Numerik für Differentialgleichungen

Malte Braack

Mathematisches Seminar

Christian-Albrechts-Universität zu Kiel

Vorlesungsskript, SoSe 2013,
Stand: 26.05.2013

Inhaltsverzeichnis

Kapitel 1

Beispiele von Differentialgleichungen

1.1 Gewöhnliche Differentialgleichungen

1.1.1 Federpendel

Das Hook'sche Gesetz besagt, dass sich die Rückstellkraft einer Feder linear zur Auslenkung $x(t)$ verhält:

$$F(t) \quad = \quad -Dx(t)\,.$$

Diese Kraft F wirkt wie eine Beschleunigung a auf das Gewicht (siehe Abb. 1.1)

$$F(t) \quad = \quad ma(t)\,.$$

Die Beschleunigung a ist die Ableitung der Geschwindigkeit $v(t)$, also $v'(t) = a(t)$, während die Geschwindigkeit die Ableitung der Ablenkung $x(t)$ ist, d.h. $x'(t) = v(t)$.

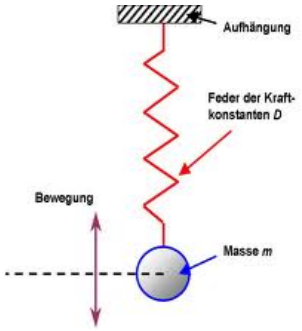

Abbildung 1.1: Federpendel mit Massse m.

Insgesamt erhalten wir

$$x''(t) \;=\; -\frac{D}{m}x(t)\,. \tag{1.1}$$

Dies ist eine lineare Differentialgleichung zweiter Ordnung. Lösungen dieser Gleichung sind von der Gestalt

$$x(t) \;=\; \alpha\sin(\omega t) + \beta\cos(\omega t)$$

mit der Frequenz $\omega = \sqrt{c/m}$. Die Parameter $\alpha, \beta \in \mathbb{R}$ sind bislang noch beliebig. Fügt man der Gleichung (1.1) noch Anfangsbedingungen hinzu, z.B.

$$x(t_0) \;=\; a\,, \qquad x'(t_0) \;=\; b\,,$$

so lassen sich aus a und b auch α, β bestimmen.

1.1.2 Populationsmodell in der Biologie

$r(t)$ = Rabbits / Kanninchen und $f(t)$ = foxes / Füchse.

1. unbeschränkter Futtervorrat für Kanninchen
2. Kaninchen einzige Nahrung für Füchse

$$\begin{aligned} r'(t) &= r(t) - \alpha r(t) f(t) \\ f'(t) &= -f(t) + \alpha r(t) f(t) \\ r(0) &= r_0 \\ f(0) &= f_0 \end{aligned}$$

Als System schreibt sich dies auch in der Form

$$\begin{pmatrix} r \\ f \end{pmatrix}' \;=\; \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} r \\ f \end{pmatrix} + \alpha r f \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Im Fall $\alpha = 0$ sind beide Größen entkoppelt und die Lösung lautet einfach

$$\begin{aligned} r(t) &= r_0 \exp(t) \qquad \text{exponentielles Wachstum} \\ f(t) &= f_0 \exp(-t) \qquad \text{exponentielles Verschwinden}\,. \end{aligned}$$

Im Fall $\alpha > 0$ dezimieren die Füchse die Kaninchen. Die Differentialgleichung wird dann nichtlinear.

1.1.3 Chemische Reaktionskinetik

Drei chemische Stoffe A_1, A_2, A_3 die gemäß des folgenden Reaktionsmechanismuses reagieren:

$$A_1 \to A_2, \quad A_2 + A_3 \to A_1 + A_3, \quad 2A_2 \to A_3$$

Diese drei Reaktionen laufen mit jeweiligen Reaktionsgeschwindigkeiten k_1, k_2 und k_3 ab. In der zweiten Reaktion wirkt A_3 wie ein Katalysator. Die zeitlichen Änderungen der jeweiligen Konzentrationen c_i, $i \in \{1, 2, 3\}$ lauten:

$$\begin{aligned} c_1'(t) &= -k_1 c_1(t) + k_2 c_2(t) c_3(t) \\ c_2'(t) &= k_1 c_1(t) - k_2 c_2(t) c_3(t) - 2k_3 c_2(t) \\ c_3'(t) &= 2k_3 c_2(t)\,. \end{aligned}$$

1.1.4 Lorenz-System

Das Lorenz-System ist abgeleitet von den Navier-Stokes Gleichungen, die erheblich komplezierter sind und Strömungsvorgänge beschreiben. Das vereinfachte System besteht aus drei skalaren, zeitabhängigen, miteinander gekoppelten Größen:

$$\begin{aligned} x'(t) &= -\sigma x(t) + \sigma y(t) \\ y'(t) &= r x(t) - y(t) - x(t) z(t) \\ z'(t) &= x(t) y(t) - b z(t)\,. \end{aligned}$$

Interessant ist dieses nichtlineare System besonders für die Parameter $\sigma = 10$, $b = 8/3$ und $r = 28$. Als Anfangswerte wählen wir $(x_0, y_0, z_0) = (1, 0, 0)$. Es zeigt sich, dass diese System eine eindeutige Lösung für beliebiges $t > 0$ besitzt. Allerdings ist diese Lösung extrem instabil in dem Sinne, dass kleinste Änderungen in den Anfangswerten zu einer starken Verstärkung führen. Für $t = 25$ besitzt der Verstärkungsfaktor die Größenordnung 10^8. Eine analytische Lösung gibt es nicht und eine numerische Berechnung (Approximation) für große Zeiten t (≥ 25) ist aufgrund dieser Instabilität praktisch nicht möglich. Man kann den zeitlichen Verlauf als Kurve im Raum darstellen indem man x, y, z als die jeweilige Koordinate betrachtet.

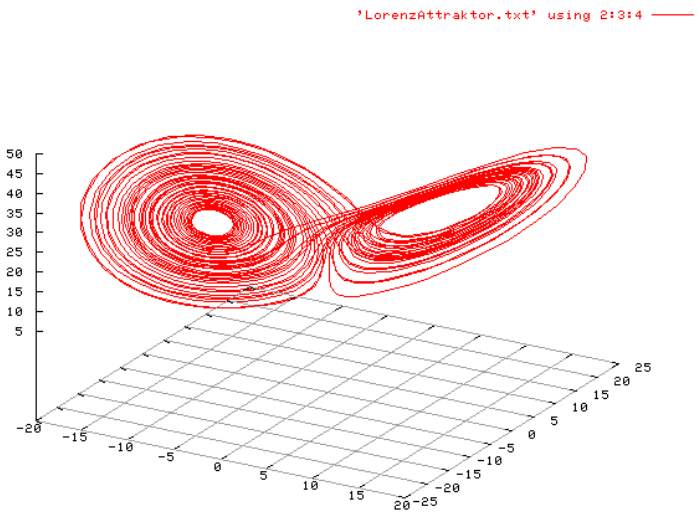

Abbildung 1.2: Trajektorie des Lorenz-Systems $(x(t), y(t), z(t))$.

1.2 Partielle Differentialgleichungen

Bei den bisherigen Beispielen hatten wir es stets mit Ableitungen bezüglich einer Variablen t (häufig die Zeit) zu tun. Unter partiellen Differentialgleichungen versteht man Differentialgleichungen, bei denen man in mehrere Richtungen ableitet.

Die eindimensionale Wärmeleitungsgleichung lautet

$$\begin{aligned} \frac{\partial}{\partial t}u(x,t) - \kappa\frac{\partial^2}{\partial x^2}u(x,t) &= f(x,t) \qquad (x,t) \in (a,b)\times(0,T)\,, \\ u(a,t) = u(b,t) &= 0 \qquad t \in [0,T]\,, \\ u(x,0) &= u_0 \qquad x \in (a,b)\,. \end{aligned}$$

Hierbei bezeichnet κ die Wärmeleitfähigkeit des Materials und f eine äußere Wärmequelle. Darüberhinaus treten hier sogenannte Randwerte an den Stellen $x = a$ und $x = b$ auf. Die Temperatur soll hier 0 betragen. Die anfängliche Wärmeverteilung ist durch u_0 gegeben. Selbstverständlich gibt es auch Verallgemeinerungen in mehrere Raumdimensionen.

Kapitel 2

Analytische Grundlagen

In diesem ganzen Abschnitt sei $\Omega \subset \mathbb{R}^n$ offen, I ein reelles Intervall und $D = I \times \Omega$.

2.1 Existenz von Lösungen bei Anfangswertaufgaben

Wir betrachten die allgemeine Anfangswertaufgabe (AWA) zu $f \in C(D)$. Gesucht ist $u \in C^1(I, \Omega)$, so dass

$$u'(t) = f(t, u(t)) \qquad t \in I\,, \tag{2.1}$$
$$u(t_0) = u_0\,. \tag{2.2}$$

Im Fall, dass f (affin) linear ist lässt sich dieses Problem darstellen in der Form $f(t, x) = A(t)x + b(t)$. In diesem Fall spricht man von einer linearen Differentialgleichung. Dies werden wir im nächsten Abschnitt behandeln.

Wir bezeichnen eine abgeschlossene Kugel mit Radius $r > 0$ um u_0 in einer Norm $\|\cdot\|$ mit $\overline{B}_r(u_0) := \{x \in \mathbb{R}^n \,\big|\, \|x - u_0\| \leq r\}$. Der folgende Satz liefert uns lokale Existenz von Lösungen.

Satz 2.1 (Existenzsatz von Peano) *Sei $f \in C(Z, \mathbb{R}^n)$ auf dem Zylinder $Z = \overline{I} \times \overline{B_r(u_0)}$. Dann existiert eine Lösung $u \in C^1(J, \overline{B}_r(u_0))$ der AWA (2.1)-(2.2) mit*

$$J := I \cap [t_0 - \epsilon, t_0 + \epsilon] \quad \textit{und} \quad \epsilon := \frac{r}{\|f\|_{L^\infty(Z)}}\,.$$

Beweis. Den Beweis erhält man beispielsweise über die Formulierung eines numerischen Verfahrens (expliziter Euler/Polygonzugverfahren), den Grenzübergang Gitterweite $h \to 0$ und den Satz von Arzela-Ascoli. □

Im allgemein Fall benötigt man für die Existenz von globalen Lösungen, also $u \in C^1(I, \mathbb{R}^n)$, zum Beispiel eine Beschränkung von f:

Satz 2.2 *Sei $f \in C(I \times \mathbb{R}^n, \mathbb{R}^n)$ mit der Wachstumsbeschränkung*

$$\|f(t,x)\| \leq \alpha(t)\|x\| + \beta(t) \qquad \forall (t,x) \in I \times \mathbb{R}^n \,,$$

wobei $\alpha, \beta \in C(I)$. Dann besitzt die AWA (2.1)-(2.2) eine globale Lösung $u \in C^1(I, \mathbb{R}^n)$.

Lineare Anfangswertaufgaben erfüllen diese Wachstumsbeschränkung. Dies werden wir im übernächsten Abschnitt behandeln.

2.2 Eindeutigkeit von Lösungen bei Anfangswertaufgaben

Die Eindeutigkeit ist i.a. aber nicht gegeben. Hierzu betrachten wir das Beispiel

$$\begin{aligned} u'(t) &= (u(t))^{1/3} \,, \qquad t \geq 0 \,, \\ u(0) &= 0 \,. \end{aligned}$$

Hier ist $u \equiv 0$ eine globale Lösung. Aber es gibt noch unendlich viele andere Lösungen, nämlich

$$u(t) = \begin{cases} 0, & t \leq c, \\ (\frac{2}{3}(t-c)^{3/2}, & t \geq c. \end{cases}$$

Hierbei handelt es sich um eine nichtlineare Differentialgleichung. Existenz und Eindeutigkeit erhält man allerdings, wenn man eine Lipschitz-Bedingung an f voraussetzt.

Definition 2.3 *$f \in C(D, \mathbb{R}^n)$ heißt lipschitz-stetig bzgl. x, wenn ein $L \in \mathbb{R}$ existiert, so dass*

$$\|f(t,x) - f(t,y)\| \leq L\|x - y\| \qquad \forall (t,x), (t,y) \in D \,.$$

Eine etwas schwächere Bedingung ist eine sogenannte Lipschitz-Bedingung:

Definition 2.4 *$f \in C(D, \mathbb{R}^n)$ erfüllt eine Lipschitz-Bedingung bzgl. x, wenn ein $L \in C(I)$ existiert, so dass*

$$\|f(t,x) - f(t,y)\| \leq L(t)\|x - y\| \qquad \forall (t,x), (t,y) \in D \,.$$

Wir bezeichnen den Raum dieser Funktionen mit $Lip(D, \mathbb{R}^n)$.

Es ist offensichtlich, dass die Lipschitz-Stetigkeit eine stärkere Eigenschaft ist als eine Lipschitz-Bedingung. Im Fall einer unbeschränkten Menge Ω, kann man die Lipschitz-Bedingung sogar noch schwächer formulieren (lokale Lipschitz-Bedingung), indem man die Funktion $L(t)$ abhängig macht von jeweils einer kompakten Menge $K \subset \Omega$, aus der x und y zu wählen sind (anstelle von ganz Ω).

Satz 2.5 (Picard-Lindelöf) *Es gelte $f \in Lip(Z, \mathbb{R}^n)$ auf dem Zylinder $Z = \overline{I} \times \overline{B}_r(u_0)$. Dann besitzt (2.1)-(2.2) auf $J := I \cap [t_0 - \epsilon, t_0 + \epsilon]$ mit $\epsilon := r/\|f\|_{L^\infty(Z)}$ eine eindeutige Lösung $u \in C^1(J, \overline{B}_r(u_0))$.*

Der Beweis dieses Satzes kann geführt werden mittels einer integralen Darstellung der Anfangswertaufgabe. Diese wird uns auch bei der Konstruktion von numerischen Schemata helfen. Integration von (2.1) liefert

$$u(t) \quad = \quad u_0 + \int_{t_0}^{t} f(s, u(s))\, ds\,.$$

Stetig differenzierbare Lösungen u dieser Integralgleichung sind offensichtlich auch Lösungen von (2.1). Durch die rechte Seite dieser Gleichung wird nun ein Operator definiert, der aufgrund des Banachschen Fixpunktsatzes einen Fixpunkt u besitzt. Dies ist die gesuchte Lösung. Da der Fixpunktsatz auch Eindeutigkeit liefert, erhält man auch die Eindeutigkeit der Lösung der Differentialgleichung.

Der Satz von Picard-Lindelöf liefert i.a. nur lokale Lösungen. Allerdings ist die Situation im Fall einer linearen AWA vom Typ (2.4) mit einer konstanten Matrix A weniger kritisch, denn hier kann man r beliebig groß wählen, ohne dass dies einen Einfluß auf $\|f\|_{L^\infty(D)} \leq c\|A\|$ hat. Man kann somit auch ϵ beliebig groß wählen, so dass wir den Beweis von Satz 2.7 im Fall von konstantem A erhalten.

Hierdurch decken wir auch lineare Anfangswertaufgaben mit $A \in C(I, \mathbb{R}^{n\times n})$ ab.

Korollar 2.6 *Im Fall $f \in Lip(I \times \mathbb{R}^n, \mathbb{R}^n)$ mit $\|f\|_{L^\infty(I\times\mathbb{R}^n)} < \infty$ existiert eine globale Lösung $u \in C^1(I, \mathbb{R}^n)$ von (2.1)-(2.2) und diese ist eindeutig.*

Beweis. Die Behauptung ergibt sich aus Satz 2.2 mit $\alpha \equiv 0$ und $\beta \equiv \|f\|_{L^\infty(I\times\mathbb{R}^n)}$. □

Diese Beschränktheit von f ist selbstverständlich eine sehr starke Voraussetzung. Nicht einmal die Identität $f(x) = x$ oder andere lineare Funktionen $f(x) = Ax$ mit einer Matrix $A \neq 0$ erfüllen diese Eigenschaft.

2.3 Lineare Differentialgleichungen

Unter einer linearen Differentialgleichungen 1. Ordnung in einem offenen, halboffenen oder abgeschlossenen Intervall $I \subset \mathbb{R}$ versteht man Gleichungen der Gestalt

$$u'(t) \quad = \quad A(t)u(t) + b(t) \quad t \in I\,, \tag{2.3}$$

mit einer stetigen Matrixfunktion $A \in C(I, \mathbb{R}^{n\times n})$ und stetigen Daten $b \in C(I, \mathbb{R}^n)$. Im Fall $b \neq 0$ ist diese Gleichung *inhomogen*, während man im Fall $b \equiv 0$, also

$$u'(t) \quad = \quad A(t)u(t) \quad t \in I\,, \tag{2.4}$$

von einer *homogenen* linearen Differentialgleichung spricht. Wir wollen kurz ein paar Lösungsprinzipien für solche Gleichungen wiederholen. Zunächst wissen wir, dass sowohl das homogene wie auch das inhomogene System zusammen mit Anfangswerten an einem Punkt $t_0 \in I$

$$u(t_0) \;=\; u_0 \tag{2.5}$$

stets eindeutig lösbar ist:

Satz 2.7 *Die Anfangswertaufgabe (2.3), (2.5) besitzt für beliebige $A \in C(I, \mathbb{R}^{n\times n})$, $b \in C(I, \mathbb{R}^n)$ lokal stets eine eindeutige Lösung $u \in C^1(J, \mathbb{R}^n)$ mit $J = I \cap [t_0 - \epsilon, t_0 + \epsilon]$ und einem $\epsilon > 0$.*

Bemerkung: Die Lösung existiert sogar global. Für den Beweis benötigen wir aber das Lemma von Gronwall, welches erst später bewiesen wird.

Beweis. Die zugehörige Funktion f lautet

$$f(t,x) \;=\; A(t)x + b(t)\,.$$

Diese erfüllt eine Lipschitz-Bedingung, denn

$$\|f(t,x) - f(t,y)\| \;=\; \|A(t)(x-y)\| \;\le\; \|A(t)\|_F \|x-y\|\,,$$

mit der Frobenius-Norm $\|\cdot\|_F$ von $A(t)$. Da Normen stetig sind und A stetig ist, folgt $f \in Lip(I \times \mathbb{R}^n, \mathbb{R}^n)$. Der Satz 2.5 von Picard-Lindelöf liefert daher die Existenz und Eindeutigkeit einer lokalen Lösung. ∎

Definition 2.8 *Unter einer* Fundamentalmatrix $X(t) \in \mathbb{R}^{n\times n}$ *versteht man eine Matrixfunktion $X \in C^1(I, \mathbb{R}^{n\times n})$, die die Differentialgleichung*

$$X'(t) \;=\; A(t)X(t) \quad \forall t \in I$$

erfüllt und an einer Stelle $t_0 \in I$ regulär ist, also $\det X(t_0) \neq 0$.

Im Fall, dass X zum Zeitpunkt t_0 die Identität st, also $X(t_0) = I$, redet man von der *Hauptfundamentalmatrix X*.

Satz 2.9 *Fundamentalmatrizen sind für beliebige $t \in I$ regulär, also*

$$X \in C^1(I, Gl(n, \mathbb{R}))\,.$$

Beweis. Die Regularität für alle $t \in I$ folgt aus Satz 2.7: Seien $u_i(t)$, $i \in \{1, \ldots, n\}$, die Spaltenvektoren (-funktionen) von $X(t)$. Diese erfüllen (2.4). Angenommen es gelte $\det X(t^*) = 0$. Dann existieren Koeffizienten a_i, so dass für die Funktion

$$w \;:=\; \sum_{i=2}^{n} a_i u_i$$

oBdA gilt $u_1(t^*) = w(t^*)$. Dann sind aber sowohl w als auch u_1 Lösungen der AWA (2.4) mit Anfangswerten $u(t^*) = w(t^*)$. Aus der Eindeutigkeit der Lösung folgt nun $u_1(t) = w(t)$ für $t \geq t^*$, also $\det X(t) = 0$ für $t \geq t^*$. Mit dem gleichen Argument (rückwärts in der Zeit t) folgt die Singularität von $X(t)$ für $t \leq t^*$. □

Satz 2.10 *Unter den gleichen Voraussetzungen wie in Satz 2.9 existiert stets eine Fundamentalmatrix $X(t)$.*

Beweis. Wir wählen in (2.5) als u_0 sukzessive eine Basis des $\mathbb{R}^n$. Da (2.3), (2.5) stets eine Lösung besitzt, erhalten wir durch Zusammenführen dieser n Lösungen in Form von Spaltenvektoren die Matrixfunktion $X(t)$. Diese liefert uns $X(t)$, wobei $X(t_0)$ gerade aus der ursprünglich gewählten Basis besteht und somit regulär ist. □
Für die *Wronski-Determinante* $W \in C(I, \mathbb{R})$,

$$W(t) \quad := \quad \det X(t)$$

einer Fundamentalmatrix gilt also $W(t) \neq 0$ für alle $t \in I$.

Kennt man eine/die Fundamentalmatrix $X(t)$, so läßt sich hieraus eine beliebige Lösung u von (2.4) konstruieren, die zudem die Anfangsbedingung (2.5) erfüllt, denn

$$u(t) \quad = \quad X(t)c$$

mit $c = X(t_0)^{-1}u_0$ ist dann eine Lösung.

Für die inhomoge Gleichung (2.3) macht man den Ansatz

$$u(t) \quad = \quad X(t)c(t)\,.$$

Nun folgt

$$u'(t) \quad = \quad X'(t)c(t) + X(t)c'(t) \quad = \quad A(t)X(t)c(t) + X(t)c'(t) \quad = \quad A(t)u(t) + X(t)c'(t)\,.$$

Man muß also $c(t)$ so bestimmen, dass $X(t)c'(t) = b(t)$ gilt. Dies lautet umgeformt, wenn $c(t)$ stetig differenzierbar ist:

$$c(t) \quad = \quad c(t_0) + \int_{t_0}^{t} c'(s)\,ds \quad = \quad c(t_0) + \int_{t_0}^{t} X(s)^{-1}b(s)\,ds\,.$$

Um auch noch die Anfangsbedingungen zu erfüllen, wählt man $c(t_0) = X(t_0)^{-1}u_0$. Insgesamt erhalten wir somit folgenden Satz:

Satz 2.11 *Bei gegebener Fundamentalmatrix $X(t)$ lautet die Lösung von (2.3) mit Anfangswert (2.5)*

$$u(t) \quad = \quad X(t)\left(X(t_0)^{-1}u_0 + \int_{t_0}^{t} X(s)^{-1}b(s)\,ds\right)\,.$$

2.3.1 Autonome lineare Systeme

Nun wollen wir eine solche Fundamentalmatrix $X(t)$ für *autonome* lineare Differentialgleichungen finden, d.h. die Matrix A in (2.3) hängt nicht mehr von t ab, sondern ist konstant. In diesem Fall lautet die Fundamentalmatrix wie im folgenden Satz gegeben:

Satz 2.12 *Die Hauptfundamentalmatrix zur homogenen linearen Differentialgleichung (2.4) für konstantes A ist für $t_0 = 0$ gegeben durch*

$$X(t) \;=\; \exp(tA) \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \,. \tag{2.6}$$

Beweis. Es gilt offensichtlich $X(0) = \exp(0A) = I$. Für jede submultiplikative Matrizennorm $\|\cdot\|$ gilt

$$\|X(t)\| \;\leq\; \sum_{k=0}^{\infty} \frac{1}{k!} \|tA\|^k \;\leq\; \exp(\|tA\|) \;=\; \exp(|t|\|A\|) \,.$$

Damit konvergiert die auftretende unendliche Reihe auf jedem kompakten Intervall $J \subset I$ gleichmäßig. Folglich dürfen wir differenzieren:

$$X'(t) \;=\; \frac{\partial}{\partial t} \exp(tA) \;=\; A \exp(tA) \;=\; AX(t) \,.$$

□

Allerdings ist $X(t)$ durch (2.6) noch nicht ohne weiteres berechenbar. Wir schränken uns weiter ein und setzen zusätzlich voraus, dass A symmetrisch ist, d.h. es existiert eine reguläre Matrix $S \in \mathbb{R}^{n\times n}$, so dass $D = S^{-1}AS \in \mathbb{R}^{n\times n}$ eine Diagonalmatrix ist. Wegen $A^k = (SDS^{-1})^k = SD^kS^{-1}$, lautet in diesem Fall die Hauptfundamentalmatrix

$$X(t) \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!} SD^kS^{-1} \;=\; S \left(\sum_{k=0}^{\infty} \frac{t^k}{k!} D^k \right) S^{-1} \;=\; S \exp(tD) S^{-1} \,.$$

Die Matrix $\exp(tD)$ berechnet sich hierbei einfach, indem man elementweise die Exponentialfunktion anwendet, also wenn λ_i die Diagonalelemente von D bezeichnet,

$$\exp(tD) \;=\; \operatorname{diag}\left(\exp(t\lambda_1), \ldots, \exp(t\lambda_n)\right) .$$

2.4 Stabilität von Anfangswertaufgaben

Wir betrachten nun die AWA (2.1) mit zwei unterschiedlichen Anfangswerten und fragen uns nach der zeitlichen Entwicklung. Dies ist beispielsweise relevant bei Störungen in den Anfangswerten. Sei also $u \in C^1(I, \Omega)$ Lösung von (2.1) mit den Anfangswerten

$$u(t_0) \;=\; u_0 \,, \tag{2.7}$$

und $v \in C^1(I, \Omega)$ Lösung von (2.1) mit den Anfangswerten

$$v(t_0) \quad = \quad v_0 \,. \tag{2.8}$$

Für den Fehler $e(t) := \|u(t) - v(t)\|$ gilt im Fall einer lipschitz-stetigen Funktion f:

$$\begin{aligned} e(t) &= e(0) + \int_{t_0}^{t} \|f(t, u(s)) - f(t, v(s))\| \, ds \\ &\leq e(0) + L \int_{t_0}^{t} e(s) \, ds \,. \end{aligned}$$

Hieraus läßt sich noch nicht direkt eine nutzbare Grenze für e ableiten, da auf der rechten Seite ebenfalls der Fehler e erscheint. Daher benötigen wir das folgende zentrale Lemma, das von extremer Wichtigkeit für die Analyse von Differentialgleichungen ist:

Lemma 2.13 (Grönwall[1]) *Seien $I = [a, b]$, $\alpha \in \mathbb{R}$, $\beta \geq 0$ und $u \in C(I)$ mit der oberen Schranke*

$$u(t) \quad \leq \quad \alpha + \beta \int_{a}^{t} u(s) \, ds \qquad \forall t \in I \,,$$

Dann gilt

$$u(t) \quad \leq \quad \alpha \, e^{(t-a)\beta} \qquad \forall t \in I \,.$$

Beweis. Wir betrachten die Funktion $\varphi \in C^1(I)$ definiert durch

$$\varphi(t) \quad := \quad e^{(a-t)\beta} \beta \int_{a}^{t} u(s) \, ds \,.$$

Da $\varphi(a) = 0$ und φ stetig differenzierbar ist, gilt nach Voraussetzung an u:

$$\begin{aligned} e^{(a-t)\beta} u(t) &\leq \alpha e^{(a-t)\beta} + e^{(a-t)\beta} \beta \int_{a}^{t} u(s) \, ds \\ &= \alpha e^{(a-t)\beta} + \varphi(t) - \varphi(a) \\ &= \alpha e^{(a-t)\beta} + \int_{a}^{t} \varphi'(s) \, ds \,. \end{aligned}$$

Den hierbei auftretende Integranden berechnen wir mit der Produktregel der Differentialrechnung und schätzen ihn nach oben ab:

$$\begin{aligned} \varphi'(t) &= e^{(a-t)\beta} \beta u(t) - \beta \varphi(t) \\ &= e^{(a-t)\beta} \beta \left(u(t) - \beta \int_{a}^{t} u(s) \, ds \right) \\ &\leq e^{(a-t)\beta} \beta \alpha \,. \end{aligned}$$

Wir setzten dies ein und berechnen das Integral per Substitution ($s \to \xi := (a-s)\beta$):

$$\begin{aligned} e^{(a-t)\beta}u(t) &\leq \alpha e^{(a-t)\beta} + \alpha\beta \int_a^t e^{(a-s)\beta}\, ds \\ &= \alpha e^{(a-t)\beta} - \alpha \int_0^{(a-t)\beta} e^{\xi}\, d\xi \\ &= \alpha e^{(a-t)\beta} - \alpha \left(e^{(a-t)\beta} - 1 \right) \\ &= \alpha\,. \end{aligned}$$

Multiplikation beider Seiten mit $e^{(t-a)\beta}$ ergibt die Behauptung. ∎

In den folgenden Lemmata sei wieder $D = I \times \Omega$. Eine obere Schranke für $\|u - v\|$ liefert uns das folgende Lemma:

Lemma 2.14 *$f \in C(D, \mathbb{R}^n)$ sei lipschitz-stetig mit Lipschitz-Konstante L. Dann gilt für die Lösungen u, v von (2.1) mit den Anfangsdaten (2.7) bzw. (2.8):*

$$\|(u-v)(t)\| \leq \|u_0 - v_0\| e^{L(t-t_0)} \qquad \forall t \in I\,.$$

Beweis. Wir wollen auf die Fehlerfunktion $e(t) := \|(u-v)(t)\|$ das Lemma 2.13 von Grönwall anwenden. Hierzu verwenden wir die Integraldarstellung für den Fehler,

$$u(t) - v(t) = u(t_0) - v(t_0) + \int_{t_0}^t (f(s,u(s)) - f(s,v(s)))\, ds\,.$$

Also folgt mit der Dreiecksungleichung und der Lipschitz-Eigenschaft von f für beliebiges $t \in I$:

$$\begin{aligned} e(t) &\leq e(t_0) + \int_{t_0}^t |f(s,u(s)) - f(s,v(s))|\, ds \\ &\leq e(t_0) + L \int_{t_0}^t e(s)\, ds\,. \end{aligned}$$

Das Lemma 2.13 liefert nun die Schranke:

$$e(t) \leq e(t_0)\, e^{L(t-t_0)}\,.$$

Dies ist gerade die Behauptung. ∎

Alternativ kann man die Situation betrachten, dass anstelle der Anfangswerte die Funktion f etwas geändert/gestört ist. Sei dazu $w \in C^1(I, \Omega)$ Lösung von

$$w'(t) = g(t, w(t)) \qquad t \in I\,, \tag{2.9}$$

mit den Anfangswerten

$$w(a) = u_0\,. \tag{2.10}$$

Lemma 2.15 *Seien $f, g \in C(D, \mathbb{R}^n)$ und f sei lipschitz-stetig mit Lipschitz-Konstante L. Dann gilt für die Lösungen u von (2.1),(2.7) und w von (2.9),(2.10):*

$$\|(u-w)(t)\| \leq \frac{\epsilon}{L}\left(e^{L(t-a)}-1\right) \qquad \forall t \in I\,,$$

wobei $\epsilon := \|f-g\|_{L^\infty[a,t]} = \max\{\|f(s,\cdot)-g(s,\cdot)\|_\Omega : a \leq s \leq t\}$.

Beweis. Wir betrachten wieder die Fehlerfunktion $e(t) := \|(u-w)(t)\|$ und wenden das Lemma von Grönwall an. Die Integraldarstellung liefert nun für $t \in I$:

$$\begin{aligned} e(t) &= \left\| \int_a^t (f(s,u(s)) - g(s,w(s)))\, ds \right\| \\ &\leq \int_a^t (\|f(s,u(s)) - f(s,w(s))\| + \epsilon)\, ds \\ &\leq \int_a^t (Le(s) + \epsilon)\, ds\,. \end{aligned}$$

Wir betrachten ferner $\widehat{e}(t) := e(t) + \epsilon/L$. Für diese Größe ergibt sich mit dieser Abschätzung:

$$\widehat{e}(t) \leq \frac{\epsilon}{L} + L\int_a^t \widehat{e}(s)\, ds\,.$$

Die Anwendung des Lemmas von Grönwall auf $\widehat{e}$ liefert:

$$\widehat{e}(t) \leq \frac{\epsilon}{L} e^{L(t-a)}\,.$$

Nun führen wir dieses Ergebnis zurück auf e und erhalten die Behauptung:

$$e(t) \leq \frac{\epsilon}{L} e^{L(t-a)} - \frac{\epsilon}{L} = \frac{\epsilon}{L}\left(e^{L(t-a)}-1\right)\,.$$

□

Nun kombinieren wir diese beiden Ergebnisse und betrachten die Kombination aus gestörten Anfangswerten und gestörter Funktion. Sei also $w \in C^1(I,\Omega)$ die Lösung von

$$w'(t) = g(t,w(t)) \qquad t \in I\,, \tag{2.11}$$

$$w(a) = w_0\,. \tag{2.12}$$

Eine obere Schranke für die zeitliche Entwicklung des Fehler fassen wir in folgendem Satz zusammen.

Satz 2.16 *Seien $f, g \in C(D, \mathbb{R}^n)$ und f sei lipschitz-stetig mit Lipschitz-Konstante L. Dann gilt für die Lösungen u von (2.1),(2.7) und w von (2.11),(2.12):*

$$\|(u-w)(t)\| \leq c_1(t)\|u_0 - w_0\| + c_2(t)\|f-g\|_{L^\infty[a,t]} \qquad \forall t \in I\,,$$

mit $c_1(t) = e^{L(t-a)}$ *und* $c_2(t) = L^{-1}\left(e^{L(t-a)}-1\right)$.

Beweis. Dreiecksungleichung aus den beiden vorherigen Sätzen. ∎

Später benötigen wir die folgende erweiterte Version des Lemmas von Grönwall:

Lemma 2.17 (Grönwall) *Sei $I = [a,b]$, $u : I \to \mathbb{R}$ stückweise stetig mit der oberen Schranke*

$$u(t) \quad \leq \quad \alpha(t) + \int_a^t \beta(s)u(s)\,ds \qquad \forall t \in I\,,$$

wobei $\alpha : I \to \mathbb{R}$ nicht-fallend und $\beta : I \to \mathbb{R}_{\geq 0}$ stückweise stetig. Dann gilt

$$u(t) \quad \leq \quad \alpha(t)\,\exp\left(\int_a^t \beta(s)\,ds\right) \qquad \forall t \in I\,.$$

Beweis. Die Funktion

$$\varphi(t) \quad := \quad \int_a^t \beta(s)u(s)\,ds$$

ist differenzierbar (aber nicht unbedingt in $C^1(I)$), denn $\varphi'(t) = \beta(t)u(t)$. Ferner sei

$$\psi(t) \quad := \quad u(t) - \varphi(t)\,.$$

Es gilt dann $\varphi'(t) = \beta(t)(\varphi(t) + \psi(t))$. Also ist φ Lösung der linearen AWA

$$\begin{aligned} \varphi'(t) &= \beta(t)\varphi(t) + \beta(t)\psi(t) \qquad t \in I\,, \\ \varphi(a) &= 0\,. \end{aligned}$$

Diese AWA ist zwar linear, hat aber nur stückweise stetige Koeffizienten. Durch stückweise Betrachtung läßt sich zeigen, dass auch hier eine eindeutige Lösung existiert (Übungsaufgabe). Die Lösung erhalten wir über die Fundamentalmatrix $X(t)$, die in diesem Fall eine skalare Funktion ist:

$$X(t) \quad = \quad \exp\left(w(t)\right) \quad \text{mit} \quad w(t) := \int_a^t \beta(s)\,ds\,.$$

Nach Satz 2.11 lautet die Lösung dieser AWA

$$\varphi(t) \quad = \quad X(t)\int_a^t X(s)^{-1}\beta(s)\psi(s)\,ds\,.$$

Wegen $\psi(s) \leq \alpha(s) \leq \alpha(t)$ folgern wir

$$\begin{aligned} \varphi(t) &\leq \alpha(t)\exp(w(t))\int_a^t \exp(-w(s))\beta(s)\,ds \\ &= \alpha(t)\exp(w(t))\int_a^t \frac{d}{ds}\left(\exp(-w(s))\right)\,ds \\ &= \alpha(t)\exp(w(t))\left[1 - \exp(-w(t))\right] \\ &\leq \alpha(t)\exp(w(t)) - \alpha(t)\,. \end{aligned}$$

Mit der Definition von φ und der Voraussetzung für u folgt nun

$$u(t) \;\leq\; \alpha(t) + \varphi(t) \;=\; \alpha(t) \exp\left\{ \int_a^t \beta(s)\, ds \right\} .$$

□

Die einfache Version (Lemma 2.13) ergibt sich aus Lemma 2.17 indem wir die konstanten Funktionen $\alpha(t) \equiv \alpha$ und $\beta(t) \equiv \beta$ wählen:

$$u(t) \;\leq\; \alpha \exp\left(\int_a^t \beta\, ds \right) \;=\; \alpha \exp((t-a)\beta) \qquad \forall t \in I\,.$$

Für die Analyse diskreter Schemata benötigen wir eine diskrete Version des Lemmas von Grönwall:

Lemma 2.18 (Diskretes Grönwall'sches Lemma) *Seien $(u_n)_{n\in\mathbb{N}_0}, (a_n)_{n\in\mathbb{N}_0}, (b_n)_{n\in\mathbb{N}_0}$ Folgen in $\mathbb{R}_{\geq 0}$, $(a_n)_{n\in\mathbb{N}_0}$ sei nicht-fallend und es gelte*

$$u_0 \;\leq\; a_0 \quad \textit{und} \quad u_n \;\leq\; a_n + \sum_{i=0}^{n-1} b_i u_i \quad \forall n \in \mathbb{N}\,.$$

Dann folgt

$$u_n \;\leq\; a_n \exp\left(\sum_{i=0}^{n-1} b_i \right) .$$

Dieses Lemma läßt sich einfach per vollständige Induktion beweisen. Wir verwenden alternativ hierzu die kontinuierliche Version von Gronwall.

