{1}{2}hA_{\frac{1}{2}} & I - \frac{1}{2}hA_{\frac{1}{2}} & \ldots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & -I - \frac{1}{2}hA_{N-\frac{1}{2}} & I - \frac{1}{2}hA_{N-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} u_0^h \\ \vdots \\ \vdots \\ u_N^h \end{pmatrix} = \begin{pmatrix} r \\ hf_{\frac{1}{2}} \\ \vdots \\ hf_{N-\frac{1}{2}} \end{pmatrix}$$

Definition 4.8:

a.) Ein Differenzenverfahren heißt konsistent von der Ordnung p, wenn für die Interpolation $I_h u$ mit $(I_h u)_n = u(t_n)$ der exakten Lösung gilt:

$$|R_h(I_n u) - r| + \max_{n=1,\ldots,N} |F_n(h,f) - (L_h(I_n u))_n| = O(h^p)$$

b.) Es heißt stabil, wenn

$$\max_{0 \le n \le N} |u_n^h| \le K \left(|R_h u^h| + \max_{1 \le n \le N} |(L_h u^h)_n| \right)$$

Bezeichnung:

Wir führen für Gitterfunktionen und unsere Diskretisierungsmatrix folgende Normen ein:

$$\|u^h\|_\infty = \max_{0 \le n \le N} |u_n^h| \text{ und } \|\mathcal{A}_h\| = \sup_{\|u^h\|_\infty = 1} \|A_h u^h\|_\infty$$

Lemma 4.9:

Die Differenzengleichung (4.7) ist genau dann stabil, wenn $\mathcal{A}_h$ regulär ist und $\sup_{h>0} \|A_h^{-1}\|_\infty < \infty$ ist.

Beweis: „⇒“

Wenn $\mathcal{A}_h u^h = 0$ ist, gilt $R_h u^h = 0$ und $L_h u^h = 0$. Unter dieser Voraussetzung gilt also

$$\|u^h\|_\infty \le K(|R_h u^h| + \max |(L_h u^h)_n|) = 0$$

und somit $u^h = 0$. Somit hat die Gleichung $A_h u^h = 0$ nur die triviale Lösung und $\mathcal{A}_h$ ist regulär! Entsprechend folgt aus $A_h u^h = F^h$, dass $F_0^h = R_h u^h$ und $F_n^h = (L_h u^h)_n$ ist.

$$\|\mathcal{A}_h^{-1} F^h\|_\infty = \|u^h\|_\infty \le \left[K(|R_h u^h) + \max_{1 \le n \le N} |(L_h u^h)_n| \right] \le 2K \|F^k\|_\infty \Rightarrow \|A_h^{-1}\|_\infty \le 2K$$

Beweis: „⇐“

Gilt umgekehrt $\|\mathcal{A}_h^{-1}\|_\infty \le C$, so folgt $\|u^h\|_\infty = \|A_h^{-1}F^h\|_\infty \le C\|F^h\|_\infty \le C\max\{|R_h u^h|, |(L_h u^h)_n|\}$. □

Satz 4.10:

Das Differenzenverfahren (4.7) sei konsistent von der Ordnung $p \ge 1$ und stabil. Dann ist es konvergent:

$\max\limits_{0\le n\le N} |u_n^h - u(t_n)| = O(h^p)$ für $h \mapsto 0$

Beweis:

Für den Fehler $e^h = u^h - I_h u$ gilt $L_h e^h = L_h u^h - L_h I_h u^h = F^h - L_h(I_h u)$ mit $F^h = F(h,f)$. Konsistenz bedeutet $\|L_h e^h\|_\infty = O(h^p)$, $|R_h e^h| = |r - R_h(I_h u)| = O(\tau^p)$.

$$\mathcal{A}_h e^h = \big(r - R_h(I_h u)F(h,f) - L_h(I_h u)\big) \Rightarrow \|e^h\|_\infty \le \|\mathcal{A}_h^{-1}\|_\infty \left\| \begin{pmatrix} r - R_h(I_h u) \\ F(h,f) - L_h(I_h u) \end{pmatrix} \right\|_\infty = O(h^p)$$

Satz 4.11:

Die Randwertaufgabe (4.2) sei eindeutig lösbar. Dann ist das Differenzenverfahren (4.7) genau dann stabil und konsistent, wenn die zugehörige Anfangswertaufgabe (mit $B_a = I$ und $B_b = 0$) stabil und konsistent ist.

