Numerik 1 – O. Ippisch / P. Bastian

11. Oktober 2013

Inhaltsverzeichnis

Lizenz und Copyright

Verbesserungen

Dieser *Mitschrieb* ist sicher nicht völlig fehlerfrei. Es handelt sich nicht um ein Skript sondern einen studentischen Mitschrieb, der Korrektur gelesen wurde. Verbesserungshinweise bitte an

peter.bastian@iwr.uni-heidelberg.de

Skript

Als Begleitung zur Vorlesung empfehle ich das Vorlesungsskript “Numerische Mathematik 1 / Numerik gewöhnlicher Differentialgleichungen” von Rolf Rannacher. Zu finden unter http://numerik.iwr.uni-heidelberg.de/~lehre/notes/

1 Motivation

1.1 Wachstumsmodelle

$y(t)\colon [a,b] \to \mathbb{R}$ „Zahl der Individuen der Population“
Annahmen:

- Anzahl ist kontinuierlich
- Räumliche Verteilung wird vernachlässigt
- Zunahme der Individuen im Zeitintervall Δt ist proportional zu Δt und zur Anzahl der Individuen

Also:

$$y(t+\Delta t) = y(t) + \lambda \Delta t y(t)$$
$$\Leftrightarrow \frac{y(t+\Delta t) - y(t)}{\Delta t} = \lambda y(t)$$
$$\lim \Delta t \to 0 \quad \boxed{\frac{dy(t)}{dt} = \lambda y(t)} \tag{1.1}$$

Gewöhnliche Differentialgleichung (GDGL).
Lösung:

$$\lambda \in \mathbb{R}: \frac{\mathrm{d}}{\mathrm{d}t}(\underbrace{e^{\lambda t}}_{y(t)}) = \lambda \underbrace{e^{\lambda t}}_{y(t)} \to e^{\lambda t} \text{ ist Lösung von (1.1)}$$

Für beliebige Lösung $y(t)$ von 1.1 gilt:

$$\frac{\mathrm{d}}{\mathrm{d}t}(y(t)e^{-\lambda t}) = \frac{\mathrm{d}y(t)}{\mathrm{d}t} \cdot e^{-\lambda t} - \lambda y(t) e^{-\lambda t} = \left(\underbrace{\frac{\mathrm{d}y(t)}{\mathrm{d}t} - \lambda y(t)}_{=0 \text{ da y Lsg von 1.1}} \right) e^{-\lambda t} = 0$$

$\Rightarrow y(t)e^{-\lambda t} = C \in \mathbb{R}$
$\Rightarrow \boxed{y(t) = Ce^{\lambda t}}$ Form aller Lösungen von 1.1
$\Rightarrow$ Festlegen der Konstante C durch eine Anfangsbedingung

Anfangswertaufgabe (AWA)

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \lambda y(t) \qquad t \in (a,b] \tag{1.2a}$$
$$y(a) = Y \tag{1.2b}$$

$$Y = y(a) = Ce^{\lambda a} \Rightarrow C = Ye^{-\lambda a}$$

also

$$y(t) = Ce^{\lambda t} = \boxed{Ye^{\lambda(t-a)}}$$

Logistisches Wachstumsmodell:

$y(t) \in [0, 1]$ (1 = Maximalpopulation)

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \lambda \cdot (1 - y(t)) \cdot y(t)$$

→ nichtlineare GDGL

Einfache numerische Lösungsverfahren

$$t_i = a + \underbrace{\frac{b-a}{N}}_{h} i = a + hi$$

Expliziter Euler:

$$\begin{aligned} y_0 &= Y & i &= 0 \\ y_i^h &= y_{i-1}^h + h\lambda y_{i-1}^h & i &= 1, \ldots, N \end{aligned}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t}(t_{i-1}) \approx \frac{y_i^h - y_{i-1}^h}{h} = \lambda y_{i-1}^h$$

Impliziter Euler:

Ansatz:

$$\frac{y_i^h - y_{i-1}^h}{h} = \lambda y_i^h$$

$\Leftrightarrow (1 - h\lambda) y_i^h = y_{i-1}^h$
$\Leftrightarrow y_i^h = \frac{1}{1-h\lambda} y_{i-1}^h \qquad i = 1, \ldots, N$
$\Rightarrow$ im allgemeinen deutlich höherer Aufwand. Frage: Lohnt das?

1.2 Chemische Reaktionen

A und B reagieren zu C ($A + B \to C$).
Sei $c_A(t)$ die Konzentration (z.B. Mol pro Volumen) von A zur Zeit t. Analog: $c_B(t), c_C(t)$.

$$c_A(t + \Delta t) = c_A(t) - k\Delta t \underbrace{c_A(t)c_B(t)}_{\substack{\text{beide Stoffe} \\ \text{müssen vorliegen}}}$$

$$\Leftrightarrow \quad \begin{aligned} \frac{\mathrm{d}c_A(t)}{\mathrm{d}t} &= -kc_A(t)c_B(t) \\ \frac{\mathrm{d}c_B(t)}{\mathrm{d}t} &= -kc_A(t)c_B(t) \\ \frac{\mathrm{d}c_C(t)}{\mathrm{d}t} &= +kc_A(t)c_B(t) \end{aligned}$$

System von GDGL + Anfangsbedingungen

$$\textit{komplexere Reaktion: } aA + bB \underset{k_2}{\overset{k_1}{\rightleftharpoons}} cC + dD$$

$c_x(t)$ für $x \in \{A, B, C, D\}$

$$R(t) = -k_1(c_A(t))^a(c_B(t))^b + k_2(c_C(t))^c(c_D(t))^d$$

$$\frac{\mathrm{d}c_A(t)}{\mathrm{d}t} = aR(t) \qquad \frac{\mathrm{d}c_B(t)}{\mathrm{d}t} = bR(t)$$
$$\frac{\mathrm{d}c_C(t)}{\mathrm{d}t} = -cR(t) \qquad \frac{\mathrm{d}c_D(t)}{\mathrm{d}t} = -dR(t)$$

chemisches Gleichgewicht:

$$\frac{\mathrm{d}c_x(t)}{\mathrm{d}t} = 0 \Leftrightarrow R(t) = 0 \Leftrightarrow \boxed{\frac{k_2}{k_1} = \frac{c_A^a c_B^b}{c_C^c c_D^d}} = K_{\text{eq}} \text{ „Massenwirkungsgesetz“}$$

1.3 Neurowissenschaft

Grafik zu einer Nervenzelle, wie z.B. die hier: http://de.wikipedia.org/w/index.php?title=Datei:Impulsfortleitung_an_der_Nervenzelle.png&filetimestamp=20060120204945. Weiteres Bild zu einem Axonabschnitt.

Unbekannte:

$v(t)$ Spannungsdifferenz zwischen innen und außen. $\underbrace{m(t)}_{\text{Natrium aktivierend}}$, $\underbrace{h(t)}_{\text{Natrium inhibierend}}$, $\underbrace{n(t)}_{\text{Kalium aktivierend}}$ ist der Anteil der Gating-Partikel die in einem Zustand sind, der die relative Leitfähigkeit erhöht (beides zwischen 0 und 1).

$$\underbrace{C_M}_{\substack{\text{Membran-}\\ \text{kapazität}[\frac{As}{Vm^2}]}} \frac{\mathrm{d}v}{\mathrm{d}t} = \underbrace{I_{\text{ext}}(v,t)}_{\substack{\text{Strominjektion}\\ \text{Synapse}[Am^{-2}]}} - g_{\text{Na}} m^3 h(v - v_{\text{Na}}) - g_K n^4 (v - v_K) - g_L(v - v_L)$$

(g_k und v_k ($k \in \{\text{Na}, \text{K}, \text{L}\}$) sind Konstanten)

$$\frac{\mathrm{d}k}{\mathrm{d}t} = \alpha_k(v)(1-k) - \beta_k(v)k \qquad k \in \{n, m, h\}$$

$\Rightarrow$ Nobelpreis 1963 Hodgin/Huxley
Betrachtung der räumlichen Ausbreitung

$$\underbrace{c_M}_{\substack{\text{spez. Membran-}\\ \text{kapazität}[\frac{As}{Vm}]}} \frac{\partial v(x,t)}{\partial t} = \frac{\partial}{\partial x}\Big(\underbrace{\varrho_m}_{\substack{\text{spez. Membran-}\\ \text{leitfähigkeit}[\Omega^{-1}m]}} \frac{\partial v}{\partial x} \Big) + i_{\text{ext}}(x,t,v(x,t)) - i_{\text{Na}}(\ldots) - i_K(\ldots) - i_L(\ldots)$$

$\rightarrow$ Erstes Beispiel für partielle DGL. Zusätzlich sind hier *Randbedingungen* anzugeben, z.B. v oder $\varrho_m \frac{\partial v}{\partial x}$ am Anfang/Ende des Axioms.
Zeichnung zu einer Nervenzelle. x_L und x_R bezeichnen die Ränder des Axons in „Stromrichtung“.

