NUMERIK III

Numerische Verfahren für Anfangs- und Anfangsrandwertaufgaben

Ulrich Langer
Institut für Mathematik
Johannes Kepler Universität Linz

Linz, Wintersemester 1994/95 & 1996/97

Vorwort

Das vorliegende Vorlesungsskriptum entstand aus Vorlesungen, die der Autor im Wintersemester 1994/95 und im Wintersemester 1996/97 an der Johannes Kepler Universität Linz gehalten hat. Die Lehrveranstaltung „Numerik III“ ist die dritte Vorlesung in einem Zyklus von drei Vorlesungen zur „Höheren Numerischen Mathematik“.
Die Vorlesung „Numerik I“ stellt das Handwerkszeug zur numerischen Behandlung linearer und nichtlinearer Operatorgleichungen in Banach- bzw. Hilbert–Räumen bereit und gibt eine Einführung in die Theorie moderner Funktionenräume (Sobolev-Räume, Räume von Distributionen) [16]. In der Vorlesung „Numerik II“ [17] werden Randwertaufgaben (RWA) für partielle Differentialgleichungen (PDgl.) und die wichtigsten Techniken (FEM, FDM, FVM) zu ihrer Diskretisierung betrachtet, sowie ein Überblick über moderne Verfahren zur Auflösung der bei der Diskretisierung entstehenden Gleichungssysteme gegeben (siehe auch Spezialvorlesung [15]). In der vorliegenden Vorlesung „Numerik III“ werden Anfangswertaufgaben (AWA) und Anfangsrandwertaufgaben (ARWA) für gewöhnliche und partielle Differentialgleichungen und die wichtigsten Methoden zu ihrer numerischen Lösung betrachtet. Zur Vorlesung gehört das Praktikum „Zeitintegrationsverfahren“ (siehe Anhang A). Im Praktikum werden Übungsaufgaben behandelt und ein praktisches Beispiel auf dem Rechner simuliert. Dieses praktische Beispiel wird durch ein Team von zwei Studenten in seiner Ganzheit (Modellierung, Analysis, numerische Analysis, Implementierung, Simulation, Ergebnisbewertung) bearbeitet. Die einzelnen Teams präsentieren ihre Ergebnisse in seminaristischer Form.

Die vorliegende Vorlesung setzt Kenntnisse aus den Grundvorlesungen zur linearen Algebra, zur Analysis und zur Numerischen Mathematik, sowie die in den Vorlesungen „Numerik I“ [16] und „Numerik II“ [17] vermittelten Lehrinhalte voraus. Zum anderen liefert die Lehrveranstaltung „Numerik III“ Vorkenntnisse für nachfolgende Spezialvorlesungen zur Numerischen und Angewandten Mathematik als auch für Spezialvorlesungen zur Industriemathematik.

Das Skriptum wurde bewußt im Stile eines Vorlesungsmanuskriptes gehalten. Im Gegensatz zu vielen Lehrbüchern wurde auf „belletristische“ Ausschmückungen verzichtet. Die Lehrinhalte sollen schnell und kompakt erfaßbar sein. Es wird eine Vielzahl von Abkürzungen eingeführt. Die Abkürzungen **Ü x.x** und **(mms)** bedeuten harte Arbeit an der Materie. Das Lösen von Übungsaufgaben und das „Mach-Mal-Selbst“ ist angesagt. Das vorliegende Skriptum ist ein Arbeitspapier, es ist kein Ersatz für den Vorlesungsbesuch und auch kein Ersatz für ein Lehrbuch, aber eine gute Vorbereitung auf die allfällige Prüfung.

Der Autor möchte an dieser Stelle Frau Doris Holzer für die Erstellung des LaTeX–Files und für die umfangreichen technischen Überarbeitungen recht herzlich danken.

Ulrich Langer

Linz, den 1. Februar 1997

Inhaltsverzeichnis

Kapitel 1

Einführung

■ **Ziel**

der Vorlesung NUMERIK III ist das Kennenlernen von Handwerkszeug zur Analysis und zur numerischen Behandlung von

- Anfangswertaufgaben (AWA) für Systeme gewöhnlicher Differentialgleichungen (Dgl.) der Art

$$
(1) \qquad \boxed{\begin{array}{l} \text{Ges. } u(t) = (u_1(t), \ldots, u_n(t))^T : \\ \text{Dgl.: } \dot{u}(t) = f\,(t, u(t))\,, \;\; t \in I = \bar{\boldsymbol{T}} = [0, T], \\ \text{AB: } u(0) = u_0 \text{ (Anfangsbedingung).} \end{array}}
$$

- Anfangsrandwertaufgaben (ARWA) für parabolische partielle Differentialgleichungen (PDgl.) der Art. (→ siehe [17] Numerik II, Pkt. 1.1):

$$
(2) \qquad \boxed{\begin{array}{l} \text{Ges. } u(x,t) : \\ \dfrac{\partial u}{\partial t} + \underbrace{L[u(x,t)]}_{\substack{\uparrow \\ \text{linearer [nichtlinearer]} \\ \text{elliptischer Operator}}} = f(x,t) \quad \forall (x,t) \in Q_T = \Omega \times \boldsymbol{T}, \\ + \text{ RB: 1. - 4. Art,} \\ + \text{ AB: } u(x,0) = u_0(x), \;\; x \in \bar{\Omega}. \end{array}}
$$

$$
(1) \xleftarrow[\text{vertikale Linienmethode}]{\text{Semidiskretisierung durch die}} (2)
$$

Die Semidiskretisierung von (2) durch die sogenannte (vertikale) Linienmethode

(= Ortsdiskretisierung durch Galerkin–Ansatz $u(x,t) = \sum_{i=1}^{n} u_i(t)p_i(x)$) führt unmittelbar auf (1) zur Bestimmung der zeitabhängigen Koeffizienten $u_i(t), \ldots, u_n(t)$.

- ARWA für hyperbolische PDgl. der Art ($\rightarrow$ siehe [17] Numerik II, Pkt. 1.2):

(3)
$$\boxed{\begin{array}{l} \text{Ges. } u(x,t): \\ \dfrac{\partial^2 u}{\partial t^2} + L\,u(x,t) = f(x,t) \quad \forall (x,t) \in Q_T \\ + \text{ RB: 1. - 4. Art} \\ + \text{ AB: } \left.\begin{array}{ll} u(x,0) & = u_0(x) \\ \dfrac{\partial u}{\partial t}(x,0) & = u_1(x) \end{array}\right\} x \in \bar{\Omega} \end{array}}$$

- Erhaltungsgleichungen (örtlich 1D):

$$\boxed{\begin{array}{l} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}[f(u)] = 0 \quad \text{in } Q_T = \mathbb{R}^1 \times \boldsymbol{T} \\ + \text{ AB: } u(x,0) = u_0(x) \quad \text{in } \mathbb{R}^1. \end{array}}$$

■ Inhalt:

- ARWA für parabolische PDgl. (Kapitel 2)
- AWA für (Systeme) gewöhnliche Dgl. (Kapitel 3)
- ARWA für hyperbolische PDgl. (Kapitel 4)

■ Wichtigste Literaturquellen:

- zu Kap. 2: [4], [20], [21].
- zu Kap. 3: [1], [5], [6], [19].
- zu Kap. 4: [4], [20].
- zu Erhaltungsgleichungen: [4], [13], [20].

■ **Praktikum:** zur Vorlesung Numerik III zum Thema „Zeitintegrationsverfahren“
$\longrightarrow$ siehe Anhang A.

Kapitel 2

Numerische Behandlung parabolischer Anfangsrandwertaufgaben

2.1 Gewichtete zweischichtige Differenzenschemata für die instationäre Wärmeleitgleichung

2.1.1 Konstruktion und Eigenschaften

- **Btr. der Einfachheit halber zunächst die örtlich 1D, instationäre Wärmeleitgleichung in differentieller Form** (siehe [17] Numerik II, Pkt. 1.1.2)
  $=$ **Ausgangsaufgabe** $\overset{\text{hier}}{=}$ **1. ARWA**:

$$\text{(1)}\qquad \boxed{\begin{array}{l} \text{Ges. } u(x,t): c\rho\dfrac{\partial u}{\partial t} - \dfrac{\partial}{\partial x}\left(\lambda \dfrac{\partial u}{\partial x}\right) + qu = f(x,t), \quad (x,t) \in Q_T = (a,b)\times(0,T) \\ \qquad\qquad + \text{ RB: } \left.\begin{array}{l} u(a,t) = g_a(t) \\ u(b,t) = g_b(t) \end{array}\right\} t \in \boldsymbol{T} = (0,T) \\ \qquad\qquad + \text{ AB: } u(x,0) = u_0(x), \quad x \in \bar{\Omega} = [a,b] \end{array}}$$

- **Btr. Spezialfall:** $c, \rho, \lambda = \text{const.} > 0, \quad q = 0.$
  - Mit $k^2 = \lambda/c\rho$ und $\bar{f} = f/c\rho$ folgt aus (1):

$$\boxed{\begin{array}{l} \dfrac{\partial u}{\partial t} - k^2\dfrac{\partial^2 u}{\partial x^2} = \bar{f}(x,t), \quad (x,t) \in Q_T \\ + \text{ RB} + \text{ AB} \end{array}}$$

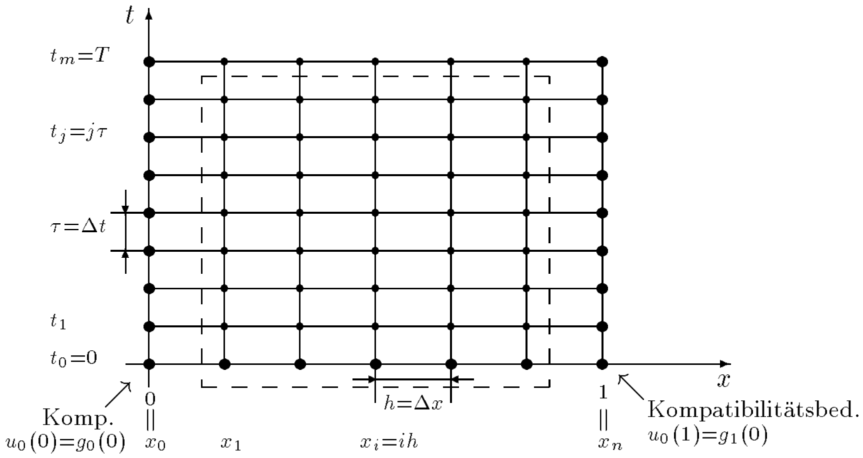
- O. B. d. Allg.: $k^2 = 1, \quad a = 0, \quad b = 1.$
  Tatsächlich:

$$l = b - a \qquad \left.\begin{aligned} \tilde{x} &= \frac{x-a}{b-a} = \frac{x-a}{l}, \quad \tilde{t} = \frac{k^2}{l^2}t \\ \frac{\partial}{\partial x} &= \frac{\partial}{\partial \tilde{x}}\frac{\partial \tilde{x}}{\partial x} = \frac{1}{l}\frac{\partial}{\partial \tilde{x}}, \quad \frac{\partial}{\partial t} = \frac{k^2}{l^2}\frac{\partial}{\partial \tilde{t}} \end{aligned}\right\} \Rightarrow \underline{\underline{\frac{\partial u}{\partial \tilde{t}} - \frac{\partial^2 u}{\partial \tilde{x}^2} = \frac{l^2}{k^2}\bar{f} =: \tilde{f}}}$$

■ **Nach eventueller Umbenennung:** $\tilde{x} \to x, \ \tilde{t} \to t, \ \tilde{f} \to f, \ \tilde{T} = \frac{k^2}{l^2}T \mapsto T$:

$$(1) \qquad \boxed{\begin{aligned} &\text{Ges. } u(x,t): \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x,t), \quad x \in (0,1), \ t \in \boldsymbol{T}, \\ &\qquad + \text{RB: } \left.\begin{aligned} u(0,t) = g_0(t) \\ u(1,t) = g_1(t) \end{aligned}\right\} \ t \in \boldsymbol{T} = (0,T), \\ &\qquad + \text{AB: } u(x,0) = u_0(x), \quad x \in [0,1]. \end{aligned}}$$

■ <u>**Gitterkonstruktion:**</u> Raum – Zeit – Gitter (vgl. [17] Numerik II, Pkt. 5.1.1)



z.B. gleichmäßig

$$\begin{aligned} \omega_{h\tau} &= \omega_h \times \omega_\tau, \\ \omega_h &= \{x_i = ih : i = \overline{1, n-1}\}, \quad h = 1/n, \\ \omega_\tau &= \{t_j = j\tau : j = \overline{1, m-1}\}, \quad \tau = T/m, \\ \bar{\omega}_\tau &= \{t_j = j\tau : j = \overline{0, m-1}\}. \end{aligned}$$

■ **Bez. von Gitterfunktion:**

$v : \bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau \longmapsto I\!R^1, \quad v_i^j = v\,(x_i, t_j);$

$v = v^j = v\,(\cdot, t_j) : \bar{\omega}_h \mapsto I\!R^1$: Gitterfkt. auf j-ter Zeitschicht;

$\quad v : \bar{\omega}_\tau \longmapsto [\bar{\omega}_h \mapsto I\!R^1]$ – abstrakte Gitterfkt.;

$\hat{v} = v^{j+1}; \;\; \check{v} = v^{j-1};$

$\bar{u} = u^{j+0.5} = u\,(\cdot, t_j + \frac{\tau}{2}) : \bar{\omega}_h \mapsto I\!R^1$, z.B. definiert für $u \in C(\bar{Q}_T)$.

■ **explizites Schema** **rein implizites Schema**

$v_t - v_{\bar{x}x} = \varphi\,(x,t)$ $\qquad \boxed{x \in \omega_h, \quad t \in \bar{\omega}_\tau}$ $\qquad v_t - \hat{v}_{\bar{x}x} = \varphi\,(x,t)$

$x = x_i = ih, \; i = \overline{1, n-1}; \qquad t = t_j = j\tau, \; j = \overline{0, m-1}$

- RB: $v_0^j = g_0(t_j), \;\; t_j = j\tau, \;\; j = \overline{1, m}$

  $v_n^j = g_1(t_j), \;\; t_j = j\tau, \;\; j = \overline{1, m}$
- AB: $v_i^0 = u_0(x_i), \;\; x_i = ih, \;\; i = \overline{0, n}$
- Wahl von φ: $\varphi = f, \;\; = \bar{f}, \;\; = (f + \hat{f})/2$?

$$\frac{v_i^{j+1} - v_i^j}{\tau} - \frac{v_{i+1}^j - 2v_i^j + v_{i-1}^j}{h^2} = \varphi_i^j$$

$i = 1, 2, \ldots, n-1$

$j = 0, 1, \ldots, m-1$

+ RB (•) + AB (•)

σ

$$\frac{v_i^{j+1} - v_i^j}{\tau} - \frac{v_{i+1}^{j+1} - 2v_i^{j+1} - v_{i-1}^{j+1}}{h^2} = \varphi_i^j$$

$i = 1, 2, \ldots, n-1$

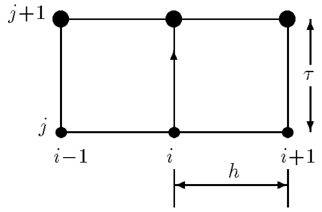
$j = 0, 1, \ldots, m-1$

+ RB (•) + AB (•)

σ-gewichtetes zweischichtiges DS
(=Einschrittverfahren)

■ **<u>Kompromiß:</u>** Familie von DS = σ-gewichtetes zweischichtiges DS

<u>im Sinne:</u>
- Approximation
- Stabilität
- Aufwand



Ges. $v : \bar{\omega}_{h\tau} \longmapsto \mathbb{R}$:

$$v_t - \sigma \hat{v}_{\bar{x}x} - (1-\sigma) v_{\bar{x}x} = \varphi(x,t), \quad (x,t) \in \overset{\mapsto}{\omega}_{h\tau} = \\ = \omega_h \times \overset{\vdash}{\omega}_\tau$$

\+ <u>RB:</u> $v_0^j = g_0(t_j), \quad v_n^j = g_1(t_j), \quad j = \overline{1,m}$

\+ <u>AB:</u> $v_i^0 = u_0(x_i), \quad i = \overline{0,n}$ (2)

$0 \le \sigma \le 1$ noch frei wählbares Gewicht

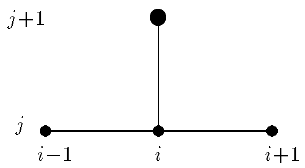
■ **<u>Bemerkung:</u>**

∗ Kompatibilitätsbedingung: $g_0(t_0) = u_0(x_0) = v_0^0$
$g_1(t_0) = u_0(x_n) = v_n^0$

∗ Im allgemeinen Fall (veränderliche Koeffizienten, 2D, 3D etc.) wird Ortsdiskretisierung mit <u>Integralbilanzmethode</u> ([17] Numerik II, Pkt. 5.2) bzw. <u>FEM–Linienmethode</u> (siehe Pkt. 2.3) vorgenommen.

■ **<u>Spezialfälle:</u>**

$\boxed{\sigma = 0}$ explizites Schema $\quad v_i^{j+1} = v_i^j + \tau \left[\varphi_i^j + \frac{1}{h^2} \left(v_{i-1}^j - 2v_i^j + v_{i+1}^j \right) \right]$

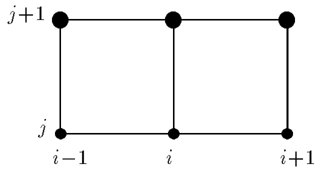


$i = \overline{1, n-1}$ $\qquad$ + RB $(i = 0, n)$

$j = 0, 1, \ldots, m-1$

AB

$\boxed{0 < \sigma \leq 1}$ implizites Schema $v_i^{j+1} - \tau\sigma v_{\bar{x}x,i}^{j+1} = v_i^j + \tau\left[\varphi_i^j + (1-\sigma)v_{\bar{x}x,i}^j\right]$

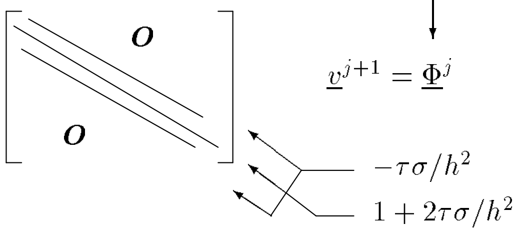


$i = \overline{1, n-1}$ + RB $(i = 0, n)$

$j = 0, 1, \ldots, m-1$

AB

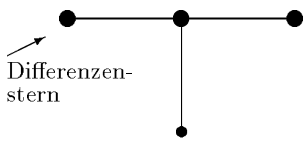
homog. der Dirichletbed.



$\underline{v}^{j+1} = \underline{\Phi}^j$

→ Direkte Auflösung mit LU-Faktorisierung (siehe [9], S. 71 ff) !

- $\sigma = 1/2$: CRANK–NICOLSON–Schema
- $\sigma = 2/3$: GALERKIN–Schema
- $\sigma = 1$: rein implizites Schema:



■ **<u>Wahl von σ:</u>** in Abhängigkeit von
(↓)
- h, τ
- Approximation (– ordnung)
- Stabilität
- Aufwand, insbes. bei örtl. mehrdim. Probl.

2.1.2 Lokale Approximationsordnung (Konsistenzordnung)

■ **<u>Fehlerschema:</u>** $z = \overset{(1)}{u} - \overset{(2)}{v} : \bar{\omega}_{h\tau} \longmapsto I\!R^1$:

(3)

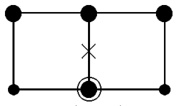
$$z_t - \sigma\hat{z}_{\bar{x}x} - (1-\sigma)z_{\bar{x}x} = (u_t - \sigma\hat{u}_{\bar{x}x} - (1-\sigma)u_{\bar{x}x}) - \varphi =: \psi(x,t)$$

$$(x,t) \in \overset{\mapsto}{\omega}_{h\tau} = \omega_h \times \overset{\vdash}{\omega}_\tau;$$

RB: $z_0^j = 0$, $z_n^j = 0$, $j = \overline{0, m}$;

AB: $z_i^0 = 0$, $i = \overline{0, n}$.

■ $\psi(x,t) \quad = \quad u_t - (\sigma\hat{u} + (1-\sigma)u)_{\bar{x}x} - \varphi =$
$(x,t) \in \bar{\omega}_{h\tau}$



$$= \quad u_t - \left(\underbrace{\frac{1}{2}(\hat{u}+u)}_{} + (\sigma - \frac{1}{2})\overbrace{\tau u_t}^{\hat{u}-u}\right)_{\bar{x}x} - \varphi =$$

$$\downarrow$$

$$= \bar{u} + O(\tau^2) \qquad u_t = \dot{\bar{u}} + O(\tau) = \dot{\bar{u}} + O(\tau^2) \qquad \text{(mms)}$$
$$\scriptstyle (0,2) \qquad\qquad\qquad (0,2) \qquad\qquad (0,3)$$

(x,t)
$\times$ Taylor

$\bar{u} = u\left(x, t + \frac{\tau}{2}\right)$

$$\boxed{\begin{array}{l} \underline{\text{Bez.:}}\ (k,l) \text{ bedeutet } u \in C^{k,l}(Q_T) \\ \dot{\bar{u}} = \overline{\left(\dfrac{\partial u}{\partial t}\right)} := \dfrac{\partial u}{\partial t}\left(x, t + \dfrac{\tau}{2}\right) \end{array}}$$

$$= \quad u_t - \underline{\bar{u}_{\bar{x}x}} - (\sigma - \tfrac{1}{2})\tau\dot{\bar{u}}_{\bar{x}x} + \underline{\left(O(\tau^2)\right)_{\bar{x}x}} - \varphi =$$
$$\downarrow$$
$$u_t = \dot{\bar{u}} + O(\tau^2) \qquad \scriptstyle (0,3)$$

<u>NR:</u> (mms)

$$\bar{u}_{\bar{x}x} = \frac{\partial^2 \bar{u}}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 \bar{u}}{\partial x^4} + O(h^4), \text{ falls } \bar{u}^{(6,0)} \quad \scriptstyle (6,0)$$
$$\left(O(\tau^2)\right)_{\bar{x}x} = O(\tau^2), \text{ falls } \left(O(\tau^2)\right)^{(2,2)}, \text{ d.h. } \bar{u}^{(2,2)} \quad \scriptstyle (2,2)$$

$$= \quad \frac{\partial \bar{u}}{\partial t} - \frac{\partial^2 \bar{u}}{\partial x^2} - \frac{h^2}{12}\frac{\partial^4 \bar{u}}{\partial x^4} - \left(\sigma - \frac{1}{2}\right)\tau\left(\frac{\partial \bar{u}}{\partial t}\right)_{\bar{x}x} - \varphi + O(\tau^2 + h^4) =$$

$$= \quad \underbrace{\frac{\partial \bar{u}}{\partial t} - \frac{\partial^2 \bar{u}}{\partial x^2}}_{\underset{\text{Dgl.}}{\overset{(1)}{=}}\ \bar{f}} - \frac{h^2}{12}\underbrace{\frac{\partial^4 \bar{u}}{\partial x^4}}_{} - \left(\sigma - \frac{1}{2}\right)\tau\left(\frac{\partial \bar{u}}{\partial t}\right)_{\bar{x}x} - \varphi + O(\tau^2 + h^4)$$

$$\frac{\partial^4 \bar{u}}{\partial x^4} = \frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 \bar{u}}{\partial x^2}\right) =$$
$$= \left(\frac{\partial^2 \bar{u}}{\partial x^2}\right)_{\bar{x}x} + O(h^2), \text{ falls } (\cdot)^{(4,0)}, \text{ d.h. } \scriptstyle (6,0)$$
$$\underset{\text{Dgl.}}{\overset{(1)}{=}} \left(\frac{\partial \bar{u}}{\partial t} - \bar{f}\right)_{\bar{x}x} + O(h^2)$$

$$= \quad \left[\bar{f} + \underbrace{\frac{h^2}{12}\bar{f}_{\bar{x}x}}_{= \frac{h^2}{12}\frac{\partial^2 \bar{f}}{\partial x^2} + O(h^4), \text{ falls } f \in C^{4,0}} - \varphi\right] - \left[\left(\sigma - \frac{1}{2}\right)\tau + \frac{h^2}{12}\right]\left(\frac{\partial \bar{u}}{\partial t}\right)_{\bar{x}x} + O(\tau^2 + h^4)$$

$$(\Rightarrow u \in C^{6,1})$$

■ **Lemma 2.1.:** Zusammenfassung der Approximationsresultate:

DS	Vor.	σ	φ	$\psi(x,t)$
Schema bester Approx.	$u \in C^{6,3}$	$\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}$	$\varphi = \bar{f} + \frac{h^2}{12}\bar{f}_{\bar{x}x}$ oder $= \bar{f} + \frac{h^2}{12}\frac{\partial^2 \bar{f}}{\partial x^2}$	$\psi = O(h^4 + \tau^2)$
CRANK–NICOLSON	$u \in C^{4,3}$	$\sigma = \frac{1}{2}$	$\varphi = \bar{f},\ = \frac{1}{2}(\hat{f} + f)$	$\psi = O(h^2 + \tau^2)$
	$u \in C^{4,2}$	$0 \leq \sigma \leq 1$	$\varphi = \bar{f},\ = f,\ = \hat{f}, \ldots$	$\psi = O(h^2 + \tau)$
explizites Schema	$u \in C^{4,2}$	$\sigma = 0$	$\varphi = \bar{f},\ = f,\ = \hat{f}, \ldots$	$\psi = O(h^2 + \tau)$

Beweis: folgt direkt aus der obigen Darstellung des Approximationsfehlers.

Bezeichnung:

Differenzierbarkeit bzgl. x
Differenzierbarkeit bzgl. t

$$C^{k,l} = C^{k,l}(Q_T)$$

q.e.d.

2.1.3 Zur Stabilitätsproblematik

■ **Btr. folgendes Schema mit homogenen RB:**

(4)

$$z_t - \sigma \hat{z}_{\bar{x}x} - (1-\sigma) z_{\bar{x}x} = \psi(x,t), \quad (x,t) \in \overset{\mapsto}{\omega}_{h\tau} = \omega_h \times \overset{\vdash}{\omega}_\tau;$$

RB: $z_0^j = 0, \quad z_n^j = 0, \quad j = 0, 1, \ldots, m;$

AB: $z_i^0 = w_i^0 \qquad , \quad i = \overline{1, n-1};$

Bemerkung 2.2.:

1. Im Fehlerschema (3): $\Longrightarrow$ AB: $z_i^0 = 0$ (homogen), aber dies setzt voraus, daß exakt gilt:
$$v_i^0 = u_0(x_i), \quad i \in \omega_h \qquad (x_i \in \omega_h).$$
praktisch: Rundungsfehler: $v_i^0 = u_0(x_i) - w_i^0, \quad i \in \omega_h$.
interessant: Untersuchungen der Fortpflanzung solcher Fehler in den AB:
$\Longrightarrow$ Stabilität bzgl. AB !

2. RB (1. Art) können o. B. d. Allg. im Fehlerschema homogen angenommen werden, denn durch Homogenisieren

$$\begin{aligned} u &= U + g_0(t) + x(g_1(t) - g_0(t)) \\ v &= V + \tilde{g}_0(t) + x(\tilde{g}_1(t) - \tilde{g}_0(t)) \end{aligned}$$

kann das Problem der Stabilität bzgl. RB auf die Stabilität bzgl. RS und Stabilität bzgl. AB zurückgeführt werden !

■ **Def. 2.3.:** (Stabilität bzgl. AB und RS)

Das DS (4) heißt stabil (bzgl. AB z^0 und RS ψ, sowie den gewählten Normen $\|\cdot\|_{(1)}$ und $\|\cdot\|_{(*,j)}$!), falls

(5) $\quad \|z^{j+1}\|_{(1)} \leq c_1\|z^0\|_{(1)} + c_2\|\Psi\|_{(*,j)} \quad \forall j = 0, 1, \ldots, m-1,$

wobei $c_1,\ c_2 = \text{const.} > 0 : c_\alpha \neq c_\alpha(h, \tau, j, \psi, z^0)$;

$\Psi = (\psi^0, \psi^1, \ldots, \psi^j)$,

$\|\psi\|_{(*,j)}$ – geeignet gewählte Norm:

z.B.: $\|\psi\|_{(*,j)} := \max\limits_{k=0,j} \|\psi^k\|_{(2)}$;

$\|\cdot\|_{(1)},\ \|\cdot\|_{(2)}$ – geeignet gewählte Normen auf den Zeitschichten, d.h. für Gitterfunktionen

$$v : \omega_h \longmapsto I\!R^1.$$

Spezialfälle: Stabilität bzgl. AB: $\psi = 0$ in (4) $\Rightarrow$ A-priori Absch. (5)
Stabilität bzgl. RS: $z^0 = 0$ in (4) $\Rightarrow$ A-priori Absch. (5)

■ **Bemerkung:** „geeignet gewählte“ Normen

- soll heißen, die Normen werden eigentlich definiert durch die vom Anwender gewünschten „Konvergenzaussagen“
- d.h. lok. Approximationsordnung $O(h^p + \tau^q)$

$? \quad \Downarrow$

globale Approximationsordnung $O(h^p + \tau^q)$,

$\oplus$ d.h. $\|\psi\|_{(*,j)} \leq c_{A,1}(u)h^p + c_{A,2}(u)\tau^q \leq c_A(u)(h^p + \tau^q)$

Stabilität Def. 2.3. $\longrightarrow \Downarrow$

diskrete Konvergenz: $\|z^{j+1}\|_{(1)} \leq c_1\|w^0\|_{(1)} + c_2 c_A(u)(h^p + \tau^q)$.

■ Wir untersuchen Stabilität in folgenden diskreten Normen:

- L_2–Norm: Fourier–Analyse (v. Neumann),
- Energienorm: Energetische Methode (siehe Pkt. 2.2),
- C–Norm: Diskretes Maximumprinzip.

■ Stabilität in diskreten $\mathbf{L_2}$ –Normen:

$\Rightarrow$ Fourier–Analyse nach den Eigenfunktionen des elliptischen Anteils:

- Bezeichnung:

  $y : \omega_h \longmapsto I\!R^1 \quad ; \quad y \in L_2(\omega_h) = \overset{\circ}{L}_2(\bar\omega_h) :$

  $\|y\|^2 = \|y\|^2_{L_2(\omega_h)} = (y,y)_h := \sum\limits_{i\in\omega_h} h\, y_i^2 = \sum\limits_{i=1}^{n-1} h\, y_i^2$

  $\|\cdot\|_{(1)} = \|\cdot\|_{(2)} := \|\cdot\| = \|\cdot\|_{L_2(\omega_h)}$ (vgl. Def. 2.3.)

- Btr. EWP (siehe auch $\boxed{\text{PI}}$)

$$\boxed{\begin{array}{l} -v_{\bar x x} = \lambda v(x), \quad x \in \omega_h \\ v_0 = v_n = 0 \end{array}} \qquad \Longrightarrow$$

EW: $\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi h}{2l}, \; k=\overline{1,n-1}$

EFkt.: $\mu_k(x) = \sqrt{\frac{2}{l}}\sin\frac{k\pi x}{l}, \; x\in\bar\omega_h, \; k=\overline{1,n-1}$

hier: $l=b-a=1$

Es gilt: $\{\mu_k\}_{k=\overline{1,n-1}}$ voll. ONS:

$$\Updownarrow$$

$$\boxed{\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & \boldsymbol{O} & \\ & \ddots & \ddots & \ddots & \\ & \boldsymbol{O} & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{bmatrix}}$$

- $(\mu_k,\mu_m)_h = \delta_{km}$
- Fourier–Zerlegung: $y(x) = \sum\limits_{k=1}^{n-1}(y,\mu_k)_h\mu_k(x)$
- PARSEVALsche Gleichung: $\|y\|^2 = \sum\limits_{n=1}^{n-1}(y,\mu_k)_h^2$

$$\approx \quad \boxed{\begin{array}{l} -u''(x) = \lambda u(x), \quad x \in (0,1) \\ u(0) = u(1) = 0 \end{array}}$$

- Entwickeln „Fehler“ $z = z(\cdot, t_j)$ auf der j-ten Zeitschicht in eine Fourierreihe nach den Eigenfkt. $\{\mu_k\}_{k=\overline{1,n-1}}$:

$$z = z(x,t_j) = \sum_{k=1}^{n-1} c_k(t_j)\mu_k(x), \quad x = x_i \in \bar\omega_h$$

Fourierkoeff.: $c_k = c_k(t_j) = (z(\cdot,t_j),\mu_k(\cdot))_h$

und analog die RS $\psi = \psi(x, t_j)$ des DS (4):

$$\psi = \psi(x, t_j) = \sum_{k=1}^{n-1} \underbrace{d_k}_{\text{Fourierkoeff.: } d_k = (\psi, \mu_k)_h = (\psi(\cdot, t_j), \mu_k(\cdot))_h}(t_j)\mu_k(x) = \sum_{n=1}^{n-1} d_k \mu_k, \quad x = x_i \in \omega_h.$$

- Btr. nun DS (4) mit diesen Fourierentwicklungen:

$$\frac{\hat{z} - z}{\tau} - \sigma \hat{z}_{\bar{x}x} - (1-\sigma) z_{\bar{x}x} = \psi \equiv \sum_{k=1}^{n-1} d_k \mu_k,$$

$$\hat{z} - \tau\sigma \hat{z}_{\bar{x}x} = z + \tau(1-\sigma) z_{\bar{x}x} + \tau \sum_{k=1}^{n-1} d_k \mu_k,$$

$$\sum_{k=1}^{n-1} (\hat{c}_k + \tau\sigma\lambda_k \hat{c}_k)\mu_k = \sum_{k=1}^{n-1} (c_k - \tau(1-\sigma)\lambda_k c_k)\mu_k + \tau \sum_{n=1}^{n-1} d_k \mu_k,$$

$$\sum_{k=1}^{n-1} [\hat{c}_k(1+\tau\sigma\lambda_k) - c_k(1-\tau(1-\sigma)\lambda_k) - \tau d_k]\mu_k = 0$$

$$\Longleftrightarrow \quad [\hat{c}_k(1+\tau\sigma\lambda_k) - c_k(1-\tau(1-\sigma)\lambda_k) - \tau d_k] = 0 \quad \forall k = 1, 2, \ldots, n-1$$

$$\Longleftrightarrow \quad \boxed{\hat{c}_k = \underbrace{\frac{1-\tau(1-\sigma)\lambda_k}{1+\tau\sigma\lambda_k}}_{= q_k} c_k + \tau \frac{d_k}{1+\tau\sigma\lambda_k} = q_k c_k + \tau \frac{d_k}{1+\tau\sigma\lambda_k}} \quad k = 1, 2, \ldots, n-1 \tag{6}$$

$$\Rightarrow \quad \hat{z} = \sum_{k=1}^{n-1} \hat{c}_k \mu_k = \sum_{k=1}^{n-1} q_k c_k \mu_k + \tau \sum_{k=1}^{n-1} \frac{d_k}{1+\sigma\tau\lambda_k} \mu_k.$$

- <u>Vor.:</u> $\boxed{|q_k| \le 1 \quad \forall k = 1, 2, \ldots, n-1}$ (7)

$$\|\hat{z}\| \underset{\text{PARSEVAL}}{=} \sqrt{\sum_{k=1}^{n-1} \hat{c}_k^2} \underset{(6)}{\le} \sqrt{\sum_{k=1}^{n-1} q_k^2 c_k^2} + \tau \sqrt{\sum_{k=1}^{n-1} \frac{d_k^2}{(1+\sigma\tau\lambda_k)^2}} \le$$

$$\underset{1+\sigma\tau\lambda_k \ge 1}{\overset{(7)}{\le}} \sqrt{\sum_{k=1}^{n-1} c_k^2} + \tau \sqrt{\sum_{n=1}^{n-1} d_k^2} \underset{\text{PARSEVAL}}{\le} \|z\| + \tau\|\psi\|.$$

- <u>Resultat:</u> $\|z^{j+1}\| \le \|z^j\| + \tau\|\psi^j\| \le$

$$
\begin{aligned}
&\le \|z^{j-1}\| + \tau\{\|\psi^{j-1}\| + \|\psi^j\|\} \le \ldots \le \\
&\le \|z^0\| + \tau \sum_{k=0}^{j} \|\psi^k\| \le \\
&\le \|z^0\| + \tau \underbrace{\sum_{k=0}^{j}}_{= (j+1)\tau = t_{j+1} \le T} \max_{l=\overline{0,j}} \|\psi^l\| \\
&\le \|z^0\| + T \underbrace{\max_{k=\overline{0,j}} \|\psi^k\|}_{=: \|\Psi\|_{(*,j)}}
\end{aligned}
$$

d.h. Stabilität bzgl. AB und RS
mit $\|\cdot\|_{(1)} = \|\cdot\|_{(2)} = \|\cdot\|, \quad c_1 = 1, \quad c_2 = T,$
falls $|q_k| \le 1 \quad \forall k = 1, 2, \ldots, n-1$
(vgl. Def. 2.3.).

- <u>Frage:</u> $|q_k| \le 1$? $\quad \forall k = 1, 2, \ldots, n-1$:

(a) $q_k = \dfrac{1 - \tau(1-\sigma)\lambda_k}{1 + \tau\sigma\lambda_k} < 1$ (gilt automatisch),

(b) $-1 \le q_k \iff 0 \le 1 + q_k = \dfrac{1 + \tau\sigma\lambda_k + 1 - \tau(1-\sigma)\lambda_k}{1 + \tau\sigma\lambda_k}$

$$\iff 2 - \tau\lambda_k + 2\tau\sigma\lambda_k \ge 0$$

$$\iff \boxed{\sigma \ge \frac{\tau\lambda_k - 2}{2\tau\lambda_k} = \frac{1}{2} - \frac{1}{\tau\lambda_k}, \quad \forall k = 1, 2, \ldots, n-1.}$$

- <u>Resultat:</u> $\boxed{\begin{array}{l} \sigma \ge \dfrac{1}{2} - \dfrac{1}{\tau\lambda_k} \\ k = 1, 2, \ldots, n-1 \end{array}} \iff \boxed{\begin{array}{c} |q_k| \le 1 \\ k = 1, 2, \ldots, n-1 \end{array}}$

- Aus $0 < \dfrac{8}{l^2} \underset{\substack{\uparrow \\ \boxed{\text{PI}}}}{\le} \lambda_1 < \lambda_2 < \ldots < \underbrace{\lambda_{n-1}}_{} = \lambda_{\max}\left(\dfrac{1}{h^2}\left[\ddots\right]\right) < \dfrac{4}{h^2}$ folgt:
$\lambda_{n-1} = \frac{4}{h^2}\sin^2\frac{(n-1)\pi h}{2}$

Lemma 2.4.: (L_2–Stabilität)

<u>Vor.:</u> $\sigma \geq \frac{1}{2} - \frac{1}{\tau\lambda_{\max}}$ (hier: $\boxed{\sigma \geq \frac{1}{2} - \frac{h^2}{4\tau}} > \frac{1}{2} - \frac{1}{\tau\lambda_{\max}}$).

<u>Bh.:</u> $\|z^{j+1}\| := \sqrt{\sum\limits_{i=1}^{n-1} h\left(z_i^{j+1}\right)^2} \leq \|z^0\| + T \max\limits_{k=0,j} \|\psi^k\|.$

- **Def. 2.5.:** (unbedingte und bedingte Stabilität)

  - DS heißt <u>unbedingt stabil</u>, wenn Stabilität vorliegt, unabhängig von den Beziehungen zwischen den Gitterparametern (h, τ).
  - Andernfalls sprechen wir von <u>bedingter Stabilität</u>, d.h. Stabilität unter Zusatzbedingungen an die Gitterparameter.

- **Bemerkung 2.6.:**

$$
\begin{aligned}
\sigma &\geq \frac{1}{2} && \geq \frac{1}{2} - \frac{h^2}{4\tau} =: \sigma_0 \Rightarrow \text{ unbedingt stabil !}\\
\sigma &= \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau} && \geq \frac{1}{2} - \frac{h^2}{4\tau} =: \sigma_0 \Rightarrow \text{ unbedingt stabil !}\\
\sigma &= 0 && \geq \frac{1}{2} - \frac{h^2}{4\tau} =: \sigma_0 \text{ , falls } \tau \leq \frac{h^2}{2} \Rightarrow \text{ bedingt stabil !}
\end{aligned}
$$

$$\downarrow$$

$$\frac{h^2}{4\tau} \geq \frac{1}{2}$$

<u>allgemein:</u>

$$0 \geq \frac{1}{2} - \frac{1}{\tau\lambda_{\max}} \Longleftrightarrow \frac{1}{\tau\lambda_{\max}} \geq \frac{1}{2} \Longleftrightarrow \boxed{\tau \leq \frac{2}{\lambda_{\max}}}$$

- **Bemerkung 2.7.:**

Die v. Neumannsche Fourier–Analyse nach den EFkt. des elliptischen Anteils ist immer dann anwendbar, wenn der elliptische Anteil <u>symmetrisch</u> und <u>zeit–unabhängig</u> ist (vgl. auch Pkt. 2.3.).
Aus der oben durchgeführten Analyse ist ersichtlich, daß sich eine Störung in den AB der Form

$$z^0 = w^0 := c\mu_k$$

mit dem Faktor q_k von Zeit– zu Zeitschicht fortpflanzt, d.h. im Falle einer homogenen RS ($\psi \equiv 0$) gilt:

$$z^{j+1} = q_k^{j+1} c\mu_k = q_k^{j+1} z^0.$$

Für eine formale Stabilitätsanalyse untersucht man deshalb die Fortpflanzung der harmonischen Störung

$$z_s^0 = e^{i\lambda sh} = \cos(\lambda sh) + i\sin(\lambda sh)$$

mit dem Ansatz

$$z_s^j = \left(e^{\alpha\tau}\right)^j e^{i\lambda sh}$$

und aus dem betr. DS (ohne RB !) zu bestimmenden α.
Das DS ist stabil, falls

$$\boxed{|e^{\alpha\tau}| \leq 1}\ .$$

Für das explizite Schema mit homogener RS und ohne Berücksichtigung der RB

$$z_s^{j+1} = (1-2\gamma) z_s^j + \gamma(z_{s-1}^j + z_{s+1}^j), \quad \gamma = \frac{\tau}{h^2},$$

erhalten wir z.B.

$$\left(e^{\alpha\tau}\right)^{j+1} e^{i\lambda sh} = (1-2\gamma)\left(e^{\alpha\tau}\right)^j e^{i\lambda sh} + \gamma\left(e^{\alpha\tau}\right)^j \left(e^{i\lambda(s-1)h} + e^{i\lambda(s+1)h}\right)$$

$$\begin{aligned} e^{\alpha\tau} &= (1-2\gamma) + \gamma(e^{-i\lambda h} + e^{i\lambda h}) = \\ &= 1 - 2\gamma + \gamma\left(\cos(-\lambda h) + i\sin(-\lambda h) + \cos(\lambda h) + i\sin(\lambda h)\right) \\ &= 1 - 2\gamma + 2\gamma\cos(\lambda h) \\ &= 1 - 2\gamma\underbrace{(1-\cos(\lambda h))}_{\substack{\| \\ 2\sin^2\frac{\lambda h}{2}}}. \end{aligned}$$

Die v. Neumannsche Stabilitätsbedingung $|e^{\alpha\tau}| \leq 1$ ist dann äquivalent zu

$$|e^{\alpha\tau}| = |1 - 2\gamma\underbrace{(1-\cos(\lambda h))}_{=2\sin^2\frac{\lambda h}{2}}| = |1 - 4\gamma\sin^2\frac{\lambda h}{2}| \leq 1, \text{ d.h.}$$

$$4\gamma\underbrace{\sin^2\frac{\lambda h}{2}}_{\leq 1} \leq 2.$$

Diese Bedingung ist für $\boxed{\gamma = \frac{\tau}{h^2} \leq \frac{1}{2}}$ erfüllt ($\forall\lambda$) !

Das allgemeinere v. Neumannsche Stabilitätskriterium

$$(*) \qquad \boxed{|e^{\alpha\tau}| \leq 1 + c\tau}$$

mit $c = \text{const.} > 0$, $\; c \neq c(\tau, h)$ läßt <u>exponentielles Wachstum</u> zu:

$$|(e^{\alpha\tau})^m| = \underset{\substack{\uparrow \\ \tau = \frac{T}{m}}}{|e^{\alpha m \tau}| \leq} \underbrace{\left(1 + \frac{cT}{m}\right)^{\frac{m}{cT}}}_{\leq e}{}^{\cdot cT} \underset{\uparrow}{\leq} e^{cT}$$

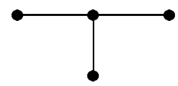
Das Kriterium $(*)$ ist <u>notwendig</u> und <u>hinreichend</u> dafür, daß $\exists$ positive Konstante $K = \text{const.} \geq 1$, $\; K \neq K(\tau)$: $|e^{\alpha\tau m}| \leq K$. $\quad (**)$

Verbleibt die Notwendigkeit zu zeigen:

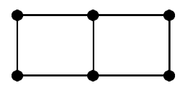
Aus $(**) \Rightarrow |e^{\alpha\tau}| \leq K^{1/m} \underset{\substack{\uparrow \\ m = T/\tau}}{=} K^{\frac{1}{T/\tau}} \leq K^{\frac{\tau}{T-\tau}} \underset{\substack{\uparrow \\ \text{Taylor} \\ \exists c = \text{const.} > 0 \\ \text{(mms)}}}{\leq} 1 + c\tau.$

- **Ü 2.1** Führen Sie die formale Stabilitätsanalyse nach v. Neumann für die folgenden DS (mit homogener RS) durch:

(a) rein implizites Schema,

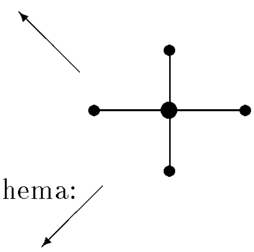


(b) CRANK–NICOLSON–Schema,



(c) Leapfrog–Schema (=Richardson–Schema)
$z_s^{j+1} - z_s^{j-1} - 2\gamma(z_{s-1}^j - 2z_s^j + z_{s+1}^j) = 0$; $\;\gamma = \tau/h^2$

(= Dreischichtiges DS = Zweischrittverfahren)



(d) duFort–Frankel–Schema (= modifiziertes Leapfrog–Schema: $2z_s^j \to z_s^{j+1} + z_s^{j-1}$):
$z_s^{j+1} - z_s^{j-1} - 2\gamma\left[z_{s-1}^j - (z_s^{j+1} + z_s^{j-1}) + z_{s+1}^j\right] = 0$

(= Dreischichtiges DS = Zweischrittverfahren).

Ü 2.2 Untersuchen Sie die lokale Approximationsordnung (= Konsistenzordnung) des Leapfrog–Schemas (c) und des duFort–Frankel–Schemas (d) für die homogene Wärmeleitgleichung

$$\boxed{\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad x \in (0,1), \quad t \in \boldsymbol{T} = (0,T), \\ &+ \text{RB: } u(0,t) = g_0(t), \quad u(1,t) = g_1(t), \quad t \in \boldsymbol{T}, \\ &+ \text{AB: } u(x,0) = u_0(x), \quad x \in [0,1]. \end{aligned}}$$

Bemerkung: Stabilitäts– und Konsistenzuntersuchungen von (c) und (d) siehe auch $\boxed{\text{P II}}$: Ü 5 – Ü 7.

■ Stabilität in diskreten C–Normen:

$\Longrightarrow$ Diskretes Maximumprinzip: $\Rightarrow$ [17] Numerik II, Pkt. 5.1.4 !

- Btr. DS (4) in der Form

$$\hat{z} - \tau\sigma\hat{z}_{\bar{x}x} = z + \tau(1-\sigma)z_{\bar{x}x} + \tau\psi =: \eta(x,t)$$

für $x = x_i \in \omega_h$ und $t = t_j \in \overset{\vdash}{\omega}_\tau$, d.h.

$$-\sigma\gamma z_{i-1}^{j+1} + (1+2\sigma\gamma)z_i^{j+1} - \sigma\gamma z_{i+1}^{j+1} = \eta_i^j, \tag{8}$$

mit $\gamma = \tau/h^2$ und

$$\eta_i^j = (1-\sigma)\gamma z_{i-1}^j + (1-2(1-\sigma)\gamma)\, z_i^j + (1-\sigma)\gamma z_{i+1}^j + \tau\psi_i^j. \tag{9}$$

- DS (8) ist offenbar <u>streng monoton</u> mit

$$D(x) := -\sigma\gamma\cdot 1 + (1+2\sigma\gamma)\cdot 1 - \sigma\gamma\cdot 1 \equiv 1.$$

Aus $\boxed{\text{Satz II.5.1}}$ aus [17] Numerik II folgt dann sofort:

$$\begin{aligned} \|z^{j+1}\|_{C(\omega_h)} &:= \max_{i=\overline{1,n-1}} |z_i^{j+1}| \leq \\ &(C(\bar{\omega}_h)) \quad (i = \overline{0,n}) \\ &\leq \frac{1}{\min\limits_{x\in\omega_h} |D(x)|}\, \|\eta^j\|_{C(\omega_h)} = \max_{i=\overline{1,n-1}} |\eta_i^j|. \end{aligned} \tag{10}$$

Setzt man

$$\boxed{1 - 2(1-\sigma)\gamma \geq 0}$$

voraus, so erhält man wegen $0 \leq \sigma \leq 1$ aus (9) und (10) die Abschätzung

$$\max_i |z_i^{j+1}| \leq \max_i |\eta_i^j| \leq \max_i |z_i^j| + \tau \max_i |\psi_i^j|,$$

deren rekursive Auswertung die gewünschte Stabilitätsabschätzung ergibt:

$$\begin{aligned}\|z^{j+1}\|_{C(\omega_h)} &\leq \|z^0\|_{C(\omega_h)} + \tau \sum_{k=0}^{j} \|\psi^k\|_{C(\omega_h)} \\ &\leq \|z^0\|_{C(\omega_h)} + T \underbrace{\max_{k=\overline{0,j}} \|\psi^k\|_{C(\omega_h)}}_{=: \|\psi\|_{(*,j)}}\end{aligned}$$

- **Lemma 2.8.:** (C–Stabilität)

> Vor.: $1 - 2\,(1-\sigma)\,\gamma \geq 0, \quad \gamma = \tau/h^2, \quad 0 \leq \sigma \leq 1$
> Bh.: $\|z^{j+1}\|_{C(\omega_h)} := \max_{i=\overline{1,n-1}} |z_i^{j+1}| \leq \|z^0\|_{C(\omega_h)} + T \max_{k=\overline{0,1}} \|\psi^k\|_{C(\omega_h)}$

- **Bemerkung 2.9.:**

$\sigma = 1 \;\Rightarrow$ rein implizites Schema ist unbedingt stabil in der diskreten C–Norm;

$\sigma = \frac{1}{2} \;\Rightarrow$ C–Stabilität für $\tau \leq h^2$, d.h. bedingt stabil;

$\sigma = 0 \;\Rightarrow$ C–Stabilität für $\tau \leq \frac{h^2}{2}$, d.h. bedingt stabil.

Stabilitätsuntersuchungen in der diskreten C–Norm sind auch für allgemeinere parabolische ARWA mit variablen Koeffizienten vom Typ

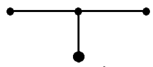
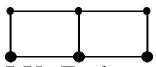
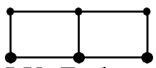
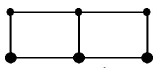
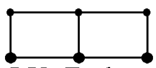
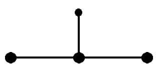
$$cg\frac{\partial u}{\partial t} - \left[\frac{\partial}{\partial x}\left(\lambda(x,t)\frac{\partial u}{\partial x}\right) + v(x,t)\frac{\partial u}{\partial t} + q(x,t)u\right] = f(x,t)$$

möglich und relativ einfach.

2.1.4 Zusammenfassung: Approximation + Stabilität $\Longrightarrow$ diskrete Konvergenz

- **Konvergenz in diskreter $\mathbf{L_2}$ –Norm:**

Satz 2.10.: $\|z\| = \|z\|_{L_2(\omega_h)} := \left(\sum\limits_{i=1}^{n-1} h\, z_i^2\right)^{1/2}$

Vor.	$0 \leq \sigma \leq 1$	Differenzenstern	Stabilitätsbedingung
		Auflösung	$\sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau}$
$u \in C^{4,2}$	$\sigma = 1$ rein impl. DS	 LU–Zerlegung	unbedingt stabil
$u \in C^{4,3}$	$\sigma = 1/2$ CRANK–NICOLSON	 LU–Zerlegung	unbedingt stabil
$u \in C^{6,3}$	$\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}$ Schema bester Approximation	 LU–Zerlegung	unbedingt stabil
$u \in C^{4,2}$	$1 \overset{(>)}{\geq} \sigma \overset{(>)}{\geq} \frac{1}{2}$ fix	 LU–Zerlegung	unbedingt stabil
$u \in C^{4,2}$	$\frac{1}{2} > \sigma > 0$	 LU–Zerlegung	bedingt stabil, d.h. h, τ: $\sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau}$
$u \in C^{4,2}$	$\sigma = 0$ explizites Schema		bedingt stabil: $\tau \leq \frac{h^2}{2}$

Vor.	rechte Seite φ	Konsistenz = lokaler Approximationsfehler ψ	diskrete Konvergenz $\max_{j=\overline{0,m}} \| \overset{(1)}{u^j} - \overset{(2)}{v^j} \| \le T \max_{k=\overline{0,m-1}} \|\psi^k\| \le$ $\le T\, c_A(u)\, (h^p + \tau^q)$
$u \in C^{4,2}$	$\varphi = \bar{f}$ $\varphi = f, \hat{f}$	$O(h^2 + \tau)$	$p = 2, \quad q = 1$
$u \in C^{4,3}$	$\varphi = \bar{f}$ $\varphi = \frac{1}{2}(\hat{f} + f)$	$O(h^2 + \tau^2)$	$p = 2, \quad q = 2$
$u \in C^{6,3}$	$\varphi = \bar{f} + \frac{h^2}{12} \bar{f}_{\bar{x}x}$ $\varphi = \bar{f} + \frac{h^2}{12} \overline{\frac{\partial^2 f}{\partial x^2}}$	$O(h^4 + \tau^2)$	$p = 4, \quad q = 2$
$u \in C^{4,2}$	$\varphi = \bar{f}$ $\varphi = f, \hat{f}$	$O(h^2 + \tau)$	$p = 2, \quad q = 1$
$u \in C^{4,2}$	$\varphi = \bar{f}$ $\varphi = f, \hat{f}$	$O(h^2 + \tau)$	$p = 2, \quad q = 1$
$u \in C^{4,2}$	$\varphi = \bar{f}$ $\varphi = f, \hat{f}$	$O(h^2 + \tau)$	$p = 2, \quad q = 1$

■ Konvergenz in der diskreten C–Norm:

- **Satz 2.11:** $\|z\| = \|z\|_{C(\omega_h)} := \max_{i=\overline{1,n-1}} |z_i|$

> Vor.: 1) $0 \le \sigma \le 1, \qquad (1-\sigma)\dfrac{\tau}{h^2} \le \dfrac{1}{2}$,
>
> 2) $u \in C^{4,2}(\bar{Q}_T)$.
>
> Bh.: $\max_{j=\overline{0,m}} \|u(\cdot, t^j) - v^j\|_{C(\bar{\omega}_h)} := \max_{j=\overline{0,m}} \max_{i=\overline{0,n}} |u(x_i, t_j) - v_i^j| \le$
>
> (1) (2)
>
> $\le T\, c_A(u)\, (h^2 + \tau)$.

- <u>**Satz 2.12:**</u>

> - Für das Crank–Nicolson–Schema ($\sigma = 1/2$) gilt speziell
> $$\max_{j=\overline{0,m}} \|\underset{(1)}{u(\cdot,t^j)} - \underset{(2)}{v^j}\|_{C(\bar{\omega}_h)} \leq T\, c_A(u)\, (h^2 + \tau^2),$$
> falls $\tau/h^2 \leq 1$ (Stabilitätsbed.) und $u \in C^{4,3}(\bar{Q}_T)$.
> - Für das Schema bester Approximation ($\sigma = \sigma_* \equiv \frac{1}{2} - \frac{h^2}{12\tau}$ mit $\sigma \in [0,1]$) gilt speziell
> $$\max_{j=\overline{0,m}} \|\underset{(1)}{u(\cdot,t^j)} - \underset{(2)}{v^j}\|_{C(\bar{\omega}_h)} \leq T\, c_A(u)\, (h^4 + \tau^2),$$
> falls $\tau/h^2 \leq 5/6$ (Stabilitätsbed.) und $u \in C^{6,3}(\bar{Q}_T)$.

2.2 Eine allgemeine Stabilitätstheorie für zweischichtige Schemata in der Energienorm (Energetische Methode)

2.2.1 Allgemeine und kanonische Form

■ Für **parabolische ARWA** (siehe [17] Numerik II, Pkt. 1.1) der Art

(11)

> $$\frac{\partial u}{\partial t} + L\, u(x,t) = f(x,t), \quad \forall (x,t) \in Q_T = \Omega \times \boldsymbol{T}$$
> + RB + AB, mit $\Omega \subset \mathbb{R}^d$ beschr. Gebiet, $\boldsymbol{T}= (0,T)$ und mit dem formal s.a. elliptischen Differentialausdruck
> $$L\, u := -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_j(x) \frac{\partial u}{\partial x_j}\right) + a(x)\, u(x,t)$$

sei ein zweischichtiges $\left(\frac{\partial u}{\partial t}\right)$ DS aufgeschrieben, dessen <u>allgemeine Form</u> offenbar folgendermaßen aussieht:

$$(12)\qquad \boxed{\begin{array}{ll} C_1\, v^{j+1} + C_0\, v^j = \tau \varphi^j(x), & j = 0,1,\ldots,m-1, \\ \text{AB: } v^0 \text{ geg.,} & x \in \omega = \omega_h \end{array}}$$

wobei $v = v^j : \bar{\omega} \longmapsto I\!R^1$ – Gitterfkt. auf j-ter Zeitschicht

$(v|_{\gamma=\gamma_1=\bar{\omega}\setminus\omega} = 0\ !)$;

$H_h^{(1)}$, $\|\cdot\|_{(1)}$, $(\cdot,\cdot)_{(1)}$ – diskreter H–Raum;

$\varphi = \varphi^j : \omega \longmapsto I\!R^1$ – rechte Seite auf j-ter Zeitschicht;

$H_h^{(2)}$, $\|\cdot\|_{(2)}$, $(\cdot,\cdot)_{(2)}$ – diskreter H–Raum;

$C_0, C_1 : H_h^{(1)} \longmapsto H_h^{(2)}$ – lineare Operatoren ($\leftarrow \overset{\text{Pkt. 2.1}}{\text{FDM}}$, $\overset{\text{Pkt. 2.3}}{\text{FEM}}$, ...);

RB – sind eingearbeitet: • RB 1. Art seien homogenisiert $\to$ RS,

• Nat. RB: $x \in \gamma_N = \gamma_2 \cup \gamma_3 \Rightarrow \omega = \overset{\circ}{\omega} \cup \gamma_N$;

τ – Zeitschritt, h – örtlicher Diskretisierungsparameter, ω – Gitter für Ω,
$\gamma = \gamma_1 = \gamma_D = \bar{\omega} \setminus \omega$ – diskreter Γ_1–Rand (Dirichlet–Rand).

■ **Kanonische Form:**

$$(13)\qquad \boxed{B\, v_t^j + A\, v^j = \varphi^j, \quad j = \overline{0, m-1};\ \ v^0 \text{ geg.}}$$

■ **Beziehung zwischen kanonischer und allgemeiner Form:**

$(13) \Longleftrightarrow B\, v^{j+1} - B\, v^j + \tau\, A\, v^j = \tau\, \varphi^j$

$$(14)\qquad \boxed{C_1 = B,\ C_0 = \tau\, A - B \text{ bzw. } B = C_1,\ A = \frac{1}{\tau}(C_0 + C_1)}$$

■ **Beispiel:** σ–gewichtetes Schema aus Pkt. 2.1:

$$\left\{ \begin{array}{l} v_t^j - \sigma v_{\bar{x}x}^{j+1} - (1-\sigma) v_{\bar{x}x}^j = \varphi^j(x), \quad x \in \omega_h \\ +\ \text{RB (hom. 1. Art: } g_0 = g_1 = 0)\ +\ \text{AB} \end{array} \right\}$$

$\bar{A} : \bar{A}v := -v_{\bar{x}x}$, $H_h^{(1)} = \overset{\circ}{L}_2(\bar{\omega}_h) = L_2(\omega_h) = H_h^{(2)} = H_h$
$\Rightarrow v^{j+1} - v^j + \tau\sigma\bar{A}v^{j+1} + \tau(1-\sigma)\bar{A}v^j = \tau\varphi^j$

$\Rightarrow$ allgemeine Form: $\underbrace{(I+\tau\sigma\bar{A})}_{=:\,C_1} v^{j+1} + \underbrace{(\tau(1-\sigma)\bar{A}-I)}_{=:\,C_0} v^j = \tau\varphi^j$

$\Rightarrow$ Kanonische Form: $B = C_1 = I + \tau\sigma\bar{A}$

$$A = \frac{1}{\tau}(C_0 + C_1) = \frac{1}{\tau}\left(\tau(1-\sigma)\bar{A} - I + I + \tau\sigma\bar{A}\right) = \bar{A}$$

- **Def. 2.13:** ($\hat{=}$ Def. 2.3): $(H_h^{(1)};\ H_h^{(1)},\ H_h^{(2)})$ – Stabilität.

> Das zweischichtige DS (12) = (13) heißt stabil (bzgl. AB v^0 und RS φ, sowie den gewählten Normen $\|\cdot\|_{(1)}$ und $\|\cdot\|_{(2)}$), wenn für evtl. hinreichend kleinen h und τ und $\forall j = 0, 1, \ldots, m-1$ gilt:
>
> (15) $$\|v^{j+1}\|_{(1)} \leq c_1\, \|v^0\|_{(1)} + c_2 \max_{0\leq k\leq j} \|\varphi^k\|_{(2)}$$
>
> mit $c_1, c_2 = \text{const.} > 0$, $c_\alpha \neq c_\alpha(h, \tau, j, \varphi, v^0)$, $\|\cdot\|_{(1)} = \|\cdot\|_{H_h^{(1)}}$, $\|\cdot\|_{(2)} = \|\cdot\|_{H_h^{(2)}}$ – gew. Normen.
>
> Spezialfälle: Stabilität bzgl. AB: $\varphi = 0$ in (12): A–priori–Absch. (15),
> Stabilität bzgl. RS: $v^0 = 0$ in (12): A–priori–Absch. (15).

2.2.2 Stabilität in der energetischen Norm $\|\cdot\|_A$

- **Standardvoraussetzungen an (13):**

H – (diskreter) reeller Hilbert–Raum ($\rightarrow$ Grundraum), i. S. Raum von Gitterfkt.: $\|\cdot\|$, $(\cdot,\cdot)$;

(1) $A, B : H \longmapsto H$ – lineare Operatoren ($\leftarrow$ L–linear !).

(2) $A = A^* > 0$ spd ($\leftarrow$ L formal s.a., ellipt. Operator !); d.h. $(A\,v, w) = (v, A\,w)$, $\forall v, w \in H$ und $(A\,v, v) > 0$ $\forall v \in H$, $v \neq \boldsymbol{O}$.

(3) $B > 0$ ($\Longrightarrow \exists\, B^{-1} \Rightarrow$ Schema eindeutig auflösbar !).

