# Numerik Partieller Differentialgleichungen

**Martin Neumüller, Institute of Computational Mathematics, JKU Linz**

martin.neumueller@jku.at www.numa.uni-linz.ac.at

# Inhaltsverzeichnis

# 2 Parabolische Differentialgleichungen

## 2.1 Variationsformulierung von parabolischen Anfangs-Randwertproblemen

Ausgehend von der klassischen Formulierung wollen wir in diesem Abschnitt eine schwache Formulierung herleiten.

### 2.1.1 Klassische Formulierung

Wir betrachten das beschränkte Gebiet $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ mit hinreichend glatten Rand $\Gamma = \partial\Omega = \Gamma_D \cup \Gamma_N$ und weiters das Zeitintervall $(0, T)$ mit $0 < T < \infty$. Wir definieren dann den Raum-Zeit-Zylinder $Q_T := \Omega \times (0, T) \in \mathbb{R}^{d+1}$. Die Problemstellung lautet dann: Gesucht ist $u : \overline{Q}_T \to \mathbb{R}$, sodass

$$\frac{\partial u}{\partial t}(x,t) + (\mathcal{L}u)(x,t) = f(x,t) \qquad \text{für } (x,t) \in Q_T, \tag{2.1}$$

$$u(x,t) = g_D(x,t) \qquad \text{für } (x,t) \in \Gamma_D \times (0,T), \tag{2.2}$$

$$(A(x)\nabla u(x,t)) \cdot \underline{n}(x) = g_N(x,t) \qquad \text{für } (x,t) \in \Gamma_N \times (0,T), \tag{2.3}$$

$$u(x,0) = u_0(x) \qquad \text{für } x \in \overline{\Omega} \tag{2.4}$$

mit dem Differentialoperator

$$(\mathcal{L}u)(x,t) := -\mathrm{div}(A(x)\nabla u(x,t)) + \underline{b}(x) \cdot \nabla u(x,t) + c(x)u(x,t),$$

den hinreichend regulären Koeffizienten

$$A(x) \in \mathbb{R}^{d\times d}, \quad \underline{b}(x) \in \mathbb{R}^d, \quad c(x) \in \mathbb{R} \qquad \text{für } x \in \overline{\Omega}$$

und den gegebenen skalaren Daten $f, g_D, g_N, u_0$.

**Beispiel 2.1** (Das $d$-dimensionale parabolische Modellproblem)**.** *Für $A = I$, $\underline{b} = \underline{0}$ und $c = 0$ erhalten wir das $d$-dimensionale Modellproblem*

$$\begin{aligned}
\frac{\partial u}{\partial t}(x,t) - \Delta u(x,t) &= f(x,t) && \textit{für } (x,t) \in Q_T,\\
u(x,t) &= g_D(x,t) && \textit{für } (x,t) \in \Gamma_D \times (0,T),\\
\frac{\partial u}{\partial n}(x,t) &= g_N(x,t) && \textit{für } (x,t) \in \Gamma_N \times (0,T),\\
u(x,0) &= u_0(x) && \textit{für } x \in \overline{\Omega}.
\end{aligned}$$

Für die Klassische Formulierung (2.1)–(2.4) müssen die Daten $g_D$ und $u_0$ die Bedingung

$$\lim_{t\to 0} g_D(x,t) = u_0(x) \qquad \text{für alle } x \in \Gamma_D$$

erfüllen, damit für die klassische Lösung $u$ die Stetigkeit gewährleistet werden kann. Weitere Diskussionen über die Funktionenräume im klassischen Sinn werden hier nicht behandelt.

**Bemerkung 2.2.** *Für den Fall $g_D \neq 0$ kann die Problemstellung (2.1)–(2.4) homogenisiert werden, falls es eine Funktion $g : \overline{Q}_T \to \mathbb{R}$ gibt für die gilt*

$$g(x,t) = g_D(x,t) \qquad \textit{für alle } (x,t) \in \Gamma_D \times (0,T).$$

Auf Grund der Bemerkung 2.2 betrachten wir im Weiteren den Fall $g_D = 0$.

### 2.1.2 Semi-Variationsformulierung

Die Herleitung dieser Variationsformulierung beruht auf den folgenden Schritten:

- Multiplikation der Differentialgleichung (2.1) mit einer Testfunktion $v : \overline{\Omega} \to \mathbb{R}$.
- Integration über das Rechengebiet $\Omega$.
- Partielle Integrations des Hauptteils.
- Einarbeitung der natürlichen Randbedingung (2.3).
- Die wesentliche Randbedingung (2.2) und die Anfangsbedingung (2.4) werden explizit gefordert.

Wir erhalten dann die Variationsformulierung: Gesucht ist $u : \overline{Q}_T \to \mathbb{R}$ mit $u(x,t) = 0$ für $(x,t) \in \Gamma_D \times (0,T)$, sodass

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)v(x) + a(u(\cdot,t),v) = \langle F(t), v\rangle,$$

für alle $v : \overline{Q}_T \to \mathbb{R}$ mit $v(x) = 0$ für $x \in \Gamma_D$ und weiters

$$u(x,0) = u_0(x) \qquad \text{für alle } x \in \overline{\Omega}$$

erfüllt ist. Dabei ist die Bilinearform für $t \in (0,T)$ gegeben durch

$$a(u(\cdot,t),v) := \int_\Omega \big[A(x)\nabla u(x,t)\cdot\nabla v(x) + \underline{b}(x)\cdot\nabla u(x,t)v(x) + c(x)u(x,t)v(x)\big]\mathrm{d}x$$

und die Linearform ist definiert als

$$\langle F(t), v\rangle := \int_\Omega f(x,t)v(x)\mathrm{d}x + \int_{\Gamma_N} g_N(x,t)v(x)\mathrm{d}s_x.$$

Zur einfacheren Darstellung schreiben wir jetzt die Funktion $u = u(x,t)$ also Funktion von $t$, deren Werte Funktionen von $x$ sind, also

$$u(x,t) = u(t)(x).$$

Unter Verwendung der schwachen Ableitungen aus Kapitel **??** und des daraus resultierenden Funktionenraums

$$V = \left\{ v \in H^1(0,1) : v = 0 \text{ auf } \Gamma_D \right\}$$

ergibt sich für klassiche Voraussetzungen bezüglich der Zeit die Semi-Variationsformulierung: Gesucht ist $u \in \mathcal{C}^1([0,T], V)$, sodass

$$\begin{aligned} (u', v)_{L^2(\Omega)} + a(u(t), v) &= \langle F(t), v \rangle && \text{für alle } v \in V \text{ und alle } t \in (0,T), \\ u(0) &= u_0 && \text{in } V \end{aligned} \tag{2.5}$$

erfüllt ist. Dabei stellen wir an die Daten die Voraussetzungen

$$f \in \mathcal{C}([0,T], L^2(\Omega)), \qquad g_N \in \mathcal{C}([0,T], L^2(\Gamma_N)), \qquad u_0 \in V,$$

beziehungsweise an die Koeffizienten

$$A \in [L^\infty(\Omega)]^{d \times d}, \quad \underline{b} \in [L^\infty(\Omega)]^d, \quad c \in L^\infty(\Omega).$$

Unter diesen Voraussetzungen ist die Bilinearform $a : V \times V \to \mathbb{R}$ beschränkt und weiters ist $F(t) \in V^*$ für alle $t \in [0,T]$ und, siehe Kapitel **??**.

Ziel ist es nun die klassischen Voraussetzungen bezüglich der Zeit auch abzuschwächen.

### 2.1.3 Abstrakte Variationsformulierung unter Verwendung von Bochner Analysis

In diesem Abschnitt wollen wir, ausgehend von einer abstrakten semi-variationellen Formulierung, unser endgültiges Variationsproblem herleiten. Dazu seien $V$ und $H$ zwei Hilbert Räume mit $V \subseteq H$. Weiters sei $a : V \times V \to \mathbb{R}$ eine auf $V$ beschränkte und elliptische Bilinearform. Dann betrachten wir die semi-variationelle Formulierung: Gesucht ist $u \in \mathcal{C}^1([0,T], V)$, sodass

$$\begin{aligned} (u'(t), v)_H + a(u(t), v) &= \langle F(t), v \rangle && \text{für alle } v \in V \text{ und alle } t \in (0,T), \\ (u(0), w) &= (u_0, w) && \text{für alle } w \in H \end{aligned}$$

erfüllt ist. Dabei sei vorerst $F \in \mathcal{C}([0,T], V^*)$.

**Definition 2.3** (Bochner-meßbar)**.** *Eine Funktion $w : (0,T) \to V$ heißt* Bochner-meßbar*, falls die Funktionen*

$$t \mapsto \langle G, w(t) \rangle$$

*für alle $G \in V^*$ auf $(0,T)$ Lebesque-meßbar sind.*

**Lemma 2.4.** *Sei $w : (0,T) \to V$ Bochner-meßbar, dann ist die Funktion*

$$t \mapsto \|w(t)\|_V$$

*Lebesgue-meßbar.*

**Beweis.** Sei $w : (0,T) \to V$ Bochner-meßbar, dann ist die Funktion

$$t \mapsto \frac{\langle G, w(t)\rangle}{\|G\|_{V^*}}$$

für alle $0 \neq G \in V^*$ Lebesgue-meßbar. Das Supremum von einer Folge von Lebesgue-meßbarer Funktionen ist wiederum Lebesgue-meßbar und somit folgt, dass

$$\|w(t)\|_V = \sup_{0\neq G\in V^*} \frac{\langle G, w(t)\rangle}{\|G\|_{V^*}}.$$

Lebesgue-meßbar ist. ∎

**Definition 2.5.** *Sei $V$ ein Banachraum, dann definieren wir die Menge der Funktionen*

$$L^2((0,T),V) := \left\{ v : (0,T) \to V : v \text{ ist Bochner-meßbar und } \|v\|_{L^2((0,T),V)} < \infty \right\},$$

*mit*

$$\|v\|_{L^2((0,T),V)} := \left[ \int_0^T \|v(t)\|_V^2 \mathrm{d}t \right]^{\frac{1}{2}}.$$

**Bemerkung 2.6.** *Man kann zeigen, dass die Menge $L^2((0,T),V)$ ausgestattet mit der Norm $\|\cdot\|_{L^2((0,T),V)}$ vollständig ist, also ein Banachraum ist.*

Für $u \in \mathcal{C}^1([0,T],V)$ gilt

$$(u'(t), v)_H = \frac{d}{dt}(u(t), v)_H \qquad \text{für alle } v \in V.$$

Dies motiviert nun einen schwachen Ableitungsbegriff für die reelle Funktion

$$t \mapsto (u(t), v)_H$$

einzuführen.