Beweis. Dieses Behauptung läßt sich auch in integraler Form formulieren. Hierzu setzen wir $I = [0, \infty)$ und stückweise stetige Funktionen $\alpha, \beta, u : I \to \mathbb{R}$:

$$\alpha(t) \;:=\; a_i\,, \quad \beta(t) \;:=\; b_i\,, \quad u(t) \;:=\; u_i\,,$$

mit $i = \mathrm{int}(t)$. α ist nicht-fallend. Nach Voraussetzung gilt nun

$$u(t) \;\leq\; \alpha(t) + \int_0^t \beta(s) u(s)\, ds \qquad \forall t \in I\,.$$

Nach dem erweiterten Grönwall'schen Lemma 2.17 folgt nun die Behauptung:

$$u_n \;=\; u(n) \;\leq\; \alpha(n) \exp\left(\int_0^n \beta(s)\, ds \right) \;=\; a_n \exp\left(\sum_{i=0}^{n-1} b_i \right) .$$

□

Gelegentlich wird auch folgende Variante benutzt:

Lemma 2.19 (Diskretes Grönwall'sches Lemma) *Sei $(u_n)_{n\in\mathbb{N}_0}$ eine Folge in $\mathbb{R}_{\geq 0}$, mit $u_n \leq a + bu_{n-1}$ für alle $n \in \mathbb{N}$, wobei $a, b \geq 0$. Dann gelten folgende Implikationen:*

$$u_n \quad \leq \quad \begin{cases} u_0 + na & \text{falls } b = 1 \\ e^{n(b-1)}u_0 + \frac{e^{n(b-1)}-1}{b-1}a & \text{falls } b > 1, \\ b^n u_0 + \frac{b^n-1}{b-1}a & \text{falls } 0 < b < 1\,. \end{cases}$$

Kapitel 3

Einschrittverfahren

Wir wählen im Intervall $I = [a, b]$ diskrete Zeitpunkte, $a = t_0 < t_1 < \ldots < t_N = b$ und setzen als Teilintervalle $I_k = [t_{k-1}, t_k)$ mit jeweiligen Längen $h_k = t_k - t_{k-1}$. Die maximale Länge bezeichnen wir mit $h = \max_{k=1,\ldots,N} h_k$. Die integrale Formulierung lautet auf solchen Teilintervallen:

$$u(t_n) \quad = \quad u(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(s, u(s))\, ds\,.$$

In Anlehnung an diese Darstellung versteht man unter einem Einschrittverfahren die rekursive Berechnung von Approximationen u_n^h von $u(t_n)$ in der Form

$$u_n^h \quad = \quad u_{n-1}^h + h_n \Phi(h_n, t_{n-1}, u_{n-1}^h, u_n^h)\,. \tag{3.1}$$

Das Φ sollte also eine Approximation an den integralen Mittelwert von f sein. Charakteristisch ist hierbei, dass nur die Werte u_{n-1}^h und u_n^h (evtl. sogar nicht einmal beide) in die Berechnung eingehen und nicht etwa die diskrete Lösung an weiter zurückliegenden Zeitpunkten t_k mit $0 \le k < n-1$. Wenn die Verfahrensfunktion $\Phi(h, t, u, v)$ unabhängig von v ist, so spricht man von einem *expliziten* Verfahren, anderenfalls von einem *impliziten* Verfahren. Bei impliziten Verfahren müssen i.d.R. nichtlineare Gleichungssysteme gelöst werden (sofern f nichtlinear ist). Wir beginnen nun mit dem einfachsten expliziten Verfahren.

3.1 Expliziter Euler

Das explizite Euler-Verfahren, oder auch Euler'sches Polygonzug-Verfahren genannt, erzeugt ausgehend von dem Startwert $u_0^h = u_0$ eine Folge $(u_n^h)_{n\in\mathbb{N}_0}$ durch die rekursive Vorschrift

$$u_n^h \quad = \quad u_{n-1}^h + h_n f(t_{n-1}, u_{n-1}^h)\,, \qquad n \ge 0\,. \tag{3.2}$$

Wir erhalten also ein explizites Einschrittverfahren mit $\Phi(h,t,x,y) = f(t,x)$. Dieses Verfahren ist motiviert durch die Integraldarstellung der AWA und einer Aproximation durch eine einfachen Quadraturformel

$$\begin{aligned} u(t_n) &= u(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(s,u(s))\,ds \\ &\approx u(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(t_{n-1},u(t_{n-1}))\,ds \\ &= u(t_{n-1}) + h_n f(t_{n-1},u(t_{n-1}))\,. \end{aligned}$$

Dieses einfache Verfahren (3.2) wird üblicherweise dazu benutzt, um den eingangs erwähnten Satz 2.1 von Peano zu beweisen. Wie wir anschließend sehen werden ist das explizite Euler-Verfahren sehr ungenau (von 1. Ordnung in der Gitterweite) und ist daher praktisch kaum relevant. Dennoch ist es an dieser Stelle interessant, weil die Fehleranalyse exemplarisch ist für eine Vielzahl anderer Verfahren. Um die Genauigkeit dieses Verfahrens zu beurteilen, schauen wir uns den sogenannten Abschneidefehler an.

3.1.1 Abschneidefehler

Unter dem Abschneidefehler (eng. *truncation error*) versteht man die Größe, die entsteht, wenn man die exakte Lösung in die Differenzenformel (3.2) einsetzt:

$$\tau_n^h := \frac{u(t_n) - u(t_{n-1})}{h_n} - f(t_{n-1},u(t_{n-1}))\,.$$

Dieser Abschneidefehler ist also eine rein lokale Größe. Eine evtl. auftretende "Fehlerakkumulation" wird durch den Abschneidefehler nicht erfasst. Man nennt den Abschneidefehler daher auch gelegentlich "lokalen Diskretisierungsfehler".

Satz 3.1 *Für die Lösung u der AWA (2.1)-(2.2) gelte $u \in C^2(I,\Omega)$. Dann ist der Abschneidefehler des expliziten Euler-Verfahrens von erster Ordnung (im Hinblick auf die Gitterweite):*

$$\|\tau_n^h\| \leq \frac{1}{2} h_n \|u''\|_{L^\infty(I_n,\Omega)}\,.$$

Beweis. Da u die AWA erfüllt, gilt

$$\begin{aligned} \|\tau_n^h\| &= h_n^{-1} \int_{t_{n-1}}^{t_n} u'(s)\,ds - u'(t_{n-1}) \\ &= h_n^{-1} \int_{t_{n-1}}^{t_n} \big(u'(s) - u'(t_{n-1})\big)\,ds \\ &\leq h_n^{-1} \max_{s\in I_n} \|u''(s)\| \int_{t_{n-1}}^{t_n} (s - t_{n-1})\,ds \\ &= h_n^{-1} \max_{s\in I_n} \|u''(s)\| \frac{1}{2} h_n^2\,. \end{aligned}$$

Dies ist genau die Behauptung. ∎

3.1.2 Globaler Fehler

In diesem Abschnitt betrachten wir den Fehler zwischen der exakten Lösung und der diskreten Approximation an den diskreten Zeitpunkten. Wir setzen: Bezeichnung ein:

$$\varepsilon_n \quad := \quad \|u(t_n) - u_n^h\| .$$

Satz 3.2 *Für die Lösung u der AWA (2.1)-(2.2) gelte $u \in C^2(I, \Omega)$. Ferner sei f lipschitzstetig mit Lipschitz-Konstante L. Dann gilt für den Fehler des expliziten Euler-Verfahrens:*

$$\varepsilon_n \quad \leq \quad \exp\{L(t_n - t_0)\} \left(\varepsilon_0 + \frac{t_n - t_0}{2} \max_{1 \leq i \leq n} (h_i \|u''\|_{L^\infty(I_i,\Omega)}) \right) \quad \forall n \in \{0, \ldots, N\} .$$

Beweis. Umformung der Definition des Abschneidefehlers liefert die Differenzengleichung für die exakte Lösung:

$$u(t_n) \quad = \quad u(t_{n-1}) + h_n f(t_{n-1}, u(t_{n-1})) + h_n \tau_n^h .$$

Hieraus ergibt sich für ε_n aufgrund der Lipschitz-Stetigkeit von f:

$$\begin{aligned} \varepsilon_n &\leq \varepsilon_{n-1} + h_n \|f(t_{n-1}, u(t_{n-1}) - f(t_{n-1}, u_{n-1}^h)\| + h_n \|\tau_n^h\| \\ &\leq \varepsilon_{n-1} + h_n L \|u(t_{n-1}) - u_{n-1}^h\| + h_n \|\tau_n^h\| . \end{aligned}$$

Per Induktion ergibt sich nun

$$\varepsilon_n \quad \leq \quad \varepsilon_0 + \sum_{i=0}^{n-1} h_{i+1} (L\varepsilon_i + \|\tau_{i+1}^h\|) \quad \leq \quad a_n + \sum_{i=1}^{n-1} b_i \varepsilon_i ,$$

mit

$$a_n \quad := \quad \varepsilon_0 + \sum_{i=1}^{n} h_i \|\tau_i^h\| \quad \text{und} \quad b_n \quad := \quad L h_{n+1} .$$

Nun verwenden wir die diskrete Version des Lemmas von Grönwall (Lemma 2.18) und der Beobachtung $\sum_{i=0}^{n-1} b_i = L(t_n - t_0)$:

$$\begin{aligned} \varepsilon_n &\leq a_n \exp\left(L(t_n - t_0)\right) \\ &= \exp\left\{L(t_n - t_0)\right\} \left(\varepsilon_0 + \sum_{i=1}^{n} h_i \|\tau_i^h\| \right) . \end{aligned}$$

Die Behauptung ergibt sich nun aus der oberen Schranke für den Abschneidefehler (Satz 3.1)

$$\begin{aligned} \sum_{i=1}^{n} h_i \|\tau_i^h\| &\leq (t_n - t_0) \max_{1 \leq i \leq n} \|\tau_i^h\| \\ &\leq \frac{1}{2} (t_n - t_0) \max_{1 \leq i \leq n} \{h_i \|u''\|_{L^\infty(I_i,\Omega)}\} . \end{aligned}$$

□

3.2 Konsistenz von Einschrittverfahren

Der Abschneidefehler eines allgemein Einschrittverfahrens ist folgendermaßen definiert:

$$\tau_n^h := \frac{u(t_n) - u(t_{n-1})}{h_n} - \Phi(h_n, t_{n-1}, u(t_{n-1}), u(t_n)) . \tag{3.3}$$

Definition 3.3 *Ein Einschrittverfahren (3.1) heißt* konsistent (mit Konsistenzordnung m)*, wenn für den Abschneidefehler bei hinreichend regulärer Lösung u für $h \to 0$ gilt:*

$$\max_{1\le i\le n} \|\tau_n^h\| = \mathcal{O}(h^m) .$$

Mit anderen Worten: Es muß eine Konstante $C \in \mathbb{R}$ und ein $h_0 > 0$ existieren, so dass

$$\max_{1\le i\le n} \|\tau_n^h\| \le Ch^m \qquad \forall h \in (0, h_0) .$$

Die Konstante darf vom Intervall I und von f aber nicht von h abhängen.

Um Verfahrensfunktionen Φ mit einer möglichst hohen Konsistenzordnung zu erstellen, entwickeln wir $u(t_{n+1})$ mit einer Taylorentwicklung. Wir nehmen hier den skalaren Fall an, $\Omega \subset \mathbb{R}$, sowie $u \in C^{m+1}(I)$:

$$u(t_n) = u(t_{n-1}) + \sum_{k=1}^{m} \frac{1}{k!} u^{(k)}(t_{n-1}) h_n^k + \frac{1}{(m+1)!} u^{(m+1)}(\xi) h_n^{m+1} ,$$

mit $\xi \in I_n$. Setzen wir dies ein in die Definition (3.3) des Abschneidefehlers, so erhalten wir

$$\tau_n^h = \sum_{k=1}^{m} \frac{1}{k!} u^{(k)}(t_{n-1}) h_n^{k-1} - \Phi(h_n, t_{n-1}, u(t_{n-1}), u(t_n)) + \mathcal{O}(h_n^m)$$

Eine Konsistenzordnung von m erfordert daher

$$\begin{aligned} \Phi(h_n, t_{n-1}, u(t_{n-1}), u(t_n)) &= \sum_{k=1}^{m} \frac{1}{k!} u^{(k)}(t_{n-1}) h_n^{k-1} + \mathcal{O}(h_n^m) \\ &= \sum_{k=0}^{m-1} \frac{1}{(k+1)!} u^{(k+1)}(t_{n-1}) h_n^k + \mathcal{O}(h_n^m) . \end{aligned} \tag{3.4}$$

Andererseits liefert die Taylor-Entwicklung von Φ im Fall einer m-mal stetigen Differenzierbarkeit bezüglich des ersten Arguments (also bzgl. h) die Darstellung:

$$\Phi(h_n, t_{n-1}, u(t_{n-1}), u(t_n)) = \sum_{k=0}^{m-1} \frac{1}{k!} h_n^k \frac{\partial^k}{\partial h^k} \Phi(0, t, u(t), u(t+h)) + \mathcal{O}(h_n^m) .$$

Kombinieren wir diese beiden Potenzreihenentwicklungen, so erhalten wir die Bedingung

$$\sum_{k=0}^{m-1} \frac{1}{(k+1)!} u^{(k+1)}(t_{n-1}) h_n^k \quad = \quad \sum_{k=0}^{m-1} \frac{1}{k!} h_n^k \phi_k + \mathcal{O}(h_n^m)\,,$$

mit der Bezeichnung $\phi_k := \frac{\partial^k}{\partial h^k}\Phi(0,t,u(t),u(t+h))$. Vergleichen wir die einzelnen Terme in der Summe, erhalten wir die Bedingung für eine Konsistenzordnung von m:

$$(k+1)\phi_k \quad = \quad u^{(k+1)}(t_{n-1}) + \mathcal{O}(h_n^m)\,.$$

Wir fassen dieses Ergebnis im folgenden Satz zusammen:

Satz 3.4 (Konsistenzkriterium) *Ein Einschrittverfahren $\Phi(h,t,x,y)$, dass m-mal stetig differenzierbar bzgl. h ist, ist genau dann konsistent von der Ordnung $m \in \mathbb{N}$, wenn für jedes $u \in C^{m+1}(I)$ und alle $k \in \{0,\ldots,m-1\}$ gilt:*

$$(k+1)\frac{\partial^k}{\partial h^k}\Phi(0,t,u(t),u(t+h)) - u^{(k+1)}(t) \quad = \quad \mathcal{O}(h^{m-k}) \qquad \forall t \in I\,.$$

Im Fall eines expliziten Einschrittverfahrens $\Phi(h,t,x)$ ist dies äquivalent zu:

$$(k+1)\frac{\partial^k}{\partial h^k}\Phi(0,t,u(t)) \quad = \quad u^{(k+1)}(t) \qquad \forall t \in I\,.$$

Beweis. Der erste Teil ist oben gezeigt. Der zweite Teil (explizite ESV) folgt aus der Beobachtung, dass die linke Seite unabhängig ist von h. □

3.3 Taylor-Methoden

Die sogenannten *Taylor-Methoden* sind motiviert durch die Darstellung (3.4) indem man die Summe bis zur Stufe s laufen läßt. Ferner verwendet man die Eigenschaft der Lösung für $0 \le k \le m$:

$$u^{(k+1)}(t) \quad = \quad \frac{d^k}{dt^k} f(t,u(t))\,.$$

Hierbei handelt es sich um die totale Zeitableitung, so dass für $k \ge 2$ auch Ableitungen von u einfließen. Dadurch erhält man die expliziten Verfahrensfunktionen

$$\Phi_{TF(s)}(h,t,x) \quad = \quad \sum_{k=0}^{s-1} \frac{h^k}{(k+1)!} \frac{d^k}{dt^k} f(t,x)\,.$$

s bezeichnet dabei die *Stufe* des Verfahrens und entspricht gemäß der Konstruktion also auch der Konsistenzordnung. Da diese Methoden explizit sind und nicht von y abhängen, haben diese Verfahrensfunktionen nur 3 Argumente.

Satz 3.5 *Die Taylor-Methode s-ter Stufe besitzt die Konsistenzordnung $m = s$.*

Beweis. Wir benutzen das Konsistenzkriterium aus Satz 3.4. Es gilt für $0 \le j \le s-1$

$$\begin{aligned}
\frac{\partial^j}{\partial h^j}\Phi_{TF(s)}(h,t,u)|_{h=0} &= \frac{\partial^j}{\partial h^j}\left(\sum_{k=0}^{s-1}\frac{h^k}{(k+1)!}u^{(k)}\right)\Big|_{h=0} \\
&= \left(\sum_{k=j}^{s-1}\frac{k!}{(k-j)!}\frac{h^{k-j}}{(k+1)!}u^{(k)}\right)\Big|_{h=0} \\
&= \frac{1}{j+1}u^{(j)}\,.
\end{aligned}$$

Hierdurch sind die Konsistenzbedingungen bis zur Ordnung s nachgewiesen. □

Bei der einstufigen Taylor-Methode ($s = 1$) erhalten wir $\Phi_{TF(1)}(h,t,x) = f(t,x)$, also das explizite Euler-Verfahren. Für $s = 2$ ergibt sich

$$\begin{aligned}
\Phi_{TF(2)}(h,t,x) &= f(t,x) + \frac{h}{2}\frac{d}{dt}f(t,x) \\
&= f(t,x) + \frac{h}{2}\left(\frac{\partial}{\partial t}f(t,x) + \frac{\partial}{\partial x}f(t,x)x'(t)\right)\,.
\end{aligned}$$

Wenden wir dieses Φ an auf die Lösung u der AWA, so können wir auch schreiben:

$$\Phi_{TF(2)}(h,t,u) = f(t,u(t)) + \frac{h}{2}\left(\frac{\partial}{\partial t}f(t,u(t)) + \frac{\partial}{\partial x}f(t,u(t))f(t,u(t))\right)\,.$$

Ab $s \ge 2$ treten also Ableitungen von f in der rechten Seite von (3.1) für die Taylor-Methoden auf. Dies wird vermieden in den expliziten Runge-Kutta Methoden, die im nächsten Abschnitt behandelt werden.

3.4 Runge-Kutta Methoden

Die Idee der Runge-Kutta Methoden besteht darin, die Ableitungen von f in der Taylor-Entwicklung durch Differenzenquotienten zu ersetzen. Es stellt sich heraus, dass hierdurch die Konsistenzordnung im Fall von skalaren Gleichungen nicht verändert wird. Wenn f vektor-wertig ist bleibt die Konsistenzordnung zumindest für $s \le 4$ erhalten.

Für $s = 2$ führen wir zunächst die folgenden zwei Approximationen durch:

$$\begin{aligned}
\frac{d}{dt}f(t,u) &\approx \frac{1}{h}(f(t+h,u(t+h)) - f(t,u(t)) \\
&\approx \frac{1}{h}(f(t+h,u(t)+hf(t,u(t))) - f(t,u(t))\,.
\end{aligned}$$

Diese Approximation wird nun in dem expliziten Runge-Kutta-Verfahren der Stufe $s = 2$ verwendet:

$$\begin{aligned}
\Phi_{RK(2)}(h,t,x) &= f(t,x) + \frac{h}{2h}(f(t+h,x+hf(t,x)) - f(t,x)) \\
&= \frac{1}{2}\Big(f(t+h,x+hf(t,x)) + f(t,x)\Big)
\end{aligned}$$

Wir erhalten also folgende rekursive Berechnung für RK(2):

$$u_n^h \;=\; u_{n-1}^h + \frac{h_n}{2}\left[f(t_n, u_{n-1}^h + h_n y_{n-1}) + y_{n-1}\right] .$$

mit $y_{n-1} = f(t_{n-1}, u_{n-1}^h)$. Die allgemeine Form einer Runge-Kutta Methode der Stufe $s \in \mathbb{N}$ lautet:

$$\Phi_{RK(s)}(h,t,x,y) \;=\; \sum_{i=1}^{s} r_i k_i \,, \tag{3.5}$$

mit rekursiv definierten (von t, h, x und f-abhängigen) Koeffizienten

$$i \in \{1,\ldots,s\}: \qquad k_i \;:=\; f\left(t + a_i h, x + b_i h\right) , \quad b_i := \sum_{j=1}^{s} \gamma_{ij} k_j \,.$$

Die noch freien Parameter a_i, r_i und γ_{ij} werden so gewählt, dass $\Phi_{RK(m)}(h,t,x)$ eine möglichst gute Approximation an den Taylor-Term $\Phi_{TF(m)}(h,t,x)$ ist, also

$$\Phi_{RK(s)}(h,t,x,y) \;=\; \Phi_{TF(s)}(h,t,x) + \mathcal{O}(h^s)$$

ist. Diese Parameter können in sogenannten Butcher-Diagrammen dargestellt werden:

a_1	γ_{11}	γ_{12}	$\cdots$	γ_{1s}
$\vdots$	γ_{21}	γ_{22}	$\ddots$	$\vdots$
$\vdots$	$\vdots$		$\ddots$	$\vdots$
a_s	γ_{s1}	$\cdots$	$\cdots$	γ_{ss}
	r_1	$\cdots$	$\cdots$	r_s

An dieser Stelle sei nocheinmal explizit darauf hingewiesen, dass die Runge-Kutta Methoden nicht ohne weiteres auf Systeme übertragbar sind.

3.4.1 Explizite Runge-Kutta Methoden

Runge-Kutta Methoden, bei denen im Butcher-Diagramm gilt $\gamma_{ij} = 0$ für $i \leq j$ heißen explizite Runge-Kutta Methoden (ERK). Für diese gilt

$$\begin{aligned} k_1 \;&:=\; f(t,x)\,, \\ i \in \{2,\ldots,s\}: \qquad k_i \;&:=\; f\left(t + a_i h, x + b_i h\right) , \quad b_i := \sum_{j=1}^{i-1} \gamma_{ij} k_j \,. \end{aligned}$$

Ein paar Beispiele hierzu sind im Folgenden aufgeführt. Die zugehörigen Butcher-Diagramme sind in Tabelle 3.1 aufgeführt. Es sei hier noch angemerkt, dass es zu gegebenem m sehr wohl mehrere Runge-Kutta-Methoden geben kann. Hier ist jeweils nur eine prominente Möglichkeit ausgewählt.

0	0
	1

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

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

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

Tabelle 3.1: Butcher-Diagramme für expliziten RK-Methoden für $1 \leq s \leq 4$ (von links nach rechts).

- $s = 1$: Expliziter Euler:

$$\Phi_{RK(1)}(h,t,x) \;=\; f(t,x)$$

- $s = 2$: Heun-Verfahren:

$$\Phi_{RK(2)}(h,t,x) \;=\; \frac{1}{2}\big(f(t,x) + f(t+h, x+f(t,x))\big)$$

$$u_n^h \;=\; u_{n-1}^h + \frac{h_n}{2}\big(f(t_{n-1},u_{n-1}^h) + f(t_n, u_{n-1}^h + f(t_{n-1},u_{n-1}^h))\big)$$

- $s = 3$: Kutta-Verfahren 3. Ordnung:

$$\begin{aligned}
\Phi_{RK(3)}(h,t,x) &= \frac{1}{6}k_1 + \frac{2}{3}k_2 + \frac{1}{6}k_3\,,\\
k_1 &= f(t,x)\,,\\
k_2 &= f(t+h/2, x+k_1h/2)\,,\\
k_3 &= f(t+h, x-k_1h+2k_2h)\,.
\end{aligned}$$

- $s = 4$: Klassisches Runge-Kutta-Verfahren:

$$\begin{aligned}
\Phi_{RK(4)}(h,t,x) &= \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4\,,\\
k_1 &= f(t,x)\,,\\
k_2 &= f(t+h/2, x+k_1h/2)\,,\\
k_3 &= f(t+h/2, x+k_2h/2)\,,\\
k_4 &= f(t+h, x+k_3h)\,.
\end{aligned}$$

  Dieses Verfahren hat eine Ähnlichkeit mit der Simpson-Regel für die numerische Quadratur.

Das folgende Lemma macht eine Aussage darüber, was passiert, wenn man eine explizite Runge-Kutta Methode auf die lineare Differentialgleichung $u' = \lambda u$ anwendet.

Lemma 3.6 *Für eine explizite Runge-Kutta Methode der Stufe s für $f(t,x) = \lambda x$ mit $\lambda > 0$ gilt*

$$y \;=\; x + h\Phi_{RK(s)}(h,t,x) \;=\; p(\lambda h)x\,,$$

mit einem Polynom $p \in \mathbb{P}_s$ von echtem Polynomgrad s.

Beweis. Die explizite Runge-Kutta Methode lautet

$$y \;=\; x + h\Phi_{RK(s)}(h,t,x) \;=\; x + h\sum_{i=1}^{s} r_i k_i\,.$$

Es genügt daher zu zeigen, dass $\lambda^{-1}k_i = q_i(\lambda h)x$, mit $q_i \in \mathbb{P}_{i-1}$, denn dann folgt, dass

$$y \;=\; x + h\sum_{i=1}^{s} r_i \lambda q_i(\lambda h)x \;=\; \left(1 + \lambda h\sum_{i=1}^{s} r_i q_i(\lambda h)\right)x \;=\; p(\lambda h)x\,,$$

mit $p \in \mathbb{P}_s$. Wir führen den Beweis per Induktion nach i. Für $i = 1$ gilt $\lambda^{-1}k_1 = \lambda^{-1}f(t,x) = x$. Der Induktionsschritt lautet nun wegen $f(t,x) = \lambda x$:

$$\begin{aligned}
\lambda^{-1}k_{i+1} &= \lambda^{-1} f\left(\cdot, x + h\sum_{j=1}^{i}\gamma_{ij}k_j\right)\\
&= x + h\sum_{j=1}^{i}\gamma_{ij}k_j\\
&= \left(1 + h\lambda\sum_{j=1}^{i}\gamma_{ij}q_j(\lambda h)\right)x\\
&=: q_{i+1}(\lambda h)x\,,
\end{aligned}$$

mit $q_{i+1} \in \mathbb{P}_i$. Es gilt ferner $\deg q_{i+1} = 1 + \deg q_i = i$. ∎

Satz 3.7 *Eine s-stufige explizite Runge-Kutta Methode besitzt die maximale Konsistenzordnung $m = s$.*

Beweis. Wir betrachten folgende konkrete lineare AWA zu beliebigem $\lambda > 0$

$$\begin{aligned}
u'(t) &= \lambda u(t) \qquad t \geq 0\,,\\
u(0) &= 1\,,
\end{aligned}$$

und zeigen hierfür, dass sich der Abschneidefehler nicht besser verhalten kann also $\mathcal{O}(h^s)$. Die exakte Lösung lautet

$$u(t) \;=\; e^{\lambda t}\,.$$

Insbesondere gilt also zum Zeitpunkt t_n auch die Darstellung

$$u(t_n) \;=\; e^{\lambda(t_{n-1}+h_n)} \;=\; e^{\lambda h_n} u(t_{n-1})\,.$$

In Lemma 3.6 wurde gezeigt, dass sich die diskrete Lösung einer expliziten Runge-Kutta Methode der Stufe s schreiben läßt in der Form

$$u_n^h \;=\; p(\lambda h_n) u_{n-1}^h\,.$$

Die bedeutet: wenn wir die Runge-Kutta Verfahrensfunktion anwenden auf die exakte Lösung zum Zeitpunkt t_{n-1}, erhalten wir

$$p(\lambda h_n) u(t_{n-1}) \;=\; u(t_{n-1}) + h_n \Phi_{RK(s)}(h_n, t_{n-1}, u(t_{n-1}))\,.$$

mit einem Polynom $p \in \mathbb{P}_s$. Damit erhalten wir für den Abschneidefehler

$$\begin{aligned}
\tau_n^h &= \frac{u(t_n) - u(t_{n-1})}{h_n} - \Phi_{RK(s)}(h_n, t_{n-1}, u(t_{n-1})) \\
&= \frac{e^{\lambda t_n}}{h_n} - \frac{1}{h_n}\left(u(t_{n-1}) + h_n \Phi_{RK(s)}(h_n, t_{n-1}, u(t_{n-1}))\right) \\
&= \frac{e^{\lambda t_n}}{h_n} - \frac{1}{h_n} p(\lambda h_n) e^{\lambda t_{n-1}} \\
&= \frac{1}{h_n} e^{\lambda t_{n-1}} \left(e^{\lambda h_n} - p(\lambda h_n)\right)\,.
\end{aligned}$$

Damit ist die Frage nach der Konsistenzordnung zurück zu führen auf die Frage, wie gut sich die Exponentialfunktion $(e^{\lambda h_n})$ durch ein Polynom $(p(\lambda h_n))$ vom maximalen Grad m approximieren läßt. Hierbei soll die Approximation für $h_n \to 0$ besonders gut sein. Die Exponentialfunktion ist eine Potenzreihe, so dass das best-approximierende Polynom nahe der Null gerade der m-ten Partialsumme entspricht. Für dieses Polynom $p^* \in \mathbb{P}_n$ gilt:

$$e^{\lambda h_n} - p^*(\lambda h_n) \;=\; \sum_{k=s+1}^{\infty} \frac{(\lambda h_n)^k}{k!}\,.$$

Insgesamt folgt daher

$$|\tau_n^h| \;=\; e^{\lambda t_{n-1}} \lambda \sum_{k=s+1}^{\infty} \frac{(\lambda h_n)^{k-1}}{k!} \;=\; c_1 h_n^s + c_2 h_n^{s+1} + \ldots$$

mit $c_1 \neq 0$. Daher gilt nicht mehr als $|\tau_n^h| = \mathcal{O}(h_n^s)$. ∎

3.4.2 Implizite Runge-Kutta Methoden

Sobald für ein Paar $i \leq j$ gilt $\gamma_{ij} \neq 0$, so handelt es sich um ein implizites Runge-Kutta Verfahren. In diesem Fall lassen sich die k_i nicht einfach sukzessive berechnen, sondern führen

auf ein gekoppeltes System von i.a. nichtlinearen Gleichungen (wenn $f(t,x)$ nichtlinear in x ist). Dadurch dass man bei impliziten Verfahren mehr Koeffizienten im Butcher Diagramm zur Verfügung hat (die Diagonaleinträge und die des rechten oberen Teils müssen nicht notwendigerweise alle verschwinden), hat man sehr viel mehr Wahlmöglichkeiten bei impliziten Verfahren. Auch ist die Konsistenzordnung i.a. nicht durch die Anzahl m an Stufen beschränkt, sondern man kann sogar eine höhere Ordnung erreichen. Letztendlich ist noch anzumerken, dass die Stabilität der impliziten Verfahren i.a. sehr viel besser sind; doch dazu später mehr.

An dieser Stelle sei doch kurz angemerkt, dass man zwischen verschiedenen Verfahrensklassen impliziter RK-Methoden unterscheidet:

- DIRK (diagonal-implizit): $\gamma_{ij} = 0$ für alle $i < j$.
- SDIRK (einfach diagonal-implizit): $\gamma_{ij} = 0$ für alle $i < j$ und $\gamma_{ii} = \gamma$ für alle $i \in \{1, \ldots, m\}$ mit einem festen $\gamma \neq 0$.
- FIRK (voll implizit): $\gamma_{ij} \neq 0$ für mindestens ein Paar $i < j$.
- LIRK (linear implizit): ein implizites Verfahren, bei denen das resultierende nichtlineare Gleichungssystem nicht komplett gelöst wird, sondern nur ein Newton-Schritt ausgeführt wird.

Beispiele von impliziten Runge-Kutta Methoden sind:

- $s = 1$: Impliziter Euler:

$$\Phi_{IRK(1)}(h,t,x,y) \;=\; f(t+h,y)$$

- $s = 2$: Trapez-Regel:

$$\Phi_{IRK(2)}(h,t,x,y) \;=\; \frac{1}{2}\big(f(t,x) + f(t+h,y)\big)$$

$$u_n^h \;=\; u_{n-1}^h + \frac{h_n}{2}\big(f(t_{n-1},u_{n-1}^h) + f(t_n,u_n^h)\big)$$

- $s = 2$: 2-stufige Runge-Kutta-Formel an Gauß-Punkten:

$$\begin{aligned}\Phi_{IRK(2)}(h,t,x,y) &= \frac{1}{2}(k_1+k_2)\,,\\ \text{bzw. } u_n^h &= u_{n-1}^h + \frac{h_n}{2}(k_1+k_2)\,,\end{aligned}$$

mit den gekoppelten Größen ($\xi_{1,2} = 1/2 \mp \sqrt{3}/6$ und $\sigma_{1,2} = 1/4 \mp \sqrt{3}/6$):

$$\begin{aligned}k_1 &= f(t+\xi_1 h, x + \tfrac{1}{4}k_1 + \sigma_1 k_2)\,,\\ k_2 &= f(t+\xi_2 h, x + \sigma_2 k_1 + \tfrac{1}{4}k_2)\,.\end{aligned}$$

1	1
	1

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

$\frac{1}{2}-c$	$\frac{1}{4}$	$\frac{1}{4}-c$
$\frac{1}{2}+c$	$\frac{1}{4}+c$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

Tabelle 3.2: Butcher-Diagramme für impliziten RK-Methoden für impliziten Euler $s = 1$, Trapez-Regel $s = 2$ und Gauß-Punkten $s = 2$ mit der Konstanten $c = \sqrt{3}/6$ (von links nach rechts).

Hier erscheint es, als ob $\Phi_{IRK(2)}(h, t, x, y)$ unabhängig von y ist. Dies ist aber nicht der Fall, da k_1 und k_2 voneinander abhängig sind und damit auch von $y = x + h/2(k_1 + k_2)$. Es gilt folgende Aussage:

Satz 3.8 *Die implizite 2-stufige Runge-Kutta Methode an Gauß-Punkten besitzt die Konsistenzordnung $m = 4$.*

Beweis. Der Beweis resultiert direkt aus der Approximationsordnung für die numerische Quadratur. Die Gauß-Quadratur mit einem linearen Polynom ($n = 1$) an zwei Stützstellen ist exakt für Polynome vom Grad $2n + 1 = 3$. Also besitzt sie die (Approximations)-Ordnung $2n + 2 = 4$. Dies entspricht auch der Konsistenzordnung:

$$\|\tau_n^h\| \;=\; \left\| \int_{t_{n-1}}^{t_n} \frac{1}{h_n} f(s, u(s))\, ds - \Phi_{IRK(2)}(h_n, t_{n-1}, u(t_{n-1}), u(t_n)) \right\| \;=\; \mathcal{O}(h_n^4)\,.$$

□

3.5 Lokale Konvergenzaussagen bei Einschrittverfahren

Definition 3.9 *Ein Einschrittverfahren heißt lipschitz-stetig, wenn die zugehörige Verfahrensfunktion $\Phi(h, t, x, y)$ lipschitz-stetig ist bzgl. x und y, d.h. $\exists L_\Phi \in \mathbb{R}$:*

$$\|\Phi(h, t, x_1, y_1) - \Phi(h, t, x_2, y_2)\| \;\leq\; L_\Phi(\|x_1 - x_2\| + \|y_1 - y_2\|)\,,$$

für alle $x_1, x_2, y_1, y_2 \in \mathbb{R}^n$ und alle $t \in I$.

Satz 3.10 (Konvergenzsatz für ESV) *Sei Φ ein konsistentes und lipschitz-stetiges explizites Einschrittverfahren mit Lipschitz-Konstante L_Φ und Abschneidefehlern (τ_n^h). Dann gilt für den Fehler $\varepsilon_n := \|u(t_n) - u_n^h\|$*

$$\varepsilon_n \;\leq\; \exp\{L_\Phi(t_n - t_0)\} \left(\varepsilon_0 + \sum_{i=1}^{n} h_i \|\tau_i^h\| \right)\,.$$

Im Fall eines impliziten ESV mit obigen Voraussetzungen gilt sofern $h \leq (2L_\Phi)^{-1}$ die Abschätzung

$$\varepsilon_n \;\leq\; 2\exp\{2L_\Phi(t_n - t_0)\} \left(\varepsilon_0 + \sum_{i=1}^{n} h_i \|\tau_i^h\| \right)\,.$$

In diesem Satz ist offensichtlich auch ein eventueller Fehler in den Anfangsbedingungen $\varepsilon_0 = \|u_0 - u_0^h\|$ berücksichtigt.

Beweis. (a) Wir führen den Beweis zunächst für explizite ESV. Die beiden Gleichungen

$$\begin{aligned} u(t_n) &= u(t_{n-1}) + h_n \Phi(h_n, t_{n-1}, u(t_{n-1})) + h_n \tau_n^h \\ u_n^h &= u_{n-1}^h + h_n \Phi(h_n, t_{n-1}, u_{n-1}^h)\,, \end{aligned}$$

ergeben durch Subtraktion und Dreiecksungleichung

$$\begin{aligned} \varepsilon_n &\leq \varepsilon_{n-1} + h_n \|\Phi(h_n, t_{n-1}, u(t_{n-1})) - \Phi(h_n, t_{n-1}, u_{n-1}^h)\| + h_n \|\tau_n^h\| \\ &\leq \varepsilon_{n-1} + L_\Phi h_n \varepsilon_{n-1} + h_n \|\tau_n^h\|\,. \end{aligned}$$

Rekursive Anwendung liefert dann

$$\varepsilon_n \leq \varepsilon_0 + \sum_{i=1}^{n} h_i \left(L_\Phi \varepsilon_{i-1} + \|\tau_i^h\| \right) = a_n + \sum_{i=0}^{n-1} b_i \varepsilon_i\,, \tag{3.6}$$

mit $a_n := \varepsilon_0 + \sum_{i=1}^n h_i \|\tau_i^h\|$ und $b_i := L_\Phi h_{i+1}$. Das diskrete Grönwall'sche Lemma liefert nun

$$\begin{aligned} \varepsilon_n &\leq a_n \exp\left\{ \sum_{i=0}^{n-1} b_i \right\} \\ &= \left(\varepsilon_0 + \sum_{i=1}^{n} h_i \|\tau_i^h\| \right) \exp\left\{ L_\Phi (t_n - t_0) \right\} \end{aligned}$$

(b) Im Fall eines impliziten Verfahrens ergibt sich analog

$$\varepsilon_n \leq \varepsilon_{n-1} + L_\Phi h_n (\varepsilon_{n-1} + \varepsilon_n) + h_n \|\tau_n^h\|\,.$$

Aufgrund der zusätzlichen Bedingung an den Zeitschritt $L_\Phi h_n \leq \frac{1}{2}$ folgt

$$\varepsilon_n \leq 2\varepsilon_{n-1} + 2 L_\Phi h_n \varepsilon_{n-1} + 2 h_n \|\tau_n^h\|\,.$$

Die rekursive Anwendung führt wieder auf eine obere Schranke wie in (3.6), aber mit $a_n := 2\varepsilon_0 + 2\sum_{i=1}^n h_i \|\tau_i^h\|$ und $b_i := 2 L_\Phi h_{i+1}$. Dies führt auf

$$\begin{aligned} \varepsilon_n &\leq a_n \exp\left\{ \sum_{i=0}^{n-1} b_i \right\} \\ &= 2 \left(\varepsilon_0 + \sum_{i=1}^{n} h_i \|\tau_i^h\| \right) \exp\left\{ 2 L_\Phi (t_n - t_0) \right\}\,. \end{aligned}$$

□

Dieser Satz führt mit der Bezeichnung $\tau^h := \max\{\|\tau_i^h\| : 1 \le i \le n\}$ und $T := t_n - t_0$ beim expliziten Verfahren auf

$$\varepsilon_n \quad \le \quad e^{L_\Phi T}\left(\varepsilon_0 + T\tau^h\right) .$$

Der globale Fehler ist also bei lipschitz-stetigen Verfahrensfunktionen von der gleichen Ordnung wie der Abschneidefehler. Allerdings beobachtet man ein starkes exponentielles Anwachsen mit der zeitlichen Länge T. Es stellt sich daher die Frage, ob man auch Abschätzungen erhalten kann, die unabhängig sind von T. Dies führen wir am Beispiel des implizites Euler-Verfahrens durch. Mit höherem Aufwand sind ähnliche Ergebnisse auch bei anderen expliziten Verfahren durchführbar.