(4) A, B – seien nicht von t_j abhängig (d.h. Koeff. von L sind zeitunabhängig !).

- **Btr. Fehlerschema:** $z \overset{(11)}{=} u - \overset{(12)}{v}$

(16)
> $B\,z_t + A\,z = \psi$, ψ – Approximationsfehler;
> AB: z^0 geg.

$z = y + w$: (17) $By_t + Ay = \psi$; $y^0 = 0$: Ziel: Stabilität bzgl. RS

RS AB: (18) $Bw_t + Aw = 0$; $w^0 = z^0$: Ziel: Stabilität bzgl. AB

$+ \Longrightarrow$ (16): Stabilität bzgl. AB und RS !

- **Die „energetische" Identität:** (im parabolischen (zweischichtigen) Falle)

((16), $2\tau z_t$): $2\tau(Bz_t, z_t) + 2\tau(A\,\boxed{z}\,, \boxed{z_t}) = 2\tau(\psi, z_t)$

$$\boxed{z = \frac{\hat z + z}{2} - \frac{\hat z - z}{2} = \frac{1}{2}(\hat z + z) - \frac{\tau}{2} z_t} \qquad \boxed{z_t = \frac{\hat z - z}{\tau}}$$

$$2\tau(Bz_t, z_t) + \underbrace{\tau\left(A(\hat z + z), \frac{\hat z - z}{\tau}\right)}_{= (A(\hat z + z), \hat z - z) = (A\hat z, \hat z) + \underbrace{(Az, \hat z) - (A\hat z, z)}_{= 0, \text{ da } A = A^*} - (Az, z)} - 2\tau\left(\frac{\tau}{2}Az_t, z_t\right) = 2\tau(\psi, z_t)$$

Resultat: Energetische Identität

$$\boxed{2\tau\left(\left(B - \frac{\tau}{2}A\right) z_t, z_t\right) + (A\hat z, \hat z) - (Az, z) = 2\tau(\psi, z_t)} \tag{19}$$

- **Lemma 2.14.:** (Stabilität bzgl. AB)

> Vor.: 1. Standardvoraussetzungen (1) – (4).
> 2. $B \geq \frac{\tau}{2}A$, i.S.: $(Bv, v) \geq \frac{\tau}{2}(Av, v) \quad \forall v \in H$.
>
> Bh.: Dann ist DS (16) stabil bzgl. AB (d.h. DS (18)) mit $c_1 = 1$ im Raum H_A (= zu H energetischer Raum):
>
> (20) $\quad \|w^{j+1}\|_A := (Aw^{j+1}, w^{j+1})^{0.5} \leq \|w^0\|_A$.

Beweis: $\psi = 0$, d.h. DS (18) ! Für (18) folgt aus (19)

$\Rightarrow (A\widehat{w}, \widehat{w}) = (Aw, w) - 2\tau \underbrace{\left(\left(B - \frac{\tau}{2}A\right) w_t, w_t\right)}_{\geq 0} \leq (Aw, w)$.

$\Rightarrow \|\widehat{w}\|_A \leq \|w\|_A$. $\nearrow$

q.e.d.

- **Beispiel:** σ–gewichtetes DS aus Pkt. 2.1.:
  $B = I + \tau\sigma\bar A$, $A = \bar A$, $\bar A v := -v_{\bar x x}$;
  $H = \overset{\circ}{L}_2(\bar\omega) = L_2(\omega)$.

  - Voraussetzungen (1) – (4) offenbar erfüllt, sogar $B = B^*$ (mms)
  - Allgemeine Stabilitätsbedingung: $B = I + \tau\sigma\bar A \geq \frac{\tau}{2}\bar A$!

$$\left.\begin{cases} \sigma \geq \frac{1}{2} & \text{: offenbar erfüllt} \\ \sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau} & \text{: ebenfalls erfüllt } \boxed{\text{Ü 2.3}} \end{cases}\right\} \Longrightarrow \boxed{\|w^{j+1}\|_A \leq \|w^0\|_A}$$

$$\boxed{\text{Ü 2.3}}\ \| >$$

$$\boxed{\|w_{\bar{x}}^{j+1}|] \leq \|w_{\bar{x}}^0|]}$$

- $\boxed{\textbf{Ü 2.3}}$ Man zeige für das gew. DS mit homog. RS und mit $1 \geq \sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau}$ die A–priori–Abschätzungen:

  (a) $\|w_{\bar{x}}^{j+1}|] := \left(\sum\limits_{i=1}^{n} h\left(w_{\bar{x},i}^{j+1}\right)^2\right)^{1/2} \leq \|w_{\bar{x}}^0|]$,

  (b) $\|w^{j+1}\|_{C(\omega_h)} \leq \frac{1}{2}\|w_{\bar{x}}^0|]$.

■ **Lemma 2.15.:** (Stabilität bzgl. RS)

> Aus der Stabilität bzgl. der AB gemäß Lemma 2.14. folgt die Stabilität bzgl. der RS ($\rightarrow$ DS (17)):
>
> (21) $\quad \| y^{j+1} \|_A \ \leq\ T \max\limits_{0 \leq k \leq j} \|B^{-1}\, \psi^k\|_A$

vgl. Def. 2.13.: $\|y^{j+1}\|_{(1)} \qquad c_2 \qquad \|\psi^k\|_{(2)} = \|\psi^k\|_{\underbrace{B^{-T} A\, B^{-1}}_{\text{spd}}}$

Beweis:

- Btr. DS (17): $By_t + Ay = \psi$, $y^0 = 0$
  $\tau B^{-1}\cdot$ (17): $\Rightarrow \hat{y} = y - \tau B^{-1}Ay + \tau B^{-1}\psi = (I - \tau B^{-1}A)y + \tau B^{-1}\psi$
- Vor. (2): $A = A^* > 0 \Rightarrow \exists\, A^{1/2}$; Setzen: $v = A^{1/2}y$, $y = A^{-1/2}v$:

$$\Rightarrow \hat{v} = \underbrace{(I - \tau A^{1/2}B^{-1}A^{1/2})}_{=:\ \boldsymbol{S}\ (\text{Übergangsoperator})} v + \tau \underbrace{A^{1/2}B^{-1}\psi}_{=:\ f}$$

$$\boxed{\hat{v} = \boldsymbol{S}\ v + \tau f}$$

- Schätzen nun $\|\ \boldsymbol{S}\ \|$ ab: $\uparrow$
  Dazu setzen wir $\psi = 0 \Rightarrow f = 0\ :\ \hat{v} = \boldsymbol{S}\, v$

  Lemma 2.14. (Beweis): $\|\hat{y}\|_A \leq \|y\|_A$

$$\Updownarrow \leftarrow y = A^{-1/2}v$$

$$\|\hat{v}\| \leq \|v\|$$

$$\Updownarrow$$

$$\|\ \boldsymbol{S}\, v\| \leq \|v\| \quad \underline{\underline{\forall v \in H}}$$

Resultat: $\boxed{\| \boldsymbol{S} \| \leq 1}$

Bem.: $\| \boldsymbol{S} \| = \|I - \tau B^{-1} A\|_A$ (mms)

$$\underbrace{\hat{v} = \boldsymbol{S}\, v + \tau f}_{\downarrow} \qquad\qquad \underbrace{\| \boldsymbol{S} \| \leq 1}_{\downarrow}$$

- $\|\hat{v}\| \leq \| \boldsymbol{S}\, v\| + \tau \|f\| \leq \| \boldsymbol{S} \|\, \|v\| + \tau\|f\| \leq \|v\| + \tau\|f\|$

$$\Longrightarrow \|v^{j+1}\| \leq \|v^j\| + \tau\|f^j\| \leq \ldots \leq \|v^0\| + \tau \sum_{k=0}^{j} \|f^k\|$$

$$\Longleftrightarrow \|y^{j+1}\|_A \leq \underbrace{\|y^0\|_A}_{=0} + \tau \sum_{k=0}^{j} \|B^{-1}\psi^k\|_A \qquad (22)$$

$$\leq \max_{0 \leq k \leq j} \|B^{-1}\psi^k\|_A \cdot \underbrace{\tau \sum_{n=0}^{j} 1}_{=(j+1)\tau \leq T} \leq T \max_{0 \leq k \leq j} \|B^{-1}\psi^k\|_A .$$

q.e.d.

■ **Satz 2.16.** (Stabilität bzgl. AB und RS)

Vor.: 1. Standardvoraussetzungen (1) – (4).
2. $B \geq \frac{\tau}{2} A$.

Bh.: Das DS (16): $Bz_t + Az = \psi$, z^0 – geg., ist stabil bzgl. AB und RS im energetischen Raum H_A (genauer: H_A; H_A, $H_{B^{-T}AB^{-1}}$):

(23) $$\|z^{j+1}\|_A \leq \|z^0\|_A + T \max_{0 \leq k \leq j} \|\psi^k\|_{B^{-T}AB^{-1}} .$$

Beweis:

- folgt sofort aus Beweis von Lemma 2.15., Abschätzung (22);
- bzw. aus $z = w + y$ Lemma 2.14. Lemma 2.15.

$$\|z^{j+1}\|_A \leq \|w^{j+1}\|_A + \|y^{j+1}\|_A \leq \ldots$$

q.e.d.

■ **Beispiel:** σ–gewichtetes DS aus Pkt. 2.1.:

$$B = I + \tau\sigma\bar{A}, \quad A = \bar{A}, \quad \bar{A}v = -v_{\bar{x}x}, \quad H = \overset{\circ}{L}_2(\bar{\omega}) = L_2(\omega) = H_h$$

vertauschbar

$$\|B^{-1}\psi\|_A = \sup_{\substack{v\in H\\ v\neq \boldsymbol{O}}} \frac{((I+\sigma\tau A)^{-1}\psi, v)_A}{\|v\|_A} = \sup_{v\in H} \frac{(A(I+\sigma\tau A)^{-1}\psi, v)}{\|v\|_A} =$$

$\uparrow$

$B = (I + \sigma\tau A)$ ist s.a. und p.d. bzgl. Skalarprodukte $\underline{\underline{(\cdot,\cdot)_A}}$ und $(\cdot,\cdot)$

$$= \sup_{v\in H} \frac{((I+\sigma\tau A)^{-1}A^{0.5}\psi, A^{0.5}v)}{\|v\|_A} \leq \underbrace{\|(I+\sigma\tau A)^{-1}\|}_{=\frac{1}{1+\sigma\tau\lambda_{\min}(A)}} \|\psi\|_A \leq 1\cdot\|\psi\|_A,$$

$\uparrow$ Satz v. Banach

d.h. $(H_A; H_A, H_A)$–Stabilität: $\boxed{\|z_{\bar{x}}^{j+1}|] \leq \|z_{\bar{x}}^0|] + T \max_{0\leq k\leq j} \|\psi_{\bar{x}}^k|]}$

■ **Ü 2.4** Man zeige, daß das σ–gewichtete DS aus Pkt. 2.1 im energetischen Raum $H_{\bar{A}} = A$ stabil ist und daß die a-priori–Abschätzung

$$\|z_{\bar{x}}^{j+1}|] \leq \|z_{\bar{x}}^0|] + T \max_{0\leq k\leq j} \|\psi_{\bar{x}}^k|]$$

(für das Fehlerschema) gilt, falls $1 \geq \sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau}$.

Hinweis: 1) Satz 2.16. anwenden.

2) $\|z\|_{\bar{A}} = \|z_{\bar{x}}|]$ (Formel der partiellen Summation).

3) Satz von Banach $\|(I+K)^{-1}\| \leq$?

2.3 Galerkin–FEM für parabolische ARWA

2.3.1 Verallgemeinerte Formulierungen parabolischer ARWA

- Btr. zunächst parabolische ARWA in <u>**klassischer Formulierung**</u> (vgl. [17] Nu II, Pkt. 1.1: Instationäre Wärmeleitprobleme bzw. Wärmeleit–Wärmetransport-Probleme):

(24)

Ges. $u(x,t) \in X = \overset{x,t}{C^{2,1}}(Q_T) \cap C^{0,0}(\bar{\Omega} \times [0,T)) \cap C^{1,0}(\Omega \cup \Gamma_2 \cup \Gamma_3 \times (0,T))$:

$$\frac{\partial u}{\partial t} - \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x,t)\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{d} a_i(x,t)\frac{\partial u}{\partial x_i} + a(x,t)\,u(x,t) = f(x,t)$$

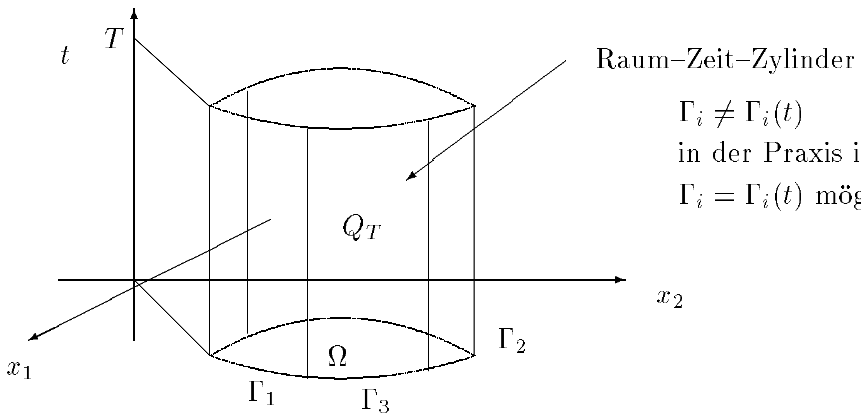
$\forall (x,t) \in Q_T = \Omega \times \boldsymbol{T}$, $\boldsymbol{T} = (0,T)$,

$\Omega \subset I\!R^d$ – beschr. Gebiet: $\Gamma = \partial\Omega \in C^2$;

o. B. d. A. <u>(Homogenisieren)</u>

$$\text{+ RB: } \left.\begin{array}{lll} u(x,t) & = g_1(x,t) \overset{\downarrow}{\equiv} 0, & \forall x \in \Gamma_1 \\ \dfrac{\partial u(x,t)}{\partial N} := \displaystyle\sum_{i,j=1}^{d} a_{ij}(x,t)\frac{\partial u}{\partial x_j}\vec{n}_i & = g_2(x,t) \quad , & \forall x \in \Gamma_2 \\ \dfrac{\partial u(x,t)}{\partial N} + \kappa(x,t)\,u(x,t) & = g_3(x,t) \quad , & \forall x \in \Gamma_3 \end{array}\right\} \ \forall\, t \in \boldsymbol{T}$$

+ AB: $u(x,0) = u_0(x) \quad \forall x \in \bar{\Omega}$.



$\Gamma_i \neq \Gamma_i(t)$
in der Praxis ist auch
$\Gamma_i = \Gamma_i(t)$ möglich !

■ **Variationsformulierung in Sobolev–Räumen der Art $\overset{\circ}{W}{}_2^{1,k}(Q_T)$ auf dem Raum–Zeit–Zylinder $Q_T := \Omega \times T$:**

Sei $\overset{\circ}{W}{}_2^{\overset{x,t}{1,k}}(Q_T) := \left\{ u \in W_2^{1,k}(Q_T) : u|_{\Gamma_1 \times [0,T]} = 0 \right\}$ – Ansatzfkt. = Lsg.–Raum,

$$\dot{W}_2^{1,l}(Q_T) := \left\{ v \in W_2^{1,l}(Q_T) \;:\; \begin{array}{l} u|_{\Gamma_1} = 0 \\ u|_{t=T} = 0 \quad (l \geq 1) \end{array} \right\} \text{ – Testfkt.–Raum,}$$

$\int\limits_{Q_T} \overset{(24)}{\text{PDgl.}} \cdot v(x,t)\,dx\,dt \quad \forall v \in \dot{W}_2^{1,l}(Q_T), \quad k + l = 1:$

(a) Nur partielle Integration im elliptischen Hauptteil:

$$\boxed{\begin{array}{l}
\text{Ges. } u \in \overset{\circ}{W}{}_2^{1,1}(Q_T): \\
\int\limits_{Q_T} \left(\dot{u}v + \sum\limits_{i,j=1}^{d} a_{ij} \dfrac{\partial u}{\partial x_j} \dfrac{\partial v}{\partial x_i} + \sum\limits_{i=1}^{d} a_i \dfrac{\partial u}{\partial x_i} v + auv \right) dx\,dt + \int\limits_0^T \int\limits_{\Gamma_3} \kappa uv\,ds\,dt = \\
\qquad = \int\limits_{Q_T} fv\,dx\,dt + \int\limits_0^T \int\limits_{\Gamma_2} g_2 v\,ds\,dt + \int\limits_0^T \int\limits_{\Gamma_3} g_3 v\,ds\,dt, \quad \forall v \in \overset{\circ}{W}{}_2^{1,0}(Q_T) \\
+ \text{ AB: } u(x,0) = u_0(x) \text{ i.S. } L_2(\Omega), \text{ d.h.} \\
\qquad \|u(\cdot,t) - u_0(\cdot)\|_{L_2(\Omega)} \longrightarrow 0 \text{ für } t \to 0.
\end{array}} \tag{25}$$

mit $\dot{u} := \frac{\partial u}{\partial t}$.

(b) Partielle Integration im elliptischen Hauptteil und im Term mit der Zeitableitung:

$$\boxed{\begin{array}{l}
\text{Ges. } u \in \overset{\circ}{W}{}_2^{1,0}(Q_T): \\
\int\limits_{Q_T} \left(-u\dot{v} + \sum\limits_{i,j=1}^{d} a_{ij} \dfrac{\partial u}{\partial x_j} \dfrac{\partial v}{\partial x_i} + \sum\limits_{i=1}^{d} a_i \dfrac{\partial u}{\partial x_i} v + auv \right) dx\,dt + \int\limits_0^T \int\limits_{\Gamma_3} \kappa uv\,ds\,dt = \\
= \int\limits_{\Omega} u_0(x)\,v(x,0)\,dx + \int\limits_{Q_T} fv\,dx\,dt + \int\limits_0^T \int\limits_{\Gamma_2} g_2 v\,ds\,dt + \int\limits_0^T \int\limits_{\Gamma_3} g_3 v\,ds\,dt \\
\forall v \in \dot{W}_2^{1,1}(Q_T).
\end{array}} \tag{26}$$

Bemerkungen:

1. Existenz–, Eindeutigkeits– und Regularitätsaussagen siehe Literatur (z.B. [14] Ladyshenskaja A.: Aufgaben der Mathematischen Physik, Nauka, Moskau 1973).
2. VF (25) und (26) sind Ausgangspunkt für die FE–Diskretisierung mit Raum-Zeit-Elementen !

■ Linienvariationsformulierung (LVF):

- $\overset{(24)}{\int\limits_\Omega \text{PDgl.}} \cdot v(x)\,dx, \quad \forall v \in V_0 = \{v \in V = W_2^1(\Omega) : v|_{\Gamma_1} = 0\}, \quad \forall$ f.ü. $t \in \boldsymbol{T}$
  $\downarrow \longleftarrow$ Schritte 1. – 5., [17] Nu II, Pkt. 3.1.2.1 !
  LVF:

  $(27) = (24)_{\text{LVF}}$

Ges. $u(x,t)$, sodaß $\forall$ f.ü. $t \in \boldsymbol{T}$, $u \in \overset{(\bar{V}_g)}{V_0}$ und $\dot{u} \in \overset{(?)}{L_2}(\Omega)$:

$$\underbrace{\int\limits_\Omega \dot{u}(x,t)v(x)dx}_{\substack{= (\dot{u},v)_0 \\ \swarrow \quad \searrow \\ :=<\dot{u},v> \;\in\; V_0^* \quad = \frac{d}{dt}(u,v)_0}} + \underbrace{\int\limits_\Omega \left(\sum_{i,j=1}^d a_{ij} \frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \sum_{i=1}^d a_i \frac{\partial u}{\partial x_i} v + auv \right) dx + \int\limits_{\Gamma_3} \kappa uv\,dx}_{\substack{= a(u,v) =: < \underbrace{A(t)\,u(t)}_{\in V_0^*}, v> \\ = a(t;u,v)}} =$$

$$= \underbrace{\int\limits_\Omega fv\,dx + \int\limits_{\Gamma_2} g_2 v\,ds + \int\limits_{\Gamma_3} g_3 v\,ds}_{=: < F(t), v > \;\in\; V_0^*}, \quad \forall v \in V_0 \;\; \forall \text{ f.ü. } t \in \boldsymbol{T}$$

+ AB: $u(x,0) = u_0(x)$ in $L_2(\Omega)$.

$\Updownarrow \longleftarrow$ Nach evtl. Homogenisierung ($u = g_1 + w$ ($\uparrow$)) der Dirichletschen RB erhalten wir eine Cauchy–Aufgabe (AWA) für eine Operatordifferentialgleichung (ODgl.)

$(24)_{\text{ODgl.}}$

Ges. $u(\cdot) : \bar{\boldsymbol{T}} \mapsto V_0$ mit $\dot{u}(\cdot) : \boldsymbol{T} \longrightarrow V_0^*$:
$\dot{u}(t) + A(t)\,u(t) = F(t)$ in V_0^*, $\quad \forall$ f.ü. $t \in \boldsymbol{T}$,
+ AB: $u(0) = u_0$ in $L_2(\Omega)$.

- Zur präzisen Formulierung von (27), sowie von Existenz– und Eindeutigkeitsaussagen benötigen wir noch einige funktionalanalytische Hilfsmittel !

■ Funktionalanalytische Formulierung und Untersuchung der LVF:

- L_2–Räume abstrakter Fkt.:

  X – Banach–Raum (bzw. Hilbert–Raum), $\boldsymbol{T} = (0,T)$.
  $L_2\,(\boldsymbol{T},X) := \{u : \boldsymbol{T} \to X : \|u\|_{L_2(\mathbf{T},X)} < \infty\}$,
  wobei $\|u\|_{L_2(\mathbf{T},X)} := \left(\int\limits_0^T \|u(t)\|_X^2\, dt\right)^{1/2}$.

  Es gilt: (mms)
  1) $L_2(\boldsymbol{T},X)$ ist selbst ein B–Raum.
  2) $L_2(\boldsymbol{T},X^*) = [L_2(\boldsymbol{T},X)]^*$,
     falls X – reflexiv, separabel; X^* = Dualraum zu X.

- Evolutionstripel (Gelfand–Dreier):

  Das Raumtripel $\{X, H, X^*\}$ heißt Evolutionstripel, falls $X \subseteq H \subseteq X^*$:
  (a) X – separabler, reflexibler B–Raum, $\|\cdot\|_X$,
  (b) H – separabler H–Raum, $(\cdot,\cdot) = (\cdot,\cdot)_H$, $\|\cdot\|_H$,
  (c) X liegt dicht in H: $\|v\|_H \leq c\|v\|_X \quad \forall v \in X$.

- Beispiel: 1) $X = \overset{\circ}{H}{}^1(\Omega)$, $H = L_2(\Omega)$, $X^* = H^{-1}(\Omega)$ (1. ARWA),
  2) $X = V_0$, $H = L_2(\Omega)$, $X^* = V_0^*$ ($\uparrow$).

- Man sagt, $u \in L_2\,(\boldsymbol{T}, X)$ besitzt eine verallgemeinerte Ableitung
  $w \equiv \dot{u} \in L_2(\boldsymbol{T}, X^*)$, falls

$$\int\limits_0^T \dot{\varphi}(t)\, u(t)\, dt = -\int\limits_0^T \varphi(t)\, w(t)\, dt \quad \text{in } X^* \quad \forall \varphi \in \dot{C}^\infty(\boldsymbol{T}),$$

  d.h. $< \int\limits_0^T \dot{\varphi}(t)\, u(t)\, dt, v > = - < \int\limits_0^T \varphi(t)\, w(t)\, dt, v > \quad \forall v \in X \quad \forall \varphi \in \dot{C}^\infty(\boldsymbol{T})$,
  wobei $< \cdot,\cdot >: X^* \times X \to I\!R^1$ – Dualitätsprodukt.

- $W_2^1(\boldsymbol{T},X) = W_2^1(\boldsymbol{T};X, H) := \{u \in L_2(\boldsymbol{T},X) : \exists\, \dot{u} \in L_2(\boldsymbol{T},X^*)\}$:
  $\|u\|_{W_2^1(\mathbf{T},X)} := \|u\|_{L_2(\mathbf{T},X)} + \|\dot{u}\|_{L_2(\mathbf{T},X^*)}$ (Grapennorm) bzw.
  $\|u\|^2_{W_2^1(\mathbf{T},X)} := \|u\|^2_{L_2(\mathbf{T},X)} + \|\dot{u}\|^2_{L_2(\mathbf{T},X^*)}$, falls X – Hilbert–Raum.

Für $u \in W_2^1(\boldsymbol{T},X)$ gilt (mms):

1) Abb. $u : \bar{\boldsymbol{T}} = [0,T] \mapsto H$ ist stetig (nach Änderung auf einer Menge vom Maße Null), d.h. $W_2^1(\boldsymbol{T},X) \hookrightarrow C(\bar{\boldsymbol{T}},H)$ und $u(0) \in H$ ist korrekt definiert.

2) Formel der partiellen Integration:

   $(u(t),v(t))_H - (u(s),v(s))_H = \int_s^t [< u'(\tau),v(\tau) > + < v'(\tau),u(\tau) >] d\tau,$

   wobei $< \cdot,\cdot >: X^* \times X \to I\!R^1$ – Dualitätsprodukt.

3) $\frac{d}{dt}(u(t),v) = < \dot{u}(t),v > \quad \forall v \in X \quad \forall$ f.ü. $t \in \boldsymbol{T}$.

- Damit läßt sich die LVF (27) wie folgt formulieren:

(27)

Ges. $u \in W_2^1(\boldsymbol{T},V_0) \equiv W_2^1(\boldsymbol{T};V_0,H)$ mit $u(0) = u_0 \in H = L_2(\Omega)$:

$$\underbrace{\frac{d}{dt}(u(t),v)}_{< \dot{u}(t),v >} + \underbrace{a(u(t),v)}_{< A(t),v >} = < F(t),v > \quad \forall v \in V_0 \quad \forall \text{ f.ü. } t \in \boldsymbol{T}$$

$\dot{u}(t) + A(t)\,u(t) = F(t)$ in V_0^* $\quad \forall$ f.ü. $t \in \boldsymbol{T}$

$\dot{u} + A\,u = F$ $\qquad$ in $L_2(\boldsymbol{T},V_0^*)$

? ↓ ?

in $C(\boldsymbol{T},V_0^*)$
mit
$u \in C^1(\boldsymbol{T},V_0^*)$ ges.

- **Satz 2.17:**

Vor.: 1) Die Bilinearform $a(\cdot,\cdot) : V_0 \times V_0 \to I\!R^1$ (der Einfachheit halber sei $a(\cdot,\cdot)$ t–unabhängig, d.h. $a_{ij} = a_{ij}(x),\cdot)$ sei V_0 – elliptisch und V_0 – beschr., d.h. $\exists\, \mu_1,\mu_2 =$ const. > 0:

a) $\mu_1\|v\|_1^2 \le a(v,v) \quad \forall v \in V_0,$
b) $|a(u,v)| \le \mu_2\|u\|_1\,\|v\|_1 \quad \forall u,v \in V_0.$

2) $F \in L_2\ (\boldsymbol{T},V_0^*)$.
3) $u_0 \in L_2(\Omega) = H$.

Bh.: 1) $\exists\,!\ u \in W_2^1(\boldsymbol{T};V_0,H)$: (27).

2) $\|u(t)\|_0 \le \|u_0\|_0 + \int_0^t \|F(\tau)\|_{V_0^*} d\tau \quad \forall t \in \bar{\boldsymbol{T}}$.

Bew.: siehe [23] Zeidler E.: Vorlesungen über nichtlineare Funktionalanalysis II. Teubner–Texte zur Mathematik,

$\boxed{\to \text{A–nichtlinear}}$ Leipzig 1977, S. 61 ff. und S. 155 ff.

- $\boxed{\textbf{Ü 2.5}}$ Man zeige die zweite Aussage (A–priori–Abschätzung) von Satz 2.17.
  Hinweis: Setzen Sie in (27) $v = u$ ein !

- Regularitätsaussagen siehe Lit., z.B. [21] Thomée, 1984.

Um Diskretisierungsfehlerabschätzungen zu erhalten, benötigen wir in der Regel höhere Glattheit von der Lsg. u als die im $\exists$! – Satz 2.17 angegebene.
Parabolische ARWA besitzen die sogenannte Glättungseigenschaft !

Z.B. gelten für die 1. ARWA
$\dot{u} - \Delta u = 0$ in Q_T, $u = 0$ auf $\partial\Omega\times \boldsymbol{T}$, Ω – glatt, $u = u_0(x)$ für $t = 0$ und für $u_0 \in L_2(\Omega)$ und $t \geq \delta > 0$ die Aussagen:

1) $u \in H^k(\Omega) \quad \forall k \geq 1$,
2) $\|u(t)\|_k \leq ct^{-\frac{1}{2}k}\|u_0\|_0 \quad \forall t > 0$.

2.3.2 Semidiskrete und volldiskrete Ersatzaufgabe

■ **Ausgangspunkt:** = LVF (27):

(27)
$$\boxed{\begin{array}{l}\text{Ges. } u \in W_2^1\ (\boldsymbol{T},V_0) \text{ mit } u(0) = u_0 \in L_2(\Omega) \text{ geg.:}\\ \dfrac{d}{dt}(u(t),v)_0 + a(u(t),v) = < F(t),v > \quad \forall v \in V_0,\ \ t \in \boldsymbol{T}\end{array}}$$

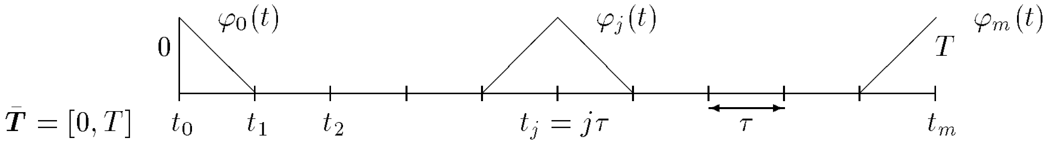
$$\boxed{\dot{u} + A\,u = F \text{ in } L_2(\boldsymbol{T},V_0^*)}$$

■ **Zwei Diskretisierungsstrategien:**

1) Die (vertikale) **Linienmethode**: Erst $\underset{\substack{\uparrow\\ \text{FEM}}}{x}$, dann $\underset{\substack{\uparrow\\ \text{ODE–Solver}}}{t}$!
Diese Methode wird in dieser Vorlesung ausführlich behandelt !

2) Die **Rothe–Methode** (horizontale Linienmethode): Erst $\underset{\substack{\uparrow\\ \text{Impl. Euler}\\ \text{Operator -}\\ \text{ODE–Solver}}}{t}$, dann $\underset{\substack{\uparrow\\ \text{FEM}\\ \text{für ellip.}\\ \text{Problem}}}{x}$!
(siehe auch Pkt. 2.3.3 „Zusammenfassung“)

- Zerlegen $\bar{\boldsymbol{T}} = [0,T] = \bigcup\limits_{\delta=0}^{m-1} [t_j, t_{j+1}]$, $\;t_j = j\tau$ (ohne Beschränkung der Allgemeinheit: gleichmäßig !), $\tau = T/m$:



- Def. $\quad \varphi_j(t) = \begin{cases} (t - t_{j-1})/\tau\,, & t \in [t_{j-1}, t_j] \\ (t_{j+1} - z)/\tau, & t \in [t_j, t_{j+1}], \quad j = \overline{1, m-1} \\ 0 & , \text{ sonst} \end{cases}$

  mit offensichtlichen Modifikationen für $j = 0$ und $j = m$.

- Die Näherung an die Lösung $u(x,t)$ wird durch die Rothe–Funktion

$$u_\tau(x,t) = \sum_{j=0}^{m} u_j(x)\,\varphi_j(t),$$

gegeben, wobei $u_{j+1} \in V_0$ ($\approx u(\cdot, t_{j+1})$) aus der Variationsgleichung

$$\begin{cases} \left(\dfrac{u_{j+1} - u_j}{\tau}, v\right)_0 + a(t_{j+1}; u_{j+1}, v) = < F(t_{j+1}), v > \quad \forall v \in V_0, \\ \qquad\qquad\qquad\qquad (?) \qquad\qquad\qquad (\text{? Stetigkeit}) \\ j = 0, 1, \ldots, m-1; \;\; u_0 \text{ geg.} \end{cases}$$

bestimmt werden.

- Literatur: [18] Rektorys K.: The Method of Discret. in Time and PDEs. Dordrecht, Boston, 1982.
  [10] Kačur J.: Method of Rothe in Evolution Equations. Teubner, Leipzig, 1985.

■ Die semidiskrete Ersatzaufgabe mittels (vertikaler) Linienmethode:

= FEM–Galerkin–Approximation mit zeitabhängigen Koeffizienten ! (vgl. [17] Nu II)

Ansatz: $u_h = u_h(x,t) := \sum\limits_{i \in \omega_h} \underbrace{u^{(i)}(t)}_{\in W_2^1(\boldsymbol{T}) \nearrow} p^{(i)}(x) \in W_2^1\,(\boldsymbol{T}, V_{0h}) \subset W_2^1\,(\boldsymbol{T}, V_0)$,

(28)

wobei $V_{0h} = \text{span}\,\{p^{(i)} : i \in \omega_h\} \subset V_0$ – FE-UR.

$(27)_h$

$$\boxed{\begin{array}{l} \text{Ges. } u_h(x,t): \dfrac{d}{dt}(u_h,v_h)_0 + \overset{a(t;u_h,v_h)}{a(u_h,v_h)} = < F(t), v_h > \quad \forall v_h \in V_{0h} \ \forall \text{ f.ü. } t \in \boldsymbol{T} \\ \qquad\qquad \underline{\text{AB:}} \text{ z.B. } L_2\text{–Projektion:} \\ \qquad\qquad\quad (u_h(\cdot,0), v_h)_0 = (u_0(\cdot), v_h)_0 \quad \forall v_h \in V_{0h}. \end{array}}$$

$\Uparrow \qquad \longleftarrow$ Ansatz (28) für u_h in $(27)_h$ einsetzen: $u_h \leftrightarrow \underline{u}_h \equiv \underline{u}_h(t)$

$\Downarrow \qquad \longleftarrow v_h = p^{(k)}, \quad k \in \omega_h \qquad\qquad \uparrow$

FE–Isom.

$$\boxed{\begin{array}{l} \text{Ges. } \underline{u}_h = [u^{(i)}(t)]_{i\in\omega_h} \in [W_2^1(0,T)]^N; \quad N = N_h = |\omega_h|: \\ \sum\limits_{i\in\omega_h} (p^{(i)}, p^{(k)})_0 \, \dot{u}^{(i)}(t) + \sum\limits_{i\in\omega_h} a(p^{(i)}, p^{(k)}) \, u^{(i)}(t) = < F(t), p^{(k)} >, \quad k \in \omega_h, \\ \underline{\text{AB:}} \sum\limits_{i\in\omega_h} (p^{(i)}, p^{(k)})_0 \, u^{(i)}(0) = (u_0, p^{(k)})_0, \quad k \in \omega_h. \end{array}}$$

$\Updownarrow$

$\underline{(27)}_h$

$$\boxed{\begin{array}{l} \text{Ges. } \underline{u}_h(t) \in [W_2^1(0,t)]^N: \quad M_h \dot{\underline{u}}_h(t) + K_h(t)\underline{u}_h(t) = \underline{f}_h(t), \quad \forall \text{ f.ü. } t \in \boldsymbol{T}, \\ \qquad\qquad\qquad\qquad\qquad M_h \underline{u}_h(0) = \underline{g}_h, \end{array}}$$

mit $M_h = \left[\int\limits_\Omega p^{(i)}(x)\, p^{(k)}(x)\, dx \right]_{k,i\in\omega_h}$ – Massenmatrix,

$\underset{\uparrow}{K_h}(t) = [\underset{\uparrow}{a}(t; p^{(i)}, p^{(k)})]_{k,i\in\omega_h}$ – Steifigkeitsmatrix,

falls Koeff. der PDgl. bzw. der RB 3. Art zeitabhängig sind !

$\underline{f}_h(t) = [< F(t), p^{(k)} >]_{k\in\omega_h} \in [L_2(0,T)]^N$ – Lastvektor,

$\underline{g}_h = \left[\int\limits_\Omega u_0(x) p^{(k)}(x)\, dx \right]_{k\in\omega_h} \in I\!R^N$ – Vektor der „Momente" der AB.

<u>Resultat:</u>

Zur Bestimmung der Vektorfkt. (Koeffizienten) $\underline{u}_h(t) = [u^{(i)}(t)]_{i\in\omega_h} \in [W_2^1(0,T)]^N$ erhalten wir das System $\underline{(27)}_h$ gewöhnlicher Dgl. mit den AB $\underline{u}_h(0) = M_h^{-1}\underline{g}_h$, d.h. eine AWA (Cauchy–Aufgabe).

■ **Probleme:**

1.) $\exists\,!\ \underline{u}_h(t) \in [W_2^1(0,T)]^N$: $\underline{(27)_h}$.
2.) Weitere Eigenschaften der AWA $\underline{(27)_h}$:
? Eventuell <u>steifes</u> Dgl.–System ?
3.) Numerische Verfahren zur Lösung von $\underline{(27)_h}$:
$\Rightarrow$ volldiskrete Ersatzaufgabe ?
? Stabilität, Approximation, Konvergenz, Aufwand ?
4.) Gesamtfehlerabschätzung.

■ **Zu Problem 1.):** $\exists\,!\ \underline{u}_h(t) \in [W_2^1(0,T)]^N$: $\underline{(27)_h}$

Satz 2.18: (Satz von Picard und Lindelöf)

<u>Vor.:</u> Btr. AWA für folgendes System gew. Dgl.

$$(29) \quad \begin{cases} \dfrac{d\underline{y}(t)}{dt} + A(t)\underline{y}(t) = \underline{f}(t) \quad \forall \text{ f.ü. } t \in \boldsymbol{T} = (0,T) \\ \underline{y}(0) = \underline{y}_0 \end{cases}$$

mit $\underline{y} = \underline{y}(t) = (y_1(t), \ldots, y_N(t))^T$ ges.,
$\underline{y}_0 = (y_{01}, \ldots, y_{0N})^T \in I\!R^N$ geg. AW,
$A = A(t) = [a_{ki}(t)]_{k,i=\overline{1,N}} : a_{ki}(\cdot) \in L_\infty(0,T)$,
$\underline{f}(t) \in [L_2(0,T)]^N$.

<u>Bh.:</u> $\exists\,!\ \underline{y}(t) \in [W_2^1(0,T)]^N$: (29).

<u>Beweis:</u> $\rightarrow$ $\boxed{\text{Ü 2.4}}$ (siehe auch [16] Nu I, Bsp. I.2.4)

$\boxed{\textbf{Ü 2.6}}$ Man beweise den Satz 2.18 von Picard und Lindelöf (für nichtstetige Daten) !
$\downarrow$
$\boxed{\textbf{PA V}}$

<u>Hinweise:</u>

1) Übergang zur äquivalenten Igl. $\int\limits_0^t (29)\ d\xi$

$$\underline{\underline{y}(t) = -\int\limits_0^t A(\xi)\ \underline{y}(\xi)\ d\xi + \int\limits_0^t \underline{f}(\xi)\ d\xi + \underline{y}_0}$$

$\underline{y} = B\underline{y}$ (= Fixpunktgleichung)
mit $B : X = [C(\bar{\boldsymbol{T}})]^N \longmapsto [W_2^1(0,T)]^N \subset X$
$\uparrow$
B–Raum !

2) Wenden Sie auf Fixpunktgleichung

$$\text{Ges. } \underline{y} \in X : \underline{y} = B\underline{y} \text{ in } X$$

den verallgemeinerten Banachschen Fixpunktsatz I.2.5 an (vgl. auch Bsp. I.2.4 !) !

<u>Anwendung von Satz 2.18 auf $\underline{(27)}_h$:</u>

- $M_h : M_h = M_h^T$ p.d. (da GRAMER–Matrix bzgl. $(\cdot,\cdot)_0$)

  $M_h \neq M_h(t)$

  $\Longrightarrow \ \exists \, M_h^{-0.5}$

  Setzen: $\underline{u}_h = M_h^{-0.5}\underline{y} \qquad \Rightarrow \quad \underline{\dot{u}}_h = M_h^{-0.5}\underline{\dot{y}}$

$$\Rightarrow \underline{(27)}_h \Longleftrightarrow \boxed{\begin{array}{l} \underline{\dot{y}}(t) + A(t)\,\underline{y}(t) = \underline{f}(t) \\ \underline{y}(0) = \underline{y}_0 \end{array}} \qquad (30)$$

mit $\underline{y} = M_h^{0.5}\underline{u}_h$

$A(t) = M_h^{-0.5} K_h(t)\, M_h^{-0.5} = [a_{ki}(t)]_{k,i\in\omega_h\ (k,i=\overline{1,N_h})}$

$\underline{f}(t) = M_h^{-0.5}\underline{f}_h(t) = [f_k(t)]_{k\in\omega_h\ (k=\overline{1,N_h})}$

$\underline{y}_0 = M_h^{-0.5}\underline{g}_h$

- <u>Offenbar:</u>

  <u>Vor.:</u> 1) Koeff. der Bilinearform $a(\cdot,\cdot) \in L_\infty(Q_T)$.

  2) $f \in L_2(Q_T)$, $\ g_i \in L_2(\Gamma_i \times \boldsymbol{T})$, $\ i = 2, 3$.

  <u>Bh.:</u> 1) $a_{ki}(\cdot) \in L_\infty(0,T)$.

  2) $f_k(\cdot) \in L_2(0,T)$.

  <u>Bew.:</u> (mms)

- Nach Satz 2.18 gilt dann:

$$\Longrightarrow \quad \boxed{\exists\,!\ \underline{y}(t) \in [W_2^1(0,T)]^N : \quad (30)}$$

$$\Updownarrow$$

$$\boxed{\exists\,!\ \underline{u}_h(t) = M_h^{0.5}\underline{y}(t) \in [W_2^1(0,T)]^N : \quad \underline{(27)}_h}$$

■ **Zu Problem 2.):** Weitere Eigenschaften der AWA $\underline{(27)_h} = (30)$:

- Seien die Voraussetzungen von Satz II.4.4 (gleichmäßig in t) erfüllt:

$$(31)\quad \begin{cases} 1)\ a(\cdot,\cdot) : V_0 \times V_0 \to I\!R^1 : V_0 - \text{elliptisch} \quad \text{gleichmäßig in } t\,! \\ \qquad\qquad\qquad\qquad\qquad\quad V_0 - \text{beschränkt} \quad \text{gleichmäßig in } t\,! \\ \qquad\qquad\qquad\qquad\qquad\quad V_0 - \text{symmetrisch.} \\ 2)\ \text{Triangularisierung sei regulär i. S. der Def. II.4.3.} \end{cases}$$

- Nach Satz II.4.4 gilt dann:

$$(32)\quad \begin{cases} \exists\, \underline{c}_E, \bar{c}_E = \text{const.} > 0 : c_E \neq c_E(h) : \\ \underline{\gamma} = \underline{c}_E h^d \leq \underset{\text{EW}}{\lambda(K_h)} \leq \bar{c}_E h^{d-2} = \overline{\gamma} \quad \Rightarrow \kappa(K_h) \leq (\bar{c}_E/\underline{c}_E)\, h^{-2}. \end{cases}$$

Diese Abschätzungen sind ordnungsgemäß scharf (vgl. $\boxed{\text{Ü II.4.6}}$), d.h. $\kappa(K_h) = O(h^{-2})$.

- $\boxed{\textbf{Ü 2.7}}$ Man zeige, daß die Massenmatrix unter der Voraussetzung (31) 2) gut konditioniert ist, d.h. $\exists\, \nu_1, \nu_2 = \text{const.} > 0 : \nu_i \neq \nu_i(h)$:

$$(33)\quad \begin{cases} \nu_1 h^d \leq \lambda(M_h) \leq \nu_2 h^d, \quad \nu_2^{-1} h^{-d} \leq \lambda(M_h^{-1}) \leq \nu_1^{-1} h^{-d}, \\ \kappa(M_h) \leq \nu_2/\nu_1 = O(1). \end{cases}$$

- $\boxed{\textbf{Ü 2.8}}$ Man zeige für $A_h = M_h^{-0.5} K_h M_h^{-0.5}$ die Abschätzungen

$$(34)\quad \begin{cases} \delta_1 = \underline{c}_E\, \nu_2^{-1} \leq \lambda(A_h) \leq \bar{c}_E \nu_1^{-1} h^{-2} = \delta_2 h^{-2}, \\ \text{d.h. } \kappa(A_h) \leq (\delta_2/\delta_1)\, h^{-2}. \end{cases}$$

Darüberhinaus zeige man, daß $\kappa(A_h) = O(h^{-2})$, d.h. die Abschätzungen sind ordnungsgemäß scharf !
Hinweis: (32) und natürlich auch (33) sind ordnungsgemäß scharf !

- Folgerung: Für kleine (?) h ist Dgl.–System $\underline{(27)_h} = (30)$ steif (?) !
(⇒ unbedingt stabile Schemata zur Zeitintegration !)

■ **Zu Problem 3.):** Numerische Verfahren zur Lösung des evtl. steifen Dgl.–Systems (AWA) $\underline{(27)_h} = (30)$: → siehe Kapitel 3 !

- z.B. σ–gewichtete Differenzenschemata aus Pkt. 2.1:

$$\bar{\boldsymbol{T}} = [0,T] \longmapsto \bar{\omega}_\tau = \{t_j = j\tau : j = \overline{0,m}, \ \ \tau = T/m\}:$$
$$0 = t_0 < t_1 < \ldots < t_j = j\tau < \ldots < t_m = T$$

Zeitschritt: $\tau = t_{j+1} - t_j = T/m$,
auch variabler Zeitschritt möglich !
→ Zeitschrittsteuerung !

- Bez.:

$$v^j = [v^{(i)}(t_j)]_{i\in\omega_h} := \underline{U}_h^j = \underline{U}_h(t_j) = [U^{(i)}(t_j)]_{i\in\omega_h} \in \mathbb{R}^{N_h} \text{ bzw. } v^j : \omega_h \to \mathbb{R}^1$$

$\updownarrow$ Gitterfkt. auf j-ter Zeitschicht

$$U_h^j = U_h(t_j) = U_h(x,t_j) = \sum_{i\in\omega_h} U^{(i)}(t_j)\, p^{(i)}(x) \in V_{0h}$$

- Resultat: Volldiskrete Ersatzaufgabe = σ–DS für AWA $\underline{(27)}_h$ = (30):

$\underline{(27)}_{h\tau}$ Ges. $v^j = \underline{U}_h^j \in \mathbb{R}^{N_h} : M_h v_t^j + \sigma K_h(t_{j+1})\, v^{j+1} + (1-\sigma)\, K_h(t_j) v^j = \varphi^j$

$j = 0, 1, \ldots, m-1,$

$\updownarrow$ $\updownarrow$ AB: $v^0 = \underline{u}_h(0) := M_h^{-1}\underline{g}_h$,

wobei z.B. $\varphi^j := \sigma \underline{f}_h(t_{j+1}) + (1-\sigma)\underline{f}_h(t_j)$,

falls $\underline{f}_h(\cdot)$ stetig ist (sonst Mittlungen).

$(27)_{h\tau}$ Ges. $U_h^j \in V_{0h} : (U_{h,t}^j, v_h)_0 + \sigma\, a(t_{j+1}; U_h^{j+1}, v_h) + (1-\sigma)\, a(t_j, U_h^j, v_h) =$

$= < \sigma F(t_{j+1}) + (1-\sigma)\, F(t_j), v_h > \quad \forall v_h \in V_{0h},$

$j = 0, 1, \ldots, m-1,$

AB: $(U_h^0, v_h)_0 = (u_0, v_h)_0 \quad \forall v_h \in V_{0h}.$

- Beispiel: $\boxed{\sigma = 0}$ Explizites Schema (= Euler vorwärts):

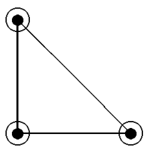
$$M_h \frac{v^{j+1} - v^j}{\tau} = \varphi^j - K_h(t_j)\, v^j \quad \| \text{ GS !}$$

Mass–Lumping, d.h.

Übergang zu einer $\longrightarrow$ $\downarrow$ $\longleftarrow$ (Vorsicht für Elemente höherer Ordnung, d.h. $k \geq 2$!)

Diagonalmatrix

$$D_h = \text{diag}\left[\sum_{i\in\omega_h} m_{ki}\right] \text{ bzw.}$$



$$M_h = [m_{ki}]_{k,i\in\omega_h}$$

lineare Ansätze (k=1):
Integrationspkt.
= Knotenpunkt

$\boxed{\sigma = 1/2}$ CRANK–NICOLSON (= Trapezregel) $\searrow$

unbed. stabil bzgl.
L_2 & Energie
($\downarrow$)

$\boxed{\sigma = 1}$ Rein implizites Schema (= Euler rückwärts) $\nearrow$

- <u>Bemerkung:</u>

  1. Mass–Lumping wird auch für $\sigma \neq 0$ durchgeführt, um für System–Matrix M-Matrix–Eigenschaft zu erhalten:

$$
\begin{array}{ccccc}
M_h & + \ \tau\,\sigma\,K_h\,(t_{j+1}) & \longmapsto & & \underbrace{D_h \ + \ \tau\,\sigma\,K_h\,(t_{j+1})} \\
\nearrow & \uparrow & & & \uparrow \\
\text{garantiert keine} & \text{sei M–Matrix} & & & \text{M–Matrix} \\
\text{M–Matrix} & \text{(unter best. Vor.)} & & & (\Rightarrow C\text{– Stabilität} \\
 & & & & \text{für } \sigma{=}1\ !)
\end{array}
$$

  ! Gesonderte Fehleranalysis notwendig (siehe [21])

  2. $M_h \underline{\dot{u}}_h(t) + K_h(t)\,\underline{u}_h(t) = \underline{f}_h(t)$

$$\Big\downarrow \ \frac{1}{\tau}\int\limits_{t_j}^{t_{j+1}} \ldots dt$$

$$M_h \, \frac{\underline{u}_h(t_{j+1}) - \underline{u}_h(t_j)}{\tau} = \frac{1}{\tau}\int\limits_{t_j}^{t_{j+1}} (\underline{f}_h(t) - K_h(t)\,\underline{u}_h(t))dt \approx$$

$$
\begin{array}{lcl}
\Big\downarrow & & \sigma = 0\text{: Rechteckregel: } t = t_j \\
 & \nearrow & \\
D_h \qquad\qquad\qquad \approx & \longrightarrow & \sigma = 1/2\text{: Trapezregel} \\
 & \searrow & \\
 & & \sigma = 1\text{: Rechteckregel: } t = t_{j+1}
\end{array}
$$

$$\underline{u}_h(t_j) \longmapsto \underline{U}_h^j$$

- <u>Anwendung der allgemeinen Stabilitätstheorie (Energienormstabilität) aus Pkt. 2.2:</u>

  <u>Vor.:</u> $K_h \neq K_h(t)$, d.h. Koeff. der Bilinearform sind t–unabhängig !
  $K_h = K_h^T$ p.d., d.h. Bilinearform ist sym., V_0–ellipt., V_0–beschränkt !

  <u>Allgem. Form:</u> $\boxed{\begin{array}{l} C_1\,v^{j+1} + C_0\,v^j = \tau\,\varphi^j \\ v^0 \text{ geg.} \end{array}}$ $\quad \begin{array}{c} D_h \text{ (Lumping !)} \\ \updownarrow \\ C_1 = M_h + \tau\,\sigma\,K_h \\ C_0 = -M_h + \tau(1-\sigma)K_h \end{array}$

  <u>Kanon. Form:</u> $\boxed{\begin{array}{l} B\,v_t^j + A\,v^j = \varphi^j \\ v^0 \text{ geg.} \end{array}}$ $\quad \begin{array}{l} B = C_1 = M_h + \tau\,\sigma\,K_h = B^T \text{ p.d.} \\ A = \frac{1}{\tau}(C_0 + C_1) = K_h = A^T \text{ p.d.} \end{array}$

  <u>Räume:</u> $H = I\!\!R^n$ mit Euklidischem Skalarprodukt $(\cdot,\cdot) := (\cdot,\cdot)_{I\!\!R^{N_h}}$ bzw. $L_2(\omega_h)$ mit FE-diskretem L_2–Skalarprodukt $(\cdot,\cdot) := (\cdot,\cdot)_{M_h} = (M_h\,\cdot,\cdot)_{I\!\!R^{N_h}}$ oder $(\cdot,\cdot) := (\cdot,\cdot)_{D_h}$ und der entsprechenden Norm $\|\cdot\| = (\cdot,\cdot)^{0.5}$ (im weiteren sei $(\cdot,\cdot) = (\cdot,\cdot)_{I\!\!R^{N_h}}$).

Stabilitätssatz 2.16:

$$\boxed{\begin{array}{l} \underline{\text{Vor.:}}\ (1)-(4) \quad \text{(sind offensichtlich erfüllt)} \\ \qquad B \geq \frac{\tau}{2} A \quad (!) \\ \underline{\text{Bh.:}}\ \|v^{j+1}\|_A \leq \|v^0\|_A + T \underline{\max}_{k=\overline{0,j}} \|B^{-1}\varphi^k\|_A \end{array}}$$

Überprüfen Stabilitätsbedingung: $\boxed{B \geq \frac{\tau}{2} A}$

$$\Rightarrow \boxed{B = M_h + \tau\,\sigma\,K_h \geq \frac{\tau}{2} A = \frac{\tau}{2} K_h} \underset{\uparrow}{\Longleftrightarrow} \boxed{\sigma \geq \frac{1}{2} - \frac{1}{\tau\,\lambda_{\max}}}$$

(mms bzw. $\boxed{\text{PA V}}$)

wobei $\lambda_{\max}$ = Maximaler EW des EWP: $K_h \underline{u}_h = \lambda M_h \underline{u}_h$.

Resultat: 1) $\sigma \geq 1/2 \Longrightarrow$ Stabilitätsbed. trivialerweise erfüllt ! $\Rightarrow$ unbed. stabil !

2) $\sigma = 0 \Longrightarrow$ bedingt stabil: $\tau \leq \frac{h^2}{2\delta_2} \leq \frac{1}{2\lambda_{\max}}$ ($\rightarrow$ $\boxed{\text{Ü 2.6}}$).

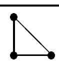
Bem.: $\|\underline{u}_h^j\|_A = \|\underline{u}_h^j\|_{K_h} \cong \|u_h^j\|_1$.

■ **Zu Problem 4.):** Gesamtfehlerabschätzung:

1) $\sigma = 1$: Pkt. 2.3.4.1:

2) $\sigma = \frac{1}{2}$: Pkt. 2.3.4.2:

$$\underline{\max}_{j=\overline{0,m}} \| \underset{\substack{(27)_{\text{LVF}} \\ \text{stetig}}}{u(t_j)} - \underset{\substack{(27)_{h\tau} \\ \text{volldiskret}}}{U_h(t_j)} \|_{\underset{(W_2^1(\Omega))}{L_2(\Omega)}} \leq c(u) \left\{ \underset{(h^k)}{h^{k+1}} + \begin{pmatrix} \tau \\ \tau^2 \end{pmatrix} \right\} \begin{array}{l} 1) \\ 2) \end{array}$$

 $k = 1$: linear

■ **Zusammenfassung:**

$(27)_{\text{LVF}}$

Ges. $u \in W_2^1(\boldsymbol{T}, V_0)$ mit AB: $u(0) = u_0 \in L_2(\Omega)$ geg.:

$$\frac{d}{dt}(u(t), v)_0 + a(t; u, v) = < F(t), v > \quad \forall v \in V_0 \;\; \forall \text{ f.ü. } t \in \boldsymbol{T}$$

$$\dot{u} + Au = F \text{ in } L_2(\boldsymbol{T}, V_0^*)$$

ARWA Ausgangs-aufgabe

(vertikale) Linienmethode $\swarrow$ Ortsdiskr. Zeitdiskr. $\searrow$ Methode von Rothe

AWA für System gew. Dgl.

$(27)_h$ Ges. $u_h = \sum\limits_{i \in \omega_h} u^{(i)}(t) p^{(i)}(x) \in W_2^1(\boldsymbol{T}, V_0)$:

$$(\dot{u}_h, v_h)_0 + a(t; u_h, v_h) = < F, v_h > \quad \forall v_h \in V_h \quad \forall \text{ f.ü. } t \in \boldsymbol{T}$$

$+$ AB: $(u_h(\cdot, 0), v_h)_0 = (u_0(\cdot), v_h)_0 \;\; \forall v_h \in V_h$

$\underline{(27)}_h$ Ges. $\underline{u}_h = \underline{u}_h(t) \in [W_2^1(0, T)]^N$:

$$M_h \underline{\dot{u}}(t) + K_h(t) \underline{u}_h(t) = \underline{f}_h(t) \;\; \forall \text{ f.ü. } t \in \boldsymbol{T}$$

$+$ AB: $M_h \underline{u}_h(0) = \underline{g}_h$

Methode von Rothe

Semi-diskrete Ersatz-aufgabe

$\searrow$ $\swarrow$

$(27)_{h\tau} \updownarrow \underline{(27)}_{h\tau}$

Ges. $U_h^{j+1} \in V_{0h} : (U_{h,t}^j, v_h)_0 + \sigma\, a(t_{j+1}; U_h^{j+1}, v_h) +$

$$+ (1 - \sigma) a(t_j; U_h^j, v_h) =$$

$$= < \sigma F(t_{j+1}) + (1 - \sigma) F(t_j), v_h > \quad \forall v_h \in V_0$$

$j = 0, 1, \ldots, m - 1$; AB: $(U_h^0, v_h)_0 = (u_0, v_h)_0 \;\; \forall v_h \in V_0$

Ges. $\underline{U}_h^{j+1} \in I\!R^{N_h} : M_h \underline{U}_{h,t}^j + \sigma K_h(t_{j+1}) \underline{U}_h^{j+1} +$

$$+ (1 - \sigma) K_h(t_j) \underline{U}_h^j = \underline{\varphi}_h^j,$$

$j = 0, 1, \ldots, m - 1$

$+$ AB: $M_h \underline{u}_h^0 = \underline{g}_h$, mit $\underline{\varphi}_h^j = \sigma \underline{f}_h(t_{j+1}) + (1 - \sigma) \underline{f}_h(t_j)$

Folge von $(m + 1)$ linearen Gleichungssystemen

Voll-diskrete Ersatz-aufgabe

Methode von Rothe

$(27)_\tau$ Ges. $U^{j+1} \in V_0 : (U^{j+1} \equiv u_{j+1}(\uparrow))$

$$(U_t^j, v)_0 + \sigma a(t_{j+1}; U^{j+1}, v) + (1 - \sigma) a(t_j; U^j, v) = < \sigma F(t_{j+1}) + (1 - \sigma) F(t_j), v > \quad \forall v \in V_0$$

$j = 0, 1, \ldots, m - 1; \quad 0 \le \sigma \le 1$

$+$ AB: $U^0 = u_0 \in L_2(\Omega)$ für $\sigma = 1$ $(\uparrow)$ bzw. $U^0 = u_0 \in V_0$ (!) sonst.

Folge von m elliptischen RWA

2.3.3 Fehlerabschätzungen für die Lösung der semidiskreten Ersatzaufgabe $(27)_h$ in der L_2- und in der W_2^1-Norm

- **Btr. ARWA $(27)_{LVF}$:**

$$(27)_{LVF} \quad \boxed{\begin{array}{l} \text{Ges. } u \in W_2^1(\boldsymbol{T},V_0) \text{ mit } \underline{\text{AB:}}\ u(0) = u_0 \in L_2(\Omega) \text{ geg.:} \\ \dfrac{d}{dt}(u(t),v)_0 + a(u(t),v) = < F(t), v > \quad \forall v \in V_0 \quad \forall \text{ f.ü. } t \in \boldsymbol{T} \end{array}}$$

unter den Standardvoraussetzungen:

$$(35) \quad \begin{cases} 1)\ a(\cdot,\cdot) : V_0 \times V_0 \mapsto \mathbb{R}^1 \text{ sei } \underline{t\text{–unabhängige}} \text{ Bilinearform:} \\ 2)\ V_0 - \text{elliptisch: } a(v,v) \geq \mu_1 \|v\|_1^2 \quad \forall v \in V_0 \subset V = W_2^1(\Omega), \\ 3)\ V_0 - \text{beschränkt: } |a(u,v)| \leq \mu_2 \|u\|_1 \, \|v\|_1 \quad \forall u, v \in V_0, \\ 4)\ V_0 - \text{symmetrisch: } a(u,v) = a(v,u) \quad \forall u, v \in V_0 \text{ (der Einfachheit halber !)} \end{cases}$$

und die semidiskrete Ersatzaufgabe $(27)_h$:

$$(27)_h \quad \boxed{\begin{array}{l} \text{Ges. } u_h(x,t) = \sum\limits_{i \in \omega_h} u^{(i)}(t)\, p^{(i)}(x) \in W_2^1(\boldsymbol{T},V_{0h}) \subset W_2^1(\boldsymbol{T},V_0): \\ \dfrac{d}{dt}(u_h, v_h)_0 + a(u_h, v_h) = < F(t), v_h > \quad \forall v_h \in V_{0h} \quad \forall \text{ f.ü. } t \in \boldsymbol{T} \\ \underline{\text{AB:}}\ (u_h(\cdot,0), v_h)_0 = (u_0(\cdot), v_h(\cdot))_0 \quad \forall v_h \in V_{0h} \end{array}}$$

unter den Voraussetzungen:

$$(36) \quad \begin{cases} 1) \text{ Triangularisierung sei regulär i.S. der Def. II.4.3,} \\ 2)\ \tau(\Delta) = \text{span}\{p^{(\alpha)}(\xi) : \alpha \in A\} \supset P_k(\Delta), \\ 3) \text{ Regularität von } u \text{ und } \dot{u} \ (\downarrow) \end{cases}$$

des <u>Approximationssatzes</u> (Satz II.4.5) aus [17] Nu II, Pkt. 4.4.2.

- **<u>Ziel:</u>** 1) $\|u(\cdot,t) - u_h(\cdot,t)\|_0 \leq .?.$ Satz 2.19/2.19*,
  2) $\|u(\cdot,t) - u_h(\cdot,t)\|_1 \leq .?.$ Satz 2.21.

- **Dazu benötigen wir den Ritz-Galerkin–Projektor** (siehe [16] Pkt. I.4.1.3)

$$(37)_0 \quad R_h : V_0 \longmapsto V_{0h} \subset V_0 : a(R_h, u, v_h) = \underbrace{a(u, v_h)}_{=: \; < F_u, v_h >,\ F_u \in V_0^* \text{ (mms)}} \quad \forall v_h \in V_{0h}.$$

siehe $\boxed{\text{Ü I.4.5}}$: $R_h (= P_R$ in Nu I) ist Orthoprojektor bzgl. energetischen Skalarproduktes $[\cdot,\cdot] := a(\cdot,\cdot)$, und es gilt in der Energienorm $|||\cdot|||^2 = [\cdot,\cdot] = a(\cdot,\cdot)$:

$(37)_1$ $$|||u - R_h u||| = \inf_{v_h \in V_{0h}} |||u - v_h|||$$

bzw. (Cea)

$(37)_2$ $$\|u - R_h u\|_1 \le \sqrt{\frac{\mu_2}{\mu_1}} \inf_{v_h \in V_{0h}} \|u - v_h\|_1 \underset{\substack{\uparrow \\ \text{Approximationssatz II.4.5}}}{\le} \bar{c}_{1,k+1}\, h^k |u|_{k+1},$$

falls $a(\cdot,\cdot)$ – symmetrisch.