**Definition 2.7.** *Eine Funktion $w' \in L^2((0,T),V^*)$ heißt* verallgemeinerte Ableitung *einer Funktion $w \in L^2((0,T),V)$ bezüglich $H$, falls für alle $v \in V$ die Funktion $t \mapsto \langle w'(t), v\rangle$ die schwache Ableitung der Funktion $t \mapsto (w(t),v)_H$ auf $(0,T)$ ist, das heißt es gilt*

$$\int_0^T \varphi(t)\langle w'(t), v\rangle \mathrm{d}t = -\int_0^T \varphi'(t)(w(t),v)_H \mathrm{d}t$$

*für alle $v \in V$ und alle $\varphi \in \mathcal{C}_0^\infty(0,T)$.*

**Bemerkung 2.8.** *Falls für eine Funktion $w \in L^2((0,T),V)$ die schwache Ableitung $w' \in L^2((0,T),V^*)$ existiert, schreiben wir auch*

$$\frac{d}{dt}(w(t),v)_H := \langle w'(t), v\rangle \qquad \textit{für alle } v \in V.$$

**Definition 2.9.** *Mit $H^1((0,T),V;H)$ bezeichnen wir den Sobolev-Raum, mit Funktionen aus $L^2((0,T),V)$ für die, die schwache Ableitung in $L^2((0,T),V^*)$ bezüglich $H$ existiert, also*

$$H^1((0,T),V;H) := \left\{ v \in L^2((0,T),V) : \text{es existiert } v' \in L^2((0,T),V^*) \text{ bezüglich } H \right\}.$$

*Weiters wird durch*

$$\|v\|_{H^1((0,T),V;H)} := \left[ \|v\|^2_{L^2((0,T),V)} + \|v'\|^2_{L^2((0,T),V^*)} \right]^{\frac{1}{2}}$$

*eine Norm auf $H^1((0,T),V;H)$ definiert.*

**Theorem 2.10.** *Die Einbettung $H^1((0,T),V;H) \subset \mathcal{C}([0,T],H)$ ist stetig. Das heißt, es existiert eine Konstante $c_{tr} > 0$, sodass*

$$\max_{t\in[0,T]} \|v(t)\|_H \leq c_{tr} \|v\|_{H^1((0,T),V;H)} \qquad \text{für alle } v \in H^1((0,T),V;H).$$

**Beweis.** Beweisskizze siehe Übungen oder siehe [Zeidler]. ∎

Theorem 2.10 garantiert, dass die Punktauswertung $u(0)$ wohldefiniert ist in $H$. Mit den oben eingeführten Räumen können wir nun das endgültige Variationsproblem aufstellen: Gesucht ist $u \in H^1((0,T),V;H)$, sodass

$$\begin{aligned} \frac{d}{dt}(u(t),v)_H + a(u(t),v) &= \langle F(t), v\rangle && \text{für alle } v \in V \text{ und fast alle } t \in (0,T), \\ u(0) &= u_0 && \text{in } H \end{aligned} \tag{2.6}$$

erfüllt ist. Dabei ist das Variationsproblem (2.6) wohldefiniert für $F \in L^2((0,T),V^*)$ und $u_0 \in H$.

**Bemerkung 2.11.** *Verwendet man den von der Bilinearform $a : V \times V \to \mathbb{R}$ induzierten Operator $A : V \to V^*$, so lässt sich das Variationsproblem (2.6) schreiben als: Gesucht ist $u \in H^1((0,T),V;H)$, sodass*

$$\begin{aligned} u'(t) + Au(t) &= F(t) && \text{in } V^* \text{ und fast alle } t \in (0,T), \\ u(0) &= u_0 && \text{in } H \end{aligned} \tag{2.7}$$

*erfüllt ist. Man spricht dabei auch von einer gewöhnlichen Differentialgleichung im Banachraum.*

**Theorem 2.12.** *Es seien $V$ und $H$ zwei seperable Hilberträume mit $V \subseteq H$. Weiters sei $V$ dicht in $H$ und es existiert eine Konstante $c > 0$, sodass*

$$\|v\|_H \leq c\|v\|_V \qquad \text{für alle } v \in V.$$

*Die Bilinearform $a : V \times V \to \mathbb{R}$ sei $V$-beschränkt und $V$-elliptisch. Dann gibt es für jedes $F \in L^2((0,T),V^*)$ und $u_0 \in H$ eine eindeutige Lösung $u \in H^1((0,T),V;H)$ vom Variationsproblem (2.6) mit*

$$\|u\|_{H^1((0,T),V;H)} \leq c_S \left[ \|u_0\|_H + \|F\|_{L^2((0,T),V^*)} \right].$$

**Beweis.** Siehe zum Beispiel [Zeidler]. ∎

**Bemerkung 2.13.** *Für das parabolische Modellproblem sind die Voraussetzungen von Theorem 2.12 für die Hilberträume $H = L^2(\Omega)$ und $V \subseteq H^1(\Omega)$ erfüllt.*

**Theorem 2.14.** *Es seien die Voraussetzungen von Theorem 2.12 erfüllt und weiters sei $F \in L^2((0,T),H)$. Dann gilt für die eindeutige Lösung $u$ von (2.6) die Abschätzung*

$$\|u(t)\|_H \leq e^{-\alpha t}\|u_0\|_H + \int_0^t e^{-\alpha(t-s)}\|F(s)\|_H \mathrm{d}s \qquad \text{für alle } t \in (0,T),$$

*mit $\alpha = \frac{c_1^a}{c} > 0$.*

**Beweis.** Siehe zum Beispiel [Zulehner]. ∎

**Korollar 2.15.** *Es seien die Voraussetzungen von Theorem 2.12 erfüllt. Für $u_0, w_0 \in H$ seien $u, w \in H^1((0,T),V;H)$ die eindeutigen Lösungen von (2.6) bezüglich den zwei Anfangswerten $u_0, w_0$ mit rechter Seite $F \in L^2((0,T),H)$. Dann gilt die Abschätzung*

$$\|u(t) - w(t)\|_H \leq e^{-\alpha t}\|u_0 - w_0\|_H \qquad \text{für alle } t \in (0,T).$$

**Beweis.** Einfach. ∎

**Bemerkung 2.16.** *Die $V$-Elliptizitätsbedingung in Theorem 2.12 kann ersetzt werden durch die Gårding-Ungleichung, also es existieren $c_1^a > 0$ und $\lambda \in \mathbb{R}$, sodass die Abschätzung*

$$a(v,v) + \lambda\|v\|_H^2 \geq c_1^a\|v\|_V^2 \qquad \text{für alle } v \in V$$

*gilt.*

## 2.2 Semi-Diskretisierung

### 2.2.1 Vertikale Linienmethode

Die Grundidee besteht darin eine Näherung für die Funktion $u = u(t)(x)$ zu finden indem man die schon bekannte Finite Elemente Methode bezüglich dem Ort anwendet. Zur Vereinfachung betrachten wir $F \in \mathcal{C}([0,T],V^*)$ und $u \in \mathcal{C}^1([0,T],V)$.

Da der Raum $V$ dicht in $H$ liegt ist, kann die Anfangsbedingung geschrieben werden als

$$\begin{aligned} & u(0) = u_0 && \text{in } H, \\ \Leftrightarrow \quad & (u(0),v)_H = (u_0,v)_H && \text{für alle } v \in H, \\ \Leftrightarrow \quad & (u(0),v)_H = (u_0,v)_H && \text{für alle } v \in V. \end{aligned}$$

Somit ist das folgende Variationsproblem der Ausgangspunkt für unsere Semi-Diskretisierung: Gesucht ist $u \in \mathcal{C}^1([0,T],V)$, sodass

$$\begin{aligned} \frac{d}{dt}(u(t),v)_H + a(u(t),v) &= \langle F(t),v\rangle && \text{für alle } v \in V \text{ und alle } t \in (0,T), \\ (u(0),v)_H &= (u_0,v)_H && \text{für alle } v \in V. \end{aligned} \tag{2.8}$$

Für die Diskretisierung betrachten wir jetzt einen endlichdimensionalen Teilraum $V_h \subset V$ und definieren dann das folgende Problem: Gesucht ist $u_h \in \mathcal{C}^1([0,T], V_h)$, sodass

$$\begin{aligned} \frac{d}{dt}(u_h(t), v_h)_H + a(u_h(t), v_h) &= \langle F(t), v_h \rangle \qquad \text{für alle } v_h \in V_h \text{ und alle } t \in (0,T), \\ (u_h(0), v_h)_H &= (u_0, v_h)_H \qquad \text{für alle } v_h \in V_h. \end{aligned} \tag{2.9}$$

Analog zu Kapitel **??** wählen wir eine Basis für den diskreten Raum $V_h$, also $V_h = \operatorname{span}\{\varphi_i\}_{i=1}^{n_h}$ und verwenden den Ansatz

$$u_h(x,t) = \sum_{j=1}^{n_h} u_j(t)\varphi_j(x)$$

mit den Zeitabhängigen Koeffizienten $u_j : [0,T] \to \mathbb{R}$, $j = 1, \dots, n_h$. Somit ist das Variationsproblem (2.9) äquivalent zu

$$\begin{aligned} \sum_{j=1}^{n_h} (\varphi_j, \varphi_i)_H u_j'(t) + \sum_{j=1}^{n_h} a(\varphi_j, \varphi_i) u_j(t) &= \langle F(t), \varphi_i \rangle \qquad \forall\, i = 1, \dots, n_h, \quad \forall t \in (0,T), \\ \sum_{j=1}^{n_h} (\varphi_j, \varphi_i)_H u_j(0) &= (u_0, \varphi_i)_H \qquad \forall\, i = 1, \dots, n_h. \end{aligned}$$

Wir definieren nun die Matrizen

$$M_h := \left[(\varphi_j, \varphi_i)_H\right]_{i,j=1}^{n_h} \qquad \text{und} \qquad K_h := [a(\varphi_j, \varphi_i)]_{i,j=1}^{n_h}$$

und die Vektoren

$$\underline{f}_h(t) := [\langle F(t), \varphi_i \rangle]_{i=1}^{n_h}, \qquad \underline{g}_h := [(u_0, \varphi_i)_H]_{i=1}^{n_h} \quad \text{und} \quad \underline{u}_h(t) := [u_j(t)]_{j=1}^{n_h}.$$

Dabei ist $K_h \in \mathbb{R}^{n_h \times n_h}$ die Steifigkeitsmatrix aus Kapitel **??** und $M_h \in \mathbb{R}^{n_h \times n_h}$ wird als Massematrix bezeichnet. Mit diesen Definitionen ist das diskretisierte Problem (2.9) äquivalent zum System gewöhnlicher Differentialgleichungen erster Ordnung: Gesucht ist $\underline{u}_h \in \mathcal{C}^1([0,T], \mathbb{R}^{n_h})$, sodass

$$\begin{aligned} M_h \underline{u}_h'(t) + K_h \underline{u}_h(t) &= \underline{f}_h(t) \qquad \text{für alle } t \in (0,T), \\ M_h \underline{u}_h(0) &= \underline{g}_h. \end{aligned} \tag{2.10}$$

**Lemma 2.17.** *Die Massematrix $M_h \in \mathbb{R}^{n_h \times n_h}$ ist symmetrisch und postiv definit.*

**Beweis.** Symmetrie folgt aus der Symmetrie Eigenschaft des Skalarprodukts, also $(\varphi_j, \varphi_i)_H = (\varphi_i, \varphi_j)_H$ und die postiv Definitheit folgt analog aus der Positivität des Skalarprodukts. ∎

**Theorem 2.18** (Picard-Lindelöf). *Sei $\underline{f} \in \mathcal{C}([0,T] \times \mathbb{R}^n, \mathbb{R}^n)$ Lipschitz-stetig bezüglich der 2. Variable, also es existiert eine Konstante $L > 0$, sodass*

$$\left\|\underline{f}(t,\underline{u}) - \underline{f}(t,\underline{v})\right\| \leq L\|\underline{u} - \underline{v}\|$$

*für alle $\underline{u}, \underline{v} \in \mathbb{R}^n$ und alle $t \in [0,T]$ erfüllt ist. Dann existiert für alle $\underline{u}_0 \in \mathbb{R}^n$ genau eine Lösung $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ des Anfangswertproblems*

$$\begin{aligned} \underline{u}'(t) &= \underline{f}(t,\underline{u}(t)) \qquad \text{für alle } t \in (0,T), \\ \underline{u}(0) &= \underline{u}_0. \end{aligned}$$

**Beweis.** Siehe gewöhnliche Differentialgleichungen. ∎

Da nach Lemma 2.17 die Massematrix $M_h$ invertierbar ist, ist die Lipschitzbedinung für die rechte Seite $\underline{f}(t,\underline{u}_h) := M_h^{-1}\left[\underline{f}_h(t) - K_h\underline{u}_h\right]$ von unserem Anfangswertproblem (2.10)

$$\left\|\underline{f}(t,\underline{u}_h) - \underline{f}(t,\underline{v}_h)\right\| = \left\|M_h^{-1}K_h(\underline{u}_h - \underline{v}_h)\right\| \leq \left\|M_h^{-1}K_h\right\|\|\underline{u}_h - \underline{v}_h\| =: L_h\|\underline{u}_h - \underline{v}_h\|$$

erfüllt mit der $h$-abhängigen Libschitz-Konstante $L_h = \left\|M_h^{-1}K_h\right\| < \infty$. Somit ist nach Theorem 2.18 das diskrete Anfangswertproblem (2.10) eindeutig lösbar.