$$c_m \frac{\partial v(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(\varrho_m \frac{\partial v(x,t)}{\partial x}\right) + i(v,t) \tag{1.7}$$

$v(x,t_0) = v_0(x)$	Anfangsbedingung
$v(x_L,t) = v_0(t)$ und $v(x_R,t) = v_R(t)$	„Randbedingung“

Stationäre Lösung ($\frac{\partial v}{\partial t} = 0$)

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\varrho_m \frac{\mathrm{d}v(x)}{\mathrm{d}x}\right) + i(v) = 0$$

$$v(x_L) = v_L$$
$$v(x_R) = v_R$$

„Randwertproblem“ 2. Ordnung
→ erster Teil der Vorlesung *nur AWA*.

$$\varrho_m \frac{\mathrm{d}v(x)}{\mathrm{d}x} = u(x)$$
$$\frac{\mathrm{d}u(x)}{\mathrm{d}x} = -i\,(v(x))$$

⇒ Reduktion GDGL höherer Ordnung ⇒ System erster Ordnung

1.4 Astrophysikalisches N-Körper-Problem

N Punktmassen der Masse m_i
Bewegung der Körper unter ihrem eigenen Schwerefeld.

$$\to x_i(t)\colon [t_0,t_1] \to \mathbb{R}^3 \qquad v_i(t)\colon [t_0,t_1] \to \mathbb{R}^3$$

Grafik: $0 \overset{x_i(t)}{\to} m_i$, $0 \overset{x_j(t)}{\to} m_j$, und $m_i \overset{x_j - x_i}{\to} m_j$ (als Dreieck)

$$F_{ij}(x_i,x_j) = G \underbrace{\frac{m_i m_j}{\|x_j - x_i\|^2}}_{\text{eukl. Norm}} \underbrace{\frac{x_j - x_i}{\|x_j - x_i\|}}_{\text{Richtung}} \tag{1.11}$$

2. Newtonsches Gesetz ($F = \frac{\mathrm{d}p}{\mathrm{d}t} = ma$)

$$m_i a_i(t) = \sum_{\substack{j=1 \\ j\neq i}}^{N} F_{ij}(x_i(t), x_j(t))$$

Mit $a_i(t) = \frac{\mathrm{d}v_i(t)}{\mathrm{d}t}$ und $\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = v_i(t)$ erhält man:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = v_i(t) \qquad x_i(t_0) = x_i^0 \tag{1.12a}$$

$$\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = G \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{m_j(x_j(t) - x_i(t))}{\|x_j - x_i\|^3} \qquad v_i(t_0) = v_i^0 \tag{1.12b}$$

$\Rightarrow 6N$ gew. DGL

Potential:

$x, y \in \mathbb{R}^3$
Setze

$$\varphi(x,y) = -\frac{1}{\|y-x\|} = -\left(\sum_{k=1}^{3}(y_k - x_k)^2\right)^{-\frac{1}{2}}$$

$$\frac{\partial}{\partial y_\ell}\varphi(x,y) = \ldots = \frac{y_\ell - x_\ell}{\|y-x\|^3} \Rightarrow \boxed{\nabla_y \varphi(x,y) = \frac{y-x}{\|y-x\|^3}}$$

also:

$$\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = G \sum_{\substack{j=1 \\ j\neq i}}^{N} m_j \nabla_y \varphi(x_i(t), x_j(t))$$

Dieses System ist:

- nicht dissipativ
- konservativ

Es gibt:

- potentielle Energie:

$$E_{\mathrm{pot}}(t) = G \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} m_i m_j \varphi(x_i(t), x_j(t))$$

- kinetische Energie:

$$E_{\mathrm{kin}} = \frac{1}{2} \sum_{i=1}^{N} m_i \|v_i(t)\|^2$$

$$\Rightarrow E_{\mathrm{tot}} = E_{\mathrm{pot}}(t) + E_{\mathrm{kin}}(t) \text{ bleibt erhalten}$$

Energieerhaltung im numerischen Verfahren erfordert spezielle Methoden.

1.5 Raketengleichung

Bewegung eines Raumfahrzeuges

$m(t)$ Masse der Rakete zur Zeit t
$x(t)$ Position
$v(t)$ Geschwindigkeit

2. Newton: ($p = m(t)v(t)$: Impuls)

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = \underbrace{F_G(t)}_{\text{Gewichtskraft}} + \underbrace{F_T(t)}_{\text{Antrieb}}$$

$$F_T(t) = \underbrace{-}_{\text{actio = reactio}} \underbrace{c(t)}_{\substack{\text{Treibstoffaus-}\\ \text{stossrate}\left[\frac{kg}{s}\right]}} \cdot \underbrace{w(t)}_{\substack{\text{Geschwindigkeit der}\\ \text{ausgestoßenen Gase}(\in\mathbb{R}^3)}}$$

ergibt zusammen das System:

$$\frac{\mathrm{d}(m(t)v(t))}{\mathrm{d}t} = F_G(t) - c(t)w(t) \qquad \text{(rechte Seite ist gegeben!)}$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v(t) \qquad \text{NB: } c(t) \geq 0 \qquad \int_{t_0}^{t_1} c(t)\,\mathrm{d}t \leq m(t_0)$$

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = -c(t) \qquad \rightarrow m(t) = m(t_0) - \int_{t_0}^{t} \underbrace{c(s)}_{\text{(Ausgestossene Masse)}}\,\mathrm{d}s$$

$\Rightarrow$ Optimierungsproblem

- Wähle $c(t), w(t)$ so dass ein Punkt im Raum mit möglichst wenig Ressourcen erreicht wird
- $c(t), w(t)$ unterliegen technischen Beschränkungen, was die Probleme numerisch sehr schwer macht, z.B. weil $c(t)$ nur stückweise glatt
- Liefert i.d.R. differentiell-algebratische Systeme.

2 Zur Theorie von gewöhnlichen Differentialgleichungen

2.1 Problemstellung

$$u'(t) = f(t, u(t)) \tag{2.1}$$

mit

$$u(t) = (u_1(t), \ldots, u_d(t)) \qquad f(t, x) = (f_1(t, x), \ldots, f_d(t, x))$$

Die Funktion f sei auf $D = I \times \Omega \subset \mathbb{R}^1 \times \mathbb{R}^d$. definiert und dort stetig.
Standardnotation:

$$(x, y) = \sum_{i=1}^{d} x_i y_i \qquad \|x\| = (x, x)^{1/2} \quad \forall\, x, y \in \mathbb{R}^d$$

$$\|A\| = \sup\left\{\|Ax\| \;\middle|\; x \in \mathbb{R}^d, \;\; \|x\| = 1\right\} \text{ (Spektralnorm)}$$

$$u'(t) = \frac{\mathrm{d}u(t)}{\mathrm{d}t} \qquad f'_t(t,x) = \frac{\partial f(t,x)}{\partial t} \qquad \partial_i f(t,x) = \frac{\partial f(t,x)}{\partial x_i}$$

$$\nabla_x f(t,x) = \left(\frac{\partial f_i}{\partial x_j}\right)_{i,j=1}^{d} \qquad \text{Jacobi-Matrix}$$

Implizite Formen

(2.1) ist Spezialfall der allgemeinen, impliziten Form

$$F(t, u(t), u'(t)) = 0 \tag{2.2}$$

Satz von der impliziten Funktion: $F(t,x,y)$ ist lokal nach y auflösbar, falls $\nabla_y F = \left(\frac{\partial F_i}{\partial y_j}\right)_{i,j=1}^{d}$ invertierbar ist.

Beispiel: Netzwerkanalyse führt immer auf DGL-System der Form

$$\begin{aligned} A_{11}u_1'(t) + B_{11}u_1(t) + B_{12}u_2(t) &= g_1(t) \\ B_{21}u_1(t) + B_{22}u_2(t) &= g_2(t) \end{aligned}$$

mit

$$u(t) = (u_1(t), u_2(t))^T \qquad u_1(t) \in \mathbb{R}^n \qquad u_2(t) \in \mathbb{R}^m \qquad m + n = d$$

Ist nicht nach $u(t)$ auflösbar, da es sich im ein DAE-System handelt.
In der Praxis kommt oft folgende Form vor:

$$\frac{\mathrm{d}m(t, u(t))}{\mathrm{d}t} = f(t, u(t)) \tag{2.3}$$

auch hier muss $m(t,x)$ nicht nach x auflösbar sein.