■ **Zunächst L_2–Abschätzung** (im Unterschied zu ellipt. RWA):

$$\| \underbrace{\overset{(27)_{\mathrm{LVF}}}{u(\cdot,t)} - \overset{(27)_h}{u_h(\cdot,t)}}_{\substack{= z(\cdot,t) \\ \text{Fehler}}} \|_0 \le .?.$$

$$\underset{\substack{= \\ z(t)}}{z(x,t)} = \overset{(27)_{\mathrm{LVF}}}{u} - \overset{(27)_h}{u_h} = \underbrace{u - R_h u}_{\substack{=: \rho = \rho(t) = \rho(x,t) \\ \text{(a)}}} - \underbrace{(u_h - R_h u)}_{\substack{=: \theta_h = \theta_h(t) = \theta_h(x,t) \in V_{0h}\ ! \\ \text{(b)}}}$$

(a) Unter den Voraussetzungen von Satz II.4.9 (L_2–Abschätzung für H^2–koerziative elliptische RWA: $\Rightarrow$ Nitsche–Trick !):

$$\left.\begin{array}{l} \text{1) Reguläre Triangularisierung} \\ \text{2) } \tau(\Delta) \supset P_k(\Delta) \\ \text{3) } u, \dot{u} \in W_2^{k+1}(\Omega) \quad \forall \text{ (f.ü.) } t \in \boldsymbol{T} \end{array}\right\} \quad (36)$$

4) W_2^2 $(= H^2)$–Koerzitivität $(\Rightarrow (37)_0)$

gilt offenbar (siehe Satz II.4.9)

$$(38)\quad \|\rho(t)\|_0 = \|u - R_h u\|_0 \le c_{0,k+1} h^{k+1}\, |u(t)|_{k+1}$$

$$\underset{\substack{\uparrow \\ u(\cdot,t) = \underset{\substack{= \\ u_0(\cdot)}}{u(\cdot,0)} + \int_0^t \underbrace{\dot{u}(\cdot,s)}_{\in V_0^* \cap H^{k+1}}\, ds \in V_0 \cap H^{k+1}}}{\le} c_{0,k+1} h^{k+1} \left\{ |u_0|_{k+1} + \int_0^t |\dot{u}(\cdot,s)|_{k+1}\, ds \right\}.$$

(b) $\|\theta_h(t)\|_0 = \|u_h(\cdot,t) - R_h u(\cdot,t)\|_0 \le .?.$

Btr. $\quad (\dot{\theta}_h, v_h)_0 + a(\theta_h, v_h) =$

$$= \underbrace{(\dot{u}_h, v_h)_0 + a(u_h, v_h)}_{\underset{(27)_h}{=} < F(t), v_h >} - ((R_h \dot{u}), v_h)_0 - \underset{\substack{\| \leftarrow (37)_0 \\ a(u, v_h)}}{a(R_h u, v_h)}$$

$$= \underbrace{< F(t), v_h > - a(u, v_h)} - ((R_h \dot{u}), v_h)_0 \qquad \boxed{a(\cdot,\cdot)\ t \text{ –unabhängig !}}$$

(27)$_{\text{LVF}}$, $V_{0h} \subset V_0 \rightarrow$ $\searrow = \dfrac{d}{dt}(u, v_h)_0 = < \dot{u}, v_h > = (\dot{u}, v_h)_0$

$$\uparrow$$

da $\dot{u} \in V_0^* \cap H^{k+1} \subset L_2(\Omega)$

$$= \left(\frac{d}{dt}(u - R_h u), v_h \right)_0 = (\dot{\rho}(t), v_h)_0$$

(39) <u>Resultat:</u>

$$\boxed{\begin{array}{l} (\dot{\theta}_h, v_h)_0 + a(\theta_h, v_h) = (\dot{\rho}, v_h)_0 \\ \forall v_h \in V_{0h}, \quad \forall \text{ f.ü. } t \in \boldsymbol{T} \end{array}}$$

Setzen hier $\boxed{v_h = \theta_h := u_h - R_h u \in V_{0h}}$

$$\Rightarrow \underbrace{(\dot{\theta}_h, \theta_h)_0}_{\frac{1}{2}\frac{d}{dt}(\theta_h, \theta_h)_0} + \underbrace{a(\theta_h, \theta_h)}_{\geq \mu_1 \|\theta_h\|_1^2 \geq \mu_1 \|\theta_h\|_0^2 \geq 0} = (\dot{\rho}, \theta_h)_0$$

$$\Rightarrow \frac{1}{2} \frac{d}{dt} \|\theta_h\|_0^2 + \mu_1 \|\theta_h\|_0^2 \leq \|\dot{\rho}\|_0 \, \|\theta_h\|_0$$

(40)

$$\boxed{\frac{d}{dt} \|\theta_h\|_0 + \mu_1 \|\theta_h\|_0 \leq \|\dot{\rho}\|_0 \quad \forall \text{ f.ü. } t \in \boldsymbol{T}}$$

(40)'

$$\boxed{\frac{d}{dt} \|\theta_h\|_0 \leq \|\dot{\rho}\|_0 \quad \forall \text{ f.ü. } t \in \boldsymbol{T}}$$

$$\int_0^t (40)' \, ds \;\Rightarrow\; \|\theta_h(t)\|_0 \leq \|\theta_h(0)\|_0 + \int_0^t \|\dot{\rho}(s)\|_0 \, ds =$$

$$= \| \underbrace{u_h(\cdot, 0) - R_h u(\cdot, 0)}_{} \|_0 + \int_0^t \|\dot{u}(\cdot, s) - R_h \dot{u}(\cdot, s)\|_0 \, ds$$

(41) $\Big\| = 0$, falls $u_h(\cdot, 0) = R_h u_0(\cdot)$, d.h. $a(u_h(\cdot, 0), v_h) = a(u_0, u_h) \quad \forall v_h \in V_{0h}$

$$\leq \|u_h(\cdot, 0) - u_0(\cdot)\|_0 + \|u_0 - R_h u_0\|_0 + \int_0^t \|\dot{u} - R_h \dot{u}\|_0 \, ds$$

$$\leq \|u_h(\cdot, 0) - u_0(\cdot)\|_0 + c_{0,k+1} h^{k+1} |u_0|_{k+1} + c_{0,k+1} h^{k+1} \int_0^t |\dot{u}(s)|_{k+1} \, ds$$

(a) + (b): Aus (38) und (41) folgt:

$$\|u(\cdot,t)-u_h(\cdot,t)\|_0 \le \underbrace{\|u_h(\cdot,0)-u_0(\cdot)\|_0}_{\le \inf\limits_{v_h\in V_{0h}} \|u_0-v_h\|_0 \le a_{0,k+1}h^{k+1}|u_0|_{k+1}} + 2c_{0,k+1}h^{k+1}\left\{|u_0|_{k+1}+\int_0^t |\dot u(s)|_{k+1}\,ds\right\}$$

$\uparrow$

$$\boxed{\begin{array}{l} u_h(\cdot,0)=P_{0h}u_0 = L_2\text{–Projektion d. AB, d.h.} \\ (u_h(\cdot,0),v_h)_0=(u_0,v_h) \quad \forall v_h\in V_{0h} \end{array}}$$

$$\le a_{0,k+1}h^{k+1}|u_0|_{k+1} + 2c_{0,k+1}h^{k+1}\left\{|u_0|_{k+1}+\int_0^t |\dot u(s)|_{k+1}\,ds\right\}.$$

Damit haben wir den folgenden Satz bewiesen:

■ **Satz 2.19:** (L_2–Konvergenz: $O(h^{k+1})$)

> Vor.: 1) Vor. (35) an Bilinearform $a(\cdot,\cdot)$ seien erfüllt.
> 2) Vor. $(36)_{1),2)}$ an FE–Diskretisierung seien erfüllt (d.h. Approximationssatz II.4.5 gilt).
> 3) W_2^2–Koerzitivität der Bilinearform (Nitsche–Trick).
> 4) $u_0 \in V_0 \cap W_2^{k+1}(\Omega)$.
> 5) $u, \dot u \in W_2^{k+1}(\Omega) \quad \forall$ (f.ü.) $t\in \boldsymbol{T}=(0,T)$ ($\downarrow$).
>
> Bh.: $$\|u(\cdot,t) - u_h(\cdot,t)\|_0 \le \|u_h(\cdot,0)-u_0(\cdot)\|_0 + 2c_{0,k+1}h^{k+1}\{|u_0|_{k+1}+\int_0^t |\dot u(s)|_{k+1}\,ds\}$$
> $$\le a_{0,k+1}h^{k+1}|u_0|_{k+1} + 2c_{0,k+1}h^{k+1}\{|u_0|_{k+1}+\int_0^t |\dot u(\cdot,s)|_{k+1}\,ds\}$$
> $\uparrow$
> $u_h(\cdot,0)=P_{0h}u_0(\cdot) \qquad \forall t\in \bar{\boldsymbol{T}}=[0,T]$.

■ **Verbesserung der L_2–Abschätzung aus Satz 2.19:**

$\longrightarrow$ exponentielles Abfallen der Fehler in den AB !

Dazu benötigen wir das Gronwall'sche Lemma:

Lemma 2.20: (Gronwall'sches Lemma)

> Vor.: 1) $f\in W_1^1(0,T)$ (d.h. absolut stetig): $f(t)\ge 0 \quad \forall t\in\bar{\boldsymbol{T}}$;
> 2) $\dfrac{d\,f(t)}{dt} \le c_1(t)\,f(t)+c_2(t) \quad \forall$ f.ü. $t\in\boldsymbol{T}=(0,T)$;
> 3) $c_i\in L_1(0,T)$ für $i=1,2$.
>
> Bh.: Dann gilt $\forall t\in\bar{\boldsymbol{T}}=[0,T]$ die Abschätzung:
> $$0\le f(t)\le e^{\int_0^t c_1(s)ds} f(0) + e^{\int_0^t c_1(s)ds}\int_0^t c_2(\xi)\,e^{-\int_0^\xi c_1(s)ds}\,d\xi.$$

Beweis:

$$\frac{d\,f(t)}{dt}\, e^{-\int\limits_0^t c_1(s)ds} \le c_1(t) f(t)\, e^{-\int\limits_0^t c_1(s)ds} + c_2(t)\, e^{-\int\limits_0^t c_1(s)ds}$$

$$\underbrace{\left[\frac{d\,f(t)}{dt} - c_1(t)\, f(t)\right] e^{-\int\limits_0^t c_1(s)ds}}_{} \le c_2(t)\, e^{-\int\limits_0^t c_1(s)ds}$$

$$= \frac{d}{dt}\left(f(t)\, e^{-\int\limits_0^t c_1(s)ds} \right)$$

$\Updownarrow \leftarrow t \to \xi$

$$\frac{d}{d\xi}(f(\xi)\, e^{-\int\limits_0^\xi c_1(s)ds}) \le c_2(\xi)\, e^{-\int\limits_0^\xi c_1(s)ds}$$

$\Downarrow \leftarrow \int\limits_0^t$

$$f(t)\, e^{-\int\limits_0^t c_1(s)ds} - f(0) \le \int\limits_0^t c_2(\xi)\, e^{-\int\limits_0^\xi c_1(s)ds} d\xi, \quad \forall t \in \boldsymbol{T}.$$

q.e.d.

Satz 2.19*: (Verbesserte L_2–Abschätzung)

Vor.: Es gelten die Voraussetzungen 1) – 5) aus Satz 2.19. ($\downarrow$)

Bh.: $(27)_{\mathrm{LVF}} \quad (27)_h$

$$\| u(\cdot,t) - u_h(\cdot,t) \|_0 \le e^{-\mu_1 t} \|u_h(\cdot,0) - u_0(\cdot)\|_0 +$$

(43) $$+ c_{0,k+1} h^{k+1} \left\{ e^{-\mu_1 t}|u_0|_{k+1} + \int\limits_0^t e^{-\mu_1(t-s)}|\dot{u}(\cdot,s)|_{k+1}\, ds + |u(\cdot,t)|_{k+1} \right\}$$

$\forall t \in \bar{\boldsymbol{T}}$.

Beweis: Analog zu Satz 2.19 mit folgenden Modifikationen:

- (38) $\quad \|\rho(t)\|_0 = \|u - R_h u\|_0 \le c_{0,k+1} h^{k+1} |u(\cdot,t)|_{k+1}.$

- Verwenden anstelle von

  (40)' $\quad \dfrac{d}{dt}\|\theta_h\|_0 \le \|\dot{\rho}\|_0 \quad \forall$ f.ü. $t \in \boldsymbol{T}$

  die verschärfte Abschätzung

  (40) $\quad \dfrac{d}{dt}\|\theta_h\|_0 \le -\mu_1\|\theta_h\| + \|\dot{\rho}\|_0 \quad \forall$ f.ü. $t \in \boldsymbol{T}$.

- Anwendung des Gronwall'schen Lemmas mit

$$f(t) = \|\theta_h(t)\|_0, \quad c_1 = -\mu_1, \quad c_2(t) = \|\dot\rho(t)\|_0$$

ergibt:

$$\|\theta_h(t)\|_0 \le e^{-\mu_1 t}\|\theta_h(0)\|_0 + \int\limits_0^t e^{-\mu_1(t-\xi)}\|\dot\rho(\xi)\|_0\, d\xi.$$

q.e.d.

■ W_2^1–Abschätzungen:

Satz 2.21: (W_2^1–Konvergenz: $O(h^k)$)

Vor.: 1) Vor. 1) – 4) aus Satz 2.19 seien erfüllt, $a(\cdot,\cdot)$ – symmetrisch,
2) $u \in V_0 \cap W_2^{k+1}(\Omega)$, $\quad \dot u \in V_0 \cap W_2^k(\Omega) \quad \forall$ f.ü. $t \in \boldsymbol{T}= (0,T)$ ($\downarrow$)

Bh.: $\boxed{\text{Ü 2.7}}$

(44) $$\| \overset{(27)_{\text{LVF}}}{u}(\cdot,t) - \overset{(27)_h}{u_h}(\cdot,t)\|_1 \le \sqrt{\frac{\mu_2}{\mu_1}}\|u_h(\cdot,0) - u_0(\cdot)\|_1 + \sqrt{\frac{\mu_2}{\mu_1}}\bar c_{1,k+1}h^k|u_0|_{k+1} +$$

$$+ \bar c_{1,k+1}h^k|u(\cdot,t)|_{k+1} + \frac{1}{\sqrt{2\mu_1}}\, c_{0,k}h^k\sqrt{\int\limits_0^t |\dot u(\cdot,s)|_k^2\, ds}.$$

Beweis: $z = \overset{(27)_{\text{LVF}}}{u} - \overset{(27)_h}{u_h} = \underbrace{u - R_h u}_{(a) =: \rho} - \underbrace{(u_h - R_h u)}_{(b) =: \theta_h \in V_{0h}}$

(a) $\|\rho(t)\|_1 = \|u - R_h u\|_1 \le c_{1,k+1}\, h^k |u(\cdot,t)|_{k+1}$

(b) Setzen in (39) $\boxed{(\dot\theta_h, v_h)_0 + a(\theta_h, v_h) = (\dot\rho, v_h)_0 \quad \forall v_h \in V_{0h}}$

$v_h = \dot\theta_h \in V_{0h}$:

$$\Longrightarrow \underbrace{(\dot\theta_h,\dot\theta_h)_0}_{=\|\dot\theta_h\|_0^2} + \underbrace{a(\theta_h,\dot\theta_h)}_{=\frac12\frac{d}{dt}a(\theta_h,\theta_h)=\frac12\frac{d}{dt}|||\theta_h|||^2} = (\dot\rho,\dot\theta_h)_0 \underset{\uparrow}{\le} \frac{\varepsilon}{2}\|\dot\rho\|_0^2 + \frac{1}{2\varepsilon}\|\dot\theta_h\|_0^2$$

Cauchy
ε–Ungleichung:
$ab \le \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2$
$\boxed{\varepsilon = \frac12}$

$$\Longrightarrow \boxed{\frac{d}{dt}|||\theta_h|||^2 \le \frac12\|\dot\rho\|_0^2} \text{ mit Bez. } |||\theta_h|||^2 = a(\cdot,\cdot).$$

$$\int\limits_0^t \Longrightarrow |||\theta_h(t)|||^2 \le |||\theta_h(0)|||^2 + \frac12\int\limits_0^t \|\dot\rho(s)\|_0^2\, ds$$

$$\Longrightarrow \ \mu_1 \|\theta_h(t)\|_1^2 \le \mu_2 \|\theta_h(0)\|_1^2 + \frac{1}{2} \int\limits_0^t \|\dot\rho(s)\|_0^2 \, ds$$

$$\Longrightarrow \ \| \theta_h(t) \|_1 \ \le \ \sqrt{\frac{\mu_2}{\mu_1}} \, \| \theta_h(0) \|_1 \ + \ \frac{1}{\sqrt{2\mu_1}} \left(\int\limits_0^t \| \dot\rho(s) \|_0^2 \, ds \right)^{1/2}$$

$$\uparrow \qquad \boxed{\begin{array}{l} \sqrt{a+b} \le \sqrt{a} + \sqrt{b} \\ a,b \ge 0 \end{array}} \qquad \underset{u_h(\cdot,0) - R_h u(\cdot,0) = u_h(\cdot,0) - R_h u_0(\cdot) = u_h - u_0 + u_0 - R_h u_0}{\|}$$

$$\le \sqrt{\frac{\mu_2}{\mu_1}} \, \|u_h(\cdot,0) - u_0(\cdot)\|_1 + \sqrt{\frac{\mu_2}{\mu_1}} \, \bar c_{1,k+1} h^k \, |u_0|_{k+1} +$$

$$+ \frac{1}{\sqrt{2\mu_1}} \, c_{0,k} h^k \left(\int\limits_0^t |\dot u(\cdot,s)|_k^2 \, ds \right)^{1/2}$$

q.e.d.

Bemerkung 2.22:

$$\boxed{L_2\text{–Abschätzung} \xrightarrow[\text{Ungleichung}]{\text{Inverse}} W_2^1\text{–Abschätzung !}}$$

Unter Verwendung der inversen Ungleichung (siehe Lemma II.4.10)

(45) $\quad \|v_h\|_{1,\Omega} \le c\, h^{-1} \|v_h\|_{0,\Omega} \quad \forall v_h \in V_{0h}$

kann aus einer L_2–Abschätzung immer auch eine W_2^1–Abschätzung abgeleitet werden:

Satz 2.19/2.19*

(46)

$$\boxed{\|u(\cdot,t) - u_h(\cdot,t)\|_1 \le (\bar a_{1,k+1} + c\, a_{0,k+1}) h^k |u(t)|_{k+1} + c\, h^{-1} \|u(\cdot,t) - u_h(\cdot,t)\|_0}$$

Interpolant in $V_{0h} : I_h u := \sum\limits_{i \in \omega_h} u(x^{(i)}, t)\, p^{(i)}(x)$!

Tatsächlich,

$$\|u - u_h\|_1 \le \|u - I_h u\|_1 + \| \underbrace{I_h u - u_h}_{\in V_{0h}} \|_1 \le$$

$$\overset{(45)}{\le} \|u - I_h u\|_1 + c\, h^{-1} \, \|I_h u - u_h\|_0 \le$$

$$\le \|u - I_h u\|_1 + c\, h^{-1} \|I_h u - u\|_0 + c\, h^{-1} \|u - u_h\|_0 \le$$

$$\le (\bar a_{1,k+1} + c\, a_{0,k+1}) h^k \, |u(t)|_{k+1} + c\, h^{-1} \|u - u_h\|_0$$

$\uparrow$

Approximationssatz II.4.5 (Beweis)

Nun kann $\|u - u_h\|_0$ sowohl mit Satz 2.19 als auch mit Satz 2.19* weiter abgeschätzt werden. Falls $u_h(\cdot,0) = P_{0h} u_0(\cdot)$, dann gilt:

(47) $\quad \|u_h(\cdot,0) - u_0(\cdot)\|_0 \le a_{0,k+1} h^{k+1} |u_0|_{k+1},$

und damit erhalten wir in beiden Fällen insgesamt eine $O(h^k)$–Abschätzung. Im Falle der Anwendung von Satz 2.19* mit Abklingverhalten $e^{-\mu_1 t}$ des Fehlers in den AB.

Ü 2.7 Man zeige, daß für die L_2–Projektion der AB,

$$u_h(\cdot,0) \;=\; P_{0h}\, u_0(\cdot), \;\text{ d.h.}$$

$$(u_h(\cdot,0), v_h(\cdot))_0 = (u_0(\cdot), v_h(\cdot))_0 \quad \forall v_h \in V_{0h},$$

die Abschätzung

$$\|u_h(\cdot,0) - u_0(\cdot)\|_1 \le (2c\, a_{0,k+1} + \bar{a}_{1,k+1}) h^k |u_0|_{k+1}$$

gilt !

Hinweis: Verwenden Sie wieder die inverse Ungleichung (45) !

Beweis: $\qquad u_h = P_{0h} u_0 \qquad\qquad$ Satz II.4.5

a) $\|u_h(\cdot,0) - u_0(\cdot)\|_0 \overset{u_h = P_{0h}u_0}{=} \inf\limits_{v_h \in V_{0h}} \|u_0 - v_h\|_0 \overset{\text{Satz II.4.5}}{\le} a_{0,k+1} h^{k+1} |u_0|_{k+1}$

b) $\|u_h(\cdot,0) - u_0(\cdot)\|_1 \le \|\underbrace{u_h - I_h u_0}_{\in V_{0h}}\|_1 + \|I_h\, u_0 - u_0\|_1$

$\le c\, h^{-1} \|u_h - I_h\, u_0\|_0 + \|I_h\, u_0 - u_0\|_1 \le$

$\le c\, h^{-1} (\|u_h - u_0\|_0 + \|u_0 - I_h\, u_0\|_0) + \|I_h\, u_0 - u_0\|_1 \le$

$\le c\, h^{-1} (a_{0,k+1} h^{k+1} |u_0|_{k+1} + a_{0,k+1} h^{k+1} |u_0|_{k+1}) + \bar{a}_{1,k+1} h^k |u_0|_{k+1} =$

$= (2c\, a_{0,k+1} + \bar{a}_{1,k+1}) h^k |u_0|_{k+1} = O(h^k).$

2.3.4 Fehlerabschätzung für volldiskrete Ersatzaufgaben in der $\mathbf{L_2}$–Norm

2.3.4.1 Rein implizites Schema (Euler rückwärts)

■ **Btr. rein implizites Schema $(27)_{h\tau} \leftrightarrow (\underline{27})_{h\tau}$ $\;(\boldsymbol{\sigma} = 1)$:**

$(27)_{h\tau}$ Ges. $U_h^{j+1} \in V_{0h} : (U_{h,t}^j, v_h)_0 + a(U_h^{j+1}, v_h) = \langle F(t_{j+1}), v_h \rangle \quad \forall v_h \in V_{0h},$

$\qquad j = 0, 1, \ldots, m-1,$

$\qquad$ AB: $(U_h^0, v_h)_0 = (u_0, v_h)_0 \quad \forall v_h \in V_{0h}.$

$\updownarrow$

$(\underline{27})_{h\tau}$ Ges. $\underline{U}_h^{j+1} \in \mathbb{R}^{N_h} : M_h \underline{U}_{h,t}^j + K_h \underline{U}_h^{j+1} = \underline{\varphi}_h^j := \underline{f}_h(t_{j+1}), \quad j = \overline{0, m-1},$

$\qquad M_h \underline{U}_h^0 = \underline{g}_h$

$V_{0h} \ni U_h^j = \sum\limits_{i \in \omega_h} U^{(i)}(t_j)\, p^{(i)}(x) \longleftrightarrow \underline{U}_h^j = [U^{(i)}(t_j)]_{i \in \omega_h} \in \mathbb{R}^{N_h}.$

■ **Satz 2.23:** (L_2–Konvergenz: $O(h^{k+1}+\tau)$ für impl. Schema)

<u>Vor.:</u> 1. Vor. 1) – 4) aus Satz 2.19.
2. $u, \dot{u} \in W_2^{k+1}(\Omega), \quad \ddot{u} \in L_2(\Omega) \quad \forall$ (f.ü.) $t \in \boldsymbol{T}$ (↓).

<u>Bh.:</u> Dann gilt die Diskretisierungsfehlerabschätzung

$$(48) \qquad \underbrace{\| u(\cdot, t_j) - U_h^j(\cdot) \|_0}_{(27)_{\text{LVF}} \quad (27)_{h\tau}} \leq \overbrace{\| U_h^0 - u_0 \|_0}^{\leq a_{0,k+1} h^{k+1} |u_0|_{k+1}} +$$
$$+2c_{0,k+1} h^{k+1} \{ |u_0|_{k+1} + \int_0^{t_j} |\dot{u}(s)|_{k+1}\, ds \} + \tau \int_0^{t_j} \|\ddot{u}(s)\|_0\, ds,$$

$j = 0, 1, \ldots, m$.

<u>Beweis:</u>

- Bez. $u(t_j) = u(\cdot, t_j) \in V_0 : (27)_{\text{LVF}}$
  $\qquad\qquad\qquad\qquad\qquad\cup$
  $U_h^j = U_h^j(\cdot) \in V_{0h} : (27)_{h\tau}$

- Btr. Fehler
$$u(t_j) - U_h^j = \underbrace{u(t_j) - R_h\, u(t_j)}_{\text{a) } =: \rho^j \in V_0} - \underbrace{(U_h^j - R_h\, u(t_j))}_{\text{b) } =: \theta^j = \theta_h^j \in V_{0h}}$$

- a) $\|\rho^j\|_0 = \|u(t_j) - R_h\, u(t_j)\|_0 \leq c_{0,k+1} h^{k+1} |u(t_j)|_{k+1} =$
$$= c_{0,k+1} h^{k+1} |u(0) + \int_0^{t_j} \dot{u}(s)\, ds|_{k+1} \leq c_{0,k+1} h^{k+1} \{ |u(0)|_{k+1} + \int_0^{t_j} |\dot{u}(s)|_{k+1}\, ds \}.$$

- b) $(\theta_t^j, v_h)_0 + a(\theta^{j+1}, v_h) =$
$$= \underbrace{(U_{h,t}^j, v_h)_0 + a(U_h^{j+1}, v_h)}_{\|} - (\underbrace{(R_h\, u)_t^j}_{\| \leftarrow a(\cdot,\cdot) \text{ zeitunabhängig}}, v_h)_0 - a(R_h\, u(t_{j+1}), v_h) =$$

$R_h\, u_t(t_j) \qquad \searrow \cancel{a(u(t_{j+1}), v_h)}$

$$< F(t_{j+1}), v_h > \; \underset{(27)_{\text{LVF}},\ V_{0h} \subset V_0}{\overset{\uparrow}{=}} \; (\dot{u}(t_{j+1}), v_h)_0 + \cancel{a(u(t_{j+1}), v_h)}$$

$$= (\underbrace{\dot{u}(t_{j+1}) - R_h\, u_t(t_j)}_{=: \tilde{\rho}^j}, v_h)_0 = (\tilde{\rho}^j, v_h)_0 \quad \forall v_h \in V_{0h},$$

Resultat:

(50)

$$\boxed{\begin{array}{l} (\theta_t^j, v_h)_0 + a(\theta^{j+1}, v_h) = (\tilde\rho^j, v_h)_0 \quad \forall v_h \in V_{0h}, \text{ mit} \\ \tilde\rho^j = \dot u(t_{j+1}) - R_h\, u_t(t_j) = \underbrace{\dot u(t_{j+1}) - u_t(t_j)}_{=:\,\tilde\rho_1^j} - \underbrace{(R_h - I)\, u_t(t_j)}_{=:\,\tilde\rho_2^j} = \tilde\rho_1^j - \tilde\rho_2^j \end{array}}$$

Setzen in (50) $\boxed{v_h = \theta^{j+1} \equiv \theta_h^{j+1} \in V_{0h}}$:

(51) $$\boxed{(\theta_t^j, \theta^{j+1})_0 + \underbrace{a(\theta^{j+1}, \theta^{j+1})}_{} = (\tilde\rho^j, \theta^{j+1})_0} \le \|\tilde\rho^j\|_0\, \|\theta^{j+1}\|_0$$

$\downarrow$

1. Variante: ≥ 0

2. Variante: $\ge \mu_1 \|\theta^{j+1}\|_1^2 \ge \mu_1 \|\theta^{j+1}\|_0^2$
$\Rightarrow$ Satz 2.23* (mms)
$\to$ Abklingverhalten !

1. Variante: $(\theta_t^j, \theta^{j+1})_0 \le \|\tilde\rho^j\|_0\, \|\theta^{j+1}\|_0$

$$\theta_t^j = \frac{\theta^{j+1} - \theta^j}{\tau}$$

$$\Longrightarrow \|\theta^{j+1}\|_0^2 - \underbrace{(\theta^j, \theta^{j+1})_0}_{\le \|\theta^j\|_0\, \|\theta^{j+1}\|_0} \le \tau \|\tilde\rho^j\|_0\, \|\theta^{j+1}\|_0$$

$$\Longrightarrow \|\theta^{j+1}\|_0^2 \le \tau \|\tilde\rho^j\|_0\, \|\theta^{j+1}\|_0 + \|\theta^j\|_0\, \|\theta^{j+1}\|_0$$

$$\Longrightarrow \|\theta^{j+1}\|_0 \le \|\theta^j\|_0 + \tau \|\tilde\rho^j\|_0$$

$\Longrightarrow$ Rekursiv anwenden:

(52)

$$\boxed{\begin{array}{l} \|\theta^j\|_0 \le \|\theta^0\|_0 + \tau \sum\limits_{k=0}^{j-1} \|\tilde\rho^k\|_0 \le \\ \qquad \le \underbrace{\|\theta^0\|_0}_{1)} + \underbrace{\tau \sum\limits_{k=0}^{j-1} \|\tilde\rho_1^k\|_0}_{2)} + \underbrace{\tau \sum\limits_{k=0}^{j-1} \|\tilde\rho_2^k\|_0}_{3)} \end{array}}$$

1) $\|\theta^0\|_0 = \|U_h^0 - R_h\, \underset{\overset{\|}{u(0)}}{u_0}\|_0 \leq \|U_h^0 - u_0\|_0 + \|u_0 - R_h\, u_0\|_0 \leq$
$\qquad\qquad \leq \|U_h^0 - u_0\|_0 + c_{0,k+1} h^{k+1} |u_0|_{k+1}$

2) $\tilde{\rho}_1^k = \dot{u}(t_{k+1}) - u_t(t_k) = \dot{u}(t_{k+1}) - \frac{1}{\tau}(u(t_{k+1}) - u(t_k)) =$

$$= \dot{u}(t_{k+1}) - \frac{1}{\tau} \int\limits_{t_k}^{t_{k+1}} \dot{u}(s)\, ds \underset{\uparrow}{=} \frac{1}{\tau} \int\limits_{t_k}^{t_{k+1}} (s - t_k)\, \overset{1\times}{\ddot{u}}(s)\, ds$$

$$\underline{\text{partielle Integration:}} = -\frac{1}{\tau} \int\limits_{t_k}^{t_{k+1}} \dot{u}(s)\, ds + \frac{1}{\tau}(s - t_k)\, \dot{u}(s)\Big|_{t_k}^{t_{k+1}}$$

$$\Longrightarrow\ \tau \sum_{k=0}^{j-1} \|\tilde{\rho}_1^k\|_0 = \sum_{k=0}^{j-1} \| \int\limits_{t_k}^{t_{k+1}} \underbrace{(s - t_k)}_{\leq \tau}\ \ddot{u}(s)\, ds\|_0 \leq$$

$$\leq \tau \sum_{h=0}^{j-1} \int\limits_{t_k}^{t_{k+1}} \|\ddot{u}(s)\|_0\, ds = \tau \int\limits_0^{t_j} \|\ddot{u}(s)\|_0\, ds$$

3) $\tilde{\rho}_2^k = (R_h - I)\, u_t(t_j) = (R_h - I) \frac{1}{\tau} \int\limits_{t_k}^{t_{k+1}} \dot{u}(s)\, ds \underset{\uparrow}{=} \frac{1}{\tau} \int\limits_{t_k}^{t_{k+1}} (R_h - I)\, \dot{u}(s)\, ds$

$a(\cdot,\cdot)$ – zeitunabhängig !

$$\Longrightarrow\ \tau \sum_{k=0}^{j-1} \|\tilde{\rho}_2^k\|_0 = \sum_{k=0}^{j-1} \| \int\limits_{t_k}^{t_{k+1}} (R_h - I)\, \dot{u}(s)\, ds\|_0 \leq$$

$$\leq \sum_{h=0}^{j-1} \int\limits_{t_k}^{t_{k+1}} \|R_h\, \dot{u}(s) - \dot{u}(s)\|_0\, ds \leq$$

$$\leq c_{0,k+1}\, h^{k+1} \int\limits_0^{t_j} |\dot{u}(s)|_{k+1}\, ds$$

$\Longrightarrow$ (49) + $(52)_{1)-3)} \Rightarrow$ (48) **q.e.d.**

2.3.4.2 CRANK–NICOLSON–Schema

■ **Btr. CRANK–NICOLSON–Schema $(27)_{h\tau} \leftrightarrow (\underline{27})_{h\tau}$ $(\boldsymbol{\sigma} = 0.5)$:**

$(27)_{h\tau}$ Ges. $U_h^{j+1} \in V_{0h} : (U_{h,t}^j, v_h)_0 + a(\frac{1}{2}(U_h^{j+1} + U_h^j), v_h) = < F(t_{j+\frac{1}{2}}), v_h > \quad \forall v_h \in V_{0h}$
$\updownarrow \qquad j = 0, 1, \ldots, m-1,$
$\qquad \underline{\text{AB:}}\ (U_h^0, v_h)_0 = (u_0, v_h)_0 \quad \forall v_h \in V_{0h}.$

$(\underline{27})_{h\tau}$ Ges. $\underline{U}_h^{j+1} \in I\!R^{N_h} : M_h U_{h,t}^j + K_h(\frac{1}{2}(\underline{U}_h^{j+1} + \underline{U}_h^j)) = \underline{f}_h(t_{j+\frac{1}{2}}), \quad j = \overline{0, m-1}$
$\qquad M_h \underline{U}_h^0 = \underline{g}_h$

■ **Satz 2.24:** (L_2–Konvergenz: $O(h^{k+1}+\tau^2)$ für CN–Schema)

Vor.: 1. Vor. 1) – 4) aus Satz 2.19.

2. $u, \dot u \in W_2^{k+1}(\Omega)$, $\quad \ddot u \in W_2^2(\Omega)$, $\quad u''' \in L_2(\Omega) \quad \forall$ f.ü. $t \in \boldsymbol{T}$ ($\downarrow$),

wobei $u''' = d^3u|dt^3$.

Bh.: Dann gilt die Diskretisierungsfehlerabschätzung

$$(53) \qquad \overset{(27)_{\mathrm{LVF}}}{\| \, u\,(\cdot, t_j)} \;-\; \overset{(27)_{h\tau}}{U_h^j\,(\cdot)\,\|_0} \le \overbrace{\| \, U_h^0 - u_0 \, \|_0}^{\le\, a_{0,k+1}\, h^{k+1} |u_0|_{k+1}} + 2c_{0,k+1}\, h^{k+1}\{|u_0|_{k+1} + \int_0^{t_j} |\dot u(s)|_{k+1}\, ds\} +$$

$$+ \left(\frac{1+2c_2}{8}\right)\,\tau^2 \left[\int_0^{t_j} (\|\ddot u(s)\|_0 + \|\ddot u(s)\|_2)\, ds\right],$$

wobei $c_2 = W_2^2$–Beschränktheitskonstante ($\Rightarrow$ partielle Integration):

$$(54) \qquad |a(w,v)| \le c_2\, \|w\|_2\, \|v\|_0 \quad \forall w \in W_2^2(\Omega) \cap V_0, \quad \forall v \in V_0,$$

$c_{0,k+1}$ = Konstante aus Satz II.4.9 (L_2–Abschätzung),

$a_{0,k+1}$ = Konstante aus Approximationssatz II.4.5.

Beweis:

- Btr. Fehler

$$z^j = u(t_j) - U_h^j = \underbrace{u(t_j) - R_h\, u(t_j)}_{\text{a) } =:\, \rho^j \,\in\, V_0} - \underbrace{(U_h^j - R_h\, u(t_j))}_{\text{b) } :=\, \theta^j \,=\, \theta_h^j \,\in\, V_{0h}}$$

- a) $\|\rho^j\|_0 \le c_{0,k+1}\, h^{k+1} |u(t_j)|_{k+1} \le c_{0,k+1}\, h^{k+1}\, \{|u_0|_{k+1} + \int_0^{t_j} |\dot u(s)|_{k+1}\, ds\}$.

- b) $(\theta_t^j, v_h)_0 + a\left(\frac{1}{2}(\theta^{j+1} + \theta^j), v_h\right) = \qquad a(\cdot,\cdot) = t$-unabhängig !

(55)

$$\underbrace{(U_{h,t}^j, v_h)_0 + a(\tfrac{1}{2}(U_h^{j+1} + U_h^j), v_h)}_{(27)_{\mathrm{LVF}},\ V_{0h}\subset V_0} - (\underbrace{(R_h\, u(t_j))_t}_{=\, R_h\, u_t(t_j)}, v_h)_0 - \underbrace{a\left(R_h\left(\tfrac{u(t_{j+1})+u(t_j)}{2}\right), v_h\right)}_{=\, a(\frac{1}{2}(u(t_{j+1}) + u(t_j)), v_h)} =$$

$$\| \qquad \downarrow$$

$$< F(t_{j+1/2}), v_h > = (\dot u(t_{j+1/2}), v_h) + a(u(t_{j+1/2}), v_h), \quad t_{j+1/2} = t_j + \tfrac{1}{2}\tau$$

$$= (\dot u(t_{j+1/2}), v_h)_0 - (R_h\, u_t(t_j), v_h)_0 + a\left(u(t_{j+1/2}) - \frac{u(t_{j+1}) + u(t_j)}{2}, v_h\right)$$

$$= (\underset{\text{(i)}}{\dot u(t_{j+1/2}) - u_t(t_j)}, v_h)_0 + (\underset{\text{(ii)}}{(I - R_h)\, u_t(t_j)}, v_h)_0 +$$

$$+ a\left(\underset{\text{(iii)}}{u(t_{j+1/2}) - \frac{1}{2}(u(t_{j+1}) - u(t_j))}, v_h\right) =: < \hat\rho^j, v_h >$$

unter Benutzung der Beziehungen

$$(56)\qquad \dot u(t_{j+1/2}) - u_t(t_j) = \frac{1}{2\tau}\left(\int\limits_{t_j}^{t_{j+1/2}} (s-t_j)^2 u'''(s)\,ds + \int\limits_{t_{j+1/2}}^{t_{j+1}} (s-t_{j+1})^2 u'''(s)\,ds\right)$$

2× partiell integrieren !

und

(57)

$$u(t_{j+1/2}) - \frac{1}{2}(u(t_{j+1}) - u(t_j)) = \frac{1}{2}\left[\int\limits_{t_{j+1/2}}^{t_{j+1}} (s-t_{j+1})\,\ddot u(s)\,ds - \int\limits_{t_j}^{t_{j+1/2}} (s-t_j)\,\ddot u(s)\,ds\right]$$

1× partiell integrieren !

erhalten wir die Abschätzungen

(i) $|(\dot u(t_{j+1/2}) - u_t(t_j), v_h)_0| \le \|\dot u(t_{j+1/2}) - u_t(t_j)\|_0\,\|v_h\|_0 \le$

$$\overset{(56)}{\le} \frac{1}{2\tau}\,\frac{\tau^2}{4}\left[\int\limits_{t_j}^{t_{j+1/2}} \|u'''(s)\|_0\,ds + \int\limits_{t_{j+1/2}}^{t_{j+1}} \|u'''(s)\|_0\,ds\right]\|v_h\|_0 =$$

$$= \frac{\tau}{8}\int\limits_{t_j}^{t_{j+1}} \|u'''(s)\|_0\,ds\cdot\|v_h\|_0.$$

(ii) $$|\,((I-R_h)\underbrace{u_t(t_j)}_{=\frac{1}{\tau}\int_{t_j}^{t_{j+1}}\dot u(s)\,ds}, v_h)_0\,| = \left|\left(\frac{1}{\tau}\int\limits_{t_j}^{t_{j+1}}(I-R_h)\,\dot u(s)\,ds, v_h\right)_0\right| \le$$

$\uparrow$ $a(\cdot,\cdot)$ t-unabhängig

$$\le \Big\|\frac{1}{\tau}\int\limits_{t_j}^{t_{j+1}}(I-R_h)\,\dot u(s)\,ds\Big\|_0\,\|v_h\|_0 \le$$

$$\le \frac{1}{\tau}\int\limits_{t_j}^{t_{j+1}}\|(I-R_h)\,\dot u(s)\,ds\|_0\,\|v_h\|_0 \le$$

$$\le \frac{1}{\tau}\,c_{0,k+1}\,h^{k+1}\int\limits_{t_j}^{t_{j+1}}|\dot u(s)|_{k+1}\,ds\cdot\|v_h\|_0.$$

(iii) $$\left|a\left(u(t_{j+1/2}) - \frac{1}{2}(u(t_{j+1}) + u(t_j)), v_h\right)\right| \overset{(57)}{=}$$

$$= \left| a\left(\frac{1}{2}\left[\int\limits_{t_{j+1/2}}^{t_{j+1}} (s-t_{j+1})\,\ddot{u}(s)\,ds - \int\limits_{t_j}^{t_{j+1/2}} (s-t_j)\,\ddot{u}(s)\,ds\right], v_h\right)\right| \overset{(54)}{\leq}$$

$$\leq c_2 \left\| \frac{1}{2}\left[\int\limits_{t_{j+1/2}}^{t_{j+1}} (s-t_{j+1})\,\ddot{u}(s)\,ds - \int\limits_{t_j}^{t_{j+1/2}} (s-t_j)\,\ddot{u}(s)\,ds\right]\right\|_2 \|v_h\|_0 \leq$$

$$\leq \frac{c_2}{4}\,\tau \int\limits_{t_j}^{t_{j+1}} \|\ddot{u}(s)\|_2\,ds \cdot \|v_h\|_0.$$

Aus (i) – (iii) folgt sofort die Abschätzung

$$(58) \qquad |<\hat{\rho}^j, v_h>| \leq \left[\frac{\tau}{8}\int\limits_{t_j}^{t_{j+1}} \|u'''(s)\|_0\,ds + \frac{1}{\tau}\,c_{0,k+1}\,h^{k+1}\int\limits_{t_j}^{t_{j+1}} |\dot{u}(s)|_{k+1}\,ds + \right.$$

$$\left. + \frac{c_2}{4}\,\tau\int\limits_{t_j}^{t_{j+1}} \|\ddot{u}(s)\|_2\,ds\right] \|v_h\|_0 = [(58)^j]\,\|v_h\|_0$$

Setzen in (55) $\quad v_h = \dfrac{1}{2}(\theta^j + \theta^{j+1}) \in V_{0h}: \;\Rightarrow$

$$\underbrace{\left(\theta_t^j, \frac{1}{2}(\theta^j+\theta^{j+1})\right)_0}_{\theta_t^j \,=\, \frac{\theta^{j+1}-\theta^j}{\tau}} + \underbrace{a\left(\frac{\theta^j+\theta^{j+1}}{2}, \frac{\theta^j+\theta^{j+1}}{2}\right)}_{\geq \mu_1 \|\frac{\theta^j+\theta^{j+1}}{2}\|_1^2 \,\geq\, \frac{\mu_1}{4}\|\theta^j+\theta^{j+1}\|_0^2 \,\geq\, 0\;!} = <\hat{\rho}^j, \frac{\theta^j+\theta^{j+1}}{2}>$$

$$\underbrace{\|\theta^{j+1}\|_0^2 - \|\theta^j\|_0^2}_{=\,(\|\theta^{j+1}\|_0 - \|\theta^j\|_0)\,(\|\theta^j\|_0 + \|\theta^{j+1}\|_0)} \overset{(58)}{\leq} \tau[(58)^j]\,(\|\theta^j\|_0 + \|\theta^{j+1}\|_0)$$

$$\Rightarrow \boxed{\|\theta^{j+1}\|_0 \leq \|\theta^j\|_0 + \tau[(58)^j] \leq \|\theta^0\|_0 + \tau\sum_{l=0}^{j}[(58)^l]}$$

Wegen $\|\theta^0\|_0 \leq \|U_h^0 - u_0\|_0 + \|u_0 - R_h\,u_0\|_0 \leq \|U_h^0 - u_0\|_0 + c_{0,k+1}\,h^{k+1}\,|u_0|_{k+1}$ erhalten wir

$$(59) \qquad \|\theta^j\|_0 \leq \|\theta^0\|_0 + \tau\sum_{l=0}^{j-1}[(58)^l] \leq$$

$$\leq \|U_h^0 - u_0\|_0 + c_{0,k+1}\,h^{k+1}|u_0|_{k+1} + \frac{\tau^2}{8}\int\limits_0^{t_j} \|u'''(s)\|_0\,ds +$$

$$+ c_{0,k+1}\,h^{k+1}\int\limits_0^{t_j} |\dot{u}(s)|_{k+1}\,ds + \frac{c_2}{4}\,\tau^2\int\limits_0^{t_j} \|\ddot{u}(s)\|_2\,ds.$$

Aus a) und (59) folgt (53). **q.e.d.**

2.3.5 Abschließende Bemerkungen

■ **Fehlerabschätzungen in anderen Normen (z.B. H^1, L_∞):**

1. Über inverse Ungleichung unter Benutzung der L_2–Abschätzungen.
   Vorsicht: z.B. liefert diese Technik für CN–Schema offenbar die folgende
   H^1–Abschätzung:
   $$\|u(\cdot,t_j) - u_h^j(\cdot)\|_1 \underset{?}{=} O(h^k + h^{-1}\,\tau^2) \underset{k=1,\ \tau=O(h)}{=} O(h)\ !$$

2. Direkte Abschätzung ($e^{-\mu_1 t}$ bzw. $e^{-\mu_1(t-s)}$ – Abklingen ?)
   a) H^1–Norm: analog zu Satz 2.21 !
   b) L_∞–Norm: siehe [4] S. 311, [21] S. 62 ff.

3. Mass–Lumping ist nur für $k = 1$ (lineare Dreieckselemente) begründet:
   siehe [4] S. 312 ff., [21] S. 166 ff.

■ **Auflösung:**

1. Explizites Schema ($\sigma = 0$): $\quad M_h \underline{U}_{h,t}^j + K_h(t_j)\,\underline{U}_h^j = \underline{f}_h(t_j)$

   (60) $\quad M_h \underline{U}_h^0 = \underline{g}_h =: \underline{d}_h^0$
   $\qquad j = 0, 1, \ldots, m-1$
   (60) $\quad \underbrace{M_h}_{\downarrow\ D_h} \underline{U}_h^{j+1} = \underline{d}_h^j := \underbrace{M_h}_{\downarrow\ D_h} \underline{U}_h^j + \tau(\underline{f}_h(t_j) - K_h(t_j)\,\underline{U}_h^j)$

   $\longleftarrow$ Mass–Lumping für $k = 1$
   $\Rightarrow$ Lösung ist explizit hinschreibbar !

   Bemerkung: $\kappa(M_h) \le \nu_2/\nu_1 = O(1)$, d.h. M_h ist gut konditioniert.
   Folglich sind obige GS mit Systemmatrix offenbar selbst mit klassischen Iterationsverfahren (z.B. Jacobi–Verfahren) schnell auflösbar. Startnäherung für (60) ist Näherung $\underline{U}_h^{j,h_j}$ von letzter Zeitschicht.

2. Implizite Schemata ($0 < \sigma \le 1$):

   (60) $\quad M_h \underline{U}_h^0 = \underline{d}_h^0 := \underline{g}_h^0$
   $\qquad j = 0, 1, \ldots, m-1$

(61) $\quad (M_h + \tau\,\sigma\,K_h(t_{j+1}))\,\underline{U}_h^{j+1} = \underline{d}_h^j := M_h \underline{U}_h^j - \tau(1-\sigma)\,K_h(t_j)\,\underline{U}_h^j + \tau\underline{\varphi}_h^j$

$$\begin{array}{ccc} \downarrow & \qquad\qquad\qquad\qquad & \downarrow \longleftarrow \\ D_h & & D_h \end{array}$$

Mass–Lumping für $k=1$
auch im Fall $\sigma \neq 0$ möglich
und sinnvoll, insbesondere
für $\sigma = 1$ ($\Rightarrow$ C–Stabilität)
! $D_h + \tau\sigma K_h(t_{j+1})$ ist für $k=1$
und $\sphericalangle \leq 90°$ M–Matrix !

Offenbar gelten für die Systemmatrizen $M_h + \tau\,\sigma\,K_h(t_{j+1})$ die folgenden Eigenwert– und Konditionsabschätzungen:

$$\begin{array}{ccc} \nu_1\,h^d + \tau\,\sigma\,\underline{c}_E\,h^d & \leq \lambda(M_h + \tau\,\sigma\,K_h) & \leq \nu_2\,h^d + \tau\,\sigma\,h^{d-2} \\ \geq & EW & \leq \\ (\nu_1 + \tau\,\sigma)\,h^d & & (\nu_2 + \sigma\,\tau\,h^{-2})\,h^d \\ \geq & & \\ \nu_1\,h^d & & \end{array}$$

Also: $\boxed{\kappa(M_h + \tau\,\sigma\,K_h) \leq \dfrac{\nu_2 + \sigma\,\tau\,h^{-2}}{\nu_1 + \sigma\,\tau} \leq \dfrac{\nu_2}{\nu_1} + \dfrac{\sigma}{\nu_1}\,\tau\,h^{-2}}$

<u>Damit gilt z.B. für $k = 1$ und $K_h \neq K_h(t_j)$ o. B. d. Allg.:</u>

a) <u>$\sigma = 1$:</u> L_2–Konvergenz: $O(h^2 + \tau)$, d.h. $\tau = O(h^2)$

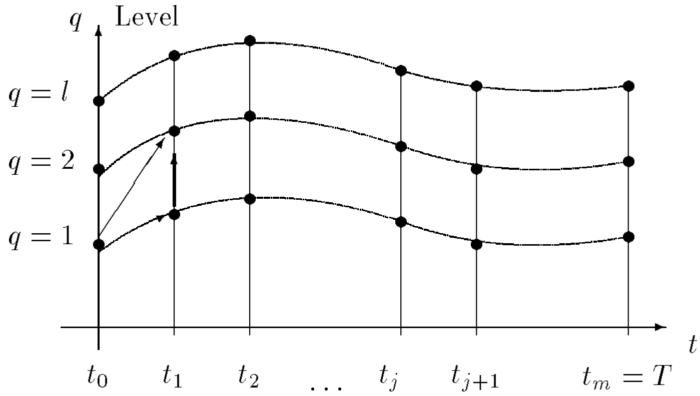
$$\underset{\tau \leq c\,h^2}{\overset{\Longrightarrow}{\uparrow}} \quad \boxed{\kappa(M_h + \tau\,K_h) \leq \frac{\nu_2}{\nu_1} + \frac{1}{\nu_1}\cdot \boldsymbol{c} = O(1)}$$

<u>Praktisch:</u> $\tau \leq c\,h^2$, wobei $c = \text{const.} > 0$ wesentlich größer sein kann als Konstante in der <u>Stabilitätsbedingung</u> für explizites Schema !

b) <u>$\sigma = 1/2$:</u> L_2–Konvergenz: $O(h^2 + \tau^2)$, d.h. $\tau = O(h)$

$$\underset{\tau \leq c\,h}{\overset{\Longrightarrow}{\uparrow}} \quad \boxed{\kappa\left(M_h + \frac{\tau}{2}\,K_h\right) \leq \frac{\nu_2}{\nu_1} + \frac{1}{2\nu_1}\cdot \boldsymbol{c}\;h^{-1} = O(h^{-1})}$$

$\Longrightarrow$ Lösung von (61) durch Multigrid–Verfahren bzw. MG–Präkonditionierte CG–Verfahren + Nested Iteration (siehe auch [17]):



■ **<u>Nichtlineare parabolische ARWA:</u>** ($\Rightarrow$ Kap. 3)

- Nichtlinearitäten: a) Koeffizienten hängen von der Temperatur ab
  b) Strahlung
- $\{V_0, H, V_0^*\}$ – sei Evolutionstripel, $\boldsymbol{T}= (0,T)$, $\frac{1}{p} + \frac{1}{q} = 1$

(62)

Ges. $u \in W_p^1(\boldsymbol{T}; V_0, H) := \{u \in L_p(\boldsymbol{T}, V_0) : \exists\, \dot{u} \in L_q(\boldsymbol{T}, V_0^*)\}$:

$\dot{u}(\cdot) + A(\cdot)\, u(\cdot) = F(\cdot)$ in $L_q(\boldsymbol{T}, V_0^*)$

<u>AB:</u> $u(0) = u_0 \in H$

mit geg. $F \in L_q(\boldsymbol{T}, V_0^*)$, $\quad u_0 \in H$,

$A(\cdot) : V_0 \to V_0^*$

Genauer: $A : W_p^1(\boldsymbol{T}; V_0, H) \longrightarrow L_q(\boldsymbol{T}, V_0^*)$

$\downarrow$ FE–Galerkin–Semidiskretisierung

<u>(62)</u>$_h$

Ges. $\underline{u}_h(t) \in [W_p^1(0,T)]^N$:

$M_h \underline{\dot{u}}_h(t) + K_h(t, \underline{u}_h(t)) = \underline{f}_h(t) \quad \forall$ f.ü. $t \in \boldsymbol{T}$

$M_h \underline{u}_h(0) = \underline{g}_h \qquad \|$ oft

$\hat{K}_h(t; \underline{u}_h(t))\, \underline{u}_h(t)$ – quasilinear

$\|$

AWA für System nichtlinearer gew. Dgl. 1. Ordnung

Ges. $u(\cdot) : \bar{\boldsymbol{T}} \to \mathbb{R}^N : \dot{u}(t) = f(t, u(t)), \quad t \in \boldsymbol{T}$

AB: $u(0) = u_0$ geg.

- <u>Literatur:</u> [2], [10], [23].

Kapitel 3

Anfangswertaufgaben für gewöhnliche Differentialgleichungen und Systeme gewöhnlicher Dgl.

■ <u>Literatur:</u>

[1] Deuflhard P., Bornemann F.: Numerische Mathematik II: Integration gewöhnlicher Differentialgleichungen.
de Gruyter Lehrbuch; Berlin · N.Y. 1994.

[5] Hairer E., Nørsett S.P., Wanner G.: Solving Ordinary Differential Equation I: Nonstiff Problems.
Springer–Verlag , Berlin · Heidelberg 1987.

[6] Hairer E., Wanner G.: Solving Ordinary Differential Equation II: Stiff and Differential–Algebraic Problems.
Springer–Verlag, Berlin · Heidelberg 1991.

[19] Stetter H.J.: Analysis of Discretization Methods for Ordinary Differential Equations.
Springer–Verlag, Berlin · Heidelberg · N.Y. 1973.

■ <u>**Bezeichnung:**</u>

$$\dot{u}(t) \;=\; u'(t) \;:=\; \frac{d}{dt}\,u(t)$$

3.1 Beispiele

■ **Bsp. 3.1:** Chemische Reaktionen (Brusselator)

Brusselator (von R. Lefever und G. Nicolis, 1971) ist ein Modell einer chemischen Reaktion, die aus den folgenden Einzelreaktionen der Substanzen A, B, D, E, X, Y besteht:

$$\boxed{\begin{array}{rcll} A & \xrightarrow{k_1} & X & \text{(monomolecular reaction)}, \\ B + X & \xrightarrow{k_2} & Y + D & \text{(bimolecular reaction)}, \\ 2X + Y & \xrightarrow{k_3} & 3X & \text{(autocatalytic trimolecular reaction)}, \\ X & \xrightarrow{k_4} & E & \text{(monomolecular reaction)}, \end{array}}$$

wobei k_i – Reaktionsgeschwindigkeitskoeffizient (reaction rate coefficient).

Aufgrund des <u>Massenwirkungsgesetzes</u> ergibt sich die folgende Reaktionskinetik (siehe [1], S. 8 ff.), beschrieben durch ein System von Dgl. für die Konzentrationen $c_A = c_A(t)$, $c_B = c_B(t)$, $c_D = c_D(t)$, $c_E = c_E(t)$, $c_X = c_X(t)$ und $c_Y = c_Y(t)$ der einzelnen Substanzen als Fkt. der Zeit t:

$$t \in \boldsymbol{T} = (0, \overset{\overset{\infty\ (!)}{\nearrow}}{T})$$

$$\begin{array}{rcl} c_A' & = & -k_1\, c_A \\ c_B' & = & -k_2\, c_B\, c_X \\ c_D' & = & +k_2\, c_B\, c_X \xleftarrow{c_B,\, c_X} \\ c_E' & = & +k_4\, c_X \xleftarrow{c_X} \\ c_X' & = & +k_1\, c_A - k_2\, c_B\, c_X + (-2k_3\, c_X^2\, c_Y + 3k_3\, c_X^2\, c_Y) - k_4\, c_X \\ c_Y' & = & +k_2\, c_B\, c_X - k_3\, c_X^2\, c_Y \end{array}$$

$\frac{c_X}{c_B}$

+ <u>AB:</u> $c_A(0)$, $c_B(0)$, $c_D(0)$, $c_E(0)$, $c_X(0)$, $c_Y(0)$ geg. $(c_A + c_B + c_D + c_E + c_X + c_Y = 1)$.

<u>Literatur:</u> [1], S. 8 – 14; [5], S. 111 f.

Ü 3.1 Der Oregonator (Zhabotinski–Belousov–Reaktion einer chemischen Oszillation) wird durch das Reaktionsschema

$$\begin{array}{rcl} BrO_3^- + Br^- & \xrightarrow{k_1} & HBrO_2 \\ HBrO_2 + Br^- & \xrightarrow{k_2} & P \\ BrO_3^- + HBrO_2 & \xrightarrow{k_3} & 2\,HBrO_2 + Ce\,(IV) \\ 2\,HBrO_2 & \xrightarrow{k_4} & P \\ Ce\,(IV) & \xrightarrow{k_5} & Br^- \end{array}$$

beschrieben, mit $k_1 = 1.34, k_2 = 1.6 \cdot 10^9, k_3 = 8.0 \cdot 10^3, k_4 = 4.0 \cdot 10^7, k_5 = 1.0$ und geg. AB für die ges. Konzentrationen $c_1 = c_{BrO_3^-}(t)$, $c_2 = c_{Br^-}(t)$, $c_3 = c_{HBrO_2}(t)$, $c_4 = c_P(t)$, $c_5 = c_{Ce(IV)}(t)$.
Stellen Sie das Differentialgleichungssystem für die ges. Konzentrationen $c_1(t)$, $c_2(t)$, $c_3(t)$, $c_4(t)$, $c_5(t)$ auf, und lösen Sie es (später) numerisch mit einem geeigneten (?) Integrationsverfahren für die Anfangskonzentrationsverteilung: $c_1(0) = 0.25$, $c_2(0) = 0.25$, $c_3(0) = 0$, $c_4(0) = 0$, $c_5(0) = 0.5$.

Ü 3.2 Die Robertson–Reaktion (1966) wird durch das folgende Reaktionsschema beschrieben:

$$\begin{array}{rcll} A & \xrightarrow{0.04} & B & \text{(langsame Reaktion)} \\ B + B & \xrightarrow{3 \cdot 10^7} & C + B & \text{(sehr schnell)} \\ B + C & \xrightarrow{10^4} & A + C & \text{(schnell)} \end{array}$$

Stellen Sie das Dgl.–System zur Bestimmung der Konzentrationen $c_A = c_A(t)$, $c_B = c_B(t), c_C = c_C(t)$ auf, und lösen Sie es (später) numerisch mit einem geeigneten (?) Verfahren für die AB: $c_A(0) = 1, c_B(0) = 0, c_C(0) = 0$
(vgl. Willoughby, 1974).

■ **Bsp. 3.2:** Semidiskrete ($\Rightarrow$ Vertikale Linienmethode: siehe Pkt. 2.3.) lineare parabolische A(R)WA:

$$\boxed{\begin{array}{ll} \text{Ges. } \underline{u}_h(t) \in [W_2^1(0,T)]^N : & M_h\, \dot{\underline{u}}_h(t) + K_h(t)\, \underline{u}_h(t) = \underline{f}_h(t), \quad \forall \text{ f.ü. } t \in \boldsymbol{T} = (0,T) \\ & +\text{AB: } M_h\, \underline{u}_h(0) = \underline{g}_h \end{array}}$$

Besonderheiten:

- $\underline{u}_h(t) = [u^{(i)}(t)]_{i \in \omega_h} : u^{(i)}(\cdot) \in W_2^1(\boldsymbol{T})$ (!)
  mit geg. Daten $\underline{f}_h(\cdot) \in [L_2(T)]^N$ und L_∞-Koeffizienten der Matrix $K_h(t)$.
- $h \to 0 \Rightarrow N = N_h = |\omega_h| \to \infty$ (keine fixierte Dimension !!).

Bemerkung: $\exists!$ (Satz 2.18), Integrationsverfahren (σ–gew. DS), Stabilität, Approximation und Fehlerabschätzungen $O(\tau^q + h^p)$ siehe Pkt. 2.3.

■ **Bsp. 3.3:** Semidiskrete ($\Rightarrow$ Vertikale Linienmethode) nichtlineare parabolische A(R)WA (z.B. nichtlineare Wärmeleitprobleme):

Ges. $u \in W_p^1(\boldsymbol{T}; V_0, H) := \{u \in L_p(\boldsymbol{T}, V_0) : \exists\, \dot{u} \in L_q(\boldsymbol{T}, V_0^*), p^{-1} + q^{-1} = 1\}$:
$\dot{u}(\cdot) + A(\cdot)\, u(\cdot) = F(\cdot)$ in $L_q(\boldsymbol{T}, V_0^*)$ + AB : $u(0) = u_0$ in H
mit geg. $F \in L_q(\boldsymbol{T}, V_0^*)$, $u_0 \in H$, $A : W_p^1(\boldsymbol{T}, V_0, H) \longrightarrow L_q(\boldsymbol{T}, V_0^*)$ nichtlinear.

$\Downarrow$ $\leftarrow$ FE–Galerkin–Semidiskretisierung (vertikale Linienmethode)

Ges. $\underline{u}_h(t) \in [W_p^1(0,T)]^{N_h}$: $M_h\, \underline{\dot{u}}_h(t) + K_h(t, \underline{u}_h(t)) = \underline{f}_h(t)$, $\forall$ f.ü. $t \in \boldsymbol{T}$
+ AB: $M_h \underline{u}_h(0) = \underline{g}_h$
oft liegt der quasilineare Fall vor: $K_h(t, \underline{u}_h(t)) = \hat{K}_h(t, \underline{u}_h(t))\, \underline{u}_h(t)$.

$\|$

AWA für System nichtlinearer gew. Dgl. 1. Ordnung:
Ges. $u(\cdot) : \overline{\boldsymbol{T}} \to \mathbb{R}^N : u'(t) = f(t, u(t))$, $t \in \boldsymbol{T}$ mit AB: $u(0) = u_0$ geg.,
die in diesem Kap. behandelt werden (allerdings in Räumen stetig diff. Fkt.)