Beweisskizze:

Betrachte $Lu = \dot u - Au = f$ mit $R^i u_i = B_a^i u(a) + B_b^i u(b) = r$ und $i = 0, 1$.

$$\mathcal{A}_h^i = \begin{pmatrix} B_a^i & 0 & \dots & 0 & B_b^i \\ C_{10}(h) & & & & \\ & \ddots & & & \\ & & & & C_{NN}(h) \end{pmatrix}$$

Zu zeigen ist:

$\|(\mathcal{A}_h^0)^{-1}\|_\infty < \infty \Leftrightarrow \|(A_h^1)^{-1}\|_\infty < \infty$

- „⇒“: $B_a^0 = I$, $B_b^0 = 0$, $B_a^1 = B_a$, $B_b^1 = B_b$

  $$D_h = \mathcal{A}_h^1 - \mathcal{A}_h^0 = \begin{pmatrix} B_a^1 - B_a^0 & 0 & B_b^1 - B_b^0 \\ 0 & 0 & 0 \end{pmatrix}$$

  Dann können wir eine Störungsrechnung machen der folgenden Form:

  $\mathcal{A}_h^1 = \mathcal{A}_h^0 + D_h = (I + D_h(\mathcal{A}_h^0)^{-1})\mathcal{A}_h^0$

  Zeige:

  i.) Die Matrix $I + D_h(\mathcal{A}_h^0)^{-1}$ ist regulär.

  ii.) Die Norm der inversen Matrix, also $\|(I + D_h(\mathcal{A}_h^0)^{-1})^{-1}\|$, ist $< C$ für $0 < h < h_0$.

  Dann gilt:

  $\|(\mathcal{A}_h^1)^{-1}\|_\infty \le \|(\mathcal{A}_h^0)^{-1}\|_\infty \|(I + D_h(\mathcal{A}_j^0)^{-1})^{-1}\|_\infty \Rightarrow \sup\limits_{0<h<h_0} \|(\mathcal{A}_h^1)^{-1}\|_\infty < \infty$

  Man benötigt weiterhin die NEUMANNsche Reihe und Störungsrechnung.

3.11.2 Modellproblem: „Fruchtfliege der Numerik“

Wir betrachten $-\ddot{u} = f$ mit $u(0) = u(1) = 0$ und nehmen an, dass $u \in C^2[0,1]$:

$$u(t) = u(0) + \int_0^t \dot{u}(s)\,\mathrm{d}s = \int_0^t \left(\dot{u}(0) + \int_0^s \ddot{u}(\theta)\,\mathrm{d}\theta \right) \mathrm{d}s = t\dot{u}(0) - \int_0^t \int_0^s -f(\theta)\,\mathrm{d}\theta\,\mathrm{d}s =$$

$$= c_1 t + \int_0^t F(s)\,\mathrm{d}s \text{ mit } c_1 := \dot{u}(0) \text{ und } F(s) := \int_0^s f(\theta)\,\mathrm{d}\theta$$

Wir wollen nun das Integral über $F(s)$ durch partielle Integration weiter auswerten:

$$\int_0^t F(s)\,\mathrm{d}s = sF(s)|_0^t - \int_0^t sf(s)\,\mathrm{d}s = t\int_0^t f(s)\,\mathrm{d}s - \int_0^t sf(s)\,\mathrm{d}s = \int_0^t (t-s)f(s)$$

Aus $u(1) = 0$ ergibt sich c_1:

$$c_1 + \int_0^1 F(s)\,\mathrm{d}s = 0 \Rightarrow c_1 = \int_0^1 (s-1)f(s)\,\mathrm{d}s$$

$$\boxed{u(t) = t\int_0^1 (s-1)f(s)\,\mathrm{d}s + \int_0^t (t-s)f(s)\,\mathrm{d}s = \int_0^1 G(t,s)f(s)\,\mathrm{d}s}$$

$G(t,s)$ bezeichnet man als Greensche Funktion.

$$\boxed{G(t,s) = \begin{cases} s(1-t) & \text{für } 0 \le s \le t \\ t(1-s) & \text{für } t \le s \le 1 \end{cases}}$$

Folgerung:

a.) $G(t,s) \ge 0$ für $s, t \in [0,1]$
Das Maximumprinzip besagt: Wenn $f \ge 0$ ist, dann ist $u \ge 0$.

b.) Für $f \in C[0,1]$ existiert genau eine Lösung.

c.) $\|u\|_\infty = \max_{t\in[0,1]} |u(t)| \le \max_{t\in[0,1]} \left| \int_0^1 G(t,s)f(s)\,\mathrm{d}s \right| \le \frac{1}{2}\int_0^1 (t(1-t))\,\mathrm{d}t \|f\|_\infty = \frac{1}{\gamma}\|f\|_\infty$
(Zeige, dass $\|u_n\|_\infty \le K\|f\|_\infty$ ist.)

Lemma 4.12:

Sei $h > 0$ und $I = [t-h, t+h]$. Dann gilt:

1.) $\dot{u}(t) = \frac{1}{h}(u(t+h) - u(t)) + hR_1(u)$ mit $|R_1(u)| \le \frac{1}{2}\|\ddot{u}\|_\infty$ $(u \in C^2(I))$

2.) $\dot{u}(t) = \frac{1}{h}(u(t) - u(t-h)) + hR_2(u)$ mit $|R_2(u)| \le \frac{1}{2}\|\ddot{u}\|_\infty$ $(u \in C^2(I))$

3.) $\dot{u}(t) = \frac{1}{2h}(u(t+h) - u(t-h)) + h^2 R_3(u)$ mit $|R_3(u) \le \frac{1}{6}\|\dddot{u}\|_\infty$ $(u \in C^3(I))$