### 2.2.2 Der Semi-Diskretisierungsfehler

Für die Näherungslösung $u_h \in \mathcal{C}^1((0,T), V_h)$ von (2.9) wollen wir den Fehler

$$\max_{t\in[0,T]} \|u(t) - u_h(t)\|_H$$

bezüglich der exakten Lösung $u \in \mathcal{C}^1((0,T), V)$ von (2.8) geeignet abschätzen. Dazu benötigen wir den folgenden Operator.

**Definition 2.19** (Ritz-Projektion). *Sei $a : V \times V \to \mathbb{R}$ eine beschränkte und elliptische Bilinearform bezüglich $V$. Die Ritz-Projektion $R_h : V \to V_h$ ist für ein $u \in V$ definiert als die Lösung des Variationsproblems: Gesucht ist $R_h u \in V_h$, sodass*

$$a(R_h u, v_h) = a(u, v_h) \qquad \text{für alle } v_h \in V_h$$

*erfüllt ist.*

Es lässt sich leicht zeigen, dass die Ritz-Projektion wohldefiniert ist und einen linearen und beschränkten Operator definiert. Weiters sieht man leicht, dass $R_h : V \to V_h$ tatsächlich eine Projektion ist, also $R_h R_h = R_h$.

Die Grundlegende Idee für die Fehlerabschätzung ist gegeben durch die Aufspaltung

$$u(t) - u_h(t) = u(t) - R_h u(t) + R_h u(t) - u_h(t) = \varrho_h(t) + \theta_h(t),$$

mit $\varrho_h(t) := u(t) - R_h u(t) \in V$ und $\theta_h(t) := R_h u(t) - u_h(t)$.

**Lemma 2.20.** *Für den Fehler* $\varrho_h(t) := u(t) - R_h u(t) \in V$ *gilt*

$$\varrho_h \in \mathcal{C}^1([0,T], V) \qquad \textit{und} \qquad \varrho_h'(t) = (I - R_h)u'(t).$$

**Beweis.** Übung. ∎

**Lemma 2.21.** *Für den Fehler* $\theta_h(t) := R_h u(t) - u_h(t) \in V_h$ *gilt*

$$\theta_h \in \mathcal{C}^1([0,t], V), \qquad \theta_h'(t) = R_h u'(t) - u_h'(t)$$

*und*

$$(\theta_h'(t), v_h)_H + a(\theta_h(t), v_h) = -(\varrho_h'(t), v_h)_H \quad \textit{für alle } v_h \in V_h \textit{ und alle } t \in [0,T].$$

**Beweis.** Die ersten beiden Behauptungen folgen analog zu Lemme 2.21. Sei $u \in \mathcal{C}^1([0,T], V)$ die Lösung von (2.8) und $u_h \in \mathcal{C}^1([0,T], V_h)$ die Näherungslösung von (2.8), dann gilt wegen $V_h \subset V$, dass

$$(u'(t), v_h)_H + a(u(t), v_h) = \langle F(t), v_h \rangle \qquad \text{für alle } v_h \in V_h \text{ und alle } t \in [0,T].$$

Daraus folgt für $v_h \in V_h$ und $t \in [0,T]$ weiters

$$\begin{aligned} 0 &= (u'(t) - u_h'(t), v_h)_H + a(u(t) - u_h(t), v_h) \\ &= (u'(t) - R_h u'(t) + R_h u'(t) - u_h'(t), v_h)_H + a(R_h u(t) - u_h(t), v_h) \\ &= (\varrho_h'(t), v_h)_H + (\theta_h'(t), v_h)_H + a(\theta_h(t), v_h). \end{aligned}$$

Und somit ist die Behauptung bewiesen. ∎

**Lemma 2.22.** *Für die obigen Fehler* $\varrho_h \in V$ *und* $\theta_h \in V_h$ *gilt die Abschätzung*

$$\|\theta_h(t)\|_H \le \|\theta_h(0)\|_H + \int_0^t \|\varrho_h'(s)\| \mathrm{d}s \qquad \textit{für alle } t \in [0,T].$$

**Beweis.** Mit Lemma 2.21 gilt für $t \in [0,T]$ mit $v_h = \theta_h$

$$(\theta_h'(t), \theta_h(t))_H + a(\theta_h(t), \theta_h(t)) = -(\varrho_h'(t), \theta_h(t))_H.$$

Unter Verwendung der Identität

$$(\theta_h'(t), \theta_h(t))_H = \frac{1}{2}\frac{d}{dt}(\theta_h(t), \theta_h(t))_H = \frac{1}{2}\frac{d}{dt}\|\theta_h(t)\|_H^2 = \|\theta_h(t)\|_H \frac{d}{dt}\|\theta_h(t)\|_H,$$

folgt dann weiters

$$\begin{aligned} \|\theta_h(t)\|_H \frac{d}{dt}\|\theta_h(t)\|_H &= -a(\theta_h(t), \theta_h(t)) - (\varrho_h'(t), \theta_h(t))_H \\ &\le \|\varrho_h'(t)\|_H \|\theta_h(t)\|_H. \end{aligned}$$

Somit gilt die Abschätzung

$$\frac{d}{dt}\|\theta_h(t)\|_H \leq \|\varrho_h'(t)\|_H \qquad \text{für alle } \|\theta_h(t)\|_H \neq 0. \tag{2.11}$$

Falls $\|\theta_h(t)\|_H = 0$ ist in einer Umgebung von $t$, dann ist auch $\frac{d}{dt}\|\theta_h(t)\|_H = 0$ in dieser Umgebung. Somit ist die Ungleichung (2.11) für fast alle $t \in [0,T]$ erfüllt. Integration über (2.11) bezüglich $(0,t), t \in [0,T]$ liefert die Behauptung

$$\|\theta_h(t)\|_H - \|\theta_h(0)\|_H = \int_0^t \frac{d}{dt}\|\theta_h(s)\|_H \mathrm{d}s \leq \int_0^t \|\varrho_h'(s)\|_H \mathrm{d}s.$$

■

**Theorem 2.23.** *Sei $u \in \mathcal{C}^1([0,T],V)$ die Lösung von (2.8) und $u_h \in \mathcal{C}^1([0,T],V_h)$ die Näherungslösung von (2.8), dann gilt für $t \in [0,T]$ die Fehlerabschätzung*

$$\|u(t) - u_h(t)\|_H \leq \|(I-R_h)u(t)\|_H + \|R_h u_0 - u_h(0)\|_H + \int_0^t \|(I-R_h)u'(s)\|_H \mathrm{d}s.$$

**Beweis.** Mit Hilfe der Lemmata 2.20–2.22 folgt die Behauptung

$$\begin{aligned}
\|u(t) - u_h(t)\|_H &= \|u(t) - R_h u(t) + R_h u(t) - u_h(t)\|_H \\
&\leq \|(I-R_h)u(t)\|_H + \|\theta_h(t)\|_H \\
&\leq \|(I-R_h)u(t)\|_H + \|\theta_h(0)\|_H + \int_0^t \|\varrho_h'(s)\| \mathrm{d}s \\
&= \|(I-R_h)u(t)\|_H + \|R_h u_0 - u_h(0)\|_H + \int_0^t \|(I-R_h)u'(s)\|_H \mathrm{d}s.
\end{aligned}$$

■

Theorem 2.23 besagt nun, dass der Fehler der Semi-Diskretisierung im wesentlichen über den Fehler der Ritz-Projektion abgeschätz werden kann. Für das parabolische Modellproblem 1D, kann unter Verwendung von

- Aubin-Nitsche Trick, siehe Theorem **??**: Es existiert eine Konstante $c > 0$, sodass für alle $w \in V \cap H^2(0,1)$ gilt

$$\|(I-R_h)w(t)\|_{L^2(0,1)} \leq ch^2 \left|w(t)\right|_{H^2(0,1)}.$$

- $L^2(0,1)$-Projektion: Sei $w \in L^2(0,1)$, dann ist $Q_h w \in V_h$ gesucht, sodass

$$(Q_h w, v_h)_{L^2(0,1)} = (w, v_h)_{L^2(0,1)} \qquad \text{für alle } v_h \in V_h$$

  gilt. Somit gilt nach (2.8) die Beziehung $u_h(0) = Q_h u_0$ und darauß folgt

$$\begin{aligned}
\|R_h u_0 - u_h(0)\|_{L^2(0,1)} &= \|R_h u_0 - Q_h u_0\|_{L^2(0,1)} \\
&\leq \|u_0 - R_h u_0\|_{L^2(0,1)} + \|u_0 - Q_h u_0\|_{L^2(0,1)} \\
&\leq ch^2 \left|u_0\right|_{H^2(0,1)}
\end{aligned}$$

eine Fehlerabschätzung gezeigt werden.

**Theorem 2.24.** *Sei $u \in \mathcal{C}^1([0,T], H^2(0,1))$ die Lösung des parabolischen Modellproblems und $u_h \in \mathcal{C}^1([0,T], V_h)$ die entsprechende Näherungslösung, dann gilt für $t \in [0,T]$ die Fehlerabschätzung*

$$\|u(t) - u_h(t)\|_{L^2(0,1)} \leq ch^2 \left[ |u(t)|_{H^2(0,1)} + |u_0|_{H^2(0,1)} + \int_0^t |u'(s)|_{H^2(0,1)} \, \mathrm{d}s \right].$$

**Beweis.** Folgt mit Theorem 2.23 unter Verwendung der obigen Ergebnisse. ∎

**Bemerkung 2.25.** *Nach Theorem 2.24 ergibt sich für das parabolische Modellproblem bezüglich der $L^2(0,1)$-Norm die gleich Konvergenzrate wie im elliptischen Fall.*

## 2.3 Explizite Runge-Kutta Verfahren für allgemeine gewöhnliche Differentialgleichungen erster Ordnung

In diesem Abschnitt betrachten wir die gewöhnlich Differentialgleichung erster Ordnung: Gesucht ist $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$, sodass

$$\begin{aligned} \underline{u}'(t) &= \underline{f}(t, \underline{u}(t)) \qquad \text{für alle } t \in (0,T), \\ \underline{u}(0) &= \underline{u}_0. \end{aligned} \tag{2.12}$$

Dabei ist $\underline{u}_0 \in \mathbb{R}^n$ und $\underline{f} \in \mathcal{C}([0,T] \times \mathbb{R}^n, \mathbb{R}^n)$. Sei $(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ ein inneres Produkt auf $\mathbb{R}^n$ mit der induzierten Norm $\|\cdot\|$. Weiters sei die Lipschitz-Bedingung

$$\left\| \underline{f}(t, \underline{u}) - \underline{f}(t, \underline{v}) \right\| \leq L \|\underline{u} - \underline{v}\| \qquad \text{für alle } \underline{u}, \underline{v} \in \mathbb{R}^n \text{ und alle } t \in [0,T]$$

erfüllt, siehe auch Theorem 2.18.