■ **Bsp. 3.4:** van der Polsche Differentialgleichung ([5], S. 107 – 111)

Die van der Polsche Dgl.

$$y''(t) - \varepsilon(1 - y^2(t))y'(t) + y(t) = 0, \quad t > 0$$
+ AB: $y(0)$, $y'(0)$ geg.

beschreibt Kippschwingungen in einem elektrischen Schaltkreis. Es handelt sich hier um eine nichtlineare Dgl. 2. Ordnung.

Ü 3.3 In [5] S. 110, Figure 16.3, wird die Lösung der van der Polschen Dgl. für $\varepsilon = 10$ und für die konkreten Anfangsbedingungen $y(0) = 0$ und $y'(0) = 0$ gezeigt. Man bestimme die Lösung mit einem geeigneten (?) Integrationsverfahren numerisch.

■ **Bsp. 3.5:** Newtonsche Himmelsmechanik ($mx'' = F(x,x'), F = -\nabla U$, U – Gravitationspotential): Das restringierte Dreikörperproblem (siehe [1], S. 3 – 8 bzw. [5], S. 127 – 129):

Die Bahn eines Satelliten in der Ebene des Erde–Mond–Systems läßt sich durch das folgende System von Dgl. 2. Ordnung beschreiben:

$$\boxed{\begin{array}{l} y_1'' = y_1 + 2y_2' - (1-\mu)\,\dfrac{y_1+\mu}{D_1} - \mu\,\dfrac{y_1-(1-\mu)}{D_2}, \quad t>0, \\ y_2'' = y_2 - 2y_1' - (1-\mu)\,\dfrac{y_2}{D_1} - \mu\,\dfrac{y_2}{D_2}, \quad t>0 \\ +\text{ AB: } y_1(0), y_1'(0), y_2(0), y_2'(0) \text{ geg.}, \\ \text{mit } D_1 = [(y_1+\mu)^2 + y_2^2]^{3/2},\ D_2 = [(y_1-(1-\mu))^2 + y_2^2]^{3/2}, \\ \qquad \mu = 0.012277471. \end{array}}$$

Bei geeigneter Setzung der AB ergeben sich periodische Lösungen !

Ü 3.4 Man bestimme die Lösung $(y_1(t), y_2(t))$ mit einem geeigneten (?) Integrationsverfahren numerisch für die Anfangswerte:

$y_1(0) = 0.994$
$y_1'(0) = 0$
$y_2(0) = 0$
$y_2'(0) = -2.00158510637908252240537862224$

Hinweis: $t_{\text{per}} = 17.0652165601579625588917206249$, $\tau = \dfrac{t_{\text{per}}}{m}$ mit $m \geq 6000$ (RK).

■ **Bsp. 3.6:** Semidiskrete ($\Rightarrow$ Vertikale Linienmethode) lineare bzw. nichtlineare hyperbolische A(R)WA (Schwingungsgleichung, Wellengleichung):

$$\boxed{\begin{array}{ll} \text{Ges. } u(x,t): & (\ddot{u}, v)_0 + a(t; u, v) = < F(t), v > \quad \forall v \in V_0 \quad \forall \text{ f.ü. } t \in \boldsymbol{T} \\ & +\text{ AB: } u(0) = u_0,\ \dot{u}(0) = u_1 \\ & \text{(vgl. auch [17] Numerik II, Pkt. 1.2.3)} \end{array}}$$

$\Big\Downarrow \longleftarrow$ FE–Galerkin–Semidiskretisierung (vertikale Linienmethode)

Ges. $\underline{u}_h(t)$:	$M_h \underline{\ddot{u}}_h(t) + K_h(t; \underline{u}_h(t)) = \underline{f}_h(t), \quad \forall$ f.ü. $t \in \boldsymbol{T}$
	+ AB: $M_h \underline{u}_h(0) = \underline{u}_0,$
	$M_h \underline{\dot{u}}_h(0) = \underline{u}_1,$
wobei	$K_h(t; \underline{u}_h(t)) := \hat{K}_h(t, \underline{u}_h(t))\underline{u}_h(t)$ im quasilinearen Fall und
	$K_h(t; \underline{u}_h(t)) := \hat{K}_h(t)\underline{u}_h(t)$ im linearen Fall.

$\Vert$

AWA für Systeme nichtlinearer gew. Dgl. 2. Ordnung:
Ges. $u(\cdot) : \bar{T} \to I\!R^N$: $\quad u''(t) = f(t, u(t)),\ t \in \boldsymbol{T} = (0, T),$
mit AB: $u(0) = u_0,\ u'(0) = u_1.$

- **Bsp. 3.7:** Evolutionsgleichungen für interne Parameter in Prozessen, die von der Prozeßgeschichte abhängen:

z.B. Elastisch–plastische Fließprobleme in der Festkörpermechanik (vgl. auch [17] Numerik II, Pkt. 3.1.2.2: lineares Elastizitätsproblem $\Rightarrow$ quasistatisch) [12]:

Ges. Verschiebungen $u \in C^1(\boldsymbol{T}, V_0)$,
Spannungen $\sigma \in C^1(\boldsymbol{T}, S)$,
Verfestigungsparameter $\kappa \in C^1(\boldsymbol{T}, H)$:

(i) das verallgemeinerte Kräftegleichgewicht

$$\int_\Omega \dot{\sigma}^T \varepsilon(v)\, dx = < F, v > := \int_\Omega \dot{f}^T v\, dx + \int_{\Gamma_2} \dot{g}^T v\, ds, \quad \forall v \in V_0,\ \ t \in \boldsymbol{T},$$

(ii) die differentiellen Spannungs–Verzerrungsbez. (Evolutionsgl.)

$$\dot{\sigma} = D\dot{\varepsilon} - D\alpha(\sigma, \kappa, \dot{\varepsilon}),$$

(iii) die Evolutionsgleichung $\dot{\kappa} = \beta(\sigma, \kappa, \dot{\varepsilon})$ für die Verfestigungsparameter,

(iv) die geometrischen Verzerrungs–Verschiebungsbeziehungen

$$\dot{\varepsilon} = \varepsilon(\dot{u}) \equiv [\varepsilon_{ij}(\dot{u})]_{ij=\overline{1,3}},\ \ \varepsilon_{ij}(\dot{u}) := \frac{1}{2}\left(\frac{\partial \dot{u}_i}{\partial x_j} + \frac{\partial \dot{u}_j}{\partial x_i}\right) = 0.5\,(\dot{u}_{i,j} + \dot{u}_{j,i}),$$

(v) die AB: $u(x, 0) \overset{\text{z.B.}}{=} 0,\ \ \sigma(x, 0) = 0,\ \ \kappa(x, 0) = 0,\ \ x \in \Omega,$

erfüllt werden.

mit $V_0 = \{v \in V = [H^1(\Omega)]^3 : v = \boldsymbol{O}$ auf $\Gamma_1\}$, $(u, v)_V = \int_\Omega \varepsilon^T(u)\, D\, \varepsilon(v)\, dx$,

$S = \{\sigma \in [L_2(\Omega)]^9 : \sigma_{ij} = \sigma_{ji}\}, \ \ (\sigma, \tau)_S = \int_\Omega \sigma^T D^{-1} \tau\, dx,$

$H = [L_2(\Omega)]^l,\ (\kappa, \lambda)_H = \int_\Omega \kappa\lambda\, dx,\ \ \|\cdot\|_X = (\cdot, \cdot)_X^{0.5},\ \ D$ – Matrix der elast. Konst.

FE-Diskretisierung von (ii) bzw. (iii) $\Rightarrow$ AWA für System gew. Dgl. !

3.2 Formulierungen und analytische Resultate

- **Im weiteren wird von folgender <u>klassischen Formulierung</u> einer <u>AWA</u> <u>für ein System gew. Dgl. 1. Ordnung</u> ausgegangen:**

  (1) Sei $I = \bar{\boldsymbol{T}} = [0,T] \subset I\!R \equiv I\!R^1$ – kompaktes Intervall, $0 < T < \infty$; $u_0 \in I\!R^N$ – geg. Vektor der Startwerte,
  $f : D \underset{(=)}{\subset} I \times I\!R^N \longmapsto I\!R^N$ – geg. stetige Fkt.

  Ges. stetig differenzierbare Fkt. $u : I \longmapsto I\!R^N$, d.h. $u \in C^1(I)$:

  $$\boxed{\begin{array}{l} u'(t) = f(t, u(t)), \quad t \in I = [0,T], \\ \text{AB: } u(0) = u_0, \end{array}}$$

  wobei N als fixiert (?) angenommen wird !

- **<u>Bez.:</u>**

  - $u \in C^1(I) \cong C^1(I, I\!R^N)$ bedeutet für Vektorfkt.
    $u = (u_1(\cdot), u_2(\cdot), ..., u_N(\cdot))^T \in [C^1(I)]^N$, d.h. $u_i(\cdot) \in C^1(I) \quad \forall i = \overline{1,N}$.
  - Analog: $f \in C(D)$ bedeutet $f = (f_1, \ldots, f_N)^T \in [C(D)]^N$.

- **<u>Bemerkung 3.8:</u>** zu verallgemeinerten Formulierungen !

  1. In den Bsp. 3.2 und 3.3 können die Daten, d.h. f, oft nicht mehr als stetig vorausgesetzt werden. Folglich kann <u>nicht</u> $u \in C^1(I)$ erwartet werden. Typischerweise gilt: <u>Bsp. 3.2:</u> $u \in [W_2^1(\boldsymbol{T})]^N$ <u>Bsp. 3.3:</u> $u \in [W_p^1(\boldsymbol{T})]^N$.
     Analoge Aussagen gelten für <u>Bsp. 3.6.</u>
  2. AWA für Operatordgl. (vgl. Bsp. 3.7):
     Ges. $u \in C^1(I,Y) : \dot{u}(t) = F(t, u(t))$ in Y, $\forall t \in I$ mit AB: $u(0) = u_0$ in Y,
     wobei Y ein Banach–Raum ist:
     $\Rightarrow$ Galerkin–Diskr. von $Y \longrightarrow$ (1) (in allen Gauß–Pkt.).

- **<u>Spezialfälle:</u>**

  1. Die Einschränkung auf Systeme gew. Dgl. 1. Ordnung ist <u>nicht</u> wesentlich, denn jedes System gew. Dgl. höherer Ordnung läßt sich auf ein System 1. Ordnung transformieren:

z.B. für allgem. Dgl. (bzw. System von Dgl.) 2. Ordnung

$$\begin{cases} u''(t) = f(t, u(t), u'(t)), \\ \text{mit AB: } u(0) = u_0, u'(0) = u_1 \end{cases}$$

erhält man mit der Setzung

$$u_1(t) = u(t), \quad u_2(t) = u'(t)$$

ein äquivalentes System 1. Ordnung

$$\begin{cases} u_1'(t) = u_2(t), \\ u_2'(t) = f(t, u_1(t), u_2(t)), \\ \text{mit AB: } u_1(0) = u_0, \quad u_2(t) = u_1. \end{cases}$$

2. Das System (1) nennt man (affine) <u>linear</u>, falls

$(1)_{\text{lin.}}$ $\quad f(t, u(t)) = A(t)\, u(t) + f(t)$

$$\underset{\text{stetig}}{A(\cdot) : I \longmapsto} \mathbb{R}^N \times \mathbb{R}^N - N \times N - \text{Matrix.}$$

Bsp. 3.2 (nach M_h^{-1}–Multiplikation) ist ein affine lineares System gew. Dgl. 1. Ordnung.

3. Hängt die rechte Seite der Dgl. <u>nicht</u> von u ab, d.h.

$(1)_{\text{integr.}}$ $\quad u'(t) = f(t),\ t \in I$ mit AB: $u(0) = u_0$,

so läßt sich die exakte Lsg. mit Hilfe des <u>Integrals</u>

$$u(t) = u_0 + \int_0^t f(s)\, ds$$

darstellen. Die <u>numerische Lsg. des AWP</u> führt dann auf die Fragestellung der <u>numerischen Integration</u> !

4. Hängt die rechte Seite der Dgl. <u>nicht explizit</u> von t ab, d. h.

$(1)_{\text{autonom}}$ $\quad u'(t) = f(u(t)),\ t \in I$ mit AB: $u(0) = u_0$,

dann nennt man die Dgl. (System v. Dgl.) <u>autonom</u>.

Die Dgl. (1) $u'(t) = f(t, u(t))$ kann durch die Setzung

$$u_1(t) = t, \quad u_2(t) = u(t)$$

immer in das äquivalente autonome System

$$\begin{cases} u_1'(t) & = 1 \\ u_2'(t) & = f(u_1(t), u_2(t)) \\ + \text{ AB:} & u_1(0) = 0 \\ & u_2(0) = u(0) \equiv u_0 \end{cases}$$

transformiert werden.

■ $\boxed{\textbf{Ü 3.5}}$ Transformieren Sie
Bsp. 3.1 – 3.3 $\longrightarrow (1)_{\text{autonom}}$,
Bsp. 3.4 – 3.6 $\longrightarrow (1) \longrightarrow (1)_{\text{autonom}}$.

■ Formulierung als Operatorgleichung:

- Sei
  $X = C^1(I, \mathbb{R}^N) \cong [C^1(I)]^N$ mit geeignet definierter Norm $\|\cdot\|_X$,

  z.B. $\|u\|_X := \max\limits_{t \in I} |u(t)| + \max\limits_{t \in I} |u'(t)|$,

  mit $|v| := \max\limits_{i=\overline{1,N}} |v_i|$ oder

  $|v|^2 := \sum\limits_{i=1}^{N} |v_i|^2$,

  $Y = C(I, \mathbb{R}^N) \times \mathbb{R}^N$ mit geeignet definierter Norm $\|\cdot\|_Y$,

  $$F : X \mapsto Y \quad : \quad F(u) := \begin{bmatrix} u'(\cdot) - f(\cdot, u(\cdot)) \\ u(0) - u_0 \end{bmatrix}.$$

  Das AWP (1) wird dann äquivalent zur Operatorgleichung

  $$(2) \qquad \boxed{\text{Ges. } u \in X \ : \ F(u) = 0 \text{ in } Y},$$

  und es können die Techniken aus [16] Numerik I (Kap. 6) angewendet werden.

- Eine andere Möglichkeit zur Formulierung des AWP (1) als Operatorgleichung in einem B-Raum basiert auf der Integralbeziehung für Lsg. von (1):

  $$(3) \qquad u(t_2) = u(t_1) + \int_{t_1}^{t_2} u'(t)\, dt = u(t_1) + \int_{t_1}^{t_2} f(t, u(t))\, dt$$

  für beliebige $t_1, t_2 \in I = \bar{\boldsymbol{T}} = [0, T]$.

  Mit der speziellen Wahl $t_1 = 0$ und $t_2 = t$ erhält man aus AB und (3):

  $$(4) \qquad u(t) = u_0 + \int_0^t f(s, u(s))\, ds, \quad \forall t \in I.$$

  Also ist die Lösung von (1) offenbar äquivalent zur Lösung der Fixpunktgleichung (= Volterrasche Integralgleichung)

  $$(4) \qquad \boxed{\text{Ges. } u \in X \ : \ u = G(u) \text{ in } X,}$$

  mit $G : X := C(I, \mathbb{R}^N) \longrightarrow X$:

  $$(5) \qquad G(u) \equiv G(u)(t) := u_0 + \int_0^t f(s, u(s))\, ds.$$

  Existenz– und Eindeutigkeitsaussagen liefern eine Verallgemeinerung des Satzes von Picard–Lindelöf auf der Basis des Banachschen Fixpunktsatzes I.6.1 [16] ($\rightarrow$ Satz 3.1: $\exists$!) und des Satzes von Peano auf der Basis des Fixpunktsatzes von Schauder ($\rightarrow \exists$: siehe [22], S. 26 – 27, 35 – 36).

Satz 3.1: (Verallgemeinerter Satz v. Picard–Lindelöf)

Vor.: 1) $f : \tilde{D} \longrightarrow Y := \mathbb{R}^N$ (B–Raum) – stetig,
wobei $\tilde{D} := \{(t,v) \in \mathbb{R} \times Y : |t| \le a \equiv T, \|v - u_0\|_Y \le b\}$,
$a, b \in (0, \infty)$,

2) $\|f(t,v) - f(t,\bar{v})\|_Y \le L\,\|v - \bar{v}\|_Y$, $\|f(t,v)\|_Y \le K$ auf $\tilde{D}$
mit festen Konstanten $L, K \in [0, \infty]$,

3) $c = \text{const.} > 0 : \qquad 0 < c < a, \;\; c\,K < b, \;\; c\,L < 1.$

Bh.: 1) Dann besitzt

$(\tilde{1})$ $u'(t) = f(t, u(t))$, $t \in [-c, c]$, AB: $u(0) = u_0$

bei festem $u_0 \in Y$ genau eine Lösung $u(\cdot)$ in der Kugel

$M := \{v \in X := C([-c,+c], Y) : \|v - u_0\|_X \le b\}$

mit $\|v\|_X := \max\limits_{t \in [-c,+c]} \|v(t)\|_Y$.
$u(\cdot)$ ist auf $[-c, +c]$ stetig differenzierbar.

2) Die Fixpunktiteration
$u_{n+1}(t) = u_0 + \int\limits_0^t f(s, u_n(s))\, ds$, $u_0(t) := u_0$,
konvergiert auf $[-c, +c]$ gleichmäßig gegen $u(\cdot)$.

3) Die Lsg. $u(\cdot)$ hängt in der Norm auf X stetig vom Anfangswert
$u_0 \in Y$ ab.

4) Ist f stetig auf $\tilde{D}$ für alle $b > 0$ und gilt Vor. 2) für alle $b > 0$ mit
einem einheitlichen L, dann besitzt $(\tilde{1})$ genau eine Lsg. auf
$C([-a, +a], Y)$, folglich auch (1) auf $C(I, \mathbb{R}^N)$.
Vor. 3) entfällt hier !

Beweis:

1. Zunächst für $N = 1$, d.h. $Y = \mathbb{R}^1$ (mms)
   Hinweis: Bh. 4) $M = X = C[-a, a]$, $\|v\|_X := \max\limits_{|t| \le a} |v(t)| e^{-L|t|}$.

2. Dann übertragen auf $Y = \mathbb{R}^N$ mit $N > 1$ und B–Raum Y.

Literatur: [22] Zeidler E.: Vorlesungen über nichtlineare Funktionalanalysis I, Fixpunktsätze.
Teubner–Texte zur Mathematik, Teubner–Verlag, Leipzig, 1976:
$Y = \mathbb{R}^1$: S. 19 – 21 und Y – B–Raum: S. 34 – 35.

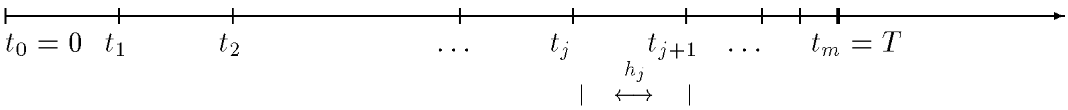
3.3 Einschrittverfahren (= zweischichtig: t_j, t_{j+1})

3.3.1 Das Eulersche Polygonenzugverfahren (EPZV)

- **Eulersches Polygonenzugverfahren:**
  = explizites Eulerverfahren ($\widehat{=}\ \sigma = 0$ in Kap. 2)
  = Euler vorwärts (vorwärtige Differenzen)
  = linksseitige Rechteckregel

- **Unterteilen Zeitintervall $I = \overline{T} = [0, T]$ i. allg. ungleichmäßig (z.B. adaptiv durch Schrittweitensteuerung) in m Teilintervalle:**

  — Schrittweitensteuerung (?) $\longrightarrow$



$0 = t_0 < t_1 < \ldots < t_j < t_{j+1} < \ldots < t_{m-1} < t_m = T$

$\uparrow$ Gitterpkt. (bei t_j)

$I_h := \{t_0, t_1, \ldots, t_m\} \equiv \bar{\omega}_h$	–	(Zeit-)Gitter,
$h_j = t_{j+1} - t_j$	–	Schrittweite im Gitterpkt. t_j $\ (h_j = \tau_j\ (\uparrow))$
$h = \vec{h} = (h_0, h_1, \ldots, h_{m-1})$	–	Schrittweitenvektor,
$h = T/m = h_j \quad \forall j = \overline{0, m-1}$	–	Konstante Schrittweite bei äquidistanter Unterteilung $(h = \tau\ (\uparrow))$

$\downarrow$

$h = |\vec{h}| := \max\limits_{j=\overline{0,m-1}} h_j$

$\uparrow$

der Einfachheit halber

Frühere Bezeichnung: $\tau_j = h_j,\ \tau = h,\ \bar{\omega}_\tau = I_h, \ldots (\uparrow)$

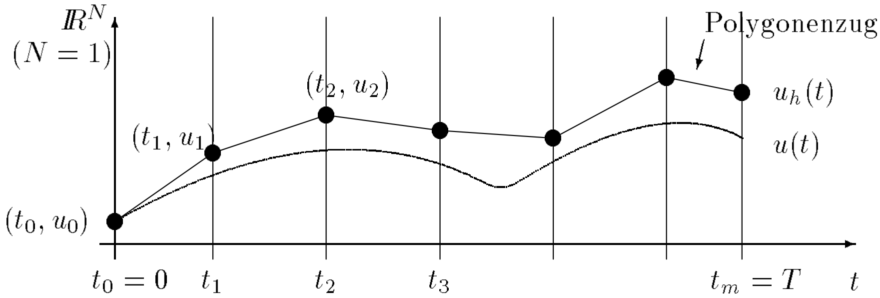
- **Methoden zur Herleitung des Eulerschen Polygonenzugverfahrens:**

  1. Taylorentwicklung ($\Rightarrow$ Polygonenzug):

  $$\underset{\text{Lsg. (1)}}{u(t)} \approx u(t_0) + (t - t_0)\, u'(t_0) \overset{(1)}{=} u(t_0) + (t - t_0)\, f(t_0, u(t_0))$$

$$
\begin{array}{lll}
\Rightarrow \quad t = t_0 = 0: & (t_0, u_0) \qquad \text{AW: } u(0) = u_0 \\
& \downarrow \\
t \in [t_0, t_1]: & u_h(t) = u_0 + (t - t_0)\, f(t_0, u_0) \\
& \downarrow \\
t = t_1: & u_1 = u_h(t_1) \approx u(t_1) \\
& (t_1, u_1) \\
& \downarrow\longleftarrow \ \text{Idee wiederholen} \\
t \in [t_1, t_2]: & u_h(t) = u_1 + (t - t_1)\, f(t_1, u_1) \\
t = t_2: & u_2 = u_h(t_2) \approx u(t_2) \\
& (t_2, u_2) \\
& \downarrow \ \text{usw.} \\
& \underline{\text{Allgemeine Vorschrift}}
\end{array}
$$

$$
(6) \qquad \boxed{\begin{array}{l} u_h(t) = u_j + (t - t_j)\, f(t_j, u_j), \quad t \in [t_j, t_{j+1}] \\ \hline u_{j+1} \equiv u_h(t_{j+1}) = u_j + h_j\, f(t_j, u_j) \\ j = 0, 1, \ldots, m-1; \text{ AB: } u_0 \text{ geg.} \end{array}}
$$



Btr. $u_h : I_h \longrightarrow I\!R^N$ – als Gitterfkt. $u_h \in X_h = \{v_h : I_h \longrightarrow I\!R^N\}$.

Mit dieser Schreibweise läßt sich das EPZV (6) als diskretes Ersatzproblem (Näherungsgleichung)

$$
(2)_h \qquad \boxed{\text{Ges. } u_h \in X_h : \ F_h(u_h) = 0 \text{ in } Y_h}
$$

für (2) $F(u) = 0$ interpretieren, wobei

$Y_h = X_h$ zumindest mengenmäßig, aber $\|\cdot\|_{X_h}, \ \|\cdot\|_{Y_h}$ (?),

$$
F_h(v_h)(t_{j+1}) := \begin{cases} D_h\, v_h(t_j) - f(t_j, v_h(t_j)), & j = 0, 1, \ldots, m-1, \\ v_h(t_0) - u_0, & j = -1, \end{cases}
$$

mit

$$
D_h\, v_h(t_j) = D_h\, v_j = v_{t,j} := \frac{v_{j+1} - v_j}{h_j}.
$$

2. Vorwärtiges (Explizites) Differenzenverfahren:

$$(1) \qquad u'(t) = f(t, u(t)) \qquad + \underline{\text{AB:}}\ u(0) = u_0$$

$$\Big|$$

ersetzen durch
vorwärtigen Differenzquotienten

$$\downarrow$$

$$(1)_h \qquad \frac{u(t+h) - u(t)}{h} \approx f(t, u(t)), \qquad t = t_j, \quad j = 0, 1, \ldots, m-1$$

$$(6) \qquad \Rightarrow t = t_j : h \longmapsto h_j, \ u \longmapsto u_h, \ u_h(t_j) = u_j, \ \approx \longmapsto = \, !$$

3. Linksseitige Rechteckregel:
Setzen in (3) $t_1 = t_j$ und $t_2 = t_{j+1} = t_j + h_j \ \Rightarrow$

$$u(t_{j+1}) = u(t_j) + \int_{t_j}^{t_{j+1}} f(s, u(s))\, ds$$

$$\approx u(t_j) + h_j\, f(t_j, u(t_j)) \quad \Rightarrow (6).$$

$$\uparrow$$

$\int \mapsto$ linksseitige Rechteckregel

■ Die Sätze von Cauchy und Peano zur Konvergenz des EPZV:

Satz 3.2: (Cauchy)

Vor.:	1) $f : D \mapsto Y = \mathbb{R}^N$ – stetig, wobei $\lvert\cdot\rvert := \lVert\cdot\rVert_{Y=\mathbb{R}^N}$, $D = \{(t,v) : t \in [0,T], \lvert v - u_0\rvert \le b\}$, $b \in (0,\infty)$ fix. ($\Rightarrow \forall b$).
	2) $\lvert f(t,v) - f(t,\bar{v})\rvert \le L\lvert v - \bar{v}\rvert$, $f(t,v) \le K$ auf D mit festen Konstanten $L, K \in [0,\infty)$.
	3) $K \cdot T \le b$ (fällt weg, falls 1) $\forall b > 0$ gilt, d.h. $b = \infty$, „ $\le$ “ $\mapsto$ „ $<$ “).
Bh.:	Dann konvergiert $u_h(\cdot)$ aus (6) gleichmäßig gegen eine stetig differenzierbare Fkt. $u(\cdot)$ für $h \to 0$. Die Fkt. $u(\cdot)$ ist eindeutige Lsg. des AWP (1) in D.

Beweis: siehe Literatur [5].

Bemerkung: Das EPZV liefert ebenfalls konstruktive Möglichkeit zum Beweis des Satzes 3.1 von Picard–Lindelöf über $\exists + \ !$ #

Satz 3.3: (Peano)

> Vor.: 1) $f : D := \{(t,v) : t \in [0,T],\ \ |v-u_0| \le b\} \to I\!R^N$ – stetig, $b \in (0,\infty)$ fix.
>
> 2) $|f(t,v)| \le K$ auf D mit festen Konstanten $K \in [0,\infty)$.
>
> 3) $K \cdot T \le b$ (fällt weg, falls 1) $\forall b > 0$ gilt, d.h. $b = \infty$, „ $\le$ “ $\mapsto$ „ $<$ “).
>
> Bh.: Dann konvergiert $u_h(\cdot)$ aus (6) gleichmäßig gegen eine Lsg. $u(\cdot)$ des AWP (1) für $h \to 0$.

Beweis: siehe Literatur [5].

Bemerkung: Das EPZV liefert ebenfalls konstruktive Möglichkeit zum Beweis der $\exists$ der Lsg. des AWP (1), falls f nicht Lipschitz–stetig ist. #

■ Fehlerbegriffe:

- globaler Diskretisierungsfehler (globaler Fehler):

$$(7) \qquad e_h(t) := u(t) - u_h(t) \begin{cases} \nearrow \ \text{SL: } t \in I_h = \{t_0, t_1, \ldots, t_m\} \\ \searrow \ \text{NL: } t \in I = [0,T] \end{cases}$$

Satz 3.2 und 3.3: $e_h \to 0$ in $C(I, I\!R^N)$ für $h \equiv |h| \to 0$,
d.h. $\|e_h\|_{C(I,I\!R^N)} := \max\limits_{t \in I} \|e_h\|_{I\!R^N} \underset{h\to 0}{\longrightarrow} 0$.

Btr. im weiteren

$$e_h(\cdot) : I_h \longrightarrow I\!R^N, \qquad e_h \in X_h$$

als Gitterfunktion, d.h. als Element $X_h := \{v_h : I_h \to I\!R^N\}$.

Wir nennen das EPZV konvergent ($\widehat{=}$ diskrete Konvergenz), falls

$$\|e_h\|_{X_h} \underset{h \to 0}{\longrightarrow} 0 \ \text{ für } \ h \to 0,$$

wobei $\|\cdot\|_{X_h}$ geeignet definierte Norm in X_h, z.B.

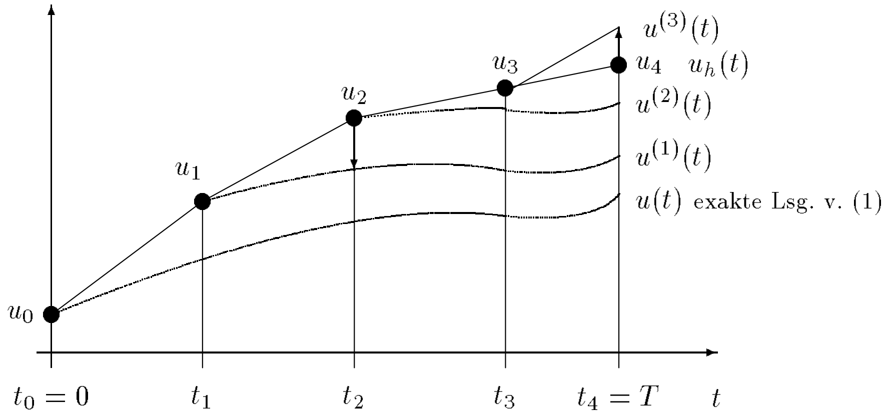
$$\|v_h\|_{X_h} := \max_{j=0,1,\ldots,m} |v_h(t_j)| + \max_{j=0,1,\ldots,m-1} |D_h\, v_h(t_j)|,$$

d.h. $X_h = C^1(I_h, I\!R^N) \,\widehat{=}\, X = C^1(I, I\!R^N)$.

Nach Sätze 3.2 u. 3.3 konvergiert das EPZV in $X_h = C(I_h, I\!R^N)$.

- Lokale Diskretisierungsfehler (lokale Fehler):

= Unterschied zwischen exakter Lsg. der Dgl. und Näherungslösung (bzw. SL) nach einem Schritt des Verfahrens bei gleichen AB im Pkt. $t = t_j$, $j = 0, 1, \ldots, m-1$:



- EPZV $t = t_1$: Lokaler Fehler = globaler Fehler = $u(t_1) - u_h(t_1)$

- Btr. AWP: $\dot{u}^{(1)} = f(t, u^{(1)}(t)), \;\; t \in [t_1, T]$ mit AB: $u^{(1)}(t_1) = u_1$.
  Dann beschreibt die Differenz
  $u^{(0)}(t) - u^{(1)}(t)$ mit $u^{(0)}(t) = u(t)$
  die Fortpflanzung des lokalen Fehlers $u(t_1) - u_h(t_1)$ nach dem ersten Schritt des EPZV durch die Dgl. Auch bei exakter Integration der Dgl. bleibt ab t_1 dieser Fehler erhalten !

- Dem nächsten Schritt des EPZV läßt sich der lokale Fehler
  $u^{(1)}(t_2) - u_h(t_2)$
  zuordnen, dessen Fortpflanzung analog durch die Lsg.
  $$u^{(2)}(t) \text{ des AWP: } \begin{cases} \dot{u}^{(2)}(t) = f(t, u^{(2)}(t)), \;\; t \in [t_2, T] \\ \text{mit AB: } u^{(2)}(t_2) = u_2 \end{cases}$$
  beschrieben wird, u.s.w. (siehe Abb.).

Fazit: Der globale Fehler setzt sich aus der Fortpflanzung der oben eingeführten lokalen Fehler durch die Dgl. zusammen (siehe Abb.).

Abschätzung des lokalen Fehlers für EPZV:

- $v(t) = u^{(j)}(t) : v'(t) = f(t, v(t)), \;\; t \in [t_j, T]$
  AB: $v(t_j) = u_j$

- Lokaler Fehler: $h = h_j$

(8)
$$\begin{aligned} u^{(j)}(t_{j+1}) - u_h(t_{j+1}) &= v(t_j + h) - u_h(t_j + h) = \\ &= v(t_j) + hv'(t_j) + \frac{h^2}{2} v''(t_j) + \ldots - u_h(t_j + h) \\ &= v(t_j) + hf(t_j, v(t_j)) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial v} f\right)(t_j, v(t_j)) + \ldots - (u_j + hf(t_j, u_j)) \\ &= \frac{h^2}{2}(f_t + f_v f)(t_j, u_j) + o(h^2) = O(h^2). \end{aligned}$$

Man spricht in diesem Fall von einer Methode der Konsistenzordnung 1 (d.h. $\widehat{=}$ lokaler Fehler $= O(h^2)$!).
Falls die Fortpflanzung der lokalen Fehler durch die Dgl. „stabil“ erfolgt, darf erwartet werden, daß der globale Fehler höchstens von der Größenordnung

$$O(h_0^2) + O(h_1^2) + \ldots + O(h_{m-1}^2) = O(h) \text{ ist !}$$

Um in diesem Sinne zumindest mit Konvergenz rechnen zu dürfen, muß für die lokalen Fehler gelten

$$u^{(j)}(t_j + h) - u_h(t_j + h) = o(h).$$

Man spricht dann von einer konsistenten Methode.

- Lokale Abschneidefehler (= lokaler Approximationsfehler ψ ($\uparrow$)):

engl.: = local truncation error:

$$(9) \qquad \underset{\substack{\uparrow \\ (2)_h}}{\tau_h(t_{j+1}) = F_h}\underset{\substack{\uparrow \\ \text{Lsg. v. (1)}}}{(u)}(t_{j+1}) \equiv \begin{cases} \frac{u(t_{j+1})-u(t_j)}{h} - f(t_j, u(t_j)) & , j = \overline{0, m-1}, \\ 0 & , j = -1, \end{cases}$$

d.h. der Abschneidefehler mißt, inwieweit die exakte Lsg. $u(\cdot)$ des AWP (1) die Näherungsgl. $(2)_h$ erfüllt.

Bemerkung: $\tau_h(t_{j+1}) = \psi_h(u)(t_{j+1}) = F_h(u)(t_{j+1}) - F(u)(t_{j+1})$ ($\uparrow$)

- Zusammenhang: lokaler Abschneidefehler $\leftrightarrow$ lokaler Fehler

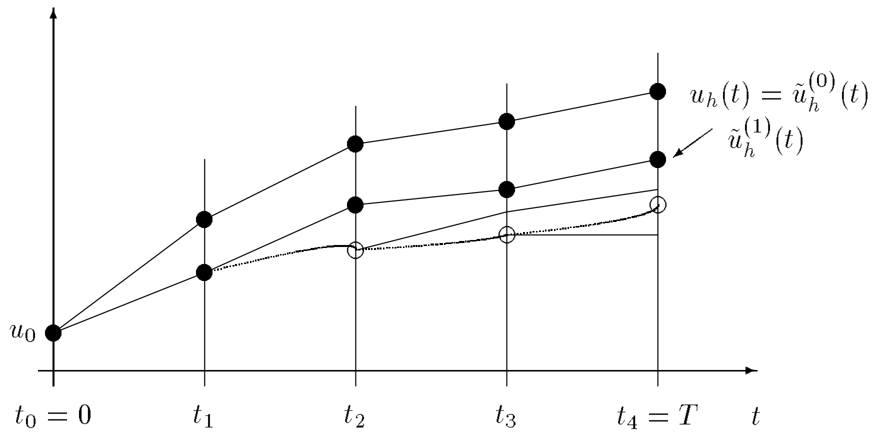
$$(10) \qquad \tau_h(t_{j+1}) = \frac{1}{h} \underbrace{\{u(t_{j+1}) - [u(t_j) + hf(t_j, u(t_j))]\}}_{\substack{= \text{ lokaler Fehler, wenn man EPZV in } t = t_j \\ \text{mit exakter Lsg. der Dgl. } u(t_j) \text{ starten würde !}}}$$

Wegen (8), $u_j \mapsto u(t_j)$, gilt: $\tau_h = O(h)$.

Im speziellen folgt

$$(11) \qquad \|\tau_h\|_{Y_h} := \max_{j=0,1,\ldots,m} |\tau_h(t_j)| = O(h) \underset{h\to 0}{\longrightarrow} 0.$$

Dies ist eine weitere Möglichkeit, die Konsistenz bzw. die Konsistenzordnung 1 zu definieren:



Der globale Fehler kann auch aus der Fortpflanzung von $u(t_{j+1}) - \tilde{u}_h^{(j)}(t_{j+1})$ durch EPZV bestimmt werden.

■ $\boxed{\textbf{Ü 3.6}}$ Man beweise den folgenden Konvergenzsatz !

Satz 3.4:

Vor.: Es gelten die Voraussetzungen des Satzes 3.2 mit $b = \infty$.

Bh.: Dann gilt die folgende Fehlerabschätzung für das EPZV auf äquidistantem Gitter $I_h = \{t_j = jh,\ j = \overline{0,m},\ h = T/m\}$:

$$|u(t_j) - u_h(t_j)| \le e^{Lt_j} \left\{ |e_0| + \frac{\tau}{L} \right\}, \quad j = 1, 2, \ldots, m,$$

wobei L – Lipschitz-Konstante,

$e_0 = u(t_0) - u_h(t_0) = u_0 - u_0 = 0$, d.h. im betrachteten Fall ist der Anfangsfehler 0, $\tau = \max\limits_{j=\overline{1,m}} |\tau_j|$,

$\tau_j = \tau_h(t_j)$ – lokale Abschneidefehler.

Hinweise zum Beweis: (siehe $\boxed{\text{P VII}}$):

1) Schreiben Sie unter Benutzung der Darstellung (τ_{j+1} nach $u(t_{j+1})$ auflösen)

$$(*) \quad u(t_{j+1}) = u(t_j) + hf(t_j, u(t_j)) + h \underbrace{\left[\frac{u(t_{j+1}) - u(t_j)}{h} - f(t_j, u(t_j)) \right]}_{=:\ \tau_h(t_{j+1}) = \tau_{j+1}}$$

und des EPZV

$$(**) \quad u_{j+1} = u_j + hf(t_j, u_j)$$

eine Rekursionsbeziehung ((∗) – (∗∗) !) für den Fehler

$$e_{j+1} = u(t_{j+1}) - u_h(t_{j+1}) = u(t_{j+1}) - u_{j+1}$$

auf, und schätzen Sie diese ab.

2) Verwenden Sie dabei die elementare Beziehung (mms)
$(1 + hL)^{j+1} \le e^{(j+1)hL} = e^{Lt_{j+1}} \le e^{LT}$.

q.e.d.

■ **Folgerung 3.5:**

Vor.:	1) Es gelten die Voraussetzungen von Satz 3.4,
	2) $f,\ \dfrac{\partial f}{\partial t},\ \dfrac{\partial f}{\partial v}$ seien auf D beschränkt.
Bh.:	Dann gilt die folgende Fehlerabschätzung für das EPZV auf äquidistantem Gitter $I_h = \{t_j = jh\}$:
	$\lvert u(t_j) - u_h(t_j)\rvert \le e^{Lt_j}\dfrac{c}{L}\,h.$

Beweis: folgt sofort aus Satz 3.4 und den Fakten

- $e_0 = 0$
- (11) $\|\tau_h\|_{Y_h} := \max\limits_{j=0,\ldots,m} \lvert\tau_h(t_j)\rvert \underset{(8),\,(10),\,\text{Vor. }2)}{\underset{\uparrow}{\le}} c\,h$

q.e.d.

■ **Siehe auch Pkt. 3.3.4:** „Allgemeine Konvergenztheorie für Einschrittverfahren" !

3.3.2 Explizite Runge–Kutta–Verfahren (ERKV)

■ **Motivation:**

- EPZV hat Konsistenzordnung 1 $\Rightarrow \|u(\cdot) - u_h(\cdot)\|_{C(I_h,\mathbb{R}^N)} = O(h)$!
- Es sind sehr kleine Schrittweiten $\{h_j\}$ notwendig, um genaue Resultate zu erzielen !
- Deshalb: Ges. sind Konstruktionsprinzipien für Verfahren höherer Konsistenzordnung !

3.3.2.1 Konstruktionsprinzip

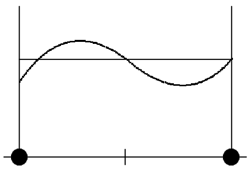
■ **Ausgangspunkt:** = dabei jene Interpretation des EPZV, die auf der linksseitigen Rechteckregel für

$$\int_t^{t+h} f(s,u(s))\,ds \approx h\,f(t,u(t))$$

beruht, d.h. $\int_t^{t+h} f(s,u(s))\,ds$ durch bessere Quadratformel (QF) ersetzen !

■ **<u>Eine genauere QF $\overset{\text{z.B.}}{=}$ Mittelpunktsregel (Gauß 1):</u>**

$$\int_t^{t+h} f(s,u(s))\,ds \approx h\,f\left(t+\frac{h}{2},\, \underset{\substack{\uparrow\\ \textbf{zunächst unbekannt !}}}{u\left(t+\frac{h}{2}\right)}\right)$$



t $\qquad$ $t+h$

t_j $\qquad$ $t_{j+1} = t_j + h_j$

<u>Idee:</u> Approximieren $u(t+0.5\,h)$ durch EPZV:

$$\boxed{u\left(t+\tfrac{h}{2}\right) \approx u(t) + \tfrac{h}{2}\,f(t,u(t))}$$

Als Resultat erhält man das folgende Verfahren (Runge, 1895):

$$\boxed{\begin{aligned} K_1 &= f(t,u(t)) \\ K_2 &= f\left(t+\tfrac{h}{2},\, u+\tfrac{h}{2}\,K_1\right) \\ u_h(t+h) &= u + h\,K_2 \end{aligned}}$$

bzw. vollständig aufgeschrieben:

$$\boxed{\begin{array}{l} \text{AB: } u_h(t_0) \equiv u_h(0) = u_0 \\ j = 0,1,\ldots,m-1 \\ K_{1,j} = f(t_j, u_j) \\ K_{2,j} = f(t_j + 0.5\,h_j,\ u_j + 0.5\,h_j\,K_{1,j}) \\ u_{j+1} \equiv u_h(t_{j+1}) = u_j + h_j\,K_{2,j} \end{array}}$$

Dieses Verfahren wird

<u>verbesserte Euler–Methode</u> bzw. <u>Euler–Cauchy–Methode</u>

genannt.

Durch <u>Taylorentwicklung</u> des <u>lokalen Fehlers</u> läßt sich nun leicht die <u>Ordnung</u> des Verfahrens bestimmen:

- $$\begin{aligned} u_h(t+h) &= u(t) + h\,f\left(t+\tfrac{h}{2},\, u(t)+\tfrac{h}{2}\,f(t,u(t))\right) \\ &= u + h\Big[f(t,u) + f_t\tfrac{h}{2} + f_u\cdot\tfrac{h}{2}f + \\ &\qquad + \tfrac{1}{2!}\left(f_{tt}\left(\tfrac{h}{2}\right)^2 + 2f_{tu}\cdot\tfrac{h}{2}\cdot\tfrac{h}{2}\,f + f_{uu}\left(\tfrac{h}{2}f\right)^2\right) + \ldots\Big] \\ &= u + h\,f(t,u) + \tfrac{h^2}{2}(f_t + f_u f)(t,u) + \tfrac{h^3}{8}(f_{tt} + 2f_{tu}f + f_{uu}f^2)(u,t) + \ldots \end{aligned}$$

  <u>Bez.:</u> $f_t = \frac{\partial f}{\partial t}$, $f_u = \frac{\partial f}{\partial u}$ usw.

- $u(t+h) \quad = u(t) + h \cdot u'(t) + \frac{h^2}{2}\, u''(t) + \frac{h^3}{6}\, u'''(t) + \ldots$

$$\overset{\text{Dgl.}\, u'(t)=f(t,u(t))}{\downarrow}$$

$$(12) \qquad = u(t) + h\, f(t,u) + \tfrac{h^2}{2}(f_t + f_u\, f)(t,u) +$$
$$+ \tfrac{h^3}{6}(f_{tt} + f_{tu} f + f_{ut} f + f_{uu} f^2 + f_u f_t + f_u f_u f)(t,u) + \ldots$$

- $u(t+h) - u_h(t+h) =$

$= \frac{h^3}{24}[4f_{tt} + 8f_{tu}f + 4f_{uu}f^2 + 4f_u f_t + 4f_u^2 f - 3f_{tt} - 6f_{tu}f - 3f_{uu}f^2](u,t) + \ldots$

$= \frac{h^3}{24}[f_{tt} + 2f_{tu}f + f_{uu}f^2 + 4(f_u f_t + f_u^2 f)](u,t) + \ldots$

$= O(h^3)$

d.h. Verfahren hat die Ordnung 2 !

■ Verallgemeinerung dieses Konstruktionsprinzips:

- $$\underbrace{\int_t^{t+h} f(s,u(s))\, ds \approx h \sum_{i=1}^{l} b_i}_{\uparrow\ l\text{–stufige QF}}\, f(t + c_i h, \underbrace{u(t + c_i h)}_{\uparrow\ \text{unbekannt !}})$$

  wobei $\{b_i\}, \{c_i\} \uparrow$ – noch frei wählbar mit $c_1 = 0$.

- Anstelle der unbekannten Fkt–Werte $u(t + c_i h)$ werden Näherungen

  $g_i \approx u(t + c_i h)$

  rekursiv durch $(i-1)$-stufige QF berechnet:

  $g_1 = u, \qquad (c_1 = 0)$
  $g_2 = u + h\, a_{21} f(t, g_1),$
  $g_3 = u + h[a_{31} f(t, g_1) + a_{32} f(t + c_2 h, g_2)],$
  $\vdots$
  $g_l = u + h[a_{l1} f(t, g_1) + a_{l2} f(t + c_2 h, g_2) + \ldots + a_{l,l-1} f(t + c_{l-1} h, g_{l-1})].$

- Setzt man

  $K_i = f(t + c_i h, g_i),$

  so erhält man die Darstellung

  $K_1 = f(t, u),$
  $K_2 = f(t + c_2 h, u + h a_{21} K_1),$
  $K_3 = f(t + c_3 h, u + h[a_{31} K_1 + a_{32} K_2]),$
  $\vdots$
  $K_l = f(t + c_l h, u + h[a_{l1} K_1 + a_{l2} K_2 + \ldots + a_{l,l-1} K_{l-1}]).$

- Die nächste Näherung hat dann die Form

$$\begin{aligned} u_h(t+h) &= u(t) + h(b_1 K_1 + b_2 K_2 + \ldots + b_l K_l). \\ &= u(t) + h \sum_{i=1}^{l} b_i f(t + c_i h, g_i). \end{aligned}$$

Man nennt diese Methode ein

l–stufiges explizites Runge–Kutta–Verfahren/Formel

und ordnet dieser Methode das folgende Tableau zu, durch das die Methode eindeutig beschrieben wird:

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & & & \\ c_l & a_{l1} & a_{l2} & \cdots & a_{l,l-1} & \\ \hline & b_1 & b_2 & \cdots & b_{l-1} & b_l \end{array}$$

■ Beispiele:

(a) EPZV: $u_h(t+h) = u + hb_1K_1, \quad K_1 = f(t,u)$
d.h. EPZV = 1-stufiges RK–Verfahren mit dem Tableau

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

(b) Verbesserte Euler–Methode:

$$\begin{aligned} u_h(t+h) = \quad & u + h(b_1 K_1 + b_2 K_2) \\ & K_1 = f(t,u),\ K_2 = f(t + \tfrac{1}{2}h, u + \tfrac{1}{2}h\,K_1),\ \ b_1 = 0,\ \ b_2 = 1, \end{aligned}$$

d.h. VEM = 2-stufiges RK–Verfahren mit dem Tableau

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

■ Koeffizienten werden aus gewünschter Konsistenzordnung bestimmt !

3.3.2.2 Konsistenzordnung der expliziten Runge–Kutta–Formeln

■ **Definition 3.6:**

> Eine Runge-Kutta-Formel hat die Konsistenzordnung p, falls
>
> (13) $$u(t+h) - u_h(t+h) = O(h^{p+1}),$$
>
> wobei $u(t+h)$: $u'(s) = f(s, u(s)),\ \ s \in [t, t+h]$
>
> $u(t)$ geg. (! $u_h(t)$ geg. anstelle von $u(t)$)
>
> $\updownarrow\longleftarrow$ Start mit gleichen Werten zum Zeitpunkt t.
>
> $u_h(t+h) = \ u(t) + h\sum\limits_{i=1}^{l} b_i K_i$ mit $K_i = f(t + c_i h, g_i)$ ($\uparrow$)

Beispiele:

(a) EPZV = ein 1-stufiges ERKV der Ordnung 1.

(b) Die verbesserte Euler–Methode = ein 2-stufiges ERKV der Ordnung 2.

■ **Ü 3.7** Verwendet man anstelle der Mittelpunktsregel die Trapezregel zur Berechnung des Integrals $\int\limits_t^{t+h} f(s, u(s))\,ds$, so erhält man das Verfahren von Heun:

$$\int\limits_t^{t+h} f(s, u(s))\,ds \overset{TR}{\approx} \frac{h}{2}[f(t, u(t)) + f(t+h, \underset{\substack{\uparrow\\ \text{EPZV}}}{u(t+h)})]$$

$$\boxed{\begin{aligned} K_1 &= f(t, u(t)) \\ K_2 &= f(t+h, u + h\,K_1) \\ u_h(t+h) &= u + \tfrac{h}{2}K_1 + \tfrac{h}{2}K_2 \end{aligned}} \qquad \boxed{u(t+h) \approx u(t) + hf(t, u(t))}$$

Damit ergibt sich für das Verfahren von Heun das folgende Tableau:

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

d.h. das Verfahren von Heun ist eine 2-stufige Runge–Kutta–Formel, und es gilt:

$$u(t+h) = u(t) + \tfrac{h}{2}[f(t, u) + f(t+h, u + hf(t, u))].$$

Man zeige, daß dieses Verfahren die Konsistenzordnung 2 hat.

■ **Frage:** Existiert eine 2-stufige Runge–Kutta–Formel

$$\begin{array}{c|cc} 0 & & \\ c_2 & a_{21} & \\ \hline & b_1 & b_2 \end{array}$$

der Ordnung 3 oder höher ?

Zur Klärung dieser Frage entwickeln wir den lokalen Fehler $u(t+h) - u_h(t+h)$ wieder in eine Taylor-Reihe:

- $u(t+h) \overset{(12)}{\overset{\downarrow}{=}} u(t) + hu'(t) + \frac{h^2}{2}u''(t) + \frac{h^3}{6}u'''(t) + o(h^3) =$

$$\overset{(12)}{\overset{\downarrow}{=}} u(t) + hf(t,u) + \tfrac{h^2}{2}(f_t + f_u f)(t,u) +$$
$$+\tfrac{h^3}{6}(f_{tt} + 2f_{tu}f + f_{uu}f^2 + f_u f_t + f_u^2 f)(t,u) + o(h^3)$$

- $u_h(t+h) = u + h[b_1 f(t,u) + b_2 f(t + c_2 h, u + ha_{21} f(t,u))] =$

$$= u + h\, b_1 f + hb_2[f + f_t c_2 h + f_u h a_{12} f +$$
$$+\tfrac{1}{2!}(f_{tt}(c_2 h)^2 + 2f_{tu}(c_2 h)(ha_{21} f) + f_{uu}(ha_{12} f)^2) + o(h^2)]$$
$$= u + h(b_1 + b_2)f + h^2 b_2(c_2 f_t + a_{12} f_u f) +$$
$$+h^3 b_2\left(\tfrac{c_2^2}{2} f_{tt} + f_{tu} c_2 a_{21} f + \tfrac{a_{12}^2}{2} f_{uu} f^2\right) + o(h^3)$$

- Koeffizientenvergleich: $\Rightarrow$ (Notwendige) Bedingungen für Ordnung 2:

$$\boxed{\begin{aligned} b_1 + b_2 &= 1 \\ b_2\, c_2 &= 1/2 \\ b_2\, a_{12} &= 1/2 \end{aligned}} \quad \overset{\text{allgem. Lsg.}}{\Rightarrow} \quad \boxed{\begin{aligned} b_1 &= 1 - b_2 \\ c_2 &= 1/2\, b_2 \\ a_{12} &= 1/2\, b_2 \end{aligned}}$$

mit beliebiger Wahl von b_2 !

$\Updownarrow$

Damit ergibt sich der lokale Fehler zu:

$$u(t+h) - u_h(t+h) = \frac{h^3}{6}\left[\left(1 - \frac{3}{4b_2}\right)(f_{tt} + 2f_{tu}f + f_{uu}f^2) + \underline{\underline{f_u f_t + f_u^2 f}}\right](t,u) + \ldots,$$

woraus sofort erkennbar ist, daß die Ordnung 3 i. a. nicht erreichbar ist.

Für

$$\boxed{b_2 \ = \ 3/4}$$

ist die Summe der Beträge der einzelnen Beiträge zum lokalen Fehler minimal. Man spricht von einer (in diesem Sinne) optimalen Formel. Falls $f_u = 0$ ($\Rightarrow$ reine Integration), dann folgt für $b_2 = 3/4$ Ordnung 3 !

■ Runge–Kutta–Verfahren der Ordnung 3:

- RK-Verfahren der Ordnung 3 muß mindestens 3-stufig sein (↑):

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \end{array}$$

- Formales Herleiten von Bedingungsgleichungen für die Koeffizienten $\{c_i\}$, $\{a_{ij}\}$, $\{b_j\}$ ohne weiteres durch Taylor–Entwicklung und Koeffizientenvergleich möglich:

$$u(t+h) \overset{\overset{(12)}{\downarrow}}{=} u + hf + \tfrac{h^2}{2}(f_t + f_u f) + \tfrac{h^3}{6}(f_{tt} + 2f_{tu}f + f_{uu}f^2 + f_u f_t + f_u^2 f_t) + \ldots \text{ usw.}$$

$$\updownarrow$$

$$\begin{aligned} u_h(t+h) &= u + h(b_1 f(t,u) + b_2 f(t + c_2 h, u + h a_{21} f(t,u)) + b_3 f(t + c_3 h, u + \ldots)) \\ &= u + h(b_1 + b_2 + b_3)f + \tfrac{h^2}{2}(\qquad) + \tfrac{h^3}{6}(\qquad) + \ldots \end{aligned}$$

$\longrightarrow$ Vorgehen ist technisch zu kompliziert !

■ Systematisches Vorgehen zur Ermittlung der Bedingungsgleichungen für die Ordnung von RK–Formeln:

- Die folgende $\boxed{\text{Ü 3.8}}$ besagt, daß man sich bei der Analyse der Ordnung von RK–Formeln auf autonome Dgl.

(14) $$u'(t) \;=\; f(u(t))$$

beschränken kann.

Ü 3.8 Falls die Bedingungen

(15) $$c_i \;=\; \sum_{j=1}^{i-1} a_{ij}, \quad i = \overline{2,l}$$

erfüllt sind, erhält man für den Fall allgem. Dgl. der Form $u'(t) = f(t, u(t))$ die gleiche Ordnung wie für den Fall autonomer Dgl. (14).

Hinweis: $$\boxed{\begin{aligned} u'(t) &= f(t, u(t)) \\ u(0) &= u_0 \end{aligned}} \iff \boxed{\begin{aligned} &u_1'(t) = 1 \\ &u_2'(t) = f(u_1(t), u_2(t)) \\ &u_1(0) = 0,\; u_2(0) = u_0 \end{aligned}}$$

- Für die weitere Diskussion können wir annehmen, daß (15) gilt und daß nur autonome Dgl. (14) betrachtet werden !
Btr. dazu RK–Formeln in der Form (für autonome Dgl. (14))

$$(16)\quad g_i \equiv g_i(h) = u + \sum_{j=1}^{i-1} a_{ij} h f(g_j(h)),\ \ i = \overline{1,l} \quad (\sum_{j=1}^{0} = 0)$$

Taylor–Entwicklung an der Stelle $\underline{\underline{h=0}}$

$$u_h(t+h) = u + \sum_{j=1}^{l} b_j \underbrace{h f(g_j(h))}_{h\varphi(h)} = \ldots$$

- **<u>Lemma 3.7:</u>** (Leibnitz)

$$\boxed{\text{Falls } \varphi \in C^q(I\!R), \text{ dann gilt: } [h\varphi(h)]^{(q)}(0) = q\,\varphi^{(q-1)}(0).}$$

<u>Beweis:</u> durch Induktion nach q (für $q = 1$ o.k.).

$$[h\varphi(h)]^{(q+1)}(0) = [(h\varphi(h))']^{(q)}(0) = [\varphi(h) + h\varphi'(h)]^{(q)}(0) =$$
$$= \varphi^{(q)}(0) + (h\varphi'(h))^{(q)}(0) = \varphi^{(q)}(0) + q\varphi'^{(q-1)}(0) = (1+q)\varphi^{(q)}(0)$$

q.e.d.

- Mit Lemma 3.7 folgt aus (16) für $g_i^{(q)}(0) := \left.\dfrac{d^q g_i}{dh^q}\right|_{h=0}$:

$\underline{q = 0:}$ $\quad g_i(0) \quad = u \equiv u(t),$

$\underline{q = 1:}$ $\quad g_i'(0) \quad = \sum_{j=1}^{i-1} a_{ij} \cdot 1 \cdot f(g_j(0)) = \sum_{j=1}^{i-1} a_{ij} f(u),$

$\underline{q = 2:}$ $\quad g_i''(0) \quad = 2 \sum_{j=1}^{i-1} a_{ij} (f \circ g_j)'(0) =$

$$\boxed{(f \circ g_j)'(h) = f'(g_j(h)) g_j'(h)}$$

$$= 2 \sum_{j=1}^{i-1} a_{ij} f'(u) g_j'(0) =$$
$$= 2 \sum_{j=1}^{i-1} a_{ij} f'(u) \sum_{k=1}^{j-1} a_{jk} f(u) =$$
$$= 2 \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} a_{ij} a_{jk} f'(u) f(u)$$

$\underline{q = 3:}$ $\quad g_i'''(0) \quad = 3 \sum_{j=1}^{i-1} a_{ij} (f \circ g_j)''(0) =$

$$\boxed{\begin{aligned}[f(g_j(h))]'' &= [f'(g_j(h)) g_j'(h)]' = \\ &= f''(g_j(h)) g_j'(h) g_j'(h) + f'(g_j(h)) g_j''(h)\end{aligned}}$$

$$= 3 \sum_{j=1}^{i-1} a_{ij} [f''(u) g_j'(0) g_j'(0) + f'(u) g_j''(0)] =$$

$$= 3 \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} \sum_{n=1}^{j-1} a_{ij} a_{jk} a_{jn} f''(u) f(u) f(u) +$$
$$+ 6 \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} \sum_{n=1}^{k-1} a_{ij} a_{jk} a_{kn} \cdot f'(u) \cdot f'(u) \cdot f(u)$$

usw.

- $u_h(t+h)$ unterscheidet sich von $g_i(h)$ nur durch eine andere Gewichtung:

  $\Rightarrow \left.\dfrac{d^q u_h(t+h)}{dh^q}\right|_{h=0} = u_h^{(q)}(t)$

  $\Rightarrow$ in obigen Formeln a_{ij} durch b_j ersetzen:

$$
\boxed{\begin{aligned}
u_h'(t) &= \sum_{j=1}^{l} b_j f(u) \\
u_h''(t) &= 2 \sum_{j=1}^{l} \sum_{k=1}^{j-1} b_j a_{jk} f'(u) f(u) \\
u_h'''(t) &= 3 \sum_{j=1}^{l} \sum_{k=1}^{j-1} \sum_{n=1}^{j-1} b_j a_{jk} a_{jn} f''(u) f(u) f(u) + \\
&\quad + 6 \sum_{j=1}^{l} \sum_{k=1}^{j-1} \sum_{n=1}^{k-1} b_j a_{jk} a_{kn} f'(u) f'(u) f(u) \\
&\text{usw.}
\end{aligned}} \tag{17}
$$

- <u>Aus der Dgl. $u'(t) = f(u(t))$ folgen die entsprchenden Abl. für u:</u>

$$
\boxed{\begin{aligned}
u'(t) &= f(u(t)) \\
u''(t) &= f'(u(t))u'(t) = f'(u(t))f(u(t)) \\
u'''(t) &= f''(u(t))u'(t)f(u(t)) + f'(u(t))f'(u(t))u'(t) = \\
&= f''(u(t))f(u(t))f(u(t)) + f'(u(t))f'(u(t))f(u(t)) \\
&\text{usw.}
\end{aligned}} \tag{18}
$$

- <u>Daraus erhält man die Taylor–Entwicklung des lokalen Fehlers:</u>

$$
\begin{aligned}
u(t+h) - u_h(t+h) &= u(t) + hu'(t) + \tfrac{h^2}{2}u''(t) + \tfrac{h^3}{6}u'''(t) + \ldots && (18) \\
&\quad -(u_h''(t) + hu_h'(t) + \tfrac{h^2}{2}u_h''(t) + \tfrac{h^3}{6}u_h'''(t) + \ldots) && (17)
\end{aligned}
$$

$$
\begin{aligned}
\overset{(17)}{\underset{(18)}{=}} \quad & h(1 - \sum_{j=1}^{l} b_j) f(u) + \\
& + \tfrac{h^2}{2}(1 - 2\sum_{j=1}^{l}\sum_{k=1}^{j-1} b_j a_{jk}) f'(u) f(u) + \\
& + \tfrac{h^3}{6}(1 - 3\sum_{j=1}^{l}\sum_{k=1}^{j-1}\sum_{n=1}^{j-1} b_j a_{jk} a_{jn}) f''(u) f(u) f(u) + \\
& + \tfrac{h^3}{6}(1 - 6\sum_{j=1}^{l}\sum_{k=1}^{j-1}\sum_{n=1}^{k-1} b_j a_{jk} a_{kn}) f'(u) f'(u) f(u) + \\
& + \ldots \text{ usw.}
\end{aligned}
$$

- Damit ergeben sich die folgenden 4 Bedingungen für RK–Verfahren der Ordnung 3:

$$
(19) \qquad
\boxed{
\begin{array}{ll|c}
 & \sum\limits_{j=1}^{l} b_j \qquad = 1 & h \\
\hline
2 & \sum\limits_{j=1}^{l} \sum\limits_{k=1}^{j-1} b_j a_{jk} \qquad = 1 & \frac{h^2}{2!} \\
\hline
3 & \sum\limits_{j=1}^{l} \sum\limits_{k=1}^{j-1} \sum\limits_{n=1}^{j-1} b_j a_{jk} a_{jn} = 1 & \frac{h^3}{3!} \\
6 & \sum\limits_{j=1}^{l} \sum\limits_{k=1}^{j-1} \sum\limits_{n=1}^{k-1} b_j a_{jk} a_{kn} = 1 & \\
\hline
\multicolumn{2}{l|}{\text{usw. für höhere Ordnung, z.B. + 4 Bd. für } \frac{h^4}{4!}} & \text{usw.}
\end{array}
}
\quad
\begin{array}{l}
\Big\downarrow \ \text{Taylor} \\[1em]
\underline{\text{Bez.:}} \\
\sum\limits_{j=1}^{0} = 0
\end{array}
$$

- Ein 3-stufiges RK–Verfahren ($l = 3$) wird durch das Tableau

$$
\begin{array}{c|ccc}
0 & & & \\
c_2 & a_{21} & & \\
c_3 & a_{31} & a_{32} & \\
\hline
 & b_1 & b_2 & b_3
\end{array}
\qquad
\begin{array}{l}
\\
\longleftarrow \sum \quad (15) \\
\longleftarrow \sum \quad (15) \\
\\
\end{array}
$$

beschrieben und hat somit 6 frei wählbare Parameter, da die Koeffizienten $\{c_i\}$ bereits durch (15) festgelegt sind.
Diese 6 Parameter $\{a_{21}, a_{31}, a_{32}, b_1, b_2, b_3\}$ lassen sich so wählen, daß die Bedingungen (19) für $l = 3$ erfüllt werden, d.h. es gibt 3-stufige RK–Formeln der Ordnung 3 !