4.) $\ddot{u}(t) = \frac{1}{h^2}(u(t+h) - 2u(t) + u(t-h)) + h^2 R_4(u)$ mit $|R_4(u) \le \frac{1}{12}\|\ddddot{u}\|_\infty$ $(u \in C^4(I))$

Beweis:

3.) $u(t \pm h) = u(t) \pm h\dot{u}(t) + \frac{1}{2}h^2\ddot{u}(t) \pm \frac{1}{6}h^3\dddot{u}(t \pm \xi_\pm)$ mit $0 < \xi_\pm < h$
Durch Differenzenbildung ergibt sich:

$$\frac{1}{2h}(u(t+h) - u(t-h)) = \dot{u}(t) + \frac{1}{12}h^2(\dddot{u}(t+\xi_+) + \dddot{u}(t-\xi_-)) \text{ mit } \left|\frac{1}{12}h^2(\dddot{u}(t+\xi_+) + \dddot{u}(t-\xi_-))\right| \le \frac{h^2}{12}2\|\dddot{u}\|_\infty \square$$

Definition:

Wir wenden $\partial_h u(t) := \frac{1}{2h}(u(t+h) - u(t-h))$ an auf:

$$Lu = -\frac{\mathrm{d}}{\mathrm{d}t}(p\dot{u}) + q\dot{u} + ru \Rightarrow L_h u = -\partial_{\frac{h}{2}}(p\partial_{\frac{h}{2}} u) + q\partial_h u + ru$$

Nun setzen wir $h = (b-a)/(N+1)$ und $t_n = a + nh$. Damit approximieren wir die Aufgabe $Lu = f$ mit $u(a) = u(b) = 0$ durch $L_h u^h(t_n) = f_n := f(t_n)$ und $u^h(a) = u_0^h = 0$, $u^h(b) = u_{N+1}^h = 0$. Gesucht ist die Matrix A_h für

$$A_h u^h = f^h \text{ mit } u^h = \begin{pmatrix} u_1^h \\ \vdots \\ u_N^h \end{pmatrix} \in \mathbb{P}^n \text{ und } f^h = \begin{pmatrix} f(t_1) \\ \vdots \\ f(t_N) \end{pmatrix}$$

Die Matrix A_h ist tridiagonal.

$$A_h = \begin{pmatrix} p_{\frac{1}{2}} + p_{\frac{3}{2}} + h^2 r_1 & -p_{\frac{3}{2}} + \frac{1}{2}hq_1 & & \\ \ddots & \ddots & & \\ & -p_{n-\frac{1}{2}} - \frac{1}{2}hq_n & p_{n-\frac{1}{2}} + p_{n+\frac{1}{2}} + h^2 r_n & -p_{n+\frac{1}{2}} + \frac{1}{2}hq_n \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

Beispiel:

Für $Lu = -\ddot{u}$ lautet die Matrix A:

$$A = \text{tridiag}(-1, 2, -1) = \begin{pmatrix} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & 0 & & \\ 0 & -1 & 2 & -1 & \ddots & \\ & 0 & -1 & \ddots & \ddots & 0 \\ & & \ddots & \ddots & \ddots & -1 \\ & & & 0 & -1 & 2 \end{pmatrix}$$

Satz 4.13:

Für $A \in \mathbb{R}^{N \times N}$ gelte:

a.) A sei stark diagonal-dominant. Das heißt

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}| \leq |a_{ii}| \text{ für } j = 1, \ldots, N$$

und es existiert ein $s \in \{1, \ldots, N\}$ mit

$$\sum_{j \neq s} |a_{sj}| < |a_{rs}|$$

b.) A ist irreduzibel. Das heißt, zu jedem Paar $s \neq r$ existiert eine Folge $s = j_1, \ldots, j_2, \ldots, j_m = r$ mit $a_{j_2 j_1} \neq 0$, $a_{j_3 j_2} \neq 0$, $\ldots$, $a_{j_m} a_{j_{m-1}} \neq 0$. (Zeichnet man einen Matrixgraphen, so sind dessen Kanten dadurch ausgezeichnet, dass die zugehörigen Einträge nicht verschwinden. Die Matrix ist also nicht in Blockmatrizen zerlegbar. Das zugehörige Problem zerfällt nicht in zwei voneinander unabhängige Probleme.)

Dann gilt:

1.) A ist regulär.

2.) Falls $a_{ii} > 0$, ist A positiv definit.

3.) Falls $a_{ii} > 0$ und $a_{ij} < 0$ (für $i \neq j$) ist $A^{-1} \geq 0$ (elementweise). (Dann folgt aus $Au = f$ und $f \geq 0$, dass $u \geq 0$ ist.)

Solche Matrizen werden als M-Matrizen bezeichnet.