Reduktion von Gleichungen höherer Ordnung

Eine DGL höherer Ordnung der Form

$$F\left(t, u(t), \frac{\mathrm{d}u(t)}{\mathrm{d}t}, \ldots, \frac{\mathrm{d}^m u(t)}{\mathrm{d}t^m}\right) = 0$$

lässt sich mittels Einführung von Hilfsvariablen als System erster Ordnung schreiben:

$$\left.\begin{array}{rl} \frac{\mathrm{d}u_0(t)}{\mathrm{d}t} & = u_1(t) \\ \vdots & \vdots \\ \frac{\mathrm{d}u_{m-2}(t)}{\mathrm{d}t} & = u_{m-1}(t) \end{array}\right\} m-1\text{-Gleichungen}$$

$$u_0(t) = u(t) \qquad u_i(t) = \frac{\mathrm{d}^i u(t)}{\mathrm{d}t^i}$$

$$F\left(t, u_0(t), u_1(t), \ldots, u_{m-1}(t), \frac{\mathrm{d}u_{m-1}(t)}{\mathrm{d}t}\right) = 0$$

Spezielle Formen der rechten Seite

Ein System von DGL der Form (2.1) heißt

- separiert, falls $f(t,x) = a(t) \cdot g(x)$
- linear, falls $f(t,x) = A(t) \cdot x + b(t)$
- autonom, falls $f(t,x) = g(x)$

Eine nicht autonome DGL der Form (2.1) lässt sich immer auf ein autonomes System reduzieren

$$\overline{u}(t) := (u(t), t)$$
$$\overline{f}(\overline{u}(t)) := (f(t,u(t)), 1)^T$$

Integralgleichungsform

Mit dem Hauptsatz der Differential- und Integralrechnung lässt sich (2.1) als Integralsgleichung schreiben:

$$u'(t) = f(t,u(t)) \Leftrightarrow \boxed{u(t) = u(t_0) + \int_{t_0}^{t} f(s,u(s))\,\mathrm{d}s} \qquad t \in I$$

Anfangswertaufgabe (AWA)

Definition 2.1
Zu einem gegebenen Punkt $(t_0, u_0) \in D$ ist eine (stetig) diffbare Funktion $u\colon I \to \mathbb{R}^d$ gesucht mit Eigenschaften

1) $\mathrm{Graph}(u) := \{(t,u(t)) : t \in I\} \subset D$
2) $u'(t) = f(t,u(t)) \quad \forall\, t \in I$
3) $u(t_0) = u_0$

Dann heißt $u(t)$ Lösung der AWA.

2.2 Existenzaussagen

→ Wir betrachten Existenz und Eindeutigkeit getrennt!

Satz 2.2 Existenz von Peano
Die Funktion $f(t,x)$ sei stetig auf dem $(d+1)$-dimensionalen Zylinder

$$D = \left\{(t,x) \in \mathbb{R}^1 \times \mathbb{R}^d : |t - t_0| \le \alpha, \|x - u_0\| \le \beta\right\}$$

Dann existiert eine Lösung $u(t)$ der AWA auf dem Intervall $I := [t_0 - T, t_0 + T]$ wobei

$$T := \min\left(\alpha, \frac{\beta}{M}\right) \qquad M := \max_{(t,x)\in D} \|f(t,x)\|$$

Beweis .
Rannacher Satz 1.1 □

Satz 2.3 Fortsetzungssatz
Die Funktion $f(t,x)$ sei stetig auf einem abgeschlossenen Bereich D des $\mathbb{R}^1 \times \mathbb{R}^d$, welcher den Punkt (t_0, u_0) enthält und es sei eine Lösung der AWA auf $I = [t_0 - T, t_0 + T]$. Dann ist die lokale Lösung u nach rechts und links auf ein „maximales" Existenzintervall $I_{\max} = (t_0 - T_*, t_0 + T^*)$ stetig differenzierbar fortsetzbar, solange der Graph von u nicht an den Rand von D stößt. Dabei kann

$$\text{Graph}(u) := \{(t, u(t)), t \in I_{\max}\}$$

unbeschränkt sein, sowohl durch $t \to t_0 + T^* = \infty$, als auch durch

$$\|u(t)\| \to \infty (t \to t_0 + T^*)$$

Beweis .
Rannacher Satz 1.2 □

Korollar 2.4 Globale Existenz
$f(t,x)$ sei auf ganz $\mathbb{R}^1 \times \mathbb{R}^d$ definiert und stetig. Für jede durch den Satz von Peano gelieferte lokale Lösung $u(t)$, gelte

$$\|u(t)\| \le \beta(t) \qquad t \in [t_0 - T, t_0 + T]$$

mit einer festen stetigen Funktion $\beta\colon \mathbb{R} \to \mathbb{R}$.
Dann lässt sich u zu einer Lösung auf ganz $\mathbb{R}$ fortsetzen.

Beweis .
Rannacher Korollar 1.1 □

Beispiele:

(i) $u'(t) = \sin(\overbrace{u(t)}^{\text{autonome DGL}}) \qquad t \ge 0 \qquad u(0) = 0$

$f(t,x) = \sin(t)$ ist stetig, also existieren nach dem Satz von Peano lokale Lösungen. Wegen

$$|u(t)| = \left| u(t_0) + \int_{t_0}^{t} \sin(u(s))\, \mathrm{d}s \right|$$

$$\le \underbrace{|u(t_0)|}_{=0} + \int_0^t \underbrace{|sin(u(s))|}_{\le 1}\, \mathrm{d}s \le t$$

$\to$ Lösung auf ganz $\mathbb{R}$ fortsetzbar nach Korollar 2.4

(ii) $u'(t) = u(t)^{1/3}$

Für beliebiges $C \ge 0$ ist jede Funktion

$$u_C(t) = \begin{cases} 0 & 0 \le t \le C \\ \left(\frac{2}{3}(t - C)\right)^{3/2} & t > C \end{cases}$$

eine Lösung.

Beweis: Für $u_C(t) = 0$ trivial, da $0^{1/3} = 0$. Sonst:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\underbrace{\left[\frac{2}{3}(t-C)\right]^{3/2}}_{u(t)}\right\} = \frac{3}{2}\left[\frac{2}{3}(t-C)\right]^{1/2}\cdot\frac{2}{3} = \left(\underbrace{\left[\frac{2}{3}(t-C)\right]^{3/2}}_{u(t)}\right)^{1/3}$$

$\rightarrow$ AWA zu $(0,0)$ hat ∞-viele Lösungen

- Euler Verfahren liefert nur die 0-Lösung
- Zum Anfangswert $u(0) = 1$ ist die Lösung eindeutig $\rightarrow$ nächste Stunde

(iii) $u'(t) = u(t)^2 \qquad 0 \le t < 1 \qquad u(0) = 1$ besitzt eine lokale Lösung der Form

$$u(t) = \frac{1}{1-t},$$

d.h. $u(t) \to \infty$ für $t \to 1$ "Blow-up" in "finite time". Dagegen hat

$$u'(t) = -200tu^2(t), \qquad t \ge -3, \qquad u(-3) = \frac{1}{901}$$

die auf ganz $\mathbb{R}$ existierende Lösung

$$u(t) = \frac{1}{1+100t^2}$$

Kleine Änderung, große Wirkung.

(iv) $u'(t) = \lambda u(t) \qquad t \ge 0 \qquad u(0) = u_0 \qquad (\lambda \in \mathbb{C},$ Modellproblem) hat die global eindeutige Lösung $u(t) = u_0 \cdot e^{\lambda \cdot t}$

$$\begin{aligned}
\operatorname{Re}\lambda < 0 &: \lim_{t\to\infty} |u(t)| = 0\\
\operatorname{Re}\lambda = 0 &: \lim_{t\to\infty} |u(t)| = |u_0|\\
\operatorname{Re}\lambda > 0 &: \lim_{t\to\infty} |u(t)| = \infty
\end{aligned}$$

Satz 2.5 Regularitätssatz
Sei u eine Lösung der AWA aus Definition 2.1 auf dem Intervall I. Ist $f \in C^m(D)$ für ein $m \ge 1$, dann ist $u \in C^{m+1}(I)$.

$$\begin{aligned}
u''(t) &= \frac{\mathrm{d}}{\mathrm{d}t}(u'(t)) = \frac{\mathrm{d}}{\mathrm{d}t}(f(t,u(t))\\
&= f_t'(t,u(t)) + \underbrace{\nabla_x f(t,u(t))}_{\text{Jacobimatrix bezüglich } x} \cdot u'(t)
\end{aligned}$$

Für $f \in C^1(D)$ folgt also $u \in C^2(I)$.
Wiederholte Anwendung liefert das für $m > 1$ □

2.3 Eindeutigkeit und Stabilität

Definition 2.6 Lipschitzbedingung

- Die Funktion $f(t,x)$ genügt auf ihrem Definitionsbereich

$$D \subseteq \mathbb{R} \times \mathbb{R}^d$$

einer (gleichmäßigen) „Lipschitz-Bedingung“, wenn mit einer stetigen Funktion $L(t) > 0$ gilt:

$$\|f(t,x) - f(t,x')\| \leq L(t)\|x - x'\| \qquad (t,x),(t,x') \in D \tag{2.5}$$

- Die Funktion $f(t,x)$ genügt einer „lokalen“ Lipschitz-Bedingung, wenn $f(t,x)$ auf jeder beschränkten Teilmenge von D eine Lipschitz-Bedingung genügt (L darf von TM abhängen)

Beispiele:

1) $d = 1$. Sei $f(t,x)$ stetig, partiell differenzierbar nach x mit beschränkter Ableitung:

$$\max_{(t,x)\in D} |\partial_x f(t,x)| \leq K$$

dann gilt

$$|f(t,x) - f(t,x')| = \left| \int_x^{x'} \underbrace{\partial_x f(t,s)}_{\leq K} \,\mathrm{d}s \right| \leq K|x - x'|$$

lässt sich erweitern auf $d > 1$.