■ Runge–Kutta–Verfahren der Ordnung 4:

- Um Ordnung 4 zu erreichen, müssen weitere 4 Bedingungen erfüllt werden.

- Es läßt sich zeigen, daß dies mit 3-stufigen Formeln nicht möglich ist !

- 4-stufige RK-Formeln werden durch das Tableau

$$
\begin{array}{r|cccc}
c_1 = 0 & & & & \\
c_2 & a_{21} & & & \\
c_3 & a_{31} & a_{32} & & \\
c_4 & a_{41} & a_{42} & a_{43} & \\
\hline
 & b_1 & b_2 & b_3 & b_4
\end{array}
$$

beschrieben und haben somit 10 freie Parameter, die tatsächlich so gewählt werden können, daß die 8 Bedingungen für Ordnung 4 erfüllt sind (Allgem. Lösung des polynominalen GS: eine 2-parametrige Lösungsschar und drei 1-parametrige Lösungsscharen).

- Beispiele 4-stufiger RK–Formeln der Ordnung 4:

(1) Das sogenannte „klassische" Runge–Kutta–Verfahren:

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

Für ein Quadraturproblem $u'(t) = f(t)$ wird aus dem klassischen RK–Verfahren die Simpson–Regel:

$$u_h(t+h) = u(t) + h\left[\frac{1}{6}f(t) + \frac{1}{3}f\left(t+\frac{h}{2}\right) + \frac{1}{3}f\left(t+\frac{h}{2}\right) + \frac{1}{6}f(t+h)\right]$$

(2) Die 3/8–Regel:

$$\begin{array}{c|cccc} 0 & & & & \\ 1/3 & 1/3 & & & \\ 2/3 & -1/3 & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array}$$

Für ein Quadraturproblem $u'(t) = f(t)$ wird aus der 3/8–Regel die sogenannte Newton–3/8–Regel:

$$u_h(t+h) = u(t) + h\left[\frac{1}{8}f(t) + \frac{3}{8}f\left(t+\frac{h}{3}\right) + \frac{3}{8}f\left(t+\frac{2}{3}h\right) + \frac{1}{8}f(t+h)\right]$$

- Mit 4-stufigen RK–Formeln läßt sich die Ordnung 5 nicht erreichen !

■ **Ergebnisse zur Analyse von s -stufigen RK–Formeln** ($s \geq 5$)
siehe [1], Punkt 4.2.3, S. 124-130.

3.3.3 Implizite Runge–Kutta–Verfahren

3.3.3.1 Beispiele

■ **Implizites Euler–Verfahren ($\widehat{=}\ \sigma = 1$ in Kap.2):**

$$J = \int_t^{t+h} f(s, u(s))\, ds$$

linksseitige Rechteckregel: $J \approx hf(t, u(t))$
$\Rightarrow$ Expliziter Euler: $u_{j+1} = u_j + h_j f(t_j, u_j)$

rechtsseitige Rechteckregel: $J \approx hf(t+h, u(t+h))$
$\Rightarrow$ Impliziter Euler: $u_{j+1} = u_j + h_j f(t_j + h_j, u_{j+1})$

Resultat: Implizites Euler–Verfahren ($\sigma = 1$)
= Euler rückwärts
= rechtsseitige Rechteckregel

(20) $$\mathbf{u_{j+1}} = u_j + h_j f(t_j + h_j, \mathbf{u_{j+1}}), j = 0, 1, \ldots, m-1;\ u_0 \text{ geg.}$$

> Zur Bestimmung von $\mathbf{u_{j+1}}$ muß i. allgem. ein nichtlineares Gleichungssystem gelöst werden, z.B. durch Fixpunktiteration oder durch Newton–Iteration mit der Startnäherung $u_{j+1}^0 = u_j$!

Bemerkung:

- Nachteile impliziter Verfahren: Nichtlineare GS zu lösen ($\uparrow$)
- Vorteile impliziter Verfahren:
  - Bessere Stabilität ($\downarrow$)
  - Genauigkeit ($\downarrow$)

Implizite Mittelpunktsregel:

$$\int_t^{t+h} f(s, u(s))\, ds \overset{\text{MR}}{\approx} hf\left(t + \tfrac{h}{2}, u\left(t + \tfrac{h}{2}\right)\right)$$

$$\Rightarrow\ u(t+h) \approx u(t) + hf\left(t + \tfrac{h}{2}, \underset{\substack{\uparrow \\ \mathbf{g_1} \approx \mathbf{u}\left(\mathbf{t} + \frac{\mathbf{h}}{\mathbf{2}}\right)}}{\mathbf{u}\left(\mathbf{t} + \tfrac{\mathbf{h}}{\mathbf{2}}\right)}\right)$$

durch implizites Euler–Verfahren

$$\mathbf{g_1} = u + \tfrac{h}{2} f\left(t + \tfrac{h}{2}, \mathbf{g_1}\right)$$

Resultat: Implizite Mittelpunktsregel

$$\boxed{\begin{array}{rcl} \mathbf{g_1} & = & u + \frac{h}{2} f\left(t + \frac{h}{2}, \mathbf{g_1}\right) \\ u_h(t+h) & = & u + hf\left(t + \frac{h}{2}, g_1\right) \\ \hline \mathbf{K_1} & \equiv & f\left(t + \frac{h}{2}, g_1\right) = f\left(t + \frac{h}{2}, u + \frac{h}{2}\mathbf{K_1}\right) \\ u_h(t+h) & = & u + h\,K_1 \end{array}}$$

g–Form

K–Form

■ **Beide Verfahren sind Beispiele von sogenannten impliziten Runge–Kutta–Verfahren** (Formeln).

3.3.3.2 Allgemeines Konstruktionsprinzip

- **Die allgemeine Form einer impliziten l -stufigen Runge–Kutta–Formel läßt sich entweder**

  - in der g–Form

$$
\boxed{
\begin{array}{l}
\mathbf{g_1} = u + h[a_{11} f(t + c_1 h, \mathbf{g_1}) + \ldots + a_{1l} f(t + c_l h, \mathbf{g_l})] \\
\mathbf{g_2} = u + h[a_{21} f(t + c_1 h, \mathbf{g_1}) + \ldots + a_{2l} f(t + c_l h, \mathbf{g_l})] \\
\vdots \\
\mathbf{g_l} = u + h[a_{l1} f(t + c_1 h, \mathbf{g_1}) + \ldots + a_{ll} f(t + c_l h, \mathbf{g_l})] \\
\hline
u_h(t+h) = u + h[b_1 f(t + c_1 h, g_1) + \ldots + b_l f(t + c_l h, g_l)
\end{array}
}
$$

  bzw.

  - in der K–Form

$$
\boxed{
\begin{array}{l}
\mathbf{K_1} = f(t + c_1 h, u + h[a_{11}\mathbf{K_1} + a_{12}\mathbf{K_2} + \ldots + a_{1l}\mathbf{K_l}]) \\
\mathbf{K_2} = f(t + c_2 h, u + h[a_{21}\mathbf{K_1} + a_{22}\mathbf{K_2} + \ldots + a_{2l}\mathbf{K_l}]) \\
\vdots \\
\mathbf{K_l} = f(t + c_l h, u + h[a_{l1}\mathbf{K_1} + a_{l2}\mathbf{K_2} + \ldots + a_{ll}\mathbf{K_l}]) \\
\hline
u_h(t+h) = u + h[b_1 K_1 + b_2 K_2 + \ldots + b_l K_l]
\end{array}
}
$$

  schreiben.

- **Das Verfahren wird durch das folgende Tableau eindeutig definiert:**

$$
(21) \qquad
\begin{array}{c|ccccc}
c_1 & a_{11} & a_{12} & \ldots & a_{1,l-1} & a_{1l} \\
c_2 & a_{21} & a_{22} & \ldots & a_{2,l-1} & a_{2l} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
c_l & a_{l1} & a_{l2} & \ldots & a_{l,l-1} & a_{ll} \\
\hline
 & b_1 & b_2 & \ldots & b_{l-1} & b_l
\end{array}
\qquad\qquad
\begin{array}{c}
\underline{\text{kompakt}} \\
\begin{array}{c|c}
c & A \\
\hline
 & b^T
\end{array}
\end{array}
$$

■ **Definition 3.8:**

> Eine durch das Tableau (21) beschriebene RK–Formel heißt
> - explizit, falls A eine echte linke untere Dreiecksmatrix ist,
> - implizit, falls A nicht explizit ist.

Bem.: Diese Def. einer expliziten RK–Formel ist etwas allgemeiner, da nicht mehr $c_1 = 0$ a–priori angenommen wird.

■ **Beispiele:**

1. Implizites Euler–Verfahren = 1-stufige implizite RK–Formel:

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

2. Implizite Mittelpunktsregel = 1-stufige implizite RK–Formel:

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

3. Implizite Trapezregel = 2-stufige implizite RK–Formel:

$u_{j+1} = u_j + \frac{h_j}{2}[f(t_j, u_j) + f(t_j + h_j, u_{j+1})]$

($\sigma = 1/2$ in Kap. 2: CRANK-NICOLSON–Verfahren)

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

3.3.3.3 Durchführbarkeit

■ **Bei allen impliziten RK–Verfahren stellt sich die Frage nach der Durchführbarkeit, d.h. nach der Lösbarkeit des i. allgem. nicht–linearen Gleichungssystems zur Bestimmung der nächsten Näherung:**

- g–Form:

  - Geg. (t, u, h)
  - Ges. $g = (g_1, \ldots, g_l)^T$ als Lösung der Fixpunktgleichung

    (22) $\boxed{\mathbf{g} = \Phi(\mathbf{g}, t, u, h)}$

    mit $\Phi = (\Phi_1, \ldots, \Phi_l)^T$, $\Phi_i = u + h \sum\limits_{j=1}^{l} a_{ij} f(t + c_j h, g_j)$
  - $u_h(t+h) = u + h \sum\limits_{j=1}^{l} b_j f(t + c_j h, g_j)$

- K–Form:

  - Geg. (t, u, h)
  - Ges. $K = (K_1, \ldots, K_l)^T$ als Lösung der Fixpunktgleichung

    (23) $\boxed{\mathbf{K} = \Psi(\mathbf{K}, t, u, h)}$

    mit $\Psi = (\Psi_1, \ldots, \Psi_l)^T$, $\Psi_i = f(t + c_i h, u + h \sum\limits_{j=1}^{l} a_{ij} K_j)$
  - $u_h(t+h) = u + h \sum\limits_{j=1}^{l} b_j K_j$

■ **Der folgende Satz zeigt unter geeigneten Voraussetzungen die Existenz und Eindeutigkeit der Lösung der Fixpunktgleichungen (22) bzw. (23), die z.B. durch die Banachsche Fixpunktiteration iterativ bestimmt werden kann:**

<u>**Satz 3.9:**</u> (Existenz, Eindeutigkeit, Konstruktion)

<u>Vor.:</u> Sei $u \in \mathbb{R}^N$ und $D = \{(t,v) : t \in I = [0,T],\ |v-u| \le b\}$ mit fixierten $b \in (0,\infty)$ bzw. $D = I \times \mathbb{R}^N$ für $b = \infty$. Desweiteren gelte:

1) $f : D \longrightarrow \mathbb{R}^N$ – stetig,

2) f sei auf D beschränkt, d.h. $|f(t,v)| \le M,\ \ \forall\,(t,v) \in D$, und f sei auf D im zweiten Argument Lipschitz–stetig, d.h.

$|f(t,v) - f(t,u)| \le L|v-u| \ \ \forall\,(t,v),(t,u) \in D$

mit fixierten Konstanten $M, L \in [0,\infty)$,

3) $h\|A\|_\infty M \le b,\ \ h\|A\|_\infty L < 1$.

<u>Bh.:</u> 1) Dann besitzt die Fixpunktgleichung

(22) $\qquad \mathbf{g} = \Phi(\mathbf{g}, t, u, h)$

für bel. (aber fixierte) (t,u,h) mit $t, t + c_i h \in [0,T]$ $(i = \overline{1,l})$ eine eindeutig bestimmte Lösung

$g \in \mathcal{K} := \{g \in (\mathbb{R}^N)^l := \mathbb{R}^N \times \ldots \times \mathbb{R}^N : |g_i - u| \le b \ \ \forall\, i = \overline{1,l}\}$.

2) Die Fixpunktiteration

(24) $\qquad \mathbf{g^{(n+1)}} = \Phi(\mathbf{g^{(n)}}, t, u, h),\ n = 0, 1, \ldots$

mit dem Startwert $g^{(0)} = (u, u, \ldots, u)^T \in \mathcal{K}$ konvergiert linear gegen diese Lösung.

<u>Beweis:</u> beruht natürlich auf Banachschem Fixpunktsatz I.6.1 [16]:

- $(\mathbb{R}^N)^l$ – B–Raum, $\|\cdot\| := \|\cdot\|_\infty := \max\limits_{i=\overline{1,l}} |\cdot|$.

- $\mathcal{K} = \bar{\mathcal{K}} \subset (\mathbb{R}^N)^l$ – abgeschlossen, nicht leer.

- $\Phi\mathcal{K} \subset \mathcal{K}$, tatsächlich

$$|\Phi_i(g) - u| = \left| \left[u + h \sum_{j=1}^{l} a_{ij} f(t + c_j h, g_j) \right] - u \right| =$$

$$= h \underbrace{\sum_{j=1}^{l} |a_{ij}|}_{\le \|A\|_\infty} \underbrace{|f(t + c_j h, g_j)|}_{\le M} \le h\|A\|_\infty M \le b \ \ \forall\, g \in \mathcal{K}.$$

- Kontraktivität von Φ:

$$\|\Phi(\bar g) - \Phi(g)\|_\infty = \max_{i=\overline{1,l}} |u - h \sum_{j=1}^{l} a_{ij} f(t + c_j h, \bar g_j) - [u - h \sum_{j=1}^{l} a_{ij} f(t + c_j h, g_j)]|$$

$$= \max_{i=\overline{1,l}} |h \sum_{j=1}^{l} a_{ij} (f(t + c_j h, g_j) - f(t + c_j h, \bar g_j))|$$

$$\leq h \max_{i=\overline{1,l}} \sum_{j=1}^{l} |a_{ij}| \, |f(t + c_j h, g_j) - f(t + c_j h, \bar g_j)|$$

$$\leq h \max_{i=\overline{1,l}} \sum_{j=1}^{l} |a_{ij}| \, L \, |\bar g_j - g_j|$$

$$\leq \underbrace{h \, \|A\|_\infty \, L}_{=: \, q \, < \, 1} \|\bar g - g\|_\infty = q \, \|\bar g - g\|_\infty .$$

- Aussagen von Satz 3.9 folgen nun unmittelbar aus Satz I.6.1.

q.e.d.

■ **Bemerkungen:**

1. Aussagen von Satz 3.9. gelten auch für Fixpunktgl. (23), (mms).
2. Durchführbarkeit impl. RK–Formeln ist somit für hinreichend klein h garantiert !
3. Fixpkt.–Iteration: $g^{(n+1)} = \Psi(g^{(n)}, t, u, h) : \|g - g^{(n)}\|_\infty \leq q^n \|g - g^{(0)}\|_\infty$ usw.
   Möglich: Newton–Iteration, falls $f \in C^1$, sowie andere Iterationsverfahren.

3.3.3.4 Konsistenzordnung impliziter Runge–Kutta–Formeln

■ **Eine RK–Formel läßt sich auch folgendermaßen aufschreiben:**

$$\boxed{\begin{aligned} & u_h(t+h) = u + h\,\varphi(t,u,h), \\ & \varphi(t,u,h) = \sum_{j=1}^{l} b_j K_j(t,u,h) = \sum_{j=1}^{l} b_j f(t + c_j h, g_j) \\ & K_j = K_j(t,u,h) : \underset{(23)}{K = \Psi(K,t,u,h)} \text{ bzw. } g_j = g_j(t,u,h) : \underset{(22)}{g = \Phi(g,t,u,h)} \end{aligned}} \tag{25}$$

■ **Fehlerbegriffe** werden analog zu den expliziten RK–Formeln definiert:

– globaler Diskretisierungsfehler: $e_h(t) = u(t) - u_h(t)$,

– lokale Diskretisierungsfehler: $u(t+h) - u_h(t+h)$
(vgl. Def. 3.6) Dgl. RK–Formel

mit gleichen Startwerten $u(t) = u_h(t)$
im Pkt. t, wobei $t \in \{t_0 = 0, t_1, \ldots, t_{m-1}\}$.

– lokaler Abschneidefehler (vgl. (9)):

$$\tau_h(t+h) = \frac{u(t+h) - u(t)}{h} - \varphi(t, u, h), \quad t \in \{t_0 = 0, t_1, \ldots, t_{m-1}\},$$

$$\tau_h(t_0 \equiv 0) = 0.$$

– Es gilt: $\tau_h(t+h) = \frac{1}{h}[u(t+h) - (u(t) + h\varphi(t, u(t), h))]$.

– Konsistenzordnung p (vgl. auch Def. 3.6)
  – entspricht der Bedingung
    $$u(t+h) - u_h(t+h) = O(h^{p+1})$$
    an den lokalen Fehler
  – bzw. der Bedingung
    $$\tau_h = O(h^p)$$
    an den lokalen Abschneidefehler.

■ **Die Konsistenzordnung impliziter RK–Formeln läßt sich analog zum expliziten Fall für autonome Dgl. durch Taylor–Entwicklung des lokalen Fehlers bestimmen** (siehe Pkt. 3.3.2.2):

- 1 Bedingung für Ordnung 1:
  $$\sum_{j=1}^{l} b_j = 1$$
- 1 zusätzliche Bed. für Ordnung 2:
  $$2 \sum_{j,k} b_j a_{jk} = 1$$
- 2 zusätzliche Bed. für Ordnung 3:
  $$3 \sum_{j,k,n} b_j a_{jk} a_{jn} = 1$$
  $$6 \sum_{j,k,n} b_j a_{jk} a_{kn} = 1$$
- 4 zusätzliche Bed. für Ordnung 4:
  $$4 \sum_{j,k,n,m} b_j a_{jk} a_{jn} a_{jm} = 1$$
  $$8 \sum_{j,k,n,m} b_j a_{jk} a_{jn} a_{nm} = 1$$
  $$12 \sum_{j,k,n,m} b_j a_{jk} a_{kn} a_{km} = 1$$
  $$24 \sum_{j,k,n,m} b_j a_{jk} a_{kn} a_{nm} = 1$$
- usw., mit $\sum_{\ldots} = \sum_{\ldots=1}^{l}$

Die für die allgem. Dgl. $u'(t) = f(t, u(t))$ notwendigen Koeffizienten $\{c_i\}_{i=\overline{1,l}}$:

(15) $$c_i = \sum_{j=1}^{l} a_{ij}.$$

- **Unter Verwendung von (15) $\boldsymbol{c_i = \sum_{j=1}^{l} a_{ij}}$ kann man die obigen Ordnungsbedingungen für den allgemeinen Fall $\boldsymbol{u'(t) = f(t, u(t))}$ umschreiben:** $(\sum_{\dots} = \sum_{\dots=1}^{l})$

$$\left.\begin{array}{l}\left.\begin{array}{l}\left.\begin{array}{l}\left.\begin{array}{ll}\sum_j b_j & = 1\end{array}\right\}\text{Ordn. 1}\\ \begin{array}{ll}\sum_j b_j c_j & = 1/2\end{array}\end{array}\right\}\text{Ordn. 2}\\ \sum_j b_j c_j^2 = 1/3\\ \sum_{j,k} b_j a_{jk} c_k = 1/6\end{array}\right\}\text{Ordn. 3}\\ \sum_j b_j c_j^3 = 1/4\\ \sum_{j,n} b_j c_j a_{jn} c_n = 1/8\\ \sum_{j,k} b_j a_{jk} c_k^2 = 1/12\\ \sum_{j,k,n,} b_j a_{jk} a_{kn} c_n = 1/24\end{array}\right\}\text{Ordn. 4}$$

usw.

- **Beispiele:**

1. Impliziter Euler ($l = 1$): $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \Rightarrow \boxed{\text{Ordnung} = 1}$

2. Implizite MP–Regel ($l = 1$): $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \Rightarrow \boxed{\text{Ordnung} = 2}$

3. Implizite Trapezregel ($l = 2$): $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \Rightarrow \boxed{\text{Ordnung} = 2}$

- **Frage:** $\exists$ 2-stufige, implizite RK–Formeln, die eine höhere Ordnung als 2 besitzen ?

  – Für $l = 2$ haben die 4 Bedingungen für Ordnung 3 eine 2–parametrige Lösungsschar !

  – Die zusätzlichen 4 Bedingungen für Ordnung 4 können durch eine geeignete Wahl der 2 freien Parameter ebenfalls noch erfüllt werden.
  Die einzige Lösung hat das folgende Tableau:

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

  $\uparrow$
  Stützstellen der
  Gauß-QF der Stufe 2
  für das Intervall [0,1]

  $\Rightarrow$ $\exists$! 2-stufige ($l = 2$) implizite RK–Formel der Ordnung 4 !
  $\longrightarrow$ RK–Formel vom Gauß–Typ !

  – Resümee: Mit l–stufigen impliziten RK–Formeln sind offenbar höhere Konsistenzordnungen erreichbar als mit den entsprechenden expliziten RK–Formeln !

- **Die folgenden beiden Sätze geben Aussagen über die mit l –stufigen, impliziten RK–Formeln erreichbare Konsistenzordnung:**

  - **Satz 3.10:**

Vor.: 1. Es seien die Voraussetzungen von Satz 3.2 erfüllt:

- $f : D = \{(t, v) : t \in [0, T], |v - u| \leq b\} \to I\!R^N$ – stetig, $b \in (0, \infty)$ fix,
- $|f(t, v) - f(t, \bar{v})| \leq L|v - \bar{v}|, |f(t, v)| \leq M$ auf D; $L, M \in [0, \infty)$,
- $M \cdot T \leq b$ (entfällt bei „$b = \infty$“).

2. $r, l \in I\!N : r \leq l$; $f \in C^r(D, I\!R^N)$; $[h\|A\|_\infty \; L < 1]$.

Bh.: Falls für eine l–stufige RK–Formel die Bedingungen

(26) $\bullet \; \sum\limits_{j=1}^{l} a_{ij} c_j^k = \dfrac{c_i^{k+1}}{k+1}, \; k = \overline{0, r-2}, \; i = 1, 2, \ldots, l$

(27) $\bullet \; \sum\limits_{j=1}^{l} b_j c_j^k = \dfrac{1}{k+1}, \; k = \overline{0, r-1}$

erfüllt sind, besitzt die Formel die Konsistenzordnung $p = r$.

<u>Beweis:</u> $\Rightarrow$ direkte Abschätzung des <u>lokalen Fehlers</u> $u(t+h) - u_h(t+h) = ?$

für <u>autonome Dgl.</u> ($\Rightarrow$ (15) $c_i = \sum\limits_{j=1}^{l} a_{ij}$) !:

- $$u(t+h) = u + \underbrace{\int\limits_t^{t+h} f(u(s))\, ds}_{\text{Integral}} = u + \underbrace{h \sum_{j=1}^{l} b_j f(u(t+c_j h))}_{\text{QF}} + \underbrace{R_0}_{\text{Restglied}}$$

  Aus der Taylorentwicklung

  $$f(u(s)) = u'(s) = \sum_{k=0}^{r-1} \frac{1}{k!} u^{(k+1)}(t)\,(s-t)^k + O(h^r), \quad s \in [t, t+h],$$

  folgt für das Restglied

  $$R_0 = \int\limits_t^{t+h} f(u(s))\, ds - h \sum_{j=1}^{l} b_j f(u(t+c_j h)) =$$

  $$= \sum_{k=0}^{r-1} \frac{1}{k!} u^{(k+1)}(t) \int\limits_t^{t+h} (s-t)^k\, ds - h \sum_{j=1}^{l} b_j \left(\sum_{k=0}^{r-1} \frac{1}{k!} u^{(k+1)}(t)\,(c_j h)^k \right) +$$

  $$+ O(h^{r+1})$$

  $$= \sum_{k=0}^{r-1} \frac{1}{k!} u^{(k+1)}(t) \frac{h^{k+1}}{k+1} - \sum_{k=0}^{r-1} \frac{1}{k!} u^{(k+1)}(t) h^{k+1} \sum_{j=1}^{l} b_j c_j^k + O(h^{r+1}) =$$

  $$= \sum_{k=0}^{r-1} \frac{1}{k!} u^{(k+1)}(t) h^{k+1} \underbrace{\left[\frac{1}{k+1} - \sum_{j=1}^{l} b_j c_j^k \right]}_{\underset{(27)}{=}\; 0} + O(h^{r+1}) = O(h^{r+1})$$

- Mit

  $$u_h(t+h) = u + h \sum_{j=1}^{l} b_j f(g_j)$$

  erhalten wir für den lokalen Fehler

  $$u(t+h) - u_h(t+h) = h \sum_{j=1}^{l} b_j [f(u(t+c_j h)) - f(g_j)] + \overbrace{O(h^{r+1})}^{=R_0}$$

  die Abschätzung

  $$(28) \qquad \begin{aligned} |u(t+h) - u_h(t+h)| &\le h \sum_{j=1}^{l} |b_j|\, |f(u(t+c_j h)) - f(g_j)| + c h^{r+1} \\ &\le h \sum_{j=1}^{l} |b_j|\, L\, |u(t+c_j h) - g_j| + c h^{r+1}. \end{aligned}$$

- Schätzen nun völlig analog wie oben

  $|u(t+c_i h) - g_i| \le \ldots$

  für $i = 1, 2, \ldots, l$ ab:

$$u(t+c_ih) = u + \int\limits_t^{t+c_jh} f(u(s))\,ds = u + h\sum_{j=1}^{l} a_{ij} f(u(t+c_jh)) + R_i.$$

Für das Restglied R_i erhalten wir ($\uparrow$):

$$R_i = \sum_{k=0}^{r-2} \frac{1}{k!} u^{(k+1)}(t) h^{k+1} \underbrace{\left[\frac{c_i^{k+1}}{k+1} - \sum_{j=1}^{l} a_{ij} c_j^k \right]}_{\underset{(26)}{=}\;0} + O(h^r) = O(h^r).$$

Mit

$$g_i = u + h\sum_{j=1}^{l} a_{ij} f(g_j)$$

erhalten wir

$$u(t+c_ih) - g_i = h\sum_{j=1}^{l} a_{ij}[f(u(t+c_jh)) - f(g_j)] + O(h^r).$$

Wegen der Lipschitz–Stetigkeit von f folgt:

$$\max_{i=\overline{1,l}} |u(t+c_ih) - g_i| \le h\,\|A\|_\infty\, L \max_{j=\overline{1,l}} |u(t+c_jh) - g_j| + ch^r,$$

und daher gilt:

$$(29) \qquad \max_{i=\overline{1,l}} |u(t+c_ih) - g_i| \le \frac{1}{1 - h\|A\|_\infty L}\, ch^r = O(h^r).$$

- Aus (28) und (29) folgt nun unmittelbar:

$$u(t+h) - u_h(t+h) = O(h^{r+1}), \quad h \to 0.$$

q.e.d.

- **Bemerkung 3.11:**

1. Für $r = l$ folgt aus Satz 3.10, daß eine l–stufige, implizite RK–Formel die Ordnung $p = l$ hat, falls

$$(26) \qquad \boxed{\sum_{j=1}^{l} a_{ij} c_j^k = \frac{c_i^{k+1}}{k+1}, \quad k = \overline{0, l-2}}$$

für alle $i = 1, 2, \ldots, l$ und

$$(27) \qquad \boxed{\sum_{j=1}^{l} b_j c_j^k = \frac{1}{k+1}, \quad k = \overline{0, l-1}}$$

gilt. Diese Bedingungen bedeuten für die dazugehörigen QF, daß sie den algebraischen Genauigkeitsgrad $l-1$ besitzt, d.h. z.B. für

$$(30)\quad \int_t^{t+h} g(s)\,ds = h\sum_{j=1}^{l} b_j g(t+c_jh),$$
$$\forall\, g(s)=(s-t)^k,\quad k=0,1,\ldots,l-1.$$

Tatsächlich,

$$\int_t^{t+h}(s-t)^k\,ds = \left.\frac{(s-t)^{k+1}}{k+1}\right|_t^{t+h} = \frac{h^{k+1}}{k+1},$$
$$h\sum_{j=1}^{l} b_j(t+c_jh-t)^k = h^{k+1}\sum_{j=1}^{l} b_jc_j^k,$$

d.h. (30) ist äquivalent zu (27).

2. Falls $\{c_j\}_{j=\overline{1,l}}$ paarweise verschieden, d.h. $c_i \neq c_j\ \forall\, i\neq j$, <u>vorgegeben</u> werden, dann lassen sich die Vektoren $\{a_{ij}\}_{j=\overline{1,l}}$ aus den l GS (26) mit $k=\overline{0,l-1}$ (!) und der Vektor $\{b_j\}_{j=\overline{1,l}}$ aus den GS (27) eindeutig bestimmen, da die Systemmatrix

$$\begin{bmatrix} 1 & \ldots\ 1 \\ c_1 & \ldots\ c_l \\ c_1^{l-1} & \ldots\ c_l^{l-1}\end{bmatrix},\ \det[\ldots] = \prod_{\substack{i,j=1\\ i>j}}^{l}(c_i-c_j)\neq 0,$$

(Van der Monde'sche Determinante)

dieser GS offenbar regular ist.

Die damit erzeugten QF beruhen auf Polynominterpolation an den Stützstellen $t+c_ih,\quad i=1,2,\ldots,l$.

3. Man kann nun durch geeignete <u>Wahl</u> der Stützstellen $\{t+c_ih\}_{i=\overline{1,l}}$ versuchen, die Ordnung weiter zu erhöhen. Das erinnert, so wie auch schon die ersten 3 Bedingungen für Ordnung 3, an die Gauß–QF (siehe Satz 3.12).

- **<u>Satz 3.12:</u>**

<u>Vor.:</u> 1. Es seien die Voraussetzungen von Satz 3.2 erfüllt (siehe auch Satz 3.10).
2. $l\in I\!N : f\in C^l(D, I\!R^N)$.

<u>Bh.:</u> Falls für eine l–stufige RK–Formel die Bedingungen

$$(26)\quad \sum_{j=1}^{l} a_{ij}c_j^k = \frac{c_i^{k+1}}{k+1},\quad k=0,1,\ldots,l-1,$$

für alle $i=1,2,\ldots,l$ und

$$(31)\quad \sum_{j=1}^{l} b_jc_j^k = \frac{1}{k+1},\quad k=0,1,\ldots,2l-1$$

erfüllt sind, besitzt die RK–Formel die Konsistenzordnung $p=2l$.

<u>Beweis:</u> siehe z.B. [3] Grigorieff R.D.: Numerik gewöhnlicher Differentialgleichungen I. Teubner–Verlag, Stuttgart 1972.

- **Bemerkung 3.13:**

  Bedingung (31) bestimmt genau die Gauß–QF der Ordnung l, die bekanntlich den algebraischen Genauigkeitsgrad $2l-1$ besitzt (vgl. auch Bem. 3.11.1). Bedingungen an die $\{a_{ij}\}$ mit $k=\overline{0,l-1}$ besitzen eine eindeutige Lsg. bei geg. $\{c_i\}$:

  $\Rightarrow$ $\boxed{\text{Runge–Kutta–Formeln vom Gauß–Typ}}$!

- **Bemerkung 3.14:**

  1. QF von Radau := QF maximaler algebr. Genauigkeit $(2l-1)$:
     $$c_1 = 0 \text{ oder } c_l = 1.$$

     $\Rightarrow$ $\boxed{\text{Runge–Kutta–Formeln vom Radau–Typ}}$:

     (32) • $\sum\limits_{j=1}^{l} b_j c_j^k = \dfrac{1}{k+1}, \quad k = 0,1,\ldots,p-1 \text{ mit } p = 2l-1.$

     (33) • a) $c_1 = 0$: $\quad a_{11} = a_{12} = \ldots = a_{1l} = 0 \quad$ (Bd. !)
     $(\Rightarrow K_1 = f(t,u)$ bzw. $g_1 = u)$

     (34) • b) $c_l = 1$: $\quad a_{1l} = a_{2l} = \ldots = a_{ll} = 0 \quad$ (Bd. !)
     $(\Rightarrow K_l$ bzw. g_l explizit berechenbar !)

     (35) • $\sum\limits_{j=1}^{l} a_{ij} c_j^k = \dfrac{c_i^{k+1}}{k+1}, \quad k = 0,1,\ldots,r-1 \text{ mit } r = \begin{cases} l, & \text{a)} \\ l-1, & \text{b)} \end{cases}$
     $i = 1,2,\ldots,l.$

     $\Rightarrow$ Konsistenzordnung = $2l-1$

  2. QF von Lobatto := QF maximaler algebr. Genauigkeit $(2l-2)$:
     $$c_1 = 0 \text{ und } c_l = 1.$$

     $\Rightarrow$ $\boxed{\text{Runge–Kutta–Formeln vom Lobatto–Typ}}$

     • (32) für $p = 2l-2$ • (33) + (34) • (35) für $r = l-1$

     $\Rightarrow$ Konsistenzordnung = $2l-2$.

3.3.4 Konvergenztheorie für Einschrittverfahren

■ **Def. 3.15:** (ESV)

Ein Verfahren zur Lösung des AWP

$$(1) \quad \begin{cases} u'(t) = f(t, u(t)), \quad t \in I = [0, T] \\ \text{AB: } u(0) = u_0 \end{cases}$$

heißt Einschrittverfahren (ESV), falls es von der Form

$$(36) \quad u_{j+1} = u_j + h_j \varphi(t_j, u_j, h_j), \quad j = 0, 1, \ldots, m-1,$$

ist, mit geg. AB u_0.

Bsp.: Die bisher behandelten RK–Formeln (expliziten wie auch impliziten) sind offenbar ESV. Tatsächlich,

EPZV: $\varphi = f$, RK: $\varphi(t_j, u_j, h_j) = \sum\limits_{i=1}^{l} b_j f(t_j + c_i h_i, g_i), \quad g = \Phi(g, t_j, u_j, h_j).$

■ **Analog zum Pkt. 3.3.1 ($\Rightarrow$ EPZV) können wir auch (36) als Operator–gleichung in der Form**

$$(36) \quad \boxed{\text{Ges. } u_h \in X_h : F_h(u_h) = 0 \text{ in } Y_h}$$

schreiben mit

$$F_h(v_h)\,(t_{j+1}) := \begin{cases} \overbrace{\frac{1}{h_j}(v_{j+1} - v_j)}^{=:D_h v_h(t_j)} - \varphi(t_j, v_j, h_j) & , j = 0, 1, \ldots, m-1, \\ v_0 - u_0 & , j = -1, \end{cases}$$

und (z.B.) $X_h = C^1(I_h)$ und $Y_h = C(I_h)$.

Der lokale Abschneidefehler (Approximationsfehler) wird dann durch (vgl. (9))

$$(37) \quad \tau_h(t_{j+1}) = F_h(u)(t_{j+1}) \equiv \begin{cases} \dfrac{u(t_{j+1}) - u(t_j)}{h_j} - \varphi(t_j, u(t_j), h_j) & , j = \overline{0, m-1}, \\ 0 & , j = -1, \end{cases}$$

eingeführt oder kurz: $\tau_h = F_h(u)$, wobei u Lsg. des AWP (1) ist.

■ **Konsistenz** (lokale und globale):

- **Definition 3.16:** (globale Konsistenz(–ordnung))

  > Ein ESV (36) heißt mit dem AWP (1) konsistent, falls
  > $$\|\tau_h\|_{Y_h} \longrightarrow 0 \text{ für } h \equiv |h| \to 0.$$
  > Ein ESV (36) hat die Konsistenzordnung p, falls
  > $$\|\tau_h\|_{Y_h} = O(h^p).$$

- Bemerkungen:
  - In den Pkt. 3.3.2.2 (explizite) und 3.3.3.4 (implizite) wurde die (lokale) Konsistenzordnung von RK–Formeln studiert.
  - RK–Formeln: $\sum\limits_{j=1}^{l} b_j = 1 \Rightarrow$ Konsistenz (Ordnung = 1, falls $f \in C^1$).
  - Für (36) bedeutet Konsistenz (vgl. auch Bew. Satz 3.17)
  $$\varphi\,(t, u, h) \longrightarrow f(t, u) \text{ für } h \to 0$$
  bzw. präziser ($\|\tau_h\|_{Y_h} := \max\limits_{j=0,m} |\tau_j| \equiv \max\limits_{j=1,m-1} |\tau_{j+1}|,\quad m = m(h)$!)
  $$(38) \qquad \max_{j=0,m-1} |\varphi(t_j, u(t_j), h_j) - f(t_j, u(t_j))| \underset{h\to 0}{\longrightarrow} 0.$$

- **Satz 3.17:**

  > Vor.: 1) $f \in C(I \times \mathbb{R}^N) \equiv [C(I \times \mathbb{R}^N)]^N$ – stetig;
  > 2) Es gelte (38).
  >
  > Bh.: Dann ist das ESV konsistent.

  Beweis: folgt sofort aus der Darstellung
  $$\tau_h(\underbrace{t+h}_{\substack{t_{j+1}\\ t=t_j}}) = \left[\frac{1}{h}(u(t+h) - u(t)) - u'(t)\right] + [f(t, u(t)) - \varphi\,(t, u(t), h)].$$

  q.e.d.

■ Stabilität:

- Btr. zunächst die Auswirkung eines Fehlers im Startwert ($\Rightarrow$ Stabilität bzgl. der AB):

$$(39)\quad \begin{cases} v_0 &= u_0 + \delta_{\mathbf{0}}, \\ v_{j+1} &= v_j + h_j\,\varphi\,(t_j, v_j, h_j), \quad j = 0, 1, \dots, m-1. \end{cases}$$

- Frage: $\;|\underset{\substack{\uparrow \\ (36)}}{u_j} - \underset{\substack{\uparrow \\ (39)}}{v_j}| \leq \;?\quad j = 1, 2, \dots, m.$

Lemma 3.18:

> Vor.: φ sei Lipschitz–stetig im zweiten Argument, d.h.
> $\exists\, L = \text{const.} \in [0, \infty)$:
>
> (40) $|\varphi\,(t, v, h) - \varphi\,(t, w, h)| \leq L\,|v - w|$
> $\forall t \in I,\;\; \forall v, w \in \mathbb{R}^N,\;\; \forall$ zulässigen $h \in \{h_0, h_1, \dots, h_{m-1}\}$.
>
> Bh.: Dann gilt für (39) die Abschätzung
>
> (41) $|u_j - v_j| \leq e^{Lt_j}\,|\delta_0|,\;\; |D_h u_j - D_h v_j| \leq L\,|u_j - v_j|,$
> mit $D_h v_j = v_{t,j} = h_j^{-1}(v_{j+1} - v_j)$.

Beweis: Btr. Fehler $z_j = u_j - v_j$:

∘ $z_{j+1} = z_j + h_j[\varphi\,(t_j, u_j, h_j) - \varphi(t_j, v_j, h_j)]$

$$\Rightarrow |z_{j+1}| \leq (1 + Lh_j)|z_j| \leq (1 + Lh_j) = \left[(1 + Lh_j)^{\frac{1}{Lh_j}}\right]^{Lh_j} \leq e^{Lh_j}|z_j|$$
$$\leq e^{Lh_j}\, e^{Lh_{j-1}} \cdots e^{Lh_0}|z_0| = e^{Lt_{j+1}}|\delta_0|.$$

∘ $D_h u_j - D_h v_j = \varphi\,(t_j, u_j, h_j) - \varphi\,(t_j, v_j, h_j)$
$\Rightarrow |D_h u_j - D_h v_j| \leq L\,|u_j - v_j|.$

q.e.d.

- Btr. nun die Fortpflanzung der einzelnen lokalen Fehler $\boldsymbol{\delta_{j+1} = h_j y_{j+1}}$ durch das ESV:

$$(42)\quad \begin{cases} v_0 &= u_0 + \mathbf{y_0} \\ v_{j+1} &= v_j + h_j\,\varphi\,(t_j, v_j, h_j) + \underbrace{\mathbf{h_j y_{j+1}}}_{\delta_{j+1}}, \quad j = \overline{0, m-1} \end{cases}$$

Frage: $\;|\underset{\substack{\uparrow \\ (36)}}{u_j} - \underset{\substack{\uparrow \\ (42)}}{v_j}| \leq \;?$

Lemma 3.19:

Vor.:	$\lvert\varphi(t,v,h)-\varphi(t,w,h)\rvert \le L\lvert v-w\rvert \quad \forall\, t; v, w; h.$
Bh:.	Dann gelten für (42) die Abschätzungen:
(43)	$\lvert u_j - v_j\rvert \le e^{Lt_j}\lvert y_0\rvert + \dfrac{e^{Lt_j}-1}{L} \max\limits_{i=\overline{1,j}} \lvert y_i\rvert,$
(44)	$\lvert D_h u_j - D_h v_j\rvert \le L\,\lvert u_j - v_j\rvert + \lvert y_{j+1}\rvert.$

Beweis:

- Def. für fixiertes $i \in \{0, 1, \ldots, m-1\}$ Gitterfkt. $v^{(i)}$:

$$(45) \quad \begin{cases} v_{j+1}^{(i)} = v_j^{(i)} + h_j\,\varphi(t_j, v_j^{(i)}, h_j), \quad j = i, i+1, \ldots, m-1, \\ v_i^{(i)} \;= v_i := \underbrace{v_{i-1} + h_{i-1}\,\varphi(t_{i-1}, v_{i-1}, h_{i-1})}_{=:v_i^{(i-1)}} + h_{i-1}y_i, \end{cases}$$

  wobei $h_{-1} = 1$ und $v_0^{(-1)} = u_0$ gesetzt wird.

- Btr. $v_j^{(i-1)}$:

$$(46) \quad \begin{cases} v_{j+1}^{(i-1)} = v_j^{(i-1)} + h_j\varphi(f_j, v_j^{(i-1)}, h_j), \quad j = i-1, i, i+1, \ldots, m-1, \\ v_{i-1}^{(i-1)} = v_{i-1}. \end{cases}$$

- Aus Lemma 3.18: folgt:

$$(47) \quad \lvert v_j^{(i-1)} - v_j^{(i)}\rvert \le e^{L(t_j - t_i)} h_{i-1}\lvert y_i\rvert, \quad j = \overline{i,m} \text{ und } i = \overline{1,m},$$

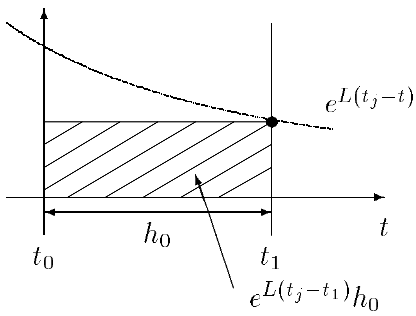
  wobei $v_m^{(m)} = v_m := v_{m-1} + h_{m-1}\varphi(t_{m-1}, v_{m-1}, h_{m-1}) + h_{m-1}y_m$.

- Für den Gesamtfehler folgt daraus:

$$\lvert u_j - v_j\rvert \le \underbrace{\lvert u_j - v_j^{(0)}\rvert}_{\substack{\le e^{Lt_j}\lvert y_0\rvert \\ \uparrow \\ \text{Lemma 3.18}}} + \lvert v_j^{(0)} - \underbrace{v_j^{(j)}}_{=v_j}\rvert$$

  Btr.

$$\begin{aligned} \lvert v_j^{(0)} - v_j^{(j)}\rvert &\le \lvert v_j^{(0)} - v_j^{(1)}\rvert + \lvert v_j^{(1)} - v_j^{(2)}\rvert + \ldots + \lvert v_j^{(j-1)} - v_j^{(j)}\rvert \\ &\le e^{L(t_j - t_1)} h_0 \lvert y_1\rvert + e^{L(t_j - t_2)} h_1 \lvert y_2\rvert + \ldots + e^{L(t_j - t_j)} h_{j-1}\lvert y_j\rvert \\ &\le [e^{L(t_j - t_1)} h_0 + e^{L(t_j - t_2)} h_1 + \ldots + e^{L(t_j - t_j)} h_{j-1}] \max_{i=\overline{1,j}} \lvert y_i\rvert \end{aligned}$$



$$\leq \left[\int\limits_{t_0=0}^{t_1} e^{L(t_j-t)}\,dt + \ldots + \int\limits_{t_{j-1}}^{t_j} e^{L(t_j-t)}\,dt\right] \max_{i=\overline{1,j}} |y_i|$$

$$= \int\limits_0^{t_j} e^{L(t_j-t)}\,dt \cdot \max_{i=\overline{1,j}} |y_i| = \frac{e^{Lt_j}-1}{L} \max_{i=\overline{1,j}} |y_i|$$

$$\Longrightarrow |u_j - v_j| \leq e^{Lt_j}|y_0| + \frac{e^{Lt_j}-1}{L} \max_{i=\overline{1,j}} |y_i|$$

◦ $D_h u_j - D_h v_j = \varphi\,(t_j, u_j, h_j) - \varphi\,(t_j, v_j, h_j) - y_{j+1}$

$\Longrightarrow |D_h u_j - D_h v_j| \leq L\,|u_j - v_j| + |y_{j+1}|.$

q.e.d.

- Das gestörte ESV (42) läßt sich kompakt in der Form

(48) $\qquad F_h\,(v_h) \; = \; y_h$

schreiben. Aus <u>Lemma 3.19</u> folgt sofort:

<u>Satz 3.20:</u>

<u>Vor.:</u> $|\varphi\,(t, v, h) - \varphi\,(t, w, h)| \leq L|v - w| \quad \forall t; v, w; h$

<u>Bh.:</u> Dann gibt es eine positive Konstante C:

(49) $\|v_h - w_h\|_{X_h} \leq C\,\|F_h(v_h) - F_h(w_h)\|_{Y_h} \quad \forall\, v_h, w_h \in X_h,$

mit $X_h = C^1(I_h) = C^1(I_h, I\!R^N)$ und $Y_h = C(I_h)$.

<u>Beweis:</u>

◦ Sei zunächst: $w_h = u_h \;:\; F_h(u_h) = 0 \qquad (36)$

$v_h \;:\; F_h(v_h) = y_h$

Dann folgt aus Lemma 3.19 sofort:

$$\max_{j=\overline{0,m}} |v_j - u_j| \overset{(43)}{\leq} C_1 \max_{j=\overline{0,m}} |y_j|.$$

mit $C_1 = e^{LT} + \left(e^{LT} - 1\right)/L$. Aus (44) und (43) folgt

$$\max_{j=\overline{0,m-1}} |D_h u_j - D_h v_j| \leq \underbrace{[Le^{LT} + e^{LT} - 1 + 1]}_{=:C_2} \max_{j=\overline{0,m}} |y_j|.$$

Daraus ergibt sich dann:

$$\|v_h - u_h\|_{X_h} := \max_{j=0,m} |v_j - u_j| + \max_{j=0,m-1} |D_h v_j - D_h u_j| \leq \overbrace{(C_1 + C_2)}^{C} \|y_h\|_{Y_h}.$$

○ Für fix. bel. $w_h \in X_h$ folgt (49) aus dem eben bew. Spezialfall angewandt auf

$\widetilde{F}_h(\cdot) := F_h(\cdot) - z_h$ mit $z_h = F_h(w_h) \Rightarrow \widetilde{F}_h(v_h) = y_h - z_h \equiv F(v_h) - F(w_h)$
$\widetilde{F}_h(w_h) = 0$

q.e.d.

- **<u>Definition 3.21:</u>** (Stabilität)

> Ein Einschrittverfahren heißt <u>stabil</u>, falls $\exists\, C = \text{const.} > 0,\ \ C \neq C(h)$:
> (50) $\quad \|v_h - w_h\|_{X_h} \leq C\|F_h(v_h) - F_h(w_h)\|_{Y_h} \quad \forall v_h, w_h \in X_h.$

- <u>Bemerkung:</u>

Seien $y_h, z_h \in Y_h$ beliebig, und erfüllen $v_h, w_h \in X_h$ die Gleichungen

$$F_h(v_h) = y_h \;\text{ und }\; F_h(w_h) = z_h.$$

Dann gilt für ein <u>stabiles</u> ESV

$$\|v_h - w_h\|_{X_h} \leq C\|y_h - z_h\|_{Y_h}.$$

Die Stabilität des ESV bedeutet also, daß sich die Lösung der Näherungsgleichung (Lipschitz–) stetig und gleichmäßig in h mit den Daten (den rechten Seiten) ändert !

Satz 3.20 zeigt also, daß die

Lipschitz–Stetigkeit von $\varphi(\cdot, \bullet, \cdot)$

ein hinreichendes Kriterium für die Stabilität des ESV ist.

■ **<u>Konvergenz:</u>** $e_h = u - u_h \longrightarrow 0$ in X_h für $h = |h| \to 0$.

- **<u>Definition 3.22:</u>** (diskrete Konvergenz)

> Ein ESV heißt <u>konvergent</u>, falls
> $\quad \|e_h\|_{X_h} \equiv \|u - u_h\|_{X_h} \overset{h\to 0}{\longrightarrow} 0.$
> Falls
> $\quad \|e_h\|_{X_h} = O(h^p), \quad h = |h|,$
> nennt man das ESV ein Verfahren der Konvergenzordnung p.
> Hierbei ist $e_h = u - u_h$ der Fehler.
> (1) (36)

- **Satz 3.23:** ($\overset{\text{Approximation}}{\text{Konsistenz}_{Y_h}}$ + Stabilität$_{(Y_h,X_h)}$ $\Rightarrow$ diskr. Konvergenz$_{X_h}$)

Vor.:	1) ESV (36) sei stabil i. S. Def. 3.21.
	2) ESV(36) sei konsistent (von der Ordnung p) i. S. Def. 3.16.
Bh.:	Dann ist das ESV (36) konvergent (von der Ordnung p).

Beweis: trivial !

- $u \in X|_{I_h} \subset X_h : F_h(u) = \tau_h$
(1)
- $u_h \in X_h : F_h(u_h) = 0$
- Stabilität: $\Rightarrow \| \underbrace{u - u_h}_{=e_h} \|_{X_h} \le C \, \|\tau_h\|_{Y_h}$

  +
- Konsistenz: $\Rightarrow$ a) $\|\tau_h\|_{Y_h} \to 0, \;\; h \to 0$

  $\Downarrow$ b) $\|\tau_h\|_{Y_h} = O(h^p)$

  diskrete Konvergenz

Konvergenz: $\Rightarrow \;\; \|e_h\|_{X_h} \to 0$, $h \to 0$

$\Rightarrow \|e_h\|_{X_h} = O(h^p)$.

q.e.d.

■ Anwendung der Theorie auf (implizite) RK–Formeln:

- RK–Formeln sind ESV mit

$$(*) \qquad \varphi\,(t,u,h) = \sum_{j=1}^{l} b_j f(t + c_j h, g_j(t,u,h)),$$

wobei $g = (g_1, \ldots, g_l)^T$ Lösung der Fixpunktgleichung

$$(**) \qquad g = \Phi\,(g,t,u,h)$$

ist, mit $\Phi = (\Phi_1, \ldots, \Phi_l)^T, \;\; \Phi_i = \Phi_i(g,t,u,h) = u + h \sum_{j=1}^{l} a_{ij} f(t + c_j h, g_j)$.

- Konsistenz ($\uparrow$):

Satz 3.24: (Konsistenz)

Vor.:	1) f sei stetig, beschränkt und erfülle die Lipschitz–Bedingung $\lvert f(t,v) - f(t,w)\rvert \le L\lvert v - w\rvert \quad \forall\, t; v, w.$
	2) Für die Gewichte b_j gelte: $\sum_{j=1}^{l} b_j = 1.$
Bh.:	Dann ist die RK–Formel konsistent mit dem AWP (1).

Beweis: $\longrightarrow$ Satz 3.17 !

- Sei $g_i = g_i(t, u(t), h)$. Dann folgt wie im Beweis von Satz 3.9 aus

$$g_i - u(t) \overset{(**)}{=} h \sum_{j=1}^{l} a_{ij} f(t + c_j h, g_j) =$$
$$= h \sum_{j=1}^{l} a_{ij} [f(t + c_j h, g_j) - f(t + c_j h, u(t))] + h \sum_{j=1}^{l} a_{ij} f(t + c_j h, u(t))$$

  die Abschätzung

$$\max_{i=\overline{1,l}} |g_i - u(t)| \leq h\|A\|_\infty \, L \, \max_{i=\overline{1,l}} |g_i - u(t)| + h\|A\|_\infty \, M$$

  mit $M = \sup\limits_{\tau, t \in I} |f(\tau, u(t))|$.

  Daher gilt für hinreichend kleine h

$$\max_{i=\overline{1,l}} |g_i - u(t)| \leq \frac{h\|A\|_\infty \, M}{1 - h\|A\|_\infty \, L} \, .$$

- Aus

$$\varphi(t, u(t), h) - f(t, u(t)) \overset{(*),\, \sum b_j = 1}{=}$$
$$= \sum_{j=1}^{l} b_j [f(t + c_j h, g_j) - f(t + c_j h, u(t))] + \sum_{j=1}^{l} b_j [f(t + c_j h, u(t)) - f(t, u(t))]$$

  folgt dann

$$|\varphi(t, u(t), h) - f(t, u(t))| \leq \frac{\overbrace{\|b\|_1}^{=\sum |b_j|} \, L h \|A\|_\infty \, M}{1 - h\|A\|_\infty \, L} + $$
$$+ \|b\|_1 \max_{j=\overline{1,l}} |f(t + c_j h, u(t)) - f(t, u(t))|$$

  und daher (beachte: $f(\tau, u(t))$ ist gleichmäßig stetig !)

$$\max_{k=\overline{0,m-1}} |\varphi(t_k, u(t_k), h_k) - f(t_k, u(t_k))| \longrightarrow 0 \text{ für } |h| \to 0.$$

- Der Rest folgt sofort aus Satz 3.17 !

q.e.d.

$\boxed{\textbf{Ü 3.9}}$ Man zeige, daß im Falle expliziter RK–Formeln die Lipschitz–Bd. für den Konsistenzbew. nicht benötigt wird !

Bem.: Falls f entsprechend oft differenzierbar ist, dann kann man Aussagen über die Konsistenzordnung von RK–Formeln gewinnen:

siehe Pkt. 3.3.2.2: explizite RK–Formeln,
Pkt. 3.3.3.4: implizite RK–Formeln.

- Stabilität:

Satz 3.25: (Stabilität)

Vor.:	f sei stetig und erfülle die Lipschitz–Bedingung. $\qquad |f(t,v) - f(t,w)| \le L|v-w| \quad \forall\, t; v, w.$
Bh.:	Dann ist die RK–Formel stabil.

Beweis: $\longrightarrow$ Satz 3.20 $\Rightarrow$ z.Z.: $|\varphi(t,v,h) - \varphi(t,w,h)| \le \widetilde{L}|v-w|$!

- Seien t, v, w, h bel.: $\Rightarrow$

$$|\varphi(t,v,h) - \varphi(t,w,h)| \overset{(*)}{\le} \sum_{j=1}^{l} |b_j|\, |f(t+c_j h, g_j(t,v,h)) - f(t+c_j h, g_j(t,w,h))| \le$$
$$\le \sum_{j=1}^{l} |b_j|\, L\, |g_j(t,v,h) - g_j(t,w,h)|.$$

- Wie im Beweis von Satz 3.9 folgt:

$$\max_{j=\overline{1,l}} |g_j(t,v,h) - g_j(t,w,h)| \overset{(**)}{\le} |v-w| + h\, \|A\|_\infty\, L \max_{j=\overline{1,l}} |g_j(t,v,h) - g_j(t,w,h)|$$
$$\Downarrow$$
$$\max_{j=\overline{1,l}} |g_j(t,v,h) - g_j(t,w,h)| \le \frac{1}{1 - h\|A\|_\infty L} |v-w|.$$

- Insgesamt folgt damit die Lipschitz–Bed. für φ:

$$|\varphi(t,v,h) - \varphi(t,w,h)| \le \underbrace{\frac{L\|b\|_1}{1 - h\|A\|_\infty L}}_{=:\widetilde{L}} |v-w|.$$

- Der Rest folgt aus Satz 3.20 !

q.e.d.

- Konvergenz:

Konsistenz+Stabilität $\Rightarrow$ Konvergenz
Aus den Sätzen 3.24, 3.25 und 3.23 folgt sofort:

Satz 3.26: ($\hat{=}$ Satz 3.23 !)

Vor.: 1) f sei stetig und erfülle die Lipschitz–Bedingung:

$$|f(t,v) - f(t,w)| \leq L|v-w| \quad \forall\, t; v, w.$$

2) Für die Gewichte b_j gelte:

$$\sum_{j=1}^{l} b_j = 1.$$

Bh.: Dann ist die RK–Formel für das AWP (1) konvergent.

Bemerkung:
Falls die RK–Formel die Konsistenzordnung p hat, und f entsprechend oft stetig differenzierbar ist, dann ist auch die Konvergenzordnung p.

Bemerkung:

- Die obigen Aussagen gelten auch unter der lokalen Lipschitz–Bd.

$$|f(t,v) - f(t,w)| \leq L|v-w| \quad \forall (t,v), (t,w) \in U,$$

wobei $U \subset I \times \mathbb{R}^N$ eine Umgebung des Graphen von f

$$\{(t, f(t, u(t))) : t \in I\}$$

ist und $u(t)$ die exakte Lsg. des AWP (1) bezeichnet.

- Für eine lokale Variante des Stabilitätssatzes 3.25 genügt ebenfalls eine lokale Lipschitz–Bed. der Art

$$|\varphi(t,v,h) - \varphi(t,w,h)| \leq L|v-w| \quad \forall (t,v), (t,w) \in U, \quad h \leq H,$$

falls zusätzlich die Konsistenz des Verfahrens vorausgesetzt wird. Man kann die Aussagen $\forall v_h, w_h \in X_h : \|F_h(v_h)\|_{Y_h}, \|F_h(w_h)\|_{Y_h} \leq \eta$ mit η – hinr. klein, zeigen.

3.3.5 Praktische Durchführung von Einschrittverfahren

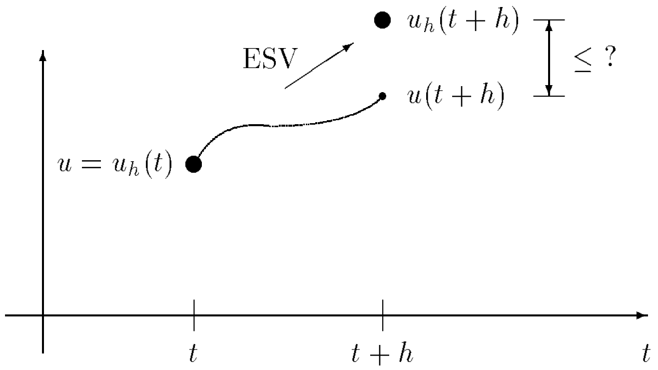
- Die Wahl der **richtigen Schrittweiten** $\{h_j\}$ ist offenbar von entscheidender Bedeutung für eine effiziente Durchführung eines ESV. Die Aufgabe einer Schrittweitensteuerung ist es, den jeweils neu entstehenden **lokalen Fehler** in einer gewünschten Größenordnung zu halten.

- Im folgenden werden zwei Möglichkeiten zur **Schätzung der lokalen Fehler** angegeben (Pkt. 3.3.5.1), darauf aufbauend eine **Schrittweitensteuerung** vorgeschlagen (Pkt. 3.3.5.2) und schließlich eine Technik zur **Berechnung von Näherungen an vorgegebenen Punkten** $t + \theta h$, die nicht von der Schrittweitensteuerung herrühren (Pkt. 3.3.5.3).

3.3.5.1 Schätzung der lokalen Fehler

- **Betrachten den lokalen Fehler**

$u(t+h) - u_h(t+h)$



wobei $\quad u(t+h) \quad$ Lösung d. AWP: $\left\{\begin{array}{lcll} u'(s) & = & f(s,u(s)), & s \geq t, \\ u(t) & = & u & \end{array}\right.$

im Pkt. $t+h$, und

$u_h(t+h) = u + h\varphi\,(t,u,h)$ – ESV.

Es wird nun vorausgesetzt, daß das ESV die Konsistenzordnung p hat, d.h.

$$u(t+h) - u_h(t+h) = O(h^{p+1}),$$

und der Fehlerterm die Gestalt

$$(51) \qquad u(t+h) - u_h(t+h) = \underbrace{c(t,u)\,h^{p+1}}_{\substack{=\text{ führender Fehlerterm} \\ \text{(engl.: principal error term)}}} + O(h^{p+2})$$

hat, wobei c unabhängig von der Schrittweite h sein soll.

Für Runge–Kutta–Formeln läßt sich diese Darstellung des lokalen Fehlers durch Taylor–Entwicklung leicht nachweisen, z.B.:

- explizites Euler–Verfahren (vgl. Pkt. 3.3.1):

$$u(t+h) - u_h(t+h) = \underbrace{\frac{1}{2}\,(f_t + f_u f)\,(t,u)}_{=:\,c\,(t,u)}\,h^2 + O(h^3),$$

- explizite Mittelpunktsregel (vgl. Pkt. 3.3.2):

$$u(t+h) - u_h(t+h) = \underbrace{\frac{1}{24}\,(f_{tt} + 2f_{tu}f + f_{uu}f^2 + 4\,(f_u f_t + f_u^2 f))\,(t,u)}_{=:\,c\,(t,u)}\,h^3 + O(h^4).$$

Im folgenden wird nur von der Existenz eines von h unabhängigen, in t und u hin–reichend glatten (!) Koeffizienten $c\,(t,u)$, nicht aber von seiner speziellen Gestalt Gebrauch gemacht !!