3.6 Evolution bei gestörten Anfangsdaten

In diesem Abschnitt verfolgen wir die Frage, ob es möglich ist, sich von der exponentiellen zeitlichen Abhängigkeit in Satz 3.10 zu lösen, sofern die Differentialgleichung selbst keine Fehlerakkumulierung beinhaltet. Im folgenden bezeichnet $\langle\cdot,\cdot\rangle$ das Skalarprodukt zur Norm $\|\cdot\|$.

Definition 3.11 *Eine Funktion $f : D \to \mathbb{R}^n$ genügt einer* einseitigen Lipschitz-Bedingung*, wenn eine stückweise stetige Funktion $l : I \to \mathbb{R}$ existiert, so dass folgende* Monotoniebedingung *gilt:*

$$\langle f(t,x) - f(t,y), x - y\rangle \quad \le \quad l(t)\|x-y\|^2 \qquad \forall x, y \in \mathbb{R}^n\,, \quad \forall t \in I\,. \tag{3.7}$$

Lemma 3.12 *Ist $f : D \to \mathbb{R}^n$ lipschitz-stetig mit Lipschitz-Konstante L, so erfüllt sie auch eine einseitige Lipschitz-Bedingung mit $l \equiv L$.*

Beweis. Aufgrund der Cauchy-Schwarz'schen Ungleichung ergibt sich:

$$\begin{aligned}\langle f(t,x) - f(t,y), x - y\rangle \quad &\le \quad \|f(t,x) - f(t,y)\|\,\|x-y\| \\ &\le \quad L\|x-y\|^2 .\end{aligned}$$

□

Die Umkehrung gilt aber i.a. nicht; insbesondere ist auch $l(t) < 0$ zugelassen, während Lipschitz-Konstanten stets nicht-negativ sind $L \ge 0$.

Die Bedingung (3.7) ist für $l(t) \le 0$ eine Verallgemeinerung der Eigenschaft "monoton fallend" bei skalarwertigen Funktionen $f : D \to \mathbb{R}$, denn dann lautet (3.7):

$$\frac{f(t,x) - f(t,y)}{x - y} \quad \le \quad l(t) \qquad \forall x, y \in \mathbb{R}\,, x \ne y\,.$$

Im Fall einer in x differenzierbaren Funktion $f(t,x)$ mit negativer Ableitung ist dies erfüllt.

Neben der AWA (2.1) im Intervall $I = [a, \infty)$ betrachten wir die AWA mit gestörten Anfangswerten zu einem Zeitpunkt:

$$v'(t) = f(t, v(t)) \qquad t \geq t_0\,, \tag{3.8}$$

$$v(t_0) = v_0\,. \tag{3.9}$$

Satz 3.13 *Die Funktion $f : D \to \mathbb{R}^n$ genüge einer* einseitigen Lipschitz-Bedingung *(3.7). Dann gilt für die Fehlerfunktion zwischen der Lösung u von (2.1) und der Lösung v von (3.8)-(3.9):*

$$\|(u-v)(t)\| \leq \exp\left\{\int_{t_0}^{t} l(s)\,ds\right\}\|u_0 - v_0\| \qquad \forall t \geq t_0\,.$$

Beweis. Die Fehlerfunktion $\varepsilon(t) = \|u(t)-v(t)\|^2$ ist stetig differenzierbar. Die Ableitung ergibt sich aufgrund der einseitigen Lipschitz-Bedingung von f zu:

$$\begin{aligned}\varepsilon'(t) &= 2(u'(t) - v'(t), u(t) - v(t)) \\ &= 2(f(t,u(t)) - f(t,v(t)), u(t) - v(t)) \\ &\leq 2l(t)\varepsilon(t)\,.\end{aligned}$$

Nun betrachten wir $G(t) := \eta(t)\varepsilon(t)$ mit der positiven Funktion

$$\eta(t) := \exp\left\{-2\int_{t_0}^{t} l(s)\,ds\right\}.$$

Für die Ableitung gilt mit obiger Abschätzung für $\varepsilon'(t)$ und der Positivität von η:

$$\begin{aligned}G'(t) &= \eta(t)\varepsilon'(t) + \eta'(t)\varepsilon(t) \\ &= \eta(t)\varepsilon'(t) - 2l(t)\eta(t)\varepsilon(t) \\ &= \eta(t)\big(\varepsilon'(t) - 2l(t)\varepsilon(t)\big) \\ &\leq 0\,.\end{aligned}$$

Also ist G monoton fallend, und damit folgt für $t \geq t_0$:

$$\varepsilon(t) = \eta^{-1}(t)G(t) \leq \eta^{-1}(t)G(t_0) = \frac{\eta(t_0)}{\eta(t)}\varepsilon(t_0) = \exp\left\{2\int_{t_0}^{t} l(s)\,ds\right\}\varepsilon(t_0)\,.$$

Durch Ziehen der Quadratwurzel auf beiden Seiten erhält man die Behauptung. □

Im Fall $l(t) \leq 0$ wachsen Fehler also nicht über den Fehler in den Anfangsdaten hinaus an. Dieses Verhalten legt folgende Definition nahe:

Definition 3.14 *Die AWA (2.1) mit $f : D \to \mathbb{R}^n$ heißt* dissipativ *(oder* nicht-expansiv*), wenn f einer einseitigen Lipschitz-Bedingung (3.7) mit $l : I \to \mathbb{R}_{\leq 0}$ genügt. Die Verallgemeinerung im Komplexen, $f : D \to \mathbb{C}^n$, lautet:*

$$\langle f(t,x) - f(t,y), \overline{x-y}\rangle \leq l(t)\|x-y\|^2 \qquad \forall x, y \in \mathbb{C}^n\,, \quad \forall t \in I\,,$$

mit $Re\, l(t) \leq 0$.

Im folgenden Satz bezeichne $\tau_n^{h,\epsilon}$ den maximalen Abschneidefehler für alle möglichen Lösungen, die in einer $\epsilon-$Umgebung um die Lösung $u(t)$ verlaufen:

$$U_\epsilon(u) \;:=\; \{v : I \to \mathbb{R}^n : \|u - v\|_{L^\infty(I)} < \epsilon\}$$

Satz 3.15 *Die AWA (2.1) sei dissipativ mit $l(t) \leq -\alpha < 0$ für $t \in I$ und f sei lipschitz-stetig. Wir betrachten ein beliebiges lipschitz-stetiges und konsistentes Einschrittverfahrens (3.1) und die zugehörige diskrete Lösung u^h. Dann existieren zu jedem $\epsilon > 0$ eine von I unabhängige Konstante $K \in \mathbb{R}$ und eine Maximalschrittweite $h_{max} > 0$, so dass für $\max_{n\in\mathbb{N}} h_n \leq h_{max}$ folgt:*

$$\max_{n\in\mathbb{N}} \|u(t_n) - u_n^h\| \;\leq\; K \max_{n\in\mathbb{N}} \|\tau_n^{h,\epsilon}\|\,.$$

Beweis. Der Beweis erfolgt in vier Schritten. Wir gehen zur Vereinfachung von einer konstanten Schrittweite h aus. Die Verallgemeinerung für variable Schrittweite ist einfach.
(a) Fehlerabschätzung für kurze Zeiten: Die Fehlerabschätzung aus Satz 3.10 liefert mit $\varepsilon_0 = 0$ die Schranke

$$\max_{0\leq n\leq m} \|u(t_n) - u_n^h\| \;\leq\; K_m \max_{1\leq n\leq m} \|\tau_n^h\|\,,$$

mit bislang noch beliebigem $m \in \mathbb{N}$ und

$$\begin{aligned} K_m &= (t_m - t_0)e^{L_\Phi(t_m - t_0)} \qquad \text{für explizite ESV,} \\ \text{bzw.}\quad K_m &= 2(t_m - t_0)e^{2L_\Phi(t_m - t_0)} \qquad \text{für implizite ESV.} \end{aligned}$$

Somit gilt die Behauptung zunächst für die endlich vielen Zeitschritte $0 \leq n \leq m$ mit festem aber beliebigem m. Wir wählen nun ein festes $m \in \mathbb{N}$, so dass für $T := t_m - t_0$ gilt:

$$\exp(-\alpha T) \;\leq\; \frac{1}{2} \tag{3.10}$$

und setzen $K := 2K_m$. Offensichtlich kann m und damit auch T beliebig groß gewählt werden. Dies ändert lediglich die Konstante K.
(b) Schrittweitenbeschränkung: Es gilt für beliebig großes N und für $0 < h \leq h_{max}$,

$$\lim_{h_{max}\to 0} \left(K \max_{1\leq k\leq N} \|\tau_k^h\| \right) \;=\; 0\,.$$

Daher gilt für hinreichend kleines $h_{max} > 0$:

$$K \max_{1\leq k\leq N} \|\tau_k^h\| \;\leq\; \delta.$$

(c) Induktionsverankerung: Nun führen wir den Beweis per Induktion. Es gilt also für ein $n \geq 1$ und $t_n - t_0 \geq T$ folgende Induktionsvoraussetzung an den diskreten Zeitpunkten t_k:

$$\max_{0\leq k\leq n} \|u(t_k) - u_k^h\| \;\leq\; K \max_{0\leq k\leq n} \|\tau_k^{h,\epsilon}\|\,.$$

Ferner folgt wegen (b) insbesondere für $0 < h \leq h_{max}$:

$$\|u(t_n) - u_n^h\| \leq \epsilon\,.$$

Nun wollen wir zeigen, dass hieraus, zusammen mit einer Schrittweitenbeschränkung, auch folgt:

$$\|u(t_{n+1}) - u_{n+1}^h\| \leq K \max_{0\leq k\leq n+1} \|\tau_k^{h,\epsilon}\|\,. \tag{3.11}$$

(d) Induktionsschritt: Wir betrachten den Zeitpunkt $t_m \leq t_n$, so dass $t_{n+1} - t_m \geq T$. Sei $v : J \to \mathbb{R}^n$ die Lösung der gestörten AWA im Intervall $J := I \cap [t_m, \infty)$

$$\begin{aligned} v'(t) &= f(t, v(t)) \qquad (t \in J)\,, \\ v(t_m) &= u_m^h\,. \end{aligned}$$

Da die Lösung u als exponentiell stabil vorausgesetzt wurde, wegen Bedingung (3.10) sowie der Induktionsvoraussetzung folgt:

$$\begin{aligned} \|(u-v)(t_{n+1})\| &\leq e^{-\alpha(t_{n+1}-t_m)}\|(u-v)(t_m)\| \leq e^{-\alpha T}\|u(t_m) - u_m^h\| \\ &\leq \frac{1}{2}\|u(t_m) - u_m^h\| \leq \epsilon\,. \end{aligned}$$

Mit der Bezeichnung $\tilde{\tau}_n^h$ für den Abschneidefehler zu v ergibt die nochmalige Anwendung von Satz 3.10:

$$\begin{aligned} \|v(t_{n+1}) - u_{n+1}^h\| &\leq (t_{n+1} - t_m)e^{L_\Phi(t_{n+1}-t_m)} \max_{m\leq k\leq n+1} \|\tilde{\tau}_k^h\| \\ &\leq Te^{L_\Phi T} \max_{m\leq k\leq n+1} \|\tilde{\tau}_k^h\| \\ &\leq \frac{1}{2}K \max_{m\leq k\leq n+1} \|\tau_k^{h,\epsilon}\|\,. \end{aligned}$$

Mittels der Dreiecksungleichung folgt

$$\begin{aligned} \|u(t_{n+1}) - u_{n+1}^h\| &\leq \|u(t_{n+1}) - v(t_{n+1})\| + \|v(t_{n+1}) - u_{n+1}^h\| \\ &\leq \frac{1}{2}\|u(t_m) - u_m^h\| + \frac{1}{2}K \max_{m\leq k\leq n+1} \|\tau_k^{h,\epsilon}\| \\ &\leq K \max_{0\leq k\leq n+1} \|\tau_k^{h,\epsilon}\|\,. \end{aligned}$$

Damit ist der Induktionsschritt (3.11) gezeigt und der Beweis vollendet. □

3.7 Implizites Euler-Verfahren

In diesem Abschnitt betrachten wir das implizite Euler-Verfahren als das "einfachste" implizite Verfahren. Hierfür zeigen wir direkt eine globale Fehlerabschätzung vom Charakter wie in Satz 3.15.

Um die Lösbarkeit der auftretenden Gleichungen beim impliziten Euler-Verfahren zu beweisen, benötigen wir zur Vorbereitung ein paar Hilfsmittel:

Definition 3.16 *Eine Abbildung $g : \mathbb{R}^n \to \mathbb{R}^n$ heißt* strikt monoton*, wenn ein $\lambda > 0$ existiert, so dass für das euklidische Skalarprodukt gilt:*

$$\langle g(x) - g(y), x - y \rangle \quad \geq \quad \lambda \|x - y\|^2 \qquad \forall x, y \in \mathbb{R}^n \,.$$

Im Fall, dass g linear ist, also $g(x) = Ax$ mit $A \in \mathbb{R}^{n \times n}$, so ist g strikt monoton offensichtlich äquivalent mit A positiv definit.

Lemma 3.17 *Sei $g : \mathbb{R}^n \to \mathbb{R}^n$ lipschitz-stetig und strikt monoton. Dann besitzt die Gleichung $g(x) = c$ für beliebiges $c \in \mathbb{R}^n$ eine eindeutige Lösung $x \in \mathbb{R}^n$.*

Beweis. Wir verwenden den Banachschen Fixpunktsatz und formulieren hierzu die Gleichung in eine Fixpunktgleichung:

$$G_\theta(x) \quad := \quad x - \theta(g(x) - c) \;=\; x \,.$$

Der Parameter $\theta > 0$ ist noch zu bestimmen. Sei L_g die Lipschitz-Konstante von g. Dann gilt aufgrund der Lipschitz-Stetigkeit und der strikten Monotonie:

$$\begin{aligned} \|G_\theta(x) - G_\theta(y)\|^2 &= \|x - \theta(g(x) - c) - y + \theta(g(y) - c)\|^2 \\ &= \|x - y + \theta(g(y) - g(x))\|^2 \\ &\leq \|x - y\|^2 + \theta^2 \|g(y) - g(x)\|^2 - 2\theta \langle x - y, g(y) - g(x) \rangle \\ &\leq (1 + \theta^2 L_g^2 - 2\theta\lambda) \|x - y\|^2 \,. \end{aligned}$$

Für die Konstante

$$K \quad := \quad 1 + \theta^2 L_g^2 - 2\theta\lambda \;=\; 1 - \theta(2\lambda - \theta L_g^2)$$

gilt im Fall $0 < \theta < \min(L_g^{-2}\lambda, (2\lambda)^{-1})$ offensichtlich $0 \leq K < 1$. Also ist G_θ für solche θ eine Kontraktion und besitzt nach dem Banachschen Fixpunktsatz in der abgeschlossenen Menge $\mathbb{R}^n$ einen eindeutigen Fixpunkt x. Dieses x ist Lösung der Gleichung $g(x) = c$. ∎

Die Verfahrensfunktionen des implizites Euler-Verfahrens lautet:

$$\Phi(h, t, x, y) \quad := \quad f(t + h, y) \,,$$

und führt somit auf

$$u_n^h \quad := \quad u_{n-1}^h + h_n f(t_n, u_n^h) \,. \tag{3.12}$$

Das diese nichtlineare Gleichung stets lösbar ist, ist nicht in jedem Fall gegeben. Wir haben aber Existenz von Lösungen im folgenden Spezialfall:

Satz 3.18 *Sei $f \in C(I \times \mathbb{R}^n, \mathbb{R}^n)$ lipschitz-stetig und erfülle die Monotonie-Bedingung (3.7). Der Startwerte u_0^h sei beliebig. Dann besitzt das implizite Euler-Verfahren (3.12) im dissipativen Fall $l \leq 0$ für beliebige Schrittweiten h_n und im Fall $l > 0$ für $h_n < l(t_n)^{-1}$ stets eine eindeutige Lösung.*

Beweis. Seien $t_n \in I$ und $h_n \in \mathbb{R}$ fest aber beliebig. Die Abbildung $g(x) := x - h_n f(t_n, x)$ ist lipschitz-stetig und es gilt die strikte Monotonie:

$$\begin{aligned}\langle g(x) - g(y), x - y\rangle &= \langle x - h_n f(t_n, x) - y + h_n f(t_n, y), x - y\rangle \\ &= \|x - y\|^2 - h_n\langle f(t_n, x) - f(t_n, y), x - y\rangle \\ &\geq \|x - y\|^2 - h_n l(t_n)\|x - y\|^2 \\ &= (1 - l(t_n)h_n)\|x - y\|^2 .\end{aligned}$$

Unter den Voraussetzungen an h_n und l, folgt die strikte Monotonie von g. Nach Lemma 3.17 besitzt $g(x) = u^h_{n-1}$ demzufolge eine eindeutige Lösung $x = u^h_n$. Dies ist gerade die gesuchte Lösung, denn

$$u^h_n \;=\; u^h_{n-1} + h_n f(t_n, u^h_n) \quad \text{bzw.} \quad u^h_{n-1} \;=\; u^h_n - h_n f(t_n, u^h_n) \;=\; g(u^h_n) .$$

□

Satz 3.19 *Sei $f \in C(I \times \mathbb{R}^n, \mathbb{R}^n)$ lipschitz-stetig und dissipativ mit zugehöriger Funktion $l : I \to \mathbb{R}_{\leq 0}$ in (3.7). Für die Lösung der AWA gelte $u \in C^2(I, \mathbb{R}^n)$. Dann gilt für das implizite Euler-Verfahren mit Schrittweiten $h_k \leq 1/|l(t_k)|$, $k = 1, \ldots, n$, die Abschätzung:*

$$\varepsilon_n \;\leq\; \varepsilon_0 + \frac{1}{2}(t_n - t_0) \max_{1\leq k\leq n} \left\{h_k \|u''\|_{L^\infty(I_k)}\right\} .$$

Beweis. Die Differenz der beiden Differenzengleichungen ergibt

$$u(t_n) - u^h_n \;\leq\; u(t_{n-1}) - u^h_{n-1} + h_n(f(t_n, u(t_n)) - f(t_n, u^h_n)) + h_n\tau^h_n .$$

Wir multiplizieren beide Seiten mit $u(t_n) - u^h_n$ und erhalten

$$\begin{aligned}\varepsilon^2_n &\leq \varepsilon_{n-1}\varepsilon_n + h_n\langle f(t_n, u(t_n)) - f(t_n, u^h_n), u(t_n) - u^h_n\rangle + h_n\|\tau^h_n\|\varepsilon_n \\ &\leq \varepsilon_{n-1}\varepsilon_n + h_n l(t_n)\varepsilon^2_n + h_n\|\tau^h_n\|\varepsilon_n .\end{aligned}$$

Hieraus folgt:

$$(1 - h_n l(t_n))\varepsilon_n \;\leq\; \varepsilon_{n-1} + h_n\|\tau^h_n\| ,$$

bzw. (mit der offensichtlichen Interpretation für $l(t_n) = 0$)

$$\begin{aligned}\varepsilon_n &\leq \frac{1}{1 - h_n l(t_n)}\varepsilon_{n-1} + \frac{h_n}{1 - h_n l(t_n)}\|\tau^h_n\| \\ &\leq \varepsilon_{n-1} + \min(|l(t_n)|^{-1}, h_n)\|\tau^h_n\| \\ &= \varepsilon_{n-1} + h_n\|\tau^h_n\| .\end{aligned}$$

Es folgt per Rekursion

$$\varepsilon_n \;\leq\; \varepsilon_0 + \sum_{k=1}^{n} h_k\|\tau^h_k\| \;\leq\; \varepsilon_0 + \left(\sum_{k=1}^{n} h_k\right) \max_{1\leq k\leq n} \|\tau^h_k\| .$$

Mit der oberen Grenze für den Abschneidefehler folgt die Behauptung. □

3.8 Schrittweitenkontrolle und adaptive Schrittweite

Da der Lösungsaufwand für eine Differentialgleichung erheblich von der Wahl der Schrittweiten h_n abhängt, ist es in vielen Anwendungen sinnvoll, diese Schrittweite angemessen und evtl. sogar variabel zu wählen. Zu Zeitpunkten, an denen sich die Lösung nur sehr leicht ändert, also $\|f(t,u(t)\|$ ist "klein", kann sollte man die Schrittweite kleiner wählen, als an Zeitpunkten mit "großer" Dynamik. Gute Strategien zielen darauf ab, den Fehler in jedem Zeitschritt möglichst konstant (klein) zu halten. Der globale Fehler

$$e_n \;:=\; u(t_n) - u_n^h$$

sollte durch geeignete Wahl von h_n also normmäßig stets von annährend gleicher Größe sein:

$$\varepsilon_n \;:=\; \|e_n\| \;\approx\; \text{const.}\; \forall n \in \mathbb{N}\,.$$

Gemäß Satz 3.10 wissen wir, dass mit einer (i.a. exponentiell wachsenden Funktion K) gilt mit $T = t_n - t_0$:

$$\max_{1\le i\le n} \varepsilon_i \;\le\; K(T) \sum_{i=1}^{n} h_i \|\tau_i^h\|\,.$$

Eine mit T wachsende Konstante ist aber der "worst case". Wir gehen in diesem Abschnitt davon aus, dass solche Abhängigkeit nicht besteht. Im Fall von dissipativen AWA haben wir gesehen, dass diese Annahme gerechtfertigt ist. Um den Fehler zu minimieren, ist es daher sinnvoll, $h_i\|\tau_i^h\|$ weitesgehend konstant (klein) zu halten. Im folgenden soll es darum gehen, diesen mit der Schrittweite gewichteten Abschneidefehler durch einen Schätzer η_n zu schätzen und diesen mit einer vorgegebenen Toleranz `tol` zu beschränken:

$$h_n\|\tau_n^h\| \;\approx\; \eta_n \;\le\; \texttt{tol}\,.$$

Für ein ESV der Konsistenzordnung m können wir von einer Darstellung

$$\eta_n \;=\; c(t_n)h_n^{m+1} + \mathcal{O}(h_n^{m+2})$$

mit einer h unabhängigen Funktion $c(t)$ ausgehen. Wir wollen zunächst davon ausgehen, dass die Information des Schätzers η_n vorliegt und beschäftigen uns mit der Frage, wie man hieraus einen neuen Zeitschritt h_n bestimmt. Sei h_n^* der Zeitschritt für den der zugehörige Schätzers η_n^* gerade der Toleranz entspricht, also

$$\texttt{tol} \;=\; c(t_n)(h_n^*)^{m+1} + \mathcal{O}((h_n^*)^{m+2})$$

Unter Vernachläßigung der Terme höherer Ordnung erhalten wir

$$\frac{\texttt{tol}}{\eta_n} \;\approx\; \left(\frac{h_n^*}{h_n}\right)^{m+1}\,.$$

```
set t := 0, n := 1, ρ = 0.9, q = 1.1
initialize u_0 and h
while t < T do
    u_n^h := u_{n-1}^h + hΦ(h, t, u_{n-1}^h, u_n^h)
    Compute estimator η
    h* := ρ (tol/η)^{1/(m+1)} h
    h := min(h_max, qh, h*, T − t)
    if (η < tol){
        t := t + h
        n := n + 1}
end
```

Tabelle 3.3: Algorithmus zur Lösung einer Differentialgleichung mit ESV der Ordnung m und adaptiver Zeitschrittweite.

Eine mögliche Strategie ist daher, mit Einführung eines Sicherheitsfaktors $\rho < 1$ den neuen Zeitschritt zu wählen als:

$$h_n^* \quad := \quad \rho \left(\frac{\texttt{tol}}{\eta_n} \right)^{\frac{1}{m+1}} h_n \, .$$

Die praktische Umsetzung könnte lauten wie im Algorithmus in Tabelle 3.3. Hierbei wurde noch ein Faktor $q > 1$ verwendet, um den Zeitschritt ggf. kontrolliert zu erhöhen.

3.9 Fehlerschätzung

Nun wollen wir uns mit der Frage beschäftigen, wie man einen Fehlerschätzer η_n erhält. Eine Möglichkeit ist, zwei unterschiedliche Verfahren Φ und $\widehat{\Phi}$ anzuwenden:

$$\begin{aligned} u_n^h &:= u_{n-1}^h + h_n \Phi(h_n, t_{n-1}, u_{n-1}^h, u_n^h) \\ \widehat{u}_n^h &:= u_{n-1}^h + \widehat{h}_n \widehat{\Phi}(\widehat{h}_n, t_{n-1}, u_{n-1}^h, \widehat{u}_n^h) \end{aligned}$$

Hierbei sind h_n und $\widehat{h}_n$ zwei möglicherweise unterschiedliche Zeitschritte. Es sind hierbei zwei Möglichhkeiten denkbar:

- $0 < \widehat{h}_n < h_n$ und $\Phi = \widehat{\Phi}$,
- $h_n = h_n$ aber $\widehat{\Phi}$ besitzt eine höhere Konsistenzordnung als Φ.

In beiden Fällen kann $\widehat{u}_n^h$ als genauer angenommen werden als u_n^h. Ein häufig benutzter Schätzer lautet dann

$$\eta_n \quad := \quad \| \widehat{u}_n^h - u_n^h \| \, . \tag{3.13}$$

Wir führen nun noch hilfsweise die folgende AWA ein:

$$\begin{aligned} v'(t) &= f(t, v(t)) \qquad t \geq t_{n-1}\,, \\ v(t_{n-1}) &= u^h_{n-1}\,, \end{aligned}$$

und bezeichnen mit $\tau(v)$ und $\widehat{\tau}(v)$ die Abschneidefehler zu diesem Problem für zwei explizites ESV $\Phi, \widehat{\Phi}$ mit gleicher Schrittweite $\widehat{h}_n = h_n$:

$$\begin{aligned} \tau(v) &:= \frac{v(t_n) - v(t_{n-1})}{h_n} - \Phi(h_n, t_{n-1}, v(t_{n-1}))\,, \\ \widehat{\tau}(v) &:= \frac{v(t_n) - v(t_{n-1})}{h_n} - \widehat{\Phi}(h_n, t_{n-1}, v(t_{n-1}))\,. \end{aligned}$$

Das folgende Lemma macht nun eine Aussage über den Schätzer η, wenn $\widehat{\Phi}$ genauer ist als Φ, bzw. der Abschneidefehler $\widehat{\tau}(v)$ kleiner ist als $\tau(v)$.

Lemma 3.20 *Sei θ der Quotient aus obigen lokalen Fehlern, $\theta := \|\widehat{\tau}(v)\|/\|\tau(v)\|$. Dann gilt für den Schätzer (3.13) eines expliziten ESV*

$$(1-\theta)h_n\|\tau(v)\| \;\leq\; \eta_n \;\leq\; (1+\theta)h_n\|\tau(v)\|\,.$$

Beweis. Es gilt:

$$\begin{aligned} v(t_n) &= u^h_{n-1} + h_n\Phi(h_n, t_{n-1}, u^h_{n-1}) + h_n\tau(v)\,, \\ u^h_n &= u^h_{n-1} + h_n\Phi(h_n, t_{n-1}, u^h_{n-1})\,, \\ \Longrightarrow \qquad v(t_n) - u^h_n &= h_n\tau(v)\,. \end{aligned}$$

Analog gilt

$$v(t_n) - \widehat{u}^h_n \;=\; h_n\widehat{\tau}(v)\,.$$

Nun können wir η_n als Differenz der Abschneidefehler ausdrücken:

$$\eta_n \;=\; \|\widehat{u}^h_n - u^h_n\| \;=\; h_n\|\widehat{\tau}(v) - \tau(v)\|\,.$$

Die Behauptung folgt nun unmittelbar aus zweimaliger Anwendung der Dreiecksungleichung:

$$\begin{aligned} (1-\theta)\|\tau(v)\| &= \|\tau(v)\| - \|\widehat{\tau}(v)\| \;\leq\; \|\tau(v) - \widehat{\tau}(v)\| \;=\; h_n^{-1}\eta_n \\ h_n^{-1}\eta_n &\leq \|\tau(v)\| + \|\widehat{\tau}(v)\| \;=\; (1+\theta)\|\tau(v)\|\,. \end{aligned}$$

□

Der Schätzer η_n ist also eine umso bessere Approximation an $h_n\|\tau(v)\|$, je kleiner θ ist.

Kapitel 4

Numerische Stabilität

4.1 Stabilitätsfunktion

Wir betrachten zu $\lambda \in \mathbb{C}$ die lineare skalare AWA bestehend aus der Dahlquist'schen Testgleichung

$$\begin{aligned} u'(t) &= \lambda u(t) \qquad t \geq 0\,, && (4.1)\\ u(0) &= 1\,. && (4.2) \end{aligned}$$

Die Lösung lautet $u(t) = e^{\lambda t}$ und deren Betrag ist $|u(t)| = e^{Re\,\lambda\, t}$, also

$$\begin{aligned} Re\,\lambda < 0 &\implies \lim_{t\to\infty} |u(t)| = 0\,,\\ Re\,\lambda = 0 &\implies |u(t)| = 1 \quad \forall t \geq 0\,,\\ Re\,\lambda > 0 &\implies \lim_{t\to\infty} |u(t)| = +\infty\,. \end{aligned}$$

Definition 4.1 *Sei Φ ein numerisches Verfahren zur approximativen Lösung der AWA (4.1)-(4.2) und dieses lasse sich darstellen in der Form*

$$u_n^h = g(\lambda h_n) u_{n-1}^h$$

mit einer Funktion $g : \mathbb{C} \to \mathbb{C}$. Dann heißt g Stabilitätsfunktion *zu Φ.*

Eine Stabilitätsfunktion liefert sozusagen die Verstärkungsfaktoren für die Dahlquist'sche Testgleichung (mit $f(t,x) = \lambda x$). Mithilfe der Stabilitätsfunktion ergibt sich durch Rekursion für das obige Modelproblem bei konstanter Schrittweite $h = h_n$:

$$u_n^h = g(\lambda h)^n\,.$$

Beispiele:

1. Für das explizite Euler-Verfahren gilt

$$\begin{aligned} u_n^h &= u_{n-1}^h + h_n f(t_{n-1}, u_{n-1}^h) \;=\; u_{n-1}^h + h_n \lambda u_{n-1}^h \\ &= (1 + h_n\lambda) u_{n-1}^h \,. \end{aligned}$$

   Also lautet die Stabilitätsfunktion des expliziten Euler-Verfahrens $g(z) = 1 + z$. Im Fall $Re\,\lambda < 0$ muss also $|g(\lambda h_n)| = |1 + \lambda h_n| < 1$ gelten, damit auch für die diskrete Lösung gilt $\lim_{n\to\infty} |u_n^h| = 0$. Dies entspricht einer Schrittweitenbeschränkung:

$$|\lambda h_n - (-1)| < 1 \quad \text{bzw.} \quad 0 < h_n < 2|\lambda|^{-1} \,.$$

2. Für das implizite Euler-Verfahren gilt

$$\begin{aligned} & u_n^h &= u_{n-1}^h + h_n f(t_{n-1}, u_n^h) \;=\; u_{n-1}^h + h_n \lambda u_n^h \,, \\ \Longrightarrow \quad & u_n^h &= (1 - h_n\lambda)^{-1} u_{n-1}^h \,. \end{aligned}$$

   Also lautet die Stabilitätsfunktion des impliziten Euler-Verfahrens $g(z) = \frac{1}{1-z}$.

3. Für die Trapez-Regel gilt

$$\begin{aligned} & u_n^h &= u_{n-1}^h + \frac{1}{2} h_n \Big(f(t_{n-1}, u_{n-1}^h) + f(t_{n-1}, u_n^h) \Big) \\ \Longrightarrow \quad & \Big(1 - \frac{1}{2} h_n\lambda\Big)\, u_n^h &= \Big(1 + \frac{1}{2} h_n\lambda\Big)\, u_{n-1}^h \,, \\ & \Longrightarrow \quad u_n^h &= \frac{2 + h_n\lambda}{2 - h_n\lambda} u_{n-1}^h \end{aligned}$$

   Also lautet die Stabilitätsfunktion der Trapez-Regel $g(z) = \frac{2+z}{2-z}$.

Lemma 4.2 *Die Stabilitätsfunktionen von expliziten Runge-Kutta-Methoden sowie die der Taylor-Methoden sind stets Polynome vom echten Grad ≥ 1.*

Beweis. Wir hatten in Lemma 3.6 gesehen, dass die Stabilitätsfunktion eines s-stufigen expliziten Runge-Kutta-Verfahrens ein Polynom vom Grad s ist, insbesondere sind sie nicht konstant. Für die Taylor-Methoden gilt

$$\begin{aligned} u_n^h &= u_{n-1}^h + h_n \Phi_{TF(s)}(h_n, t_{n-1}, u_{n-1}^h) \\ &= u_{n-1}^h + h_n \sum_{k=0}^{s-1} \frac{1}{(k+1)!} h_n^k \frac{d^k}{dt^k} (\lambda x(t)) \Big|_{x=u_{n-1}^h} \\ &= u_{n-1}^h + \sum_{k=0}^{s-1} \frac{1}{(k+1)!} h_n^{k+1} \lambda (\lambda^k x(t)) \Big|_{x=u_{n-1}^h} \\ &= \sum_{k=0}^{s} \frac{1}{k!} h_n^k \lambda^k u_{n-1}^h \end{aligned}$$

Also ist $g(z) = \sum_{k=0}^{s} \frac{1}{k!} z^k$ und somit ebenfalls ein Polynom vom Grad s. ∎

Lemma 4.3 *Die Stabilitätsfunktion eines s-stufigen impliziten Runge-Kutta-Verfahrens (3.5) ist eine rationale Funktion der Form*

$$g(z) \;=\; 1 + \langle r, (I - zA)^{-1}\mathbb{1}\rangle z\,,$$

wobei $r \in \mathbb{R}^s$ der Vektor der Runge-Kutta-Koeffizienten r_i ist, $A = (\gamma_{ij}) \in \mathbb{R}^{s\times s}$ und der Bezeichnung $\mathbb{1} := (1,\ldots,1)^T \in \mathbb{R}^s$.

Beweis. Wir betrachten zunächst das Gleichungssystem für die RK-Koeffizienten k_i für unsere spezielle Dahlquist'schen Testgleichung:

$$k_i \;=\; f\left(t + a_i h_i, x + \sum_{j=1}^{s} \gamma_{ij} k_j h\right) \;=\; \lambda\left(x + \sum_{j=1}^{s} \gamma_{ij} k_j h\right)\,.$$

Dieses lineare Gleichungsystem läßt sich durch Umsortierung auch schreiben in der Form

$$k_i - \lambda h \sum_{j=1}^{s} \gamma_{ij} k_j \;=\; \lambda x \qquad \forall k \in \{1,\ldots,s\}\,.$$

In kompakter Matrixschreibweise lautet das lineare Gleichungsystem also:

$$(I - \lambda h A)k \;=\; \lambda x \mathbb{1}\,.$$

Damit lautet die Verfahrensfunktion des RK(s)-Verfahrens

$$\Phi_{RK(s)}(h,t,x,y) \;=\; \langle r, k\rangle \;=\; \langle r, (I - \lambda h A)^{-1}\mathbb{1}\rangle\, \lambda x.$$

Die Lösung y zum neuen Zeitschritt ergibt sich also aus dem vorherigen Wert x aus:

$$\begin{aligned} y \;&=\; x + h\Phi_{RK(s)}(h,t,x,y) \;=\; x + h\langle r, (I - \lambda h A)^{-1}\mathbb{1}\rangle\, \lambda x \\ &=\; (1 + \langle r, (I - \lambda h A)^{-1}\mathbb{1}\rangle \lambda h)x \\ &=\; g(\lambda h)x\,, \end{aligned}$$

mit der Stabilitätsfunktion g wie im Lemma angegeben. □

4.2 Stabilitätsgebiet und A-Stabilität

Definition 4.4 *Unter dem* Stabilitätsgebiet *(oder auch* Gebiet absoluter Stabilität*) eines ESV Φ mit Stabilitätsfunktion g versteht man die Menge*

$$S_\Phi \;:=\; \{z \in \mathbb{C} \,:\, |g(z)| < 1\}\,.$$

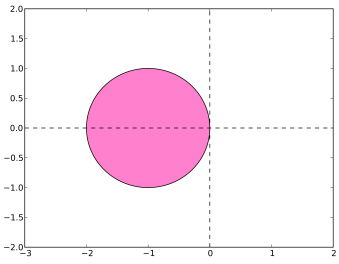

Abbildung 4.1: Stabilitätsgebiet des expliziten Eulers.