2) $f(t,x) = x^{1/3}$ nicht Lipschitz-stetig bei $x = 0$ aber in $[\varepsilon, \infty)$ mit $\varepsilon > 0$. → dies wird sich als Grund der Nichteindeutigkeit erweisen.

3) $f(t,x) = x^2 (d = 1)$ ist lokal Lipschitz-stetig:

$$|x^2 - y^2| = |(x+y)(x-y)| = |x+y||x-y| \leq L \cdot |x-y|$$

für $L = \max\{|x+y| \mid x,y \in D\}$

Satz 2.7 Lokaler Stabilitätssatz
Wir betrachten zwei AWA:

$$u'(t) = f(t,u(t)) \qquad t \in I \qquad u(t_0) = u_0 \tag{2.6a}$$

$$v'(t) = g(t,v(t)) \qquad t \in I \qquad v(t_0) = v_0 \tag{2.6b}$$

mit f, g stetig. Die Funktion $f(t,x)$ genüge der (gleichmäßigen) Lipschitz-Bedingung (2.5) auf D. Dann gilt für zwei beliebige Lösungen u von (2.6a) und v von (2.6b).

$$\|u(t) - v(t)\| \leq e^{L(t-t_0)} \left\{ \|u_0 - v_0\| + \int_{t_0}^{t} \varepsilon(s)\,\mathrm{d}s \right\} \qquad t \in I$$

mit

$$\varepsilon(t) = \sup_{x\in\Omega} \|f(t,x) - g(t,x)\|$$

Beweis. Rannacher □

Mit der Integraldarstellung gilt:

$$u(t) - v(t) = u(t_0) + \int_{t_0}^{t} f(s, u(s))\,\mathrm{d}s - v(t_0) - \int_{t_0}^{t} g(s, v(s))\,\mathrm{d}s$$
$$= \int_{t_0}^{t} f(s, u(s)) - f(s, v(s))\,\mathrm{d}s + \int_{t_0}^{t} f(s, v(s)) - g(s, v(s))\,\mathrm{d}s + u_0 - v_0$$

Für vektorwertiges $e(s)$ und jede beliebige Vektornorm $\|\cdot\|$ gilt

$$\| \int_{t_0}^{t} \underbrace{e(s)}_{\begin{pmatrix} \int_{t_0}^{t} e_1(s)\,\mathrm{d}s \\ \vdots \\ \int_{t_0}^{t} e_d(s)ds \end{pmatrix}} \mathrm{d}s\| = \left\| \lim_{N\to\infty} \sum_{i=1}^{N} e(S_i)\cdot(S_i - S_{i-1}) \right\| \overset{\text{stetig}}{=} \lim_{N\to\infty} \left\| \sum_{i=1}^{N} e(S_i)(S_i - S_{i-1}) \right\|$$

$$\leq \lim_{N\to\infty} \sum_{i=1}^{N} \|e(S_i)\|(S_i - S_{i-1}) = \int_{t_0}^{t} \|e(s)\| ds$$

Dann gilt für $e(t) = u(t) - v(t)$:

$$\|e(t)\| \leq \int_{t_0}^{t} \|f(s, u(s)) - f(s, v(s))\|\,\mathrm{d}s + \int_{t_0}^{t} \|f(s, v(s)) - g(s, v(s))\|\,\mathrm{d}s + \|u_0 - v_0\|$$

$$\leq L \int_{t_0}^{t} \|e(s)\|\,\mathrm{d}s + \int_{t_0}^{t} \varepsilon(s)\,\mathrm{d}s + \|u_0 - v_0\|$$

Lemma 2.8 Gronwall
Die stückweise stetige Funktion $w(t) \geq 0$ genüge der Integralgleichung

$$w(t) \leq \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s + b(t) \qquad t \geq t_0$$

mit $a(t) \geq 0$ integrierbar und $b(t) \geq 0$ nicht fallend. Dann gilt

$$w(t) \leq \exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s \right) b(t) \qquad t \geq t_0$$

Beweis. Setze

$$\varphi(t) \coloneqq \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s$$

$$\psi(t) \coloneqq w(t) - \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s \leq b(t)$$

$\varphi(t)$ erfüllt

$$\varphi'(t) = a(t)\cdot w(t) \qquad \varphi(t_0) = 0.$$

Weiter ist

$$a(t)\cdot\psi(t) = \underbrace{a(t)\cdot w(t)}_{\varphi'(t)} - a(t)\underbrace{\int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s}_{\varphi(t)} = \varphi'(t) - a(t)\cdot\varphi(t)$$

Für gegebenes $\psi(t)$ erfüllt $\varphi(t)$ also die AWA:

$$\varphi'(t) = a(t)\cdot\varphi(t) + a(t)\cdot\psi(t) \qquad \varphi(t_0) = 0 \qquad t \geq t_0$$

Diese AWA hat die (eindeutige) Lösung:

$$\begin{aligned}
\varphi(t) &= \exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)\cdot\int_{t_0}^{t}\exp\left(-\int_{t_0}^{s} a(r)\,\mathrm{d}r\right)a(s)\psi(s)\,\mathrm{d}s \\
&\leq b(t)\exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)\int_{t_0}^{t}\left(-\frac{\mathrm{d}}{\mathrm{d}s}\exp\left(-\int_{t_0}^{s} a(r)\,\mathrm{d}r\right)\right)\mathrm{d}s \\
&= b(t)\exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)\left[-\exp\left(-\int_{t_0}^{s} a(r)\,\mathrm{d}r\right)\right]_{t_0}^{t} \\
&= b(t)\cdot\exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)\left(-\exp\left(-\int_{t_0}^{t} a(r)\,\mathrm{d}r\right)+1\right) \\
&= b(t)\cdot\exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right) - b(t)
\end{aligned}$$

Einsetzen in die Integralgleichung

$$w(t) \leq \underbrace{\int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s}_{=\varphi(t)} + b(t) \leq b(t)\cdot\exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)$$

□

Damit weiter im Satz von 2.7:

$$w(t) := \|e(t)\| \qquad a(t) = L \qquad b(t) = \underbrace{\int_{t_0}^{t} \varepsilon(s)\,\mathrm{d}s + \|u_0 - v_0\|}_{\geq 0,\ \text{nicht fallend}}$$

also nach Gronwall:

$$\|e(t)\| \leq e^{L(t-t_0)} \left\{ \int_{t_0}^{t} \varepsilon(s)\,\mathrm{d}s + \|u_0 - v_0\| \right\}$$

Korollar 2.9 Eindeutigkeitssatz
Erfüllt $f(t,x)$ eine Lipschitz-Bedingung, so ist die durch den Existenzsatz von Peano gelieferte Lösung eindeutig bestimmt.

Beweis. Seien $u(t), v(t)$ zwei verschiedene Lösungen derselben AWA.

$$u'(t) = f(t,u(t)) \qquad u(t_0) = u_0$$
$$v'(t) = f(t,v(t)) \qquad v(t_0) = v_0 = u_0$$

Dann gilt nach dem Stabilitätssatz:

$$\|u(t) - v(t)\| \leq e^{L(t-t_0)} \left\{ \underbrace{\|u_0 - v_0\|}_{=0} + \int_{t_0}^{t} \sup_{x\in\Omega} \underbrace{\|f(s,x) - f(s,x)\|}_{=0,\ \text{da } g=f}\,\mathrm{d}s \right\} = 0 \ \lightning$$

□

Korollar 2.10
Betrachte die DGL d-ter Ordnung

$$u^{(d)}(t) = f(t, u(t), \ldots, u^{(d-1)}(t)), t \geq t_0$$

$f\colon I \times \mathbb{R}^d$ sei Lipschitz-stetig bzgl. der letzten d Argumente. Dann existiert für jeden Satz von d Werten $u_0, \ldots, u_{d-1} \in \mathbb{R}$ genau eine Lösung, die der Anfangsbedingung $u^{(i)}(t_0) = u_i$ genügt.