Beweis:

Da A irreduzibel ist, muss $\sum_{j=1}^{N} |a_{ij}| > 0$ sein für $i = 1, \ldots, N$. Insbesondere folgt aus Voraussetzung (a), dass $|a_{ii}| \neq 0$ ist. Zu $A = D + L + R$ mit $D = \text{diag}(A)$, $L = \text{lower}(A)$ und $R = \text{upper}(A)$ gilt $J := D^{-1}(L + R) = I - D^{-1}A$. Behauptung: Es gilt $\varrho(J) < 1$. Sei $\mu \in \mathbb{C}$ Eigenwert von J, also $J\omega = \mu\omega$ mit $\omega \neq 0$. Ohne Einschränkung nehmen wir an, dass $\|\omega\|_\infty = 1 = |\omega_r|$ ist. Damit gilt:

$$|\mu| = |\mu\omega_r| = \left|a_{rr}^{-1} \sum_{j\neq i} a_{rj}\omega_j\right| \leq |a_{rr}^{-1}| \sum_{j\neq i} |a_{rj}| \underset{\text{(a)}}{\leq} 1$$

Wir nehmen nun an, dass $|\mu| = 1$ ist. Wähle $j_1 = s$ aus (a) bis $j_m = r$ mit $|\omega_r| = 1$.

$$|\omega_s| = |\mu\omega_s| \leq |a_{ss}^{-1}| \sum_{j\neq s} |a_{sj}||\omega_j| < 1$$

$$|\omega_{j_2}| = |\mu\omega_{j_2}| \leq a_{j_2 j_2}^{-1} | \left(\sum_{\substack{j\neq j_2 \\ j\neq s}} |a_{j_2 j}||\omega_j| + \underbrace{|a_{j_2 s}|}_{\neq 0} |\omega_s| \right) < 1$$

$$1 = |\omega_r| = |\mu\omega_r| \leq |a_{rr}^{-1}| \left(\sum_{\substack{j\neq r \\ j\neq j_{m-1}}} |a_{rj}||\omega_j| + \underbrace{|a_{rj_{m-1}}|}_{\neq 0} |\omega_{s_{m-1}}| \right) < 1$$

Dies ist ein Widerspruch und damit ist $\varrho(J) < 1$.

$$(I - J)^{-1} = \sum_{k\geq 0} J^k$$

konvergiert als Neumannsche Reihe. Damit ist A^{-1} invertierbar

$$(I - J)^{-1} = (D^{-1}A)^{-1} = A^{-1}D \Rightarrow A^{-1} = (I - J)^{-1}D$$

und A regulär.

3.12 Variationsmethoden

Wir betrachten die Sturm-Liouville-Randwertaufgabe:

$$Lu(t) = -\frac{\mathrm{d}}{\mathrm{d}t}(p(t)\dot{u}(t)) + q(t)\dot{u}(t) + r(t)u(t) = f(t) \text{ mit } t \in I = [a, b]$$

Dann gilt für jede Lösung $u \in C^2(I)$:

$$\int_a^b Lu\varphi \,\mathrm{d}t = \int_a^b f\varphi \,\mathrm{d}t$$

Definiere $\hat{V} = \{\varphi \in C(I) : \varphi(a) = \varphi(b) = 0, \varphi \text{ stückweise differenzierbar}\}$.

Beispiel:

Lineare Splines sind in $\hat{V}$. Zu $\varphi \in \hat{V}$ wähle eine Zerlegung $a = t_0 < t_1 < \ldots < t_N = b$, so dass $\varphi \in C^1(t_{n-1}, t_n)$. Hieraus folgt:

$$-\int_a^b \frac{\mathrm{d}}{\mathrm{d}t}(p\dot{u})\varphi \,\mathrm{d}t = -\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \frac{\mathrm{d}}{\mathrm{d}t}(p\dot{u}) \,\mathrm{d}t = \sum_{n=1}^{N} \left(-(p\dot{u}\varphi)|_{t_{n-1}}^{t_n} + \int_{t_{n-1}}^{t_n} p\dot{u}\dot{\varphi} \,\mathrm{d}t \right) = p(a)\dot{u}(a)\varphi(a) - p(b)\dot{u}(b)\varphi(b) + \int_a^b p\dot{u}\dot{\varphi} \,\mathrm{d}t$$

$$\int_a^b Lu\varphi \,\mathrm{d}t = \int_a^b (p\dot{u}\dot{\varphi} + q\dot{u}\varphi + ru\varphi) \,\mathrm{d}t$$

Wir führen folgende Bezeichnungen ein:

$$\|v\| = \left(\int_a^b (v(t))^2 \, \mathrm{d}t \right)^{\frac{1}{2}}$$

Diese Norm bezeichnet man auch als $L_2(I)$-Norm. Die Bilinearform auf $\hat{V}$ ist gegeben durch

$$a(u,v) = \int_a^b (p\dot{u}\dot{v} + q\dot{u}v + uv) \, \mathrm{d}t$$

und die Linearform auf $\hat{V}$ durch

$$l(v) = \int_a^b fv \, \mathrm{d}t$$

Satz 4.15:

a.) Er gilt die Poincare-Ungleichung:

$$\|v\| \leq \frac{b-a}{2} \|\dot{v}\| \text{ mit } v \in \hat{V}$$

b.) $l(\bullet)$ ist stetig bezüglich $\|\dot{v}\|$: $|l(v)| \leq C\|\dot{v}\|$ mit $v \in \hat{V}$

c.) $a(\bullet,\bullet)$ ist stetig: $|a(u,v) \leq C\|\dot{u}\|\|\dot{v}\|$ mit u, $v \in \hat{V}$

d.) Sei zusätzlich nun $p(t) \geq \varrho > 0$ und $\varrho + ((b-a)/2)^2 \min(r - 1/2\dot{q}) > 0$. Dann ist $a(\bullet,\bullet)$ elliptisch, das heißt: $a(v,v) \geq c\|\dot{v}\|^2$ mit $v \in \hat{V}$.