Damit ein numerisches Verfahren Φ für die einfache (und gutartige) Dahlquist'sche Testgleichung (4.1)-(4.2) mit $Re\,\lambda < 0$ eine asymptotisch korrekte diskrete Lösung liefert muss also gelten $h_n\lambda \in S_\Phi$. Für den expliziten und impliziten Euler gilt

$$\begin{aligned} S_{ExEuler} &= \{z \in \mathbb{C} : |z+1| < 1\} = B_1(-1)\,, \\ S_{ImEuler} &= \{z \in \mathbb{C} : |z-1| > 1\} = \mathbb{C} \setminus \overline{B_1(1)}\,. \\ S_{Trapez} &= \{z \in \mathbb{C} : |z+2| < |z-2|\} = \{z \in \mathbb{C} :\ Re\, z < 0\}\,. \end{aligned}$$

Hierbei ist $B_r(x)$ die offene Kugel in $\mathbb{C}$ um x mit Radius r.

Definition 4.5 *Ein numerischen Verfahrens Φ heißt* A-stabil *(oder auch* absolut stabil*), wenn dessen Stabilitätsgebiet S_Φ die gesamte linke komplexe Halbebene enthält:*

$$\{z \in \mathbb{C} :\ Re\, z < 0\} \quad \subseteq \quad S_\Phi\,.$$

Ein A-stabiles Verfahren erwirkt also für das obige Modellproblem mit $Re\,\lambda < 0$ das qualitativ gleiche Verhalten der Lösung, nämlich für konstante aber beliebig große Schrittweiten $h > 0$ die Eigenschaft

$$\lim_{n\to\infty} |u_n^h| \quad = \quad 0\,.$$

Lemma 4.6 *Ein ESV Φ mit Stabilitätsfunktion g ist genau dann* A-stabil*, wenn für alle $z \in \mathbb{C}$ gilt:*

$$Re\, z < 0 \quad \Longrightarrow \quad |g(z)| < 1\,.$$

Damit sieht man unmittelbar, dass das explizite Euler-Verfahren nicht A-stabil ist. Dies gilt sogar für alle expliziten Runge-Kutta-Methoden:

Korollar 4.7 *Explizite Runge-Kutta Methoden sind niemals A-stabil.*

Beweis. Nach Lemma 4.2 sind die zugehörigen Stabilitätsfunktionen nicht-konstante Polynome. Daher gilt

$$\lim_{|z|\to\infty} |g(z)| \quad = \quad \infty\,.$$

□

Beispiele für A-stabile Verfahren sind:

1. das implizite Euler-Verfahren,
2. die Trapezregel, denn für $Re\, z \leq 0$ gilt $|z+2| = |z-(-2)| \leq |z-2|$.

4.3 Exponentiell wachsende Lösungen

An dieser Stelle wollen wir uns den Fall $Re\,\lambda > 0$ ansehen. Das skalare Modellproblem (4.1)-(4.2) erzeugt jetzt exponentiell wachsende Lösungen $u(t)$. Das implizite Euler-Verfahren für Schrittweiten $h > 0$ mit $\lambda h \in S_{ImEuler}$ erzeugt hingegen beschränkte Lösungen. Diese Eigenschaft entspricht $|\lambda h - 1| > 1$ und ist für hinreichend große Schrittweite stets erfüllt. Große Schrittweiten können also auch bei dem impliziten Euler-Verfahren zu physikalisch unsinnigen Lösungen führen.

4.4 Stabilität bei linearen Systemen

Nun wollen wir den allgemeineren Fall eines linearen Systems betrachten. Es wird sich herausstellen, dass man sich im Fall von diagonalisierbaren Systemen auf den skalaren Fall zurückziehen kann. Sei u die Lösung von

$$\begin{aligned} u'(t) &= Au(t) \quad t \geq 0\,, \\ u(0) &= 1\,, \end{aligned}$$

mit einer diagonalisierbaren Matrix $A \in \mathbb{C}^{n\times n}$. Sei $Q \in Gl(n,\mathbb{C})$, so dass $D := QAQ^{-1}$ eine Diagonalmatrix ist, bestehend aus den Eigenwerten $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ von A. Hier bezeichnet man zu einem ESV Φ eine Stabilitätsfunktion $G : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ als eine rationale Funktion, so dass

$$u_n^h \quad = \quad G(hA)u_{n-1}^h\,.$$

Wir nehmen zudem an, dass sich die Stabilitätsfunktion schreiben läßt in der Form

$$G(A) \quad = \quad \sum_{j=-s}^{s} \alpha_j A^j\,,$$

mit skalaren Koeffizienten $\alpha_i \in \mathbb{C}$. Für solche G folgt, dass $G(B) = QG(A)Q^{-1}$ gilt für ähnliche Matrizen A, B mit $B = QAQ^{-1}$, sowie

$$G\left(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}\right) = \begin{pmatrix} g(\lambda_1) & & \\ & \ddots & \\ & & g(\lambda_n) \end{pmatrix},$$

mit $g(z) = \sum_{j=-s}^{s} \alpha_j z^j$. Sei nun $y_n^h := Qu_n^h$. Dann folgt

$$\begin{aligned} y_n^h &= Qu_n^h = QG(hA)Q^{-1}y_{n-1}^h = G(hQAQ^{-1})y_{n-1}^h \\ &= G(hD)y_{n-1}^h = G(hD)^n y_0^h \,. \end{aligned}$$

Komponentenweise führt dies auf:

$$y_{n,i}^h = g(h\lambda_i)^n y_{0,i}^h \,.$$

Für lineare Systeme mit diagonalisierbaren Matrizen genügt es daher, Stabilitätsbetrachtungen für die skalare Dahlquist'sche Testgleichung durchzuführen.

4.5 Starke A-Stabilität, L-Stabilität und B-Stabilität

Es gibt noch zwei stärkere Formen als A-Stabilität, die sich aber ebenso auf das lineare Dahlquist'sche Problem beziehen:

Definition 4.8 *Ein A-stabiles ESV Φ heißt* stark A-stabil*, wenn für dessen Stabilitätsfunktion g gilt:*

$$\lim_{Re\, z \to -\infty} |g(z)| < 1 \,.$$

Ein stark A-stabiles ESV Φ heißt L-stabil*, wenn sogar*

$$\lim_{Re\, z \to -\infty} |g(z)| = 0, \,. \tag{4.3}$$

Die Trapezregel ist offensichtlich nicht stark A-stabil, denn für $g(z) = (2+z)/(2-z)$ existiert eine Folge (z_n) mit $\lim Re\, z_n = -\infty$ und $\lim |g(z_n)| = 1$. Die Folge $z_n := -n + in$ leistet dies beispielsweise.

Eine Verallgemeinerung auf eine größere Klasse von Problemen ist die B-Stabilität nach Burrage und Butcher (1979):

Definition 4.9 *Ein ESV Φ heißt* B-stabil*, wenn für beliebige dissipative*[1] *AWA gilt:*

$$\|u_1^h - v_1^h\| < \|u_0^h - v_0^h\| \,.$$

Hierbei bezeichnen u_1^h und v_1^h die Lösungen eines Zeitschritts unter Φ mit Anfangsbedingungen u_0^h und v_0^h.

[1] Hierbei bezieht sich der Ausdruck "dissipativ" auf die Def. 3.14 aber mit der Verallgemeinerung auf komplexe $\lambda \in \mathbb{C}$ mit $Re\, \lambda < 0$.

Lemma 4.10 *B-stabile ESV sind auch A-stabil.*

Beweis. Man betrachtet die spezielle AWA mit $f(t,x) := \lambda x$ mit $Re\,\lambda < 0$. Dieses AWA ist dissipativ, denn

$$\langle f(t,x) - f(t,y), \overline{x-y} \rangle \;=\; \lambda \|x-y\|^2 .$$

Setzen wir $v_0^h = 0$ so gilt auch $v_1^h = 0$. Mit $x = u_0^h$ und $y = u_1^h$ folgt nun aufgrund der B-Stabilität

$$\|y\| \;=\; \|g(h\lambda)x\| \;=\; \|g(h\lambda)x - g(h\lambda)v_0^h\| \;<\; \|x - v_0^h\| \;=\; \|x\| .$$

Also gilt für die Stabilitätsfunktion g von Φ, $|g(z)| < 1$ für $Re\,z < 0$. □

Satz 4.11 *Ein RK(s)-Verfahren ist B-stabil, wenn alle $r_i \geq 0$ $(1 \leq i \leq s)$ und die Matrix $M = (m_{ij})$ mit $m_{ij} = r_i\gamma_{ij} + r_j\gamma_{ji} - r_i r_j$ positiv semi-definit ist.*

Beweis. Für einen Beweis verweisen wir auf die Originalarbeit [2] oder das Buch [4]. □
Für die Trapezregel und das Gauß-Verfahren der Stufe $s = 2$ ergeben sich die Matrizen

$$M_{Trapez} \;=\; \frac{1}{4}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad M_{Gauss(2)} \;=\; \frac{1}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Da M_{Trapez} indefinit und $M_{Gauss(2)}$ positiv definit ist, ist die Trapezregel nicht B-stabil, das Gauß-Verfahren ist jedoch B-stabil.

4.6 Steife Differentialgleichungen

Bei skalaren Differentialgleichungen ist eine Schrittweitenbeschränkung an h zur Erreichung der numerischen Stabilität nicht wirklich dramatisch, da man aus Genauigkeitsgründen häufig eine kleine Schrittweite wählen muß. Kritischer wird es hingegen im Fall von Systemen, denn hier kann es vorkommen, dass einige Komponenten sehr schnell abklingen während andere ein langsames Abklingen (oder sogar Wachstum) zeigen. Die schnell abklingenden Komponenten erzwingen bei der Wahl von nicht A-stabilen Verfahren eine extrem kleine Schrittweite. Dies mag dann dazu führen, dass der langsamen Entwicklung anderer Komponenten nicht Rechnung getragen wird. Um dies präziser zu fassen, formulieren wir den Begriff der Steifheit bei Differentialgleichungen.

Definition 4.12 *Bei einer AWA (2.1) mit Lösung u und Jacobimatrix*

$$J(t) \;:=\; \frac{\partial}{\partial x} f(t,x)|_{x=u(t)} ,$$

versteht man unter der Steifigkeitsrate $\kappa(t)$ *die Größe*

$$\kappa(t) \;:=\; \frac{\max_{\lambda\in\sigma(t)} |Re\,\lambda|}{\min_{\lambda\in\sigma(t)} |Re\,\lambda|}$$

wobei $\sigma(t) := \{\lambda \in \mathbb{C} : \lambda$ *ist Eigenwerte von* $J(t)$ *und* $Re\,\lambda(t) < 0\}$ *das Spektrum von* $J(t)$ *mit negativem Realteil ist.*

Für die Steifigkeitsrate sind demnach nur die Realteile der Eigenwerte ausschlaggebend. Der Imaginärteil führt zu einem oszilatorischen Verhalten der Lösung und ist für das Wachstumsverhalten unbedeutend. Ferner sind nur die Eigenwerte relevant, deren Realteil negativ ist, denn bei positivem Realteil hat man es eh mit exponentiell wachsendem Verhalten zu tun, so dass der Zeitschritt schon aus Gründen der Genauigkeit klein sollte.

Definition 4.13 *Man spricht von einer* steifen *Differentialgleichung, wenn die Steifigkeitsrate groß ist, also* $\kappa(t) \gg 1$.

Merke: Der Begriff "Steifheit" macht nur bei Systemen von Differentialgleichungen Sinn. Skalare Gleichungen sind per se niemals steif. Ob eine Differentialgleichung steif ist hängt bei nichtlinearen Problemen u.a. von der Lösung selbst ab.

4.6.1 Beispiel: Steifheit bei gewöhnlichen Differentialgleichungen

Als Illustration betrachten wir das autonome lineare System

$$\begin{aligned} u'(t) &= Au(t) + b \qquad t \geq 0\,, \\ u(0) &= u_0\,, \end{aligned}$$

mit einem Vektor $b \in \mathbb{R}^3$, der 3×3-Matrix

$$A \;=\; \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{pmatrix}$$

und den Anfangsbedingungen $u_0 = (1, 0, -1)^T$. Die Eigenwerte der Matrix A lauten $\lambda_1 = -2$, $\lambda_{2,3} = -40 \pm 40i$ und besitzen somit alle negative Realteile. Für die Steifigkeitsrate gilt hier also $\kappa(t) = 20$. Die Lösung des homogenen Systems, also $b \equiv 0$, bezeichnen wir mit v. Diese ist in Abb. 4.2 abgebildet und lautet

$$\begin{aligned} v_1(t) &= \frac{1}{2}\left(e^{-2t} + e^{-40t}(\cos(40t) + \sin(40t)\right)\,, \\ v_2(t) &= \frac{1}{2}\left(e^{-2t} - e^{-40t}(\cos(40t) + \sin(40t)\right)\,, \\ v_3(t) &= -e^{-40t}(\cos(40t) - \sin(40t)\,. \end{aligned}$$

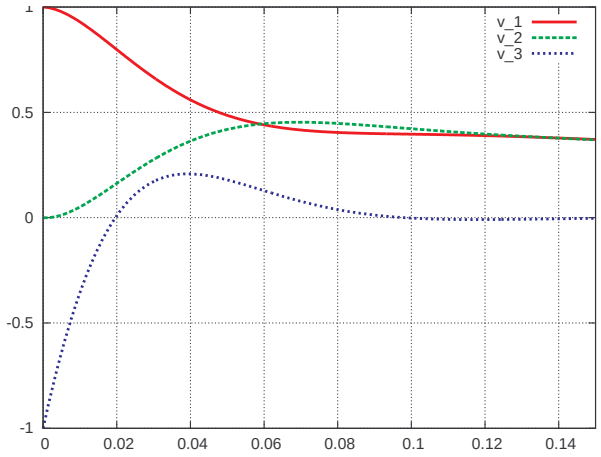

Abbildung 4.2: Lösung der homogenen Differentialgleichung aus Abschnitt 4.6.1.

Die dritte Komponente v_3 relaxiert also sehr viel schneller gegen 0 als die anderen beiden Komponenten.

Im inhomogenen Fall $b \neq 0$ und im Falle der Konvergenz $\lim_{t\to\infty} u'(t) = 0$ sei $u_\infty := \lim_{t\to\infty} u(t)$. Es gilt somit

$$Au_\infty \quad = \quad -b\,.$$

Daher lautet die Lösung des inhomogenen Systems $u(t) = v(t) + u_\infty$, denn

$$u'(t) \quad = \quad v'(t) \quad = \quad Av(t) \quad = \quad Au(t) - Au_\infty \quad = \quad Au(t) + b\,.$$

4.6.2 Beispiel: Steifheit bei partiellen Differentialgleichungen

Bei der Diskretisierung von partiellen Differentialgleichungen treten häufig steife gewöhnliche Differentialgleichungen auf. Ein Beispiel ist die Wärmeleitungsgleichung. Diese lautet im eindimensionalen Gebiet $\Omega := (0,1)$ und Zeitintervall $I = [0,T]$ zusammen mit sogenannten homogenen Dirichlet-Randbedingungen und Anfangsbedingungen:

$$\begin{aligned} \frac{\partial}{\partial t}u(t,x) - \frac{\partial^2}{\partial x^2}u(t,x) &= g(t,x) && \forall x \in \Omega\,, \forall t \in I\,, \\ u(t,0) = u(t,1) &= 0 && \forall t \in I\,, \\ u(0,x) &= u_0 && \forall x \in \Omega\,. \end{aligned}$$

Die Verwendung von zentralen Differenzenquotienten zur Approximation der zweiten räumlichen Ableitung auf einem Gitter der Weite $h = 2^{-l}$,

$$\frac{\partial^2}{\partial x^2}u(t,x) \quad \approx \quad \frac{1}{h^2}(u(t,x-h) - 2u(t,x) + u(t,x+h))$$

führt auf folgendes System gewöhnlicher Differentialgleichungen:

$$\begin{aligned} v'(t) &= Av(t) + f(t) \qquad \forall t \in I\,, \\ v(0) &= v_0 \end{aligned}$$

mit der Tridiagonalmatrix $A \in \mathbb{R}^{n\times n}$ mit $n = h^{-1} - 1 = 2^l - 1$ und

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

Hier ist die $k-$te Komponente v_k eine Approximation an u am k-ten Gitterpunkt, $v_k(t) \approx u(t, kh)$. Diese Matrix ist symmetrisch und besitzt somit nur reelle Eigenwerte. Diese lauten

$$\lambda_k = -\left(\frac{\sin(k\pi h/2)}{h/2}\right)^2 \qquad k = 1, \ldots, n\,.$$

Der minimale ($k = 1$) und maximale Eigenwert ($k = n$) ergeben sich daher zu

$$\begin{aligned} \lambda_{min} &= \lambda_1 = -\left(\frac{\sin(\pi h/2)}{h/2}\right)^2, \\ \lambda_{max} &= \lambda_n = -\left(\frac{\sin((1-h)\pi/2)}{h/2}\right)^2. \end{aligned}$$

Die Steifigkeitsrate ist also gegeben durch

$$\kappa = \frac{|\lambda_{max}|}{|\lambda_{min}|} = \left(\frac{\sin((1-h)\pi/2)}{\sin(\pi h/2)}\right)^2.$$

Für kleine räumliche Gitterweiten $h \ll 1$ gelten die Näherungen:

$$\begin{aligned} \sin((1-h)\pi/2 &\approx \sin(\pi/2) = 1\,, \\ \sin(\pi h/2) &\approx \pi h/2\,. \end{aligned}$$

Dann gilt die Näherung

$$\kappa \approx \left(\frac{2}{\pi h}\right)^2 = \mathcal{O}(h^{-2})\,.$$

Die Differentialgleichung wird demzufolge mit kleiner werdender räumlicher Gitterweite h stets steifer. Bei expliziten Verfahren müßte der Zeitschritt (hier bezeichnet mit Δt) dann an die Gitterweite h gekoppelt werden. Wie wir im vorherigen Abschnitt gesehen haben,

muss beispielsweise für das explizite Euler-Verfahren $\Delta t \leq 2/|\lambda|$ gelten. Da dies für jeden Eigenwert gelten muss, ergibt sich die stärkste Restriktion bei dem betragsmäßig größten Eigenwert, also

$$\Delta t \quad \leq \quad \frac{2}{|\lambda_{max}|} \quad = \quad \frac{1}{2}h^2 \,.$$

Aus diesem Grund kommen für diese Klasse von Problemen nur implizite Zeitschrittverfahren in Betracht.

Es sei an dieser Stelle noch erwähnt, dass es in der Literatur nicht nur diesen Steifheitsbegriff gibt sondern noch zahlreich andere. Beispielsweise spricht man auch von steifen AWA, wenn bei der Verwendung eines expliziten Euler-Verfahrens (stellvertretend für alle expliziten RK-Methoden) aus Stabilitätsgründen eine sehr viel kleinere Schrittweite gewählt werden muss, als es aus Genauigkeitsgründen erforderlich wäre.

4.7 Differential-algebraische Gleichungen

Die allgemeine Form einer differential-algebraischen Gleichung (DAE) lautet

$$F(t, u, u'(t)) \quad = \quad 0 \qquad t \in I \,, \tag{4.4}$$

$$u(t_0) \quad = \quad u_0 \,. \tag{4.5}$$

Hierbei ist $F : I \times \Omega \times \mathbb{R}^n \to \mathbb{R}^n$, $(t, x, y) \mapsto F(t, x, y)$ eine lipschitz-stetige Funktion, die bzgl. y differenzierbar ist aber mit singulärer Ableitung, also

$$\det \left(\frac{\partial}{\partial y} F(t, u(t), y)\Big|_{y=u'(t)} \right) \quad = \quad 0 \,.$$

Wir betrachten zur Illustration zwei Beispiele.

Beispiel 1: Differential-algebraische Gleichungen können auftreten als Extremfälle steifer Gleichungen. Im Beispiel aus Abschnitt 4.6.1 betrachten wir die Situation, dass die letzte (schnelle) Komponente nach (kurzer) Zeit t_1 auf den stabilen Zustand $(u_\infty)_3$ relaxiert ist. Diese Komponente ist dann zeitlich konstant, so dass wir für $t \geq t_1$ folgendes System erhalten:

$$\begin{pmatrix} u_1' t) \\ u_2' t) \\ 0 \end{pmatrix} \quad = \quad Au(t) + b \qquad t \geq t_1 \,.$$

Diese Gleichung kann in etwas allgemeinerer Form auch geschrieben werden als

$$M(t, u(t))u'(t) \quad = \quad g(t, u(t)) \qquad t \in I \,,$$

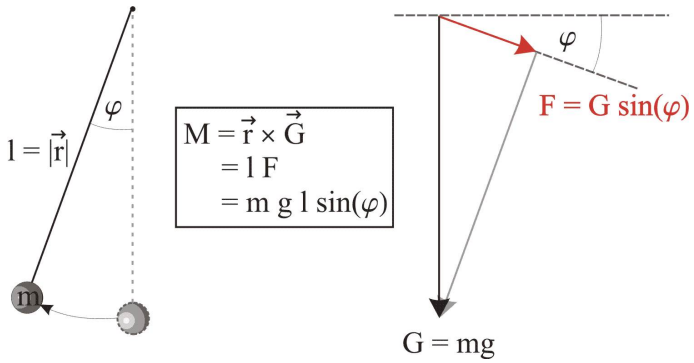

Abbildung 4.3: Ein mathematisches Pendel aus Beispiel 2 für eine DAE vom Index 3.

mit einer singulären Matrixfunktion $M : I \times \Omega \to \mathbb{R}^{n\times n}$. Im Fall einer regulären Matrixfunktion erhält man hingegen wieder die Standardform (2.1) einer AWA mit der Funktion $f(t,x) = M(t,x)^{-1}g(t,x)$.

Wir gehen nun davon aus, dass M eine konstante Matrix ist und sich diese zudem schreiben läßt in der Form

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \qquad M = \begin{pmatrix} M_{11} & 0 \\ 0 & 0 \end{pmatrix}$$

mit $M_{11} \in \mathbb{R}^{m\times m}$. Die Aufspaltung der Funktion $u = (u_1, u_2)$ mit $u_1 : I \to \mathbb{R}^m$ und $u_2 : I \to \mathbb{R}^{n-m}$ führt dann auf ein sogenanntes differential-algebraisches System (DAE)

$$\begin{aligned} M_{11}u_1' &= g_1(t, u_1, u_2) \\ 0 &= g_2(t, u_1, u_2) \end{aligned}$$

Zeitliche Ableitungen der Komponenten von u_2 fließen hierbei offensichtlich nicht ein.

Beispiel 2: Als zweites Beispiel betrachten wir ein Fadenpendel in der 2-dimensionalen Ebene (siehe Abb. 4.3). Die Position des Pendels sei beschrieben durch die Funktion $x : I \to \mathbb{R}^2$, die Geschwindigkeit durch $v : I \to \mathbb{R}^2$. Die Gravitation $g \in \mathbb{R}^2$ ist zeitlich konstant. Die Länge des Fadens wird bezeichnet mit $l > 0$. Die Rückstellkraft $F : I \to \mathbb{R}^2$ zeigt stets in Richtung der Aufhängung und deren (normalisierte) Stärke sei mit $\lambda(t) := \|F(t)\|/\|x\|$ bezeichnet, also $F(t) = -\lambda(t)x$. Die Beschreibung der Pendelbewegung wird nun durch folgende Gleichungen beschrieben:

$$x' = v\,, \tag{4.6}$$
$$v' = -\lambda x + g\,, \tag{4.7}$$
$$\|x\| = l\,. \tag{4.8}$$

Während die ersten beiden dieser drei Gleichungen normale Differentialgleichungen sind, ist die dritte eine rein algebraische Gleichung, in der die Ableitung von λ nicht auftritt.

Wir wollen nun versuchen die letzte Gleichung (4.8) in die Gestalt

$$\lambda' \;=\; f(x, v, \lambda)$$

umzuformen. Hierzu leiten wir (4.8) nach t ab und erhalten unter Nutzng von (4.7):

$$0 \;=\; \frac{d}{dt}\|x\|^2 \;=\; 2\langle x, x'\rangle \;=\; 2\langle x, v\rangle\,.$$

Diese Gleichung dividieren wir durch 2 und leiten sie nocheinmal ab:

$$0 \;=\; \frac{d}{dt}\langle x, v\rangle \;=\; \langle v, v\rangle + \langle x, v'\rangle$$

Unter Verwendung von (4.7) und (4.8) ergibt sich

$$0 \;=\; \|v\|^2 + \langle x, -\lambda x + g\rangle \;=\; \|v\|^2 - \lambda l^2 + \langle x, g\rangle$$

Ein drittes mal ableiten dieser Gleichung liefert

$$\begin{aligned} 0 \;&=\; \frac{d}{dt}\left(\|v\|^2 - \lambda l^2 + \langle x, g\rangle\right) \;=\; 2\langle v', v\rangle - \lambda' l^2 + \langle v, g\rangle \\ &=\; 2\langle -\lambda x + g, v\rangle - \lambda' l^2 + \langle v, g\rangle \\ &=\; -\lambda' l^2 + 3\langle v, g\rangle\,. \end{aligned}$$

Dies ergibt die gesuchte Gleichung

$$\lambda' \;=\; \frac{3}{l^2}\langle v, g\rangle\,,$$

so dass das Gesamtsystem geschrieben werden kann in der Form

$$\begin{pmatrix} x' \\ v' \\ \lambda' \end{pmatrix} \;=\; \begin{pmatrix} v \\ -\lambda x + g \\ \frac{3}{l^2}\langle v, g\rangle \end{pmatrix}$$

Diese System konnte mittels dreimaligen differenzierens nach t erreicht werden. Wir sprechen daher von einer DAE vom Index 3.

Definition 4.14 *Bei einer DAE der Form (4.4) bezeichnet man den* (Differentiations)-Index*, als die kleinste Zahl $k \in \mathbb{N}$, so dass durch k-maliges differenzieren der Gleichung diese umgeformt werden kann in die Form (2.1).*

Wie lassen sich nun DAEs der Form

$$\begin{aligned} u' \;&=\; f(t, u, v) \qquad t \geq 0 && (4.9) \\ 0 \;&=\; g(t, u, v) \qquad t \geq 0 && (4.10) \\ u(0) \;&=\; u_0 \end{aligned}$$

lösen ? Die algebraische Gleichung (4.10) muß in jedem Fall implizit gelöst werden, beispielsweise durch eine einfache Fixpunktiteration beginnend mit dem Startwert $v^h_{n,0} = v^h_{n-1}$ und dann für $k > 0$:

$$v^h_{n,k} \;=\; v^h_{n,k-1} + \theta g(t_n, u^h_n, v^h_{n,k-1})\,,$$

mit so gewähltem Dämpfungsparameter $\theta > 0$, dass $\|I + \theta J^g_v\| < 1$ mit der Jacobimatrix

$$J^g_v \;:=\; \frac{\partial g}{\partial v}(t_n, u^h_n, v^h_{n,k-1})\,.$$

Eine andere Möglichkeit ist eine (gedämpfte) Newton-Iteration

$$v^h_{n,k} \;=\; v^h_{n,k-1} - \theta (J^g_v)^{-1} g(t_n, u^h_n, v^h_{n,k-1})\,.$$

Dies ist für $k > 1$ solange durchzuführen bis das Residuum der Gleichung (4.10) für $v^h_{n,k}$ hinreichend klein ist, also $\|g(t_n, u^h_n, v^h_{n,k})\| < \texttt{tol}$. Bei beiden Verfahren ist aber die Kenntnis von u^h_n erforderlich. Abhängig davon, ob die Differentialgleichung (4.9) steif ist oder nicht, kann hier ein implizites oder explizites Verfahren verwendet werden. Im Fall des impliziten Euler-Verfahrens lautet das Gesamtsystem

$$\begin{aligned} u^h_n &= u^h_{n-1} + h_n f(t_n, u^h_n, v^h_n)\,, \\ 0 &= g(t_n, u^h_n, v^h_n)\,. \end{aligned}$$

Der Einsatz des Newton-Verfahrens auf das gekoppelte System erfordert die Invertierung (bzw. das Lösen mit) der Matrix

$$\begin{pmatrix} I - h_n J^f_u & -h_n J^f_v \\ J^g_u & J^g_v \end{pmatrix}.$$

Der Block $(I - h_n J^f_u)$ ist für hinreichend kleines $h_n > 0$ stets positiv definit. Für viele iterative Löser ist es aber zudem erforderlich, dass auch der Block J^g_v positiv definit ist, was nicht immer der Fall ist. In der Praxis tritt sogar der Fall $J^g_v = 0$ auf. Man spricht dann von einem Sattelpunktproblem.

Ein weiterer heikler Punkt ist auch gelegentlich die Wahl eines konsistenten Anfangswertes v_0 bzw. v^h_0, denn der ist (im Gegensatz zu u_0) nicht immer direkt gegeben, sondern muss erst durch die Gleichung (4.10) bestimmt werden:

$$0 \;=\; g(0, u_0, v_0)\,.$$

Kapitel 5

Lineare Mehrschrittverfahren

Im Gegensatz zu den Einschrittverfahren aus den vorhergehenden Abschnitten berechnet sich u_n^h nicht nur aus der Kenntnis von u_{n-1}^h sondern aus mehreren vorhergehenden Werten $u_{n-k}^h, \ldots, u_n^h$. Hierbei ist $k \geq 1$ (für $k = 1$ erhält man also ein ESV). Um die ersten k Werte $u_0^h, \ldots, u_{k-1}^h$ zu bestimmen, benötigt man in einer sogenannten *Anlaufrechnung* ein anderes Verfahren, beispielsweise ein ESV. Die nachfolgenden Werte für $n \geq k$ ergeben sich dann aus einer Vorschrift

$$u_n^h \quad = \quad \Psi(h, t_n; u_{n-k}^h, \ldots, u_n^h)$$

Die Schrittweite $h = t_n - t_{n-1}$ sei hierbei als konstant angesehen, also unabhängig von n, denn variable Schrittweiten gehen bei Mehrschrittmethoden mit weiteren technischen Schwierigkeiten einher, die wir hier nicht im Detail diskutieren wollen.

Wir werden sehen, dass sich mit der Bezeichnung $f_i := f(t_i, u_i^h)$ für die Funktionswerte, eine ganze Reihe von Verfahren in folgender Form darstellen lassen:

Definition 5.1 *Unter einem* linearen k-Schritt-Verfahren *(LMSV) zur Lösung von (2.1) mit Schrittweite $h > 0$ versteht man ein Schema der Gestalt:*

$$\sum_{j=0}^{k} \alpha_j u_{n-j}^h \quad = \quad h \sum_{j=0}^{k} \beta_j f_{n-j}\,, \tag{5.1}$$

mit Koeffizienten $\alpha_0, \ldots, \alpha_k, \beta_0, \ldots, \beta_k \in \mathbb{R}$ und $\alpha_0 \neq 0$.

Man spricht von einem linearen Mehrschrittverfahren, da die Funktionswerte f_i linear eingehen. Nichtsdestotrotz führt dies im impliziten Fall $\beta_0 \neq 0$ i.a. auf nichtlinerae Gleichungssysteme. Für $\beta_0 = 0$ sind die Verfahren offensichtlich explizit.

Die konkrete Gestalt eines linearen Mehrschrittverfahrens (LMSV) ergibt sich zu festem $\sigma \in \mathbb{N}$, $\sigma \leq k$, aus von der Darstellung

$$u(t_n) \quad = \quad u_{t_{n-\sigma}} + \int_{t_{n-\sigma}}^{t_n} f(s, u(s))\, ds\,. \tag{5.2}$$

Das auftretende Integral wird nur über eine interpolatorische Quadraturformel approximiert. Hierzu können sogar Funktionswerte für $t \leq t_{n-\sigma}$ eingehen.

5.1 Adams-Verfahren

Bei den Verfahren nach Adams wird das Integral in (5.2) nur im letzten Teilintervall ausgewertet, also $\sigma = 1$:

$$u(t_n) \;=\; u(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(s, u(s))\, ds\,.$$

Wir verwenden ein Interpolationspolynoms $p_m \in \mathbb{P}_m$ vom Grad $m \in \mathbb{N}$ an äquidistanten $m+1$ Stützstellen. Man kann hier zwischen der expliziten Variante (Adams-Bashforth) und einer impliziten Variante (Adams-Moulton) unterscheiden.

5.1.1 Adams-Bashforth-Verfahren

Bei den Adams-Bashforth-Verfahren wird das Interpolationspolynom p_m durch die bereits bekannten Werte $t_{n-k}, \ldots, t_{n-1}$ gelegt. Damit ergibt sich der Polynomgrad $m = k-1$ und das folgende Interpolationspolynom

$$p_m(t) \;:=\; \sum_{j=1}^{k} f_{n-j} L_j(t)\,,$$

mit dem Lagrange-Polynom $L_j \in \mathbb{P}_m$,

$$L_j(t) \;:=\; \prod_{i=1, i\neq j}^{k} \frac{t - t_{n-i}}{t_{n-j} - t_{n-i}}$$

Wir erhalten das diskrete Schema:

$$u_n^h \;=\; u_{n-1}^h + \sum_{j=1}^{k} f_{n-j} \int_{t_{n-1}}^{t_n} L_j(t)\, dt$$

Um dies in die Gestalt (5.1) zu bringen, ordnen wir um:

$$-u_{n-1}^h + u_n^h \;=\; h \sum_{j=1}^{k} \beta_j f_{n-j}\,.$$

mit den Koeffizienten $\alpha_0 = 1$, $\alpha_1 = -1$, $\alpha_j = 0$ für $2 \leq j \leq k$ sowie

$$\begin{aligned}\beta_j \;&=\; \frac{1}{h}\int_{t_{n-1}}^{t_n} L_j(t)\, dt \;=\; \frac{1}{h^k}\int_{t_{n-1}}^{t_n} \prod_{i=1, i\neq j}^{k} \frac{t - t_{n-i}}{i-j}\, dt \\ &=\; \frac{1}{h^k}\int_0^h \prod_{i=1, i\neq j}^{k} \frac{(i-1)h + s}{i-j}\, ds \;=\; \int_0^1 \prod_{i=1, i\neq j}^{k} \frac{i-1+t}{i-j}\, dt\,.\end{aligned}$$

Insbesondere erhält man:

- $k = 1$ (expliziter Euler):

$$u_n^h := u_{n-1}^h + h f_{n-1} .$$

- $k = 2$:

$$u_n^h := u_{n-1}^h + \frac{h}{2} \left(-f_{n-2} + 3 f_{n-1} \right) .$$

- $k = 3$:

$$u_n^h := u_{n-1}^h + \frac{h}{12} \left(5 f_{n-3} - 16 f_{n-2} + 23 f_{n-1} \right) .$$

- $k = 4$:

$$u_n^h := u_{n-1}^h + \frac{h}{24} \left(-9 f_{n-4} + 37 f_{n-3} - 59 f_{n-2} + 55 f_{n-1} \right) .$$

5.1.2 Adams-Moulton-Verfahren

Hier wählt man wieder den Polynomgrad $m = k - 1$ aber verwendet die Werte von u^h an den Stellen $t_{n-m}, \ldots, t_n$. Dies führt auf das Interpolationspolynom

$$p_m(t) := \sum_{j=0}^{k-1} f_{n-j} L_j(t) ,$$

Da f_n eingeht, erhält man implizite Verfahren:

$$-u_{n-1}^h + u_n^h = h \sum_{j=0}^{k-1} \beta_j f_{n-j} .$$

Wieder erhalten wir die Koeffizienten $\alpha_0 = 1$, $\alpha_1 = -1$, $\alpha_j = 0$ für $2 \leq j \leq k$ sowie

$$\beta_j = \int_0^1 \prod_{i=0, i \neq j}^{k-1} \frac{i - 1 + t}{i - j} \, dt , \qquad 0 \leq j \leq k - 1 .$$

Insbesondere ergeben sich:

- $k = 1$ (impliziter Euler):

$$u_n^h := u_{n-1}^h + h f_n .$$

- $k = 2$ (Trapez-Regel):

$$u_n^h := u_{n-1}^h + \frac{h}{2} \left(f_{n-1} + f_n \right) .$$

- $k = 3$:

$$u_n^h \;:=\; u_{n-1}^h + \frac{h}{12}\left(-f_{n-2} + 8f_{n-1} + 5f_n\right) .$$

- $k = 4$:

$$u_n^h \;:=\; u_{n-1}^h + \frac{h}{24}\left(f_{n-3} - 5f_{n-2} + 19f_{n-1} + 9f_n\right) .$$

Häufig wird zunächst ein Schritt mit dem expliziten Adams-Bashforth-Verfahren verwendet, um einen Prädiktor u_n^{h*} zu erhalten. Anschliessend kann beispielsweise durch eine Fixpunktiteration u_n^h iterativ berechnet werden. Mithilfe des Banachschen Fixpunktsatzes kann für hinreichend kleine Schrittweite h die Konvergenz hierfür gezeigt werden.

5.2 Nyström- und Milne-Formeln

Im Gegensatz zu den Adams-Formel kann auch über die letzten beiden Intervalle integriert werden:

$$u(t_n) \;=\; u(t_{n-2}) + \int_{t_{n-2}}^{t_n} f(s, u(s))\, ds\,.$$

Die gleichen Techniken wie zuvor führen jetzt auf die Nyström-Formeln im expliziten Fall und auf Milne-Formeln im impliziten Fall. Beispielsweise

- $k = 1$ (Nyström-Formel):

$$u_n^h \;:=\; u_{n-2}^h + 2h f_{n-1}\,.$$

- $k = 2$ (Milne-Formel / Keplersche Fassregel):

$$u_n^h \;:=\; u_{n-2}^h + \frac{h}{3}(f_{n-2} + 4f_{n-1} + f_n)\,.$$

5.3 BDF-Verfahren

Im Gegensatz zur Integration kann man auch die Differentialgleichung direkt im Interpolationspolynom verwenden. Wir bilden das Interpolationspolynom $q \in \mathbb{P}_m$, $m = k$, durch die Punkte $(t_{n-k}, u_{n-k}), \ldots, (t_n, u_n)$. Da u_n noch nicht bekannt ist, benötigt man eine weitere Bedingung. Man fordert daher, dass q die Differentialgleichung (2.1) im Punkt $t = t_n$ erfüllt:

$$q'(t_n) \;=\; f(t_n, q(t_n)) \;=\; f(t_n, u_n) \;=\; f_n\,.$$

In der Lagrange-Darstellung für q ergibt sich

$$\sum_{j=0}^{k} u_{n-j} L_j'(t_n) \quad = \quad f_n \,.$$

Dies entspricht wieder der Standardform (5.1) mit $\alpha_j = hL_j'(t_n)$ und $b_0 = 1$, $b_j = 0$ für $1 \leq j \leq k$:

$$\sum_{j=0}^{k} \alpha_j u_{n-j}^h \quad = \quad hf_n \,,$$

Per Konstruktion sind die BDF-Verfahren also implizit und damit gut geeignet für steife Differentialgleichungen.