Beweis. Reduktion auf System $y'(t) = F(t, y(t))$. Zeige, dass F Lipschitz-stetig ist. □

Beispiele:

- $f(t,x) = x^2$ $(d = 1)$ ist lokal Lipschitz-stetig
- Eindeutige Lösung auf beschränktem Intervall
- solange die Lösung existiert ist sie eindeutig

Korollar 2.11 Globale Existenz bei „linear beschränkter“ Nichtlinearität
$f(t,x)$ sei stetig auf $D = \mathbb{R}^1 \times \mathbb{R}^d$ und genüge der Wachstumsbedingung

$$\|f(t,x)\| \leq \alpha(t)\|x\| + \beta(t) \qquad (t,x) \in D$$

mit stetigem $\alpha(t), \beta(t) \geq 0$. Dann besitzt die AWA eine „globale“ Lösung. Erfüllt f eine Lipschitzbedingung, ist die Lösung eindeutig.

Beweis. $u(t)$ sei Lösung nach Peano auf $I = [t_0, t_0 + I]$ zum Startpunkt (t_0, u_0).

$$\underbrace{\|u(t)\|}_{=:w(t)} = \left\| u_0 + \int_{t_0}^{t} f(s, u(s))\, \mathrm{d}s \right\| \leq \|u_0\| + \int_{t_0}^{t} \{\alpha(s)\|u(s)\| + \beta(s)\}\ \mathrm{d}s$$

Gronwall:

$$\|u(t)\| \leq \exp\left[\int_{t_0}^{t} \alpha(s)ds\right]\left\{\|u_0\| + \int_{t_0}^{t} \beta(s)ds\right\}, \quad t \in I,$$

d.h.

$$\|u(t)\| \leq G(T, \alpha, \beta)$$

also endlich (kein blow-up).
Fortsetzungssatz: Lösung kann fortgesetzt werden bis zum Rand von D, also auf ganz $\mathbb{R}^1 \times \mathbb{R}^d$, also existiert u für alle $t \geq t_0$. □

Eindeutigkeit folgt aus Korollar 2.9.
Falls f global Lipschitz-stetig, existiert die Lösung der AWA global.

$$\|f(t,x)\| = \|f(t,x) - f(t,0) + f(t,0)\| \leq \|f(t,x) - f(t,0)\| + \|f(t,0)\| \leq L\|x\| + \|f(t,0)\|$$

Korollar 2.12 Lineare AWA
Gegeben sind die stetigen

$$\text{Matrixfunktion: } A\colon [t_0, \infty) \to \mathbb{R}^{d\times d}$$
$$\text{Vektorfunktion: } b\colon [t_0, \infty) \to \mathbb{R}^{d}$$

Dann hat die lineare AWA:

$$u'(t) = A(t)u(t) + b(t) \qquad t \geq t_0 \qquad u(t_0) = u_0$$

eine eindeutige globale Lösung auf $I = [t_0, \infty]$.

Beweis .

$$\|f(t,x) - f(t,x')\| = \|A(t)x + b(t) - A(t)x' - b(t)\| = \|A(t)(x - x')\| \leq \underbrace{\|A(t)\|}_{L}\|x - x'\|$$

□

2.4 Globale Stabilität

- bisherige Abschätzung im Stabilitätssatz sehr pessimistisch
- für $T \to \infty$ wächst Abstand bei Störung (schnell) über alle Grenzen (lokale Stabilität)

Definition 2.13 Monotone AWA
Die Funktion $f(t, x)$ genügt einer „Monotoniebedingung“ wenn mit

$$\lambda(t) > 0 \text{ und } \lambda := \inf_{t\in I} \lambda(t) > 0$$

gilt:

$$-(f(t,x) - f(t,y), x - y) \geq \lambda(t)\|x - y\|^2 \qquad (t,x),(t,y) \in D$$

Verallgemeinerung von „monoton fallend"für vektorwertige f:

1) f sei *skalar*, also $d = 1$

$$-(f(x) - f(y), x - y) = -(f(x) - f(y))(x - y) \geq \lambda(x - y)^2$$

$$\Leftrightarrow -\frac{f(x) - f(y)}{x - y} \geq \lambda$$

$$\overset{\cdot(-1)}{\Leftrightarrow} \underbrace{\frac{f(x) - f(y)}{x - y}}_{\text{Steigung}} \leq -\lambda < 0$$

bzw. $f' < 0$ für f differenzierbar.

2) Beispiel für $d > 1$ lineares $f(t, x) = A(t)x + b(t)$

$$(A(t)x + b(t) - A(t)y - b(t), x - y) = \left(-A(t)\underbrace{(x - y)}_{z}, \underbrace{(x - y)}_{z}\right) \geq \lambda(t)\underbrace{\|x - y\|^2}_{z^2} \quad \forall\, x, y$$

d.h. f erfüllt in diesem Fall die Monotoniebedingung genau dann wenn $-A(t)$ positiv definit bzw. wenn $A(t)$ negativ definit.

Definition 2.14 Exponentielle Stabilität
Eine globale Lösung $u(t)$ einer AWA heißt „exponentiell stabil“, wenn es $\delta, \alpha, A > 0$ gibt, so dass gilt: Zu jedem Zeitpunkt $t_* \geq t_0$ und jedem $w_* \in \mathbb{R}^d$ mit $\|w_*\| \leq \delta$ hat die gestörte AWA

$$v'(t) = f(t, v(t)) \qquad t \geq t_* \geq t_0 \qquad v(t_*) = u(t_*) + w_*$$

ebenfalls eine globale Lösung $v(t)$ für die gilt

$$\|v(t) - u(t)\| \leq Ae^{-\alpha(t - t_*)}\|w_*\| \qquad t \geq t_*$$

Bemerkungen:

- gestörte Lösung läuft exponentiell auf die ungestörte Lösung zurück
- Anwendung: Untersuchung von Fixpunkten autonomer Systeme f , d.h. $f(t, x) = f(x)$. Sei $u^e \in \mathbb{R}^d$ so gewählt, dass $f(u^e) = 0$ so gilt für

$$u'(t) = f(u(t)) \qquad t \geq t_0 \qquad u(t_0) = u^e$$

$u(t) = u^e$ für $t \geq t_0$. u^e heißt „Fixpunkt“.
- Es gibt weitere Stabilitätsdefinitionen, die schwächer sind. Lyapunov-Stabilität: Zu jedem U gibt es $V \subset U$. Für jeden Startwert in V bleibt die Lösung innerhalb von U.

Satz 2.15 Globaler Stabilitätssatz
Alle Lösungen einer L-stetigen und monotonen AWA sind global und exponentiell stabil mit

$$\delta \text{ beliebig} \qquad \alpha = \lambda \qquad A = 1$$

Im Fall

$$\sup_{t > t_0} \|f(t, 0)\| < \infty$$

sind alle Lösungen gleichmäßig (d.h. unabhängig von t) beschränkt.

Beweis .

(i) Existenz globaler und eindeutiger Lösung ist Folge der Lipschitz-Bedingung

(ii) Exponentielle Stabilität

$$\begin{aligned} \text{ungestörtes Problem: } u'(t) = \quad & f(t,u(t)) \quad && t \geq t_0 && u(t_0) = u_0 \\ \text{gestörtes Problem: } v'(t) = \quad & f(t,v(t)) \quad && t \geq t_* \geq t_0 && v(t_*) = u(t_*) + w_* \end{aligned}$$

Setze

$$w(t) = v(t) - u(t)), \text{ also } w'(t) = v'(t) - u'(t)$$

also

$$w'(t) = (f(t,v(t)) - f(t,u(t)) = 0 \mid \text{multipliziere Skalar mit } w(t)$$
$$\underbrace{(w'(t), w(t))} - (f(t,v(t) - f(t,u(t)), w(t)) = 0$$

Nebenrechnung: $(u^2)' = (u'u + uu') = 2u'u$

$$\Leftrightarrow 0 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|w(t)\|^2 \underbrace{- \Big(f(t,\underline{v(t)})\Big) - f\Big(t,\underline{u(t)}, v(t) - u(t)\Big)} = 0$$

Monotoniebedingung:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|w(t)\|^2 + \lambda\|w(t)\|^2 \leq 0$$

multipliziere beide Seiten mit $2e^{2\lambda(t-t_*)}$

$$e^{2\lambda(t-t_*)}\frac{\mathrm{d}}{\mathrm{d}t}\|w(t)\|^2 + 2\lambda e^{2\lambda(t-t_*)} \leq 0$$

$$\Leftrightarrow \frac{\mathrm{d}}{\mathrm{d}t}\left[e^{2\lambda(t-t_*)}\|w(t)\|^2\right] \leq 0$$

Integral über $[t_0, t]$

$$e^{2\lambda(t-t_*)}\|w(t)\|^2 - \|w(t_*)\|^2 \leq 0$$

Wurzelziehen:

$$\|w(t)\| \leq e^{-\lambda(t-t_*)}\|w_*\|$$

(iii) Beschränktheit: Rannacher (Beweis zu Satz 1.7 (ii))

□

Korollar 2.16

$$A(t)\colon [t_0,\infty) \to \mathbb{R}^{d\times d}$$

sei gleichmäßig negativ definit und $b(t)\colon [t_0,\infty) \to \mathbb{R}^d$ sei beschränkt dann hat die lineare AWA

$$u'(t) = A(t)u(t) + b(t) \qquad t \geq t_0 \qquad u(t_0) = u_0$$

eine eindeutige globale Lösung

$$u\colon [t_0,\infty) \to \mathbb{R}^d$$

die beschränkt und exponentiell stabil ist.