### 2.3.1 Das Euler-Verfahren

Wir betrachten die Zerlegung des Intervalls $[0,T]$

$$0 = t_0 < t_1 < \ldots < t_m = T$$

mit den Knoten/Zeitpunkten $I_\tau := \{t_0, t_1, \ldots, t_m\}$ und den Zeitschrittweiten

$$\tau_k := t_{k+1} - t_k, \quad k = 0, 1, \ldots, m-1, \qquad \tau := \max_{k=0,\ldots,m-1} \tau_k.$$

Für äquidistante Zerlegungen gilt $t_k = k\tau$, mit $\tau = \frac{T}{m}$. Zur Herleitung des Euler-Verfahrens integrieren wir die Gleichung $\underline{u}'(t) = \underline{f}(t, \underline{u}(t))$ über das Intervall $(t, t+\tau)$ und erhalten

$$\underline{u}(t+\tau) = \underline{u}(t) + \int_t^{t+\tau} \underline{f}(s, \underline{u}(s)) \mathrm{d}s \approx \underline{u}(t) + \tau \underline{f}(t, \underline{u}(t)).$$

Dies motiviert nun die folgende Vorschrift:

$$\begin{aligned} &\underline{u}_0 \in \mathbb{R}^n \text{ gegeben} \\ &\underline{u}_{k+1} = \underline{u}_k + \tau_k \underline{f}(t_k, \underline{u}_k) \qquad \text{für } k = 0, 1, \ldots, m-1. \end{aligned} \tag{2.13}$$

Die Methode (2.13) wird *explizites Euler-Verfahren* genannt.

### 2.3.2 Eine Verallgemeinerung: Einschrittverfahren

**Definition 2.26.** *Ein Schema zur Zeitdiskretisierung der Form*

$$\begin{aligned} &\underline{u}_0 \in \mathbb{R}^n \text{ gegeben} \\ &\underline{u}_{k+1} = \underline{u}_k + \tau_k \phi(t_k, \underline{u}_k, \tau_k) \qquad \text{für } k = 0, 1, \ldots, m-1 \end{aligned} \tag{2.14}$$

*mit der Verfahrensfunktion $\phi : [0,T] \times \mathbb{R}^n \times [0,\tau] \to \mathbb{R}^n$ wird als* Einschrittverfahren *bezeichnet.*

Beispiele sind:

- Das Eulerverfahren:
$$\phi(t, \underline{u}, \tau) = \underline{f}(t, \underline{u}).$$
- Das verbesserte Eulerverfahren/Mittelpunktsregel:
$$\phi(t, \underline{u}, \tau) = \underline{f}(t + \frac{\tau}{2}, \underline{u} + \frac{\tau}{2}\underline{f}(t, \underline{u})) \approx \underline{f}(t + \frac{\tau}{2}, \underline{u}(t + \frac{\tau}{2})).$$

### 2.3.3 Konvergenzanalyse für Einschrittverfahren

Die Iterationsvorschrift (2.14) liefert für gegebene Gitterpunkte $I_\tau = \{t_0, t_1, \ldots, t_m\}$ eine Näherung aus der Menge aller Gitterfunktionen

$$X_\tau := \{\underline{v}_\tau : I_\tau \to \mathbb{R}^n\}.$$

**Definition 2.27** (Globaler Fehler, Konvergenz)**.** *Sei $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ die exakte Lösung von (2.12). Weiters sei $u_\tau \in X_\tau$ die Gitterfunktion, die durch das Einschrittverfahren (2.14) für eine gegebene Verfahrensfunktion $\phi$ erzeugt wird. Wir definieren dann den* globalen Fehler *als*

$$\underline{e}_\tau : I_\tau \to \mathbb{R}^n, \qquad t_k \mapsto \underline{e}_k := \underline{u}(t_k) - \underline{u}_k.$$

*Wir messen den globalen Fehler $\underline{e}_\tau \in X_\tau$ bezüglich einer gegebene Norm $\|\cdot\|$ auf $\mathbb{R}^n$ mit*

$$\|\underline{e}_\tau\|_{X_\tau} := \max_{k=0,\ldots,m} \|\underline{e}_k\|.$$

*Weiters bezeichnen wir das Einschrittverfahren (2.14) als konvergent bezüglich der Norm $\|\cdot\|_{X_\tau}$, falls*

$$\|\underline{e}_\tau\|_{X_\tau} \to 0 \qquad \text{für } \tau \to 0.$$

*Das Einschrittverfahren (2.14) konvergiert mit der Ordnung $p \in \mathbb{N}$, falls*

$$\|\underline{e}_\tau\|_{X_\tau} = \mathcal{O}(\tau) \qquad \text{für } \tau \to 0.$$

**Definition 2.28** (Lokaler Fehler)**.** *Sei $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ die exakte Lösung von (2.12). Dann ist für das Einschrittverfahren (2.14) der* lokale Fehler *$\underline{d}_\tau \in X_\tau$ gegeben durch*

$$\begin{aligned} \underline{d}_0 &:= \underline{0}, \\ \underline{d}_{k+1} &:= \underline{u}(t_{k+1}) - [\underline{u}(t_k) + \tau_k \phi(t_k, \underline{u}(t_k), \tau_k)] \qquad \text{für } k = 0, 1, \ldots, m-1. \end{aligned}$$

**Definition 2.29** (Konsistenzfehler)**.** *Für das Einschrittverfahren (2.14) und einer gegebenen Gitterfunktion $\underline{v}_\tau \in X_\tau$ definieren wir $\underline{\psi}_\tau(\underline{v}_\tau) \in X_\tau$ als*

$$\begin{aligned}\underline{\psi}_0(\underline{v}_\tau) &:= \underline{v}_0 - \underline{u}_0,\\ \underline{\psi}_{k+1}(\underline{v}_\tau) &:= \frac{1}{\tau_k}\left[\underline{v}_{k+1} - \underline{v}_k\right] - \phi(t_k, \underline{v}_k, \tau_k) \qquad \text{für } k = 0, 1, \dots, m-1.\end{aligned}$$

*Sei $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ wiederum die exakte Lösung von (2.12) und $\underline{\tilde{u}}_\tau \in X_\tau$ die Gitterfunktion mit $\underline{\tilde{u}}_k = \underline{u}(t_k)$, $k = 0, \dots, m$, dann bezeichnen wir*

$$\underline{\psi}_\tau(\underline{\tilde{u}}_\tau) \in X_\tau$$

*als den* Konsistenzfehler*.*

**Bemerkung 2.30.** *Für den Konsistenzfehler gilt*

$$u(t_{k+1}) = u(t_k) + \tau_k\left[\phi(t_k, \underline{u}(t_k), \tau_k) + \psi_{k+1}(\underline{\tilde{u}}_\tau)\right], \qquad \text{für } k = 0, 1, \dots, m-1.$$

*Das heißt, die exakte Lösung im Punkt $t_{k+1}$ kann geschrieben werden durch die Anwendung einer Iteration des Einschrittverfahrens (2.14), wobei die Verfahrensfunktion $\phi$ durch den Konsistenzfehler $\psi_{k+1}(\underline{\tilde{u}}_\tau)$ korrigiert werden muss. Weiters gilt der Zusammenhang mit dem lokalen Fehler $\underline{d}_\tau \in X_\tau$*

$$\psi_{k+1}(\underline{\tilde{u}}_\tau) = \frac{1}{\tau_k}\underline{d}_{k+1} \qquad \text{für } k = 0, 1, \dots, m-1.$$

Die Konvergenzuntersuchung basiert auf den zwei Schritten

- Konsistenzanalyse: Die Abschätzung des Konsistenzfehlers $\psi_{k+1}(\underline{\tilde{u}}_\tau) \in X_\tau$.
- Stabilitätsanalyse: Die Analyse wie Störungen propergiert werden.

Für die Konsistenzanalyse betrachten wir eine weitere Norm $\|\cdot\|_{Y_\tau}$ auf $X_\tau$, die wir später genauer fixieren werden.

**Definition 2.31** (Konsistenz)**.** *Sei $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ die exakte Lösung von (2.12) und $\underline{\tilde{u}}_\tau \in X_\tau$ die Gitterfunktion mit $\underline{\tilde{u}}_k = \underline{u}(t_k)$, $k = 0, \dots, m$. Ein Einschrittverfahren (2.14) heißt* konsistent *bezüglich $\|\cdot\|_{Y_\tau}$, falls für den Konsistenzfehler $\underline{\psi}_\tau(\underline{\tilde{u}}_\tau) \in X_\tau$ gilt, dass*

$$\left\|\underline{\psi}_\tau(\underline{\tilde{u}}_\tau)\right\|_{Y_\tau} \to 0 \qquad \text{für } \tau \to 0.$$

*Wir sprechen weiters von der Konsistenzordnung $p \in \mathbb{N}$, falls eine Konstante $K > 0$ existiert mit*

$$\left\|\underline{\psi}_\tau(\underline{\tilde{u}}_\tau)\right\|_{Y_\tau} \leq K\tau^p.$$

**Definition 2.32** (Stabilität)**.** *Sei $\underline{u}_\tau \in X_\tau$ die Gitterfunktion die von dem Einschrittverfahren (2.14) erzeugt wird. Das Einschrittverfahren (2.14) heißt* stabil *bezüglich $\|\cdot\|_{X_\tau}$ und $\|\cdot\|_{Y_\tau}$, falls eine Konstante $c_S > 0$ existiert, sodass*

$$\|\underline{u}_\tau - \underline{v}_\tau\|_{X_\tau} \leq c_S\left\|\underline{\psi}_\tau(\underline{v}_\tau)\right\|_{Y_\tau} \qquad \text{für alle } \underline{v}_\tau \in X_\tau.$$

**Lemma 2.33.** *Ein konsistentes und stabiles Einschrittverfahren (2.14) mit der Konsistenzordnung $p \in \mathbb{N}$ ist konvergent bezüglich $\|\cdot\|_{X_\tau}$ mit der Konvergenzordnung $p$.*

**Beweis.** Sei $\underline{u} \in \mathcal{C}^1([0,T],\mathbb{R}^n)$ die exakte Lösung von (2.12) und $\tilde{\underline{u}}_\tau \in X_\tau$ die Gitterfunktion mit $\tilde{\underline{u}}_k = \underline{u}(t_k)$, $k = 0,\dots,m$. Dann gilt

$$\|\underline{e}_\tau\|_{X_\tau} = \|\underline{u}_\tau - \tilde{\underline{u}}_\tau\|_{X_\tau} \leq c_S \|\psi_\tau(\tilde{\underline{u}}_\tau)\|_{Y_\tau} \leq c_S K \tau^p.$$

Somit gilt $\|\underline{e}_\tau\|_{X_\tau} \to 0$ für $\tau \to 0$ mit der Konvergenzordnung $p$. ∎

Das heißt Konsistenz und Stabilität impliziert Konvergenz. Die Konsistenz eines Verfahrens, lässt sich meistens über eine Taylorreihenentwicklung zeigen. Um die Stabilität zeigen zu können, ist folgendes Theorem hilfreich.