- $k = 1$ (impliziter Euler):

$$u_n^h - u_{n-1}^h \quad = \quad hf_n \,.$$

- $k = 2$ (BDF-2):

$$\frac{3}{2} u_n^h - 2u_{n-1}^h + \frac{1}{2} u_{n-2}^h \quad = \quad hf_n \,.$$

- $k = 3$ (BDF-3):

$$\frac{11}{6} u_n^h - 3u_{n-1}^h + \frac{3}{2} u_{n-2}^h - \frac{1}{3} u_{n-3}^h \quad = \quad hf_n \,.$$

5.4 Abschneidefehler und Konsistenz bei LMSV

Definition 5.2 *Unter dem Abschneidefehler (*engl.: truncation error*) eines LMSV (5.1) mit Schrittweite $h > 0$ versteht man die Größe:*

$$\tau_n^h \quad := \quad \frac{1}{h} \sum_{j=0}^{k} \alpha_j u(t_{n-j}) - \sum_{j=0}^{k} \beta_j f(t_{n-j}, u(t_{n-j}))$$

Die Begriffe *Konsistenz* und *Konsistenzordnung* übertragen sich direkt aus denen bei ESV, siehe Definition 3.3.

Lemma 5.3 *Für den Abschneidefehler eines LMSV zu einer AWA mit lipschitz-stetigem f gilt für den Fehler $e_n := u(t_n) - u_n^h$ bei exakten "Startwerten" $u_{n-j}^h = u(t_{n-j})$ für $j \in \{1, \ldots, k\}$:*

$$\frac{\|e_n - \alpha_0^{-1} h \tau_n^h\|}{\|e_n\|} \quad = \quad \mathcal{O}(h) \,.$$

Beweis. Nach der Definition des Abschneidefehlers und aufgrund der vorausgesetzten Exaktheit an den Stellen t_{n-j} gilt:

$$\begin{aligned} h\tau_n^h &= \sum_{j=0}^{k} \alpha_j u(t_{n-j}) - h\sum_{j=0}^{k} \beta_j f(t_{n-j}, u(t_{n-j})) \\ &= \sum_{j=0}^{k} \Big(\alpha_j u_{n-j}^h - h\beta_j f(t_{n-j}, u_{n-j}^h)\Big) \\ &\quad +\alpha_0(u(t_n) - u_n^h) + h\beta_0(f(t_n, u(t_n)) - f(t_n, u_n^h)) \\ &= \alpha_0 e_n + h\beta_0(f(t_n, u(t_n)) - f(t_n, u_n^h)) \end{aligned}$$

Hieraus ergibt sich mit der Lipschitz-Stetigkeit von f und Lipschitz-Konstante L:

$$\begin{aligned} \|e_n - \alpha_0^{-1} h\tau_n^h\| &= h\frac{\beta_0}{\alpha_0}\|f(t_n, u(t_n)) - f(t_n, u_n^h)\| \\ &\leq h\frac{\beta_0}{\alpha_0} L\|e_n\| \,. \end{aligned}$$

Dies impliziert die Behauptung. □

Für LMSV sind die folgenden Koeffizienten von Interesse:

$$\begin{aligned} c_0 &= \sum_{j=0}^{k} \alpha_j \,, \\ c_i &= \sum_{j=0}^{k} \left(\frac{1}{i!}(k-j)^i \alpha_j - \frac{1}{(i-1)!}(k-j)^{i-1}\beta_j\right) \,, \qquad i \in \{1, \ldots, k\}. \end{aligned}$$

Satz 5.4 *Ein lineares k-Schritt-Verfahren (LMSV) ist genau dann konsistent mit Konsistenzordnung $p \in \mathbb{N}$, wenn für dessen Koeffizienten gilt:*

$$c_0 = \ldots = c_p = 0 \quad \textit{und} \quad c_{p+1} \neq 0 \,.$$

Beweis. Der Abschneidefehler läßt sich ausdrücken in der Form:

$$\tau_n^h := \sum_{j=0}^{k} \left(\frac{1}{h}\alpha_j u(t_{n-j}) - \beta_j u'(t_{n-j}))\right) \,.$$

Für die Konsistenzordnung können wir u als beliebig glatt voraus setzen. Wir benutzen nun die Taylor-Entwicklungen von $u(t_{n-j})$ und $u'(t_{n-j})$:

$$\begin{aligned} u(t_{n-j}) &= \sum_{i=0}^{\infty} \frac{1}{i!}((k-j)h)^i u^{(i)}(t_{n-k}) \,, \\ u'(t_{n-j}) &= \sum_{i=0}^{\infty} \frac{1}{i!}((k-j)h)^i u^{(i+1)}(t_{n-k}) \,. \end{aligned}$$

Setzen wir dies in obige Formel des Abschneidefehlers ein, so erhalten wir mit der Konvention $\frac{1}{(-1)!} := 0$:

$$\begin{aligned}\tau_n^h &:= \sum_{j=0}^{k}\sum_{i=0}^{\infty}\left(\frac{1}{h}\alpha_j\frac{1}{i!}((k-j)h)^i - \beta_j\frac{1}{(i-1)!}((k-j)h)^{i-1}\right)u^{(i)}(t_{n-k}) \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{k}\left(\alpha_j\frac{(k-j)^i}{i!} - \beta_j\frac{(k-j)^{i-1}}{(i-1)!}\right)h^{i-1}u^{(i)}(t_{n-k}) \\ &= \sum_{i=0}^{\infty} c_i h^{i-1}u^{(i)}(t_{n-k})\end{aligned}$$

Hieraus ergibt sich für $c_0 = \ldots = c_p = 0$ und $c_{p+1} \neq 0$, dass $\tau_n^h = \mathcal{O}(h^p)$. Die Gegenrichtung folgt ebenfalls. □

Für die Konsistenz genügt demzufolge die Bedingung $c_0 = c_1 = 0$ bzw.

$$\sum_{j=0}^{k}\alpha_j = 0 \quad \text{und} \quad \sum_{j=0}^{k}(k-j)\alpha_j = \sum_{j=0}^{k}\beta_j\,.$$

Dies führt auf die folgende Definition:

Definition 5.5 *Zu einem LMSV heißen*

$$\rho(x) := \sum_{j=0}^{k}\alpha_{k-j}x^j \qquad \textit{und} \quad \sigma(x) := \sum_{j=0}^{k}\beta_{k-j}x^j \tag{5.3}$$

das zugehörige erste bzw. zweite charakteristische Polynom.

Aus obiger Überlegung ergibt sich also das folgende Resultat:

Korollar 5.6 *Ein LMSV ist genau dann konsistent, wenn für die charakteristischen Polynome gilt:*

$$\rho(1) = 0 \quad \textit{und} \quad \sigma(1) = \rho'(1)\,.$$

5.5 Null-Stabilität bei LMSV

Wir werden sehen, dass bei LMSV die Konsistenz allein nicht ausreicht, um Konvergenz im obigen Sinne zu erreichen. Um eine notwendige Bedingung für konvergente LMSV herzuleiten, betrachten wir an dieser Stelle die sehr simple Differentialgleichung (2.1) mit $f \equiv 0$. Von einem konvergenten LMSV erwarten wir, dass für dieses f die diskrete Lösung u_n^h für $n \to \infty$ beschränkt bleibt, und zwar bei beliebigen Anfangswerten $u_0^h, \ldots, u_k^h$. Die Werte u_n^h ergeben sich in diesem speziellen Fall (und der Normierung $\alpha_0 = 1$) aus

$$u_n^h = -\sum_{j=1}^{k}\alpha_j u_{n-j}^h\,, \qquad n \geq k\,. \tag{5.4}$$

Definition 5.7 *Ein LMSV heißt* null-stabil *(oder auch* asymptotisch stabil *oder* D-stabil[1]*), wenn alle Lösungen $u^h_{n-k},\dots,u^h_n$ der homogenen Differenzengleichung (5.4) unabhängig von h beschränkt sind.*

Wir wollen untersuchen, ob es Lösungen der Form $u^h_n = \lambda^n$ mit $\lambda \in \mathbb{C}^*$ geben kann. Setzen wir dies ein in (5.4), so folgt

$$0 \;=\; \sum_{j=0}^{k} \alpha_j \lambda^{n-j} \;=\; \lambda^{n-k} \sum_{j=0}^{k} \alpha_{k-j} \lambda^j \;=\; \lambda^{n-k} \rho(\lambda)\,.$$

Eine Bedingung dafür, dass $u^h_n = \lambda^n$ eine Lösung ist, ist also, dass λ eine Nullstelle des ersten charakteristischen Polynoms aus (5.3) ist:

$$\rho(\lambda) \;=\; \sum_{j=0}^{k} \alpha_{k-j} \lambda^j \;=\; 0\,.$$

Satz 5.8 *Seien $\lambda_i \in \mathbb{C}$, $i \in \{1,\dots,\nu\}$ Nullstellen des ersten charakteristischen Polynoms ρ mit der jeweiligen Vielfachheit μ_i, so dass $\sum_{i=1}^{\nu} \mu_i = k$. Dann sind die k Folgen $(y_n^{(i,j)})_{n\in\mathbb{N}}$ mit $1 \le i \le \nu$ und $1 \le j \le \mu_i$, deren Glieder gegeben sind durch*

$$y_n^{(i,j)} \;:=\; \frac{n!}{(n+1-j)!} \lambda_i^n\,, \quad n \in \mathbb{N}_0\,,$$

linear unabhängige Lösungen (sogenannte Fundamentallösungen*) der Differenzengleichung (5.4).*

Auch hier ist die Fakultät von negativen Zahlen als Null definiert: $(-j)! = 0$ für $j \in \mathbb{N}$.

Beweis. (a) Zunächst ist zu überprüfen, dass hierdurch tatsächlich Lösungen gegeben sind. Sei hierzu $\lambda \in \mathbb{C}$ eine Nullstelle von ρ mit Vielfachheit μ_j. Wir untersuchen die Folge $(y_n)_{n\in\mathbb{N}}$ mit

$$y_n \;:=\; \frac{n!}{(n+1-j)!} \lambda^n\,, \quad n \in \mathbb{N}_0\,.$$

Um (5.4) zu überprüfen betrachten wir:

$$D \;:=\; \sum_{l=0}^{k} \alpha_l y_{n-l} \;=\; \sum_{l=0}^{k} \alpha_l \frac{(n-l)!}{(n-l+1-j)!} \lambda^{n-l}\,.$$

Für $j = 1$ ergibt sich:

$$D \;=\; \sum_{l=0}^{k} \alpha_l \lambda^{n-l} \;=\; \lambda^{n-k} \sum_{l=0}^{k} \alpha_l \lambda^{k-l} \;=\; \lambda^{n-k} \sum_{l=0}^{k} \alpha_{k-l} \lambda^l \;=\; \lambda^{n-k} \rho(\lambda) \;=\; 0\,.$$

[1] zu Ehren von Germund Dahlquist, 1925-2005, schwedischer Numeriker.

Für $j \geq 2$ ergibt sich:

$$\begin{aligned} D &= \sum_{l=0}^{k} \alpha_l (n-l) \cdot \ldots \cdot (n-l+2-j) \lambda^{n-l} \\ &= \lambda^{j-1} \sum_{l=0}^{k} \alpha_l (n-l) \cdot \ldots \cdot (n-l+2-j) \lambda^{n-l+1-j} \\ &= \frac{d^{\mu-1}}{d\lambda^{\mu-1}} (\lambda^{n-k} \rho(\lambda)) \,. \end{aligned}$$

Nun betrachten wir die Funktion:

$$\psi(x) \;:=\; \lambda^{n-k}\rho(x) \;=\; \sum_{l=0}^{k} \alpha_{n-l} x^{n-k+l} \;=\; \sum_{l=0}^{k} \alpha_l x^{n-l} \,.$$

Es gilt:

$$\begin{aligned} \frac{d}{dx}\psi(x) &= \sum_{l=0}^{k} \alpha_l (n-l) x^{n-l-1} \,, \\ &\vdots \\ \frac{d^{j-1}}{dx^{j-1}}\psi(x) &= \sum_{l=0}^{k} \alpha_l (n-l) \ldots (n-l+2-j) x^{n-l+1-j} \end{aligned}$$

Somit gilt:

$$D \;=\; \lambda^{j-1} \frac{d^{j-1}}{dx^{j-1}} \psi(\lambda)$$

Da $\rho^{(j-1)}(\lambda) = 0$ folgt $\psi^{(j-1)}(\lambda) = 0$ und damit $D = 0$. Also ist y_n Lösung (5.4).
(b): Lineare Unabhängigkeit: Für die lineare Unabhängigkeit betrachten wir die Matrix, die sich aus den Fundamentallösungen an den ersten k Zeitschritten bildet:

$$A \;:=\; \begin{pmatrix} y_0^{(1,1)} & \cdots & y_{k-1}^{(1,1)} \\ \vdots & & \vdots \\ y_0^{(i,j)} & \cdots & y_{k-1}^{(i,j)} \\ \vdots & & \vdots \\ y_0^{(\nu,\mu_\nu)} & \cdots & y_{k-1}^{(\nu,\mu_\nu)} \end{pmatrix} \;=\; \begin{pmatrix} \lambda_1^0 & \cdots & \lambda_1^{k-1} \\ \vdots & & \vdots \\ \frac{1}{(1-j)!}\lambda_i^0 & \cdots & \frac{(k-1)!}{(k-j)!}\lambda_i^{k-1} \\ \vdots & & \vdots \\ \frac{1}{(1-\mu_\nu)!}\lambda_\nu^0 & \cdots & \frac{(k-1)!}{(k-\mu_\nu)!}\lambda_\nu^{k-1} \end{pmatrix} .$$

Diese Matrix entspricht aber genau der, die sich aus dem folgenden hermiteschen Interpolationproblems ergibt: Finde Interpolationspolynom $q \in P_{k-1}$ mit vorgegebenen Werten für $q^{(j)}(\lambda_i)$, $0 \leq j \leq \mu_i$. Daher ist diese Matrix regulär. ∎

Satz 5.9 (Wurzelbedingung) *Ein LMSV ist genau dann null-stabil, wenn für die Nullstellen des zugehörigen ersten charakteristischen Polynoms* $N := \rho^{-1}(\{0\}) \subset \mathbb{C}$ *gilt:*

$$\begin{aligned} &(a) \qquad \lambda \in N \Longrightarrow |\lambda| \leq 1 \\ &(b) \qquad \lambda \in N \text{ und } |\lambda| = 1 \Longrightarrow \lambda \text{ einfach.} \end{aligned}$$

Beweis. (a) Wir nehmen zunächst an, dass eine Nullstelle $\lambda \in N$ existiert mit $|\lambda| > 1$. Dann ist $u_n^h = \lambda^n$ Lösung der Differenzengleichung (5.4). Nun ist für beliebiges $T > 0$ die Bedingung $0 \leq t_n \leq T$ gleichbedeutend mit $0 \leq n \leq Th^{-1}$. Damit folgt aber

$$\lim_{h\to 0} \max_{0\leq t_n\leq T} |u_n^h| \;=\; \lim_{h\to 0} \max_{0\leq n\leq Th^{-1}} |\lambda^n| \;=\; \lim_{n\to\infty} |\lambda|^n \;=\; \infty\,.$$

Also ist die Lösung nicht beschränkt und damit das LSMV nicht null-stabil.
(b) Im Fall eines Eigenwertes $\lambda \in N$ mit $|\lambda| = 1$ und Vielfachheit größer als 1, gilt $\rho(\lambda) = \rho'(\lambda) = 0$. Dann ist aber auch $v_n^h = n\lambda^n$ eine diskrete Lösung von (5.4), denn

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j v_{n-j}^h &= \sum_{j=0}^{k} \alpha_j (n-j)\lambda^{n-j} \\ &= n\lambda^{n-k} \sum_{j=0}^{k} \alpha_j \lambda^{k-j} - \lambda^{n-k+1} \sum_{j=0}^{k} \alpha_j j \lambda^{k-j-1} \\ &= n\lambda^{n-k}\rho(\lambda) - \lambda^{n-k+1}\rho'(\lambda) \\ &= 0\,. \end{aligned}$$

Für diese Lösung gilt nun ebenfalls

$$\lim_{h\to 0} \max_{0\leq t_n\leq T} |v_n^h| \;=\; \lim_{h\to 0} \max_{0\leq n\leq Th^{-1}} n|\lambda|^n \;=\; \lim_{n\to\infty} n \;=\; \infty\,.$$

Also ist die Lösung nicht beschränkt und das LSMV nicht null-stabil.
(c) Die umgekehrte Richtung folgt aus Satz 5.8. Jede Lösung läßt sich durch die k unabhängigen Lösungen linear kombinieren. All diese Lösungen sind monoton fallend für $|\lambda| < 1$ und n hinreichend groß, oder aber im Fall $|\lambda| = 1$ aufgrund der Vielfachheit 1 von der Form $y_n^{(i,1)} = \lambda^n$ und damit beschränkt. ∎

Satz 5.10 *Die Verfahren von Adams-Bashforth, Adams-Moulton, Nyström, Milne-Verfahren sowie die BDF(k)-Verfahren für* $1 \leq k \leq 6$ *sind null-stabil.*

Beweis. Für die Verfahren von Adams-Bashforth und Adams-Moulton gilt $\alpha_0 = 1$ und $\alpha_1 = -1$. Daher erhalten wir das folgenden ersten charakteristischen Polynom:

$$\rho_{Adams}(\lambda) \;=\; \lambda^k - \lambda^{k-1} \;=\; \lambda^{k-1}(\lambda - 1)\,.$$

Die Nullstellen sind also $\lambda = 0$ (mit Vielfachheit $k-1$) und die einfache Nullstelle $\lambda = 1$. Damit gilt Null-Stabilität nach Satz 5.9. Für die Nyström- und Milne-Formel gilt $\alpha_0 = 1$, $\alpha_1 = 0$ und $\alpha_2 = -1$. Folglich gilt:

$$\rho_{Nys/Mil}(\lambda) \;=\; \lambda^k - \lambda^{k-2} \;=\; \lambda^{k-2}(\lambda^2 - 1)\,.$$

Die Nullstellen sind also $\lambda = 0$ (mit Vielfachheit $k-2$) und die einfachen Nullstelle $\lambda = \pm 1$. Damit gilt Null-Stabilität nach Satz 5.9. Für BDF(2) mit der Skalierung $\alpha_0 = 1$ gilt:

$$\rho_{BDF(2)}(\lambda) \;=\; \lambda^2 - \frac{4}{3}\lambda + \frac{1}{3}\,.$$

Die Nullstellen ergeben sich zu $\lambda = 1$ und $\lambda = \frac{1}{3}$. Die BDF-Verfahren für $3 \le k \le 6$ lassen wir als Übungaufgabe. □

5.6 Konvergenz

Wir führen jetzt den Konvergenzbegriff für LMSV ein:

Definition 5.11 *Ein k-Schritt-Verfahren heißt* konvergent*, wenn für AWA (2.1) mit lipschitzstetigem f aus der Konvergenz der Startwerte*

$$\lim_{h\to 0} \max_{0\le i\le k} \|u(t_0) - u_i^h\| \;=\; 0$$

auch die Konvergenz für alle beschränkten Zeitintervalle $[t_0, t_0 + T]$ folgt:

$$\lim_{h\to 0} \max_{t_0\le t_i\le t_0+T} \|u(t_i) - u_i^h\| \;=\; 0\,.$$

Auch hier ist komplexe Arithmetik zugelassen.

Satz 5.12 (Dahlquist) *Eine LMSV ist genau dann konvergent, wenn es null-stabil und konsistent ist. Die Konvergenzordnung entspricht dann der Konsistenzordnung.*

Beweis. Der Beweis erfolgt in drei Teilen:
(a) Konvergenz ⇒ Null-Stabilität: Wir nehmen an, das LMSV wäre nicht null-stabil und betrachten die AWA mit $f \equiv 0$ und Anfangswerten $u_0 = 0$. Wir hatten im Beweis von Satz 5.9 gesehen, dass dann diskrete Lösungen u_n^h existieren, die nicht beschränkt sind. Insbesondere können diese für $h \to 0$ auch nicht gegen die exakte Lösung $u \equiv 0$ konvergieren. Wichtig ist hierbei allerdings, dass wir die Lösungen mit h skalieren, also $u_j^h = h\lambda^j$ für $|\lambda| > 1$ bzw. $u_j^h = hj\lambda^j$für $|\lambda| = 1$. Für $h \to 0$ und $0 \le j \le k$ konvergieren diese Startwerte gegen $u_0 = 0$.
(b) Konvergenz ⇒ Konsistenz: Man zeigt die Kriterien aus Korollar 5.6, also $\rho(1) = 0$ und $\sigma(1) = \rho'(1)$. Hierzu betrachten wir wieder $f \equiv 0$ aber mit den Anfangsbedingungen

$u_0 = 1$. Die Startwerte wählen wir unabhängig von h, nämlich $u_j^h = 1$ für $0 \le j \le k$. Die diskrete Lösung ist damit unabhängig von h und aufgrund der Konvergenz gilt:

$$u_n^h \;=\; 1 \quad \forall n \in \mathbb{N},.$$

Mit der Differenzengleichung (5.4) folgt:

$$\alpha_0 \;=\; -\sum_{j=1}^{k} \alpha_j \,.$$

Hieraus folgt unmittelbar $\rho(1) = 0$. Für die zweite Bedingung betrachten wir $f \equiv 1$ und $u_0 = 0$. Die Differenzengleichung für das LMSV lautet nun

$$\sum_{j=0}^{k} \alpha_j u_{n-j}^h \;=\; h \sum_{j=0}^{k} \beta_j \;=\; h\sigma(1) \,.$$

Andererseits ist $\rho'(1) = \sum_{j=0}^{k} \alpha_j (k-j)$. Als Startwerte setzen wir $u_j^h = hj\sigma(1)/\rho'(1)$, $0 \le j \le k$. Diese Startwerte sind für $h \to 0$ konsistent mit $u_0 = 0$. Man prüft durch Induktion leicht nach, dass die folgenden Werte ($j > k$) ebenfalls gegeben sind durch $u_j^h = hj\sigma(1)/\rho'(1)$. Aufgrund der Konvergenz folgt nun für festes $t_i \in [0,T]$, also $i(h) = t_i/h$:

$$t_i \;=\; u(t_i) \;=\; \lim_{h\to 0} u_{i(h)}^h \;=\; \lim_{h\to 0} hi(h)\sigma(1)/\rho'(1) \;=\; \frac{\sigma(1)}{\rho'(1)} \lim_{h\to 0} t_i \,.$$

Es folgt $\sigma(1) = \rho'(1)$ und damit die Konsistenz.
(c) Konsistenz+Null-Stabilität ⇒ Konvergenz: Wir betrachten die allgemeine AWA (2.1). Wir fassen die k Werte $u_n^h, \ldots, u_{n-k+1}^h$ jeweils in einem Vektor $\mathbf{y}_n$ zusammen, $\mathbf{y}_n := (u_n^h, \ldots, u_{n-k+1}^h)^T$. Dann gilt (mit der Normierung $\alpha_0 = 1$):

$$\mathbf{y}_n \;=\; A\mathbf{y}_{n-1} + h\gamma_n e_1$$

mit der *Übergangs-Matrix*, der skalaren Größe

$$A \;:=\; \begin{pmatrix} -\alpha_1 & \ldots & \ldots & -\alpha_k \\ 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 1 & 0 \end{pmatrix}, \quad \gamma_n \;:= \sum_{j=0}^{k} \beta_j f_{n-j}$$

und dem Vektor $e_1 = (1, 0 \ldots, 0)^T$. Für den Vektor $\mathbf{w}_n$ mit den exakten Werten $(\mathbf{w}_n)_j := u(t_j)$ gilt entsprechend

$$\mathbf{w}_n \;=\; A\mathbf{w}_{n-1} + h\widetilde{\gamma}_n e_1 \,,$$

mit

$$\widetilde{\gamma}_n \;:=\; \sum_{j=0}^{k} \beta_j f(t_{n-j}, u(t_{n-j})) + \tau_n^h \,.$$

Für den Vektor mit den Diskretisierungsfehlern, $\mathbf{e}_n := \mathbf{w}_n - \mathbf{y}_n$, gilt daher

$$\mathbf{e}_n \;=\; A\mathbf{e}_{n-1} + h(\widetilde{\gamma}_n - \gamma_n)e_1 \,.$$

Somit erhalten wir zunächst für eine beliebige Matrixnorm $\|\!|\cdot|\!\|$ und verträglicher Vektornorm $\|\!|\cdot|\!\|$ die Abschätzung:

$$\|\!|\mathbf{e}_n|\!\| \;\leq\; \|\!|A|\!\|\,\|\!|\mathbf{e}_{n-1}|\!\| + h|\widetilde{\gamma}_n - \gamma_n|\epsilon \,.$$

Hier haben wir die Bezeichnung $\epsilon := \|\!|e_1|\!\|$ verwendet. Um $\|\!|A|\!\|$ zu beschränken, bestimmen wir die Eigenwerte von A. Es gibt nämlich für beliebiges $\delta > 0$ stets eine Matrixnorm mit

$$\varrho(A) \;\leq\; \|\!|A|\!\| \;\leq\; \varrho(A) + \delta \,.$$

Im Fall von $\varrho(A) = 1$ und einfacher Vielfachheit der betragsmäßig größten Eigenwerte gilt dies sogar für $\delta = 0$. Hierzu suchen wir die Nullstellen des charakteristische Polynoms p_A:

$$p_A(\lambda) \;=\; \det(A - \lambda I) \;=\; \det \begin{pmatrix} -\alpha_1 - \lambda & \dots & \dots & -\alpha_k \\ 1 & -\lambda & \dots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 1 & -\lambda \end{pmatrix} .$$

Durch Entwicklung nach der ersten Zeile ergibt sich wegen $\alpha_0 = 1$:

$$\begin{aligned} p_A(\lambda) &= -(\alpha_1 + \lambda)(-\lambda)^{k-1} + \sum_{j=2}^{k} (-1)^j \alpha_j (-\lambda)^{k-j} \\ &= (-\lambda)^k + (-1)^k \alpha_1 \lambda^{k-1} + (-1)^k \sum_{j=2}^{k} \alpha_j \lambda^{k-j} \\ &= (-1)^k \sum_{j=0}^{k} \alpha_j \lambda^{k-j} \\ &= (-1)^k \rho(\lambda) \,. \end{aligned}$$

Also sind die Eigenwerte von A gerade die Nullstellen des ersten charakteristischen Polynoms ρ. Diese sind nach Voraussetzung stets $|\lambda| \leq 1$ und die Nullstellen mit $|\lambda| = 1$ sind einfach. Man kann zeigen, dass eine natürliche Matrizennorm $\|\!|\cdot|\!\|$ existiert, so dass $\|\!|A|\!\| = \varrho(A) \leq 1$. Wir bezeichnen die zugehörige Vektornorm ebenfalls mit $\|\!|\cdot|\!\|$. Dann folgt:

$$\|\!|\mathbf{e}_n|\!\| \;\leq\; \|\!|\mathbf{e}_{n-1}|\!\| + h|\widetilde{\gamma}_n - \gamma_n|\epsilon \,.$$

Wir schätzen jetzt noch $|\widetilde{\gamma}_n - \gamma_n|$ mithilfe der Lipschitz-Stetigkeit von f ab:

$$\begin{aligned}
\epsilon|\widetilde{\gamma}_n - \gamma_n| &\leq \epsilon \sum_{j=0}^{k} |\beta_j||f(t_{n-j}, \mathbf{w}_n) - f(t_{n-j}, u(t_{n-j}))| + \epsilon|\tau_n^h| \\
&\leq \epsilon L_f \sum_{j=0}^{k} |\beta_j||u_{n-j}^h - u(t_{n-j})| + \epsilon|\tau_n^h| \\
&\leq c(|||\mathbf{w}_{n-1}||| + |||\mathbf{w}_n|||) + \epsilon|\tau_n^h|\,,
\end{aligned}$$

mit einer von den Parametern β_j, ϵ und L_f abhängigen Konstanten c. Insgesamt erhalten wir somit

$$(1 - hc)|||\mathbf{e}_n||| \leq (1 + hc)\,|||\mathbf{e}_{n-1}||| + h\epsilon|\tau_n^h|\,.$$

Für $0 < h < c^{-1}$ dividieren wir beide Seiten durch $1 - hc$ und erhalten

$$|||\mathbf{e}_n||| \leq a + b|||\mathbf{e}_{n-1}|||\,,$$

mit

$$\begin{aligned}
a &:= \max_{0\leq n\leq N} \left(\frac{h\epsilon}{1 - hc}|\tau_n^h| \right) = \mathcal{O}(h^{p+1})\,, \\
b &:= \frac{1 + hc}{1 - hc}\,,
\end{aligned}$$

wobei p die Konsistenzordnung des Verfahrens bezeichnet. Das diskrete Gronwall-Lemma in der Version von Satz 2.19 liefert nun

$$|||\mathbf{e}_n||| \leq e^{n(b-1)}|||\mathbf{w}_0||| + \frac{e^{n(b-1)} - 1}{b - 1} a\,.$$

Nun untersuchen wir den Grenzübergang $h \to 0$. Der erste Faktor ist wegen $n = Th^{-1}$ und für $h < 1/(2c)$ beschränkt:

$$e^{n(b-1)} \leq \exp\left(\frac{2cT}{1 - hc} \right) \leq \exp(4cT)\,.$$

Die Startwerte verhalten sich wie die Konsistenzordnung, also $|||\mathbf{w}_0||| = \mathcal{O}(h^p)$.

$$\frac{e^{n(b-1)} - 1}{b - 1} = \frac{(1 - hc)(e^{n(b-1)} - 1)}{2hc} \leq \frac{e^{n(b-1)} - 1}{2hc}\,.$$

Da wie oben gezeigt $n(b - 1) \leq 4Tc$ erhält man insgesamt

$$|||\mathbf{e}_n||| = c_1\mathcal{O}(h^p) + c_2h^{-1}\mathcal{O}(h^{p+1}) = \mathcal{O}(h^p)\,.$$

Hiermit ist die Konvergenz und die Konvergenzordnung p gezeigt. ∎

Korollar 5.13 *Die k-stufigen Verfahren von Adams-Bashforth, Adams-Moulton, Nyström, Milne-Verfahren sowie für* $1 \leq k \leq 6$ *sind die BDF(k)-Verfahren konvergent. Deren Konvergenzordnungen sind k für Adams-Bashforth, Nyström, und BDF-Verfahren. Das k-stufige Adams-Moulton und Milne-Verfahren besitzen Konvergenzordnung* $k+1$.

Beweis. Die Konvergenz ergibt sich mit vorherigem Satz aus der Konsistenz und der Null-Stabilität. Die Konvergenzordnung entspricht gerade der Konsistenzordnung. Diese jeweils über das Kriterium von Satz 5.4 nachzuweisen, überlassen wir als Übungsaufgabe. □

Satz 5.14 *Ein null-stabiles lineares k-Schritt-Verfahren besitzt die* maximale *Konvergenzordnung* $k+1$*, falls k ungerade, und* $k+2$ *falls k gerade ist.*

Beweis. Siehe Dahlquist. □

5.7 A-Stabilität bei linearen Mehrschrittverfahren

Bei der A-Stabilität von Einschrittverfahren haben wir für $q \in \mathbb{C}$ das Dahlquist'sche Testproblem betrachtet:

$$\begin{aligned} u'(t) &= qu(t) \quad t \geq 0\,, \\ u(0) &= 1\,. \end{aligned}$$

Bei LMSV sind die diskreten Lösungen hierzu gegeben durch folgende Differenzengleichung:

$$\sum_{j=0}^{k} \alpha_j u_{n-j}^h = qh \sum_{j=0}^{k} \beta_j u_{n-j}^h\,,$$

bzw.

$$\sum_{j=0}^{k} (\alpha_j - qh\beta_j)\, u_{n-j}^h = 0\,. \tag{5.5}$$

Lösungen hiervon liefert uns das folgende Lemma.

Lemma 5.15 *Sei* Φ *ein LMSV und* $\lambda \in \mathbb{C}$ *Nullstelle des* Stabilitätspolynoms

$$p_\Phi(z; qh) := \rho(z) - qh\sigma(z)\,.$$

Dann ist $u_n^h = \lambda^n$ *eine Lösung der Differenzengleichung (5.5). Im Fall einer Vielfachheit* ≥ 2 *ist auch* $u_n^h = n\lambda^n$ *eine Lösung von (5.5).*

Beweis. Der Ansatz $u_n^h = \lambda^n$ liefert eine Lösung der Differenzengleichung (5.5) genau dann, wenn

$$0 \;=\; \sum_{j=0}^{k} (\alpha_j - hq\beta_j)\,\lambda^{n-j} \;=\; \lambda^{n-k} \sum_{j=0}^{k} \Big(\alpha_j \lambda^{k-j} - qh\beta_j \lambda^{k-j}\Big) \;=\; \lambda^{n-k} p_\Phi(\lambda; qh)\,.$$

Für $\lambda = 0$ ist die Differenzengleichung also stets erfüllt. Für $\lambda \neq 0$ ist notwendigerweise λ eine Nullstelle von $p_\Phi(\cdot; qh)$. Und ungekehrt liefern Nullstellen von $p_\Phi(\cdot; qh)$ Lösungen der besagten Differenzengleichung. Der Fall einer Vielfachheit ≥ 2 ergibt $u_n^h = n\lambda^n$ Lösungen gemäß der Argumente im Beweis der Wurzelbedingung (Satz 5.9). ∎

Definition 5.16 *Ein LMSV Φ heißt* stabil *für $qh \in \mathbb{C}$, wenn für jede Nullstelle $\lambda \in \mathbb{C}$ des Stabilitätspolynoms $p_\Phi(\cdot; qh)$ gilt:*

$$|\lambda| < 1 \quad \textit{oder} \quad (|\lambda| = 1 \textit{ und } \lambda \textit{ einfach}).$$

Unter dem Stabilitätsgebiet *versteht man*

$$S_\Phi \;:=\; \{z \in \mathbb{C} \;:\; \Phi \textit{ ist stabil für } z\}\,.$$

Φ heisst A-stabil *(oder* absolut stabil*), wenn $\{z \in \mathbb{C} : Re(z) < 0\} \subseteq S_\Phi$.*

Im Gegensatz zu ESV muss man also bei LMSV alle Nullstellen λ des Stabilitätspolynoms $p_\Phi(\cdot; qh)$ für $Re(qh) < 0$ auf ihren Betrag λ (und ggf. auf ihre Vielfachheit) untersuchen. Die Nullstabilität enspricht somit der Stabilität für $z = 0$.

Satz 5.17 *LMSV Φ mit maximaler Konvergenzordnung gemäß Satz 5.14 sind nicht A-stabil, insbesondere besitzen sie das triviale Stabilitätsgebiet $S_\Phi = \{0\}$.*

Beweis. Wir verweisen auf das Buch [3]. ∎

Also sind gemäß Corollar 5.13 das k-stufige Adams-Moulton und Milne-Verfahren für k ungerade nicht A-stabil.

Beispiel: Bei der Milne-Formel für $k = 2$ gilt ($w := qh$):

$$\begin{aligned}
\rho(z) &= z^2 - 1\,, \\
\sigma(z) &= \frac{1}{3}(z^2 + 4z + 1)\,, \\
p(z; w) &= \rho(z) - w\sigma(z) \;=\; z^2 - 1 - \frac{1}{3}w(z^2 + 4z + 1) \\
&= (1 - \tfrac{1}{3}w)z^2 - \tfrac{4}{3}wz - \tfrac{1}{3}w - 1\,.
\end{aligned}$$

Die Nullstellen λ von $p(z; w)$ in Abhängigkeit von w zu finden, ist also nicht ganz einfach. Wir nehmen hier aber nun einmal an, dass die Nullstellen in differenzierbarer Art und Weise von w abhängen. Dies motiviert eine Entwicklung der Form

$$\lambda \;=\; a_0 + a_1 w + \mathcal{O}(w^2)\,.$$

Der Koeffizient a_0 ergibt sich aus den Nullstellen für $w = 0$, also $p(\lambda; 0) = \lambda^2 - 1 = 0$, bzw. $\lambda = \pm 1$. Es gibt also die beiden Möglichkeiten $a_0 = 1$ und $a_0 = -1$. Setzen wir $a_0 = 1$ ein, so erhalten wir

$$(1 - \tfrac{1}{3}w)(1 + a_1 w + \mathcal{O}(w^2))^2 - \tfrac{4}{3}w(1 + a_1 w + \mathcal{O}(w^2)) - \tfrac{1}{3}w - 1 \quad = \quad 0\,.$$

Die linearen (in w) Terme ergeben:

$$2a_1 w - \tfrac{1}{3}w - \tfrac{4}{3}w - \tfrac{1}{3}w \quad = \quad 0\,,$$

also $a_1 = 1$. Für $a_0 = -1$ ergibt sich entsprechend $a_1 = 1/3$. Die Eigenwerte haben also die Form

$$\begin{aligned} \lambda_1 &= 1 + w + \mathcal{O}(w^2)\,, \\ \lambda_2 &= -1 + \frac{w}{3} + \mathcal{O}(w^2)\,. \end{aligned}$$

Für sehr kleine Schrittweite h kann man den w^2-Einfluß vernachlässigen, so dass sich die beiden relevanten Eigenwerte verhalten wie $\lambda_1 \approx 1 + hq$ und $\lambda_2 \approx -1 + hq/3$. Nun sieht man, dass die Bedingung $|\lambda_{1,2}| \leq 1$ nur für den Punkt $hq = 0$ erreicht werden kann. Dies ist ein Indiz dafür, dass auch hier gilt $S_\Phi = \{0\}$. Die Milne-Formel ist nicht A-stabil.