Beweis .

(i) Lipschitzbedingung → eindeutige, globale Lsg.

(ii) $A(t)$ gleichmässig negativ definit → AWA ist monoton (siehe oben)

(iii)
$$\sup_{t\in[t_0,\infty)} \|f(t,0)\| = \sup_{t\in[t_0,\infty)} \|b(t)\| \underbrace{< \infty}_{\text{nach Vor.}}$$
somit ist die Lösung nach Satz 2.15 beschränkt.

□

3 Einschrittverfahren

Wir betrachten die AWA

$$u'(t) = f(t,u(t)) \qquad t \in [t_0, t_0+T] = I \qquad u(t_0) = u_0 \tag{3.1}$$

Die Funktion f sei stetig und erfülle eine L-Bedingung. (3.2)

⇒ globale, eindeutige Lösung
Zeitgitter: Unterteile $I = [t_0, t_0+T]$ in

$$t_0 < t_1 < \ldots < t_N = t_0 + T \qquad T < \infty \qquad N \in \mathbb{N}$$

und setze:

$$I_n := [t_{n-1}, t_n] \qquad h_n := t_n - t_{n-1} \qquad h := \max_{1\le n\le N} h_n$$

„nicht-äquidistantes Gitter“

3.1 Das explizite Euler-Verfahren

liefert für jedes $N \in \mathbb{N}$ die endliche Folge:

$$y_n^h = y_{n-1}^h + h_n f(t_{n-1}, y_{n-1}^h) \qquad 1 \le n \le N \tag{3.3}$$

und

$$y_0^h = y_0 \qquad (y_0 \text{ nicht unbedingt } u_0, \text{ z.B. } \mathrm{rd}(u_0))$$

Setze

$$y^h = \left((y_1^h)^T, \ldots, (y_N^h)^T\right)^T \in \mathbb{R}^{Nd}$$

und definiere den „Differenzenoperator“

$$L_h \colon \mathbb{R}^{Nd} \times \mathbb{R}^d \to \mathbb{R}^{Nd} \qquad \left(L_h(y^h, y_0^h)\right)_n = h_n^{-1}\left(y_n^h - y_{n-1}^h\right) - f(t_{n-1}, y_{n-1}^h) \qquad 1 \le n \le N \tag{3.4}$$

y^h ist Lösung der Gleichung:

$$L_h(y^h, y_0^h) = 0.$$

$$y^h = \left((y_1^T), (y_2^T), \ldots, (y_N^T)\right) \in \mathbb{R}^{Nd}$$

Damit schreibt sich 3.3 formal als:

$$\boxed{\text{Finde } y^h \in \mathbb{R}^{Nd} \text{ so dass } L_h(y^h, y_0^h) = 0}$$

Sei $u(t)$ die Lösung der AWA (3.1), dann setze

$$u_n^h := u(t_n)$$

und

$$u^h := \left((u_1^h)^T, \ldots, (u_N^h)^T\right)^T$$

Lokaler Diskretisierungsfehler

Die folgende Größe

$$\tau_n^h := \left(L_h(u^h, u_0)\right)_n = h_n^{-1}\left(u_n^h - u_{n-1}^h\right) - f(t_{n-1}, u_{n-1}^h)$$

heißt „Abschneidefehler“ oder „lokaler Diskretisierungsfehler“.
Für den Abschneidefehler gilt:

$$\begin{aligned}
\tau_n^h &= h_n^{-1}(u_n^h - u_{n-1}^h) - f(t_{n-1}, u_{n-1}^h) = h_n^{-1}\int_{t_{n-1}}^{t_n} u'(t)\,\mathrm{d}t - u'(t_{n-1}) \\
&\underbrace{=}_{part.Integr.} h_n^{-1}\left\{[tu'(t)]_{t_{n-1}}^{t_n} - \int_{t_{n-1}}^{t_n} tu''(t)\,\mathrm{d}t\right\} - u'(t_{n-1}) \\
&= h_n^{-1}\left\{t_n u'(t_n) - t_{n-1}u'(t_{n-1}) - \int_{t_{n-1}}^{t_n} tu''(t)\,\mathrm{d}t\right\} - \underbrace{\frac{1}{t_n - t_{n-1}}}_{h_n^{-1}}(t_n - t_{n-1})\,u'(t_{n-1}) \\
&= h_n^{-1}\left\{t_n u'(t_n) - \cancel{t_{n-1}u'(t_{n-1})} - (t_n - \cancel{t_{n-1}})u'(t_{n-1}) - \int_{t_{n-1}}^{t_n} tu''(t)\,\mathrm{d}t\right\} \\
&= h_n^{-1}\left\{t_n[\underbrace{u'(t_n) - u'(t_{n-1})}] - \int_{t_{n-1}}^{t_n} tu''(t)\,\mathrm{d}t\right\} \\
&= h_n^{-1}\left\{t_n\int_{t_{n-1}}^{t_n} u''(t)\,\mathrm{d}t - \int_{t_{n-1}}^{t_n} tu''(t)\,\mathrm{d}t\right\} = h_n^{-1}\int_{t_{n-1}}^{t_n}(t_n - t)u''(t)\,\mathrm{d}t \\
\|\tau_n^h\| &= \|h_n^{-1}\int_{t_{n-1}}^{t_n}(t_n - t)u''(t)\,\mathrm{d}t\| \le h_n^{-1}\int_{t_{n-1}}^{t_n}(t_n - t)\|u''(t)\|\,\mathrm{d}t \\
&\le h_n^{-1}\max_{t\in I_n}\|u''(t)\|\underbrace{\int_{t_{n-1}}^{t_n} t_n - t\,\mathrm{d}t}_{=\frac{1}{2}h_n^2} = \frac{1}{2}h_n\max_{t\in I_n}\|u''(t)\| \qquad (3.5)
\end{aligned}$$

- Diskretisierung „erster Ordnung“
- erfordert höhere „Regularität“ der Lösung

Globaler Diskretisierungsfehler

$e_n^h = y_n^h - u_n^h$ heißt globaler Fehler

$$\begin{aligned} e_n^h = y_n^h - u_n^h &= \underbrace{y_{n-1}^h + h_n f(t_{n-1}, y_{n-1}^h)}_{y_n^h} - \underbrace{(u_n^h \overbrace{- u_{n-1}^h - h_n f(t_{n-1}, u_{n-1}^h)}^{\text{dazugefügt}})}_{h_n \tau_n^h} - u_{n-1}^h - h_n f(t_{n-1}, u_{n-1}^h) \\ &= \underbrace{y_{n-1}^h - u_{n-1}^h}_{e_{n-1}^h} + h_n(\underbrace{f(t_{n-1}, y_{n-1}^h) - f(t_{n-1}, u_{n-1}^h)}) - h_n \tau_n^h \end{aligned}$$

Norm nehmen, Lipschitzbedingung, Dreiecksungleichung:

$$\|e_n^h\| \leq \|e_{n-1}^h\| + h_n L \|e_{n-1}^h\| + h_n \|\tau_n^h\|$$

Abspulen der Rekursion:

$$\|e_n^h\| \leq \underbrace{\|e_0^h\|}_{=y_0-u_0} + L \sum_{\nu=0}^{n-1} h_{\nu+1} \|e_\nu^h\| + \sum_{\nu=1}^{n} h_\nu \|\tau_\nu\| \tag{3.6}$$

Lemma 3.1 Diskretes Gronwall-Lemma
Seien $(w_n)_{n\geq 0}, (a_n)_{n\geq 0}$ und $(b_n)_{n\geq 0}$ Folgen nicht negativer Zahlen für die gilt

$$w_0 \leq b_0 \text{ und } w_n \leq \sum_{\nu=0}^{n-1} a_\nu w_\nu + b_n \qquad n \geq 1$$

Ist die Folge $(b_n)_{n\geq 0}$ nicht fallend, dann gilt:

$$w_n \leq \exp\left(\sum_{\nu=0}^{n-1} a_\nu\right) b_n \qquad n \geq 1$$

Beweis .
Definiere

$$\begin{aligned} S_n &= \sum_{\nu=0}^{n-1} a_\nu w_\nu + b_n & S_0 &= b_0 \\ d_n &= S_n - w_n & d_0 &= b_0 - w_0 \end{aligned}$$