Beweis:

a.) Für $v \in \hat{V}$ gilt $v(a) = v(b) = 0$ und

$$v(t) = v(a) + \int_a^t \dot{v}(s) \, \mathrm{d}s = \int_a^t \dot{v}(s) \, \mathrm{d}s = -\int_a^b \dot{v}(s) \, \mathrm{d}s$$

$$|v(t)| \leq \left(\int_a^t 1 \, \mathrm{d}s \right)^{\frac{1}{2}} \left(\int_a^t (\dot{v}(s))^2 \, \mathrm{d}s \right)^{\frac{1}{2}}$$

Dies ist die Cauchy-Schwarzsche Ungleichung. Hieraus folgt $|v(t)|^2 \leq (t-a)\|\dot{v}\|^2(a+b)/2$.

$$\int_a^{\frac{a+b}{2}} |v(t)|^2 \, \mathrm{d}t \leq \|\dot{v}\|^2 \int_a^{\frac{a+b}{2}} (t-a) \, \mathrm{d}t = \|\dot{v}\|^2 \frac{1}{2} \left(\frac{b-a}{2} \right)^2 \Rightarrow \|v\|^2 \leq \left(\frac{b-a}{2} \right)^2 \|\dot{v}\|^2$$

b.) Mit der gleichen Überlegung folgern wir $C = (b-a)/2\|f\|$.

c.) Wir machen folgende Abschätzung:

$$a(u,v)| \leq \|p\|_\infty \|\dot{u}\|\|\dot{v}\| + \|g\|_\infty \|\dot{u}\|\|v\| + \|r\|_\infty \|u\|\|v\| \leq C\|\dot{u}\|\|\dot{v}\| \text{ mit } C = \|p\|_\infty + \frac{b-a}{2}\|q\|_\infty + \frac{(b-a)^2}{4}\|r\|_\infty$$

d.) Mittels partieller Integration und der Poincare-Ungleichung ergibt sich folgendes. Wir wählen ein $\varrho_0 \in (0, \varrho)$ mit $\varrho_0 + (b-a)^2 \min(r - 1/2\dot{q}) > 0$. Hieraus folgt:

$$a(v,u) = \int_a^b (p|\dot{v}|^2 + q\dot{v}v + r|v|^2) \, \mathrm{d}t = \int_a^b \left[(p-\varrho_0)|\dot{v}|^2 + \varrho_0|\dot{v}|^2 - \frac{1}{2}\dot{q}|v|^2 + r|v|^2 \right] \mathrm{d}t \geq (\varrho - \varrho_0)\|\dot{v}\|^2 + \varrho_0 \underbrace{\left(\|\dot{v}\|^2 - \left(\frac{2}{b-a} \right)^2 \|v\|^2 \right)}_{\geq 0 \text{ (Poincare-Ungleichung)}} +$$

Dies ist eine gleichmäßige Schranke für $\|\dot{v}\|^2$ und damit ist nach Definition die Elliptizität gewährleistet.
□

Folgerung 4.16:

Unter der Bedingung 4.15 (d) hat die STURM-LIOUVILLE-Randwertaufgabe eine eindeutige Lösung $u \in C^2(I)$.

Beweis:

Aus der FREDHOLMschen Alternative folgt: Wenn die homogene Aufgabe $Lu \equiv 0$ mit $u(a) = u(b) = 0$ nur die triviale Lösung $u \equiv 0$ besitzt, so ist $Lu = f$ mit $u(a) = u(b) = 0$ für jede rechte Seite eindeutig lösbar. Sei also $Lu = 0$. Dann folgt hieraus, dass auch die Bilinearform $a(u, u)$ verschwindet.

$$0 = a(u,u) \geq C\|\dot{u}\| \Rightarrow \dot{u} = 0 \xrightarrow{\text{Satz 4.15 (a)}} \|u\| = 0 \Rightarrow u \equiv 0$$ □

3.12.1 Galerkin-Verfahren

Wir wählen einen endlichdimensionalen Teilraum $V_h \subset \hat{V}$ mit Basis φ_1^h, ..., φ_N^h. Außerdem definieren wir die diskretisierte Matrix A_h (Massenmatrix)

$$A_h = \left(a(\varphi_n^h, \varphi_m^h)\right)_{n,m=1,\ldots,N} \in \mathbb{R}^{N,N}$$

und $f_h = (l(\varphi_n^h)) \in \mathbb{R}^N$. Das heißt $u_h = \sum_{n=1}^N \vec{u}_n \varphi_n^h \in V_h$ löst $a(u_h, v_h) = l(v_h)$ für alle $v_h \in V_h$. Also ist das lineare Gleichungssystem $A_h \vec{u} = f_h$ zu lösen.