■ Schätzung durch Extrapolation:

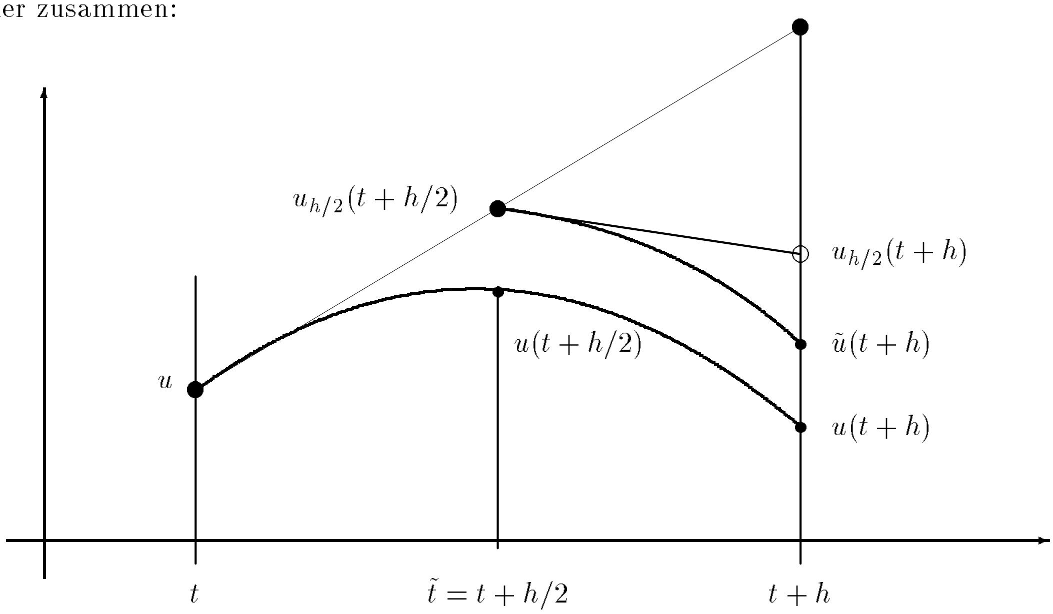
- Zuerst wird ein Schritt mit der Schrittweite h durchgeführt. Man erhält eine Näherung u_h im nächsten Gitterpunkt $t+h$ und wegen (51) gilt:

$$u(t+h) - u_h(t+h) = c\,(t,u)\,h^{p+1} + O(h^{p+2}).$$

- Zwei Schritte mit halber Schrittweite $h/2$ führen ebenfalls zum Gitterpunkt $t+h$, erzeugen aber eine andere Näherung $u_{h/2}(t+h)$ im Punkt $t+h$. Für den ersten Schritt gilt wegen (51):

$$u(t+h/2) - u_{h/2}(t+h/2) = c\,(t,u)\left(\frac{h}{2}\right)^{p+1} + O(h^{p+2}).$$

  Der Fehler nach dem zweiten Schritt setzt sich aus der Fortpflanzung des ersten (↑) lokalen Fehlers (= Fehler in den AB: siehe Lemma 3.18) und dem neuen lokalen Fehler zusammen:



$$
\begin{aligned}
(52)\qquad & u(t+h) - u_{h/2}(t+h) = \\
& = [u(t+h) - \tilde{u}(\tilde{t}+h/2)] + [\tilde{u}(t+h/2) - u_{h/2}(t+h)] = \\
& \overset{\text{(mms)}}{\downarrow} \\
& = (1+O(h))\, c\,(t,u) \left(\tfrac{h}{2}\right)^{p+1} + (c\,(t,u)+O(h)) \left(\tfrac{h}{2}\right)^{p+1} + O(h^{p+1}) = \\
& = 2c\,(t,u) \left(\tfrac{h}{2}\right)^{p+1} + O(h^{p+1})
\end{aligned}
$$

- Aus (51) und (52) kann nun $c \equiv c\,(u,t)$ offenbar eliminiert werden:

$$
\begin{array}{r|lclclcll}
(-1) & u(t+h) & - & u_h(t+h) & = & ch^{p+1} & + & O(h^{p+2}) & (51) \\
2^p * & u(t+h) & - & u_{h/2}(t+h) & = & 2c\left(\frac{h}{2}\right)^{p+1} & + & O(h^{p+2}) & (52) \\
\hline
+ & \multicolumn{8}{l}{(2^p-1)\,u\,(t+h) - 2^p u_{h/2}(t+h) + u_h(t+h) = O(h^{p+2})}
\end{array}
$$

Daraus folgt sofort, daß

$$\hat{u}_h(t+h) := u_{h/2}(t+h) + \frac{u_{h/2}(t+h) - u_h(t+h)}{2^p - 1}$$

eine bessere Näherung für die Lösung

$$u(t+h) = \hat{u}_h(t+h) + O(h^{p+2})$$

ist, d.h. mit einem um eine Ordnung verbesserten Fehler.

- Eine Schätzung err ($\rightarrow$ führender Fehlerterm) des lokalen Fehlers erhält man nun, indem man $u(t+h)$ durch den extrapolierten Wert $\hat{u}_h(t+h)$ ersetzt:

$$err := \frac{|\hat{u}_h(t+h) - u_h(t+h)|}{d} \;,$$

wobei

$$d = \begin{cases} 1 & \text{für den absoluten Fehler,} \\ |u_{h/2}(t+h)| & \text{für den relativen Fehler,} \\ \max\{|u_{h/2}(t+h)|, 1\} & \text{für ein gemischtes Fehlermaß} \end{cases}$$

ein Skalierungsfaktor ist. Gebräuchlich ist auch eine komponentenweise Anwendung der obigen Skalierungsfaktoren.

- Um den Mehraufwand zur Schätzung des lokalen Fehlers auch zumindestens zum Teil für das Verfahren zu nutzen, betrachtet man entweder $u_{h/2}(t+h)$ oder sogar den extrapolierten Wert $\hat{u}_h(t+h)$ als die Näherung im nächsten Gitterpunkt $t+h$ ($\rightarrow$ lokale Extrapolation).
Der berechnete Fehler err kann jedoch nur als Schätzung des auf die ursprünglichen Werte definierten Verfahrens interpretiert werden. Es darf allerdings erwartet werden, daß die Verwendung von $\hat{u}_h(t+h)$ anstelle von $u_h(t+h)$ die Genauigkeit des Verfahrens insgesamt eher erhöht. Die konsequente Verfolgung dieser Idee führt zu den Extrapolationsverfahren (siehe z.B. [1], [5]).

■ Schätzung mittels eingebetteter Runge–Kutta–Formeln:

- Grundidee:

  Neben der das ESV bestimmende RK–Formel der Ordnung p wird eine weitere $\widehat{\text{RK}}$–Formel mindestens der Ordnung $p+1$ konstruiert. Die erste Formel liefert einen lokalen Fehler
  $$u(t+h) - u_h(t+h) = O(h^{p+1}),$$
  für die zweite Formel gilt
  $$u(t+h) - \hat{u}_h(t+h) = O(h^{p+2}),$$
  und somit folgt
  $$u(t+h) - u_h(t+h) = \hat{u}_h(t+h) - u_h(t+h) + O(h^{p+2}).$$
  Daher bietet sich für den lokalen Fehler folgende Schätzung an:
  $$err = \frac{|\hat{u}_h(t+h) - u_h(t+h))}{d}$$
  mit $d = 1$ – absoluter Fehler, $d = |\hat{u}_h(t+h)|$ – relativer Fehler,
  $d = \max\{|\hat{u}_h(t+h)|, 1\}$ – gemischtes Fehlermaß.

- Eingebettete RK–Formeln:

  Um den Aufwand zur Berechnung von $\hat{u}_h(t+h)$ möglichst gering zu halten, wählt man die gleichen Werte für c und A in den beiden Tableaus. Man spricht dann von einer eingebetteten Runge–Kutta–Formel (engl.: embedded RK–formula) und stellt die Koeffizienten in einem gemeinsamen Tableau dar, z.B. im Fall expliziter Formeln:
  $$\begin{array}{c|ccccc}
  0 & & & & & \\
  c_2 & a_{21} & & & & \\
  c_3 & a_{31} & a_{32} & & & \\
  \vdots & \vdots & \vdots & \ddots & & \\
  c_l & a_{l1} & a_{l2} & \dots & a_{l,l1} & \\
  \hline
   & b_1 & b_2 & \dots & b_{l-1} & b_l \\
  \hline
   & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_{l-1} & \hat{b}_l
  \end{array}$$
  Die beiden Näherungen sind durch
  $$u_h(t+h) = u + h(b_1 K_1 + b_2 K_2 + \ldots + b_l K_l)$$
  und
  $$\hat{u}_h(t+h) = u + h(\hat{b}_1 K_1 + \hat{b}_2 K_2 + \ldots + \hat{b}_l K_l)$$
  gegeben.

- **Beispiel:** Runge–Kutta–Fehlberg Formel [RKF 2(3)]: Ausgangspunkt ist das allgemeine (expl.) Tableau für $l = 3$:

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \\ \hline & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 \end{array}$$

Die Bedingung, daß die erste (eigentliche) Formel die Ordnung 2 besitzt, lautet (siehe Pkt. 3.3.2.2: $c_i = \sum a_{ij}$):

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2\, c_2 + b_3\, c_3 &= 1/2. \end{aligned}$$

Damit die zweite (zur Schätzung benutzte) Formel die Ordnung 3 besitzt, muß gelten (siehe Pkt. 3.3.2.2: $c_i = \sum a_{ij}$):

$$\begin{aligned} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 &= 1, \\ \hat{b}_2\, c_2 + \hat{b}_3\, c_3 &= 1/2, \\ \hat{b}_2\, c_2^2 + \hat{b}_3\, c_3^2 &= 1/3, \\ \hat{b}_3\, a_{32}\, c_2 &= 1/6. \end{aligned}$$

Für die Wahl

$c_2 = 1,\ \ c_3 = 1/2,\ \ b_3 = 0$

erhält man eine sogenannte Runge–Kutta–Fehlberg–Formel mit der Kurzbezeichnung **RKF 2(3)**. Das Symbol **2(3)** bringt zum Ausdruck, daß das eigentliche Verfahren von der Ordnung 2 ist und daß für die Schätzung des lokalen Fehlers ein Verfahren der Ordnung 3 verwendet wird. Entsprechend ist ein Symbol $p(q)$ zu interpretieren. Das Tableau für RKF 2(3) lautet:

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 1/2 & 1/4 & 1/4 & \\ \hline & 1/2 & 1/2 & 0 \\ \hline & 1/6 & 1/6 & 4/6 \end{array} \quad \begin{array}{l} \\ \\ \\ \leftarrow \text{ Gewichte der Trapezregel} \\ \leftarrow \text{ Gewichte der Simpsonregel} \end{array}$$

- Bemerkungen:
  1. Vom Blickwinkel der Effizienz sind eingebettete Runge–Kutta–Formeln mit
     $a_{li} = b_i,\ \ i = 1, 2, \ldots, l \quad (\curvearrowright\ b_l = a_{ll} = 0)$
     besonders günstig (= Fehlberg–Trick), da gilt:

$$\begin{aligned} K_1(t+h) &\equiv f(t+h, u_h(t+h)) \\ &= f(t + c_l h,\ u_h(t) + h \textstyle\sum a_{li}(t)\, K_i(t)) \equiv K_l(t) \end{aligned}$$

$c_l = \sum a_{li} = \sum b_i = 1 \qquad$ (Konsistenzbed.)

$u_h(t) + h \sum a_{li}(t)\, K_i(t) = u_h + h \sum b_i\, K_i = u_h(t+h)$

2. Am Beispiel der 3-stufigen, expliziten, eingebetteten RK–Formeln sieht man, daß die Lösung keineswegs eindeutig ist. Es wurde daher versucht, die frei wählbaren Parameter so zu wählen, daß der führende Fehlerterm von $u_h(t+h)$ möglichst klein wird. Das führt allerdings dazu daß err den lokalen Fehler eher unterschätzt.

3. Es liegt nahe, nicht $u_h(t+h)$, sondern den genaueren Wert $\hat{u}_h(t+h)$ für das (eigentliche) Verfahren und $u_h(t+h)$ nur für die Schätzung des lokalen Fehlers zu verwenden.
   Man versucht dann natürlich, die frei wählbaren Parameter so zu wählen, daß der führende Fehlerterm von $\hat{u}_h(t+h)$ möglichst klein wird. Dann darf damit gerechnet werden, daß err den lokalen Fehler (für die erste Formel) eher überschätzt.

4. Ein sehr wichtiges, 7-stufiges, eingebettetes Runge–Kutta–Verfahren 4(5) mit den Eigenschaften aus Bemerkung 1 und 3 wurde von J. R. Dormand und P. J. Prince konstruiert [7], [8]:

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}
$$

   In der Literatur hat dieses eingebettete Runge–Kutta–Verfahren die Kurzbezeichnung **DOPRI 4(5)**.

3.3.5.2 Schrittweitensteuerung

■ **Idee:**

Es wird nun davon ausgegangen, daß eine Schätzung err des lokalen Fehlers vorliegt, die auf einer gewählten Schrittweite h basiert und für die gilt (vgl. Pkt. 3.3.5.1):

$$err = ch^{p+1}.$$

Eigentliches Ziel ist es, daß der lokale Fehler einen vorgegebenen Wert tol nicht überschreitet. Also wäre die „optimale" Schrittweite h_{neu} durch die Bedingung

$$tol = ch_{\text{neu}}^{p+1}$$

definiert. Aus diesen beiden Bedingungen läßt sich die Unbekannte c eliminieren, und man erhält:

$$(*) \qquad h_{\text{neu}} := h \cdot \sqrt[p+1]{\frac{tol}{err}}\ .$$

Daraus leitet sich sofort der folgende Algorithmus zur Schrittweitensteuerung ab.

■ **Algorithmus zur Schrittweitensteuerung:**

1. Es wird ein Schritt mit einer vorgegebenen Schrittweite h durchgeführt und gleichzeitig eine Schätzung err für den lokalen Fehler berechnet.
2. Falls $err \le tol$, wird dieser Schritt akzeptiert und mit der neuen Schrittweite h_{neu} das Verfahren fortgesetzt.
3. Andernfalls wird der Schritt verworfen und noch einmal mit der neuen Schrittweite h_{neu} gestartet.

■ **Praktische Hinweise:**

1. Um sicher zu sein, daß die neue Schrittweite auch tatsächlich den lokalen Fehler unterhalb des Wertes tol hält, wird die optimale Schrittweite um einen Sicherheitsfaktor fac reduziert (z.B. $fac = 0.8$).
2. Weiters ist es zweckmäßig, zwischen h und h_{neu} einen maximalen Vergrößerungsfaktor $facmax$ und einen minimalen Verkleinerungsfaktor $facmin$ einzuführen, um allzu dramatische Änderungen der Schrittweite zu verhindern. Mit diesen Modifikationen wird aus $(*)$ die Formel

   $$h_{\text{neu}} = h \cdot \min\left\{facmax, \max\left\{facmin, fac \cdot \left(\tfrac{tol}{err}\right)^{1/(p+1)}\right\}\right\}.$$

3. Außerdem ist es ratsam, nach einem verworfenen Schritt zwei erfolgreiche Schritte abzuwarten, bevor wieder eine Vergrößerung der neuen Schrittweite zugelassen wird.

3.3.5.3 Berechnung von Näherungen an vorgegebenen Punkten

■ **Problematik:**

Häufig ist es erwünscht, Näherungen an vorgegebenen Punkten zu berechnen (z.B. für eine graphische Ausgabe mit gegebener Feinheit). Berücksichtigt man solche Punkte bereits bie der Schrittweitensteuerung (z.B. indem man grundsätzlich höchstens bis zum nächsten gegebenen Punkt fortschreitet), so verliert man unter Umständen an Effizienz.

■ **Kontinuierliche Runge–Kutta–Formeln:**

- Eine Möglichkeit, Näherungen an vorgegebenen Punkten effizient zu berechnen, bieten die sogenannten kontinuierlichen (engl.: continuous) RKF. Diese Formeln beinhalten einen Parameter $\theta \in (0,1]$ und liefern Näherungen für $u(t+\theta h)$.
  Für $\theta = 1$ erhält man also die eigentliche RKF. Durch spezielle Setzungen für θ kann man Näherungen an vorgegebenen Punkten mitrechnen, ohne die Schrittweitensteuerung zu beeinflussen.

- Aus Effizienzgründen fordert man, daß die Werte von c und A im Tableau unabhängig von θ sind:

$$\begin{array}{c|c} c & A \\ \hline & b(\theta) \end{array}$$

- **Beispiel:** Damit eine explizite, 3-stufige RKF für alle $\theta \in (0,1]$ von der Ordnung 3 ist, müssen folgende Bedingungen gelten (vgl. Pkt. 3.3.2.2: $c_i = \sum a_{ij}$):

$$\begin{aligned} b_1 + b_2 + b_3 &= \theta, \\ b_2\, c_2 + b_3\, c_3 &= \tfrac{\theta^2}{2}, \\ b_2\, c_2^2 + b_3\, c_3^2 &= \tfrac{\theta^3}{3}, \\ b_3\, a_{32}\, c_2 &= \tfrac{\theta^3}{6}. \end{aligned}$$

  Man sieht leicht, das es nicht möglich ist, c_2, c_3 und a_{32} unabhängig von θ zu wählen. Man fordert daher Ordnung 3 nur für $\theta = 1$ und begnügt sich sonst mit Ordnung 2. Für $c_2 = 1/2$ und $c_3 = 1$ erhält man dann folgendes Tableau einer kontinuierlichen RKF:

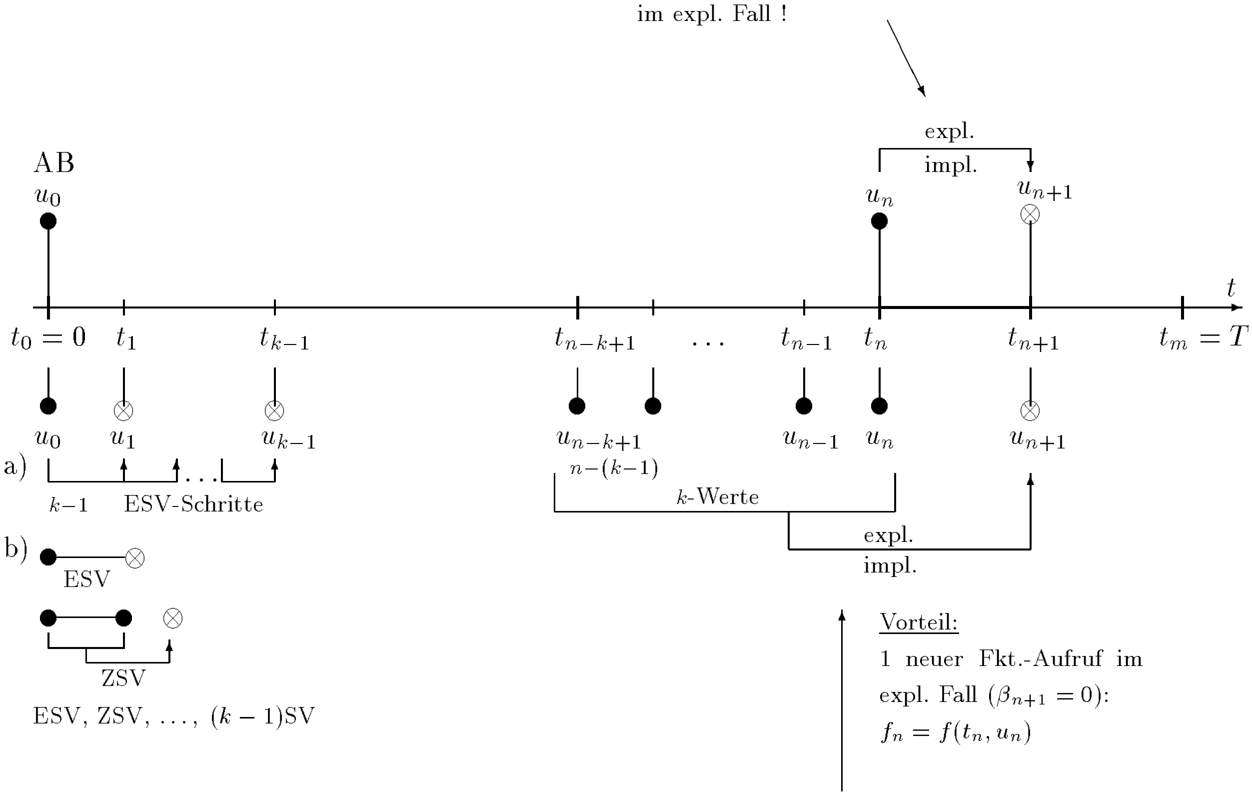
$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & \theta\left(1-\theta\left(\frac{8}{3}-\frac{11}{6}\,\theta\right)\right) & \theta^2\left(2-\frac{4}{3}\,\theta\right) & \theta^2\left(\frac{2}{3}-\frac{1}{2}\,\theta\right) \\ \theta = 1 & 1/6 & 2/3 & 1/6 \end{array}$$

3.4 Mehrschrittverfahren (MSV)

■ $\boxed{\textbf{ESV}}$ z.B. RK–Verfahren: $u_{n+1} = u_n + h_n \sum\limits_{j=1}^{l} b_j f(\underbrace{t_n + c_j h_n}, g_j), \quad n = \overline{0, m-1}$

$$g = \Phi(g, t_n, u_n, h_n)$$
$$g = (g_1, \ldots, g_l)^T$$

Nachteil:
Funktionsauswertung in den Zwischenpkt. $t_n + c_j h_n$ auch im expl. Fall !



$\boxed{\textbf{MSV}}$ z.B. lineare MSV ($h_n = h$): $\sum\limits_{j=n-k+1}^{n+1} \alpha_j u_j = h \sum\limits_{j=n-k+1}^{n+1} \beta_j f_j$

$$f_j = f_j(t_j, u_j)$$

■ **Bemerkungen:**

1. Bez. der Näherungen im MSV: $\overbrace{\underset{n-(k-1)}{u_{n-k+1}}, \ldots, u_{n-1}, u_n}^{k-\text{Werte}} \longrightarrow u_{n+1}$

$$u_n, \ldots, u_{n+k-2}, u_{n+k-1} \longrightarrow u_{n+k}$$

2. Im Startschritt: $\overset{\text{AB}}{u_0}\,, \underbrace{u_1, \ldots, u_{k-1}}_{\text{unbekannt}} \Big| \; u_k$ MSV:

   a) $k-1$ ESV–Schritte entsprechender ($\hat{=}$ MSV) Genauigkeit,
   b) ESV, ZSV, ..., $(k-1)$SV entsprechender ($\hat{=}$ MSV) Genauigkeit.

3.4.1 Spezielle Mehrschrittverfahren

■ **Ausgangspkt.:**

1) Dgl.: $\underline{u'(t)} = f(t, u(t))$

   $\uparrow$
   Numerische Differentiation eines
   Interpolationspolynoms $\Rightarrow t = t_{n+1}$

2) Igl.: $u(t) = u(t-s) + \underline{\int\limits_{t-s}^{t} f(\tau, u(\tau))\, d\tau}$

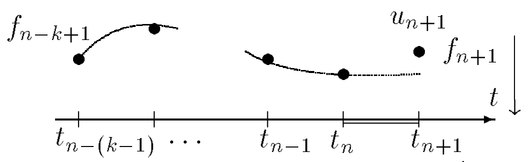
   $\uparrow$
   Numerische Integration
   beruhend auf Interpolation
   bzw. Extrapolation von f
   $\Rightarrow t = t_{n+1}$

■ **Btr. folgende 5 Klassen von MSV für $h = h_n$, $n = \overline{0, m-1}$:**

1. Adams–Bashforth–Formeln (Pkt. 3.4.1.1)

   Ausgangspunkt = 2)

   Idee: $u(t_{n+1}) = u(t_n) + \int\limits_{t_n}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$
   $p \in P_{k-1}$



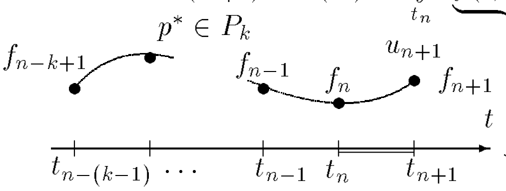
Extrapolationsquadratur mit
- Stützstellen: t_i
- Stützwerten: $f_i = f(t_i, u_i)$
  $i = n-k+1, \ldots, n-1, n$

$$u_{n+1} = u_n + h \int\limits_0^1 p(t_n + sh)\, ds$$

2. Adams–Moulton–Formeln (Pkt. 3.4.1.2)

   Ausgangspunkt = 2)

   Idee: $u(t_{n+1}) = u(t_n) + \int\limits_{t_n}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$



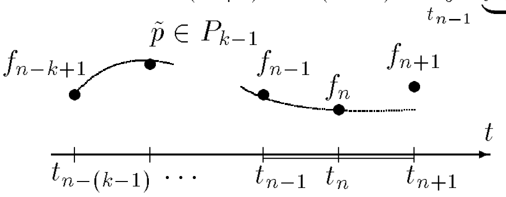
   Interpolationsquadratur mit
   - Stützstellen: t_i
   - Stützwerten: $f_i = f(t_i, u_i)$
     $i = n-k+1, \ldots, n, n+1$

$$u_{n+1} = u_n + h \int\limits_0^1 p^*(t_n + sh)\, ds$$

3. Nyström–Formeln (Pkt. 3.4.1.3)

   Ausgangspunkt = 2)

   Idee: $u(t_{n+1}) = u(t_{n-1}) + \int\limits_{t_{n-1}}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$



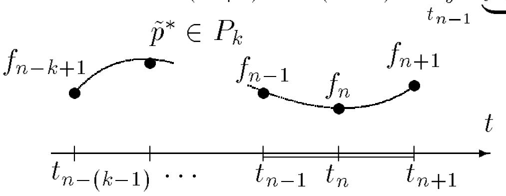
   Interpolations/Extrapolationsquadr. mit
   - Stützstellen: t_i
   - Stützwerten: $f_i = f(t_i, u_i)$
     $i = n-k+1, \ldots, n-1, n$

$$u_{n+1} = u_{n-1} + h \int\limits_{-1}^{+1} \tilde{p}(t_n + sh)\, ds$$

4. Milne–Simpson–Formeln (Pkt. 3.4.1.4)

   Ausgangspunkt = 2)

   Idee: $u(t_{n+1}) = u(t_{n-1}) + \int\limits_{t_{n-1}}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$



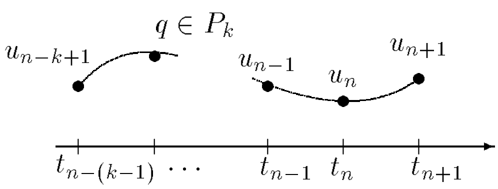
   Interpolationsquadratur mit
   - Stützstellen: t_i
   - Stützwerten: $f_i = f(t_i, u_i)$
     $i = n-k+1, \ldots, n_1, n, n+1$

$$u_{n+1} = u_{n-1} + h \int\limits_{-1}^{1} \tilde{p}^*(t_n + sh)\, ds$$

5. BDF–Verfahren (Pkt. 3.4.1.5)

   Ausgangspunkt = 1)

   Idee: $\underbrace{u'(t_{n+1})} = f(t_{n+1}, u(t_{n+1}))$



   Interpolation von $u(\cdot)$ mit
   - Stützstellen: t_i
   - Stützwerten: u_i
     $i = n-k+1, \ldots, n-1, n, n+1$

$$q'(t_{n+1}) = f_{n+1} := f(t_{n+1}, u_{n+1})$$

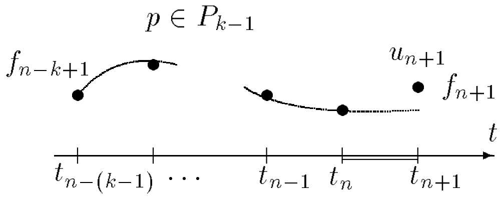
3.4.1.1 Die Adams–Bashforth–Formeln

- **Idee:** Extrapolationsquadratur mit
  - Stützstellen t_i
  - Stützwerten $f_i = f(t_i, u_i)$

  $\Big\}\ i = n-k+1, \ldots, n-1, n$

$\big\downarrow$ aus Newton-Darstellung des Interpolationspolynoms mit Hilfe dividierter Differenzen $\Rightarrow$

$$u(t_{n+1}) = u(t_n) + \int\limits_{t_n}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$$



$$p(t_n + sh) = \sum_{j=0}^{k-1} (-1)^j \tbinom{-s}{j} \nabla^j f_n$$

mit $\nabla f_n = f_n - f_{n-1}$, $\nabla^{j+1} = \nabla^j \nabla$, $\nabla^0 = I$,

$$\binom{x}{j} = \frac{x(x-1)\cdot \ldots \cdot (x-j+1)}{1 \cdot 2 \cdot \ldots \cdot j} = \frac{x!}{(x-j)!j!}$$

$$u_{n+1} = u_n + h \int\limits_0^1 p\,(t_n + sh)\, ds$$

- **Resultat:** Adams–Bashforth–Formeln ($k = 1, 2, \ldots$)

$$u_{n+1} = u_n + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f_n \quad \text{mit} \quad \gamma_j = (-1)^j \int\limits_0^1 \binom{-s}{j}\, ds$$

j	0	1	2	3	4	5	6	7	8
γ_j	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$	$\frac{1070017}{3628800}$

- **Beispiele:**

  $k = 1 : u_{n+1} = u_n + hf_n \quad \Rightarrow$ Explizite Euler–Methode,

  $k = 2 : u_{n+1} = u_n + h\left[\frac{3}{2} f_n - \frac{1}{2} f_{n-1}\right]$,

  $k = 3 : u_{n+1} = u_n + h\left[\frac{23}{12} f_n - \frac{16}{12} f_{n-1} + \frac{5}{12} f_{n-2}\right]$.

- **Bemerkungen:**

  1. Adams–Bashforth–Formeln sind explizite MSV.
  2. Konsistenzordnung $= k$ (siehe Pkt. 3.4.2).
  3. Stabilität: 0–stabil (siehe Pkt. 3.4.3).

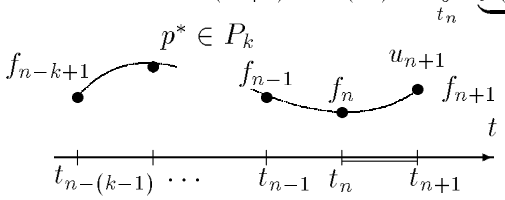
3.4.1.2 Die Adams–Moulton–Formeln

- **<u>Idee:</u>** Interpolationsquadratur mit
  - Stützstellen t_i
  - Stützwerten $f_i = f(t_i, u_i)$

  $\left.\right\}$ $i = n-k+1, \ldots, n-1, n, \mathbf{n+1}$

  $\downarrow$ aus Newton-Darstellung des Interpolationspolynoms mit Hilfe dividierter Differenzen $\Rightarrow$

$$u(t_{n+1}) = u(t_n) + \int\limits_{t_n}^{t_{n+1}} \underbrace{f(t, u(t))} \, dt$$



$$p^*(t_n + sh) = \sum_{j=0}^{k} (-1)^j \binom{-s+1}{j} \nabla^j f_{n+1}$$

$\downarrow$ mit $\nabla f_{n+1} = f_{n+1} - f_n$, $\nabla^{j+1} = \nabla^j \nabla$, $\nabla^0 = I$,

$$\binom{x}{j} = \frac{x(x-1)\cdot \ldots \cdot (x-j+1)}{1 \cdot 2 \cdot \ldots \cdot j} = \frac{x!}{(x-j)!j!}$$

$$u_{n+1} = u_n + h \int\limits_0^1 p^*(t_n + sh)\, ds$$

- **<u>Resultat:</u>** Adams–Moulton–Formeln ($k = 0, 1, \ldots$)

$$u_{n+1} = u_n + h \sum_{j=0}^{k} \gamma_j^* \nabla^j f_{n+1} \ \text{ mit } \ \gamma_j^* = (-1)^j \int\limits_0^1 \binom{-s+1}{j} ds$$

j	0	1	2	3	4	5	6	7	8
γ_j^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$	$-\frac{275}{24192}$	$\frac{33953}{3628800}$

- **<u>Beispiele:</u>**

  $k = 0: u_{n+1} = u_n + hf_{n+1} \quad \Rightarrow$ Implizite Euler–Methode,

  $k = 1: u_{n+1} = u_n + h\left[\frac{1}{2}f_{n+1} + \frac{1}{2}f_n\right] \quad \Rightarrow$ CRANK–NICOLSON,

  $k = 2: u_{n+1} = u_n + h\left[\frac{5}{12}f_{n+1} + \frac{8}{12}f_n - \frac{1}{12}f_{n-1}\right]$.

- **<u>Bemerkungen:</u>**

  1. Adams–Moulton–Formeln sind <u>implizite</u> MSV: Für hinr. kleines h ist Auflösbarkeit nach u_{n+1} durch Fixpunktiteration garantiert (f = Lipschitz-stetig) !
  2. Konsistenzordnung $= k + 1$ (siehe Pkt. 3.4.2).
  3. Stabilität: 0–stabil (siehe Pkt. 3.4.3).

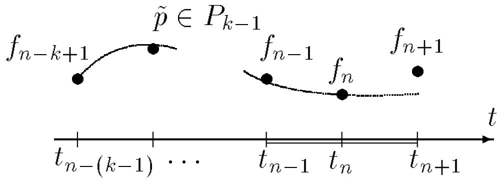
3.4.1.3 Die Nyström–Formeln

- **Idee:** Interpolations/Extrapolationsquadratur mit
  - Stützstellen t_i
  - Stützwerten $f_i = f(t_i, u_i)$

  $\left.\begin{array}{l} \\ \\ \end{array}\right\} i = n-k+1, \ldots, n-1, n$

  $\downarrow$ aus Newton-Darstellung des Interpolationspolynoms mit Hilfe dividierter Differenzen $\Rightarrow$

$$u(t_{n+1}) = u(t_{n-1}) + \int\limits_{t_{n-1}}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$$



$$\tilde{p}(t_n + sh) = \sum_{j=0}^{k-1} (-1)^j \binom{-s}{j} \nabla^j f_n$$

$\downarrow$ mit $\nabla f_n = f_n - f_{n-1}$, $\nabla^{j+1} = \nabla^j \nabla$, $\nabla^0 = I$,

$$\binom{x}{j} = \frac{x(x-1)\cdot \ldots \cdot (x-j+1)}{1 \cdot 2 \cdot \ldots \cdot j} = \frac{x!}{(x-j)!j!}$$

$$u_{n+1} = u_{n-1} + h \int\limits_{-1}^{1} \tilde{p}\,(t_n + sh)\, ds$$

- **Resultat:** Nyström–Formeln ($k = 1, 2, \ldots$)

$$u_{n+1} = u_{n-1} + h \sum_{j=0}^{k-1} \kappa_j \nabla^j f_n \quad \text{mit} \quad \kappa_j = (-1)^j \int\limits_{-1}^{+1} \binom{-s}{j}\, ds$$

j	0	1	2	3	4	5	6	7	8
κ_j	2	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{29}{90}$	$\frac{14}{45}$	$\frac{1139}{3780}$		

- **Beispiele:**

  $k = 1 : u_{n+1} = u_{n-1} + 2hf_n \quad (\widehat{=}\, k = 2) \quad \Rightarrow$ Explizite Mittelpunktsregel,

  $k = 3 : u_{n+1} = u_{n-1} + h \left[\frac{7}{3} f_n - \frac{2}{3} f_{n-1} + \frac{1}{3} f_{n-2} \right].$

- **Bemerkungen:**

  1. Nyström–Formeln sind <u>explizite</u> MSV !
  2. Konsistenzordnung $= k$ (siehe Pkt. 3.4.2).
  3. Stabilität: 0–stabil (siehe Pkt. 3.4.3).

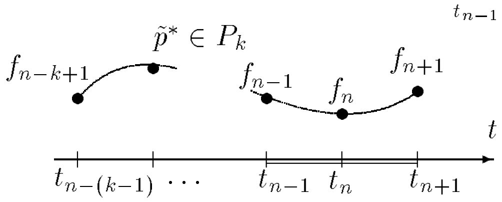
3.4.1.4 Die Milne–Simpson–Formeln

- **<u>Idee:</u>** Interpolationsquadratur mit
  - Stützstellen t_i
  - Stützwerten $f_i = f(t_i, u_i)$

  $\Big\}\ i = n-k+1, \dots, n-1, n, \mathbf{n+1}$

  $\downarrow$ aus Newton-Darstellung des Interpolationspolynoms mit Hilfe dividierter Differenzen $\Rightarrow$

$$u(t_{n+1}) = u(t_{n-1}) + \int\limits_{t_{n-1}}^{t_{n+1}} \underbrace{f(t, u(t))}\, dt$$



$$p^*(t_n + sh) = \sum_{j=0}^{k} (-1)^j \binom{-s+1}{j} \nabla^j f_{n+1}$$

$\downarrow$ mit $\nabla f_{n+1} = f_{n+1} - f_n,\ \nabla^{j+1} = \nabla^j \nabla,\ \nabla^0 = I$,

$$\binom{x}{j} = \frac{x(x-1)\cdot \ldots \cdot (x-j+1)}{1 \cdot 2 \cdot \ldots \cdot j} = \frac{x!}{(x-j)!j!}$$

$$u_{n+1} = u_{n-1} + h \int\limits_{-1}^{1} p^*\,(t_n + sh)\, ds$$

- **<u>Resultat:</u>** Milne–Simpson–Formeln $(k = 0, 1, 2, \dots)$

$$u_{n+1} = u_{n-1} + h \sum_{j=0}^{k} \kappa_j^* \nabla^j f_{n+1} \ \text{ mit } \ \kappa_j^* = (-1)^j \int\limits_{-1}^{+1} \binom{-s+1}{j}\, ds$$

j	0	1	2	3	4	5	6	7	8
κ_j^*	2	-2	$\frac{1}{3}$	0	$-\frac{1}{90}$	$-\frac{1}{90}$	$-\frac{37}{3780}$		

- **<u>Beispiele:</u>**

  $k = 0 : u_{n+1} = u_{n-1} + 2hf_{n+1} \quad \Rightarrow$ Impliziter Euler mit doppeltem Schritt,

  $k = 1 : u_{n+1} = u_{n-1} + 2hf_n \quad \Rightarrow$ Explizite Mittelpunktsregel $(2h)$,

  $k = 2 : u_{n+1} = u_{n-1} + h\left[\frac{1}{3}f_{n+1} + \frac{4}{3}f_n + \frac{1}{3}f_{n-1}\right] \quad \Rightarrow$ Simpson–Regel.

- **<u>Bemerkungen:</u>**

  1. Milne–Simpson–Formeln sind <u>implizite</u> MSV ! $\longrightarrow$ Für hinr. kleines h ist Auflösbarkeit nach u_{n+1} durch Fixpunktiteration garantiert: Startnäherung $u_{n+1}^{(0)} := u_n$ bzw. $u_{n+1}^{(0)} :=$ Nyström–Näherung.
  2. Konsistenzordnung $= k + 1$ (siehe Pkt. 3.4.2).
  3. Stabilität: 0–stabil (siehe Pkt. 3.4.3).

3.4.1.5 Die BDF–Verfahren (Backward Differencing Formula)

■ **Idee:** Numerische Differentiation des Interpolationspolynoms mit

$$\left.\begin{array}{ll} \bullet\ \text{Stützstellen} & t_i \\ \bullet\ \text{Stützwerten} & u_i \end{array}\right\} \quad i = n-k+1, \dots, n-1, n, \mathbf{n+1}$$

$\downarrow$ aus Newton-Darstellung des Interpolationspolynoms mit Hilfe dividierter Differenzen $\Rightarrow$

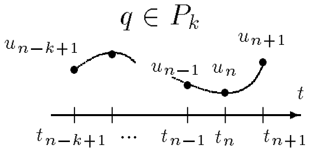
$$\underbrace{u'(t_{n+1})} = f(t_{n+1}, u(t_{n+1}))$$

$$q(t_n + sh) = \sum_{j=0}^{k} (-1)^j \tbinom{-s+1}{j} \nabla^j u_{n+1}$$

mit $\nabla u_{n+1} = u_{n+1} - u_n$, $\nabla^{j+1} = \nabla^j \nabla$, $\nabla^0 = I$,

$$\binom{x}{j} = \frac{x(x-1)\cdot \ldots \cdot (x-j+1)}{1 \cdot 2 \cdot \ldots \cdot j} = \frac{x!}{(x-j)!j!}$$

$\downarrow$ $q \in P_k$



$$q'(t_{n+1}) = f_{n+1} := f(t_{n+1}, u_{n+1})$$

■ **Resultat:** BDF–Verfahren ($k = \not{0}, 1, 2, \dots$)

$$\boxed{\begin{array}{c} \displaystyle\sum_{j=0}^{k} \delta_j^* \nabla_j u_{n+1} = h f_{n+1} \ \text{ mit } \ \delta_j^* = (-1)^j \frac{d}{ds} \binom{-s+1}{j} \Bigg|_{s=1} \\ \hline \delta_0^* = 0, \quad \delta_j^* = \dfrac{1}{j} \ \text{ für } j \geq 1 \end{array}}$$

■ **Beispiele:**

$k = 1 : u_{n+1} - u_n = h f_{n+1} \quad \Rightarrow \quad$ Impliziter Euler,

$k = 2 : \frac{3}{2} u_{n+1} - 2u_n + \frac{1}{2} u_{n-1} = h f_{n+1}$,

$k = 3 : \frac{11}{6} u_{n+1} - 3u_n + \frac{3}{2} u_{n-1} - \frac{1}{3} u_{n-2} = h f_{n+1}$.

■ **Bemerkungen:**

1. BDF–Methoden sind <u>implizite</u> MSV !
2. Konsistenzordnung $= k$ (siehe Pkt. 3.4.2).
3. Stabilität: 0–stabil für $k \leq 6$, <u>nicht</u> 0–stabil für $k \geq 7$ (siehe Pkt. 3.4.3).

3.4.2 Konsistenz linearer MSV

- **Lineare MSV:**

Alle bisher betrachteten MSV sind von der Form

$$
(53)\quad \begin{cases} \qquad\overset{\displaystyle n+1}{\uparrow} \qquad\qquad\qquad\qquad \overset{\displaystyle n-k+1}{\uparrow} \\ \alpha_k \mathbf{u_{n+k}} + \alpha_{k-1} u_{n+k-1} + \ldots + \alpha_0 u_n = h \sum\limits_{j=0}^{k} \beta_j f_{n+j}, \; n = 0, 1, \ldots, m-k, \\ \text{mit idealisierter Startphase (o. B. d. Allg.):} \qquad \downarrow \\ u_j = u(t_j), \; j = 0, 1, \ldots, k-1. \qquad\qquad f(t_{n+j}, u_{n+j}) \end{cases}
$$

MSV der Form (53) heißen lineare MSV (k–Schrittverfahren).

Durch eine Normierungsbedingung

$$(54)\qquad \alpha_k = 1 \quad \text{oder} \quad \sum_{j=0}^{k} \beta_j = 1$$

wird das MSV eindeutig festgelegt. Betrachten im folgenden nur lineare MSV !

- **Lokaler (Diskretisierungs–)Fehler, Konsistenz, Konsistenzordnung:**

$$(55)\quad u(s): \quad \begin{array}{l} u'(s) = f(s, u(s)), \quad s \geq t \\ u(t) \text{ geg.} \end{array}$$

$$(56)\quad u_h(t+kh): \quad \begin{array}{l} \sum\limits_{j=0}^{k} \alpha_j u_h(t+jh) = h \sum\limits_{j=0}^{k} \beta_j f(t+jh, u_h(t+jh)) \\ u_h(t+jh) = u(t+jh), \quad j = 0, 1, \ldots, k-1 \end{array}$$

- **Definition 3.27:** (Konsistenz von MSV)

> Ein MSV heißt konsistent, falls
> $$u(t+kh) - u_h(t+kh) \longrightarrow 0 \text{ für } h \to 0 \quad \forall t.$$
> Falls
> $$u(t+kh) - u_h(t+kh) = O(h^{p+1}) \text{ für } h \to 0 \quad \forall t,$$
> dann ist die Konsistenzordnung p.

Dabei wird $u(\cdot)$ immer als hinreichend glatt vorausgesetzt.

- **Lemma 3.28:** (Darstellung des lokalen Fehlers)

> Vor.: f sei stetig differenzierbar.
>
> Bh.: Dann $\exists\, \delta \in (0,1)$:
>
> (57) $\quad u(t+kh) - u_h(t+kh) = [\alpha_k I - h\beta_k f_u(t+kh,\nu)]^{-1}\, L(u;t,h),$
>
> $\qquad$ mit $\nu = \nu(\delta) = u_h(t+kh) + \delta[u(t+kh) - u_h(t+kh)],$
>
> (58) $\quad L(u;t,h) := \sum\limits_{j=0}^{k} [\alpha_j u(t+jh) - h\beta_j u'(t+jh)].$
>
> Im Fall von Systemen ist (57) komponentenweise zu verstehen mit $\delta_i \in (0,1)$ und $\nu_i = \nu(\delta_i)$, $\;i = \overline{1,N}$.

Beweis:

Aus (56) folgt sofort die Beziehung

$$(59) \qquad \sum_{j=0}^{k-1} [\alpha_j u(t+jh) - h\beta_j \underbrace{f(t+jh, u(t+jh))}_{u'(t+jh)}] + \alpha_k\, u_h(t+kh) - \\ -h\beta_k\, f(t+kh, u_h(t+kh)) = 0.$$

Aus (58) und (59) ergibt sich dann

$$L(u,t,h) = \alpha_k\, u(t+kh) - h\beta_k\, f(t+kh, u(t+kh)) + \\ + \underbrace{\sum_{j=0}^{k-1} [\alpha_j\, u(t+jh) - h\,\beta_j\, u'(t+jh)]}_{\overset{(59)}{=} -[\alpha_k u_h(t+kh) - h\beta_k f(t+kh, u_h(t+kh))]} =$$

$$= \alpha_k[u(t+kh) - u_h(t+kh)] - h\beta_k[f(t+kh, u(t+kh)) - f(t+kh, u_h(t+kh))]$$

$$\overset{[16]\ (\text{Satz I.5.5})}{=} [\alpha_k I - h\beta_k f_u(t+kh,\nu)]\,(u(t+kh) - u_h(t+kh)).$$

q.e.d.

- Bemerkung: Für explizite MSV ($\beta_k = 0$) und Normierungsbed. ($\alpha_k = 1$):
  $\Rightarrow\; u(t+kh) - u_h(t+kh) = L(u,t,h)$, sonst $= O(L(u;t,h))$.

■ Näherungsgleichung und lokaler Abschneidefehler:

- Schreiben MSV (53) als Operatorgleichung in der Form

$$(60) \qquad \boxed{\text{Ges. } u_h \in X_h : F_h(u_h) = 0 \text{ in } Y_h}$$

mit $X_h = C^1(I_h)$, $\;Y_h = C(I_h)$ und

$$
(61)\quad \begin{cases} F_h(v_h)(t_i+kh) := \frac{1}{h}\sum\limits_{j=0}^{k}\alpha_j v_h(t_i+jh) - \sum\limits_{j=0}^{k}\beta_j f(t_i+jh, v_h(t_i+jh)), \\ \hfill i=\overline{0,m-k} \\ \text{und idealisierter Startphase} \\ F_h(v_h)(t_0+jh) := v_h(t_0+jh) - u(t_0+jh), \quad j=0,1,\dots,k-1. \end{cases}
$$

Dabei ist es jetzt zweckmäßig, (o. B. d. Allg.) von der Normierungsbed.

(62) $\sum\limits_{j=0}^{k} b_j = 1$ auszugehen.

- Für den <u>lokalen Abschneidefehler</u>

(63) $\tau_h := F_h(u)$, wobei u Lsg. von (1) ist,

gilt offenbar die Beziehung

(64) $\tau_h(t+kh) = \frac{1}{h}L(u;t,h)$.

Der damit verbundene Begriff der <u>Konsistenzordnung</u> p

(65) $\|\tau_h\|_{Y_h=C(I_h)} = O(h^p)$

entspricht wegen <u>Lemma 3.28</u> genau dem in <u>Def. 3.27</u> eingeführten Begriff der Konsistenzordnung über die Kleinheit des lokalen Fehlers.

■ <u>Zur Bestimmung der Konsistenzordnung linearer MSV:</u>

- Ordnen den Koeffizienten eines linearen MSV Polynome zu:

(66) $\rho(z) = \alpha_k z^k + \alpha_{k-1}z^{k-1} + \ldots + \alpha_0$; (charak. Polynom)
$\sigma(z) = \beta_k z^k + \beta_{k-1}z^{k-1} + \ldots + \beta_0$.

- **<u>Satz 3.29:</u>**

<u>Vor.:</u> f sei hinreichend oft stetig differenzierbar.

<u>Bh.:</u> Ein MSV der Form (53) besitzt genau dann die Konsistenzordnung p, falls

$$
(67)\quad \begin{cases} \sum\limits_{j=0}^{k}\alpha_j = 0, \\ \sum\limits_{j=0}^{k}\alpha_j j^q = q\sum\limits_{j=0}^{k}\beta_j j^{q-1}, \quad q=1,2,\dots,p. \end{cases}
$$

Beweis:

$$L(u,t,h) \;=\; \sum_{j=0}^{k} [\alpha_j \underbrace{u\,(t\,+\,jh)}_{\sum\limits_{q=0}^{p}\frac{j^q}{q!}u^{(q)}(t)h^q+O(h^{p+1})} \;-\; h\,\beta_j \;\underbrace{u'(t+jh)}_{\sum\limits_{r=0}^{p-1}\frac{j^r}{r!}u^{(r+1)}(t)h^r+O(h^p)}\;] \;\overset{\substack{\text{Taylor}\\ \downarrow}}{=}$$

$$= u(t)\sum_{j=0}^{k}\alpha_j + \sum_{q=1}^{p}\frac{u^{(q)}(t)}{q!}h^q\sum_{j=0}^{k}\alpha_j j^q - \sum_{q\equiv r+1=1}^{p}\frac{u^{(q)}(t)q}{(q-1)!q}h^q\sum_{j=0}^{k}\beta_j j^{q-1} + O(h^{p+1}) =$$

$$= u(t)\underbrace{\sum_{j=0}^{k}\alpha_j}_{=0} + \sum_{q=1}^{p}\frac{h^q}{q!}u^{(q)}(t)\underbrace{\left[\sum_{j=0}^{k}\alpha_j j^q - q\sum_{j=0}^{k}\beta_j j^{q-1}\right]}_{=0 \text{ für } q=1,2,\ldots,p} + O(h^{p+1}) =$$

$$= O(h^{p+1})$$ **q.e.d.**

- Bemerkung:

  Für den Fall $p = 1$ (Konsistenzordnung 1) lautet (67):

$$(68)\qquad \begin{cases} \rho(1) = 0 & \left[\sum\limits_{j=0}^{k}\alpha_j = 0\right], \\[2ex] \rho'(1) = \sigma(1) & \left[\sum\limits_{j=0}^{k} j\alpha_j = \sum\limits_{j=0}^{k}\beta_j\right]. \end{cases}$$

- **Folgerung 3.30:**

  Die höchste erreichbare Konsistenzordnung eines k–Schrittverfahrens $= 2k$.

  Beweis: mms.

- **Folgerung 3.31:** (Konsistenzordnung der MSV aus Pkt. 3.4.1)

MSV (k–Schritt)	Konsistenzordnung	MSV ist exakt für $u'(t) = qt^{q-1}$
Adams–Bashforth	k	$q = 0, 1, \ldots, k$
Adams–Moulton	$k + 1$	$q = 0, 1, \ldots, k + 1$
Nyström	k	$q = 0, 1, \ldots, k$
Milne–Simpson	$k + 1$	$q = 0, 1, \ldots, k + 1$
BDF	k	$q = 0, 1, \ldots, k$

Beweis: Unter Benutzung von (57) aus Lemma 3.28 und der Exaktheit des MSV für $u'(t) = qt^{q-1}$, d.h. für $u(t) = t^q + c$ erhalten wir sofort die Identität

$$
\begin{aligned}
0 &\overset{\substack{\text{Lemma 3.28, (57)}\\ \downarrow}}{=} [\alpha_k I - 0](u(t+kh) - u_h(t+kh)) = L(t^q + c, t, h) = \\
&= \sum_{j=0}^{k} [\alpha_j[(t+jh)^q + c] - h\beta_j q(t+jh)^{q-1}] = \\
&= c \underbrace{\sum_{j=0}^{k} \alpha_j}_{=0} + \underbrace{\sum_{j=0}^{k} [\alpha_j (t+jh)^q - h\beta_j q(t+jh)^{q-1}]} \\
&\qquad\qquad \underset{\substack{\uparrow \\ t=0}}{=} h^q \underbrace{\sum_{j=0}^{k} [\alpha_j j^q - q\beta_j j^{q-1}]}_{=0}
\end{aligned}
$$

q.e.d.

3.4.3 Stabilität linearer MSV

■ **Im Gegensatz zu den RK–Formeln (Satz 3.25: f – Lipschitz–stetig $\Rightarrow$ Stabilität) ist die Untersuchung der Stabilität für MSV komplizierter !**

Ü 3.10 Man zeige zunächst, daß das explizite Zweischrittverfahren ($k = 2$)

(∗) $$u_{n+2} + 4u_{n+1} - 5u_n = h(4f_{n+1} + 2f_n)$$

die höchstmögliche Konsistenzordnung $p = 3$ hat !

Tatsächlich: $\rho(z) = 1 \cdot z^2 + 4z - 5$
$\sigma(z) = 4z + 2$

			p
• $\sum_{j=0}^{2} \alpha_j$	$=$	$1 + 4 - 5 = 0$	$0\surd$
• $\sum_{j=0}^{2} \alpha_j j$	$\equiv$	$0 \cdot (-5) + 1 \cdot 4 + 2 \cdot 1 \equiv 6 \overset{!}{=} 1 \sum \beta_j j^0 \equiv 6$	$1\surd$
• $\sum_{j=0}^{2} \alpha_j j^2$	$=$	$4 \cdot 1^2 + 1 \cdot 2^2 \equiv 8 \overset{!}{=} 2 \sum_{j=0}^{2} \beta_j j^1 \equiv 2(4 \cdot 1) = 8$	$2\surd$
• $\sum \alpha_j j^3$	$=$	$4 \cdot 1^3 + 1 \cdot 2^3 \equiv 12 \overset{!}{=} 3 \sum \beta_j j^2 = 3(4) = 12$	$3\surd$
• $\sum \alpha_j j^4$	$=$	$4 \cdot 1^4 + 1 \cdot 2^4 = 20 \overset{!}{\neq} 4 \sum \beta_j j^3 = 4(4) = 16$	4 ↯

Wendet man dieses Verfahren auf das AWP

$$u'(t) = f(t, u(t)) := u(t), \quad t \geq 0; \text{ Lsg. sei } \underline{u(t) = e^t}$$

mit der exakten Startphase $\underline{u(0) = 1 \text{ und } u(h) = e^h}$ an, so erhält man völlig unbrauchbare Resultate, auch und insbesondere für $h \to 0$, obwohl $f(t,u) := u$ Lipschitz-stetig mit der Lipschitz–Konstanten $L = 1$ ist !

Bestätigen Sie diese Behauptung durch numerische Experimente, und begründen Sie die Behauptung <u>theoretisch</u> !

<u>Hinweis:</u> • <u>theoretisch:</u> $(*)$ mit Ansatz: $u_j = \zeta^j$
$\Rightarrow$ Charakteristisches Polynom: $\zeta^2 + 4(1-h)\zeta - (5+2h) = 0$
$\Rightarrow$ Wurzeln: $\zeta_1(h) = 1 + h + O(h^2), \quad \zeta_2(h) = -5 + O(h)$
$$u_n = a_1 \zeta_1^n(h) + a_2 \zeta_2^n(h)$$

■ <u>0–Stabilität von MSV:</u>

- Untersuchen Stabilität von linearen MSV zunächst für die triviale RS $\boxed{f \equiv 0}$ („$h \to 0$“):

(69) $$\alpha_k u_{n+k} + \alpha_{k-1} u_{n+k-1} + \ldots + \alpha_0 u_n = 0$$

als Näherungsgleichung für die triviale Dgl.

(70) $$u'(t) = 0.$$

- Der nächste Satz liefert die allgem. Lsg. der Differenzengleichung (69):

<u>Satz 3.32:</u>

<u>Vor.:</u> Seien $\zeta_1, \zeta_2, \ldots, \zeta_l$ die Wurzeln von $\rho(z) = \sum\limits_{j=0}^{k} \alpha_j z^j$ mit den Vielfachheiten $m_1, m_2, \ldots, m_l$ $\left(\sum\limits_{j=0}^{l} m_j = k,\ l \le k\right)$.

<u>Bh.:</u> Dann ist die allgemeine Lösung von (69) durch

(71) $$u_n = p_1(n)\zeta_1^n + p_2(n)\zeta_2^n + \ldots + p_l(n)\zeta_l^n$$

gegeben, wobei $p_j(\cdot)$ bel. Polynome vom Grad $m_j - 1$ sind.

<u>Beweis:</u> siehe auch [5] S. 328:

◦ $l = k$, d.h. $\rho(z)$ hat nur einfache Nullstellen $(m_j = 1\ \forall j)$
$\Rightarrow u_n = a_1 \zeta_1^n + \ldots + a_k \zeta_k^n$ ist offenbar allgem. Lsg. von (69) !

◦ $l < k$, d.h. $\rho(z)$ hat mehrfache Nullstellen $(\exists\, m_j > 1)$:

1) Sei ζ eine Nullstelle der Vielfachheit m. Dann sind $u_n = \binom{n}{i}\zeta^n$ für jedes $i = 0, 1, \ldots, m-1$ spezielle Lsg. von (71).

Tatsächlich, wegen der Identität $\binom{n+j}{i} = \sum\limits_{l=0}^{i} \binom{n}{i-l}\binom{j}{l}$ (mms) gilt:

$$\sum_{j=0}^{k} \alpha_j u_{n+j} = \sum_{j=0}^{k} \alpha_j \binom{n+j}{i} \zeta^{n+j} = \zeta^n \sum_{l=0}^{i} \binom{n}{i-l} \sum_{j=0}^{k} \alpha_j \binom{j}{l} \zeta^j =$$
$$= \zeta^n \sum_{l=0}^{i} \binom{n}{i-l} l! \zeta^l p^{(l)}(\zeta) = 0$$

2) Superposition: $u_n = \sum\limits_{j=1}^{l} \sum\limits_{i=0}^{m_j-1} a_{ij} \binom{n}{i} \zeta_j^n = \sum\limits_{j=1}^{l} p_j(n) \zeta_j^n$

mit frei wählbaren Koeffizienten $\{a_{ij}\}_{j=\overline{1,l};\ i=\overline{0,m_j-1}}$.

q.e.d.

- Durch die Vorgabe der k Startwerte

$$u_0, u_1, \ldots, u_{k-1}$$

wird die Lsg. (71) von (69) eindeutig definiert.

- Damit Störungen δ_i in der Startphase, d.h. $u_i + \delta_i$, $i = \overline{0, k-1}$, nicht zu unbe–schränkten Lösungen führen, muß für jede Nullstelle ζ von $\rho(\cdot)$ offenbar gelten:

(72) $$\begin{cases} \bullet\ |\zeta| \leq 1\ , \text{ falls } \zeta - \text{ einfache Nullstelle,} \\ \bullet\ |\zeta| < 1\ , \text{ falls } \zeta - \text{ mehrfache Nullstelle.} \end{cases}$$

Das führt auf die folgende Definition:

Definition 3.33: (0–Stabilität)

> Das MSV (53) heißt 0–stabil, falls jede Nullstelle ζ des charakteristischen Polynoms $\rho(\cdot)$ die Bedingung (72) erfüllt.

- Beispiele:
  1. Adams–Bashforth und Adams–Moulton: 0–stabil, da
     $\rho(z) = z^k - z^{k-1} \Rightarrow \zeta_1 = 0\ (k-1)$–fach; $\zeta_2 = 1$ (einfach).
  2. Nyström und Milne–Simpson–Formeln: 0–stabil, da
     $\rho(z) = z^k - z^{k-2} \Rightarrow \zeta_1 = 0\ (k-2)$–fach; $\zeta_{2,3} = \pm 1$ (einfach).
  3. BDF–Formeln: $k \leq 6 \Rightarrow$ 0–stabil; $k \geq 7 \Rightarrow$ nicht 0–stabil !
     Analyse komplizierter: $\sum\limits_{j=1}^{k} \frac{1}{j} \nabla_j u_{n+1} \mapsto \rho(z) = \ldots$? siehe [5] S. 329 ff.

- Die erste Dahlquist–Barriere (1956):

Für die höchste, erreichbare Ordnung p eines 0–stabilen k–Schrittverfahrens gilt:

(73) $$p \leq \begin{cases} k+2\ , & \text{falls } k - \text{ungerade,} \\ k+1\ , & \text{falls } k - \text{gerade,} \\ k & \text{, falls } \beta_k/\alpha_k \leq 0 \text{ (also insbes. für explizite MSV).} \end{cases}$$

Beweis: siehe [5] S. 332 ff.

3.4.4 Konvergenz linearer MSV

- **Idee:** [*Butcher*, 1966]

$$
\begin{array}{lcl}
\text{MSV} & \xrightarrow{\text{Übergang zu Produkträumen}} & \text{ESV} \\
\downarrow & & \downarrow \\
\bullet\ \text{Konsistenz (Satz 3.29)} & \xrightarrow{\text{Satz 3.34}} & \bullet\ \text{Konsistenz (Satz 3.17)} \\
 & & + \\
\left.\begin{array}{l}\bullet\ \text{0–Stabilität (Def. 3.33)} \\ \\ \bullet\ f = L\text{-stetig}\end{array}\right\} & \xrightarrow{\text{Satz 3.36}} & \begin{array}{l}\bullet\ \text{Stabilität (Satz 3.20)} \\ \quad \| \\ \bullet\ \text{Konvergenz (Satz 3.23)}\end{array} \\
\bullet\ \boxed{\text{Konvergenz}} & \xleftarrow{\text{Satz 3.37; Folgerung 3.38}} &
\end{array}
$$

- **Schreiben MSV in $I\!R^N$ als ESV in $(I\!R^N)^k$:**

  - Sei $f(\cdot,\cdot)$ stetig und im 2. Arg. Lipschitz–stetig, d.h.

    (74) $\qquad |f(t,v) - f(t,w)| \leq L|v-w|, \quad \forall\, t \in I, \quad \forall v, w \in I\!R^N.$

    Dann gibt es für bel. (fix. !) Werte

$$t_n, u_n, u_{n+1}, \ldots, u_{n+k-1}$$

und hinreichend kleinen h eine eindeutig bestimmte Lösung

$$u_{n+k} = g(t_n; u_n, u_{n+1}, \ldots, u_{n+k-1}; h)$$

des MSV (53)

$$\sum_{j=0}^{k-1} \alpha'_j u_{n+j} + \mathbf{u}_{n+k} = \underbrace{h \sum_{j=0}^{k-1} \beta'_j f_{n+j} + h\beta'_k f(t_n + kh, \mathbf{u}_{n+k})}_{h\,\psi};$$

$$\alpha'_j = \tfrac{\alpha_j}{\alpha_k}, \quad \beta'_j = \tfrac{\beta_j}{\beta_k}.$$

Beweis: Banachscher Fixpunktsatz angewandt auf
$\mathbf{u}_{n+k} = h\beta'_k f(t_n + kh, \mathbf{u}_{n+k})$ + Rest ! (mms)

Damit läßt sich die Funktion ψ definieren:

$$\psi(t_n; u_n, u_{n+1}, \ldots, u_{n+k-1}; h) := \sum_{j=0}^{k-1} \beta'_j f_{n+j} +$$
$$+ \beta'_k f(t_n + kh, \underbrace{g(t_n; u_n, u_{n+1}, \ldots, u_{n+k-1}; h)}_{=u_{n+k}})$$

Damit erhält das MSV (53) die Form:

(75) $$u_{n+k} = -\sum_{j=0}^{k-1} \alpha'_j u_{n+j} + h\psi(t_n; \underbrace{u_n, u_{n+1}, \ldots, u_{n+k-1}}_{=:U_n}; h).$$

Definieren

$$U_n = \begin{bmatrix} u_{n+k-1} \\ u_{n+k-2} \\ \vdots \\ u_n \end{bmatrix}, \quad A = \begin{bmatrix} -\alpha'_{k-1} I & -\alpha'_{k-2} I & \ldots & -\alpha'_0 I \\ I & \boldsymbol{O} & \ldots & \boldsymbol{O} \\ & \ddots & \ddots & \\ & & I & \boldsymbol{O} \end{bmatrix}, \quad e_1 = \begin{bmatrix} I \\ \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \end{bmatrix}$$

und können damit das MSV (75) $\equiv$ (53) als ESV

(76)
$$\boxed{\begin{array}{l} U_{n+1} = A\,U_n + h\,\Phi(t_n, U_n, h), \quad n = \overline{0, m-k} \\ \text{AB: } U_0 \text{ geg. (ESV aus Pkt. 3.3: } A = I\text{ !)} \end{array}}$$

im Produktraum $(\mathbb{R}^N)^k = \mathbb{R}^N \times \ldots \times \mathbb{R}^N$ schreiben, wobei

$$\Phi(t, U, h) = \psi(t, U, h)\, e_1,$$

$$\mathbf{X}_h = C^1(\tilde{I}_h, (\mathbb{R}^N)^k), \quad \mathbf{Y}_h = C(\tilde{I}_h, (\mathbb{R}^N)^k), \quad \tilde{I}_h = \{t_0, \ldots, t_{m-k+1}\}.$$

■ **Konsistenz:** $\boxed{\text{MSV (53)} = (75) \quad \widehat{=} \quad \text{ESV (76)}}$

Satz 3.34:

> Aus der Konsistenz(–ordnung) des MSV (53) = (75) gemäß Def. 3.27 folgt die Konsistenz(–ordnung) des zum MSV äquivalenten ESV (76) und umgekehrt.