Zeige per Induktion

$$S_n \leq \exp\left(\sum_{\nu=0}^{n-1} a_\nu\right) b_n \qquad n \geq 0$$

1) $S_0 = b_0 = e^0 b_0$

2) $n-1 \to n$

$$S_n - S_{n-1} = \sum_{\nu=0}^{n-1} a_\nu w_\nu + b_n - \sum_{\nu=0}^{n-2} a_\nu w_\nu - b_{n-1} = a_{n-1} w_{n-1} + b_n - b_{n-1}$$

also

$$\begin{aligned} S_n &= S_{n-1} + a_{n-1} \underbrace{w_{n-1}}_{\leq S_{n-1}} + b_n - b_{n-1} \leq (1 + a_{n-1}) S_{n-1} + b_n - b_{n-1} \\ &= \underbrace{(1 + a_{n-1}) \exp\left(\sum_{\nu=0}^{n-2} a_\nu \right)}_{\geq 1, \text{ da } a_\nu \geq 0} b_{n-1} + \underbrace{b_n - b_{n-1}}_{\geq 0, (b_n) \text{ nichtfallend}} \\ &\leq \underbrace{(1 + a_{n-1})}_{\exp(a_{n-1})} \exp\left(\sum_{\nu=0}^{n-2} a_\nu \right) (\cancel{b_{n-1}} + b_n - \cancel{b_{n-1}}) \\ &\leq \exp\left(\sum_{\nu=0}^{n-1} a_\nu \right) b_n \end{aligned}$$

□

damit weiter für den globalen Fehler:

$$\|e_n^h\| \leq \sum_{\nu=0}^{n-1} \underbrace{L h_{\nu+1}}_{a_\nu} \underbrace{\|e_\nu^h\|}_{w_\nu} + \underbrace{\sum_{\nu=1}^{n} h_\nu \|\tau_n^h\| + \|e_0^h\|}_{b_n}$$

$$\overset{\text{Lemma 3.1}}{\leq} \underbrace{\exp\left(\sum_{\nu=0}^{n-1} L h_{\nu+1} \right)}_{= L(t_n - t_0) = LT} \left\{ \|e_0^h\| + \underbrace{\sum_{\nu=1}^{n} h_\nu}_{\leq T} \underbrace{\|\tau_n^h\|}_{\leq \frac{h}{2} \max_{t \in I} \|u''(t)\|} \right\}$$

$$\max_{1 \leq n \leq N} \|e_n^h\| \leq e^{LT} \left\{ \|e_0^h\| + \frac{T}{2} h \cdot \max_{t \in I} \|u''(t)\| \right\}$$

- Globale Konvergenzordnung = lokale Konvergenzordnung
- e^{LT} ist sehr pessimistisch in der Regel

Kurzer Ausflug: impliziter Euler

$$e_n^h = y_n^h - u_n^h = \underbrace{y_{n-1}^h + h_n f(t_n, y_n^h)}_{\text{impliziter Euler}} - \underbrace{(u_n^h - \underbrace{u_{n-1}^h - h_n f(t_n, u_n^h)}_{\text{hinzugefügt}})}_{h_n \tau_n^h} - u_{n-1}^h - h_n f(t_n, u_n^h)$$

$$\begin{aligned} &= y_{n-1}^h - u_{n-1}^h + h_n (f(t_n, y_n^h) - f(t_n, u_n^h)) - h_n \tau_n^h \\ \|e_n^h\| &\leq \|e_{n-1}^h\| + h_n L \|e_n^h\| + h_n \|\tau_n^h\| \\ &\leq \sum_{\nu=1}^{n} h_\nu L \|e_\nu^h\| + \|e_0^h\| + \sum_{\nu=1}^{n} h_\nu \|\tau_\nu^h\| \end{aligned}$$

Man erhält also eine implizite Summengleichung:

$$w_n \leq \sum_{\nu=0}^{n} a_\nu w_\nu + b_n \qquad n \geq 1$$
$$w_0 \leq b_0 \qquad n = 0$$

Sei $a_n < 1$, dann gilt

$$w_n \leq a_n w_n + \sum_{\nu=0}^{n-1} a_\nu w_\nu + b_n$$

$$\underbrace{(1-a_n)w_n}_{\tilde{w}_n} \leq \sum_{\nu=0}^{n-1} \underbrace{\frac{a_\nu}{1-a_\nu}}_{\tilde{a}_\nu} \underbrace{(1-a_\nu)w_\nu}_{\tilde{w}_\nu} + b_n$$

jetzt diskretes Gronwall-Lemma anwendbar:

$$\tilde{w}_n \leq \exp\left(\sum_{\nu=0}^{n-1} \tilde{a}_\nu\right) b_n$$
$$\Leftrightarrow (1-a_n)w_n \leq \exp\left(\sum_{\nu=0}^{n-1} \frac{a_\nu}{1-a_\nu}\right) b_n$$
$$w_n \leq \underbrace{\frac{1}{1-a_n}}_{\leq \exp(\frac{a_n}{1-a_n})} \exp\left(\sum \frac{a_\nu}{1-a_\nu}\right)$$

$$\boxed{w_n \leq \exp\left(\sum_{\nu=0}^{n} \frac{a_\nu}{1-a_\nu}\right) b_n} \text{ wobei } \boxed{a_\nu < 1}$$

zurück zum impliziten Euler: Bedingung $a_\nu < 1$ bedeutet Konvergenz falls $h_\nu L < 1$ klein genug.

3.2 Taylor und Runge-Kutta Verfahren

Ziel: Konstruktion von Einschritt-Verfahren höherer Ordnung
Idee:

(a) Taylorentwicklung um $t - h$:

$$u(t) = \sum_{r=0}^{R} \frac{h^r}{r!} u^{(r)}(t-h) + \frac{h^{R+1}}{(R+1)!} u^{(R+1)}(\xi) \qquad \xi \in [t-h, t]$$

(b) Differenzieren der DGL:

$$u^{(r)}(t) = \frac{\mathrm{d}^{r-1}}{\mathrm{d}t^{r-1}} f(t, u(t)) =: f^{r-1}(t, u(t))$$

R-stufiges "Taylor-Verfahren" (Setze b) in a) ein)

$$u(t) = u(t-h) + \sum_{r=1}^{R} \frac{h^r}{r!} f^{r-1}(t-h, u(t-h)) + \frac{h^{R+1}}{(R+1)!} u^{(R+1)}(\xi)$$

Weglassen des Restgliedes ergibt das Taylorverfahren der Stufe R

$$y_n^h = y_{n-1}^h + h_n \underbrace{\sum_{r=1}^{R} \frac{h_n^{r-1}}{r!} f^{r-1}(t_{n-1}, y_{n-1}^h)}_{=:F(h_n, t_{n-1}, y_{n-1}^h, y_n^h)} \tag{3.8}$$

Ein Verfahren der Form

$$y_n^h = y_{n-1}^h + h_n F(h_n, t_{n-1}, y_{n-1}^h, y_n^h) \tag{3.9}$$

heißt allgemeines Einschrittverfahren.

Beachte: Auch "implizite" Verfahren sind möglich.
Wie würde man das praktisch machen?
Ableitungen von f ausrechnen

$$f^1(t,u(t)) = \frac{\mathrm{d}}{\mathrm{d}t} f(t,u(t)) = f_t(t,u(t)) + f_x(t,u(t)) \cdot \underbrace{f(t,u(t))}_{u'(t)} = (f_t + f_x f)(t,u(t))$$

$$f^2(t,u(t)) = \frac{\mathrm{d}}{\mathrm{d}t} f^1(t,u(t)) = \left[f_{tt} + f_{tx} f + (f_{xt} + f_{xx}) f + f_x (f_t + f_x f)\right](t,u(t))$$

$\Rightarrow$theoretisch möglich, aber unpraktisch, besonders im vektorwertigen Fall ($d > 1$) oder by höherer Ordnung (kombinatorische Explosion)

Definition 3.2 Konsistenz
Die Einschrittmethode

$$y_n^h = y_{n-1}^h + h_n F(h_n, t_{n-1}, y_{n-1}^h, y_n^h)$$

heißt "konsistent" (mit der AWA) bzw. "konsistent mit Konsistenzordnung m", wenn für den Abschneidefehler

$$\tau_n^h := (L_h u^h)_n = h_n^{-1} \left\{ u_n^h - u_{n-1}^h \right\} - F(h_n, t_{n-1}, u_{n-1}^h, u_n^h) \tag{3.10}$$

gilt

$$\max_{t_n \in I} \|\tau_n^h\| \to 0 \qquad \text{bzw.} \qquad \max_{t_n \in I} \|\tau_n^h\| = O(h^m) \qquad (h \to 0)$$