Satz 4.17:

Sei $a(\bullet, \bullet)$ eine elliptische Bilinearform. Dann ist die Matrix A_h positiv definit und damit regulär. Also existiert eine eindeutige Lösung $u_h \in V_h$ von der Variationsgleichung $a(u_h, v_h) = l(v_h)$ für alle $v_h \in V_h$.

Beweis:

Für $v_h = \sum_{n=1}^N \vec{v}_n \varphi_n^h$ folgt aus $v_h \neq 0$, dass $a(v_h, v_h) > 0$ ist. Damit ist $\vec{v}^\intercal A_h \vec{v} > 0$ und damit A_h positiv definit, also auch regulär. Damit können wir den Vektor $\vec{u}$ explizit angeben. $\vec{u} = A_h^{-1} f_h$ löst die Variationsgleichung eindeutig. □

Satz 4.18:

Sei $u \in C^2(I)$ Lösung der STURM-LIOUVILLE-Randwertaufgabe und sei $u_h \in V_h$ Lösung von 4.17. Dann gilt für den Fehler $e_h = u - u_h$ die „GALERKIN-Orthogonalität“ $a(e_h, v_h) = 0$ für $v_h \in V_h$. (Der Fehler bezüglich der Bilinearform steht also senkrecht auf allen Testfunktionen, also implizit die Bestapproximation in diesem Raum darstellt. Siehe auch WEISERSTRASSscher Approximationsatz)

Beweis:

Aus $Lu = f$ folgt $a(u, v_h) = l(v_h)$ und $a(u_h, v_h) = l(v_h)$. Durch Subtraktion ergibt sich direkt:

$$a(u - u_h, v_h) = a(e_h, v_h) = l(v_h) - l(v_h) = 0$$ □

3.12.2 Ceas-Lemma

Satz 4.19:

Sei eine Bilinearform $a(\bullet, \bullet)$ beschränkt und elliptisch, also $|a(u,v)| \leq C\|\dot{u}\|\|\dot{v}\|$, $a(v,v) \geq C\|\dot{v}\|^2$. Dann gilt für den GALERKIN-Fehler $e_h = u - u_h$ die Ungleichung

$$\|\dot{e}_h\| \leq \frac{C}{c} \inf_{v_h \in V_h} \|\dot{u} - \dot{v}_h\|$$

(Es macht also Sinn, die Elemente aus dem Testraum gut mit u verträglich sind. Desto kleine ist dann der Fehler und desto besser die Lösung.

Beweis:

Wir beginnen mit der Elliptizität:

$$c\|\dot{e}_h\|^2 \leq a(u-u_h, u-u_h) = a(u-u_h, u) = a(u-u_h, u-v_h) \leq \inf_{v_h \in V_h} C\|\dot{u}-\dot{u}_h\|\|\dot{u}-\dot{v}_h\|$$

Dies gilt für alle Testfunktionen $v_h \in V_h$, weil diese die GALERKIN-Orthogonalität erfüllen. Hieraus ergibt sich:

$$c\|\dot{e}\| \leq C \int_{v_h \in V_h} \|\dot{u}-\dot{v}_h\|$$

Anwendung: Wir konstruieren den Raum $V_h \subset \hat{V}$ und definieren eine Interpolation I_h: $\hat{V} \mapsto V_h$ mit $\|\mathrm{d}/\mathrm{d}t(U - I_h u)\| = O(h^p)$. Dann gilt:

$$\|\dot{u}-\dot{u}_h\| \leq \frac{C}{c}\left\|\frac{\mathrm{d}}{\mathrm{d}t}(u-I_h u)\right\| = O(h^p)$$

3.12.3 Beispiel: Finite Differenzen

$-\ddot{u} + ru = f$, $Lu = -\ddot{u} + ru$ für $r \geq 0$

Wir wollen den Differentialoperator mit homogenen Randwerten betrachten, also $u(a) = u(b) = 0$. Nun definieren wir den diskreten Differentialoperator, welcher auf Gitterfunktionen operiert:

$$L_h u(t) = \frac{1}{2h^2}\left[-u(t-h) + 2u(t) - u(t+h)\right] + r(t)u(t)$$

Sei u_h die Gitterfunktion:

$$u_h\ \Delta \mapsto \mathbb{R},\ \vec{u}^h = \begin{pmatrix} u_h(t_1) \\ \vdots \\ u_h(t_N) \end{pmatrix}$$

$A_h \vec{u}^h = \vec{f}^h$ mit $A_h = \mathrm{tridiag}(-1, 2+r_n, -1)$ und $f^h = (f(t_n))_{n=1,\ldots,N}$

1.) Konsistenz:

$$\max|(L_h u_h - L_h(I_h u))(t_n)| = \max|-\ddot{u}(t_n) - r_n u(t_n) - L_h(I_h u)(t_n)| \leq \frac{1}{12}h^2\|\dddot{u}\|_\infty$$

Nur wenn vierte Ableitungen existieren, können wir eine solche Aussage machen.