Beweis:

- Sei $u(s)$ die exakte Lösung des AWP (1) und

  (77) $\quad U(s) = [u(s + (k-1)h), u(s + (k-2)h), \ldots, u(s)]^T$

- Die Näherungslösung an der Stelle $t + h$, die man aus dem ESV (76) mit dem Startwert $U(t)$ gewinnt, sei $U_h(t + h)$!

- Dann gilt offenbar:

$$U(t+h) - U_h(t+h) = \begin{bmatrix} u(t+kh) \\ u(t+(k-1)h) \\ \vdots \\ u(t+h) \end{bmatrix} - \begin{bmatrix} u_h(t+kh) \\ u(t+(k-1)h) \\ \vdots \\ u(t+h) \end{bmatrix} =$$

$$= \begin{bmatrix} u(t+kh) - u_h(t+kh) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Daraus folgt:

(78) $\quad \|U(t+h) - U_h(t+h)\|_{\mathbf{X}_h} = \|u(t+kh) - u_h(t+kh)\|_{X_h} = O(h^{p+1}).$

q.e.d.

■ **Stabilität:** $\boxed{\text{0–Stabilität des MSV} + L\text{–Stetigkeit von } f \Rightarrow \text{Stabilität v. (76)}}$

- **Lemma 3.35:**

> Vor.: Das MSV sei 0–stabil (Def. 3.33).
>
> Bh.: Dann gibt es eine Vektornorm in $(\mathbb{R}^N)^k$, sodaß für die zugeordnete Matrixnorm gilt:
> $$\|A\| \leq 1.$$

Beweis: siehe [5] S. 343 f.

- Bemerkung: Lineares MSV ist 0–stabil, genau dann, wenn $\exists\, c = \text{const.} \geq 0$: $\|A^n\| \leq c \quad \forall n \in \mathbb{N}$.

- **Satz 3.36:** (Stabilität des ESV (76) i. S. der Def. 3.21)

> Vor.: 1) MSV (53) = (75) sei 0–stabil.
>
> 2) $f(\cdot,\cdot)$ erfülle die Lipschitz–Bedingung
> $$|f(t,v) - f(t,w)| \leq L|v-w| \quad \forall v,w \in \mathbb{R}^N \quad \forall t \in I.$$
>
> Bh.: Dann ist das zum MSV zugeordnete ESV (76) stabil i. S. der Def. 3.21.

Beweisskizze:

- Aus der Lipschitz–Stetigkeit von $f(\cdot,\cdot)$ folgt (mms):

$$\|\Phi(t,V,h) - \Phi(t,W,h)\| \leq L^*\|V-W\| \quad \forall V,W \in (\mathbb{R}^N)^k \ \forall t, \ \forall h \leq H.$$

- Unter Benutzung von Lemma 3.35 ($\Rightarrow \|A\| \le 1$) zeigen wir zunächst Stabilität bzgl. der AB (Startphase) analog zu Lemma 3.18:

$$\begin{array}{rcl} Z_j & = & \overset{(77)}{U_j} - \overset{(76)}{V_j} \\ Z_{n+1} & = & AZ_n + h[\Phi(t_n, U_n, h) - \Phi(t_n, V_n, h)] \\ \|Z_{n+1}\| & \le & \|A\|\,\|Z_n\| + L^* h\,\|Z_n\| \le (1 + L^* h)\|Z_n\| \\ & & \qquad\qquad\qquad\qquad \le e^{Lt_{n+1}} \end{array}$$

usw.

- Daraus folgt analog zu Lemma 3.19 und Satz 3.20 die Stabilität:

$$\|V_h - W_h\|_{\mathbf{X}_h} \le C\,\|F_h(V_h) - F_h(W_h)\|_{\mathbf{Y}_h}.$$

q.e.d.

■ **Konvergenz:** $\boxed{\text{Stabilität} + \text{Konsistenz} = \text{Konvergenz}}$

- **Satz 3.37:** (Konvergenz des ESV (76))

> Vor.: 1) MSV habe Konsistenzordnung p.
> 2) f sei entsprechend oft differenzierbar.
> 3) f sei Lipschitz–stetig im 2. Arg.
> 4) MSV sei 0–stabil.
>
> Bh.: Dann ist das ESV (76) konvergent von der Ordnung p.

Beweis: analog zum Beweis von Satz 3.23 unter Benutzung der Sätze 3.34 und 3.36.

q.e.d.

- **Folgerung 3.38:** (Konvergenz des MSV)

> Unter den Voraussetzungen von Satz 3.37 ist das MSV (53) konvergent von der Ordnung p.

Beweis: MSV (53) $\equiv$ (75) $=$ ESV (76).

q.e.d.

3.5 Steife Differentialgleichung

■ **Btr. Beispiel** $\boxed{\textbf{E 6}}$ **aus** $\boxed{\textbf{P IX}}$ **:**

$$(80) \qquad \begin{cases} u'(t) = -50(u(t) - \cos(t)), \quad t \in [0,T] \\ \underline{\text{AB:}}\ u(0) = 0 \end{cases}$$

Das explizite und das implizite Euler–Verfahren verhalten sich völlig unterschiedlich (siehe P IX):

- expliziter Euler: sinnvolle Resultate erst für $h = \frac{1}{40}, \frac{1}{30} \leq \frac{2}{50}$!
- impliziter Euler: keine Probleme auch für größere h !,

obwohl beide Verfahren

- Konsistenzordnung 1 haben und
- stabil im Sinne der Def. 3.21 sind.

Offenbar reicht der bisherige Stabilitätsbegriff nicht aus, um das Verhalten der Verfahren für größere („endliche") Schrittweiten h zur Lösung steifer AWP (1), d.h. große Lipschitz–Konstante L bzw. große Kondition der Jacobi–Matrix $J = f_u(t, u(t))$, zu erklären:

- Lemma 3.18: $(1 + Lh_j) \overset{?}{\leq} e^{Lh_j} \Rightarrow e^{Lh_j} \cdot \ldots \cdot e^{Lh_0} = e^{Lt_{j+1}} \leq e^{LT}$,
- Beweis Satz 3.25: h hinreichend (?) klein.

■ **Zur Untersuchung der Stabilitätsverhalten eines Verfahrens mit „endlicher" Schrittweite wird das Verhalten des Verfahrens für das Testproblem (81) mit $\boldsymbol{\lambda} \in \mathbb{C}$ betrachtet:**

$$(81) \quad \boxed{\begin{aligned} u'(t) &= \lambda\, u(t) \\ u(0) &= 1 \end{aligned}} \quad \rightsquigarrow \quad \boxed{\begin{aligned} \tilde{u}'(t) &= \lambda\, \tilde{u}(t) \\ \tilde{u}(0) &= 1 + \delta \end{aligned}} \quad (82)$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

$$u(t) = e^{\lambda t} \qquad\qquad \tilde{u}(t) = (1+\delta)e^{\lambda t}$$

$$\searrow \qquad \swarrow$$

Fehler

$$z(t) = \tilde{u}(t) - u(t) = \delta\, e^{\lambda t} \overset{\lambda=\lambda_1+i\lambda_2,\ \lambda_1=Re\,\lambda,\ \lambda_2=Im\,\lambda}{=} \delta\, e^{Re\,\lambda\cdot t}\, e^{i\cdot Im\,\lambda\cdot t}$$

$$\Rightarrow \qquad \boxed{|z(t)| \;=\; \delta\, e^{Re\,\lambda\cdot t} \;\leq\; \delta, \quad \text{falls } Re\,\lambda \;\leq\; 0}$$

o.K., aber bekommt man so Aussagen über den allgemeinen Fall

$$u'(t) = f(t, u(t))\ ?$$

■ Motivation für das Testproblem (81):

(1) • $\left.\begin{array}{l} u'(t) = f(t, u(t)) \\ u(0) = u_0 \end{array}\right\} \Rightarrow u(t)$ exakte Lösung

(83) • $\left.\begin{array}{l} \tilde{u}'(t) = f(t, \tilde{u}(t)) \\ \tilde{u}(0) = u_0 + \delta \end{array}\right\} \Rightarrow \tilde{u}(t)$ gestörte Lösung

- Linearisieren gestörtes Problem (83) in der Näherung der exakten Lösung $u(t)$:

  $\Rightarrow \tilde{u}'(t) = \underbrace{f(t, u(t))}_{=u'(t)} + f_u(t, u(t))\,(\tilde{u}(t) - u(t)) + o(\tilde{u}(t) - u(t))$

  $\Rightarrow$ Erhalten also für den Fehler $z(t) = \tilde{u}(t) - u(t)$:

  $$z'(t) = \underbrace{f_u(t, u(t))}_{\substack{\text{Einfrieren der Jacobi–Matrix} \\ J = f_u(t_*, u(t_*)) = [\ldots]_{N\times N},\ \text{z.B. für } t_* = 0}} z(t) + o\Big((z(t))\Big), \quad z(0) = \delta$$

  Dann ist die Dgl.

  (84) $\quad z'(t) = J\,z(t) \qquad$ (vgl. Kap. 2: $J = M_h^{-0.5}\,K_h\,M_h^{-0.5}$)

  ein Modell für die Fehlerausbreitung.

- Sei J diagonalisierbar, d.h.

  ∘ $\exists$ vollst. System v. EV : $v_1, \ldots, v_N$
  ∘ EW : $\lambda_1, \ldots, \lambda_N$.

  Ansatz:

  $$z(t) = \sum_{i=1}^{N} \eta_i(t) v_i$$

  Eingesetzt in (84) ergibt das:

  $$z'(t) \equiv \sum_{i=1}^{N} \eta_i'(t) v_i = Jz(t) \equiv \sum_{i=1}^{N} \eta_i(t)\lambda_i v_i,$$

  d.h. $\boxed{\eta_i'(t) = \lambda_i \eta_i(t), \quad \forall\ i = \overline{1, N}}$ $\ \widehat{=}\ (81)$

- Schlußfolgerung:

  Verhalten des Integrationsverfahrens (ESV) für (81) mit $\lambda \in \sigma\left(f_u(t, u(t))\right) =$ $\{$ EW der Jacobi–Matrix $\}$ gibt Aufschluß über Stabilität des Verfahrens !

■ **Wenden nun expliziten und impliziten Euler auf unser Testproblem (81) an:**

- expliziter Euler:

$$u_{n+1} = u_n + h\lambda u_n \equiv R(h\lambda)u_n, \text{ mit}$$
$$R(z) = 1 + z$$

- impliziter Euler:

$$u_{n+1} = u_n + h\lambda u_{n+1}, \text{ also}$$
$$u_{n+1} = \frac{1}{1-h\lambda}u_n \equiv R(h\lambda)u_n, \text{ mit}$$
$$R(z) = \frac{1}{1-z}$$

■ **Wendet man nun ein ESV auf ein „stabiles" AWP (81) an,**

$\longrightarrow$ d.h. mit $Re\,\lambda \leq 0$ (Störungen in den AW vergrößern sich nicht !),
dann sollte das ESV

$$u_{n+1} = R(h\lambda)u_n$$

die gleiche Eigenschaft haben, d.h.

$$|R(h\lambda)| \leq 1,$$

falls $\lambda \in \mathbb{C}: Re\,\lambda \leq 0$ und $\lambda \in \sigma\,(f_u(t, u(t)))$.

Diese Stabilitätsinterpretation führt zum Begriff der A–Stabilität (Pkt. 3.5.1) !

3.5.1 A–Stabilität ($\longrightarrow$ ESV)

■ **Stabilitätsbereich:**

- Ein ESV zur Lösung des Testproblems

  (81) $\qquad u'(t) = \lambda\, u(t)$

  habe die Form

  (85) $\qquad u_{n+1} = R(h\lambda)u_n$

  mit der Stabilitätsfunktion $R(z)$.

- **Definition 3.40:** (Stabilitätsbereich)

> Die Menge
>
> $$S = \{z \in \mathbb{C}: |R(z)| \leq 1\}$$
>
> heißt Stabilitätsbereich des ESV.

- Bemerkung:
  1. Für $z = h\lambda \in S$ produziert ESV (81) beschränkte (d.h. stabile) Lösungen !
  2. Fordern daher, Schrittweite h so zu wählen, daß

  $$h\,\lambda \;\in\; S$$

  für alle „relevanten" ($\uparrow$) $\lambda : Re\,\lambda \leq 0$.

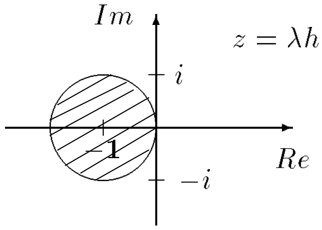
- Beispiel: Expliziter Euler: $R(z) = 1 + z$

  $\Rightarrow\; S = \{z \in \mathbb{C} : |z - (-1)| \leq 1\}$

  Für das Bsp. (80) aus $\boxed{\text{P IX}}$ $\Rightarrow\; \lambda = -50 = Re\,\lambda$

  $\Rightarrow$ Stabilitätsforderung: $-50h \in S$, d.h. $-1 \leq -50h + 1 \leq 1$

  $$\iff \quad \boxed{h \leq 2/50} \quad \text{(vgl. mit num. Resultaten !)}$$



■ A–Stabilität für ESV

- Ideal: $\mathbb{C}^- := \{z : Re\,z \leq 0\} \subset S$

  $\Rightarrow\;\; \forall\, \lambda \in \mathbb{C} : Re\,\lambda \leq 0$ hätte man Stabilität im obigen Sinne unabhängig von der Schrittweite h !

- **Definition 3.41:** (A–Stabilität für ESV)

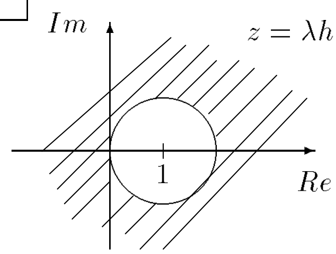
  $$\boxed{\text{Ein ESV heißt } \underline{A\text{–stabil}}\text{, falls } \mathbb{C}^- \subset S := \{z \in \mathbb{C} : |R(z)| \leq 1\}}$$

- Beispiel: Impliziter Euler: $R(z) = 1/(1 - z)$

  $$\begin{aligned} \Rightarrow \qquad S &= \{z \in \mathbb{C} : |1 - z|^{-1} \leq 1\} = \\ &= \{z \in \mathbb{C} : 1 \leq |z - 1|\} \end{aligned}$$

  $\Rightarrow\;\; \mathbb{C}^- \subset S$, d.h. impliziter Euler ist A–stabil !

  - d.h. impliziter Euler produziert beschränkte Näherungen, falls Testproblem (81) beschränkte Lösung hat !
  - d.h. Fehler im AW werden nicht vergrößert, falls $Re\,\lambda \leq 0$!



■ A–Stabilität von Runge–Kutta–Verfahren:

- Wenden allgemeines RK–Verfahren (o. B. d. Allg.: $N = 1,\ h_n = h$)

$$\begin{array}{c|c} c & A \\ \hline & b \end{array} \qquad \begin{aligned} u_{n+1} &= u_n + h \sum_{j=1}^{l} b_j f(t_n + c_j h, g_j) \\ g_i &= u_n + h \sum_{j=1}^{l} a_{ij} f(t_n + c_j h, g_j), \quad i = \overline{1, l} \end{aligned}$$

auf das Testproblem (81) $u'(t) = \underline{\underline{\lambda\, u(t) \equiv f}}$ an:

(86) $\begin{cases} g = u_n e + h\lambda A g, \quad e = (1,1,\ldots,1)^T, \text{ i. allgem. } (1 \to I), \\ u_{n+1} = u_n + h\lambda b^T g, \end{cases}$

$\Rightarrow (I - h\lambda A)g = u_n e$, d.h. $g = (I - h\lambda A)^{-1}(u_n e)$.

Also gilt die Darstellung

(87) $\quad u_{n+1} = u_n + h\lambda b^T (I - h\lambda A)^{-1} e u_n = R(h\lambda) u_n$

mit der Stabilitätsfunktion

(88) $\quad R(z) = 1 + z b^T (I - zA)^{-1} e.$

- Anwendung der <u>Cramerschen Regel</u> auf das zu (86) äquivalente System ($z = \lambda h$)

$$\begin{bmatrix} (I - zA) & 0 \\ -zb^T & 1 \end{bmatrix} \begin{bmatrix} g \\ \mathbf{u}_{n+1} \end{bmatrix} = u_n \begin{bmatrix} e \\ 1 \end{bmatrix}, \quad z = \lambda h,$$

liefert eine weitere Darstellung der Stabilitätsfunktion

(88)' $\quad R(z) = \dfrac{\det \begin{pmatrix} I - zA & e \\ -zb^T & 1 \end{pmatrix}}{\det (I - zA)} \underset{\uparrow}{=} \dfrac{\det (I - zA + zeb^T)}{\det (I - zA)}$

$$\det \begin{pmatrix} I - zA & e \\ -zb^T & 1 \end{pmatrix} = \det \begin{pmatrix} I - zA + zeb^T & 0 \\ -zb^T & 1 \end{pmatrix} = \det (I - zA + zeb^T)$$

- Betr. explizite RK–Verfahren, d.h.

$$A = \begin{bmatrix} 0 & & & \\ a_{21} & 0 & & \\ a_{31} & a_{32} & 0 & \\ \vdots & \ddots & \ddots & \\ a_{l1} & \ldots\ldots & a_{l,l-1} & 0 \end{bmatrix} = \text{ echte untere Dreiecksmatrix.}$$

$\Longrightarrow \det (I - zA) = \det \begin{bmatrix} 1 & \boldsymbol{O} \\ & \ddots & \\ x & & 1 \end{bmatrix} = 1.$

$\Longrightarrow |R(z)| = |\det (I - zA + zeb^T)| \to \infty$ für $|z| \to \infty$!

$\Longrightarrow$ $\boxed{\text{Explizite RK–Formeln können nicht } A\text{–stabil sein !}}$

$\Longrightarrow$ Damit gewinnen <u>implizite RK–Formeln</u> für <u>steife Dgl.</u> an Bedeutung !

- <u>Beispiele A–stabiler impliziter RK–Formeln:</u>
  1. Impliziter Euler (↑)
  2. Implizite Mittelpunktsregel: $R(z) = 1 + \dfrac{z}{1 - \frac{z}{2}} = \dfrac{2 + z}{2 - z}$

$f = \lambda u: u_{n+1} = u_n + h\lambda g$ mit $g = u_n + \frac{h}{2}\lambda g$

$$u_{n+1} = u_n + h\lambda \left(1 - \tfrac{h\lambda}{2}\right)^{-1} u_n = \left(1 + h\lambda\left(1 - \tfrac{h\lambda}{2}\right)^{-1}\right) u_n$$

$\Longrightarrow S := \{z \in \mathbb{C} : |R(z)| \leq 1\} = \mathbb{C}^-$ (mms)

3. Implizite Trapezregel (CN–Schema):

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

$$\uparrow$$

$$f = \lambda u: u_{n+1} = u_n + \frac{h\lambda}{2}(u_n + u_{n+1}) \Rightarrow u_{n+1} = \frac{(1+\frac{z}{2})}{(1-\frac{z}{2})} u_n, \quad z = \lambda h$$

$\Longrightarrow S = \mathbb{C}^-$ (mms)

4. Implizite l–stufige RK–Formeln vom Gauß–Typ (mms).

3.5.2 L–Stabilität

■ **<u>Motivation</u>**

- Für die implizite Mittelpunktsregel und für die implizite Trapezregel (CN–Schema) gilt z.B.

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

$\Longrightarrow |R(iy)| = 1 \quad \forall\, y \in \mathbb{R}^1$

$\Longrightarrow \lim\limits_{z\to-\infty} R(z) = \lim\limits_{z\to+\infty} R(z) = \lim\limits_{y\to\infty} R(iy)$, da $R(\cdot,\cdot)$ – rationale Funktion

$\Longrightarrow z \in \mathbb{C} : Re\, z << 0 \Rightarrow |R(z)| \underset{\leq}{\approx} 1$,

während exakte Lsg. e^z $(z = \lambda t)$ sehr klein ist.

$\Longrightarrow$ Steife Komponenten werden <u>langsam gedämpft</u> $(\uparrow)$

■ **<u>Definition 3.42:</u>** (L–Stabilität)

> Ein ESV mit der Stabilitätsfunktion $R(z)$ heißt <u>L–stabil</u>, falls es A–stabil ist und
>
> $$\lim_{z\to\infty} R(z) = 0.$$

■ **Beispiel:** Implizites Euler–Verfahren ist L–stabil, da

$$\lim_{z\to\infty} R(z) = \lim_{z\to\infty} \frac{1}{1-z} = 0.$$

Impl. MP–Regel und impl. TR sind zwar A–stabil, aber nicht L–stabil (mms) !

■ **Satz 3.43:**

Vor.: Für eine implizite, l–stufige RK–Formel $\begin{array}{c|c} c & A \\ \hline & b \end{array}$ mit regulärer Matrix A sei eine der folgenden beiden Bedingungen erfüllt:

(89) $\quad a_{lj} = b_j, \quad j = 1, 2, \ldots, l \quad$ bzw.

(90) $\quad a_{i1} = b_1, \quad i = 1, 2, \ldots, l.$

Bh.: Dann gilt:

$$\lim_{z\to\infty} R(z) = 0.$$

Beweis:

- Aus Darstellung (88) folgt:

$$\begin{aligned} R(z) &= 1 + zb^T(I - zA)^{-1}e = 1 + zb^T((-zA)(-\tfrac{1}{z}A^{-1} + I))^{-1}e \\ &= 1 - zb^T\left(I - \tfrac{1}{z}A^{-1}\right)^{-1} \tfrac{1}{z}A^{-1}e \end{aligned}$$

$$\Rightarrow \lim_{z\to\infty} R(z) = 1 - b^T A^{-1} e.$$

- Bed. (89): $\quad A^T e_l = b$

$$\Rightarrow 1 - b^T A^{-1} e = 1 - e_l^T A A^{-1} e = 1 - 1 = 0.$$

- Bed. (90): $\quad A\, e_1 = b_1 e \qquad \frac{1}{b_1} e_1 = A^{-1} e \quad (b_1 \neq 0$, da A reg. !)

$$\Rightarrow 1 - b^T A^{-1} e = 1 - b^T \tfrac{1}{b_1} e_1 = 1 - \tfrac{b_1}{b_1} = 1 - 1 = 0.$$

q.e.d.

3.5.3 B–Stabilität

■ Zur Behandlung **stark nichtlinearer steifer Probleme** benötigt man Verfahren mit noch stärkeren Stabilitätseigenschaften, wie etwa die ***B*–Stabilität:**

- Dazu betr. wir eine Klasse nichtlinearer Testprobleme

$$(91) \quad \left\{ \begin{array}{l} u'(t) = f(t, u(t)) \text{ mit der Eigenschaft} \\ (f(t,v) - f(t,w), v - w) \ \leq\ 0 \quad \forall\, v, w \in \mathbb{R}^N \quad \forall\, t \in I. \end{array} \right.$$

Solche Dgl.–Systeme nennt man dissipativ.

- Seien $v(t)$ und $w(t)$ zwei Lösungen des dissipativen Systems (91). Dann gilt:

  (93) $\quad \|v(t) - w(t)\| \leq \|v(t_0) - w(t_0)\| \quad \forall t \geq t_0,$

  d.h. insbesondere, daß Fehler im Startpunkt $t_0 = 0$ nicht zunehmen.

  Beweis: (mms)
  Btr. $m(t) := \|v(t) - w(t)\|^2 \quad \Rightarrow$ monoton fallend, da
  $$m'(t) = 2\left(v'(t) - w'(t),\ v(t) - w(t)\right) =$$
  $$= 2\left(f(t, v(t)) - f(t, w(t)),\ v(t) - w(t)\right) \leq 0 \quad \#$$

- Für ein B–stabiles Verfahren (ESV) fordert man nun das gleiche qualitative Verhalten !

■ **Definition 3.44:** (B–Stabilität)

> Ein ESV $u_{n+1} = u_n + h_n \varphi(t_n, u_n, t_n)$ heißt B–stabil, falls aus der Dissipativität
>
> (94) $\quad (f(t,v) - f(t,w), v - w) \leq 0 \quad \forall\, v, w \in \mathbb{R}^N \quad \forall\, t \in I$
>
> folgt, daß
>
> (95) $\quad \|v_1 - w_1\| \leq \|v_0 - w_0\|$
>
> für alle Schrittweiten $h \underset{(=)}{>} 0$ gilt, mit
> $v_1 = v_0 + h\varphi(t_0, v_0, h)$ und $w_1 = w_0 + h\varphi(t_0, w_0, h)$.

■ **Satz 3.45:** Aus der B–Stabilität folgt die A–Stabilität.

Beweis: (mms).

■ **Die folgenden 2 Bedingungen (algebraische Stabilitätsbed.) garantieren die B–Stabilität von (impliziten) RK–Formeln:**

1. $b_i \geq 0 \quad \forall\, i = \overline{1, l}$

2. Matrix $M = [m_{ij} := b_i a_{ij} + b_j a_{ji} - b_i b_j]$ ist positiv semi–definit.

Beweis: siehe [5] S. 193 f. #

- **Beispiele:** B–stabiler RK–Formeln (mms)

1. Impliziter Euler: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$

2. Implizite Mittelpunktsregel: $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$

3. Implizite RK–Formeln vom Gauß–Typ
(siehe [5] S. 194 ff auch weiter B–stabile RK–Formeln).

3.5.4 A-Stabilität von MSV

- **Wendet man lineares MSV**

(53) $\alpha_k u_{n+k} + \alpha_{k-1} u_{n+k-1} + \ldots + \alpha_0 u_n = h(\beta_k f_{n+k} + \ldots + \beta_0 f_n)$

auf das Testproblem

(81) $u'(t) = \lambda u(t) =: f(t, u(t)) \equiv f(u(t))$

an, so erhält man die Differenzengleichung

(96) $(\alpha_k - \mu\beta_k) u_{n+k} + (\alpha_{k-1} - \mu\beta_{k-1}) u_{n+k-1} + \ldots + (\alpha_0 - \mu\beta_0) u_n = 0$

mit $\mu = h\lambda$.

Die charakteristische Gleichung der Differenzengleichung (96) lautet:

(97) $\rho(z) \; - \; \mu\,\sigma(z) \; = \; 0.$

Im Stabilitätsbereich S werden nun diejenigen Werte $\mu \in \mathbb{C}$ zusammengefaßt, die stabile (d.h. beschr.) Lösungen der Differenzengleichung (96) zur Folge haben:

(98) $S := \{\mu \in \mathbb{C}:$ Alle Nullstellen ζ von $\rho(z) - \mu\,\sigma(z)$ erfüllen die Bed. (72),
d.h. $|\zeta| \leq 1$, falls einfach und $|\zeta| < 1$, falls mehrfach $\}$

Um das Testproblem stabil zu lösen, muß gelten:

(99) $h\,\lambda \; \in \; S.$

- **Bemerkung:** MSV (53) ist 0–stabil, genau dann, wenn $0 \in S$.

- **Definition 3.46:** (A–Stabilität von MSV)

Ein lineares MSV heißt A–stabil, falls $\mathbb{C}^- \subset S$.

■ **Satz 3.47:** (2. Dahlquist – Barriere)

> Ein A–stabiles lineares MSV muß die Ordnung
> $$p \;\leq\; 2$$
> haben.

Beweis: siehe Literatur [1], Satz 7.36, S. 318-321. #

■ **Beispiel:** Impl. TR = CN = Adams-Moulton ($k = 1$)
Die implizite Trapezregel (CRANK–NICOLSON)

$$u_{n+1} - u_n = h\left(\tfrac{1}{2}f_{n+1} + \tfrac{1}{2}f_n\right) \equiv \mu\left(\tfrac{1}{2}u_{n+1} + \tfrac{1}{2}u_n\right)$$

$$\left(1 - \mu\tfrac{1}{2}\right)z + \left(-1 - \mu\tfrac{1}{2}\right) = 0 \qquad \text{(charakteristische Gl.)}$$

$$\zeta = \frac{1 + \frac{\mu}{2}}{1 - \frac{\mu}{2}} = \frac{2+\mu}{2-\mu} \qquad \text{– einfache Nullstelle}$$

$$|\zeta| = \left|\frac{2+\mu}{2-\mu}\right| \leq 1 \text{ gdw. } \mu \in \mathbb{C}^- = S,$$

ist A–stabil ($\uparrow$ o.k.) und hat unter allen A–stabilen MSV mit der Konsistenzordnung $p \leq 2$ den kleinsten führenden Fehlerterm: $\Rightarrow$ lineare MSV sind für steife Dgl. uninteressant !

Kapitel 4

Numerische Behandlung hyperbolischer ARWA

4.1 Dreischichtige Differenzenschemata für hyperbolische ARWA

4.1.1 Allgemeine und kanonische Form dreischichtiger Schemata

■ **Allgemeine Form:**

(1) $$\boxed{\begin{array}{l} C_1 v^{j+1} + C_0 v^j + C_{-1} v^{j-1} = \tau \varphi^j, \quad j = 1, 2, \ldots, m-1, \\ \text{AB: } v^0, v^1 \text{ geg.,} \end{array}}$$

wobei $v = v^j : \omega_h \longrightarrow \mathbb{R}^1$ – Gitterfkt. auf j–ter Zeitschicht,
$\varphi = \varphi^j : \omega_h \longrightarrow \mathbb{R}^1$ – rechte Seite auf j–ter Zeitschicht,
C_{-1}, C_0, C_1 – lineare Operatoren (RB–eingearbeitet),
τ – Zeitschritt $(m\tau = T)$, h – Ortsdikretisierungsparameter,
$\omega \equiv \omega_h$ – Gitter für Ω (vgl. Pkt. 2.2.1).

■ **Entsteht z.B. bei der Diskretisierung:**

– hyperbolischer ARWA: $\dfrac{\partial^2 u}{\partial t^2} \longmapsto u_{\bar{t}t}$ (x: FDM, FEM)

$$\frac{\partial^2 u}{\partial t^2} + L\,u(x,t) = f(x,t), \ (x,t) \in Q_T = \Omega \times \boldsymbol{T}, \ \boldsymbol{T} = (0,T);$$

$+$ RB: 1. – 4. Art;

$+$ AB: $u(x,0) = u_0(x), \dfrac{\partial u}{\partial t}(x,0) = u_1(x), \ x \in \bar{\Omega};$

L–elliptischer Differentialausdruck (siehe [17] Nu II).

– parabolische ARWA: $\dfrac{\partial u}{\partial t} \longrightarrow u_{\overset{\circ}{t}}$ $\quad$ (x: FDM, FEM)
(siehe Kap. 2, insbesondere Pkt. 2.2).

■ **Kanonische Form:**

$$(2) \qquad \boxed{\begin{array}{l} B v^j_{\overset{\circ}{t}} + \tau^2 R v^j_{\bar{t}t} + A v^j = \varphi^j, \quad j = 1, 2, \ldots, m-1, \\ \text{AB: } v^0, v^1 \text{ geg.} \end{array}}$$

■ **Beziehung zwischen kanonischer und allgemeiner Form:**

$$(2) \Leftrightarrow B \frac{v^{j+1} - v^{j-1}}{2\tau} + \tau^2 R \frac{v^{j+1} - 2v^j - v^{j-1}}{\tau^2} + A v^j = \varphi^j \mid \cdot \tau$$

$$\Leftrightarrow \underbrace{\left(\frac{1}{2} B + \tau R\right)}_{=C_1} v^{j+1} + \underbrace{(\tau A - 2\tau R)}_{=C_0} v^j + \underbrace{\left(-\frac{1}{2} B + \tau R\right)}_{=C_{-1}} v^{j-1} = \tau \varphi^j$$

$$(3) \qquad \boxed{\begin{array}{l} C_1 = \frac{1}{2} B + \tau R \ , \ C_0 = \tau(A - 2R), \qquad C_{-1} = -\frac{1}{2} B + \tau R \\ \hline B \ = C_1 - C_{-1} \, , \ R \ = \frac{1}{2\tau}(C_1 + C_{-1}), \quad A = \frac{1}{\tau}(C_1 + C_0 + C_{-1}) \end{array}}$$

4.1.2 Ein allgemeines Stabilitätsresultat für dreischichtige Schemata

■ **Standardvoraussetzungen an (2):**

H – (diskreter) reeller Hilbert-Raum ($\to$ Grundraum)
i. S. Raum von Gitterfkt.: $\|\cdot\|$, $(\cdot,\cdot)$.

$$(4) \qquad \begin{cases} 1) & B, R, A : H \mapsto H - \text{lineare Operatoren;} \\ 2) & A = A^*, \quad R = R^*; \\ 3) & A > 0, \quad B + 2\tau R > 0 \quad (\Rightarrow 2C_1 > 0 \Rightarrow \exists\, C_1^{-1}\ !); \\ 4) & B, R, A - \text{zeitunabhängig.} \end{cases}$$

■ **Idee:** Rückführung des 3–schichtigen DS in H auf ein 2–schichtiges DS in $H \times H$ und Anwendung der Resultate aus Pkt. 2.2 !

Sei m – ungerade, $m \geq 3$:

$$\begin{array}{lll}
 & 2j+1 & \\
 & & \widehat{V} \equiv V^{j+1} = \binom{v^{2j+1}}{v^{2j}} \in \boldsymbol{\mathcal{H}} \\
\text{x} & 2j & \\
 & & \Phi \equiv \Phi^j = \binom{\varphi^{2j}}{\varphi^{2j-1}} \in \boldsymbol{\mathcal{H}} = H \times H \\
\text{x} & 2j-1 & \\
 & & V \equiv V^j = \binom{v^{2j-1}}{v^{2j-2}} \in \boldsymbol{\mathcal{H}} \\
 & 2j-2 & \\
 & \vdots & \\
\text{AB} & \left\{ \begin{array}{l} 1 \\ 0 \end{array} \right. & V^1 = \binom{v^1}{v^0} \in \boldsymbol{\mathcal{H}} = H \times H
\end{array}$$

$$j = \overline{1, \frac{m-1}{2}}$$

■ **Bezeichnungen:**

- Vorwärtige Differenzen:
  $V_t \equiv V_t^j := \frac{1}{2\tau}(\widehat{V} - V) \equiv \frac{1}{\bar{\tau}}(\widehat{V} - V)$ mit $\bar{\tau} = 2\tau$,

- Skalarprodukt und Norm im Faktorraum $\boldsymbol{\mathcal{H}} = H \times H$:
  $$(Y, Z) = \left(\begin{pmatrix} y^1 \\ y^2 \end{pmatrix}, \begin{pmatrix} z^1 \\ z^2 \end{pmatrix} \right) := (y^1, z^1) + (y^2, z^2) \quad \forall\, Y, Z \in \boldsymbol{\mathcal{H}} ,$$
  $$\|Y\|^2 = (Y, Y) = \|y^1\|^2 + \|y^2\|^2 \quad \forall\, Y = \begin{pmatrix} y^1 \\ y^2 \end{pmatrix} \in \boldsymbol{\mathcal{H}} .$$

- Energetisches Skalarprodukt:
  $\|Y\|^2_{\boldsymbol{\mathcal{A}}} := (\boldsymbol{\mathcal{A}}\, Y, Y)$ für $\boldsymbol{\mathcal{A}} : \boldsymbol{\mathcal{H}} \mapsto \boldsymbol{\mathcal{H}}$ – linear, s.a., p.d.

■ **Resultat:**

3–schichtiges DS (2) in $H: \;^{2j}_{2j-1}$	$\equiv$	$\boldsymbol{\mathcal{L}}\, V_t + \boldsymbol{\mathcal{A}}\, V = \Phi,\ V^1$ geg. / 2–schichtiges DS in $\boldsymbol{\mathcal{H}} = H \times H$	(5)

$\uparrow$ (mms)
Ansatz mit unbekannten Koeffizienten ($\boldsymbol{\mathcal{L}}$, $\boldsymbol{\mathcal{A}}$) und Koeffizientenvergleich:

mit $\boldsymbol{\mathcal{L}} = \begin{bmatrix} B+2\tau R & 2\tau(A-2R) \\ \boldsymbol{O} & B+2\tau R \end{bmatrix}$

$\boldsymbol{\mathcal{A}} = \begin{bmatrix} 2R & A-2R \\ A-2R & 2R \end{bmatrix}$

$$\boldsymbol{\mathcal{L}}\, V^{j+1} + (\bar\tau \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{L}})V^j = \bar\tau \Phi^j$$

(5)allg. Form $\Updownarrow$ $(2)^{2j}_{2j-1}$

$$\begin{aligned}
&\boldsymbol{\mathcal{L}}_{11} v^{2j+1} + \boldsymbol{\mathcal{L}}_{12} v^{2j} + (2\tau\boldsymbol{\mathcal{A}}_{11} - \boldsymbol{\mathcal{L}}_{11})v^{2j-1} + (2\tau\boldsymbol{\mathcal{A}}_{12} - \boldsymbol{\mathcal{L}}_{12})v^{2j-2} = 2\tau\varphi^{2j} \\
&\boldsymbol{\mathcal{L}}_{21} v^{2j+1} + \boldsymbol{\mathcal{L}}_{22} v^{2j} + (2\tau\boldsymbol{\mathcal{A}}_{21} - \boldsymbol{\mathcal{L}}_{21})v^{2j-1} + (2\tau\boldsymbol{\mathcal{A}}_{22} - \boldsymbol{\mathcal{L}}_{22})v^{2j-2} = 2\tau\varphi^{2j-1} \\
&(B+2\tau R)v^{2j+1} + 2\tau(A-2R)v^{2j} + (-B+2\tau R)v^{2j-1} = 2\tau\varphi^{2j} \\
&(B+2\tau R)v^{2j} + 2\tau(A-2R)v^{2j-1} + (-B+2\tau R)v^{2j-2} = 2\tau\varphi^{2j-1}
\end{aligned}$$

$$\begin{aligned}
\longrightarrow\ &\boldsymbol{\mathcal{L}}_{11} = B+2\tau R,\ \boldsymbol{\mathcal{L}}_{12} = 2\tau(A-2R),\ (2\tau\boldsymbol{\mathcal{A}}_{11} - \boldsymbol{\mathcal{L}}_{11}) = -B+2\tau R,\ (2\tau\boldsymbol{\mathcal{A}}_{12} - \boldsymbol{\mathcal{L}}_{12}) = 0 \\
&\boldsymbol{\mathcal{L}}_{21} = 0,\ \boldsymbol{\mathcal{L}}_{22} = B+2\tau R,\ (2\tau\boldsymbol{\mathcal{A}}_{21} - \boldsymbol{\mathcal{L}}_{21} = 2\tau(A-2R),\ (2\tau\boldsymbol{\mathcal{A}}_{22} - \boldsymbol{\mathcal{L}}_{22}) = -B+2\tau R
\end{aligned}$$

#

■ **Satz 4.1:**

Vor: 1. Standardvoraussetzungen (4): 1) – 4).
2. $R > \frac{1}{4}A$.
3. $B \geq 0$.

Bh.: Dann ist das 3–schichtige Schema (1) $\equiv$ (2) ($\Leftrightarrow$ 2–schichtiges Schema (5)) stabil (bzgl. AB u. RS) und $\forall\, j: (2j+1)\tau \leq T$ gilt die A-priori-Abschätzung:

(6) $\|V^{j+1}\|_{\boldsymbol{\mathcal{A}}} \leq \|V^j\|_{\boldsymbol{\mathcal{A}}} + T \max\limits_{1\leq k\leq j} \|\boldsymbol{\mathcal{L}}^{-1}\Phi^k\|_{\boldsymbol{\mathcal{A}}}$

Beweis: folgt aus Satz 2.16 angewandt auf das 2–schichtige DS (5):

- Standardvor. 1) - 4) an 2–schichtige DS aus Kap. 2, Pkt. 2.2.2:

1) $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{L}} : \boldsymbol{\mathcal{H}} \longrightarrow \boldsymbol{\mathcal{H}}$ – lineare Operatoren: o.k.

2) $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^* > 0$ folgt sofort aus 2) + 3) von (4) und $R > \frac{1}{4}A$:
$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^*$ (mms) unter Benutzung von 2): o.k.
$\boldsymbol{\mathcal{A}}$ p.d.:

$$(\boldsymbol{\mathcal{A}}\, V, V) = \left(\begin{bmatrix} 2R & A-2R \\ A-2R & 2R \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \end{bmatrix}, \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} \right) =$$

$$= (2Rv^1, v^1) + (2Rv^2, v^2) + \underline{((A-2R)v^2, v^1) + ((A-2R)v^1, v^2)}$$

$$= (2Rv^1, v^1) + (2Rv^2, v^2) + ((A-2R)(v^2 - v^1), v^1 - v^2) +$$
$$+((A-2R)v^1, v^1) + ((A-2R)v^2, v^2)$$

$$= \underbrace{(Av^1, v^1) + (Av^2, v^2)} + ((2R - A)(v^2 - v^1), v^2 - v^1) =$$
$$= \tfrac{1}{2}(A(v^1 + v^2), v^1 + v^2) + \tfrac{1}{2}(A(v^2 - v^1), v^2 - v^1)$$

$$= \tfrac{1}{2}(A(v^1 + v^2), v^1 + v^2) + (\underbrace{\left(2R - \frac{1}{2}A\right)}_{=2\left(R - \frac{1}{4}A\right) > 0}(v^2 - v^1), v^2 - v^1) > 0$$

$$\forall\, V = \binom{v^1}{v^2} \neq \boldsymbol{O}$$

3) $\boldsymbol{\mathcal{L}} > 0$: Folgt aus der noch zu beweisenden Stabilitätsbedingung
$\boldsymbol{\mathcal{L}} \geq \frac{\bar{\tau}}{2}\boldsymbol{\mathcal{A}} = \tau\boldsymbol{\mathcal{A}} > 0$.

4) $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{L}}$ – seien zeitunabhängig folgt aus (4) 4).

- Stabilitätsbed. $\boldsymbol{\mathcal{L}} \geq \frac{\bar{\tau}}{2}\boldsymbol{\mathcal{A}} = \tau\boldsymbol{\mathcal{A}}$ 2–schichtiges DS:

$$((\boldsymbol{\mathcal{L}} - \tau\boldsymbol{\mathcal{A}})V, V) =$$

$$= \left(\left[\begin{pmatrix} B + 2\tau R & 2\tau(A-2R) \\ \boldsymbol{O} & B + 2\tau R \end{pmatrix} - \tau \begin{pmatrix} 2R & A-2R \\ A-2R & 2R \end{pmatrix} \right] \begin{bmatrix} v^1 \\ v^2 \end{bmatrix}, \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} \right)$$

$$= \left(\begin{bmatrix} B & \tau(A-2R) \\ -\tau(A-2R) & B \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \end{bmatrix}, \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} \right) =$$

$$= (Bv^1, v^1) + (Bv^2, v^2) \geq 0, \text{ da } B \geq 0.$$

q.e.d.

4.1.3 Beispiel: Gewichtete dreischichtige Differenzenschemata für die Saitenschwingungsgleichung

■ **Betr. ARWA für Saitenschwingungsgleichung:**

(7)
$$\frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = f(x,t),\ x \in \Omega = (0,1),\ t \in \boldsymbol{T} = (0,T),$$

AB: $\left.\begin{array}{l} u(x,0) = u_0(x),\ x \in \bar{\Omega} = [0,1], \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x),\ x \in \bar{\Omega} = [0,1], \end{array}\right\}$ Konsistenzbed.: $\underbrace{u_0(z) = u_1(z) = 0 \quad z = 0,1}$

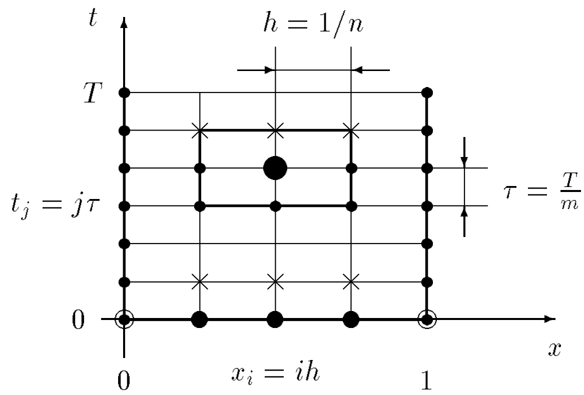
RB: $\boxed{1.}$ – 4. Art; z.B. fest eingespannt: $u(0,t) = u(1,t) = 0$

- AB: 1. $v_i^0 = u_0(x_1),\ i = \overline{0,n}$

  2. $u(x_i,\tau) = u(x_i,0) + \int\limits_0^\tau \dfrac{\partial u(x_i,t)}{\partial t}\,dt$

  $\approx u_0(x_i) + \tau u_1(x_i)$

  Also:
  $v_i^1 = u_0(x_i) + \tau u_1(x_i)$ bzw.
  $v_t^0(x) = u_1(x), x \in \bar{\omega}_h$ bzw. unter
  Benutzung $\frac{\partial^2 u(x,0)}{\partial t^2} = \ldots O(\tau^3)$



- Bem.: $\begin{array}{l}\Omega{=}(0,l) \\ \boldsymbol{T}{=}(0,T)\end{array} \vdash \boxed{\begin{array}{l} x'{=}x/l \\ t'{=}t \cdot a/l \end{array}} \longrightarrow \boxed{\begin{array}{l} \Omega'{=}(0,1), a' = 1 \\ T'{=}T \cdot a/l \end{array}} \longrightarrow \ldots$

■ **Gewichtetes, 3–schichtiges DS für (7):** $(0 \le \sigma \le 1)$

(8)
$$v_{\bar{t}t}^j - \sigma a^2 v_{\bar{x}x}^{j+1} - (1-2\sigma) a^2 v_{\bar{x}x}^j - \sigma a^2 v_{\bar{x}x}^{j-1} = \varphi^j(x),\ x \in \omega_h,\ j = \overline{1,m-1}$$

AB: $v_i^0 = u_0(x_i), \quad v_i^1 = u_0(x_i) + \tau u_1(x_i), i = \overline{0,n}$ bzw. $i = \overline{1,n-1}$

RB: $v_0^j = v_n^j = 0, \quad j = \overline{0,m}$

■ **Formulierung von (8) als Operatorgleichung:**

Def. diskreten Hilbert–Raum

$$H = L_2(\omega_h) = L_2^0(\bar{\omega}_h) \ni v : \bar{\omega}_h \to I\!R^1 : v|_{\gamma_h = \bar{\omega}_h \setminus \omega_h} = 0$$

und den Differenzenoperator

$$\bar{A}v := -a^2 v_{\bar{x}x}, \quad v \in H.$$

Dann läßt sich (8) als dreischichtiges Schema in allgemeiner und kanonischer Form schreiben:

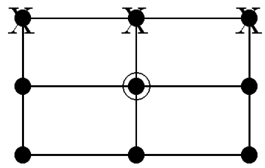
(9) $\widehat{=}$ allgem. Form (1)

$$\boxed{\begin{array}{l}\underbrace{\left(\frac{1}{\tau}I + \tau\sigma\bar{A}\right)}_{=:C_1} v^{j+1} + \underbrace{\left(-\frac{2}{\tau} + (1-2\sigma)\tau\bar{A}\right)}_{=:C_0} v^j + \underbrace{\left(\frac{1}{\tau}I + \tau\sigma\bar{A}\right)}_{=:C_{-1}} v^{j-1} = \tau\varphi \\ + \text{AB: } v^0, v^1 \text{ geg.}\end{array}}$$

$$\Updownarrow \quad \begin{array}{ll}(3) & B = C_1 - C_{-1} = 0 \\ & R = \frac{1}{2\tau}(C_1 + C_{-1}) = \frac{1}{\tau^2}I + \sigma\bar{A} \\ & A = \frac{1}{\tau}(C_1 + C_0 + C_{-1}) = \bar{A}\end{array}$$

(10) $\widehat{=}$ kanon. Form (2)

$$\boxed{\begin{array}{l}(I + \tau^2\sigma\bar{A})v_{\bar{t}t} + \bar{A}v = \varphi \\ + \text{AB: } v^0, v^1 \text{ geg.}\end{array}}$$



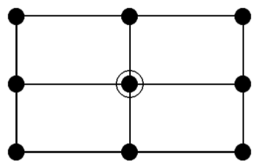
$j = \overline{1, m-1}$

$i = \overline{1, n-1}$

■ **Resultat:** Approximation + Stabilität $\Longrightarrow$ diskr. Konvergenz

1. Approximation (Konsistenz): (mms)

   $\psi = \psi(x,t) = u^j_{\bar{t}t} - \sigma a^2 u^{j+1}_{\bar{x}x} - (1-2\sigma)a^2 u^j_{\bar{x}x} - \sigma a^2 u^{j-1}_{\bar{x}x} - \varphi^j(x,t) = .?.$

   – Methode: Taylor–Entwicklung im Pkt. $(x,t) = (x_i, t_j)$
   (→ siehe Pkt. 2.1.2 für 2–schichtige DS)

   – Resultate: $\psi = O(\tau^2 + h^2) \quad \forall \sigma \in [0,1]$

   $\psi = O(\tau^2 + h^4)$ falls $\sigma = \sigma_* := \bar{\sigma} - \frac{h^2}{12\tau^2} \in [0,1]$,
   $\varphi = f - \sigma_* \tau^2 \frac{\partial^2 f}{\partial x^2}$,
   $\bar{\sigma} \neq \bar{\sigma}(h,\tau)$ bel. Konst. (Stabil.)
   $(a^2 = 1)$

2. Stabilität:

   – Methode: a) Fourier–Analyse nach Eigenfkt. von $\bar{A}$
   (→ siehe Pkt. 2.1.3 für 2–schichtige DS !)

   b) Allgemeine Stabilitätstheorie aus Pkt. 4.1.2:
   Satz 4.1: ⇒ Voraussetzungen überprüfen !
   1. Standardvor. (4): 1) – 4) o.k.
   2. $R \equiv \frac{1}{\tau^2}I + \sigma\bar{A} > \frac{1}{4}A \equiv \frac{1}{4}\bar{A}$!
   3. $B \equiv \boldsymbol{O} \geq 0$ o.k.

   – Resultate: 2. ?

$\sigma \geq \dfrac{1}{4} \Longrightarrow$ unbedingt stabil: $\Rightarrow$ A-priori–Abschätzung (6).

$\sigma = 0 \Longrightarrow$ bedingt stabil: $\Rightarrow$ A-priori–Abschätzung (6).

$$\Downarrow \qquad\qquad\qquad \Uparrow$$

(11) CFL–Bed. $\boxed{\dfrac{\tau}{h} \leq \dfrac{1}{a}}$ = klassische Stabilitätsbed. für das explizite DS von Courant, Friedrichs und Levi

Tatsächlich:

$$\underset{\substack{\uparrow \\ \Lambda v := -v_{\bar{x}x}}}{\tfrac{1}{4} A} = \tfrac{1}{4} a^2 \Lambda \;\leq\; \tfrac{1}{4} a^2 \lambda_{\max}(\Lambda)\, I \underset{\substack{\nearrow \\ \text{Pkt. 2.1.3: } \lambda_{\max}(\Lambda) < \frac{4}{h^2}}}{<} \tfrac{1}{4} a^2 \tfrac{4}{h^2} I \underset{\substack{\Updownarrow \\ \frac{a^2}{h^2} \leq \frac{1}{\tau^2}}}{\leq} \tfrac{1}{\tau^2} I \;!\; (\sigma = 0)$$

– **Ü 4.1** Untersuchen Sie die Stabilität des $O(\tau^2 + h^4)$–Schemas !

4.2 Zurückführung auf AWA mittels Semidiskretisierung

- **Btr. zunächst hyperbolische ARWA in klassischer Formulierung** (vgl. [17] Nu II, Pkt. 1.2):

(12)

Ges. $u(x,t) \in X = \overset{x,t}{C^{2,2}}(Q_T) \cap C^{0,1}(\bar{\Omega} \times [0,T)) \cap C^{1,0}(\Omega \cup \Gamma_2 \cup \Gamma_3 \times (0,T))$:

$$\frac{\partial^2 u}{\partial t^2} \underbrace{- \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x,t)\frac{\partial u}{\partial x_j}\right) + a(x,t)\, u(x,t)}_{\text{ellipt. Anteil}} = f(x,t),$$

$$\forall\ (x,t) \in Q_T = \Omega \times (0,T)$$

+ RB: o. B. d. Allg.

$$\left.\begin{aligned} u(x,t) = g_1(x,t) &\overset{\downarrow}{\equiv} 0, \quad x \in \Gamma_1 \\ \frac{\partial u}{\partial N} := \sum_{i,j=1}^{d} a_{ij}(x,t)\frac{\partial u}{\partial x_j} n_i &= g_2(x,t), \quad x \in \Gamma_2 \\ \frac{\partial u}{\partial N} + \kappa(x,t) u(x,t) &= g_3(x,t), \quad x \in \Gamma_3 \end{aligned}\right\} t \in \boldsymbol{T} = (0,T)$$

+ AB: $u(x,0) = u_0(x), \quad \dfrac{\partial u}{\partial t}(x,0) = u_1(x), \quad x \in \bar{\Omega}$

■ **Linienvariationsformulierung (vgl. Pkt. 2.3.1):**

(13)

$$\boxed{\begin{array}{l} \text{Ges. } u \in C(\bar{\boldsymbol{T}}, V_0) \text{ mit } \dot{u} \in L_\infty\ (\boldsymbol{T}, V_0) \cap C(\bar{\boldsymbol{T}}, L_2(\Omega)) \text{ und} \\ \ddot{u} \in L_\infty\ (\boldsymbol{T}, L_2(\Omega)): \\ (\ddot{u}(t), v)_0 + a(t; u(t), v) = < F(t), v > \quad \forall v \in V_0, \ \ \forall \text{ f.ü. } t \in \boldsymbol{T} \\ \ + \text{AB:} \quad u(0) = u_0 \text{ in } V_0, \\ \qquad\qquad \dot{u}(0) = u_1 \text{ in } L_2(\Omega) \\ \text{mit } V_0 := \{v \in V = H^1(\Omega) : v = 0 \text{ auf } \Gamma_1\}, \quad F \in C(\boldsymbol{T}, V_0^*). \end{array}}$$

Bemerkung: Existenz–, Eindeutigkeits– und Regularitätsaussagen siehe [10] Method of Rothe in Evolution Equations. Teubner–Verlag, Leipzig 1985.

■ **Galerkin–FEM–Semidiskretisierung von (13):**

Ansatz: $u_h(x,t) = \sum\limits_{i\in\omega_h} u^{(i)}(t)\, p^{(i)}(x) \in V_{0h} = \text{span}\ \{p^{(i)} : i \in \omega_h\} \subset V_0$ (14)

$(13)_h$

$$\boxed{\begin{array}{ll} \text{Ges. } u_h(x,t): & \dfrac{d^2}{dt^2}(u_h, v_h)_0 + a(t; u_h, v_h) = < F, v_h > \quad \forall\ v_h \in V_{0h}\ \ \forall \text{ f.ü. } t \in \boldsymbol{T} \\ & + \text{AB:} \quad \text{z.B. } L_2\text{–Projektion:} \\ & \qquad (u_h(\cdot,0), v_h)_0 = (u_0(\cdot), v_h)_0 \quad \forall\ v_h \in V_{0h}, \\ & \qquad (\dot{u}_h(\cdot,0), v_h)_0 = (u_1(\cdot), v_h)_0 \quad \forall\ v_h \in V_{0h}. \end{array}}$$

$\Updownarrow$

$(\underline{13})_h$

$$\boxed{\begin{array}{ll} \text{Ges. } \underline{u}_h(t) := [u^{(i)}(t)]_{i\in\omega_h}: & M_h \underline{\ddot{u}}_h(t) + K_h(t) \underline{u}_h(t) = \underline{f}_h(t), \quad t \in \boldsymbol{T} \\ & + \text{AB:} \quad \underline{u}_h(0) = M_h^{-1} \underline{u}_{0h}, \\ & \qquad\quad \dot{\underline{u}}_h(0) = M_h^{-1} \underline{u}_{1h}. \end{array}}$$

$\Updownarrow$ System gewöhnlicher Dgl. 2. Ordnung

$(\underline{13})''_h$

$$\boxed{\begin{array}{l} \text{Ges. } y(t) := M_h^{0.5} \underline{u}_h(t) : I \equiv \bar{\boldsymbol{T}} \longmapsto \mathbb{R}^{N=N_h} : \\ y''(t) = f(t, y(t)) := -M_h^{-0.5} K_h M_h^{-0.5} y(t) + M_h^{-0.5} \underline{f}_h(t), \quad t \in I \\ + \text{AB: } y(0) = y_0 := M_h^{-0.5} \underline{u}_{0h}, \quad y'(0) = v_0 := M_h^{-0.5} \underline{u}_{1h}. \end{array}}$$

$$\Updownarrow \quad U := \begin{bmatrix} y \\ v \end{bmatrix} \text{ mit } v(t) = y'(t)$$

(13)'$_h$

Äquivalentes System gew. Dgl. 1. Ordnung	
$U'(t) = F(t, u(t)), \quad t \in I$	$y'(t) = v(t)$
$U(0) = (y_0, y_1)^T$	$v'(t) = f(t, y(t))$
mit $F(t, u(t)) := \begin{bmatrix} v(t) \\ f(t, y(t)) \end{bmatrix}$	$y(0) = y_0, \quad v(0) = y_1$

■ **Volldiskretisierung mittels gew. 3–schichtiger DS:**

(13)$_{h\tau}$

> Ges. $\underline{v}^j = \underline{v}_h^j := [v^{(i)}(t_j)]_{i \in \omega_h} \in \mathbb{R}^{N_h}$:
> $M\underline{v}_{\bar{t}t}^j + \sigma K(t_{j+1})\underline{v}^{j+1} + (1-2\sigma)K(t_j)\underline{v}^j + \sigma K(t_{j-1})\underline{v}^{j-1} = \underline{\varphi}^j$
> $+$ AB: $\underline{v}^0$, $\underline{v}^1$ geg.
> mit $M = M_h, \quad K = K_h(t), \quad \underline{\varphi}^j \overset{\text{z.B.}}{=} \underline{f}_h(t_j)$.

Allgemeine Form: $C_1\underline{v}^{j+1} + C_0\underline{v}^j + C_{-1}\underline{v}^{j-1} = \tau\underline{\varphi}^j$

$K \neq K(t)$ mit $C_1 = \frac{1}{\tau}M + \tau\sigma K, \quad C_0 = -\frac{2}{\tau}M + (1-2\sigma)K, \quad C_{-1} = \frac{1}{\tau}M + \tau\sigma K$

Kanonische Form: $B = C_1 - C_{-1} = \boldsymbol{O}$, $R = \frac{1}{2\tau}(C_1 + C_{-1}) = \frac{1}{\tau^2}M + \sigma K$,
$A = \frac{1}{\tau}(C_1 + C_0 + C_{-1}) = K$

$$\boxed{\begin{array}{l} (M + \tau^2\sigma K)\underline{v}_{\bar{t}t} + K\underline{v} = \varphi \\ + \text{AB: } \underline{v}^0, \underline{v}^1 \text{ geg.} \end{array}}$$

Stabilität: $R \equiv \frac{1}{\tau^2}M + \sigma K > \frac{1}{4}A \equiv \frac{1}{4}K$

- $\sigma \geq \frac{1}{4} \Rightarrow$ unbedingt stabil !
- $\sigma = 0 \Rightarrow$ bedingt stabil: CFL–Bedingung !

$$\frac{4}{\tau^2} > \max_{\underline{v}} \frac{(K\underline{v}, \underline{v})}{(M\underline{v}, \underline{v})} = \lambda_{\max}(M^{-1}K) = \lambda_{\max}(M^{-0.5}KM^{-0.5}) = \underset{\substack{\uparrow \\ \boxed{\text{siehe Ü 2.6}}}}{\frac{c^2}{h^2}}$$

$$\Rightarrow \boxed{\begin{array}{c} \tau^2 < \dfrac{4}{\lambda_{\max}(M_h^{-1}K_h)} \\ \hline \dfrac{\tau}{h} < \dfrac{2}{c} \end{array}} = \text{CFL–Bed. !}$$

$\Rightarrow \tau < 2/\sqrt{\lambda_{\max}(M_h^{-1}K_h)}$.

■ **Volldiskretisierung mittels A–stabiler Einschrittformeln angewandt auf ($\underline{13}$)'$_h$: $U'(t) = F(t, U(t))$, $U(o)$ geg.:**

z.B. implizite Trapez–Regel (A–stabil $\Rightarrow$ unbedingt stabil):

$$U_{j+1} = U_j + \frac{\tau}{2}(F_j + F_{j+1}), \quad U_0 = \begin{bmatrix} y_0 \\ v_0 \end{bmatrix} \text{ geg.}$$

$$\begin{cases} y_{j+1} = y_j + \frac{\tau}{2}(v_j + v_{j+1}) \\ v_{j+1} = v_j + \frac{\tau}{2}(f(t_j, y_j) + f(t_{j+1}, y_{j+1})) \end{cases}$$

$$\boxed{\begin{aligned} \underline{u}_h^{j+1} &= \underline{u}_h^j + \frac{\tau}{2}(\underline{\dot{u}}_h^j + \underline{\dot{u}}_h^{j+1}) \\ M_h \underline{\dot{u}}_h^{j+1} &= M_h \underline{\dot{u}}_h^j + \frac{\tau}{2}(\underline{f}_h^j + \underline{f}_h^{j+1} - K_h \underline{u}_h^j - K_h \underline{u}_h^{j+1}) \\ \underline{\text{AB:}} &\quad \underline{u}_h^0 := M_h^{-1} \underline{u}_{0h}, \quad \underline{\dot{u}}_h^0 := M_h^{-1} \underline{u}_{1h} \end{aligned}}$$

$\widehat{=}$ Newmark–Schema (siehe P VI und [11], S. 165-169) mit $\beta = \frac{1}{4}$ und $\gamma = \frac{1}{2}$ (mms)

$\downarrow$

$$\boxed{\begin{aligned} & M_h \, \underline{\ddot{u}}_h^{j+1} + \cancel{C_h \, \underline{\dot{u}}_h^{j+1}} + K_h \, \underline{u}_h^{j+1} = \underline{f}_h^{j+1} \\ & \underline{u}_h^{j+1} = \underline{u}_h^j + \tau \, \underline{\dot{u}}_h^j + \frac{\tau^2}{2} \{(1 - 2\beta) \, \underline{\ddot{u}}_h^j + 2\beta \, \underline{\ddot{u}}_h^{j+1}\} \\ & \underline{\dot{u}}_h^{j+1} = \underline{\dot{u}}_h^j + \tau \{(1 - \gamma) \, \underline{\ddot{u}}_h^j + \gamma \, \underline{\ddot{u}}_h^{j+1}\} \end{aligned}}$$

Literaturverzeichnis

[1] P. Deuflhard and F. Bornemann. *Numerische Mathematik II: Integration gewöhnlicher Differentialgleichungen.* de Gruyter Lehrbuch, Berlin, New York, 1994.