Satz 5.18 *Es gibt kein A-stabiles explizites LMSV. A-stabile implizite LMSV besitzen die maximale Konvergenzordnung $p = 2$. Die Trapezregel (Adams-Moulton $k = 2$) ist ein A-stabiles LMSV mit Konvergenzordnung $p = 2$. Zudem besitzt diese die minimalste Fehlerkonstante.*

5.8 A(α)-Stabilität

Da die A-Stabilität für LMSV offensichtlich zu restriktiv ist, wurde schwächere Stabilitätseigenschaften formuliert. Hierzu gehört die A(α)-Stabilität:

Definition 5.19 *Ein konvergentes LMSV Φ heißt ($A(\alpha)$-stabil mit $0 < \alpha \leq \pi/2$, wenn sein Stabilitätsgebiet S_Φ den Sektor $S_\alpha := \{z \in \mathbb{C} : |\pi - arg(z)| < \alpha\}$ enthält, $S_\alpha \subseteq S_\Phi$. Es heißt A(0)-stabil, wenn es $A(\alpha)$-stabil ist für ein $\alpha > 0$.*

Also ist $A(\pi/2)$-Stabilität äquivalent mit A-Stabilität. Bei A(0)-stabilen Verfahren ist zumindest die reelle Halbachse $S_0 = \mathbb{R}_{<0}$ im Stabilitätsgebiet S_Φ enthalten. Die $A(\alpha)$-Stabilität erfordert es, dass für $w \in S_\alpha$ das Stabilitätspolynom $p_\Phi(z; w) = \rho(z) - w\sigma(z)$ nur Nullstellen λ besitzt, deren Betrag kleiner gleich 1 ist. Die Übertragung des Begriffs der L-Stabilität in (4.3) bedeutet, dass für $Re\, w \to -\infty$ die Nullstellen λ von $p_\Phi(\cdot; w)$ gegen Null konvergieren. Die Eigenschaft $p_\Phi(\lambda_w; w) = 0$ ist für $w \neq 0$ äquivalent mit

$$\frac{\rho(\lambda_w)}{w} \quad = \quad \sigma(\lambda_w)\,.$$

Tabelle 5.1: Ungefähre Winkel α für die $A(\alpha)$-Stabilität der BDF(k)-Verfahren.

k	1	2	3	4	5	6
α	90^o	90^o	86^o	73^o	52^o	18^o

Es folgt dann wegen der Beschränktheit von p im beschränkten Einheitskreis:

$$0 \quad = \quad \lim_{Re\, w \to -\infty} \left| \frac{p(\lambda_w)}{w} \right| \quad = \quad \lim_{Re\, w \to -\infty} \sigma(\lambda_w)\,.$$

Für $Re\, w \to -\infty$ entsprechen die Nullstellen von $p_\Phi(\cdot; w)$ also denen von σ. Um das Kriterium (4.3) zu erfüllen, darf σ also nur Nullstellen im Nullpunkt besitzen, d.h. das zweite charakteristische Polynom ist von der Form

$$\sigma(z) \quad = \quad \beta_0 z^k\,,$$

mit $\beta_0 \neq 0$. Dies leisten gerade die BDF-Verfahren, denn hier gilt für beliebiges k:

$$\sigma_{BDF(k)}(z) \quad = \quad z^k\,.$$

Die BDF-Verfahren sind daher $A(\alpha)$-stabil. Der Winkel α hängt aber von der Schrittweite k ab. Mit wachsendem k wird α kleiner. Siehe hierzu die Tabelle 5.1.

Kapitel 6

Unstetige Galerkin-Verfahren

6.1 Variationelle Formulierung

Wir formulieren die AWA (2.1) um, indem wir die ganze Gleichung mit sogenannten Testfunktionen φ multiplizieren und anschließend über das Zeitintervall I integrieren. Gesucht ist also ein $u \in C^1(I)$ mit $u(t_0) = u_0$ und

$$\int_I \langle u'(s) - f(s, u(s)), \varphi(s) \rangle ds \quad = \quad 0 \qquad \forall \varphi \in C(I)\,. \tag{6.1}$$

Hierbei und im Folgenden nehme wir stets den skalaren Fall $n = 1$ an, da sich alles problemlos auf den vektorwertigen Fall übertragen läßt. Während die Lösung u von (2.1) *klassische Lösung* oder auch *starke Lösung* heisst, heissen die Lösungen von (6.1) *variationelle Lösungen.*

Satz 6.1 *Eine klassische Lösung u von (2.1) ist auch eine variationelle Lösung von (6.1). Umgekehrt ist eine variationelle Lösung von (6.1) auch klassische Lösung u von (2.1).*

Beweis. Offensichtlich ist eine klassische Lösung auch eine variationelle Lösung, da der Term $u'(s) - f(s, u(s))$ punktweise verschwindet. Die Umkehrung gilt auch wie folgende Überlegung zeigt. Wir beschränken uns auf beschränkte Intervalle I. Sei u eine variationelle Lösung von (6.1) und $t \in I$ beliebig. Wir wählen eine Folge (φ_ϵ) von Funktionen in $C(I)$ für $\epsilon > 0$, $\epsilon \to 0$ mit

$$\varphi_\epsilon \geq 0\,, \quad \int_{\mathbb{R}} \varphi_\epsilon(s)\, ds = 1\,, \quad \int_{t-\epsilon}^{t+\epsilon} \varphi_\epsilon(s)\, ds \geq 1 - \epsilon\,.$$

Solche sogenannten Dirac-Folgen existieren, z.B.:

- $\varphi_\epsilon(x) := \epsilon^{-1}\phi((x-t)/\epsilon)$ mit einer stetigen Funktion $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ mit $1 = \phi(0) = \int_{\mathbb{R}} \phi\, ds$.

- Normalverteilungen für $t = 0$:

$$\varphi_\epsilon(x) \;:=\; \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{x^2}{2\epsilon}\right)$$

Für eine solche Folge ergibt sich :

$$\begin{aligned} 0 &= \lim_{\epsilon\to 0} \int_I \langle u'(s) - f(s,u(s)), \varphi_\epsilon(s)\rangle ds \\ &= \lim_{\epsilon\to 0} \left(\int_{|s-t|\geq\epsilon} \langle u'(s) - f(s,u(s)), \varphi_\epsilon(s)\rangle ds + \int_{t-\epsilon}^{t+\epsilon} \langle u'(s) - f(s,u(s)), \varphi_\epsilon(s)\rangle ds \right). \end{aligned}$$

Für das erste auftretende Integral gilt:

$$\begin{aligned} \int_{|s-t|\geq\epsilon} \langle u'(s) - f(s,u(s)), \varphi_\epsilon(s)\rangle ds &\leq \|u' - f(\cdot,u)\|_{L^\infty(I)} \int_{|s-t|\geq\epsilon} \varphi_\epsilon(s)\, ds \\ &\leq \|u' - f(\cdot,u)\|_{L^\infty(\overline{I})} \epsilon\,. \end{aligned}$$

Im Grenzübergang $\epsilon \to 0$ verschwindet dieses Integral daher. Für das zweite auftretende Integral gilt mit dem Mittelwertsatz der Integralrechnung für ein ξ_ϵ mit $|t - \xi_\epsilon| < \epsilon$:

$$\int_{t-\epsilon}^{t+\epsilon} \langle u'(s) - f(s,u(s)), \varphi_\epsilon(s)\rangle ds \;=\; (u'(\xi_\epsilon) - f(\xi_\epsilon, u(\xi_\epsilon))) \int_{t-\epsilon}^{t+\epsilon} \varphi_\epsilon(s)\, ds\,.$$

Da $1 - \epsilon \leq \int_{t-\epsilon}^{t+\epsilon} \varphi_\epsilon(s)\, ds \leq 1$ und $\lim_{\epsilon\to 0} \xi_\epsilon = t$ folgt

$$\lim_{\epsilon\to 0} \int_{t-\epsilon}^{t+\epsilon} \langle u'(s) - f(s,u(s)), \varphi_\epsilon(s)\rangle ds \;=\; u'(t) - f(t,u(t))\,.$$

Insgesamt folgt damit die Behauptung. ∎

Nun unterteilen wir das Zeitintervall $t_0 < \ldots < t_N = t_0 + T$ und verwenden die Teilintervalle $I_k = (t_{k-1}, t_k]$ für $k \in \{1,\ldots,N\}$. Diese Unterteilung bezeichnen wir mit $\mathcal{T}_h$ wobei h die Schrittweite bezeichnet. Zu dieser Unterteilung betrachten wir den Raum der stückweise stetig differenzierbaren Funktionen

$$V(\mathcal{T}_h) \;:=\; \{\varphi : I \to \mathbb{R} \,:\, \varphi|_{I_n} \in C^1(I_n) \cap C(\overline{I}_n)\,,\ \forall n \in \{1,\ldots,N\}\}$$

Hierbei bezeichnet $C^1(I_n) \cap C(\overline{I}_n)$ den Raum der Funktionen, die auf den halboffenen Intervall I_n stetig-differenzierbar sind und zum linken Randpunkt von I_n stetig fortsetzbar sind. Für Funktionen $v \in V(\mathcal{T}_h)$ definieren wir die jeweiligen Grenzwerte und Sprünge

$$v_n^+ \;:=\; \lim_{t\searrow t_n} v(t)\,, \quad v_n^- \;:=\; \lim_{t\nearrow t_n} v(t)\,, \quad [v]_n \;:=\; v_n^+ - v_n^-\,.$$

Die zweite variationelle Formulierung (für stückweise glatte Funktionen), die wir betrachten, lautet: Gesucht ist ein $u \in V(\mathcal{T}_h)$, so dass $u(t_0) = u_0$ und

$$\sum_{n=1}^{N} \int_I \langle u'(s) - f(s,u(s)), \varphi(s)\rangle ds + \sum_{n=1}^{N-1} \langle [u]_n, \varphi_n^+\rangle \;=\; 0 \qquad \forall \varphi \in V(\mathcal{T}_h)\,. \tag{6.2}$$

Satz 6.2 *Jede klassische Lösung u von (2.1) ist auch eine variationelle Lösung von (6.2). Umgekehrt ist jede variationelle Lösung u von (6.2) auch klassische Lösung von (2.1).*

Beweis. Da $C^1(I) \subset V(\mathcal{T}_h)$ und $[u]_n = 0$ für $u \in C^1(I)$, ist eine Lösung von (6.1) auch Lösung von (6.2). Sei nun $u \in C^1(I)$ eine Lösung von (6.2). Für Zeitpunkte $t \in I$, die kein Gitterpunkt darstellen, folgt $u'(t) = f(t, u(t))$ nach dem gleichen Argument wie bei der stetigen variationellen Formulierung. Es genügt daher Gitterpunkte t_k zu betrachten. Wir wählen eine Folge von Testfunktionen φ_ϵ mit $\varphi_\epsilon(t_k) = 1$ und $\lim_{\epsilon\to 0} \varphi_\epsilon(t) = 0$ für alle $t \neq t_k$. Aufgrund von (6.2) folgt dann $[u]_k = 0$ und somit die Stetigkeit von u auf dem ganzen Intervall I. Die Differenzierbarkeit von u auf den Gitterpunkten folgt aus der Stetigkeit von f und der Gültigkleit von $u' = f(\cdot, u)$ auf den Teilintervallen I_n. Folglich ist u auch Lösung von der stetigen Variante (6.1). Mit Satz 6.1 folgt die Behauptung. □
Wir definieren nun die Semilinearform

$$A(u;\varphi) \;:=\; \sum_{n=1}^{N} \int_{I_n} \langle u'(s) - f(s,u(s)), \varphi(s)\rangle ds + \sum_{n=0}^{N-1} \langle [u]_n, \varphi_n^+ \rangle \,. \tag{6.3}$$

Diese ist im zweiten Argument φ linear. Bauen wir nun noch die Anfangswerte in den Raum ein, $V(\mathcal{T}_h, u_0) := \{u \in V(\mathcal{T}_h) \,:\, u(t_0) = u_0\}$, so lautet die unstetige variationelle Formulierung (6.2) in kompakter Form

$$u \in V(\mathcal{T}_h, u_0): \qquad A(u;\varphi) \;=\; 0 \qquad \forall \varphi \in V(\mathcal{T}_h)\,.$$

6.2 Die DG-Verfahren

Wir wählen jetzt anstelle von $V(\mathcal{T}_h)$ endlich-dimensionale Unterräume und formulieren dadurch diskrete Lösungen. Hierzu wählen wir stückweise Polynome vom Grad $r \geq 0$:

$$V_r(\mathcal{T}_h) \;:=\; \{\varphi \in V(\mathcal{T}_h) \,:\, \varphi|_{I_n} \in P_r \,\forall n \in \{1,\ldots,n\}\} \;=\; \bigoplus_{n=1}^{N} P_r(I_n)\,. \tag{6.4}$$

Entsprechend ist $V_r(\mathcal{T}_h, u_0) := V_r(\mathcal{T}_h) \cap V(\mathcal{T}_h, u_0)$. Gesucht ist nun

$$U \in V_r(\mathcal{T}_h, u_0): \quad A(U;\varphi) \;=\; 0 \qquad \forall \varphi \in V_r(\mathcal{T}_h)\,. \tag{6.5}$$

Diese Methode nennt man *discontinuous Galerkin* Verfahren (kurz DG-Verfahren).

Satz 6.3 *Das Verfahren (6.5) ist konsistent in dem Sinne, dass jede klassische Lösung u für die gilt $u \in V_r(\mathcal{T}_h)$ auch Lösung von (6.5) ist.*

Beweis. Jede klassische Lösung u ist nach Satz 6.2 auch eine variationelle Lösung von (6.2) und wegen $V_r(\mathcal{T}_h) \subset V(\mathcal{T}_h)$ dann auch Lösung von (6.5). □

Die Idee besteht nun ferner darin, dass man für den Raum $V_r(\mathcal{T}_h)$ eine endliche Basis $\mathcal{B}_{h,r}$ wählt und die diskrete Lösung U in dieser Basis ausdrückt. Hierbei ist die Darstellung

von $V_r(\mathcal{T}_h)$ als direkte Summe (6.4) wichtig, denn dies erlaubt es die Funktionen in $\mathcal{B}_{h,r}$ so zu wählen, dass sie jeweils nur einen lokalen Trägern besitzen:

$$\mathcal{B}_{h,r} \;=\; \bigcup_{n=1}^{N}\bigcup_{i=0}^{r}\{\varphi_{n,i}\}\,, \quad \text{wobei } supp\,\varphi_{n,i} \subset \overline{I_n}\ \forall i = 0,\ldots,r\,.$$

Die diskrete Lösung U ausgedrückt in dieser Basis lautet:

$$U(t) \;:=\; \sum_{n=1}^{N}\sum_{j=0}^{r} U_{n,j}\varphi_{n,j}(t)\,.$$

Die $U_{n,j}$ sind hierbei reelle Koeffizienten. Die Gleichung (6.5) muss nur für alle Basisfunktionen gelten, denn A ist im zweiten Argument linear:

$$A\left(\sum_{n=1}^{N}\sum_{j=0}^{r} U_{n,j}\varphi_{n,j};\varphi_{m,i}\right) \;=\; 0 \qquad \forall i\in\{0,\ldots,r\}\,\forall m\in\{1,\ldots,N\}\,. \tag{6.6}$$

Dies entspricht $M = N(r+1)$ (i.a. nichtlinearen) Gleichungen für M Unbekannten Koeffizienten. Aufgrund der lokalen Träger der Testfunktionen lassen sich diese aber auch darstellen als eine sukzessive Folge von nichtlinearen Gleichungen mit weniger Unbekannten. Wählt man nämlich eine Testfunktion $\varphi = \varphi_{i,n}$, dessen Träger sich ja auf das Intervall I_n beschränkt, so ist (6.6) für dieses φ äquivalent mit der *lokalen Galerkin Gleichung*

$$\int_{t_{n-1}}^{t_n} \langle U'(s) - f(s,U(s)),\varphi(s)\rangle ds + \langle U_{n-1}^{+},\varphi_{n-1}^{+}\rangle \;=\; \langle U_{n-1}^{-},\varphi_{n-1}^{+}\rangle\,, \tag{6.7}$$

denn hier gilt $\varphi_n^+ = \varphi(t_n)$. Kennt man U_{n-1}^-, so fließt in diese Gleichung nur noch Informationen von U auf I_n ein (beachte: auch U_{n-1}^+ ist bestimmt durch U auf I_n). Mit anderen Worten: mit der Kenntnis von U_{n-1}^+ läßt sich (6.7) auffassen als ein System von $(r+1)$ nichtlinearen Gleichungen für die $(r+1)$ Koeffizienten $U_{n,0},\ldots,U_{n,r}$. Diese kleineren Systeme lassen sich sukzessive für $n = 1,\ldots,N$ abarbeiten.

6.2.1 DG(0)-Verfahren

Wir wollen hier die Fälle $r = 0$ und $r = 1$ diskutieren, denn hierbei ergeben sich Varianten von bereits behandelten ESV:

Im Fall $r = 0$ betrachtet man auf jedem Intervall nur konstante Funktionen. Die Basis besteht dann aus den charakteristischen Funktionen $\varphi_n = \chi_{I_n}$. Es gilt dann $U' = 0$ im Innern eines jeden Intervalls sowie $U_{n-1}^+ = U_n$ und $U_{n-1}^- = U_{n-1}$. Daher lautet (6.7) dann

$$\textbf{DG(0):} \qquad U_n - \int_{t_{n-1}}^{t_n} f(s,U_n)ds \;=\; U_{n-1}\,. \tag{6.8}$$

Dies entspricht einer Variante des impliziten Euler-Verfahrens, bei dem das Integral eine etwas andere gestalt hat. Wenn man dieses Integral nun durch eine Quadraturformel approximiert, z.B. durch

$$\int_{t_{n-1}}^{t_n} f(s, U_n)ds \quad \approx \quad h_n f(t_n, U_n)\,,$$

so erhält man genau das bereits bekannte implizite Euler-Verfahren:

$$U_n \quad = \quad U_{n-1} + h_n f(t_n, U_n)\,.$$

Lemma 6.4 *Das DG(0)-Verfahren (6.8) ist L-stabil.*

Beweis. (a) Wir zeigen zunächst die A-Stabilität: Zur Untersuchung der Stabilität von DG(0) wählen wir für f speziell die lineare Funktion $f(t, u) = \lambda u$. In diesem Fall ergibt sich

$$U_n - h\lambda U_n \quad = \quad U_{n-1}$$

Man erhält daher die Stabilitätsfunktion ($z = h\lambda$):

$$g(z) \quad = \quad \frac{1}{1-z}\,.$$

Das Stabilitätsgebiet $S_{DG(0)}$ ist gegeben durch die $z \in \mathbb{C}$ mit $|g(z)| \leq 1$, also

$$S_{DG(0)} \quad = \quad \mathbb{C} \setminus B_1(1)\,.$$

Somit ist die negative reelle Halbebene in $S_{DG(0)}$ enhalten und die ist A-Stabilität.
(b) Die starke A-Stabilität ist auch offensichtlich.
(c) L-Stabilität: Es gilt

$$\lim_{Re\, z \to -\infty} |g(z)| \quad = \quad \lim_{Re\, z \to -\infty} \frac{1}{|1-z|} \quad = \quad 0\,.$$

□

6.2.2 DG(1)-Verfahren

Im Fall $r = 1$ haben wir einen linearen Ansatz auf den Teilintervallen. U läßt sich daher auf I_n darstellen durch seine beiden Endwerte U_{n-1}^+ und U_n^-:

$$U(t) \quad = \quad U_{n-1}^+ + h_n^{-1}(t - t_{n-1})(U_n^- - U_{n-1}^+) \qquad \text{für } t \in I_n\,.$$

Die Ableitung U' ist auf I_n konstant, $U'|_{I_n} = h_n^{-1}(U_n^- - U_{n-1}^+)$. Das Test von (6.7) mit der konstanten Funktion $\varphi \equiv 1$ ergibt:

$$U_n^- - \int_{t_{n-1}}^{t_n} f(s, U(s))ds \quad = \quad U_{n-1}^-\,.$$

Das Testen von (6.7) mit der linearen Funktion $\varphi(t) = (t-t_{n-1})/h_n$ ergibt wegen $\varphi_{n-1}^+ = 0$:

$$(U_n^- - U_{n-1}^+)h_n^{-1} \int_{t_{n-1}}^{t_n} (s - t_{n-1})ds - h_n^{-1} \int_{t_{n-1}}^{t_n} f(s, U(s))(s - t_{n-1})ds \;=\; 0\,.$$

Die Berechnung des ersten Integrals führt auf die Gleichung

$$\textbf{DG(1):} \qquad U_n^- - U_{n-1}^+ - \frac{2}{h_n} \int_{t_{n-1}}^{t_n} f(s, U(s))(s - t_{n-1})ds \;=\; 0\,. \tag{6.9}$$

Diese beiden auftretenden Integrale approximieren wir jetzt durch die Trapezregel:

$$\begin{aligned} \int_{t_{n-1}}^{t_n} f(s, U(s))ds &\approx \frac{h_n}{2}(f(t_{n-1}, U_{n-1}^+) + f(t_n, U_n^-))\,, \\ \int_{t_{n-1}}^{t_n} f(s, U(s))(s - t_{n-1})ds &\approx \frac{1}{2} f(t_n, U_n^-)h_n^2\,. \end{aligned}$$

Dies ergibt die beiden Gleichungen:

$$\begin{aligned} U_n^- - \frac{h_n}{2}(f(t_{n-1}, U_{n-1}^+) + f(t_n, U_n^-)) &= U_{n-1}^-\,, \\ U_n^- - U_{n-1}^+ - f(t_n, U_n^-)h_n &= 0\,. \end{aligned}$$

Die Eliminierung von U_{n-1}^+ ergibt:

$$\begin{aligned} U_n^- &= U_{n-1}^- + \frac{h_n}{2}(k_1 + k_2)\,, \\ k_2 &= f(t_n, U_n^-) \;=\; f(t_n, U_{n-1}^- + \frac{h_n}{2}(k_1 + k_2))\,, \\ k_1 &= f(t_{n-1}, U_{n-1}^+) \;=\; f(t_{n-1}, U_n^- - h_n k_2) \;=\; f(t_{n-1}, U_{n-1}^- + h_n(\tfrac{1}{2}k_1 - \tfrac{1}{2}k_2))\,. \end{aligned}$$

Dies entspricht gerade einem impliziten 2-stufigen Runge-Kutta-Verfahren zum Butcher-Tableau: Man rechnet schnell nach, dass diese RK-Methode die Konsistenzornung 2 besitzt.

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

Allerdings besitzt die DG(1)-Methode (exakte Auswertung des Integrals oder hinreichend genaue Quadratur) sogar die Ordnung 3 an den Gitterpunkten.

Lemma 6.5 *Das DG(1)-Verfahren (6.9) ist stark A-stabil, aber nicht L-stabil.*

Beweis. Das DG(1)-Verfahren (6.9) lautet für $f(t,u) = \lambda u$:

$$\begin{aligned} 0 &= U_n^- - U_{n-1}^+ - \frac{2}{h_n}\lambda \int_{t_{n-1}}^{t_n} U(s)(s - t_{n-1})ds \\ &= U_n^- - U_{n-1}^+ - \frac{2}{h_n}\lambda \int_0^{h_n} (U_{n-1}^+ + h_n^{-1} s(U_n^- - U_{n-1}^+))s\, ds \\ &= U_n^- - U_{n-1}^+ - \frac{2}{h_n}\lambda(\tfrac{1}{2}h_n^2 U_{n-1}^+ + \tfrac{1}{3}h_n^2(U_n^- - U_{n-1}^+)) \\ &= (1 - \tfrac{2}{3}\lambda h_n)U_n^- + (-1 - \lambda h_n + \tfrac{2}{3}\lambda h_n)U_{n-1}^+ . \end{aligned}$$

Umstellung dieser Gleichung ergibt

$$U_n^- = \left(1 + \frac{\lambda h_n}{1 - \frac{2}{3}\lambda h_n}\right) U_{n-1}^+ .$$

Also lautet die Stabilitätsfunktion

$$g(z) = 1 + \frac{z}{1 - \frac{2}{3}z} = \frac{3+z}{3-2z} .$$

Die Bedingung $|g(z)| \le 1$ ist daher äquivalent zu $|3+z| \le |3-2z|$. Dies ist für $Re\, z \le 0$ stets erfüllt, was die A-Stabilität impliziert. Da

$$\lim_{Re\, z \to \infty} |g(z)| = \lim_{Re\, z \to \infty} \frac{|3+z|}{|3-2z|} = \lim_{Re\, z \to \infty} \frac{|z|}{|2z|} = \frac{1}{2} ,$$

folgt die starke A-Stabilität und die nicht-Gültigkeit der L-Stabilität. □

6.3 Lösbarkeit der nichtlinearen Gleichungen

Lemma 6.6 (Young'sche Ungleichung) *Für alle $a, b \in \mathbb{C}$ und $p, q \in (1, \infty)$ mit $1/p + 1/q = 1$ gilt:*

$$|ab| \le \frac{1}{p}|a|^p + \frac{1}{q}|b|^q .$$

Beweis. Es genügt den Fall $a, b \in \mathbb{R}^+$ zu betrachten. Hier benutzen wir die Tatsache, dass der Logarithmus in $\mathbb{R}^+$ monoton wachsend ist, so dass die Behauptung äquivalent ist zu

$$\ln\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) \ge \ln(ab) .$$

Diese Ungleichung ergibt sich aber, da der Logarithmus zudem konkav ist:

$$\ln\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) \ge \frac{1}{p}\ln(a^p) + \frac{1}{q}\ln(b^q) = \ln(a) + \ln(b) = \ln(ab) .$$

□

Häufig wird diese Ungleichung verwendet, um ein Produkt zweier Größen nach oben durch die gewichtete Summe der Quadrate abschätzt. Dabei kann ein Summand beliebig klein gemacht werden.

Korollar 6.7 *Seien $a, b \in \mathbb{C}$. Für beliebiges $\epsilon > 0$ gilt:*

$$|ab| \;\leq\; \frac{\epsilon}{2}|a|^2 + \frac{1}{2\epsilon}|b|^2 \,.$$

Beweis. Man benutzt die Young'sche Ungleichung für $a' := \sqrt{\epsilon}a$ und $b' := b/\sqrt{\epsilon}$. □

Satz 6.8 *[Diskrete Sobolev'sche Ungleichung] Es existiert eine Konstante C, so dass für beliebige $a, b \in \mathbb{R}$, $a < b$, und alle $u \in P_r$ für die Supremumsnorm gilt:*

$$\|u\|_{L^\infty(a,b)} \;\leq\; C \left(\int_a^b |u'(t)|^2 (t-a)\, dt + |u(b)|^2 \right)^{1/2} .$$

Beweis. Wir betrachten zunächst die folgenden beiden Normen auf dem Raum P_r:

$$\begin{aligned} \|\widehat{u}\|_{L^\infty(0,1]} &= \sup_{x\in(0,1]} |\widehat{u}(x)| \,, \\ \|\widehat{u}\| &= \left(\int_0^1 |\widehat{u}'(t)|^2 t\, dt + |\widehat{u}(1)|^2 \right)^{1/2} . \end{aligned}$$

Dass dies tatsächlich Normen auf P_r sind, verifiziere man in einer Übungsaufgabe. Da nun P_r endlich-dimensional ist, sind diese beiden Normen äquivalent, also

$$\|\widehat{u}\|_\infty \;\leq\; C\|\widehat{u}\| \qquad \forall \widehat{u} \in P_r(0,1] \,,$$

mit einer Konstanten C. Nun erfolgt ein sogenanntes *Skalierungsargument*, bei dem wir diese Ungleichungauf ein allgemeines Intervall $(a, b]$ übertragen. Hierzu führen wir die affin-lineare Transformation $\Phi : P_r \to P_r$. $\widehat{u} \mapsto \Phi\widehat{u} =: u$ ein ($h := b - a$):

$$u(t) \;=\; \widehat{u}((t-a)/h) \quad \text{bzw.} \quad \widehat{u}(\widehat{t}) \;=\; u(a + h\widehat{t}) \,.$$

Es gilt mit $t = a + h\widehat{t}$:

$$\widehat{u}'(\widehat{t}) \;=\; \frac{d}{d\widehat{t}}\widehat{u}(t) \;=\; \frac{d}{d\widehat{t}}u(a+h\widehat{t}) \;=\; u'(a+h\widehat{t})h \;=\; u'(t)h \,.$$

Mit Substitution $dt = h_n d\widehat{t}$ folgt daher

$$\begin{aligned} \|u\|_{L^\infty(a,b)} &= \|\widehat{u}\|_{L^\infty(0,1]} \;\leq\; C\|\widehat{u}\| \\ &= C \left(\int_0^1 |\widehat{u}'(\widehat{t})|^2 \widehat{t}\, d\widehat{t} + |u(b)|^2 \right)^{1/2} \\ &= C \left(\int_a^b |u'(t)|^2 h^2 (t-a) h^{-2}\, dt + |u(b)|^2 \right)^{1/2} . \end{aligned}$$

Dies ist gerade die Behauptung. □

Satz 6.9 *Sei f lipschitz-stetig mit Lipschitz-Konstante L_f. Dann liefern die DG(r)-Gleichungen (6.5) für $h < (\gamma L_f)^{-1}$ stets eine eindeutige Lösung $U \in V_r(\mathcal{T}_h)$. Hierbei ist $\gamma > 0$ eine nur von r abhängige Konstante.*

Wie der folgende Beweis zeigt, kann $\gamma := \sqrt{C(1+C)}$ gewählt werden mit der Konstante C aus der diskreten Sobolevschen Ungleichung (Satz 6.8).

Beweis. Es genügt zu zeigen, dass jeder einzelne Zeitschritt (6.7) wohldefiniert ist. Hierzu gehen wir in mehreren Schritten vor.

Schritt 1: Wir formulieren die Abbildung $\Psi : P_r(I_n) \to: P_r(I_n)$, $V \mapsto U = \Psi(V)$, die gegeben ist als Lösung des folgenden linearen Gleichungssystems:

$$\int_{I_n} \langle U', \varphi \rangle \, dt + \langle U^+_{n-1}, \varphi^+_{n-1} \rangle \;=\; \int_{I_n} \langle f(t, V(t)), \varphi \rangle \, dt + \langle V^-_{n-1}, \varphi^-_{n-1} \rangle \qquad \forall \varphi \in P_r(I_n)\,.$$

Dieses lineare System ist genau dann eindeutig lösbar, wenn es für das homogene System

$$\int_{I_n} \langle W', \varphi \rangle \, dt + \langle W^+_{n-1}, \varphi^+_{n-1} \rangle \;=\; 0 \qquad \forall \varphi \in P_r(I_n)$$

nur die triviale Lösung $W \equiv 0$ gibt. Um dies zu sehen, wählen wir $\varphi = W$ und erhalten

$$\int_{I_n} \langle W', W \rangle \, dt + |W^+_{n-1}|^2 \;=\; 0\,.$$

Wegen

$$\int_{I_n} \langle W', W \rangle \, dt \;=\; \frac{1}{2} \int_{I_n} \frac{d}{dt} |W|^2 \, dt \;=\; \frac{1}{2}(|W^-_n|^2 - |W^+_{n-1}|^2)\,,$$

erhalten wir

$$\frac{1}{2}(|W^-_n|^2 + |W^+_{n-1}|^2) \;=\; 0\,.$$

Daher folgt zunächst $W^-_n = W^+_{n-1} = 0$. Jetzt wählen wir als Testfunktion $\varphi := (t - t_{n-1})W'$ und erhalten wegen $\varphi^+_{n-1} = 0$:

$$0 \;=\; \int_{I_n} \langle W', (t - t_{n-1})W' \rangle \, dt \;=\; \int_{I_n} |W'|^2 (t - t_{n-1}) \, dt\,.$$

Da $t - t_{n-1}$ 0 für $t > t_{n-1}$, folgt $W' \equiv 0$. Insgesamt gilt damit $W \equiv 0$.

Schritt 2: Diese Abbildung Ψ definiert uns nun eine Fixpunktiteration

$$U^{(k+1)} \;:=\; \Psi(U^{(k)})\,.$$

Sofern diese Iteration konvergiert, liefert sie als Grenzwert $U := \lim_{k\to\infty} U^{(k)}$ eine Lösung von (6.7) auf dem Intervall I_n. Die Konvergenz gegen einen eindeutigen Fixpunkt weisen wir jetzt mittels des Banachschen Fixpunktsatzes nach. Dazu genügt es zu zeigen, dass

Ψ auf $\mathbb{R}^{r+1}$ bei festem V_{n-1}^- eine Kontraktion ist. Hierzu muß man die Schrittweite h_n hinreichend klein wählen.

Schritt 3: Wir zeigen nun für $V, \widetilde{V} \in P_r(I_n)$ mit $V_{n-1}^- = \widetilde{V}_{n-1}^-$ und $U := \Psi(V) - \Psi(\widetilde{V})$ in der Maximumsnorm auf I_n

$$\|U\|_\infty \;=\; \|\Psi(V) - \Psi(\widetilde{V})\|_\infty \;\leq\; h_n \gamma L \|V - \widetilde{V}\|_\infty \,,$$

so dass die Kontraktionseigenschaft aus der Bedingung an h_n folgt. Das diagonale Testen mit $\varphi := U$ (ähnlich zum Schritt 1) führt nun bei dem inhomogenen System aufgrund der Lipschitz-Stetigkeit von f auf:

$$\begin{aligned}
\frac{1}{2}(|U_n^-|^2 + |U_{n-1}^+|^2) &= \int_{I_n} \langle f(t, V(t) - f(t, \widetilde{V}(t)), U\rangle \, dt \\
&\leq L \int_{I_n} |V(t) - \widetilde{V}(t)||U| \, dt \\
&\leq L h_n \|V - \widetilde{V}\|_\infty \|U\|_\infty \\
&\leq \frac{1}{2\epsilon} L^2 h_n^2 \|V - \widetilde{V}\|_\infty^2 + \frac{\epsilon}{2} \|U\|_\infty^2 \,.
\end{aligned}$$

Im letzten Schritt haben wir die Young'schen Ungleichung aus Lemma 6.6 benutzt und gilt daher für beliebiges $\epsilon > 0$. Wieder wie im Schritt 1 testen wir nun mit $\varphi := (t - t_{n-1})U'$. Dies liefert nun mit nochmaliger Anwendung der Young'schen Ungleichung:

$$\begin{aligned}
\int_{I_n} |U'|^2 (t - t_{n-1}) \, dt &\leq \int_{I_n} L |V - \widetilde{V}||U'| (t - t_{n-1}) \, dt \\
&\leq \frac{1}{2} \int_{I_n} \left(L^2 |V - \widetilde{V}|^2 + |U'|^2 \right) (t - t_{n-1}) \, dt \,.
\end{aligned}$$

Umordnen und Multiplikation mit 2 ergibt:

$$\begin{aligned}
\int_{I_n} |U'|^2 (t - t_{n-1}) \, dt &\leq L^2 \int_{I_n} |V - \widetilde{V}|^2 (t - t_{n-1}) \, dt \\
&\leq \frac{1}{2} L^2 h_n^2 \|V - \widetilde{V}\|_\infty^2 \,.
\end{aligned}$$

Kombination dieser Resultate ergibt

$$\frac{1}{2}|U_n^-|^2 + \int_{I_n} |U'|^2 (t - t_{n-1}) \, dt \;\leq\; \frac{1}{2} \left(1 + \epsilon^{-1}\right) L^2 h_n^2 \|V - \widetilde{V}\|_\infty^2 + \frac{\epsilon}{2} \|U\|_\infty^2 \,.$$

Die diskrete Sobolevsche Ungleichung von Satz 6.8 liefert nun mit der Wahl $\epsilon = C^{-1}$:

$$\begin{aligned}
\|U\|_\infty^2 &\leq C \left(\frac{1}{2} |U_n^-|^2 + \int_{I_n} |U'|^2 (t - t_{n-1}) \, dt \right) \\
&\leq C \frac{1}{2} (1 + C) L^2 h_n^2 \|V - \widetilde{V}\|_\infty^2 + \frac{1}{2} \|U\|_\infty^2 \,.
\end{aligned}$$

Dies impliziert mit $\gamma^2 := C\,(1+C)$

$$\|U\|_\infty \;\leq\; \gamma L h_n \|V - \widetilde{V}\|_\infty \,.$$

□

Der folgende Satz liefert die Existenz von Lösungen für den Fall von dissipativen AWA gemäß Definition 3.14 für beliebig grosse Schrittweiten h:

Satz 6.10 *Die AWA (2.1) sei dissipativ mit $l(t) \leq -\alpha < 0$ für $t \in I$ und f sei lipschitzstetig. Dann besitzt das DG-Verfahren (6.5) für beliebige Schrittweiten $h_n > 0$ stets eine eindeutige Lösung.*

Beweis. Ausgangspunkt ist die lokale Galerkin-Gleichung (6.7). Wir definieren

$$g : V_r(I_n) \to V_r(I_{n-1})\,, \qquad U_n \;\mapsto\; g(U_n) :\equiv U_{n-1}^- \,.$$

$g(U_n)$ ist also ein konstantes Polynom, dessen Wert gegeben ist durch die lokale Galerkin-Gleichung

$$\int_{t_{n-1}}^{t_n} \langle U'(s) - f(s,U(s)), \varphi(s)\rangle\, ds + \langle U_{n-1}^+, \varphi_{n-1}^+\rangle \;=\; \langle U_{n-1}^-, \varphi_{n-1}^+\rangle \,.$$

Andersherum ist zu gegebenem U_{n-1}^-, U_n Lösung der Gleichung $g(U_n) = U_{n-1}$. Wir zeigen nun für g die strikte Monotonieeigenschaft gemäß Definition 3.16 bezüglich des folgenden Skalarproduktes auf $V_r(I_n)$:

$$\langle U, V\rangle_{I_n} \;:=\; \int_{t_{n-1}}^{t_n} \langle U(s), V(s)\rangle\, ds + \langle U_{n-1}^+, V_{n-1}^+\rangle \,.$$

Aufgrund der einseitigen Lipschitz-Bedingung für f erhalten wir für $x, y \in V_r(I_n)$ und $e := x - y$:

$$\begin{aligned}
\langle g(x) - g(y), e\rangle_{I_n} &= \int_{t_{n-1}}^{t_n} \langle e' + f(s,y) - f(s,x), e\rangle\, ds + \|e_{n-1}^+\|^2 \\
&= \frac{1}{2}\int_{t_{n-1}}^{t_n} \frac{d}{dt}\|e\|^2\, ds - \int_{t_{n-1}}^{t_n} \langle f(s,x) - f(s,y), e\rangle\, ds + \|e_{n-1}^+\|^2 \\
&\geq \frac{1}{2}\int_{t_{n-1}}^{t_n} \frac{d}{dt}\|e\|^2\, ds - \int_{t_{n-1}}^{t_n} l(t)\|e\|^2\, ds + \|e_{n-1}^+\|^2 \\
&\geq \frac{1}{2}\left(\|e_n^-\|^2 - \|e_{n-1}^+\|^2\right) + \alpha\int_{t_{n-1}}^{t_n} \|e\|^2\, ds + \|e_{n-1}^+\|^2 \\
&\geq \frac{1}{2}\left(\|e_n^-\|^2 + \|e_{n-1}^+\|^2\right) + \alpha\int_{t_{n-1}}^{t_n} \|e\|^2\, ds \,.
\end{aligned}$$

Der verbleibende Ausdruck auf der rechten Seite stellt eine Norm auf dem Raum $V_r(I_n)$ dar. Da in diesem endlich-dimensionalen Raum alle Normen äquivalent sind, folgt die Monotonie

$$\langle g(x) - g(y), e\rangle_{I_n} \quad \geq \quad \gamma \|e\|^2_{V_r(I_n)} \,,$$

für jede beliebige Normen auf $V_r(I_n)$. Die (von der jeweiligen Norm abhängige) Konstante γ ist stets positiv, $\gamma > 0$. Die eindeutige Lösbarkeit folgt nun aus Satz 3.17. □

6.4 A priori Fehleranalyse

6.4.1 Das DG(0)-Verfahren zur Berechnung von Stammfunktionen

Wir betrachten hier exemplarisch eine AWA, bei der die rechte Seite f unabhängig ist von u, also:

$$\begin{aligned} u'(t) &= f(t) \qquad t \in I = (0,T] \,, \\ u(0) &= u_0 \,. \end{aligned}$$

In diesem Fall ist u die Stammfunktion von f. Diese AWA ist dissipativ gemäß Definition 3.14 mit Konstante $l = 0$, denn $\langle f(t) - f(t), x - y\rangle = 0$. Daher liefert die Verwendung des impliziten Euler Verfahrens mit Satz 3.19 folgende Abschätzung für den Fehler an den Gitterpunkten, $e_n := u(t_n) - u_n$,

$$\max_{1\leq n\leq N} \|e_n\| \quad \leq \quad \frac{1}{2} T \max_{1\leq n\leq N} \left\{ h_n \|u''\|_{L^\infty(I_n)} \right\} \,. \tag{6.10}$$

Wir erhalten also für das impliziten Euler Verfahren eine obere Schranke für den Fehler, der linear mit der Zeit T wächst und in den die zweiten Ableitungen von u, bzw. die ersten Ableitungen von f, einfließen.