2.) Stabilität:
Es ist folgendes zu zeigen:

$$\max|u_h(t_n)| \leq C \max|L_h u_h(t_n)|$$

- 1.Schritt: Diskretes Maximumsprinzip
  Seien a_n, b_n, $c_n > 0$ $(n = 1, \ldots, N)$ mit $b_n \geq a_n + c_n$ und für $u_n \in \mathbb{R}$ $(n = 0, \ldots, N+1)$ gelte

  $$-a_n u_{n-1} + b_n u_n - c_n u_{n+1} \leq 0 \text{ für } n = 1, \ldots, N$$

  Dann gilt $u_n \leq K := \max\{0, u_0, u_{N+1}\}$.

  Beweis:

  Sei $u_k = \max\{u_0, \ldots, u_{N+1}\}$. Ohne Einschränkung gelte $u_h > 0$ für $h \neq 0,\ N+1$. Zeige: Dann ist u_k konstant. Aus $u_k \geq u_{k-1}$ und $u_k \geq u_{k+1}$ folgt:

  $$b_k u_k \leq a_k u_{k-1} + c_k u_{k+1} \leq (a_k + c_k)u_k \leq b_k u_k \Rightarrow u_{k-1} = u_{k+1} = u_k$$

  Induktiv ergibt sich dann $u_0 = u_{N+1} = u_k$. □
- 2.Schritt: L_h ist invers monoton. Dann folgt aus $L_h u_h \geq 0$, dass $u_h \geq 0$ ist.

Beweis:

Für $a_n = c_n = 1$ gilt $b_n = 2 + v_n \geq a_n + b_n$. Aus $(L_h u_h) \geq 0$ folgt dann:

$$-a_n(-u_h(t_{n-1})) + b_n(-u_h(t_n)) - c_n(-u_h(t_{n+1})) \leq 0 \Rightarrow \max(-u_n(t_n)) = 0 \Rightarrow u_h \geq 0$$ □

✳ 3.Schritt: L_h^{-1} ist beschränkt. Wir definieren eine Gitterfunktion $w_h(t_n) = 1/2(t_n - a)(b - t_n)$.

$L_h w_h(t_n) = 1 + r_h w_h(t_0) \geq 1$

Also gilt für $L_u u_h = f_h$ und $v_h = \|f_h\|_\infty w_h - u_h$.

$L_h v_h = \|f_h\|_\infty L_h w_h - L_h u_h \geq \|f_h\|_\infty - f_h \geq 0$

Da der Operator invers monoton ist, folgt $v_h \geq 0$ und damit $u_h \leq \|f_h\|_\infty w_h$. Analog gilt $-u_h \leq \|f_h\|_\infty w_h$. Hieraus resultiert:

$$\max|u_h(t_n)| \leq \frac{1}{8}(b-a)^2 \max|L_h u_h(t_n)|$$ □

Nun können wir unsere Konstante C explizit angeben. Wir lesen $C = 1/8(b-a)^2$ ab. (Dies gilt unabhängig von h.) Aus Stabilität und Konsistenz folgt nun Konvergenz:

$$\|u - u_h\|_\infty = \|L_h^{-1} L_h(u - u_n)\|_\infty \leq \frac{1}{8}(b-a)^2 \|L_h(u - u_h)\|_\infty \leq \frac{1}{96}(b-a)^2 h^2 \|\ddot{u}\|_\infty$$

Satz:

Sei I_h: $\widehat{V} \mapsto V_h$ die lineare Interpolation und sei $u \in C^2(I)$. Dann gilt:

a.) $\|u - I_h u\| \leq h^2 \|\ddot{u}\|$

b.) $\left\|\frac{\mathrm{d}}{\mathrm{d}t}(u - I_h u)\right\| \leq h\|\ddot{u}\|$

c.) $\|u - I_h u\|_\infty \leq \frac{1}{2} h^2 \|\ddot{u}\|_\infty$

Beweis:

b.) Sei $w(t) = u(t) - U_h u(t)$. Hieraus folgt $w(t_n) = 0$ und damit existiert ein $\xi \in (t_{n-1}, t_n)$ mit $\dot{w}(\xi_n) = 0$. Für $t \in [t_{n-1}, t_n]$ folgt:

$$|\dot{w}(t)| = \left|\int_{\xi_n}^{t} \ddot{w}(s)\,\mathrm{d}s\right| \leq \left(\left|\int_{\xi_n}^{t} 1^2\,\mathrm{d}s\right|\right)^{\frac{1}{2}} \left(\left|\int_{\xi_n}^{t} |\ddot{w}(s)|^2\,\mathrm{d}s\right|\right)^{\frac{1}{2}}$$

$$\int_{t_{n-1}}^{t_n} |\dot{w}(t)|^2\,\mathrm{d}t = \int_{t_{n-1}}^{t_n} |t - \xi_n| \left|\int_{\xi_n}^{t} |\ddot{w}(s)|^2\,\mathrm{d}s\right| \leq h^2 \int_{t_{n-1}}^{t_n} |\ddot{w}(t)|^2\,\mathrm{d}t \Rightarrow \|\dot{w}\| \leq h\|\ddot{w}\| = h\|\ddot{u}\|$$