[2] H. Gajewski, K. Gröger, and K. Zacherias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen.* Akademie-Verlag, Berlin, 1974.

[3] R. Grigorieff. *Numerik gewöhnlicher Differentialgleichungen I.* B.G. Teubner, Stuttgart, 1972.

[4] C. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen.* B.G. Teubner, Stuttgart, 1992.

[5] E. Hairer, S. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer Verlag, Berlin, Heidelberg, 1987.

[6] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems.* Springer Verlag, Berlin, Heidelberg, 1991.

[7] P. J. P. J. R. Dormand. *A Family of embedded Runge–Kutta formulae.* J. Comp. Appl. Math. 6, 19 – 26, 1980.

[8] P. J. P. J. R. Dormand. *Higher order embedded Runge–Kutta formulae.* J. Comp. Appl. Math. 7, 67 – 75, 1981.

[9] M. Jung and U. Langer. *Skriptum zur Vorlesung FEM (Eine Einführung für Ingenieurstudenten).* TU Chemnitz (Fakultät für Mathematik) und Johannes Kepler Universität (Institut für Mathematik), Chemnitz und Linz, 1993.

[10] J. Kacur. *Method of Rothe in Evolution Equation.* B.G. Teubner, Leipzig, 1985.

[11] N. Kikuchi. *Finite Element Method in Mechanics.* Cambridge University Press, Cambridge, 1986.

[12] V. Korneev and U. Langer. *Approximate Solution of Plastic Flow Theory Problems.* B.G. Teubner, Leipzig, 1984.

[13] D. Kröner. *Numerical schemes for conservation laws.* B.G. Teubner, Stuttgart, 1996.

[14] O. Ladyschenskaja. *Aufgaben der Mathematischen Physik.* Nauka, Moskau, 1973. (in Russisch).

[15] U. Langer. *Skriptum zur Vorlesung MULTIGRID – METHODEN.* Johannes Kepler Universität, Institut für Mathematik, Linz, 1996.

[16] U. Langer. *Skriptum zur Vorlesung NUMERIK I (Operatorgleichungen).* Johannes Kepler Universität, Institut für Mathematik, Linz, 1996.

[17] U. Langer. *Skriptum zur Vorlesung NUMERIK II (RWA).* Johannes Kepler Universität, Institut für Mathematik, Linz, 1996.

[18] K. Rektorys. *The Method of Discretization in Time and PDEs.* Dordrecht, Bosten, 1982.

[19] H. Stetter. *Analysis of Discretization Methods for Ordinary Differential Equations.* Springer Verlag, Berlin, Heidelberg, New York, 1973.

[20] J. Thomas. *Numerical Partial Differential Equations: Finite Difference Methods.* Springer, New York, 1995.

[21] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems.* Springer Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[22] E. Zeidler. *Vorlesungen über nichtlineare Funktionalanalysis I: Fixpunktsätze.* Teubner-Texte zur Mathematik. B.G. Teubner, Leipzig, 1976.

[23] E. Zeidler. *Vorlesungen über nichtlineare Funktionalanalysis II: Monotone Operatoren,* volume 9 of *Teubner-Texte zur Mathematik.* B.G. Teubner, Leipzig, 1977.

1

Kapitel 5

Praktikum

“Zeitintegrationsmethoden“

(Gundolf Haase)

zur Vorlesung

“Numerik III (Anfangs- und Anfangsrandwertaufgaben)“

P R A K T I K U M

"Zeitintegrationsmethoden"

(Gundolf Haase)

zur Vorlesung

"Numerik III (Anfangs- und Anfangsrandwertaufgaben)"

$\boxed{\textbf{P 0}}$ Einführungspraktikum[1]

5.1 Einführung

- Im Praktikum "Zeitintegrationsverfahren" werden
  - Übungsaufgaben $\boxed{\textbf{Ü 1}}$ – $\boxed{\textbf{Ü 22}}$ gelöst,
  - kleinere Programmieraufgaben mit numerischen Experimenten $\boxed{\textbf{E 1}}$ – $\boxed{\textbf{E 9}}$ durchgeführt und
  - von einem Team (in der Regel 2 Studenten) ähnlich zum Praktikum in Numerik II eine größere Praktikumsaufgabe $\boxed{\textbf{P x}}$ gelöst.
- Für die größere Praktikumsaufgabe stehen die parabolischen bzw. hyperbolischen Analoga zu den im Praktikum zur Vorlesung Numerik II vorgestellten Aufgaben zur Wahl :

[1] Für jedes Praktikum Px stehen 45 Minuten zur Verfügung.

Px	RWA	Numerik II Team	Numerik III Team	parabol. ARWA	hyperbol. ARWA	Bem.
1	Kolben			x		
3	Schellbach				x	nicht-linear
4	Courant				x	
7	Kanal 1			x		
9	Kammer				x	
11	Wärmeaus-tauscher			x		
12	TWD			x		
13	E-Magnet				x	
14	Chip			x		
15	Draht			x		Konvek-tion
16	Tetraeder			x		3D-FEM

Abrechnung : 1. Vortrag (15 - 20 min)
2. Belegarbeit

- Bewertung : Ü : 40 %
E : 20 %
P : 40 %

P I Praktikum I

5.2 Numerische Behandlung parabolischer ARWA

5.2.1 Differenzenverfahren

5.2.1.1 Ein spezielles diskretes Eigenwertproblem

stetig	diskret
Ges. $u \in C^2(0,\ell) \cap C[0,\ell]$: $u \not\equiv 0$ und $\lambda \in \mathbf{R}$: $$\begin{aligned} -u''(x) &= \lambda u(x)\,,\ x \in \Omega = (0,\ell) \qquad (5.2.1) \\ u(0) &= u(\ell) = 0 \end{aligned}$$	Ges. $v : \overline{\omega}_h \mapsto \mathbf{R}^1$: $v \not\equiv 0$ und $\lambda \in \mathbf{R}$, $x \in \omega_h = \{x_\imath = \imath h : \imath = \overline{1, n-1}\}$: $$\begin{aligned} -v_{\overline{x}x}(x) &= \lambda v(x) \qquad (1_h) \\ v_0 &= v_n = 0 \qquad h = \ell/n \end{aligned}$$
Homog. lin. Dgl. mit konst. Koeff. ! $$-u'' - \lambda\, u = 0$$	Matrixeigenwertproblem $$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{bmatrix}$$
1. Eigenfunktionen	
$$u_k(x) = \sqrt{\frac{2}{\ell}}\sin\frac{k\pi x}{\ell}\,, \quad x \in [0,\ell]$$ $k = 1, 2, \ldots$	$$\mu_k(x) = \sqrt{\frac{2}{\ell}}\sin\frac{k\pi x}{\ell}\,, \quad x \in \overline{\omega}_h$$ $k = 1, 2, \ldots, n-1$
2. Eigenwerte	
$$\lambda_k = \left(\frac{k\pi}{\ell}\right)^2 \qquad k = 1, 2, \ldots$$ $0 < \left(\frac{\pi}{\ell}\right)^2 = \lambda_1 < \lambda_2 < \lambda_3 < \longrightarrow \infty$	$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi h}{2\ell} \quad k = 1, 2, \ldots, n-1$$ $\frac{8}{\ell^2} \leq \lambda_1 < \lambda_2 < \ldots < \lambda_k < \ldots < \lambda_{n-1} < \frac{4}{h^2} \longrightarrow \infty$
3. Orthonormalität der Eigenfunktionen	
$$(u_k, u_m)_{L_2(\Omega)} := \int\limits_0^\ell u_k(x) u_m(x)\,dx = \delta_{k,m}$$ $k, m = 1, 2, \ldots$	$$(\mu_k, \mu_m)_h := \sum_{x \in \omega_h} h \mu_k(x) \mu_m(x) = \delta_{k,m}$$ $k, m = 1, 2, \ldots, n-1$

4. Ableitungen (bzw. Differenzen) der Eigenfunktionen

$u_k'(x) = \sqrt{\lambda_k}\sqrt{\dfrac{2}{\ell}}\cos\dfrac{k\pi x}{\ell}\,, \quad x\in[0,\ell]$ $k=1,2,\ldots$	$\mu_{k,\overline{x}}(x) = \sqrt{\lambda_k}\sqrt{\dfrac{2}{\ell}}\cos\dfrac{k\pi(x-h/2)}{\ell}\,, \quad x\in\overline{\omega}_h$ $k=1,2,\ldots,n-1$

5. Orthogonalität der Ableitungen (bzw. Differenzen) der Eigenfunktionen

$\displaystyle\int_0^\ell u_k'(x)u_m'(x)dx = \lambda_k\delta_{k,m}$ $k=1,2,\ldots$	$\displaystyle(\mu_{k,\overline{x}},\mu_{m,\overline{x}}]_h = \sum_{x\in\vec{\omega}_h} h\mu_{k,\overline{x}}(x)\mu_{m,\overline{x}}(x) = \lambda_k\delta_{k,m}$ $k=1,2,\ldots,n-1$ $\vec{\omega}_h = \{x_\imath = \imath h \,:\, \imath=\overline{1,n}\}$

6. Über die Vollständigkeit des Systems der Eigenfunktionen

Sei $f\in L_2(\Omega)$, $\Omega=(0,\ell)$. Dann gilt für die Fourierreihe : $\displaystyle f(x) \;\text{“}=\text{“}\; \sum_{k=1}^{\infty} f_k\, u_k(x)$ $\downarrow$ $f_k = (f,u_k)_{L_2(0,\ell)}$ • i.S. $L_2(0,\ell)$ falls $f\in L_2(0,\ell)$, d.h. $\displaystyle\Vert f-\sum_{k=1}^{m} f_k\,u_k(x)\Vert_{L_2} \overset{m\to\infty}{\longrightarrow} 0.$ • i.S. $\overset{o}{W}{}_2^1(\Omega)$, falls $f\in\overset{o}{W}{}_2^1(\Omega)$. • i.S. $C(\overline{\Omega})$, falls $f\in\overset{o}{W}{}_2^1(\Omega)\cap C^2(\overline{\Omega})$.	$\forall f(\cdot): \omega_h \longmapsto \mathbf{R}^1$ -Gitterfunktionen gilt : $\displaystyle f(x) = \sum_{k=1}^{n-1} f_k\,\mu_k(x)\,,\; x\in\omega_h$ $\downarrow$ Fourierkoeffizienten $\displaystyle f_k = (f,\mu_k)_h \equiv \sum_{x\in\omega_h} hf(x)\mu_k(x)$

7. Die Parseval'sche Gleichung

$\displaystyle\Vert f\Vert^2_{L_2(0,\ell)} := \int_0^\ell f(x)^2dx = \sum_{k=1}^{\infty} f_k^2$	$\displaystyle\Vert f\Vert^2_{L_2(\omega_h)} = \sum_{k=1}^{n-1} f_k^2$

Ü1 Man löse das diskrete EWP (1_h) bzw. man beweise, daß die angegebenen Eigenfunktionen (1.) und Eigenwerte (2.) richtig sind !

○ Hinweis : Ansatz $v(x)=c\cdot\sin(\alpha x)$ oder exp-Ansatz.

Ü2 Man beweise die Ungleichungen :

a) $\left(\dfrac{2k}{\ell}\right)^2 < \lambda_k < \left(\dfrac{\pi k}{\ell}\right)^2 \qquad \forall k=\overline{1,n-1}$

b) $\dfrac{8}{\ell^2} < \lambda_1 < \lambda_2 < \ldots < \lambda_k < \ldots < \lambda_{n-1} < \dfrac{4}{h^2}$!

$\boxed{\text{Ü3}}$ Man zeige, daß die Eigenfunktionen $\overline{\mu}_k = c \cdot \sin \dfrac{k\pi}{\ell} x$, $x \in \omega_h$, $k = \overline{1, n-1}$ für $c = \sqrt{2/\ell}$ bzgl. des Skalarproduktes $(\cdot, \cdot)_h$ orthonormal sind !

$\boxed{\text{Ü4}^*}$ Man zeige die Beziehungen 4. und 5. für die rückwärtigen Differenzen der diskreten Eigenfunktionen !

P II Praktikum II

5.2.1.2 Konsistenz- und Stabilitätsuntersuchungen

Betrachten das instationäre 1D-Wärmeleitproblem

$$\begin{array}{rrcll} \text{Ges. } u(x,t) : & \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} & = & f(x,t)\ , & x \in (0,1)\ , t \in \mathbf{T} = (0,T) \\ & + \text{ RB} : \ u(0,t) & = & g_0(t)\ , & t \in \mathbf{T} \\ & u(1,t) & = & g_1(t)\ , & t \in \mathbf{T} \\ & + \text{ AB} : \ u(x,0) & = & u_0(x)\ , & x \in [0,1] \end{array} \tag{5.2.2}$$

und approximieren es durch das Leapfrog-Schema (RICHARDSON-Schema) :

$$\begin{array}{l} \text{Ges. } v : \overline{\omega}_{h\tau} \longmapsto \mathbf{R}^1 : \\ \qquad v_{\overset{\circ}{t}} - v_{\overline{x}x} = \varphi(x,t)\ , \qquad (x,t) \in \overline{\omega}_{h\tau} \qquad \text{bzw.} \\ \qquad \dfrac{v_i^{j+1} - v_i^{j-1}}{2\tau} - \dfrac{v_{i-1}^j - 2v_i^j + v_{i+1}^j}{h^2} = \varphi_i^j\ , \qquad {\scriptstyle i=\overline{1,n-1}\ ,\ j=\overline{1,m-1}} \\ + \text{ RB} : \ v_0^j = g_0(t_j)\ , \qquad v_n^j = g_1(t_j) \qquad {\scriptstyle j=\overline{1,m}} \\ + \text{ AB} : \ v_i^0 = u_0(x_i) \qquad {\scriptstyle i=\overline{0,n}} \\ + \text{ geeignetes (?) Einschrittverfahren zur Bestimmung von } v_i^1,\ {\scriptstyle i=\overline{1,n}} \ !? \\ \text{mit } \varphi_i^j = f(x_i,t_j)\ . \end{array} \tag{5.2.3}$$

Ü5 Man untersuche die lokale Approximationsordnung (= Konsistenzordnung) des Leapfrog-Schemas (5.2.3) für (5.2.2) und gebe die führenden Terme des lokalen Approximationsfehlers (= lokaler Abschneidefehler = local truncation error) $\psi = .\,?\,.\, h^p + .\,?\,.\, \tau^q + 0(h^p + \tau^q) = O(h^p + \tau^q)$ an !
Welches Einschrittverfahren schlagen Sie zur Bestimmung von v_i^1 vor ?

Ü6 Untersuchen Sie die Stabilität (im V. NEUMANNschen Sinne) des Leapfrog-Schemas mit homogener rechter Seite !

○ Hinweis :

- Ansatz : $v_s^j = (e^{\alpha\tau})^j\, e^{i\lambda sh}$, $(i^2 = -1)$
  (Fortpflanzung der harmonischen Anfangsstörung)
- Stabilitätskriterium : $|e^{\alpha\tau}| \le 1$
- Allgemeines Stabilitätskriterium : $|e^{\alpha\tau}| \le 1 + c\tau$ mit $c = \text{const.} \ne c(\tau,h)$
  (exponentielles Wachstum ist zugelassen !)

$\boxed{\text{Ü7}}$ Untersuchen Sie das DU-FORT-FRANKEL-Schema (= Modifikation des Leapfrog-Schemas (5.2.3) mit $v_i^j \longrightarrow \frac{1}{2}(v_i^{j+1} + v_i^{j-1})$) :

$$\boxed{\begin{array}{l}
\text{Ges. } v : \overline{\omega}_{h\tau} \longmapsto \mathbf{R}^1 : \\[4pt]
\dfrac{v_i^{j+1} - v_i^{j-1}}{2\tau} - \dfrac{v_{i-1}^j - (v_i^{j+1} + v_i^{j-1}) + v_{i+1}^j}{h^2} = \varphi_i^j \, , \qquad {\scriptstyle i=\overline{1,n-1}\, ,\, j=\overline{1,m-1}} \\[6pt]
+ \text{ RB} : \; v_0^j = g_0(t_j) \quad , \qquad v_n^j = g_1(t_j) \qquad {\scriptstyle j=\overline{1,m}} \\
+ \text{ AB} : \; v_i^0 = u_0(x_i) \qquad\qquad\qquad\qquad\quad {\scriptstyle i=\overline{0,n}} \\
+ \text{ geeignetes (?) Einschrittverfahren zur Bestimmung von } v_i^1, \; {\scriptstyle i=\overline{1,n}} \;\; !? \\
\text{mit } \varphi_i^j = f(x_i, t_j) \, .
\end{array}} \tag{5.2.4}$$

zur Diskretisierung von (5.2.2) auf Konsistenz (lokaler Approximationsfehler) und Stabilität (im V. NEUMANNschen Sinne).

$\boxed{\textbf{P III}}$ Praktikum III

5.2.1.3 Numerische Experimente

$\boxed{\text{E1}}$ Die parabolische ARWA

$$\begin{array}{rcll} \text{Ges.:} \quad u(x,t) & \text{mit} & x \in [a,b]\,,\ t \in [0,T]\ : & \\ \dfrac{\partial u(x,t)}{\partial t} - \dfrac{\partial^2 u(x,t)}{\partial x^2} & = & f(x,t) \ := \ t^{s-1}\,(\,s(x-a)(x-b) - 2t\,) & (5.2.5) \\ +\ \text{AB:} \quad u(x,0) & = & x \qquad \forall x \in [a,b] & \\ +\ \text{RB:} \quad u(a,t) & = & a \qquad \forall t \in (0,T] & \\ u(b,t) & = & b \qquad \forall t \in (0,T] & \end{array}$$

besitzt für $s \geq 1$ die exakte Lösung

$$u(x,t) \ = \ x + (x-a)(x-b)\,t^s \quad . \tag{5.2.6}$$

Man diskretisiere (5.2.5) mit dem σ-gewichteten Differenzenschema auf einem gleichmäßigen Gitter der Form $\overline{\omega}_{h\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$, $\overline{\omega}_h = \{x_\imath = x_0 + \imath h \ : \ {\scriptstyle \imath = \overline{0,n}}\}$, $\overline{\omega}_\tau = \{t_\jmath = \jmath\tau \ : \ {\scriptstyle \jmath=\overline{0,m}}\}$ mit $x_0 = a$, $h = (b-a)/n$, $\tau = T/m$.
Implementieren Sie die Familie von Differenzenschemata und führen Sie Rechnungen für verschiedene

- Ausgangsdaten : $a = 0$, $b = 1$, $T = 1$, $s = 2$
- Gewichte : $\sigma = 1$, $\sigma = \frac{1}{2}$, $\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}$, $\sigma = 0$
- Diskretisierungen : 1) $h = \frac{1}{10}$, $\tau = \frac{1}{10}$
  2) $h = \frac{1}{10}$, $\tau = \frac{1}{200}$
  3) $h = \frac{1}{4}$, $\tau = \frac{1}{16}$

durch. Beachten Sie, daß die rechte Seite entsprechend der Parameterwahl diskretisiert werden sollte (Satz 2.10). Vergleichen Sie die diskrete Lösung mit der exakten Lösung (5.2.6) !

$\boxed{\text{E2}}$ Man diskretisiere die parabolische ARWA (vgl. $\boxed{\text{E1}}$)

$$\begin{array}{rcll} \text{Ges.:} \quad u(x,t) & \text{mit} & x \in [0,1]\,,\ t \in [0,1]\ : & \\ \dfrac{\partial u(x,t)}{\partial t} - \dfrac{\partial^2 u(x,t)}{\partial x^2} & = & t\,(\,2x(x-1) - 2t\,) & x \in (0,1)\ t \in (0,1] \\ +\ \text{AB:} \quad u(x,0) & = & x \qquad \forall x \in [0,1] & \\ +\ \text{RB:} \quad u(0,t) & = & 0 \qquad \forall t \in (0,T] & \\ u(1,t) & = & 1 \qquad \forall t \in (0,T] & \\ \text{Lösung :} \quad u(x,t) & = & x + x(x-1)\,t^2 & \end{array}$$

mit dem <u>DU-FORT-FRANKEL-Schema</u> auf einem gleichmäßigen Gitter der Form $\overline{\omega}_{h\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$, $\overline{\omega}_h = \{x_\imath = x_0 + \imath h \ : \ {\scriptstyle \imath=\overline{0,n}}\}$, $\overline{\omega}_\tau = \{t_\jmath = \jmath\tau \ : \ {\scriptstyle \jmath=\overline{0,m}}\}$ mit $x_0 = a$, $h = (b-a)/n$, $\tau = T/m$ (vgl. $\boxed{\text{Ü7}}$).
Wählen Sie ein geeignetes Einschrittverfahren im Startschritt. Implementieren Sie das Differenzenschema und führen Sie Testrechnungen für folgende Fälle durch :

1) $h = \frac{1}{10}$, $\tau = \frac{1}{10}$

2) Wählen Sie h und τ so, daß der Fehler in der Größenordnung 10^{-2} liegt bei möglichst geringen Aufwand !

$\boxed{\text{E3}}$ Die parabolische ARWA

$$
\begin{array}{rrcll}
 & \text{Ges.:}\quad u(x,t) & \text{mit} & x \in [0,1]\,,\ t \in [0,1]\ : & \\
5\,\dfrac{\partial u(x,t)}{\partial t} - & \dfrac{\partial^2 u(x,t)}{\partial x^2} & = & f(x,t) \ = \ 10t - x & \\
+\ \text{AB:} & u(x,0) & = & x^3 & \forall x \in [0,1] \\
+\ \text{RB:} & u(0,t) & = & t^2 & \forall t \in (0,1] \\
 & u(1,t) & = & t^2 + t + 1 & \forall t \in (0,1]
\end{array}
\tag{5.2.7}
$$

besitzt die exakte Lösung

$$
u(x,t) \ = \ x^3 + xt + t^2 \quad . \tag{5.2.8}
$$

Man diskretisiere (5.2.7) mit dem σ-gewichteten Differenzenschema auf einem gleichmäßigen Gitter der Form $\overline{\omega}_{h\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$, $\overline{\omega}_h = \{x_\imath = x_0 + \imath h \ : \ {\scriptstyle \imath=\overline{0,n}}\}$, $\overline{\omega}_\tau = \{t_\jmath = \jmath\tau \ : \ {\scriptstyle \jmath=\overline{0,m}}\}$ mit $x_0 = a$, $h = (b-a)/n$, $\tau = T/m$.
Wie sieht die L_2-Stabilitätsbedingung aus ?
Implementieren Sie die Familie von Differenzenschemata und führen Sie Rechnungen für verschiedene

– Gewichte :	$\sigma = 1$, $\sigma = \frac{1}{2}$, $\sigma = \sigma_* = \frac{1}{2} - ?$, $\sigma = 0$	
– Diskretisierungen :	1)	$h = \frac{1}{10}$, $\tau = \frac{1}{10}$
	2)	Fehler 10^{-2} für $\sigma = 0$, $h = \ldots$, $\tau = \ldots$
	3)	$\ldots\ldots$

durch. Beachten Sie, daß die rechte Seite entsprechend der Parameterwahl diskretisiert werden sollte (Satz 2.10). Vergleichen Sie die diskrete Lösung mit der exakten Lösung (5.2.8) !

$\boxed{\text{E4}}$ Randbedingungen der Form $\quad \dfrac{\partial u}{\partial x}(0,t) \ = \ f(t)$

können bei der numerischen Behandlung der homogenen Wärmeleitgleichung

$$
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} \ = \ 0
$$

durch die Approximation

$$
\begin{array}{rcll}
u(0,t+\tau) & \approx & u(0,t) + \dfrac{2\tau}{h^2}\left[u(h,t) - u(0,t) - h\cdot f(t)\right] & ,\text{d.h.} \\
v_0^{\jmath+1} & = & v_0^\jmath + \dfrac{2\tau}{h^2}\left[v_1^\jmath - v_0^\jmath - h\cdot f(t_\jmath)\right] & {\scriptstyle \jmath=0,1,\ldots,m-1}
\end{array}
$$

berücksichtigt werden (Warum ?).

Man diskutiere auf dieser Basis die ARWA

$$\boxed{\begin{array}{rrcll} \text{Ges. } u(x,t)\ : & \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} & = & 0\,, & \text{in } Q_T \ = \ \underbrace{(0,1)}_{\Omega} \times \underbrace{(0,1)}_{\mathbf{T}} \\ + \text{ RB}: & \dfrac{\partial u(0,t)}{\partial x} & = & 10^6\, t\,, & t \in \overline{\mathbf{T}} \\ & u(1,t) & = & 0\,, & t \in \overline{\mathbf{T}} \\ + \text{ AB}: & u(x,0) & = & 0\,, & x \in \overline{\Omega} \end{array}}$$

mit dem expliziten Euler-Verfahren ($\sigma = 0$) unter Verwendung der Schrittweiten

a) $h = 0.2$, $\tau = 0.01$

b) $h = 0.1$, $\tau = 0.01$

bis $T = 0.08$ $(m = 8)$. Vergleichen Sie die numerischen Ergebnisse !

$\boxed{\textbf{P IV}}$ Praktikum IV

5.2.1.4 Stabilität des σ-gewichteten Differenzenschemas in der Energienorm

$\boxed{\text{Ü8}}$ Man zeige, daß für beliebige Gitterfunktionen

$$u\,,\,v\ :\ \overline{\omega}_h\ =\ \{x_\imath = x_0 + \imath h\ :\ \imath = \overline{0,n}\,,\ h = \tfrac{b-a}{n}\} \mapsto \mathbf{R}^1$$

mit $x_0 = a$, $x_n = b$, $b > a$, die folgenden Formeln gelten :

a) Die diskrete Produktregel :

$$\boxed{(u\cdot v)_{x,\imath}\ =\ u_{x,\imath}v_{\imath+1}\ +\ u_\imath v_{x,\imath}} \tag{5.2.9}$$

b) Die Formel der partiellen Summation (= diskrete Formel der part. Integration) :

$$\boxed{(v_x, u)\ =\ -(v, u_{\overline{x}}]\ +\ u_n v_n\ -\ u_0 v_1} \tag{5.2.10}$$

mit $\quad (u,v)\ :=\ \sum\limits_{\imath=1}^{n-1} hv_\imath u_\imath \quad , \quad (u,v]\ :=\ \sum\limits_{\imath=1}^{n} hv_\imath u_\imath \quad .$

$\boxed{\text{Ü9}}$ Man zeige, daß für beliebige Gitterfunktionen $u\,,\,v\ :\ \overline{\omega}_h \mapsto \mathbf{R}^1$ mit $u_0 = u_n = 0$, $v_0 = v_n = 0$ die folgenden Formeln gelten :

a) Die diskrete Greensche Formel :

$$\boxed{-(u_{\overline{x}x}, v)\ =\ (u_{\overline{x}}, v_{\overline{x}}]} \qquad \forall u,v \in \overset{o}{W}{}^1_2(\omega_h) \tag{5.2.11}$$

d.h. für $u = v$ gilt also insbesondere

$$-(u_{\overline{x}x}, u)\ =\ \|\ u_{\overline{x}}\,\|]^2\ :=\ \sum_{\imath=1}^{n} hu^2_{\overline{x},i} \quad .$$

b) Die diskrete Friedrichs-Ungleichung :

$$\boxed{\|\ u\ \|\ \leq\ c_F\ \|\ u_{\overline{x}}\,\|]} \tag{5.2.12}$$

mit $\| u \|^2 = \sum_{\imath=1}^{n-1} h u_\imath^2 = (u,u)$ und $c_F = \dfrac{b-a}{\sqrt{2}}$.

<u>Hinweis :</u> $u_\imath = \sum_{\jmath=1}^{\imath} u_{\overline{x},\jmath} h = \sum_{\jmath=1}^{\imath} (u_\jmath - u_{\jmath-1}) = u_\imath - u_0 = u_\imath$

c) <u>Die diskrete C-Einbettung (1D-Fall !!) :</u>

$$\boxed{\| u \|_{C(\omega_h)} = \max_{\imath=\overline{1,n-1}} |u_\imath| \leq \sqrt{b-a}\, \| u_{\overline{x}} |]} \tag{5.2.13}$$

$\boxed{\text{Ü10}^*}$ Man gebe die optimale (kleinste !) Konstante c_F in der diskreten Friedrichsungleichung (5.2.12) an !

$\boxed{\text{Ü11}}$ Man zeige, daß das σ-gewichtete Differenzenschema (Fehlerschema)

$$z_{t,\imath}^{\jmath} - \sigma\, z_{\overline{x}x,\imath}^{\jmath+1} - (1-\sigma)\, z_{\overline{x}x,\imath}^{\jmath} = \psi_\imath^\jmath \quad , \quad {\scriptstyle \imath=\overline{1,n-1}\, , \jmath=\overline{0,m-1}} \tag{5.2.14}$$

RB : $\quad z_0^\jmath = z_n^\jmath = 0 \qquad \forall_{\jmath=\overline{1,m}}$

AB : $\quad z_\imath^0 = w_\imath^0 \qquad \forall_{\imath=\overline{0,n}} \qquad w^0$ (ist geg. Störung der AB)

im energetischen Raum H_A ($H = L_2^0(\overline{\omega}_h) = L_2(\omega_h)$, $\overline{\omega}_h = \{x_\imath = \imath h \,:\, {\scriptstyle \imath=\overline{1,n}}\}$ $A = \overline{A}$, $\overline{A}v := -v_{\overline{x}x}$) <u>stabil</u> bzgl. AB und RS ist und daß die a-priori-Abschätzungen

a) $\| z_{\overline{x}}^{\jmath+1} |] \leq \| z_{\overline{x}}^{0} |] + T \max_{k=\overline{0,\jmath}} \| \psi_{\overline{x}}^{k} |]$

b) $\| z^{\jmath+1} \|_{C(\omega_h)} := \max_{\imath=\overline{1,n-1}} |z_\imath^{\jmath+1}| \leq \| z_{\overline{x}}^{0} |] + T \max_{k=\overline{0,\jmath}} \| \psi_{\overline{x}}^{k} |]$

gelten, falls $1 \geq \sigma \geq \sigma_0 := \dfrac{1}{2} - \dfrac{h^2}{4\tau}$ gilt.

<u>Hinweis :</u> 1) Satz 2.16 aus der Vorlesung anwenden !

2) Beziehungen (5.2.11) und (5.2.13) benutzen !

3) Zeigen, daß

$$\| (I+\sigma\tau A)^{-1} \| \overset{!}{=} \text{Spektralnorm}\left([I+\sigma\tau A]^{-1}\right) \leq 1 \qquad !$$

P V Praktikum V

5.2.2 FEM-Galerkin-Verfahren für parabolische ARWA

Ü12 Man beweise den <u>Satz 2.8 von PICARD und LINDELÖF :</u>

<u>Vor.:</u> Betrachten AWA für folgendes System gewöhnlicher Differentialgleichungen

$$\left.\begin{aligned} \frac{d\underline{y}(t)}{dt} + A(t)\,\underline{y}(t) &= \underline{f}(t) \qquad \text{f.ü. } t \in \mathbf{T} = (0,T)\ , \\ \underline{y}(0) &= \underline{y}_0 \end{aligned}\right\} \tag{5.2.15}$$

mit gesuchter Vektorfunktion $\underline{y} = \underline{y}(t) = (y_1(t), \ldots, y_N(t))^T$,
gegebenen Anfangswerten $\underline{y}_0 = (y_{01}, \ldots, y_{0N})^T \in \mathbf{R}^N$,
und gegebenen Daten $A = A(t) = [a_{ki}(t)]_{k,i=\overline{1,N}}$: $a_{ki}(\cdot) \in L_\infty(0,T)$
sowie $\underline{f}(t) = (f_1(t), \ldots, f_N(t))^T \in [L_2(0,T)]^N$.

<u>Bh.:</u> $\exists!\ \underline{y}(t) \in \left[W_2^1(0,T)\right]^N$: (5.2.15)

<u>Hinweise :</u>

1) Übergang zur äquivalenten Igl.: $\int\limits_0^t (\ )\, d\xi$, d.h.

$$\underline{y}(t) = -\int\limits_0^t A(\xi)\,\underline{y}(\xi)\,d\xi + \int\limits_0^t \underline{f}(\xi)\,d\xi + \underline{y}_0$$

$$\underline{y} = B\,\underline{y} \qquad (= \text{Fixpunktgleichung})$$

mit $B : \mathcal{X} = \left[C(\overline{T})\right]^N \mapsto \left[W_2^1(0,T)\right]^N \subset \mathcal{X}$
($\mathcal{X}$ - Banachraum $\stackrel{?}{\Longrightarrow} \|\cdot\|_{\mathcal{X}}$).

2) Wenden Sie auf die Fixpunktgleichung

$$\text{Ges.: } \underline{y} \in \mathcal{X} \ : \quad \underline{y} = B\,\underline{y} \quad \text{in } \mathcal{X}$$

den verallgemeinerten Banachschen Fixpunktsatz I.2.5 an (vgl. auch Bsp. I.2.4) !

Ü13 Man zeige am Beispiel linearer Dreieckselemente, daß die Massenmatrix M_h gut konditioniert ist, falls die Triangulation im Sinne der Definition II.4.3 (vgl. Ü.II.4.4) regulär ist, d.h. $\kappa(M_h) = \lambda_{max}(M_h)/\lambda_{min}(M_h) = O(1)$!

<u>Hinweise :</u>

1) Definition der Massenmatrix :

$$(M_h\underline{u}_h, \underline{v}_h) = \int\limits_\Omega u_h(x)\,v_h(x)\,dx \qquad \forall\,\underline{u}_h, \underline{v}_h \in \mathbf{R}^{N_h}\ ,$$

$$\mathbf{R}^{N_h} \ni \underline{u}_h, \underline{v}_h \quad \longleftrightarrow \quad u_h, v_h \in \mathbf{V}_{0h} \quad \text{(FE-Raum)},$$

wobei $(\cdot,\cdot)$ das Euklidische Skalarprodukt im R^{N_h} ist.

2) Wdh. Def. II.4.3 (Reguläre Triangulation) :

Wir nennen die Familie $\{\tau_h\}_{h\in\Theta}$ der Triangulation $\tau_h = \{\delta_r : r \in \mathbf{R}_h\}$ regulär, falls positive, h-unabhängige Konstanten $\underline{c}_1, \overline{c}_1, c_2, c_3 = \text{const.} > 0$ existieren :

$$\begin{aligned} \underline{c}_1 h^d \;\le\; |J_{\delta_r}| &\;\le\; \overline{c}_1 h^d && \forall \xi \in \overline{\triangle}\ \forall r \in \mathbf{R}_h\ \forall h \in \Theta \\ \| J_{\delta_r} \| := \sqrt{\lambda_{max}(J_{\delta_r}^T J_{\delta_r})} &\;\le\; c_2 h && \forall \xi \in \overline{\triangle}\ \forall r \in \mathbf{R}_h\ \forall h \in \Theta \\ \| J_{\delta_r}^{-T} \| := \sqrt{\lambda_{max}(J_{\delta_r}^{-1} J_{\delta_r}^{-T})} &\;\le\; c_3 h^{-1} && \forall \xi \in \overline{\delta_r}\ \forall r \in \mathbf{R}_h\ \forall h \in \Theta\ , \end{aligned}$$

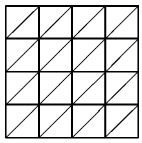
wobei $J_{\delta_r} = \dfrac{dx_{\delta_r}(\xi)}{d\xi}$, $J_{\delta_r}^{-1} = \dfrac{d\xi_{\delta_r}(x)}{dx}$, $\delta_r \stackrel{\xi=\xi_{\delta_r}(x)}{\longrightarrow} \triangle$ und $\delta_r \stackrel{x=x_{\delta_r}(\xi)}{\longleftarrow} \triangle$.

3) Zu zeigen ist also : $\exists \gamma_1, \gamma_2 = \text{const.} > 0 \;:\; \gamma_i \neq \gamma_i(h)$:

$$\gamma_1 h^d \le \lambda_{min}(M_h) \qquad \text{und} \qquad \lambda_{max}(M_h) \le \gamma_2 h^d\ .$$

Gehen Sie analog zum Beweis von Satz II.4.4 vor, in dem die Eigenwerte der Steifigkeitsmatrix K_h abgeschätzt werden.

4) Geben sie für das Beispiel der gleichmäßigen Triangularisierung des Einheitsquadrates γ_1 und γ_2 an.



Ü14 Man zeige, daß die Stabilitätsbedingung

$$B \;=\; M_h + \tau\,\sigma\,K_h \;\ge\; \frac{\tau}{2} A \;=\; \frac{\tau}{2} K_h$$

i.S. $(B\underline{u},\underline{u}) \ge \frac{\tau}{2}(A\underline{u},\underline{u}) \quad \forall \underline{u} \in \mathbf{R}^N$ für das Schema $(\underline{27})_{h\tau}$ aus der Vorlesung NuIII, Pkt. 2.3.2, äquivalent ist zur Bedingung

$$\sigma \;\ge\; \frac{1}{2} - \frac{1}{\tau\,\lambda_{max}} \quad ,$$

wobei λ_{max} = Maximaler Eigenwert des verallgemeinerten Eigenwertproblems $K_h\,\underline{u}_h = \lambda\, M_h\,\underline{u}_h$,

M_h = Massenmatrix,

K_h = Steifigkeitsmatrix ist.

Die Steifigkeitsmatrix wurde hier als t-unabhängig vorausgesetzt !

$\boxed{\textbf{P VI}}$ Praktikum VI

5.2.3 Konsultation zur Praktikumsaufgabe $\boxed{\text{Px}}$

5.2.3.1 Konsultation zu den parabolischen ARWA

- Theorie : siehe Vorlesung

5.2.3.2 Hyperbolische ARWA

$\boxed{\text{Ü15}}$ Geben Sie analog zu den parabolischen ARWA die Linienformulierung der hyperbolischen ARWA der nichtgedämpften Schwingungen einer fest eingespannten Membran

$$\boxed{\begin{array}{lrcl} \text{Ges. } u(x,t) : & \dfrac{\partial^2 u}{\partial t^2} - \triangle\, u(x,t) & = & f(x,t)\,, \quad x \in \Omega \subset \mathbf{R}^2\,, t \in \mathbf{T} = (0,T) \\ & + \text{RB}: \ u(0,t) & = & 0\,, \quad \forall x \in \partial\Omega\,; \forall t \in \mathbf{T} \\ & + \text{AB}: \ u(x,0) & = & u_0(x)\,, \quad x \in \overline{\Omega} \qquad \text{(Kompatibilität !)} \\ & \dfrac{\partial u}{\partial t}(x,0) & = & u_1(x)\,, \quad x \in \overline{\Omega} \quad . \end{array}}$$

an und leiten Sie mit der vertikalen Linienmethode (FE-Ortsdiskretisierung) die semidiskrete Ersatzaufgabe her !

$\boxed{\text{Ü16}^*}$ Betrachten Sie das System gewöhnlicher Differentialgleichungen 2. Ordnung

$$\boxed{\begin{array}{l} \text{Ges. } u(t) \;=\; [\,u_1(t),\ldots,u_N(t)\,]^T \in \left[C^2(I)\right]^N \quad : \\ \qquad\qquad M\,\ddot{u}(t) \,+\, C\,\dot{u}(t) \,+\, K\,u(t) \;=\; f(t) \quad , \quad t \in I = \overline{\mathbf{T}} = [0,T] \\ \quad \text{mit geg. AB}: \qquad\qquad u(0) \;=\; u_0 \quad , \quad \dot{u}(0) \;=\; u_1\,. \end{array}}$$

Hierbei sind $f \in [C(I)]^N$ die gegebene rechte Seite und

M, C, K	- $(N \times N)$ - Matrizen :
$M = M^T$ p.d.	- Massenmatrix,
C	- Dämpfungsmatrix,
$K = K^T$ p.d.	- Steifigkeitsmatrix.

Schreiben Sie die folgenden Zeitintegrationsschemata auf und motivieren Sie diese :

a) σ-gewichtetes dreischichtiges Differenzenschema (siehe Vorlesungsmanuskript)

b) NEWMARK-β-Methode (Kikuchi[1], S. 165-169),

c) WILSON-Θ-Methode (Kikuchi[1], S. 165-169),

d) Rückführung auf ein System 1. Ordnung und Anwendung von Integrationsverfahren für Systeme 1. Ordnung soweit bekannt.

$\boxed{\textbf{P VII}}$ Praktikum VII

5.3 Numerische Behandlung von AWA für gewöhnliche Dgl.

5.3.1 Einschrittverfahren

$\boxed{\text{Ü17}}$ Der Oregonator wird durch das Reaktionsschema

$$\boxed{\begin{array}{rcl} BrO_3^- + Br^- & \xrightarrow{K_1} & HBrO_2 \\ HBrO_2 + Br^- & \xrightarrow{K_2} & P \\ BrO_3^- + HBrO_2 & \xrightarrow{K_3} & 2HBrO_2 + Ce(IV) \\ 2HBrO_2 & \xrightarrow{K_4} & P \\ Ce(IV) & \xrightarrow{K_5} & Br^- \end{array}}$$

mit gegebenen Reaktionsgeschwindigkeitskoeffizienten K_1, K_2, K_3, K_4, K_5 beschrieben. Stellen Sie das Differentialgleichungssystem für die gesuchten Konzentrationen

$$\begin{array}{rcl} c_1(t) & = & c_{\mathrm{BrO}_3^-}(t) \\ c_2(t) & = & c_{\mathrm{Br}^-}(t) \\ c_3(t) & = & c_{\mathrm{HBrO}_2}(t) \\ c_4(t) & = & c_{\mathrm{P}}(t) \\ c_5(t) & = & c_{\mathrm{Ce(IV)}}(t) \end{array}$$

auf, und stellen Sie sinnvolle Anfangsbedingungen zum Zeitpunkt $t = t_0 = 0$!

○ Bemerkung :

Im Praktikum $\boxed{\text{P IX}}$ (18.12.1996) werden wir das abgeleitete Dgl.-System numerisch integrieren und damit den Oregonator numerisch simulieren.

$\boxed{\text{Ü18}}$ Die Robertson–Reaktion wird durch das Reaktionsschema

$$\boxed{\begin{array}{rclll} A & \xrightarrow{0.04} & B & & \text{(langsame Reaktion)} \\ B + B & \xrightarrow{3.10^7} & C + B & & \text{(sehr schnell)} \\ B + C & \xrightarrow{10^4} & A + C & & \text{(schnell)} \end{array}}$$

beschrieben. Stellen Sie das Dgl.-System zur Bestimmung der Konzentrationen $c_A(t), c_B(t), c_C(t)$ auf, und schreiben Sie geeignete Anfangsbedingungen zum Zeitpunkt $t = t_0 = 0$ vor!

○ Bemerkung :

Im Praktikum $\boxed{\text{P IX}}$ (18.12.1996) werden wir das abgeleitete Dgl.-System numerisch integrieren und damit die Robertson–Reaktion numerisch simulieren.

$\boxed{\text{Ü 19}}$ Man beweise die folgende Konvergenzaussage zum Eulerschen Polygonzugverfahren (EPZV) auf äquidistantem Gitter (Satz 3.4 aus der Vorlesung):

<u>Vor.:</u>

$$
\begin{array}{llrcl}
1.) & u(t): & u'(t) & = & f(t,u(t)),\ t \in I = [0,\ T] \\
 & & u(0) & = & g_0 \\
2.) & u_h: & I_h & = & \{t_j = jh\ :\ j = 1, \dots m\} \to \mathbb{R}^N : \\
 & & u_{j+1} & = & u_j + h\, f(t_j, u_j),\ j = 0,\ 1, \dots, m-1 \\
 & & u_h(0) & = & u_0 = g_0 + \delta \\
 & & u_h(t_j) & = & u_j,\ h = T/m.
\end{array}
$$

3.) $f : D \to \mathbb{R}^N$, wobei
$D = I \times \mathbb{R}^N = \{(t,v)\ :\ t \in I = [0,\ T],\ \mid v - u_0 \mid < \infty = b\}$

4.) $\mid f(t,v) - f(t,\bar{v}) \mid \leq L \mid v - \bar{v} \mid\ \forall\ (t,v),\ (t,\bar{v}) \in D.$

<u>Bh.:</u>

Dann gilt die Fehlerabschätzung

$$\mid u(t_j) - u_h(t_j) \mid \leq\ e^{Lt_j} \left[\mid \delta \mid + \frac{\tau}{L} \right],$$

mit $\tau = \max_{j=1,\dots,m} \mid \tau_j \mid$ und den lokalen Abschneidefehlern (siehe Vorlesung)

$$\tau_j \equiv \tau_h(t_j) = \frac{1}{h} \left[\frac{u(t_j) - u(t_{j-1})}{h} - f\left((t_{j-1}, u(t_{j-1})\right) \right]$$

○ <u>Hinweis :</u>

a.) Schreiben Sie unter Benutzung der Darstellung (*)

$$u(t_{j+1}) = u(t_j) + hf(t_j, u(t_j)) + h \underbrace{\left[\frac{u(t_{j+1}) - u(t_j)}{h} - f\left((t_j, u(t_j)\right) \right]}_{=:\tau_h(t_{j+1}) = \tau_{j+1}}$$

und des EPZV (**)

$$u_{j+1} = u_j + h\ f(t_j,\ u_j)$$

eine Rekursionsbeziehung ((*) - (**)!) für den Fehler

$$e_{j+1} = u(t_{j+1}) - u_h(t_{j+1}) = u(t_{j+1}) - u_{j+1}$$

auf, und schätzen Sie diese ab.

b.) Verwenden Sie dabei die elementare Beziehung

$$(1 + hL)^{j+1}\ \leq\ e^{(j+1)hL}\ =\ e^{Lt_{j+1}}\ \leq\ e^{LT}$$

c.) Interpretieren Sie die zu beweisende Behauptung !

$\boxed{\textbf{P VIII}}$ Praktikum VIII

$\boxed{\text{Ü20}}$ Verwendet man anstelle der Mittelpunktsregel die Trapezregel (TR) zur Berechnung des Integrals $\int\limits_t^{t+h} f(s,u(s))\,ds$ so erhält man das Verfahren von HEUN :

$$\int\limits_t^{t+h} f(s,u(s))\,ds \overset{\text{TR}}{\approx} \frac{h}{2}\left[f(t,u(t)) + f(t+h,u(t+h))\right]$$

$$\boxed{\begin{array}{rcl} K_1 & = & f(t,u) \\ K_2 & = & f(t+h, u + hK_1) \\ u_h(t+h) & = & u + \frac{h}{2}K_1 + \frac{h}{2}K_2 \end{array}}$$

$$\begin{array}{c} \Uparrow \\ \text{EPZV} \\ u(t+h) \approx u(t) + hf(t,u(t)) \end{array}$$

Damit ergibt sich für das Verfahren von HEUN das folgende Tableau :

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

d.h. das Verfahren von HEUN ist eine 2-stufige Runge-Kutta-Formel und es gilt :

$$u(t+h) = u(t) + \frac{h}{2}\left[f(t,u) + f(t+h, u+hf(t,u)\right] .$$

Man zeige durch Taylor-Entwicklung des lokalen Fehlers $u(t+h) - u_h(t+h)$, daß das Verfahren von HEUN die Konsistenzordnung 2 besitzt !

$\boxed{\text{Ü21}}$ Für die autonome Differentialgleichung

$$\begin{cases} v'(t) = g(v(t)) \\ v(0) = 0 \end{cases}$$

sei eine ℓ-stufige Runge-Kutta-Formel durch das Tableau (Koeffizienten $\{c_i\}_{i=\overline{2,\ell}}$ kommen im Falle autonomer Dgl. nicht vor !)

$$\begin{array}{|cccc} a_{21} & & & \\ a_{31} & a_{32} & & \\ \vdots & & & \\ a_{\ell 1} & a_{\ell 2} & \dots & a_{\ell,\ell-1} \\ \hline b_1 & b_2 & \dots & b_\ell \end{array} \qquad (*)$$

gegeben und habe die Ordnung p.

Für eine allgemeine Differentialgleichung der Form

$$\begin{cases} u'(t) = f(t,u(t)) \\ u(0) = u_0 \end{cases}$$

ist dann eine ℓ-stufige Runge-Kutta-Formel der Ordnung p durch das Tableau

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ c_3 & a_{31} & a_{32} & & \\ & \vdots & & & \\ c_\ell & a_{\ell 1} & a_{\ell 2} & \dots & a_{\ell,\ell-1} \\ \hline & b_1 & b_2 & \dots & b_\ell \end{array} \qquad (**)$$

gegeben, falls

$$c_i \;=\; \sum_{j=1}^{i-1} a_{ij} \quad , \quad i=\overline{2,\ell} \;. \qquad (5.3.16)$$

gilt. Zu zeigen ist also, daß die Runge-Kutta-Formel (*) für die autonome Dgl. dieselben Näherungen liefert wie (**) für eine allg. Dgl., falls (5.3.16) gilt.

$\boxed{\text{Ü22}}$ Man zeige, daß das klassische Runge-Kutta-Verfahren

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

die Konsistenzordnung 4 besitzt !

P IX Praktikum IX

5.3.2 Praktische Durchführung von Einschrittverfahren

$\boxed{\times}$ Man studiere den gleichnamigen Punkt 3.3.5 im Vorlesungsskriptum Numerik III.

5.3.3 Einfache numerische Experimente mit Einschrittverfahren

E5 Wir betrachten das Anfangswertproblem

$$\left.\begin{array}{rcl} u'(t) & = & B(u+A) \quad , \quad t \in I = [0,1] \\ u(0) & = & 0 \end{array}\right\} \tag{5.3.17}$$

mit $B = 3$ und $A = 1 \cdot (e^B - 1)$.

a) Man bestimme die exakte Lösung von (5.3.17) analytisch !

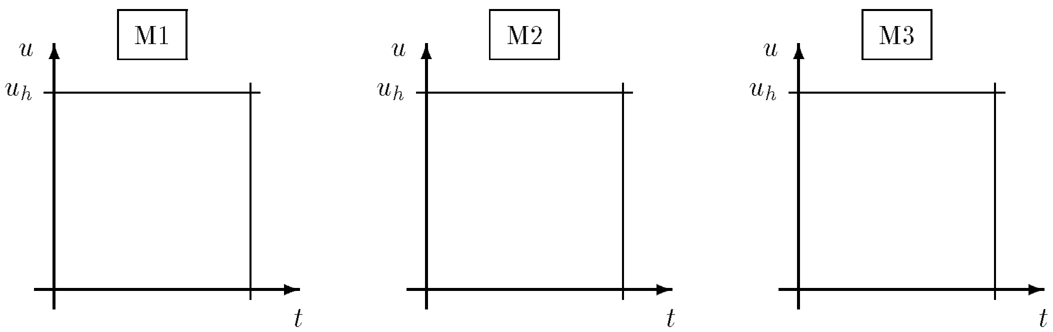
b) Man löse (5.3.17) mit den folgenden expliziten Runge-Kutta-Verfahren $\boxed{\text{Mx}}$ unter Verwendung der Schrittweiten $h = 1 \, , \, \frac{1}{2} \, , \, \frac{1}{4} \, , \, \frac{1}{8}$ und $\frac{1}{16}$;

M1 Explizites Euler-Verfahren,

M2 Verbessertes Euler-Verfahren,

M3 „klassisches“ Runge-Kutta-Verfahren der Ordnung 4 (siehe $\boxed{\text{Ü22}}$).

Man veranschauliche die Ergebnisse graphisch durch Vergleich der numerischen Lösungen für $h = 1 \, , \, \frac{1}{2} \, , \, \frac{1}{4} \, , \, \frac{1}{8} \, , \, \frac{1}{16}$ mit der analytischen Lösung :



Expliziter Euler — Verbesserter Euler — Klassisches Runge-Kutta

c) Man trage in die folgende Tabelle $\boxed{u_h(1)}$ als Approximation von $u(1) = \ldots$ für die betrachteten Verfahren $\boxed{\text{Mx}}$ ein :

Mx \ h $\longrightarrow$		1	1/2	1/4	1/8	1/16	$u(1)$
M1	(*)						
	(E)					–	
M2	(*)						
	(E)					–	
M3	(*)						
	(E)					–	
MxE							

wobei (*) = ursprüngliches Verfahren

(E) = Verbesserung der Werte durch globale Extrapolation (siehe S. 115/116 im Skriptum) nach der Formel

$$\widehat{u}_h(t) \;=\; u_{h/2}(t) \;+\; \frac{u_{h/2}(t) \;-\; u_h(t)}{2^p - 1}$$

mit p =Ordnung des Verfahrens.

(MxE) = Verfahren Mx für ein x$\in \{1, 2, 3\}$ mit lokaler Extrapolation (siehe S. 115/116 im Skriptum).

$\boxed{\text{E6}}$ Wir betrachten das Anfangswertproblem

$$\left.\begin{array}{rcl} u'(t) & = & -\,50\,(u(t) \,-\, \cos(t)) \quad , \quad t \in I \,=\, [0, 1.5] \\ u(0) & = & 0 \quad . \end{array}\right\} \qquad (5.3.18)$$

a) Man bestimme die exakte Lösung von (5.3.18) analytisch !

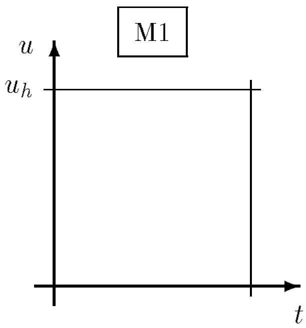
b) Man löse (5.3.18) mit den folgenden Runge-Kutta-Verfahren $\boxed{\text{Mx}}$ unter Verwendung der Schrittweiten $h \,=\, \frac{1}{20}\,,\; \frac{3}{80}\,,\; \frac{1}{30}\,,\; \frac{1}{40}\,;$

$\boxed{\text{M1}}$ Explizites Euler-Verfahren,

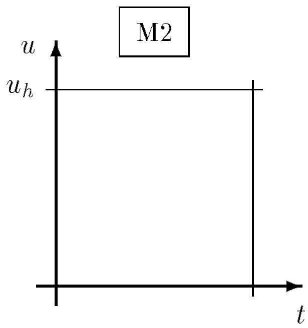
$\boxed{\text{M2}}$ Verbessertes Euler-Verfahren,

$\boxed{\text{M3}}$ Implizites Euler-Verfahren.

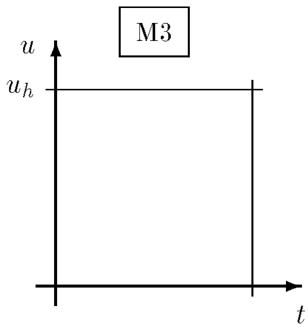
Man veranschauliche die Ergebnisse graphisch durch Vergleich der numerischen Lösungen für $h \,=\, \frac{1}{20}\,,\; \frac{3}{80}\,,\; \frac{1}{30}\,,\; \frac{1}{40}$ mit der analytischen Lösung :



Expliziter Euler



Verbesserter Euler



Impliziter Euler

c) Man starte die Explizite Eulermethode bei $t_0 = 1/2$ mit der exakten Lösung $u(t_0) \equiv u(1/2)$ unter Verwendung der Schrittweite $h = 1/20$ und stelle das Resultat wieder im Vergleich mit der analytischen Lösung graphisch dar.
Ab welcher Schrittweite muß das Verfahren stabil werden ?

d) Lösen Sie (5.3.18) nochmals mit dem impliziten Euler-Verfahren und der Schrittweite $h = 1/2$. Stellen Sie das Resultat wieder im Vergleich mit der analytischen Lösung graphisch dar.

e) Man trage in die folgende Tabelle $\boxed{u_h(1.5)}$ als Approximation von $u(1.5) = \ldots$ für die betrachteten Verfahren $\boxed{\text{Mx}}$ ein :

h Mx \	$\longrightarrow$ 1/20	3/80	1/30	1/40	$u(1.5)$
M1					
M2					
M3					

P X Praktikum X

5.3.4 Numerische Lösung komplizierter Anfangswertprobleme

E7 Die Bahn eines Satelliten in der Ebene des Erde-Mond-Systems läßt sich durch das folgende System von Dgl. 2. Ordnung beschreiben (vgl. Bsp. 3.5 der Vorlesung Numerik III und [2, 3]) :

$$\boxed{\begin{array}{lrcl} & y_1'' &=& y_1 + 2y_2' - (1-\mu)\,\dfrac{y_1+\mu}{D_1} - \mu\,\dfrac{y_1-(1-\mu)}{D_2} \quad , \\ & y_2'' &=& y_2 - 2y_1' - (1-\mu)\,\dfrac{y_2}{D_1} - \mu\,\dfrac{y_2}{D_2} \quad , t \in [0,T]\,, \\ +\text{ AB :} & y_1(0) &=& 0.994 \\ & y_1'(0) &=& 0 \\ & y_2(0) &=& 0 \\ & y_2'(0) &=& -2.001\,585\,106\,379\,082\,522\,405\,378\,622\,24 \quad , \\ \text{mit} & D_1 &=& \left[(y_1+\mu)^2 + y_2^2\right]^{3/2} \quad , \\ & D_2 &=& \left[(y_1-(1-\mu))^2 + y_2^2\right]^{3/2} \quad , \\ & \mu &=& 0.012277471 \end{array}} \tag{5.3.19}$$

Für diese Daten ergibt sich eine periodische Lösung mit der Periode

$$t_{per} = t = \underline{17.065\,216\,560\,157\,96}2\,558\,891\,720\,6249 \quad .$$

1) Man überführe (5.3.19) in eine äquivalente AWA für ein System gewöhnlicher Dgl. 1. Ordnung ! Ist dieses System autonom ?

2) Man löse dieses System numerisch mit den folgenden Integrationsverfahren :

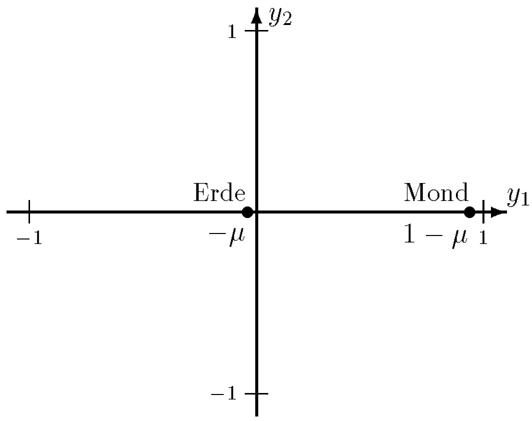
M1 Klassisches Runge-Kutta-Verfahren mit dem Tableau

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

mit $h = T/m$, $m = 6000$ und $m =?$ (eigene Wahl).

M2 Verfahren eigener Wahl !

3) Man stelle die Lösungstrajektoren $(y_{1h}(t), y_{2h}(t))$, $t \in I_h = \{t_0, t_1, \ldots, t_m\}$, linear interpoliert, grafisch dar !



E8 Die Robertson–Reaktion [4] wird durch das Dgl.-System

$$
\boxed{\begin{array}{rcl}
c_A'(t) & = & -\,0.04\,c_A(t) \,+\, 10^4\,c_B(t)c_C(t) \quad , \\
c_B'(t) & = & 0.04\,c_A(t) \,-\, 3\cdot 10^7\,c_B^2(t) \,-\, 10^4\,c_B(t)c_C(t) \quad , \\
c_C'(t) & = & 3\cdot 10^7\,c_B^2(t) \qquad ,t\in[0,T]\,,\ T=0.3 \\
+\,\mathrm{AB}: & & c_A(0) \;=\; 1 \qquad c_B(0) \;=\; 0 \qquad c_C(0) \;=\; 0
\end{array}}
\tag{5.3.20}
$$

beschrieben (vgl. Ü18). Aufgrund der Größenordnungsunterschiede in den Reaktionsgeschwindigkeitskoeffizienten ist zu erwarten, daß das Dgl.-System (5.3.20) steif ist. Lösen Sie die AWA (5.3.20) mit einem geeigneten Integrationsverfahren und stellen Sie den zeitlichen Verlauf der Konzentrationen $c_A(t)$, $c_B(t)$ und $c_C(t)$ (getrennt) grafisch dar !

E9 Der Oregonator [2] wird durch das Dgl.-System

$$
\boxed{\begin{array}{rcl}
c_1'(t) & = & -\,k_1c_1c_2 \,-\, k_3c_1c_3 \\
c_2'(t) & = & -\,k_1c_1c_2 \,-\, k_2c_2c_3 \,+\, k_5c_5 \\
c_3'(t) & = & k_1c_1c_2 \,-\, k_2c_2c_3 \,+\, k_3c_1c_3 \,-\, 2k_4c_3^2 \\
c_4'(t) & = & k_2c_2c_3 \,+\, 2k_4c_3^2 \\
c_5'(t) & = & k_2c_2c_3 \,-\, k_5c_5 \qquad ,t\in[0,T]\,, \\
+\,\mathrm{AB}: & & c_1(0) \;=\; 1/2 \qquad c_2(0) \;=\; 1/2 \qquad c_3(0) \;=\; 0 \\
 & & c_4(0) \;=\; 0 \qquad c_5(0) \;=\; 0\,, \\
\text{wobei} & & k_1 \;=\; 1.34 \qquad k_2 \;=\; 1.6\cdot 10^9 \qquad k_3 \;=\; 8.0\cdot 10^3 \\
 & & k_4 \;=\; 4.0\cdot 10^7 \qquad k_5 \;=\; 1.0
\end{array}}
\tag{5.3.21}
$$

beschrieben (vgl. Ü17). Aufgrund der Größenordnungsunterschiede in den Reaktionsgeschwindigkeitskoeffizienten ist zu erwarten, daß das Dgl.-System (5.3.21) steif ist.

Lösen Sie die AWA (5.3.21) mit einem geeigneten (?) Integrationsverfahren und stellen Sie den zeitlichen Verlauf der Konzentrationen $c_1(t)$, $c_2(t)$, $c_3(t)$, $c_4(t)$ und $c_5(t)$ (getrennt) grafisch dar !

Literaturverzeichnis

[1] N. Kikuchi. *Finite Element Method in Mechanics.* Cambridge University Press, 1986.

[2] P. Deuflhard. *Numerische Mathematik II: Integration gewöhnlicher Differentialgleichungen.* de Gruyter, Berlin, 1994. ISBN 3-11-013977-5.

[3] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer, 1993.

[4] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Equations*, volume 14 of *Springer Series in Computational Mathematics.* Springer, 1996. ISBN 3-540-60452-9.

Kapitel 6

Ergebnisse der numerischen Experimente

Für die Überlassung der folgenden Ergebnisgraphiken danken wir :

Gerald Fehringer, Franz Gruber, Ferdinand Kickinger, Roswitha Kroiss, Joachim Schöberl