Wir vergleichen dies nun mit dem DG(0)-Verfahren. Bei exakter Integration (was in der Praxis nicht immer möglich ist) gilt (mit der Konvention $U_n = U_n^-$)

$$U_n \quad = \quad U_{n-1}^- + \int_{I_n} f(s)\, ds \,.$$

Somit ergibt sich rekursiv an allen Gitterpunkten

$$e_n \quad = \quad u_n - U_n \;=\; u_{n-1} + \int_{I_n} f(s)\, ds - U_n \;=\; u_{n-1} - U_{n-1} \;=\; u_0 - U_0 \;=\; e_0 \,.$$

Wenn der Anfangsfehler verschwindet, $e_0 = 0$, so verschwindet der Fehler also an allen Gitterpunkten. Diese (schöne) Eigenschaft gilt allerdings nur unter der Annahme der obigen rechten Seite f, die unabhängig von u ist. Später werden wir aber sehen, dass sich an den Gitterpunkten eine sogenannte *Superkonvergenz* einstellt, d.h. an diesen speziellen

Punkten ist die Ordnung höher als die bloße Interpolation erwarten lässt. Innerhalb der Intervalle, $t \in I_n$, gilt jedoch

$$u(t) - U(t) \;=\; u_n - \int_t^{t_n} f(s)\,ds - U_n \;=\; -\int_t^{t_n} f(s)\,ds\,,$$

und somit gilt bei exakter Integration

$$\|e(t)\| \;\leq\; h_n \|f\|_{L^\infty(I_n)} \;=\; h_n \|u'\|_{L^\infty(I_n)}\,.$$

Im Vergleich zum Fehler beim Euler-Verfahren (6.10) erhalten wir zwar die gleiche Konvergenzordnung in Bezug auf die Schrittweite h_n (nämlich 1. Ordnung), aber es geht der lineare Faktor der Zeit nicht ein. Bei einer numerischen Quadratur des Integrals gilt dies nicht mehr, da stets ein Fehleranteil auftauchen wird, der mit der Zeit anwachsen kann. Ein weiterer Unterschied zwischen dem impliziten Euler und DG(0) ist, dass nur die erste Ableitung von u anstelle der zweiten Ableitung einfließen. Dies gilt auch für das DG(0)-Verfahren mit numerischer Quadratur. Bei der Verwendung der Mittelpunktregel:

$$U_n \;=\; U_{n-1}^- + h_n f(t_{n-1/2})\,,$$

mit $t_{n-1/2} := (t_{n-1} + t_n)/2$, lautet der Quadraturfehler bekanntermaßen

$$\left\| \int_{I_n} f(s)\,ds - h_n f(t_{n-1/2}) \right\| \;\leq\; \frac{1}{24} h_n^3 \|f''\|_{L^\infty(I_n)}\,.$$

Damit ergibt sich

$$\|e_n\| \;\leq\; \|e_{n-1}\| + \frac{1}{24} h_n^3 \|u'''\|_{L^\infty(I_n)}\,.$$

Durch Rekursion erhält man an den Gitterpunkten

$$\begin{aligned} \|e_n\| &\leq \|e_0\| + \frac{1}{24} \sum_{n=1}^{N} h_n^3 \|u'''\|_{L^\infty(I_n)} \\ &\leq \|e_0\| + \frac{1}{24} T \max_{1\leq n\leq N} \left(h_n^2 \|u'''\|_{L^\infty(I_n)} \right)\,. \end{aligned}$$

In Kombination mit dem Fehler innerhalb der Teilintervalle ergibt sich damit:

$$\|e\|_{L^\infty(I)} \;\leq\; \|e_0\| + \max_{1\leq n\leq N} h_n \left(\|u'\|_{L^\infty(I_n)} + \frac{1}{24} T h_n \|u'''\|_{L^\infty(I_n)} \right)\,.$$

Für verschwindenden Anfangsfehler und hinreichend kleine Schrittweite,

$$h_n \;\leq\; \frac{24 \|u'\|_{L^\infty(I_n)}}{T \|u'''\|_{L^\infty(I_n)}}\,,$$

erhält man insbesondere

$$\|e\|_{L^\infty(I)} \;\leq\; 2 \max_{1\leq n\leq N} \left(h_n \|u'\|_{L^\infty(I_n)} \right)\,,$$

also wieder die Ordnung $\mathcal{O}(h)$, aber mit geringerer Regularitätsanforderung an f als für Einschrittverfahren.

6.4.2 Galerkin Orthogonalität

Satz 6.11 *Sei $u \in V(\mathcal{T}_h)$ die Lösung des (kontinuierlichen) Variationsproblems (6.2) und $U \in V_r(\mathcal{T}_h)$ die (diskrete) DG-Lösung von (6.5). Dann gilt die Galerkin Orthogonalität:*

$$A(u;\varphi) - A(U;\varphi) \;=\; 0 \qquad \forall \varphi \in V_r(\mathcal{T}_h)\,.$$

Im Fall einer linearen AWA, also $f(t,u) = \mathcal{A}(t)u(t) + b(t)$, mit einer Matrix-wertigen Funktion $\mathcal{A}(t)$, gilt sogar

$$A(u-U;\varphi) \;=\; 0 \qquad \forall \varphi \in V_r(\mathcal{T}_h)\,.$$

Beweis. (a) Dies ist eine unmittelbare Folgerung aus $V_r(\mathcal{T}_h) \subset V(\mathcal{T}_h)$ und der Subtraktion der beiden Gleichungen (6.2) und (6.5).
(b) Im Fall einer linearen AWA gilt

$$f(t,u(t)) - f(t,U(t)) \;=\; \mathcal{A}(t)(u(t)-U(t))\,.$$

Die Behauptung ergibt sich nun aus (a) und der Bilinearform A in (6.3). ∎

6.4.3 A priori Abschätzung für nicht dissipative Probleme

Zunächst benötigen wir ein Hilfresultat, dass wir im folgenden auf jedes Teilintervall anwenden werden, also $J = I_n$.

Satz 6.12 *Sei $J = [a,b]$, $r \in \mathbb{N}_0$ und $u \in C^{r+1}(J)$. Es existiert genau ein Polynom $p \in P_r$ mit der Eigenschaft*

$$p(b) \;=\; u(b) \quad \textit{und} \quad \int_J (p-u)q\,dx \;=\; 0 \quad \forall q \in P_{r-1}\,.$$

Ferner besitzt der Fehler $u-p$ in J mindestens $r+1$ Nullstellen und erfüllt die Abschätzung

$$\|u-p\|_{L^\infty(a,b)} \;\leq\; \frac{1}{(r+1)!}(b-a)^{r+1}\|u^{(r+1)}\|_{L^\infty(a,b)}\,.$$

Im Fall $r = 0$ ist die Orthogonalitätsbedingung keine tatsächliche Bedingung, da P_{-1} nur das Nullpolynom enthält.

Beweis. (a) Die Behauptung ist für $r = 0$ trivial, womit wir im folgenden $r \geq 1$ annehmen können. Ausgehend von einer Basis von P_r führt die Orthogonalitätsbedingung auf ein lineares Gleichungssystem mit $r+1$ Freiheitsgraden und r Bedingungen. Die Normierungsbedingung $p(b) = u(b)$ ist eine weitere lineare Gleichung, so dass man insgesamt ein quadratisches LGS erhält. Für Existenz und Eindeutigkeit genügt es daher zu zeigen, dass das homogene System, also $u \equiv 0$, nur die triviale Lösung $p \equiv 0$ erlaubt. Sei also p eine

solche Lösung. Dieses Polynom besitzt dann einer Nullstelle im Punkt b. Folglich können wir diese Abspalten, so dass ein Polynom $\widetilde{p} \in P_{r-1}$ existiert mit der Darstellung

$$p(x) \;=\; \widetilde{p}(x)(b-x)\,.$$

Nun wählen wir $q := \widetilde{p}$ in der Orthogonalitätsbedingung, also

$$0 \;=\; \int_J pq\,dx \;=\; \int_a^b (b-x)\widetilde{p}^2(x)\,dx\,.$$

Da der Integrand nicht-negativ und stetig ist, folgt $(b-x)\widetilde{p}^2(x) = 0$ für alle $x \in [a,b]$. Dies impliziert $\widetilde{p} \equiv 0$ und in seiner Konsequenz auch $p \equiv 0$.
(b) Analog zu (a) folgern wir zunächst, dass $p-u$ mindestens $r+1$ Nullstellen in J besitzt: Gebe es lediglich $k \le r$ Nullstellen $\xi_1, \ldots, \xi_r \in J$, so ließe sich $p-u$ in J darstellen in der Form

$$(p-u)(x) \;=\; \prod_{i=1}^{k}(x-\xi_i)g(x)\,,$$

mit einer oEdA stetigen positiven Funktion $g \ge 0$. Das Polynom $q(x) = \prod_{i=1}^{k}(x-\xi_i)$ liegt in P_{r-1}, so dass aus

$$0 \;=\; \int_J (p-u)q\,dx \;=\; \int_J \prod_{i=1}^{k}(x-\xi_i)^2 g(x)\,dx$$

$g \equiv 0$ und folglich $p = u$ folgt.
(c) Aus (b) folgt, dass p auch als Knoteninterpolierende von u an $r+1$ Punkten aufgefasst werden kann. Für polynomiale Knoteninterpolation kennen wir die Approximationsgüte:

$$\|u-p\|_{L^\infty(a,b)} \;\le\; \frac{1}{(r+1)!}(b-a)^{r+1}\|u^{(r+1)}\|_{L^\infty(a,b)}\,.$$

□

Satz 6.13 *Im Fall einer lipschitz-stetigen AWA mit Lipschitz-Konstante L_f gilt für den Fehler $e = u-U$ des DG(r)-Verfahrens für hinreichend kleine Schrittweite $h_n < (2C^2\gamma L_f)^{-1}$:*

$$\|e\|_{L^\infty(I)} \;\le\; C_{L_f,r,T} \max_{1\le n\le N}\left(h_n^{r+1}\|u^{(r+1)}\|_{L^\infty(I_n)}\right)\,,$$

mit einer nur von L_f,r und T abhängigen Konstante $C_{L_f,r,T}$.

Beweis. Wie für Galerkin-Verfahren typisch spalten wir den Fehler e auf in einen Interpolationsfehler $\xi := u - I_h u$ und einen sogenannten *Projektionsfehler* $\eta := I_h u - U$. Aufgrund der Dreiecksungleichung gilt

$$\|e\|_{L^\infty(I)} \;\le\; \|\xi\|_{L^\infty(I)} + \|\eta\|_{L^\infty(I)}\,.$$

Als Interpolation wählen wir $I_h u \in V_r(\mathcal{T}_h)$ folgendermaßen: $I_h u|_{I_n}$ ist gerade das eindeutig definierte Polynom aus Satz 6.12. Somit gilt

$$\begin{aligned}\|\xi\|_{L^\infty(I)} &\leq \max_{1\leq n\leq N} \|u - I_h u\|_{L^\infty(I_n)} \\ &\leq \max_{1\leq n\leq N} \left(\frac{1}{(r+1)!} h_n^{r+1} \|u^{(r+1)}\|_{L^\infty(I_n)} \right) .\end{aligned}$$

Es genügt daher den Projektionsfehler $\|\eta\|$ noch entsprechend zu beschränken. Zur Behandlung des Projektionsfehlers ist wichtig, dass er im diskreten Ansatzraum liegt, also $\eta \in V_r(\mathcal{T}_h)$. Wir verwenden nun die diskrete Sobolev'sche Ungleichung aus Satz 6.8 auf jedem Teilintervall (mit der Notation $\eta_n^- := \eta(t_n)^-$):

$$\|\eta\|_{L^\infty(I_n)} \leq C \left(\int_{I_n} |\eta'(t)|^2 (t-a)\, dt + |\eta_n^-|^2 \right)^{1/2} .$$

(a) Abschätzung von $|\eta_n^-|^2$: Um die beiden auftretenden Terme auf der rechten Seite entsprechend zu kontrollieren, verwenden wir zweimal partielle Integration sowie die Tatsache $\xi_n^- = 0$:

$$\begin{aligned}\int_{I_n} \langle (I_h u)', \varphi \rangle\, dt + \langle (I_h u)_{n-1}^+, \varphi_{n-1}^+ \rangle &= -\int_{I_n} \langle I_h u, \varphi' \rangle\, dt + \langle (I_h u)_n^-, \varphi_n^- \rangle \\ &= -\int_{I_n} \langle u, \varphi' \rangle\, dt + \langle (I_h u)_n^-, \varphi_n^- \rangle \\ &= \int_{I_n} \langle u', \varphi \rangle\, dt - \langle \xi_n^-, \varphi_n^- \rangle + \langle u_{n-1}, \varphi_{n-1}^+ \rangle \\ &= \int_{I_n} \langle f(t,u), \varphi \rangle\, dt + \langle u_{n-1}, \varphi_{n-1}^+ \rangle .\end{aligned}$$

Die Galerkin Gleichung (6.7) besagt gerade

$$\int_{I_n} \langle U', \varphi \rangle\, dt + \langle U_{n-1}^+, \varphi_{n-1}^+ \rangle = \int_{I_n} \langle f(t,U), \varphi \rangle\, dt + \langle U_{n-1}^-, \varphi_{n-1}^+ \rangle .$$

Die Subtraktion dieser beiden Gleichungen liefert dann wegen $u_{n-1} - U_{n-1}^- = (I_h u)_{n-1}^- - U_{n-1}^- = \eta_{n-1}^-$:

$$\int_{I_n} \langle \eta', \varphi \rangle\, dt + \langle \eta_{n-1}^+, \varphi_{n-1}^+ \rangle = \int_{I_n} \langle f(t,u) - f(t,U), \varphi \rangle\, dt + \langle \eta_{n-1}^-, \varphi_{n-1}^+ \rangle .$$

Testen dieser Gleichung mit $\varphi := \eta$ ergibt

$$\begin{aligned}\int_{I_n} \langle \eta', \eta \rangle\, dt + |\eta_{n-1}^+|^2 &= \int_{I_n} \langle f(t,u) - f(t,U), \eta \rangle\, dt + \langle \eta_{n-1}^-, \eta_{n-1}^+ \rangle \\ &\leq L_f \int_{I_n} |u-U||\eta|\, dt + \frac{1}{2}|\eta_{n-1}^-|^2 + \frac{1}{2}|\eta_{n-1}^+|^2 .\end{aligned}$$

Wegen $\int_{I_n}\langle\eta',\eta\rangle = \frac{1}{2}\int_{I_n}\frac{d}{dt}(|\eta|^2) = \frac{1}{2}(|\eta_n^-|^2 - |\eta_{n-1}^+|^2)$ folgt

$$|\eta_n^-|^2 \;\leq\; 2L_f\int_{I_n}|u-U||\eta|\,dt + |\eta_{n-1}^-|^2\,.$$

Per Rekursion gelangen wir für beliebiges $\epsilon > 0$ zu

$$\begin{aligned}|\eta_n^-|^2 &\leq 2L_f\int_0^{t_n}|u-U||\eta|\,dt \;\leq\; 2L_f\sum_{k=1}^{n}h_k\|u-U\|_{L^\infty(I_k)}\|\eta\|_{L^\infty(I_k)}\\ &\leq \epsilon^{-1}L_f^2\sum_{k=1}^{n}h_k^2\|u-U\|^2_{L^\infty(I_k)} + \epsilon\|\eta\|^2_{L^\infty(t_0,t_n)}\,.\end{aligned}$$

(b) Abschätzung von $\int_{I_n}|\eta'(t)|^2(t-a)\,dt$: Wir setzen $\varphi := \eta'(t-t_{n-1})$ und erhalten nun wegen $\varphi_{n-1}^+ = 0$ aus () und der Young'schen Ungleichung:

$$\begin{aligned}\int_{I_n}|\eta'|^2(t-t_{n-1})\rangle\,dt &= \int_{I_n}\langle f(t,u)-f(t,U),\eta'(t-t_{n-1})\rangle\,dt\\ &\leq \int_{I_n}L_f|u-U||\eta'|(t-t_{n-1})\,dt\\ &\leq \frac{1}{2}\int_{I_n}L_f^2|u-U|^2(t-t_{n-1})\,dt + \frac{1}{2}\int_{I_n}|\eta'|^2(t-t_{n-1})\,dt\,.\end{aligned}$$

Also:

$$\begin{aligned}\int_{I_n}|\eta'|^2(t-t_{n-1})\rangle\,dt &\leq L_f^2\int_{I_n}|u-U|^2(t-t_{n-1})\,dt\\ &\leq \frac{1}{2}L_f^2h_n^2\|u-U\|^2_{L^\infty(I_n)}\,.\end{aligned}$$

(c) Die Kombination aus (a) und (b) ergibt:

$$\|\eta\|^2_{L^\infty(I)} \;\leq\; C^2\left(\frac{1}{2}L_f^2\sup_{1\leq n\leq N}h_n^2\|u-U\|^2_{L^\infty(I_n)} + \epsilon^{-1}L_f^2\sum_{k=1}^{N}h_k^2\|u-U\|^2_{L^\infty(I_k)} + \epsilon\|\eta\|^2_{L^\infty(I)}\right).$$

Durch die Wahl $\epsilon := \frac{1}{2}C^{-2}$ erhalten wir

$$\|\eta\|^2_{L^\infty(I)} \;\leq\; C^2L_f^2\left(\sup_{1\leq k\leq N}h_k^2\|u-U\|^2_{L^\infty(I_k)} + C^2\sum_{k=1}^{N}h_k^2\|u-U\|^2_{L^\infty(I_k)}\right).$$

Für Schrittweite $h_k < (2CL_f)^{-1}$

$$\|\eta\|^2_{L^\infty(I)} \;\leq\; \frac{1}{4}\|u-U\|^2_{L^\infty(I)} + C^4L_f^2\sum_{k=1}^{N}h_k^2\|u-U\|^2_{L^\infty(I_k)}\,.$$

(d) Für den Gesamtfehler erhalten wir

$$\begin{aligned}\|u-U\|^2_{L^\infty(I)} &\leq 2\|\xi\|^2_{L^\infty(I)} + 2\|\eta\|^2_{L^\infty(I)} \\ &\leq 2\|\xi\|^2_{L^\infty(I)} + \frac{1}{2}\|u-U\|^2_{L^\infty(I)} + 2C^4L_f^2\sum_{k=1}^{N} h_k^2\|u-U\|^2_{L^\infty(I_k)}\,.\end{aligned}$$

bzw. unter der Bedingung $h_k < (2C^2L_f)^{-1}$:

$$\begin{aligned}\|u-U\|^2_{L^\infty(I_n)} &\leq 4\|\xi\|^2_{L^\infty(I)} + 4C^4L_f^2\sum_{k=1}^{N} h_k^2\|u-U\|^2_{L^\infty(I_k)} \\ &\leq 4\|\xi\|^2_{L^\infty(I)} + 4C^4L_f^2\sum_{k=1}^{n-1} h_k^2\|u-U\|^2_{L^\infty(I_k)} + \frac{1}{2}\|u-U\|^2_{L^\infty(I_n)}\,.\end{aligned}$$

Nun wenden wir noch das diskrete Gronwall'sche Lemma an und erhalten mit $h := \max h_k$:

$$\begin{aligned}\|u-U\|^2_{L^\infty(I)} &\leq 8\exp\left(8C^4L_f^2Th\right)\|\xi\|^2_{L^\infty(I)} \\ &\leq 8\exp\left(4C^2L_fT\right)\|\xi\|^2_{L^\infty(I)}\,.\end{aligned}$$

Zusammen mit der Abschätzung des Interpolationsfehlers $\|\xi\|_{L^\infty(I)}$ erhalten wir die Behauptung mit der Konstanten $C_{L_f,r,T} := \frac{3}{(r+1)!}\exp\left(2C^2L_fT\right)$. ∎

6.5 A posteriori Fehlerkontrolle

6.5.1 Duales Problem

Sei u die Lösung der AWA (2.1) und U die zugehörige DG(r)-Lösung. Zudem bezeichne $Df_x(t,x)$ die Jacobi-Matrix von f bzgl. x an der Stelle (t,x). Der Diskretisierungsfehler sei wir gewohnt bezeichnet mit $e := u-U$. Wir definieren uns zunächst die Matrix-wertige Funktion $B_{u,U}(t)$ mittels

$$B_{u,U}(t)W(t) := \int_0^1 Df_x(t,U(t)+\lambda e(t))W(t)\,d\lambda\,.$$

Lemma 6.14 *Es gilt:*

$$B_{u,U}(t)e(t) = f(t,u(t)) - f(t,U(t))\,.$$

Beweis. Sei $t \in I$ fest aber beliebig. Dann definieren wir

$$\begin{aligned}g(\lambda) &:= Df_x(t,U(t)+\lambda e(t))e(t)\,, \\ G(\lambda) &:= f(t,U(t)+\lambda e(t))\,.\end{aligned}$$

Offensichtlich ist G eine Stammfunktion von f, denn $G'(\lambda) = g(\lambda)$. Dann folgt nach dem Hauptsatz der Integral- und Differentialrechnung

$$B_{u,U}(t)e(t) \;=\; \int_0^1 g(\lambda)d\lambda \;=\; G(1) - G(0) \;=\; f(t,u(t)) - f(t,U(t))\,.$$

□

Nun betrachten wir die folgende lineare AWA, die eigentlich ein Rückwärtsproblem darstellt, denn es sind Enddaten gegeben:

$$\begin{aligned} -z'(t) &= B_{u,U}(t)^* z\,, \qquad 0 \le t \le T\,, && (6.11)\\ z(T) &= e_N^-/\|e_N^-\|\,. && (6.12) \end{aligned}$$

Die Enddaten sind also gegeben durch den Fehler des Vorwärts-Problems zum Endzeitpunkt. Um diese AWA auch diskret zu formulieren betrachten wir die Bilinearform

$$L_{u,U}(W,\varphi) \;:=\; \sum_{n=1}^{N} \int_{I_n} \langle W' - B_{u,U}W, \varphi\rangle\, dt + \langle W_0^+, \varphi_0^+\rangle + \sum_{n=1}^{N-1} \langle [W]_n, \varphi_n^+\rangle\,.$$

Die adjungierte Bilinearform lautet nun nach partieller Integration

$$\begin{aligned} & L_{u,U}^*(Z,\varphi)\\ =\;& L_{u,U}(\varphi, Z)\\ =\;& \sum_{n=1}^{N} \int_{I_n} \langle Z, \varphi' - B_{u,U}\varphi\rangle\, dt + \langle \varphi_0^+, Z_0^+\rangle + \sum_{n=1}^{N-1} \langle [\varphi]_n, Z_n^+\rangle\\ =\;& \sum_{n=1}^{N} \left(\int_{I_n} \langle -Z' - B_{u,U}^* Z, \varphi\rangle\, dt + \langle Z_n^-, \varphi_n^-\rangle - \langle Z_{n-1}^+, \varphi_{n-1}^+\rangle \right) + \langle Z_0^+, \varphi_0^+\rangle + \sum_{n=1}^{N-1} \langle [\varphi]_n, Z_n^+\rangle\\ =\;& \sum_{n=1}^{N} \int_{I_n} \langle -Z' - B_{u,U}^* Z, \varphi\rangle\, dt + \langle Z_1^-, \varphi_1^-\rangle + \sum_{n=1}^{N-1} \left(\langle Z_{n+1}^-, \varphi_{n+1}^-\rangle - \langle \varphi_n^-, Z_n^+\rangle \right)\\ =\;& \sum_{n=1}^{N} \int_{I_n} \langle -Z' - B_{u,U}^* Z, \varphi\rangle\, dt - \sum_{n=1}^{N-1} \langle [Z_n], \varphi_n^-\rangle + \langle Z_N^-, \varphi_N^-\rangle\,. \end{aligned}$$

Die variationelle Formulierung von (6.11)-(6.12) lautet nun folgendermaßen: Gesucht $z \in V(\mathcal{T}_h)$ mit $z_N^+ = e_N^-/\|e_N^-\|$ und

$$L_{u,U}^*(z,\varphi) \;=\; \langle z_N^+, \varphi_N^-\rangle \qquad \forall \varphi \in V(\mathcal{T}_h)\,. \tag{6.13}$$

Die zugehörige diskrete DG(r)-Lösung Z von (6.11)-(6.12) ist gegeben durch $Z \in V_r(\mathcal{T}_h)$ mit $Z_N^+ = e_N^-/\|e_N^-\|$ und die Gleichungen

$$L_{u,U}^*(Z,\varphi) \;=\; \langle Z_N^+, \varphi_N^-\rangle \qquad \forall \varphi \in V_r(\mathcal{T}_h)\,.$$

6.5.2 A posteriori Fehlerdarstellung

Lemma 6.15 *Sei $z \in V(\mathcal{T}_h, u_0)$ die Lösung des dualen Problems (6.13). Dann gilt für den Fehler $e = u - U$ zum Endzeitpunkt im primalen Problem*

$$|e_N^-| \;=\; -A(U; z - I_h z)\,,$$

für eine beliebige Interpolierende $I_h z \in V_r(\mathcal{T}_h)$.

Beweis. Wir wählen als Testfunktion $\varphi := e$ in (6.13) und erhalten

$$\begin{aligned}
|e_N^-| &= \langle z_N^+, e_N^- \rangle \;=\; L^*_{u,U}(z,e) \;=\; L_{u,U}(e,z) \\
&= \sum_{n=1}^{N} \int_{I_n} \langle z' - B_{u,U} z, \varphi \rangle \, dt + \langle e_0^+, \varphi_0^+ \rangle + \sum_{n=1}^{N-1} \langle [e]_n, z_n^+ \rangle \\
&= \sum_{n=1}^{N} \int_{I_n} \langle e' - f(t,u) - f(t,U), z \rangle \, dt + \langle e_0^+, z_0^+ \rangle + \sum_{n=1}^{N-1} \langle [e]_n, z_n^+ \rangle \\
&= A(u;z) - A(U;z)\,.
\end{aligned}$$

Im letzten Schritt haben wir $e_0^- = 0$ ausgenutzt. Da $A(u;z) = 0$ und mit der Galerkin-Orthogonalität folgt die Behauptung. ∎

Dieser Satz besagt also, dass wir den Fehler berechnen können mittels Kenntnis von U und z. Hierbei sei angemerkt, dass z im Fall von einem nichtlinearen f auch von u abhängt.

6.5.3 A posteriori Fehlerschranke

Nun wollen wir eine obere Schranke für den Fehler erstellen. Hierzu verwenden wir für das Residuum der AWA die Bezeichnung

$$\varrho(U)(t) \;:=\; f(t,U(t)) - U'(t)\,.$$

Außerdem verwenden wir zwei Arten von L^2-Projektionen $\Pi_r, \widetilde{\Pi}_r : C(I) \to V_r(\mathcal{T}_h)$. Diese sind dadurch defininiert, dass auf jedem Teilintervall I_n gilt:

$$\begin{aligned}
&\int_{I_n} \langle u - \Pi_r u, q \rangle \;=\; 0 \qquad \forall q \in P_r\,, \\
(\widetilde{\Pi}_r u)_{n-1}^+ = u_{n-1}^+ \quad \text{und} \quad &\int_{I_n} \langle u - \widetilde{\Pi}_r u, q \rangle \;=\; 0 \qquad \forall q \in P_{r-1}\,,
\end{aligned}$$

Satz 6.16 *Im Fall einer lipschitz-stetigen AWA gilt für den Fehler $e = u - U$ zwischen der klassischen Lösung u und der DG(r)-Lösung U:*

$$\begin{aligned}
&(a) \qquad |e_N^-| \;\le\; \sum_{n=1}^{N} \|\varrho(U) - \Pi_{r-1}\varrho(U)\|_{L^\infty(I_n)} \int_{I_n} |z - \widetilde{\Pi}_r z| \, dt\,, \\
&(b) \qquad |e_N^-| \;\le\; C \max_{n=1\ldots N} \left(h_n^{r+1} \|\varrho(U) - \Pi_{r-1}\varrho(U)\|_{L^\infty(I_n)} \right)\,, \\
&(c) \qquad |e_N^-| \;\le\; C \max_{n=1\ldots N} h_n^r \left(h_n \|\varrho(U) - \Pi_r \varrho(U)\|_{L^\infty(I_n)} + |[U]_{n-1}| \right)\,,
\end{aligned}$$

mit von r, L_f und T abhängigen Konstanten C.

Beweis. (a) Wir gehen aus von der Fehlerdarstellung aus Lemma 6.15 und wählen als Interpolation $I_h := \widetilde{\Pi}_r$:

$$\begin{aligned} |e_N^-| &= -A(U; z - \widetilde{\Pi}_r z) \\ &= \sum_{n=1}^{N} \int_{I_n} \langle \varrho(U), z - \widetilde{\Pi}_r z \rangle \, dt + \sum_{n=0}^{N-1} \langle [U]_n, (z - \widetilde{\Pi}_r z)_n^+ \rangle \, . \end{aligned}$$

Aufgrund von $z_n^+ = \widetilde{\Pi}_r z_n^+$ gilt und der Orthonalitätsforderung an $\widetilde{\Pi}_r$ erhalten wir:

$$\begin{aligned} |e_N^-| &= \sum_{n=1}^{N} \int_{I_n} \langle \varrho(U) - \Pi_{r-1}\varrho(U), z - \widetilde{\Pi}_r z \rangle \, dt \\ &\leq \sum_{n=1}^{N} \|\varrho(U) - \Pi_{r-1}\varrho(U)\|_{L^\infty(I_n)} \int_{I_n} |z - \widetilde{\Pi}_r z| \, dt \, . \end{aligned}$$

(b) Man kann zeigen:

$$\int_{I_n} |z - \widetilde{\Pi}_r z| \, dt \;\leq\; C_I h_n^{r+1} \int_{I_n} |z^{(n+1)}| \, dt \;\leq\; C_I C_S h_n^{r+2} \, .$$

Hierbei bezeichnet C_I eine Interpolationskonstante und C_S eine Stabilitätskonstante.
(c) folgt durch Verwendung von Lemma 6.15 und der Interpolation $I_h := \Pi_r$:

$$\begin{aligned} |e_N^-| &= -A(U; z - \Pi_r z) \\ &= \sum_{n=1}^{N} \int_{I_n} \langle \varrho(U), z - \Pi_r z \rangle \, dt + \sum_{n=0}^{N-1} \langle [U]_n, (z - \Pi_r z)_n^+ \rangle \, . \end{aligned}$$

Wegen $\langle \Pi_r \varrho(U), z - \Pi_r z \rangle = 0$ und ... folgt

$$|e_N^-| \;\leq\; \sum_{n=1}^{N} \|\varrho(U) - \Pi_r \varrho(U)\|_{L^\infty(I_n)} \int_{I_n} |z - \Pi_r z| \, dt + \sum_{n=0}^{N-1} |[U]_n| |(z - \Pi_r z)_n^+| \, .$$

Die analoge Approximationseigenschaften von Π_r wie in (b):

$$\begin{aligned} \int_{I_n} |z - \Pi_r z| \, dt &\leq \widetilde{C}_I C_S h_n^{r+2} \, , \\ |(z - \Pi_r z)_n^+| &\leq \|z - \Pi_r z\|_{L^\infty(I_{n+1})} \;\leq\; \widetilde{C}_I C_S h_{n+1}^{r+1} \, , \end{aligned}$$

führen nun auf

$$\begin{aligned} |e_N^-| &\leq \widetilde{C}_I C_S \sum_{n=1}^{N} h_n^{r+1} \Big(h_n \|\varrho(U) - \Pi_r \varrho(U)\|_{L^\infty(I_n)} + |[U]_{n-1}| \Big) \\ &\leq \widetilde{C}_I C_S T \max_{1 \leq n \leq N} h_n^r \Big(h_n \|\varrho(U) - \Pi_r \varrho(U)\|_{L^\infty(I_n)} + |[U]_{n-1}| \Big) \, . \end{aligned}$$

□

Durch genauere Analyse kann man zeigen, dass $h_n \| \varrho(U) - \Pi_r \varrho(U) \|_{L^\infty(I_n)}$ asymptotisch sehr viel kleiner ist als $|[U]_{n-1}|$. Daher wird in der Praxis häufig der vereinfachte Fehlerschätzer

$$|e_N^-| \quad \approx \quad \eta \; := \; \widetilde{C}_I C_S \sum_{n=0}^{N-1} |[U]_n|$$

verwendet. Eine adaptive Schrittweitensteuerung basiert in diesem Fall einfach auf der Auswertung der Sprungterme $|[U]_n|$. Die Konstanten C_I und C_S sind für die lokale Schrittweitensteuerung dann unerheblich.

Kapitel 7

Randwertaufgaben

In den Anwendungen treten neben den Anfangswertaufgaben und Differential-Algebraischen Gleichungen noch andere Arten von Differentialgleichungen auf, insbesondere Randwertaufgaben. Hiervon gibt es zahlreiche Typen, die jede für sich charakteristische Schwierigkeiten beinhalten. Im Rahmen dieser Vorlesung können wir nur auf eine kleine Klasse solcher Randwertaufgaben eingehen, nämlich den Sturm-Liouville-Problemen.

7.1 Sturm-Liouville-Probleme

Unter einem Sturm-Liouville-Problem versteht man ein Randwertproblem 2. Ordnung in einer Raumdimension in einem Intervall I, dass wir hier der Einfachheit halber als $I := (0,1) \subset \mathbb{R}$ setzen, der Form

$$-(au')' + bu' + cu = f \qquad x \in I\,, \tag{7.1}$$

$$u(0) = \alpha\,, \quad u(1) = \beta\,. \tag{7.2}$$

Hierbei bezeichnen $a \in C^1(I)$ und $b, c, f \in C(I)$ Funktionen mit $a \geq a_0 > 0$ und $c \geq 0$. Die Parameter $\alpha, \beta \in \mathbb{R}$ sind beliebig. Gesucht ist ein $u \in C^2(I) \cap C(\overline{I})$, dass diese Gleichung löst.