Nach dem Ceas-Lemma gilt:

$$\|\dot{u} - \dot{u}_h\| \leq \frac{C}{c}\left\|\frac{\mathrm{d}}{\mathrm{d}t}(u - U_h)\right\| \leq \frac{C}{c} h\|\ddot{u}\|$$ □

Satz 4.22:

Sei $a(\bullet, \bullet)$ elliptisch. Dann gilt für die Lösung $u \in C^2(I)$ der Randwertaufgabe $\|\ddot{u}\| \leq C\|f\|$.

Beweis:

Mit $\|u\| \leq C_p\|\dot{u}\|$ (siehe Übungen) gilt:

$$\|\dot{u}\|^2 \leq \frac{1}{c}a(\bullet,\bullet) = \frac{1}{c}l(u) \leq \frac{1}{c}\|f\|\|u\| \leq \frac{C_p}{c}\|f\|\|\dot{u}\| \Rightarrow \boxed{\|\dot{u}\| \leq \frac{C_p}{c}\|f\|}$$

Betrachten wir weiter:

$$-\frac{\mathrm{d}}{\mathrm{d}t}(p\dot{u}) + q\dot{u} + ru = f \Rightarrow -p\ddot{u} - \dot{p}\dot{u} + q\dot{u} + ru = f$$

Mit $p(t) \geq \varrho > 0$ und der Dreiecksungleichung folgt dann:

$$\|\ddot{u}\| = \left\|\frac{f - ru - q\dot{u} + \dot{p}\dot{u}}{p}\right\| \leq \frac{1}{\varrho}\left(\|f\| + \|r\|_\infty\|u\| + \|q\|_\infty\|\dot{u}\| + \|\dot{\varrho}\|_\infty\|\dot{u}\|\right) \leq$$

$$\leq \underbrace{\frac{1}{\varrho}\left[1 + \frac{C_p}{c}\left(C_p\|r\|_\infty + \|\dot{p} - q\|_\infty\right)\right]}_{\equiv C}\|f\| \qquad \square$$

Satz 4.23:

Unter der Voraussetzung (4.19) gilt für die Galerkin-Lösung $u_h \in V_h$:

a.) $\|\dot{u} - \dot{u}_h\| \leq Ch\|f\|$

b.) $\|u - u_h\| \leq Ch^2\|f\|$ (da $u \in C^2(I)$!)

Beweis:

b.) Betrachte die adjungierte Aufgabe zu $g \in C(I)$:

$a(v_h, w_h) = \int_a^b g v_h \,\mathrm{d}t$ für alle $v_h \in V_h$ und $w_h \in V_h$

Analog zu (4.21) gilt für $w \in \hat{V}$ mit der Eigenschaft

$a(v, \hat{w}) = \int_a^L g v \,\mathrm{d}t$ für alle $v \in \hat{V}$

gilt $w \in C^2(I)$ mit $\|\dot{w} - \dot{w}_h\| \leq Ch\|g\|$. Setze $g = u - u_h$:

$$\|g\|^2 = \|u - u_h\|^2 = \int_a^b g\underbrace{(u - u_h)}_{\leq v \in \widehat{V}}\,\mathrm{d}t = a(u - u_h, w) \overset{\text{Galerkin}}{=} a(u - u_j, w - w_h) \overset{4.19\,(\mathrm{d})}{\leq}$$

$$\leq C\|\dot{u} - \dot{u}_h\|\|\dot{w} - \dot{w}_h\| \overset{(\mathrm{a})}{\leq} C'(h\|f\|)(h\|g\|)$$

$$\Rightarrow \|u - u_h\| \leq C'h^2\|f\| \qquad \square$$

3.12.4 Dualer Fehlerschätzer

Sei $J(\bullet)$ ein „Fehlerfunktional“ und sei $z \in \widehat{V}$ duale Lösung $a(\varphi, z) = J(z)\ \forall\ \varphi \in \widehat{V}$. Setze $\varphi = e_h = u - u_h$. Dann gilt für $z_h \in V_h$:

$$a(e_h, z) = a(e_h, z - z_h) = l(z - z_h) - a(u_h, z - z_h) = \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n} h(f - L_h u_h)h^{-1}(z - z_h) +$$

$$+ \underbrace{\text{Terme durch partielles Integrieren in den Teilschritten}}_{=0} \overset{\text{Cauchy–Schwarz}}{=}$$

$$= \left(\sum_{n=1}^{N} h^2 \int_{t_n - 1}^{t} (f - L_h u_h)^2\right)\left(h^{-2}\|z - I_h z\|^2\right)$$