**Theorem 2.34.** *Für die Verfahrensfunktion $\phi$ des Einschrittverfahrens (2.14) gelte die Lipschitz-Bedingung*

$$\|\phi(t_k,\underline{u},\tau_k) - \phi(t_k,\underline{v},\tau_k)\| \leq \Lambda \|\underline{u} - \underline{v}\| \tag{2.15}$$

*für alle $\underline{u},\underline{v} \in \mathbb{R}^n$, alle $t_k \in [0,T]$ und alle $\tau_k \in [0,\tau]$ mit einer Konstante $\Lambda > 0$, dann ist das Einschrittverfahren (2.14) stabil bezüglich der Norm*

$$\|\underline{v}_\tau\|_{Y_\tau} := \|\underline{v}_0\| + \sum_{k=1}^{m} \tau_{k-1} \|\underline{v}_k\|, \qquad \text{für alle } \underline{v}_\tau \in X_\tau$$

*mit der Stabilitätskonstante $c_S = e^{T\Lambda}$. Im Speziellen gilt somit*

$$\|\underline{e}_\tau\|_{X_\tau} \leq e^{T\Lambda} \left\| \underline{\psi}_\tau(\tilde{\underline{u}}_\tau) \right\|_{Y_\tau}.$$

**Beweis.** Induktiv, siehe zum Beispiel [Zulehner]. ∎

Für das explizite Euler-Verfahren wollen wir nun die Konsistenz und die Stabilität überprüfen. Für die Stabilität überprüfen wir die Lipschitz-Bedingung der Verfahrensfunktion $\underline{\phi}(t_k,\underline{u},\tau_k) = \underline{f}(t_k,\underline{u})$, also

$$\left\|\underline{\phi}(t_k,\underline{u},\tau_k) - \underline{\phi}(t_k,\underline{v},\tau_k)\right\| = \left\|\underline{f}(t_k,\underline{u}) - \underline{f}(t_k,\underline{v})\right\| \leq L\|\underline{u} - \underline{v}\|.$$

Somit ist nach Theorem 2.34 die Stabilitätskonstante gegeben durch $c_S = e^{TL}$. Weiters gilt für den Konsistenzfehler die Abschätzung

$$\left\|\underline{\psi}_\tau(\tilde{\underline{u}}_\tau)\right\|_{Y_\tau} = \left\|\underline{\psi}_0(\tilde{\underline{u}}_\tau)\right\| + \sum_{k=0}^{m-1} \tau_k \left\|\underline{\psi}_{k+1}(\tilde{\underline{u}}_\tau)\right\| \leq \tau \sum_{k=0}^{m-1} \left\|\underline{\psi}_{k+1}(\tilde{\underline{u}}_\tau)\right\|.$$

Für $\underline{u} \in \mathcal{C}^2([0,T],\mathbb{R}^n)$ gilt mit Hilfe der Taylor-Entwicklung, dass

$$\underline{\psi}_{k+1}(\tilde{\underline{u}}_\tau) = \frac{1}{\tau_k}\left[\underline{u}(t_{k+1}) - \underline{u}(t_k)\right] - \underline{\phi}(t_k,\underline{u}(t_k),\tau_k) = \frac{1}{\tau_k}\left[\underline{u}(t_{k+1}) - \underline{u}(t_k)\right] - \underline{f}(t_k,\underline{u}(t_k))$$

$$
\begin{aligned}
&= \frac{1}{\tau_k}\left[\underline{u}(t_{k+1}) - \underline{u}(t_k)\right] - \underline{u}'(t_k) \\
&= \frac{1}{\tau_k}\left[\underline{u}(t_k) + \underline{u}'(t_k)\tau_k + \int_{t_k}^{t_{k+1}} (t_{k+1}-t)u''(t)\mathrm{d}t - \underline{u}(t_k)\right] - \underline{u}'(t_k) \\
&= \frac{1}{\tau_k}\int_{t_k}^{t_{k+1}} (t_{k+1}-t)u''(t)\mathrm{d}t.
\end{aligned}
$$

Dies ergibt die Abschätzung

$$
\left\|\underline{\psi}_{k+1}(\tilde{\underline{u}}_\tau)\right\| \leq \int_{t_k}^{t_{k+1}} \|\underline{u}''(t)\|\mathrm{d}t.
$$

Somit gilt die Konsistenzabschätzung

$$
\left\|\underline{\psi}_{\tau}(\tilde{\underline{u}}_\tau)\right\|_{Y_\tau} \leq \tau \sum_{k=0}^{m-1} \left\|\underline{\psi}_{k+1}(\tilde{\underline{u}}_\tau)\right\| \leq \tau \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \|\underline{u}''(t)\|\mathrm{d}t = \tau \int_0^T \|\underline{u}''(t)\|\mathrm{d}t
$$

mit der Konsistenzordnung $p = 1$. Für den Fehler des Euler-Verfahrens gilt somit nach Lemma 2.33 die Abschätzung

$$
\max_{k=0,\ldots,m} \|\underline{u}(t_k) - \underline{u}_k\| \leq \tau e^{TL} \int_0^T \|\underline{u}''(t)\|\mathrm{d}t.
$$

### 2.3.4 Explizite $s$-stufige Runge-Kutta Verfahren

Ziel ist es nun Verfahren mit möglichst hoher Konsistenzordnung $p \in \mathbb{N}$ zu konstruieren. Die Idee besteht darin das Integral mit $s \in \mathbb{N}$ Integrationspunkten zu approximieren, also

$$
\int_t^{t+\tau} \underline{f}(\sigma, \underline{u}(\sigma))\mathrm{d}\sigma \approx \tau \sum_{i=1}^{s} b_i \underline{f}(t + c_i\tau, \underline{u}(t + c_i\tau)).
$$

Dabei sind $b_i$, $i = 1, \ldots, s$ gegebene Gewichte mit

$$
b_i \in [0,1] \qquad \text{und} \qquad \sum_{i=1}^{s} b_i = 1.
$$

Weiters bezeichnen wir $c_i$, $i = 1, \ldots, s$ als Koeffizienten, wobei

$$
c_1 = 0 \qquad \text{und} \qquad c_i \in (0,1] \text{ für } i = 2, \ldots, s
$$

gilt. Da die Funktionen $\underline{u}(t + c_i\tau) \in \mathbb{R}^n$ für $i = 2, \ldots, s$ unbekannt sind, führen wir für $i = 2, \ldots, s$ die weitere Approximationen

$$
\underline{u}(t + c_i\tau) = u(t) + \int_t^{t+c_i\tau} \underline{f}(\sigma, u(\sigma))\mathrm{d}\sigma \approx u(t) + \tau \sum_{j=1}^{i-1} a_{ij} \underline{f}(t + c_j\tau, \underline{u}(t + c_j\tau))
$$

ein. Dabei sind $a_{ij}$, $j = 1, \ldots, i-1$ für $i = 2, \ldots, s$ wiederum gegebene Gewichte. Diese Approximationen führen auf folgendes Schema:

Für eine gegebene Näherung $\underline{u}_k \in \mathbb{R}^n$ im Punkt $t_k$ berechne die nächste Näherung $\underline{u}_{k+1} \in \mathbb{R}^n$ im Punkt $t_{k+1}$ als

$$\begin{aligned}
\underline{g}_1 &= \underline{u}_k, \\
\underline{g}_2 &= \underline{u}_k + \tau_k a_{21} \underline{f}(t_k + c_1\tau, \underline{g}_1), \\
\underline{g}_3 &= \underline{u}_k + \tau_k \left[ a_{31} \underline{f}(t_k + c_1\tau, \underline{g}_1) + a_{32} \underline{f}(t_k + c_2\tau_k, \underline{g}_2) \right], \\
&\vdots \\
\underline{g}_s &= \underline{u}_k + \tau_k \left[ a_{s1} \underline{f}(t_k + c_1\tau, \underline{g}_1) + \ldots + a_{s,s-1} \underline{f}(t_k + c_{s-1}\tau_k, \underline{g}_{s-1}) \right], \\
\underline{u}_{k+1} &= \underline{u}_k + \tau_k \sum_{i=1}^{s} b_i \underline{f}(t + c_i\tau_k, \underline{g}_i).
\end{aligned} \tag{2.16}$$

Das Schema (2.16) wird als $s$-stufiges Runge-Kutta Verfahren bezeichnet. Dabei können je nach Wahl von $a_{ij}$, $b_i$ und $c_i$ für $j = 1, \ldots, i-1$ und $i = 1, \ldots, s$ die verschiedensten Verfahren konstruiert werden. Diese werden oft im sogenannten *Butcher-Tableau* zusammengefasst

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & & \ddots & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array} \qquad \text{bzw.} \qquad \begin{array}{c|c}
\underline{c} & A \\
\hline
 & \underline{b}^\top
\end{array}$$

**Beispiel 2.35.** *1.) Explizites Euelr-Verfahren:*

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

*2.) Verbessertes Euler-Verfahren:*

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

*3.) Klassisches 4-stufiges Runge-Kutta Verfahren mit Konsistenzordnung $p = 4$*

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

**Bemerkung 2.36.** *Die* Stabilitätsanalyse *von s-stufigen Runge-Kutta Verfahren verläuft wiederum über Theorem 2.34, indem man ausgehend von der Lipschitz-Stetigkeit von $\underline{f} : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ die Lipschitz-Stetigkeit der Verfahrensfunktion $\underline{\phi} : [0,T] \times \mathbb{R}^n \times [0,\tau] \to \mathbb{R}^n$ zeigt. Die* Konsistenzanalyse *erfolgt wiederum über Taylorreihenentwicklung.*

**Bemerkung 2.37.** *Nicht jedes s-stufige Runge-Kutta Verfahren hat die Konsistenzordnung $p = s$. Dabei gilt folgendes:*

| Stufe $s$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $s \geq 10$ |
|---|---|---|---|---|---|---|---|---|---|---|
| max. Konsistenzordnung $p$ | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | $p \leq s-3$ |

## 2.4 Steife Differentialgleichungen

Für das parabolische Modellproblem erhalten wir nach der Anwendung der Finiten Elemente Methode das System von gewöhnlichen Differentialgleichungen erster Ordnung

$$\begin{aligned} \underline{u}_h'(t) &= \underline{f}_h(t, \underline{u}_h(t)) := M_h^{-1} \left[ \underline{f}_h(t) - K_h \underline{u}_h(t) \right] \qquad \text{für } t \in (0,T), \\ \underline{u}_h(0) &= M_h^{-1} \underline{g}_h. \end{aligned}$$

Dabei geht für Einschrittverfahren die Lipschitz-Konstante $L$ der Funktion $\underline{f}_h : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ in die Fehlerabschätzung ein. Für unser parabolisches Modellproblem 1D gilt dabei folgendes

$$\left\| \underline{f}_h(t, \underline{u}_h) - \underline{f}_h(t, \underline{v}_h) \right\|_{M_h} \leq L_h \|\underline{u}_h - \underline{v}_h\|_{M_h}$$

mit der Lipschitzkonstante

$$L_h = \lambda_{\max}(M_h^{-1} K_h) \leq \frac{12}{h^2}.$$

**Lemma 2.38.** *Für das parabolische Modellproblem 1D gilt für eine gleichmäßige Zerlegung des Intervalls $(0,1)$ mit der Maschenweite $h > 0$ die Abschätzung*

$$\frac{3}{h^2} \leq \lambda_{\max}(M_h^{-1} K_h) \leq \frac{12}{h^2}.$$

**Beweis.** Übung. ∎

Somit ergibt sich für das explizite Euler-Verfahren die Stabilitätskonstante $c_S = e^{12Th^{-2}}$. Weiter folgt dann für den Fehler die Abschätzung

$$\max_{k=0,\dots,m} \|\underline{u}_h(t_k) - \underline{u}_k\| \leq \tau e^{12Th^{-2}} \int_0^T \|\underline{u}_h''(t)\| \mathrm{d}t.$$

Das heißt, die Zeitschrittweite $\tau$ muss nach dieser Abschätzung extrem klein gewählt werden ($\tau = e^{\mathcal{O}(h^{-2})}$)), um vernünftige Näherungslösungen zu erhalten.