7.2 Variationelle Formulierung

Wir setzen zunächst *homogene* Dirichletdaten voraus:

$$u(0) = u(1) = 0\,. \tag{7.3}$$

Wir benutzen nun wie gewohnt eine Unterteilung von I in N (jetzt offene) Teilintervalle $I_n = (t_{n-1}, t_n)$ und definieren hierzu den Funktionenraum der stetigen und stückweise stetig differenzierbaren Funktionen

$$\widetilde{V} := \{v \in C(I) : v|_{I_n} \in C^1(\overline{I}_n),\ v(0) = v(1) = 0\}\,.$$

Die Ableitungen können also an den Randpunkten der Teilintervalle stetig fortgesetzt werden. Auf diesem Raum definieren wir das Skalarprodunkt

$$(u,v) \;:=\; \int_I uv\,dx$$

sowie die Bilinearform

$$A(u,\varphi) \;:=\; (au',\varphi') + (bu' + cu, \varphi)\,.$$

Dadurch, dass die Ableitungen an den Teilintervallgrenzen jeweils (von einer Seite betrachtet stetig) fortgesetzt werden können, ist die Bilinearform für $u, \varphi \in \widetilde{V}$ wohldefiniert, selbst wenn die Ableitungen u' und φ' an den Teilintervallgrenzen nicht stetig sind. Die zugehörige variationelle Formulierung lautet:

$$u \in \widetilde{V}: \quad A(u,\varphi) \;=\; (f,\varphi) \qquad \forall \varphi \in \widetilde{V}\,. \tag{7.4}$$

Satz 7.1 *Jede klassische Lösung $u \in C^2(\overline{I})$ von (7.1)-(7.3) ist auch eine Lösung der variationellen Formulierung (7.4) und umgekehrt ist jede hinreichend reguläre variationelle Lösung $u \in C^2(\overline{I})$ von (7.4) auch klassische Lösung von (7.1)-(7.3).*

Beweis. Sei zunächst $u \in C^2(\overline{I})$ klassische Lösung. Dann gilt auch $u \in \widetilde{V}$. Durch partielle Integration und wegen $\varphi(0) = \varphi(1) = 0$ erhalten wir

$$\begin{aligned}
(-(au')', \varphi) &= -\int_I \langle (au')', \varphi\rangle\, dx \\
&= \int_I \langle au', \varphi'\rangle\, dx - a(1)u'(1)\varphi(1) + a(0)u'(0)\varphi(0) \\
&= \int_I \langle au', \varphi'\rangle\, dx \\
&= (au', \varphi')\,.
\end{aligned}$$

Zusammen mit der Multiplikation mit einer Testfunktion und Integration über I für die verbleibenden Terme erhalten wir (7.4).
Sei nun $u \in C^2(\overline{I})$ eine variationelle Lösung. Wir wählen jetzt Testfunktionen $\varphi \in C_0^\infty(I) \subset \widetilde{V}$ und erhalten wieder per partieller Integration

$$\int_I \langle -(au')' + bu' + cu, \varphi\rangle\, dx \;=\; -\int_I \langle f, \varphi\rangle\, dx \qquad \forall \varphi \in C_0^\infty(I)\,.$$

Da $C_0^\infty(I)$ dicht liegt in $C(\overline{I})$ bzgl. der L^2-Norm, folgt hieraus die Differentialgleichung (7.1) punktweise für alle $x \in I$. ∎

7.3 Schwache Ableitungen und der Sobolevraum $H^1(0,1)$

Wir wollen nun einen erweiterten Ableitungsbegriff einführen. Unter dem Träger $supp\,\varphi$ einer Funktion $\varphi : I \to \mathbb{R}$ für ein (nicht notwendigerweise beschränktes) reelles Intervall I verstehen wir die abgeschlossene Menge

$$supp\,\varphi \;:=\; \overline{\{x \in I \,:\, \varphi(x) \neq 0\}}\,.$$

Die Menge der $\mathcal{C}^\infty$-Funktionen mit kompaktem Träger bezeichnen wir im folgenden mit

$$\mathcal{D}(I) \;:=\; C_0^\infty(I) \;=\; \{u \in \mathcal{C}^\infty(I) \,:\, supp\,u \text{ kompakt}\}$$

Definition 7.2 *Unter der Menge der lokal L^1-integrierbaren Funktionen über eine Menge $\Omega \subset \mathbb{R}^n$ versteht man*

$$L^1_{loc}(\Omega) \;:=\; \{f : \Omega \to \mathbb{R} \,:\, f\chi_K \in L^1(\Omega) \quad \forall K \subset \Omega \text{ kompakt.}\}\,,$$

wobei χ_K die charakteristische Funktion auf der Menge K bezeichnet.

Definition 7.3 *Zu $u \in L^1_{loc}(I)$ heißt eine Funktion $w \in L^1_{loc}(I)$* verallgemeinerte/schwache Ableitung *von u, wenn*

$$\int_I w\varphi\,dx \;=\; -\int_I u\varphi'\,dx \qquad \forall \varphi \in \mathcal{D}(I).$$

Eine solche Funktion w bezeichnen wir dann mit u'.

Lemma 7.4 *Für $u \in \mathcal{C}^1(I)$ ist die verallgemeinerte Ableitung identisch mit der klassichen Ableitung.*

Beweis. Ergibt sich unmittelbar aus partieller Integration. □

Definition 7.5 *Unter dem Sobolev-Raum $H^1(I)$ versteht man*

$$H^1(I) \;:=\; \{u \in L^2(I) \,:\, \exists u' \in L^2(I)\}.$$

In dieser Definition ist u' selbstverständlich als schwache Ableitung zu verstehen.

Satz 7.6 *Der Raum $H^1(I)$ wird zusammen mit dem Skalarprodukt*

$$\langle u, v\rangle_{H^1(I)} \;:=\; (u,v) + (u',v')$$

zu einem Hilbertraum. Der Raum $C^1[0,1]$ liegt dicht in $H^1(I)$.

Die zugehörige Norm lautet

$$\|u\|_{H^1(I)} \;=\; \langle u,u\rangle_{H^1(I)}^{1/2} \;=\; \Big(\|u\|_{L^2(I)}^2 + \|u'\|_{L^2(I)}^2\Big)^{1/2}$$

Die variationelle Formulierung (7.4) lässt sich auf Funktionen aus $H^1(I)$ ebenso anwenden. Die Dirichletbedingungen an u erfordern es jedoch, dass man Punktwerte von u besitzt. Zunächst ist aber nicht klar, dass diese in H^1 überhaupt existieren, denn L^2-Funktionen besitzen nicht notwendigerweise Punktwerte. Man kann jedoch zeigen, dass der *Spuroperator* $\gamma : C^1[0,1] \to \mathbb{R}$, $\gamma u = u(0)$, folgende Abschätzung erfüllt

$$|\gamma u| \;\le\; C\|u\|_{H^1(I)} \qquad \forall u \in C^1[0,1]\,.$$

Da $C^1[0,1]$ in $H^1(I)$ dicht liegt, existiert eine eindeutige stetige Fortsetzung

$$\gamma : H^1(I) \to \mathbb{R}\,.$$

Die Stetigkeit besagt dann gerade

$$|u(0)| := |\gamma u| \;\le\; C\|u\|_{H^1(I)} \qquad \forall u \in H^1(I)\,.$$

Der Wert $u(0)$ ist daher zu verstehen als das eindeutige Bild dieses Spuroperators. Das analoge gilt für den rechten Randwert $u(1)$. Folglich macht es auch Sinn, den folgenden Unterraum zu betrachten:

$$V \;=\; H_0^1(I) \;:=\; \{u \in H^1(I) : u(0) = u(1) = 0\}\,.$$

Satz 7.7 (Ungleichung von Friedrich) *Es existiert eine Konstante $C_p > 0$, so dass für alle $u \in H_0^1(0,1)$ gilt*

$$\|u\|_{L^2(0,1)} \;\le\; C_F\|u'\|_{L^2(0,1)}\,.$$

Beweis. Wir zeigen die Behauptung zunächst für $C^1[0,1]$-Funktionen u. Wie können aufgrund der Stetigkeit des Spuroperators $u(0) = 0$ annehmen. Es gilt dann die Darstellung

$$u(x) \;=\; \int_0^x u'(y)\,dy\,.$$

Hieraus folgt

$$\|u\|_{L^\infty(0,1)} \;=\; \max_{x\in[0,1]} |u(x)| \;\le\; \int_0^1 |u'(y)|\,dy \;=\; \|u'\|_{L^1(0,1)} \le\; \|u'\|_{L^2(0,1)}\,.$$

Die letzte Ungleichung folgt aus der Tatsache, dass $C^1[0,1] \subset L^2(0,1)$, und der Cauchy-Ungleichung. Integration auf beiden Seiten ergibt:

$$\|u\|_{L^2(0,1)}^2 \;\le\; \int_0^1 \|u\|_{L^\infty(0,1)}^2\,dx = \|u\|_{L^\infty(0,1)}^2 \;\le\; \|u'\|_{L^2(0,1)}^2\,.$$

Damit ist die Behauptung für C^1-Funktionen gezeigt. Die Behauptung für $u \in H_0^1(0,1)$ folgt nun wieder aufgrund der Dichtheit von $C^1[0,1] \cap H_0^1(0,1)$ in $H_0^1(0,1)$ und der Stetigkeit der auftretenden Ausdrücke. Im Fall des Einheitsintervalls $(0,1)$ ist die Friedrich-Konstante also sogar $C_F = 1$. □

Korollar 7.8 *Die Halbnorm $|\cdot|_{H^1(0,1)}$ ist eine Norm auf $H_0^1(0,1)$, die äquivalent ist zu $\|\cdot\|_{H^1(0,1)}$. Ferner ist*

$$(u,v)_{H_0^1(0,1)} \quad := \quad \int_0^1 u'v'\,dx$$

ein Skalarprodukt auf $H_0^1(0,1)$, dass die Norm $|\cdot|_{H^1(0,1)}$ induziert.

Beweis. Aufgrund von der Friedrich'schen Ungleichung (Satz 7.7) gilt

$$\begin{aligned} \|u\|_{H^1(0,1)}^2 &= \|u\|_{L^2(0,1)}^2 + \|u'\|_{L^2(0,1)}^2 \\ &\leq (1 + C_F^2)\|u'\|_{L^2(0,1)}^2 \\ &\leq (1 + C_F^2)|u|_{H^1(0,1)}^2 . \end{aligned}$$

Insgesamt folgt $\|u'\|_{L^2(0,1)} \leq \|u\|_{H^1(0,1)} \leq C\|u'\|_{L^2(0,1)}$ und somit die Äquivalenz der Normen. Das $(\cdot,\cdot)_{H_0^1(0,1)}$ ein Skalarprodukt darstellt ist einfach zu überprüfen. Ferner sieht man unmittelbar

$$(u,u)_{H_0^1(0,1)} = \|u'\|_{L^2(0,1)}^2 .$$

□

Satz 7.9 *Der Raum $H_0^1(I)$ wird zusammen mit dem Skalarprodukt*

$$\langle u,v\rangle_{H_0^1(I)} \quad := \quad (u',v')$$

zu einem Hilbertraum. Der Raum $C_0^1[0,1]$ liegt dicht in $H^1(I)$.

Beweis. Hierbei sind zwei Dinge zu beweise: (a) $\langle\cdot,\cdot\rangle_{H_0^1(I)}$ bildet tatsächlich ein Skalarprodukt und (b) $H_0^1(I)$ ist vollständig. Die Eigenschaft (a) ist eine direkte Folgerung aus der Ungleichung von Friedrichs. Die Vollständigkeit ist eine Folgerung aus der Vollständigkeit von $H^1(I)$ und der Stetigkeit des Spuroperators. □

Dies wird der Raum sein, in dem wir letztendlich die variationelle Formulierung betrachten:

$$u \in V: \quad A(u,\varphi) \;=\; (f,\varphi) \qquad \forall \varphi \in V\,. \tag{7.5}$$

7.4 Existenz und Eindeutigkeit von Lösungen

Der Einfachheit betrachten wir zunächst den Fall $a \equiv 1$, $b = c \equiv 0$. Die Bilinearform lautet dann

$$A(u,\varphi) \;=\; (u',\varphi')\,. \tag{7.6}$$

Hierbei fällt auf, dass diese Bilinearform gerade das V-Skalarprodunkt darstellt:

$$A(u,\varphi) \;=\; \langle u,\varphi\rangle_{H_0^1(I)}\,.$$

Somit lässt sich das Problem auch folgermaßen formulieren:

$$u\in V: \quad \langle u,\varphi\rangle_V \;=\; (f,\varphi) \qquad \forall\varphi\in V\,. \tag{7.7}$$

Lemma 7.10 *Sei H ein reeller Hilbertraum und $f\in H'$. Dann sind die folgenden Probleme äquivalent:*

$$\begin{aligned} &(a) \qquad u\in H: \quad (u,\phi)_H = \langle f,\phi\rangle \qquad \forall\phi\in H\\ &(b) \qquad u\in H: \quad J(u) = \min_{v\in H} J(v)\,, \end{aligned}$$

zum Funktional $J(v) := \frac{1}{2}\|v\|_H^2 - \langle f,v\rangle$.

Beweis. Eine Lösung u des Problems (b) ist charakterisiert durch

$$J(u) \;\le\; J(u+tw)$$

für alle $w\in H$ und alle $t\in[0,1]$. Nun gilt aber

$$J(u+tw) \;=\; J(u) + \frac{t^2}{2}\|w\|_V^2 + t(u,w) - t\langle f,w\rangle\,.$$

Daher ist das Problem (b) äquivalent zu

$$u\in H: \qquad \frac{t}{2}\|w\|_V^2 + (u,w) - \langle f,w\rangle \;\le\; 0 \qquad \forall w\in H\,,\ \forall t\in[0,1]\,,$$

bzw. zu

$$u\in H: \qquad (u,w) \;\le\; \langle f,w\rangle \qquad \forall w\in H\,.$$

Da mit $w\in H$ auch $-w\in H$, ist dies genau dann der Fall, wenn Problem (a) erfüllt ist. □

Satz 7.11 (Riesz'scher Darstellungssatz) *Sei V ein reeller Hilbertraum und $f\in V'$. Dann besitzt das Problem (7.7) eine eindeutige Lösung $u\in V$ und es gilt $\|u\|_V = \|f\|_{V'}$.*

Beweis. (a) Existenz: Das Funktional J aus Lemma 7.10 ist nach unten beschränkt:

$$J(v) \;\geq\; \frac{1}{2}\|v\|_V^2 - \|f\|_{V'}\|v\|_V \;= \frac{1}{2}(\|v\|_V - \|f\|_{V'})^2 - \frac{1}{2}\|f\|_{V'}^2 \geq\; -\frac{1}{2}\|f\|_{V'}^2 .$$

Daher existiert eine Minimalfolge $(u_k)_{k\in\mathbb{N}}$ in V mit

$$\lim_{k\to\infty} J(u_k) \;=\; \inf_{v\in V} J(v) \;>\; -\infty .$$

Ferner bildet diese Folge eine Cauchyfolge, denn die Parallelogrammgleichung liefert uns

$$\begin{aligned}\|u_m - u_n\|_V^2 &= 2\|u_m\|^2 + 2\|u_n\|^2 - \|u_m + u_n\|^2 \\ &= 4J(u_m) + 4J(u_n) - 8J((u_m + u_n)/2) \\ &\leq 4J(u_m) + 4J(u_n) - 8\inf_{v\in V} J(v) .\end{aligned}$$

Der Grenzübergang $m, n \to \infty$ ergibt

$$\lim_{m,n\to\infty} \|u_m - u_n\|_V^2 \;\leq\; 0$$

und somit die Cauchy-Eigenschaft. Da V als Hilbertraum vollständig ist, konvergiert diese Folge gegen ein $u = \lim_{n\to\infty} u_n \in V$. Aufgrund der Stetigkeit von J folgt

$$J(u) \;=\; \lim_{n\to\infty} J(u_n) \;=\; \inf_{v\in V} J(v) .$$

Nach Lemma 7.10 ist u eine Lösung des betrachteten Variationsproblems (7.7).
(b) Eindeutigkeit: Sei nun $\tilde{u}$ eine weitere Lösung von (7.7). Dann erfüllt die Differenz $u - \tilde{u}$ die Gleichung

$$(u - \tilde{u}, \phi)_V \;=\; 0 \qquad \forall \phi \in V .$$

Setzen wir insbesondere $\phi = u - \tilde{u}$ folgt $\|u - \tilde{u}\|_H^2 = 0$, also $u = \tilde{u}$.
(c) $\|u\|_V = \|f\|_{V'}$:

$$\|f\|_{V'} \;=\; \sup_{0\neq v\in V} \frac{\langle f, v\rangle}{\|v\|_V} \;=\; \sup_{0\neq v\in V} \frac{(u, v)_V}{\|v\|_V} \;=\; \|u\|_V .$$

Hierbei folgt die letzte Gleichung aus der Cauchy-Ungleichung:

$$(u, v)_V \;\leq\; \|v\|_V \, \|u\|_V$$

sowie mittels der Wahl $v := u$:

$$\sup_{0\neq v\in V} (u, v)_V \;\geq\; \|u\|_V^2 .$$

□

Satz 7.12 *Die variationelle Formulierung (7.5) für die Bilinearform (7.6) besitzt für beliebige rechte Seiten $f \in L^2(I)$ stets eine eindeutige Lösung $u \in H_0^1(I)$.*

Beweis. Die Behauptung folgt unmittelbar aus dem Riesz'scher Darstellungssatz, da man f als Funktional auf V auffassen kann, also $f \in V'$. □

7.5 Galerkin Methode

Sei nun V ein unendlich-dimensionaler Hilbertraum, $f \in V'$ ein lineares Funktional und $A : V \times V \to \mathbb{R}$ eine Bilinearform. Die Grundidee einer *Galerkin-Methode* zur Lösung eines variationellen Problems der Form

$$u \in V : \quad A(u,v) \;=\; (f,v) \qquad \forall v \in V$$

besteht darin, dass man den Hilbertraum V ersetzt durch einen endlich-dimensionalen Raum $V_h \subset V$. Hierbei bezeichnet h einen Parameter, der die Feinheit der Diskretisierung beschreibt. Die approximative diskrete Lösung bezeichnen wir mit u_h:

$$u_h \in V_h : \quad A(u_h,v_h) \;=\; (f,v_h) \qquad \forall v_h \in V_h \,. \tag{7.8}$$

Charakteristisch ist für eine Galerkin-Methode die sogenannte Galerkin-Orthogonalität

$$A(u-u_h,v_h) \;=\; 0 \qquad \forall v_h \in V_h \,. \tag{7.9}$$

Man beachte, dass hier als Testfunktionen v_h nur Funktionen aus V_h eingesetzt werden dürfen. Unter *Finiten Elementen* versteht man nun eine besondere Klasse von endlich-dimensionalen Räumen V_h. Zum einen soll eine Basis von V_h existieren, die einen kleinen Träger besitzt; zum anderen sollen die Funktionen lokal eine einfache Struktur haben. Dies legt es nahe, stückweise Polynome zu verwenden. Wir werden uns nun zunächst Finite Elemente ansehen, die aus stückweise linearen Funktionen bestehen. Dies entspricht gewissermaßen der einfachsten Klasse von Finiten Elementen.

Definition 7.13 *Unter einer* Finite Elemente Methode *versteht man ein Variationsproblem (7.8) mit einem endlich-dimensionalen Raum V_h, der aus stückweisen Polynomen besteht. Man spricht von* konformen *Finiten Elementen, wenn für den Finite Elemente Raum V_h gilt, $V_h \subset V$. Anderenfalls spricht man von einer* nicht-konformen *Finite Elemente Methode.*

Existenz und Eindeutigkeit liefert das folgende Korollar:

Korollar 7.14 *Das Variationsproblem (7.8) mit der Bilinearform A aus (7.6) und mit konformen Finiten Elementen besitzt stets eine eindeutige Lösung.*

Beweis. Da $V_h \subset V$ als endlich-dimensionaler Teilraum immer vollständig ist, ist auch V_h ein Hilbertraum. Die Existenz und Eindeutigkeit folgt nun wieder aus dem Darstellungssatz von Riesz (Satz 7.11). ∎

7.6 Lineare Finite Elemente in 1D

Wir zerlegen das Intervall $[0,1]$ in n Teilintervalle $I_k = [x_{k-1}, x_k]$ mit $h_k := x_k - x_{k-1}$,

$$0 = x_0 < x_1 < x_2 < \ldots < x_n = 1\,.$$

Wir bezeichnen den Raum der linearen Funktionen auf einem Intervall I mit $P_1(I)$. Unser Finite Elemente Raum wird nun aus stetigen und stückweisen linearen Funktionen bestehen:

$$P_{h,1} \;:=\; \{v \in C[0,1] \,:\, v|_{I_k} \in P_1(I_k)\ \forall k = 1,\ldots,n\}\,.$$

Der diskrete Raum kann also beispielsweise als $V_h = P_{h,1}$ gewählt werden. Diese Finiten Elemente werden auch *Courant-Elemente* genannt.

Lemma 7.15 *Der Raum $P_{h,1}$ ist ein H^1-konformer Finite Elemente Raum.*

Beweis. Für den Nachweis der Konformität ist $P_{h,1} \subset H^1(I)$ zu zeigen. Wegen $C[0,1] \subset L^2(I)$ folgt $P_{h,1} \subset L^2(I)$. Sei $u_h \in P_{h,1}$. Dann ist u_h' stückweise konstant und damit auf I L^2-integrierbar. Also $u_h \in H^1(I)$. □

Ein $v \in V_h$ ist eindeutig definiert durch seine *Knotenwerte* $v(x_k)$, $k = 0,\ldots,n$. Umgekehrt kann man zu beliebigen $n+1$ Knotenwerte genau eine Finite Elemente Funktion $v \in V_h$ finden. Es gilt $\dim V_h = n+1$.

Entsprechend ist der zugehörige Finite Elemente Raum V_{hg} mit stückweisen linearen Polynomen zum affinen Raum $V_g = g + H_0^1(0,1)$:

$$\begin{aligned} V_{hg} \;&:=\; P_{h,1} \cap V_g \\ &=\; \{v \in C[0,1] \,:\, v(x_0) = g_0,\ v(x_n) = g_1\,,\ v|_{I_k} \in P_1(I_k)\ \forall k = 1,\ldots,n\}\,. \end{aligned}$$

Zum Raum V_h existieren verschiedene *Basen*. Unter der *Lagrangebasis* versteht man in diesem Fall die "Hütchenfunktionen", die an den Knoten x_j nur die Werte 1 oder 0 annehmen. Die "inneren" Basisfunktionen lauten

$$\phi_k(x_j) \;=\; X_{I_k}(x)\frac{x - x_{k-1}}{x_k - x_{k-1}} + X_{I_{k+1}}(x)\frac{x_{k+1} - x}{x_{k+1} - x_k} \quad k = 1,\ldots,n-1\,,$$

während die an den Rändern durch

$$\phi_0(x) \;=\; X_{I_1}(x)\frac{x_1 - x}{x_1 - x_k} \quad \text{und} \quad \phi_n(x) \;=\; X_{I_n}(x)\frac{x - x_{n-1}}{x_n - x_{n-1}}$$

gegeben sind. Hierbei bezeichnet X_J die charakteristische Funktion zum Intervall J. Für den Träger dieser Funktionen gilt

$$\operatorname{supp}\phi_k \;=\; I_k \cup I_{k+1} \qquad k = 1,\ldots,n-1\,,$$

sowie $\operatorname{supp}\phi_0 = I_1$ und $\operatorname{supp}\phi_n = I_n$. Die Lösung u_h können wir nun durch diese Basis darstellen

$$u_h(x) \;=\; \sum_{k=0}^{n} u_k \phi_k(x)\,.$$

Die Koeffizienten u_k sind gerade die Knotenwerte $u_k = u_h(x_k)$. Das diskrete Variationsproblem (7.8) mit homogenen Dirichletwerten läßt sich nun auch formulieren in der Form

$$\sum_{j=1}^{n-1} A(\phi_j,\phi_i)u_j \;=\; (f,\phi_i) \qquad \forall i = 1,\ldots,n-1\,,$$

da die Knotenwerte am Anfang und Ende des Intervalls zu Null gesetzt werden müssen, also $u_0 = u_n = 0$. Dies entspricht einem linearen Gleichungssystem

$$\mathcal{A}U \;=\; b\,,$$

mit dem Knotenvektor $U = (u_1,\ldots,u_{n-1})^T$, rechter Seite $b = (b_1,\ldots,b_{n-1})^T$, $b_k = (f,\phi_k)$ und der *Steifigkeitsmatrix* $\mathcal{A} = (a_{ij})_{i,j=1,\ldots,n-1}$.

7.6.1 Steifigkeitsmatrix

Die Koeffizienten der Steifigkeitsmatrix sind

$$a_{ij} \;=\; A(\phi_j,\phi_i)\,.$$

Da die Träger der Basisfunktionen ϕ_k lokal sind, ist die Steifigkeitsmatrix dünn besetzt. Insbesondere gilt für $|i-j|>1$ im Fall der Bilinearform (7.6):

$$a_{ij} \;=\; \int_0^1 \phi_j'(x)\phi_i'(x)\,dx \;=\; 0\,.$$

Da wir hier stückweise lineare Finite Elemente betrachten, sind die Ableitungen ganz besonders einfach, nämlich stückweise konstant:

$$\phi_i'|_{I_i} \;=\; h_i^{-1}\,, \quad \phi_i'|_{I_{i+1}} \;=\; -h_{i+1}^{-1}\,.$$

Für die Diagonaleinträge $1 \le i \le n-1$ ergibt sich daher

$$\begin{aligned} a_{ii} \;&=\; \int_0^1 \phi_i'(x)\phi_i'(x)\,dx \;=\; \int_{x_{i-1}}^{x_i} \phi_i'(x)\phi_i'(x)\,dx + \int_{x_i}^{x_{i+1}} \phi_i'(x)\phi_i'(x)\,dx \\ &=\; h_i h_i^{-2} + h_{i+1}(-h_{i+1})^{-2} \;=\; h_i^{-1} + h_{i+1}^{-1}\,. \end{aligned}$$

Die Nebendiagonaleinträge ergeben sich zu

$$\begin{aligned} a_{i,i+1} \;&=\; \int_0^1 \phi_i'(x)\phi_{i+1}'(x)\,dx \;=\; \int_{x_i}^{x_{i+1}} \phi_i'(x)\phi_{i+1}'(x)\,dx \\ &=\; h_{i+1}(-h_{i+1}^{-1}h_{i+1}^{-1}) \;=\; -h_{i+1}^{-1}\,. \end{aligned}$$

Aus Symmetriegründen gilt in diesem speziellen Fall $a_{i+1,i} = a_{i,i+1} = -h_{i+1}$. Insgesamt erhalten wir daher folgende $(n+1) \times (n+1)$ Matrix

$$\mathcal{A} = \begin{pmatrix} h_1^{-1} & -h_1^{-1} & 0 & \cdots & \cdots \\ -h_1^{-1} & h_1^{-1} + h_2^{-1} & -h_2^{-1} & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \vdots \\ 0 & \dots & & -h_n^{-1} & h_n^{-1} \end{pmatrix}.$$

Im Fall einer äquidistanten Unterteilung $h = h_1 = \ldots = h_n$ erhalten wir

$$\mathcal{A} = \frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & & 2 & -1 \\ 0 & \cdots & \cdots & -1 & 1 \end{pmatrix}.$$

Man beachte, dass in diesen Matrizen noch keine (Dirichlet-) Randbedingungen enthalten sind. Im Vergleich zur Finiten-Differenzen Matrix erhalten wir also bis auf die Skalierung mit h dieselbe Matrix,

$$\mathcal{A}_{FEM} = h\mathcal{A}_{FDM}\,.$$

7.6.2 Lastvektor

Die rechte Seite $b = (b_1, \ldots, b_{n-1})$ besteht aus den Komponenten

$$\begin{aligned} b_k &= \int_0^1 f(x)\phi_k(x)\,dx \\ &= h_k^{-1} \int_{x_{k-1}}^{x_k} f(x)(x - x_{k-1})\,dx + h_{k+1}^{-1} \int_{x_k}^{x_{k+1}} f(x)(x_{k+1} - x)\,dx\,. \end{aligned}$$

Diese Integrale lassen sich für allgemeines f nicht exakt numerisch integrieren. Man muß sich daher Integrationsformel bedienen. Hierbei ist eine ausreichende Genauigkeit zu bedenken. Wir wenden nun einmal exemplarisch die zusammengesetzte Trapezregel an, die exakt ist, wenn f eine lineare Funktion ist. Da ϕ_k an den Integrationspunkten x_{k-1} und x_{k+1} bereits verschwindet und im Punkt x_k den Wert 1 annimmt, reduziert sich dann die Approximation für $k = 1, \ldots, n-1$ auf

$$\begin{aligned} b_k &\approx h_k^{-1}\frac{1}{2}h_k(0 + f(x_k)(x_k - x_{k-1}) + h_{k+1}^{-1}\frac{1}{2}h_{k+1}(f(x_k)(x_{k+1} - x_k + 0) \\ &= \frac{1}{2}f(x_k)\,(h_k + h_{k+1})\,. \end{aligned}$$

Im Fall einer äquidistanten Gitterweite h also insbesondere

$$b_k \approx hf(x_k) = hb_k^{FDM}.$$

Auch die rechte Seite entspricht also bis auf Skalierung mit h der Finite-Differenzen rechte Seite b^{FDM}. Insgesamt sind also die FDM und die FEM Methode mit linearen Finiten Elementen auf äquidistanten Gittern äquivalent, sofern wir die rechte Seite mit der Trapezregel integrieren. Wir sind aber hier in der Lage, genauere Integrationsformel zu nehmen, etwa dann, wenn f stark fluktuiert.

7.6.3 Massematrix

Neben der Steifigkeitsmatrix $\mathcal{A}$ tritt auch häufiger die sogenannte *Massematrix* auf. Diese beschreibt die Steifigkeitsmatrix für 0.ten Ordnungsterme, wie sie beispielsweise in der Gleichung $-u'' + u = f$ auftauchen:

$$M = (m_{ij})_{i,j=1,\dots,n}, \quad m_{ij} = (\phi_j, \phi_i).$$

Die Massematrix ist stets symmetrisch. Wie man leicht nachprüft, lautet sie im Fall von P_1-Finiten Elementen:

$$M_h = \frac{1}{6}\begin{pmatrix} 2(h_1+h_2) & h_2 & & & & \\ h_2 & 2(h_2+h_3) & h_3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_n & \\ & & & h_{n-1} & 2h_n & \end{pmatrix}.$$

Im Fall eines äquidistanten Gitters reduziert sich dies zur Tridiagonalmatrix

$$M_h = \frac{h}{6}\,\mathrm{tridiag}(1,4,1).$$

7.7 P_r-Elemente

Wir hatten bislang nur lineare Finite Elemente betrachtet. Es ist naheliegend auch stückweise Polynome von höherem Grad zu verwenden. Dies führt auf die P_r-Elemente, wobei $r \geq 1$ den lokalen Polynomgrad angibt.

$$P_{h,r} := \{v \in C[0,1] : v|_{I_k} \in P_r(I_k)\ \forall k = 1,\dots,n\}.$$

Wir wollen für quadratische Elemente, $r = 2$, die Lagrangebasis angeben. Normalisiert auf das Intervall $[0,1]$ lauten die drei Basisfunktionen (vgl. auch Abb. 7.1):

$$\begin{aligned} \phi_1(x) &= 2\left(x - \frac{1}{2}\right)(x-1), \\ \phi_2(x) &= -4x(x-1), \\ \phi_3(x) &= 2x\left(x - \frac{1}{2}\right). \end{aligned}$$

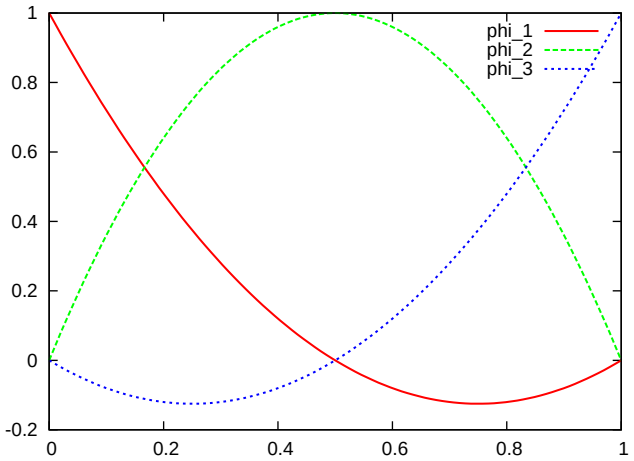

Abbildung 7.1: Lagrangebasis der quadratischen Finiten Elemente P_2 normiert auf das Einheitsintervall.

7.8 A priori Abschätzung

Zunächst wollen wir den Diskretisierungsfehler zwischen der kontinuierlichen Lösung $u \in V$ und der diskreten Lösung $u_h \in V_h$ in Verhältnis setzen zum sogenannten *Approximationsfehler*

$$\inf_{v_h \in V_h} \|u - v_h\|_V \,.$$

Der Approximationsfehler ist also das beste was wir erhoffen könnnen. Wir können in diesem Abschnitt $h \leq 1$ voraussetzen, da man i.d.R. an dem Fehlerverhalten für kleine Gitterweiten interessiert ist. Das folgende Lemma besagt, dass wir durch eine Galerkin-Methode diesen Approximationsfehler bis auf eine Konstante erreichen, sofern die Bilinearform V-beschränkt und V-elliptisch ist.

Lemma 7.16 (Cea's Lemma) *Die Bilinearform $A : V \times V \to \mathbb{R}$ auf einem Hilbertraum V erfülle folgende Eigenschaften:*

- *H-*beschränkt*, d.h. es existiert eine Konstante $\alpha_1 > 0$, so dass*

$$|A(u,v)| \quad \leq \quad \alpha_1 \|u\|_V \|v\|_V \qquad \forall u, v \in V \,,$$

- *H-*elliptisch*, d.h. es existiert eine Konstante $\alpha_2 > 0$, so dass*

$$A(u,u) \quad \geq \quad \alpha_2 \|u\|_V^2 \qquad \forall u \in V \,.$$

Ferner sei $V_h \subset V$ ein Teilraum. Dann gilt für den Diskretisierungsfehler zwischen der kontinuierlichen Lösung $u \in V$ und der "diskreten" Lösung $u_h \in V_h$:

$$\|u - u_h\|_V \quad \leq \quad \frac{\alpha_1}{\alpha_2} \inf_{v_h \in V_h} \|u - v_h\|_V \,.$$

In der Regel ist V_h ein endlich dimensionaler Teilraum von V. Daher nennen wir die Lösung u_h diskret und sprechen von Diskretisierungsfehler.

Beweis. Aufgrund der Elliptizität gilt

$$\alpha_2 \|u - u_h\|_V^2 \quad \leq \quad A(u - u_h, u - u_h) \,.$$

Nun verwenden wir die Galerkin-Orthogonalität (7.9)

$$A(u - u_h, w_h) \quad = \quad 0 \qquad \forall w_h \in V_h \,.$$

Zu gegebenem aber beliebigem $v_h \in V_h$ wählen wir $w_h := u_h - v_h \in V_h$ und erhalten aufgrund der Linearität von $A(\cdot, \cdot)$ im zweiten Argument

$$\begin{aligned} \alpha_2 \|u - u_h\|_V^2 \quad &\leq \quad A(u - u_h, u - u_h) + A(u - u_h, u_h - v_h) \\ &= \quad A(u - u_h, u - v_h) \;\leq\; \alpha_1 \|u - u_h\|_V \|u - v_h\|_V \,. \end{aligned}$$

Teilen wir nun beide Seiten durch $\alpha_2 \|u - u_h\|_V$, so erhalten wir die Behauptung. □

Der Approximationsfehler ist i.d.R. nur sehr schwer zu bestimmen. Eine obere Grenze liefert aber der *Interpolationsfehler* $\|u - P_h u\|_V$, mit einem Interpolationsoperator

$$P_h : V \to V_h \,.$$

Es gilt selbstverständlich

$$\inf_{v_h \in V_h} \|u - v_h\|_V \quad \leq \quad \|u - P_h u\|_V \,.$$

Nun kommen verschiedene Interpolationen in Betracht, die wir später auch diskutieren werden. Mithilfe der sogenannte Knoteninterpolierenden können wir aber folgenden Satz zeigen.

Satz 7.17 *Unter den gleichen Voraussetzungen an die Bilinearform $A : V \times V \to \mathbb{R}$ wie im Cea's Lemma (7.16) mit $V = H^1(0,1)$ und der zusätzlichen Bedingung an die kontinuierliche Lösung $u \in H^2(0,1)$ gilt für den Diskretisierungsfehler mit linearen Finiten Elementen*

$$\|u - u_h\|_{H^1(0,1)} \quad \leq \quad Ch \|u''\|_{L^2(0,1)} \,. \tag{7.10}$$

Die Konstante hängt von den Beschränktheits- und Elliptizitätskonstanten in der Form $C \sim \alpha_1/\alpha_2$ ab.

Beweis. Nach Cea's Lemma genügt es zu zeigen, dass ein v_h existiert mit

$$\|u - v_h\|_{H^1(0,1)} \quad \leq \quad C_I h \|u''\|_{L^2(0,1)} \,.$$

Die Existenz eines solchen v_h folgt aus dem Interpolationsfehler der sogenannte Knoteninterpolierende $v_h = I_h u$, die für $u \in H^2(0,1)$ folgendes erfüllt:

$$\|u - I_h u\|_{L^2(0,1)} + h|u - I_h u|_{H^1(0,1)} \quad \leq \quad C_I h^2 \|u''\|_{L^2(0,1)} \,.$$

Den Nachweis dieser Eigenschaft überlassen wir der Vorlesung "Finite Elemente". □

Literaturverzeichnis

[1] S. Börm. Numerische Verfahren für Differentialgleichungen. Technical report, Universität Kiel, `http://www.informatik.uni-kiel.de/~sb/data/NumDgl.pdf`, 2010.

[2] K. Burrage and J. Butcher. Stability criteria for implicit Runge-Kutta methods. *SIAM J. Numer. Anal.*, 16:46–57, 1979.

[3] G. Dahlquist. *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations.* Tekn. Högskol. Handl. 130, 1959.

[4] E. Hairer and G. Wanner. *Solving ordinary differential equations II: Stiff and differential-algebraic problems.* Berlin etc.: Springer-Verlag, 1991.

[5] M. Hermann. *Numerik gewöhnlicher Differentialgleichungen.* Oldenbourg, 446 p., 2004.

[6] G. Lube. Numerische Mathematik II. Technical report, Universität Göttingen, `http://www.num.math.uni-goettingen.de/lube/NM2-2010.pdf`, 2010.

[7] R. Rannacher. Numerische Methoden für Gewöhnliche Differentialgleichungen (Numerische Mathematik 1). Technical report, Universität Heidelberg, `http://ganymed.iwr.uni-heidelberg.de/~lehre/notes/num1/Numerik1.pdf`, 2011.