Es stellt sich nun die Frage, ob wir diese Abschätzung verbessern können? Beziehungsweise, ob wir die Stabilitätskonstante $c_S$ besser abschätzen können?

### 2.4.1 Dissipative Systeme

**Lemma 2.39.** *Für die gewöhnliche Differentialgleichung*

$$\begin{aligned} \underline{u}'(t) &= \underline{f}(t, \underline{u}(t)) \qquad \text{für } t \in (0,T), \\ \underline{u}(0) &= \underline{u}_0 \end{aligned} \tag{2.17}$$

*gelte die* einseitige Lipschitz-Bedingung

$$\big(\underline{f}(t,\underline{u}) - \underline{f}(t,\underline{v}), \underline{u} - \underline{v}\big) \leq \nu \|\underline{u} - \underline{v}\|^2 \quad \text{für alle } \underline{u}, \underline{v} \in \mathbb{R}^n \text{ und alle } t \in [0,T]. \tag{2.18}$$

*Weiters seien $\underline{u}, \underline{w} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$ die zwei Lösungen von (2.17) bezüglich den zwei Anfangswerten $\underline{u}_0, \underline{w}_0 \in \mathbb{R}^n$. Dann gilt*

$$\|\underline{u}(t) - \underline{w}(t)\| \leq e^{t\nu} \|\underline{u}_0 - \underline{w}_0\| \qquad \text{für alle } \ t \in [0,T].$$

**Beweis.** Es seien $\underline{u}_0, \underline{w}_0 \in \mathbb{R}^n$. Unter Verwendung der Identität

$$\begin{aligned} \|\underline{u}(t) - \underline{w}(t)\| \frac{d}{dt} \|\underline{u}(t) - \underline{w}(t)\| &= (\underline{u}'(t) - \underline{w}'(t), \underline{u}(t) - \underline{w}(t)) \\ &= \big(\underline{f}(t, u(t)) - \underline{f}(t, w(t)), \underline{u}(t) - \underline{w}(t)\big) \\ &\leq \nu \|\underline{u}(t) - \underline{w}(t)\|^2, \end{aligned}$$

siehe dazu auch den Beweis von Lemma 2.22, folgt

$$\frac{d}{dt} \|\underline{u}(t) - \underline{w}(t)\| \leq \nu \|\underline{u}(t) - \underline{w}(t)\| \qquad \text{für fast alle } t \in [0,T].$$

Dies impliziert

$$\begin{aligned} 0 &\geq e^{-t\nu} \left[ \frac{d}{dt} \|\underline{u}(t) - \underline{w}(t)\| - \nu \|\underline{u}(t) - \underline{w}(t)\| \right] \\ &= \frac{d}{dt} \left[ e^{-t\nu} \|\underline{u}(t) - \underline{w}(t)\| \right] \end{aligned}$$

für fast alle $t \in [0,T]$. Integration dieser Ungleichung über $(0,t)$ mit $t \in [0,T]$ ergibt die Behauptung des Lemmas

$$0 \geq e^{-t\nu}\|\underline{u}(t) - \underline{w}(t)\| - e^{-0\,\nu}\|\underline{u}(0) - \underline{w}(0)\| = e^{-t\nu}\|\underline{u}(t) - \underline{w}(t)\| - \|\underline{u}_0 - \underline{w}_0\|.$$

■

Es ist einfach zu sehen, dass die Lipschitz-Bedinung

$$\big\|\underline{f}(t,\underline{u}) - \underline{f}(t,\underline{v})\big\| \leq L\|\underline{u} - \underline{v}\| \qquad \text{für alle } \underline{u},\underline{v} \in \mathbb{R}^n \text{ und alle } t \in [0,T]$$

die einseitige Lipschitz-Bedingung (2.18) mit $\nu = L$ impliziert. Jedoch gibt es eine große Klasse von gewöhnlichen Differentialgleichungen mit der Eigenschaft

$$\nu \ll L.$$

Diese Klasse von gewöhnlichen Differentialgleichungen werden als *steif* bezeichnet.

**Definition 2.40** (Dissipative)**.** *Die gewöhnliche Differentialgleichung (2.17) wird als dissipativ bezeichnet, falls die einseitige Lipschitz-Bedingung (2.18) mit $\nu = 0$ erfüllt ist.*

Für dissipative Differentialgleichungen gilt nach Lemma 2.39 für zwei unterschiedliche Startwerte $\underline{u}_0, \underline{w}_0 \in \mathbb{R}^n$, dass

$$\|\underline{u}(t) - \underline{w}(t)\| \leq \|\underline{u}_0 - \underline{w}_0\| \qquad \text{für alle } \ t \in [0,T].$$

Das heißt die unterschiedlichen Lösungen driften nicht auseinander.

**Lemma 2.41.** *Das Anfangswertproblem (2.10) ist dissipativ bezüglich dem von der Massematrix $M_h \in \mathbb{R}^{n_h \times n_h}$ induziertem Skalaprodukt.*

**Beweis.** Die rechte Seite von (2.10) ist gegeben durch

$$\underline{f}_h(t,\underline{u}_h) = M_h^{-1}\left[\underline{f}_h(t) - K_h\underline{u}_h\right] \qquad \text{für } t \in [0,T] \text{ und } \underline{u}_h \in \mathbb{R}^{n_h}.$$

Somit gilt

$$\begin{aligned}\Big(\underline{f}_h(t,\underline{u}_h) - \underline{f}_h(t,\underline{v}_h), \underline{u}_h - \underline{v}_h\Big)_{M_h} &= -(K_h(\underline{u}_h - \underline{v}_h), \underline{u}_h - \underline{v}_h)_{\ell^2} \\ &= -a(u_h - v_h, u_h - v_h) \leq 0.\end{aligned}$$

■

### 2.4.2 Kontraktion und A-Stabilität

Für dissipative Differentialgleichungen wünschen wir uns nun, dass die aus den Einschrittverfahren resultierenden Approximationen ähnliche Eigenschaften aufweisen

**Definition 2.42** (Kontraktiv)**.** *Ein Einschrittverfahren (2.14) heißt* kontraktiv *für eine gegebene Zerlegung $I_\tau = \{t_0, t_1, \dots, t_m\}$ falls*

$$\left\| \left[\underline{u} + \tau_k \underline{\phi}(t_k, \underline{u}, \tau_k)\right] - \left[\underline{w} + \tau_k \underline{\phi}(t_k, \underline{w}, \tau_k)\right] \right\| \leq \|\underline{u} - \underline{w}\| \tag{2.19}$$

*für alle $\underline{u}, \underline{w} \in \mathbb{R}^n$ und alle $k = 0, \dots, m-1$ gilt.*

**Lemma 2.43.** *Ein kontraktives Einschrittverfahren (2.14) ist stabil mit der Stabilitätskonstante $c_S = 1$.*

**Beweis.** Durch rekursive Anwendung der Kontraktionseigenschaft (2.19). ∎

Falls nun ein konsistentes Verfahren für eine Folge von Zerlegungen $I_\tau$ für $\tau \to 0$ kontraktiv ist, dann ist dieses Verfahren konvergent und die auftretenden Konstanten hängen nur mehr von der Konsistenzkonstante $K$ ab.

Es stellt sich nun die Frage, wann ein Verfahren kontraktiv ist? Wir wollen dies für das Anfangswertproblem: Gesucht ist $\underline{u} \in \mathcal{C}^1([0,T], \mathbb{R}^n)$, sodass

$$\begin{aligned} \underline{u}'(t) &= J\underline{u}(t) + \underline{f}(t) \qquad \text{für } t \in (0,T), \\ \underline{u}(0) &= \underline{u}_0 \end{aligned} \tag{2.20}$$

erfüllt ist, untersuchen. Dabei ist $J = XDX^\top \in \mathbb{R}^{n\times n}$ eine konstante und normale Matrix. Das Problem (2.20) ist offensichtlich äquivalent zu: Gesucht ist $\hat{\underline{u}} \in \mathcal{C}^1([0,T], \mathbb{C}^n)$, sodass

$$\begin{aligned} \hat{\underline{u}}'(t) &= D\hat{\underline{u}}(t) + X^\top \underline{f}(t) \qquad \text{für } t \in (0,T), \\ \hat{\underline{u}}(0) &= X^\top \underline{u}_0 \end{aligned} \tag{2.21}$$

erfüllt ist. Da $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{C}^{n\times n}$ eine Diagonalmatrix ist, sind die Gleichungen in (2.21) entkoppelt. Es genügt also die Kontraktion unserer Einschrittverfahren für die Dahlquistsche Testgleichung: Gesucht ist $u \in \mathcal{C}^1([0,T], \mathbb{C})$, sodass

$$u'(t) = \lambda u(t) \quad \text{für } t \in (0,T), \qquad u(0) = u_0 \in \mathbb{C} \tag{2.22}$$

zu untersuchen. Das System (2.20) ist genau dann dissipativ wenn gilt

$$\begin{aligned} & (J\underline{u} - J\underline{v}, \underline{u} - \underline{v}) \leq 0 && \text{für alle } \underline{u}, \underline{v} \in \mathbb{R}^n \\ \Leftrightarrow \quad & (J\underline{v}, \underline{v}) \leq 0 && \text{für alle } \underline{v} \in \mathbb{R}^n \\ \Leftrightarrow \quad & \operatorname{Re}(\lambda_i(J)) \leq 0 && \text{für alle } i = 1, \dots, n. \end{aligned}$$

Das heißt zu untersuchen ist die Kontraktion unserer Einschrittverfahren für die Dahlquistsche Testgleichung (2.22) für den Fall $\operatorname{Re}(\lambda) \leq 0$. Wir wenden also nun ein belibieges $s$-stufiges Runge-Kutta Verfahren auf die skalare Gleichung (2.22) an und erhalten

$$g_i = u_k + \tau_k \sum_{j=1}^{i-1} a_{ij} \lambda g_j \qquad \text{für } i = 1, \dots, s,$$

$$u_{k+1} = u_k + \tau_k \sum_{i=1}^{s} b_i \lambda g_i.$$

Dies ist gleichbedeutent mit

$$\underline{g} = u_k \underline{e} + \tau_k \lambda A \underline{g} \qquad \text{und} \qquad u_{k+1} = u_k + \tau_k \lambda \underline{b}^\top \underline{g},$$

wobei $\underline{g} := \{g_1, \ldots, g_s\}^\top \in \mathbb{C}^s$ und $\underline{e} := \{1, \ldots, 1\}^\top \in \mathbb{R}^s$. Somit kann die neue Näherung berechnet werden als

$$u_{k+1} = R(\tau_k \lambda) u_k, \qquad \text{mit } R(z) := 1 + z\underline{b}^\top \left[I - zA\right]^{-1} \underline{e}, \quad \text{für } z \in \mathbb{C}.$$

Die motiviert nun folgende Definition.

**Definition 2.44** (A-Stabilitätsfunktion, A-Stabilitätsgebiet)**.** *Für ein s-stufiges Runge-Kutta Verfahren mit $A \in \mathbb{R}^{s\times s}$ und $\underline{b}, \underline{c} \in \mathbb{R}^s$ wird die Funktion*

$$R : \mathbb{C} \to \mathbb{C}, \qquad z \mapsto 1 + z\underline{b}^\top \left[I - zA\right]^{-1} \underline{e}$$

*als* A-Stabilitätsfunktion *bezeichnet. Weiters definieren wir für ein Einschrittverfahren das* A-Stabilitätsgebiet *als*

$$S := \left\{z \in \mathbb{C} : |R(z)| \leq 1\right\}.$$

**Theorem 2.45.** *Ein s-stufiges Runge-Kutta Verfahren mit der Stabilitätsfunktion $R : \mathbb{C} \to \mathbb{C}$ ist genau dann kontraktiv für das dissipative Problem (2.20), also* $\mathrm{Re}(\lambda_i(J)) \leq 0$ *für alle $i = 1, \ldots, n$, wenn*

$$|R(\tau_k \lambda)| \leq 1 \qquad \textit{für alle } \lambda \in \sigma(J) \textit{ und alle } k = 0, \ldots, m-1$$

*gilt.*

**Beweis.** Sei $t_k \in I_\tau$ und $\underline{u}, \underline{v} \in \mathbb{R}^n$ mit $\underline{u} \neq \underline{v}$, dann gilt

$$\begin{aligned}
& \left\| \left[\underline{u} + \tau_k \underline{\phi}(t_k, \underline{u}, \tau_k)\right] - \left[\underline{v} + \tau_k \underline{\phi}(t_k, \underline{v}, \tau_k)\right] \right\| \leq \|\underline{u} - \underline{v}\|, \\
\Leftrightarrow \quad & \|R(\tau_k\lambda)\underline{u} - R(\tau_k\lambda)\underline{v}\| \leq \|\underline{u} - \underline{v}\| && \text{für alle } \lambda \in \sigma(J), \\
\Leftrightarrow \quad & |R(\tau_k\lambda)| \, \|\underline{u} - \underline{v}\| \leq \|\underline{u} - \underline{v}\| && \text{für alle } \lambda \in \sigma(J), \\
\Leftrightarrow \quad & |R(\tau_k\lambda)| \leq 1 && \text{für alle } \lambda \in \sigma(J).
\end{aligned}$$

■

**Definition 2.46** (A-stabil)**.** *Ein Einschrittverfahren heißt* A-stabil*, falls das A-Stabilitätsgebiet die ganze linke komplexe Halbebene $\mathbb{C}^-$ beinhaltet, also*

$$|R(z)| \leq 1 \qquad \textit{für alle } z \in \mathbb{C} \quad \textit{mit} \quad \mathrm{Re}(z) \leq 0.$$

Falls nun ein Einschrittverfahren A-stabil ist, dann ist die Kontraktion und somit nach Lemma 2.43 die Stabilität mit der Konstante $c_S = 1$ für dissipative Systeme der Form (2.20) sichergestellt.

**Beispiel 2.47.** *Die Anwendung des expliziten Euler-Verfahrens auf die Dahlquistsche Testgleichung (2.22) ergibt*

$$u_{k+1} = u_k + \tau_k \lambda u_k = (1 + \tau_k \lambda) u_k = R(\tau_k \lambda) u_k,$$

*mit $R(z) := 1 + z$. Das A-Stabilitätsgebiet ist somit gegeben durch*

$$S = \{z \in \mathbb{C} : |1 + z| \leq 1\}.$$

*Um nun die Kontraktion des Euler-Verfahrens für das parabolische Modellproblem 1D, also für das System (2.10) sicher zu stellen, muss gelten, dass*

$$\begin{aligned} & |1 + \tau_k \lambda| \leq 1 && \text{für alle } \lambda \in \sigma(-M_h^{-1} K_h) \\ \Leftrightarrow \quad & \tau_k \leq \frac{2}{|\lambda|} && \text{für alle } \lambda \in \sigma(-M_h^{-1} K_h). \end{aligned}$$

*Da $\lambda_{\max}(M_h^{-1} K_h) \leq 12 h^{-2}$ gilt, ist durch die Forderung an die Schrittweiten*

$$\tau_k \leq \frac{2}{12 h^{-2}} = \frac{1}{6} h^2 \qquad \text{für alle } k = 0, \ldots, m-1$$

*eine hinreichende Bedinung gegeben, die die Kontraktion und somit die Stabilität gewährleistet. Das heißt, es werden sehr kleine Zeitrschrittweiten $\tau = \mathcal{O}(h^2)$ im Vergleich zu den Maschenweiten $h$ benötigt um noch "vernünftige" Näherungslösungen zu erhalten.*

**Bemerkung 2.48.** *Die Stabilitätsfunktion eines beliebigen expliziten $s$-stufigen Runge-Kutta Verfahrens, also*

$$R(z) = 1 + z \underline{b}^\top [I - zA]^{-1} \underline{e}$$

*ist ein Polynom vom Grad kleiner gleich $s$. Somit folgt, dass kein expliziten $s$-stufiges Runge-Kutta Verfahren A-stabil, da Polynome unbeschränkt sind für $z \to -\infty$.*

## 2.5 Implizite Runge-Kutta Verfahren

Ziel ist es nun A-stabile Verfahren zu konstruieren. Dazu approximatieren wir das Integral vorerst mit Hilfe der rechten Rechtecksregel und erhalten

$$\underline{u}(t + \tau) = \underline{u}(t) + \int_t^{t+\tau} \underline{f}(s, \underline{u}(s)) \mathrm{d}s \approx \underline{u}(t) + \tau \underline{f}(t + \tau, \underline{u}(t + \tau)).$$

Dies motiviert dann die Vorschrift: :

$$\begin{aligned} & \underline{u}_0 \in \mathbb{R}^n \text{ gegeben} \\ & \underline{u}_{k+1} = \underline{u}_k + \tau_k \underline{f}(t_{k+1}, \underline{u}_{k+1}) \qquad \text{für } k = 0, 1, \ldots, m-1. \end{aligned} \tag{2.23}$$

Das Verfahren (2.23) wird als *implizites Euler-Verfahren* bezeichnet. Dabei ist die Näherung $\underline{u}_{k+1} \in \mathbb{R}^n$ implizit gegeben, das heißt in jedem Schritt muss nach $\underline{u}_{k+1}$ aufgelöst werden. Ein weiteres implizites Verfahren ist zum Beispiel gegben durch die implizite Mittelpunktsregel:
Sei $\underline{u}_0 \in \mathbb{R}^n$ gegeben, dann berechne für $k = 0, 1, \ldots, m-1$

$$\underline{g}_1 = \underline{u}_k + \frac{\tau_k}{2} \underline{f}(t_k + \frac{\tau_k}{2}, \underline{g}_1),$$
$$\underline{u}_{k+1} = \underline{u}_k + \tau_k \underline{f}(t_k + \frac{\tau_k}{2}, g_1).$$

Allgemeine $s$-stufige implizite Runge-Kutta Verfahren sind gegeben durch die Vorschrift: Für eine gegebene Näherung $\underline{u}_k \in \mathbb{R}^n$ im Punkt $t_k$ berechne die neue Näherung $\underline{u}_{k+1} \in \mathbb{R}^n$ im Punkt $t_{k+1}$ als

$$\underline{g}_i = \underline{u}_k + \tau_k \sum_{j=1}^{s} a_{ij} \underline{f}(t_k + c_j \tau_k, \underline{g}_j) \qquad \text{für } i = 1, \ldots, s, \tag{2.24}$$

$$\underline{u}_{k+1} = \underline{u}_k + \tau_k \sum_{i=1}^{s} b_i \underline{f}(t_k + c_i \tau_k, \underline{g}_i). \tag{2.25}$$

Dabei ist (2.24) im Allgemeinen ein gekoppeltes System von $s \in \mathbb{N}$ Gleichungen, welches nach $\underline{g} := \{\underline{g}_1^\top, \ldots, \underline{g}_s^\top\}^\top \in \mathbb{R}^{s\,n}$ aufgelöst werden muss, um die neue Näherung $\underline{u}_{k+1} \in \mathbb{R}^n$ berechnen zu können. Eine kompakte Schreibweise für die Gleichungen (2.24)–(2.25) ist wiederum über das *Butcher-Tableau* gegeben:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \qquad \text{bzw.} \qquad \begin{array}{c|c} \underline{c} & A \\ \hline & \underline{b}^\top \end{array}$$

**Definition 2.49.** *Ein $s$-stufige Runge-Kutta Verfahren mit $A \in \mathbb{R}^{s\times s}$ und $\underline{b}, \underline{c} \in \mathbb{R}^s$ heißt* explizit*, falls die Matrix $A$ eine strikt untere Dreiecksmatrix ist. Anderenfalls bezeichnen wir dieses Schema als* implizit*.*

**Beispiel 2.50.** *1.) Implizites Euler-Verfahren:*

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

*2.) Implizite Mittelpunktsregel:*

*3.) $\theta$-Methode mit $\theta \in [0, 1]$*

Für die $\theta$-Methode ergeben sich die Sonderfälle:

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

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

- $\theta = 0$: explizite Euler-Verfahren.
- $\theta = 1$: implizites Euler-Verfahren.
- $\theta = \frac{1}{2}$ : sogenannte implizite Trapezregel.

Für das System (2.24) muss im jeden Schritt $k = 0, \ldots, m-1$ die Lösbarkeit nach $\underline{g} = \{\underline{g}_1^\top, \ldots, \underline{g}_s^\top\}^\top \in \mathbb{R}^{s\,n}$ gewährleistet sein. Dazu schreiben wir die Gleichungen (2.24) als Fixpunktgleichung

$$\underline{g} = G(\underline{g}, t_k, \underline{u}_k, \tau_k). \tag{2.26}$$

**Bemerkung 2.51.** *Falls die Funktion $\underline{f} : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ Lipschitz-stetig im zweiten Argument ist, dann lässt sich mit dem Banachschen-Fixpunktsatz zeigen, dass die Fixpunktgleichung (2.26) eindeutig nach $\underline{g} \in \mathbb{R}^{s\,n}$ aufgelöst werden kann, falls die Schrittweite $\tau_k$ klein genug gewählt wird.*

*Für dissipative Systeme lässt sich zeigen, dass die Fixpunktgleichung (2.26) für beliebige Schrittweiten $\tau_k > 0$ immer nach $\underline{g} \in \mathbb{R}^{s\,n}$ aufgelöst werden kann.*

Falls die Fixpunktgleichung (2.26) bzw. das System (2.24) nun eindeutig Lösbar sind, können wir die Vectoren $\underline{g}_i$ für $i = 1, \ldots, s$ bezüglich den gegebenen Daten $t_k \in \mathbb{R}$, $\underline{u}_k \in \mathbb{R}^n$ und $\tau_k$ berechnen und wir schreiben

$$\underline{g}_i = \gamma_i(t_k, \underline{u}_k, \tau_k) \qquad \text{für } i = 1, \ldots, s.$$

Somit ergibt sich die neue Näherung $\underline{u}_{k+1} \in \mathbb{R}^n$ als

$$\begin{aligned} \underline{u}_{k+1} &= \underline{u}_k + \tau_k \sum_{i=1}^{s} b_i \underline{f}(t_k + c_i\tau_k, \gamma_i(t_k, \underline{u}_k, \tau_k)) \\ &=: \underline{u}_k + \tau_k \underline{\phi}(t_k, \underline{u}_k, \tau_k). \end{aligned}$$

Das heißt, alle $s$-stufigen Runge-Kutta Verfahren (implizit oder explizit) sind Einschrittverfahren der Form (2.14). Somit können alle gezeigten Aussagen für allgemeine Einschrittverfahren auf die impliziten $s$-stufigen Runge-Kutta Verfahren angewendet werden.

### 2.5.1 Konsistenz impliziter Runge-Kutta Verfahren

Wie für die expliziten Verfahren lässt sich die Konsistenz von impliziten Verfahren über Taylorreihenentwicklungen zeigen. Dabei gilt folgendes:

- Implizetes Euler-Verfahren: Konsistenzordnung $p = 1$.
- Implizite Mittelpunktsregel: Konsistenzordnung $p = 2$.
- $\theta$-Methode:
  - $\theta \neq \frac{1}{2}$: Konsistenzordnung $p = 1$.
  - $\theta = \frac{1}{2}$: Konsistenzordnung $p = 2$.

Allgemein lässt sich mit einem implizitem $s$-stufigen Runge-Kutta Verfahren die maximale Konsistenzordnung $p = 2s$ erreichen. Solche Verfahren heißen auch Runge-Kutta Verfahren vom Gauß-Typ.

### 2.5.2 A-Stabilität von Runge-Kutta Verfahren

Für implizite Runge-Kutta Methoden lassen sich die Begriffe wie *A-Stabilitätsfunktion*, *A-Stabilitätsgebiet* und *A-Stabilität* anolog eingeführt werden. Dabei gilt wiederum

$$R(z) = 1 + z\underline{b}^\top \left[I - zA\right]^{-1} \underline{e},$$

mit $\underline{e} = \{1, \ldots, 1\}^\top$. Dabei ist die Matrix $I - zA \in \mathbb{R}^{s\times s}$ genau dann invertierbar, falls

$$\frac{1}{z} \notin \sigma(A).$$

Weiters lässt sich zeigen, dass die A-Stabilitätsfunktion von $s$-stufigen Runge-Kutta Verfahren die Form

$$R(z) = \frac{P(z)}{Q(z)}, \qquad \text{wobei } P, Q \in \mathbb{P}_s \quad \text{mit} \quad P(0) = Q(0) = 1$$

besitzt. Für explizite Runge-Kutta Verfahren gilt also $Q(z) = 1$. Wiederum ist das A-Stabilitätsgebiet gegeben durch

$$S = \left\{z \in \mathbb{C} : |R(z)| \leq 1\right\}.$$

Ein implizites Runge-Kutta Verfahren heißt A-stabil, falls $\mathbb{C}^- \subset S$ gilt.

**Beispiel 2.52.** • *Für das implizite Euler-Verfahren gilt für* $f(t, u(t)) = \lambda u(t)$

$$u_{k+1} = u_k + \tau_k \lambda u_{k+1}.$$

*Dies impliziert*

$$u_{k+1} = \frac{1}{1 - \tau_k\lambda} u_k = R(\tau_k\lambda)u_k, \quad \textit{mit} \quad R(z) = \frac{1}{1-z} \textit{ für } z \in \mathbb{C}.$$

*Das Stabilitätsgebiet ist somit gegeben durch*

$$S = \left\{ z \in \mathbb{C} : \frac{1}{|1-z|} \leq 1 \right\} = \{ z \in \mathbb{C} : 1 \leq |1-z| \}.$$

*Es gilt also $\mathbb{C}^- \subset S$, wodurch die A-Stabilität des impliziten Euler-Verfahrens gezeigt wurde.*

- *Analog ergibt sich für die implizite Mittelpunktsregel die A-Stabilitätsfunktion*

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

*Wodurch auch die A-Stabilität gewährleistet ist.*

- *Für die $\theta$-Methode gilt*

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

*Somit ist die $\theta$-Method A-stabil für $\theta \geq \frac{1}{2}$.*

Aus der $A$-Stabilität folgt für dissipative Systeme der Form (2.20) die Kontraktion unabhängig von der Schrittweite $\tau_k > 0$, siehe Theorem 2.45. Somit ist nach Lemma 2.43 die Stabilität mit der Stabilitätskonstante $c_S = 1$ erfüllt. Ist ein Verfahren zusätzlich konsistent, so folgt nach Lemma 2.33 die Konvergenz.

### 2.5.3 Allgemeine dissipative Systeme

**Definition 2.53** (B-Stabilität)**.** *Ein Einschrittverfahren der Form (2.14) heißt* B-stabil*, falls für alle dissipativen Differentialgleichungen der Form*

$$\begin{aligned} \underline{u}'(t) &= \underline{f}(t, \underline{u}(t)) \qquad \text{für } t \in (0,T), \\ \underline{u}(0) &= \underline{u}_0, \end{aligned}$$

*die Kontraktivität erfüllt ist.*

**Lemma 2.54.** *Das implizite Euler-Verfahren ist B-stabil.*

**Beweis.** Für $\underline{u}, \underline{w} \in \mathbb{R}^n$ seien $\underline{u}_+, \underline{w}_+ \in \mathbb{R}^n$ die Lösungen von

$$\begin{aligned} \underline{u}_+ &= \underline{u} + \tau_k \underline{\phi}(t_k, \underline{u}_+, \tau_k) = \underline{u} + \tau_k \underline{f}(t_k, \underline{u}_+), \\ \underline{w}_+ &= \underline{w} + \tau_k \underline{f}(t_k, \underline{w}_+). \end{aligned}$$

Dann gilt

$$\begin{aligned} \left\| \underline{u}_+ - \underline{w}_+ \right\|^2 &= \left( \underline{u}_+ - \underline{w}_+, \underline{u}_+ - \underline{w}_+ \right) \\ &= \left( \underline{u} - \underline{w}, \underline{u}_+ - \underline{w}_+ \right) + \tau_k \left( \underline{f}(t_k, \underline{u}_+) - \underline{f}(t_k, \underline{w}_+), \underline{u}_+ - \underline{w}_+ \right) \\ &\leq \left( \underline{u} - \underline{w}, \underline{u}_+ - \underline{w}_+ \right) \leq \|\underline{u} - \underline{w}\| \left\| \underline{u}_+ - \underline{w}_+ \right\|. \end{aligned}$$

Und somit folgt die Kontraktivität

$$\left\| \underline{u}_+ - \underline{w}_+ \right\| \leq \|\underline{u} - \underline{w}\|.$$

■

## 2.6 Voll-diskretisierte parabolische Differentialgleichungen

Die Semi-Diskretisierung des parabolischen Modellproblems 1D führt auf das System von gewöhnlichen Differentialgleichungen: Gesucht ist $\underline{u}_h \in \mathcal{C}^1([0,T], R^n)$, sodass

$$\begin{aligned} \underline{u}_h'(t) &= M_h^{-1}\left[f_h(t) - K_h\underline{u}_h(t)\right] \qquad \text{für } t \in (0,T), \\ \underline{u}_h(0) &= M_h^{-1}\underline{g}_h \end{aligned} \tag{2.27}$$

erfüllt ist. Die Anwendung des impliziten Euler-Verfahrens auf die gewöhnliche Differentialgleichung (2.27) führt auf das Schema

$$\begin{aligned} &\underline{u}_{h,k+1} = M_h^{-1}\underline{g_h}, \\ &\underline{u}_{h,k+1} = \underline{u}_{h,k} + \tau_k M_h^{-1}\left[f_h(t_{k+1}) - K_h\underline{u}_{h,k+1}\right] \qquad \text{für } k = 0,\ldots,m-1. \end{aligned}$$

Somit muss in jedem Schritt ein lineares Gleichungssystem gelöst werden, im Speziellen

$$\left[M_h + \tau_k K_h\right]\underline{u}_{h,k+1} = M_h\underline{u}_{h,k} + \tau_k\underline{f}_h(t_{k+1}) \qquad \text{für } k = 0,\ldots,m-1.$$

Dabei kann das Gleichungssystem mit den in Kapitel **??** vorgestellten Methoden gelöst werden. Aus den berechneten Näherungslösungen $\underline{u}_{h,k}$, $k = 0,\ldots,m$ für das System (2.27) ergeben sich mit

$$u_{h,k}(x) = \sum_{j=1}^{n_h} \underline{u}_{h,k}[j]\varphi_j(x)$$

Näherungen für die Gitterpunkte $t_k \in I_\tau$. Es stellt sich nun die Frage, wie gut diese Näherungen sind, also wir wollen den Fehler

$$u(t_k) - u_{h,k} \qquad \text{für } t_k \in I_\tau$$

geeignet abschätzen. Sei $\underline{u}_h \in \mathcal{C}^2([0,T], R^n)$ die Lösung von (2.27) mit der entsprechenden Funktion $u_h \in \mathcal{C}^2([0,T], V_h)$. Dann gilt

$$\|u(t_k) - u_{h,k}\|_{L^2(0,1)} \le \|u(t_k) - u_h(t_k)\|_{L^2(0,1)} + \|u_h(t_k) - u_{h,k}\|_{L^2(0,1)}.$$

Der erste Term ist der Fehler der Semi-Diskretisierung, der mit Hilfe von Theorem 2.24 unter gewissen Regularitätsannahmen abgeschätzt werden kann durch

$$\|u(t_k) - u_{h,k}\|_{L^2(0,1)} \le ch^2\left[|u(t_k)|_{H^2(0,1)} + |u_0|_{H^2(0,1)} + \int_0^{t_k} |u'(s)|_{H^2(0,1)}\,\mathrm{d}s\right].$$

Für das impliziten Euler Verfahren gilt die Konsistenzabschätzung

$$\left\|\underline{\psi}_\tau(\underline{u}_\tau)\right\| \le \tau\int_0^T \|\underline{u}''(t)\|\mathrm{d}t$$

und somit folgt wegen der A-Stabilität, für den zweiten Term die Abschätzung

$$\|u_h(t_k) - u_{h,k}\|_{L^2(0,1)} = \left\|\underline{u}_h(t_k) - \underline{u}_{h,k}\right\|_{M_h}$$

$$\leq \tau \int_0^T \|\underline{u}_h''(t)\|_{M_h} \mathrm{d}t = \tau \int_0^T \|u_h''(t)\|_{L^2(0,1)} \mathrm{d}t.$$

Insgesamt gilt für $t_k \in I_\tau$ unter geeigneten Regularitätsannahmen die Abschätzung

$$\|u(t_k) - u_{h,k}\|_{L^2(0,1)} \leq ch^2 \left[ |u(t_k)|_{H^2(0,1)} + |u_0|_{H^2(0,1)} + \int_0^{t_k} |u'(s)|_{H^2(0,1)} \, \mathrm{d}s \right] + \tau \int_0^T \|u_h''(t)\|_{L^2(0,1)} \mathrm{d}t.$$

**Bemerkung 2.55.** *In der obigen Abschätzung ist noch die Funktion $u_h \in \mathcal{C}^2([0,T], V_h)$ enthalten. Dieser Ausdruck lässt sich noch weiter abschätzen, sodass sich der Fehler insgesamt wie*

$$\|u(t_k) - u_{h,k}\|_{L^2(0,1)} = \mathcal{O}(h^2 + \tau)$$

*verhält.*