Nach Konstruktion gilt für das R-stufige Taylorverfahren gerade die Konsistenzordnung m=R

$$\begin{aligned}
\tau_n^h &= h_n^{-1} \{u(t_n) - u(t_{n-1})\} - \sum_{r=1}^{R} \frac{h_n^{r-1}}{r!} f^{r-1}(t_{n-1}, u(t_{n-1})) \\
&= h_n^{-1} \left\{ \underbrace{\cancel{u(t_{n-1})} + \sum_{r=1}^{R} \frac{h_n^r}{r!} u^{(r)}(t_{n-1}) + \frac{h_n^{R+1}}{(R+1)!} u^{(R+1)}(\xi)}_{u(t_n)} \cancel{- u(t_{n-1})} \right\} - \cancel{\sum_{r=1}^{R} \frac{h_n^{r-1}}{r!} f^{r-1}(t_{n-1}, u(t_{n-1}))} \\
&= h_n^R \frac{u^{(R+1)}(\xi)}{(R+1)!} \qquad \xi \in [t_{n-1}, t_n].
\end{aligned}$$

Runge-Kutta Verfahren

Idee: Ersetze Ableitungen (von f) durch numerische Approximation (Differenzenquotienten).
Beispiel: $R = 2$

$$\begin{aligned}
f^{(1)}(t-h,u(t-h)) &= \frac{1}{h}\left(f(t,u(t)) - f(t-h,u(t-h))\right) + O(h) \\
&= \frac{1}{h}\left(f(t,\underbrace{u(t-h) + hf(t-h,u(t-h)) + O(h^2)}) - f(t-h,u(t-h))\right) + O(h) \\
&= \frac{1}{h}\left(f(t,u(t-h) + hf(t-h,u(t-h))) + O(h^2) - f(t-h,u(t-h))\right) + O(h) \\
&= \frac{1}{h}\left[f(t,u(t-h) + hf(t-h,u(t-h))) - f(t-h,u(t-h))\right] + O(h)
\end{aligned}$$

und damit für das 2-stufige Taylor-Verfahren:

$$\begin{aligned}
u(t) &= u(t-h) + hf(t-h,u(t-h)) \\
&\quad + \frac{h^2}{2}\frac{1}{h}\left[f(t,u(t-h) + hf(t-h,u(t-h))) - f(t-h,u(t-h))\right] + O(h^3)
\end{aligned}$$

somit erhält man das folgende Verfahren mit Konsistenzordnung 2:

$$y_n^h = y_{n-1}^h + \frac{h}{2}\underbrace{f(t_{n-1},y_{n-1}^h)}_{=:k_1} + \frac{h}{2}\overbrace{f(t_n,y_{n-1}^h + h\underbrace{f(t_{n-1},y_{n-1}^h)}_{=:k_1})}^{=:k_2}$$

Man nennt dies das Verfahren von Heun.
Das Beispiel motiviert den allgemeinen Ansatz für Runge-Kutta Verfahren:

$$\begin{aligned}
y_n^h &= y_{n-1}^h + h_n(b_1k_1 + \ldots + b_sk_s) \text{ mit} \\
k_1 &= f(t_{n-1},y_{n-1}^h) \\
k_i &= f(t_{n-1} + c_ih_n, y_{n-1}^h + h_n\sum_{j=1}^{i-1} a_{ij}k_j) \qquad i = 2,\ldots,s
\end{aligned}$$

mit freien Parametern $s \in \mathbb{N}$ (Stufenanzahl), b_i, c_i, a_{ij}.
Hinweis: die Bezeichnungen sind hier anders als im Rannacher, da sich Rannacher nicht an die Standardnotation hält, die man auch in sonstiger Literatur findet.
Darstellung im sogenannten „Butcher-Tableau“:

$$\begin{array}{r|cccc}
0 = c_1 & 0 & & & \\
c_2 & a_{21} & 0 & & \\
\vdots & & \ddots & \ddots & \\
c_s & & & a_{s,s-1} & 0 \\
\hline
 & b_1 & \ldots & b_{s-1} & b_s
\end{array} \qquad \rightarrow \qquad \begin{array}{c|c} C & A \\ \hline & b^t \end{array}$$

A strikte unter Dreiecksmatrix → explizit
A untere Dreiecksmatrix → diagonal implizit
A voll → voll implizit

$s = 1$

$$\begin{aligned} k_1 &= f(t_{n-1}, y^h_{n-1}) \\ y^h_n &= y^h_{n-1} + h_n b_1 k_1 = y^h_{n-1} + h_n b_1 f(t_{n-1}, y^h_{n-1}) \\ \text{Taylor: } u(t_n) &= u(t_{n-1}) + h_n f(t_{n-1}, u(t_{n-1})) + O(h^2) \\ &\Rightarrow b_1 = 1 \end{aligned}$$

Es existiert nur ein Verfahren mit Konsistenzordnung 1 (Expliziter Euler)!

$s = 2$ **(skalar)**

$$\begin{aligned} k_2 &= f(t_{n-1} + c_2 h_n, y^h_{n-1} + h_n a_{21} k_1) \\ &= f(t_{n-1}, y^h_{n-1}) + h_n c_2 f_t(t_{n-1}, y^h_{n-1}) + h_n a_{21} \underbrace{k_1 f_x}_{=f f_x} (t_{n-1} y^h_{n-1}) + O(h_n^2) \end{aligned}$$

Einsetzen: (Argument ist immer (t_{n-1}, y_{n-1})):

$$\begin{aligned} y^h_n &= y^h_{n-1} + h_n(b_1 k_1 + b_2 k_2) \\ &= y^h_{n-1} + h_n(b_1 f + b_2 f + b_2 h_n c_2 f_t + b_2 h_n a_{21} f f_x) + O(h_n^3) \\ &= y^h_{n-1} + h_n(b_1 + b_2) f + h_n^2 (b_2 c_2) f_t + h_n^2 b_2 a_{21} f f_x + O(h_n^3) \end{aligned}$$

vergleiche mit

$$u(t_n) = u(t_{n-1}) + h_n f + h_n^2 \frac{1}{2} f_t + h_n^2 \frac{1}{2} f f_x + O(h_n^3)$$

$$\Rightarrow b_1 + b_2 = 1 \qquad b_2 c_2 = \frac{1}{2} \qquad b_2 a_{21} = \frac{1}{2}$$

also drei Bedingungen für die vier Koeffizienten c_2, a_{21}, b_1, b_2
Also ein nicht-lineares unterbestimmtes GLS, in der Regel schwer zu lösen.

Mögliche Lösungen:

Heun

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

Modifizierter Euler

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

Heun:

$$y^h_n = y^h_{n-1} + \frac{h_n}{2} \underbrace{f(t_{n-1}, y^h_{n-1})}_{k_1} + \frac{h_n}{2} \overbrace{f(t_n, y_{n-1} + h_n \underbrace{f(t_{n-1}, y^h_{n-1})}_{k_1})}^{k_2}$$

Modifizierter Euler:

$$y^h_n = y^h_{n-1} + h_n f(t_{n-1} + \frac{h_n}{2}, y^h_{n-1} + \frac{h_n}{2} f(t_{n-1}, y_{n-1}))$$

s=3

8 Parameter, 6 Gleichungen:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & a_{21} & 0 & 0 \\ c_3 & a_{31} & a_{31} & 0 \\ \hline & b_1 & b_2 & b_3 \end{array}$$

Heun 3. Ordnung:

$$\begin{array}{c|ccc} \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

Runge-Kutta 3. Ordnung:

$$\begin{array}{c|ccc} \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array}$$

s=4

13 Parameter, 11 Gleichungen
Runge-Kutta 4. Ordnung:

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

Bemerkung 3.3

- Taylormethode auch auf Systeme übertragbar
- RK-Ansatz bedingt auf Systeme übertragbar
  Man stellt fest [*R*]
  - $m \leq 4 \rightarrow$ Ordnung überträgt sich auf Systeme
  - $m > 4 \rightarrow$ Ordnung in der Regel reduziert
- maximal erreichbare Ordnung für explizite, s-stufige RK-Verfahren (skalar):

s	1	2	3	4	5	6	7	8
m	1	2	3	4	4	5	6	6

  Daher ist RK-4 besonders beliebt
- Konstruktion der Verfahren
  - Butcher-Bäume (systematische Darstellung der Ableitungen)
  - Computer-Algebra-Systeme
- Implizite Verfahren (zwei Beispiele)
  - Trapezregel: $y_n^h = y_{n-1}^h + \frac{1}{2}h_n \left\{ f(t_{n-1}, y_{n-1}^h) + f(t_n, y_n^h) \right\}$
  - Mittelpunktsregel: $y_n^h = y_{n-1}^h + h_n f(t_n + \frac{h_n}{2}, \frac{1}{2}(y_{n-1} + y_n^h))$ (eigentlich kein RK-Schema)

  beide $m = 2$.

3.3 Konvergenz allgemeiner Einschrittverfahren

ginge im Prinzip genauso wie beim expliziten/impliziten Eulerverfahren. Hier anderer Weg

Satz 3.4
Sei

$$y_n^h = y_{n-1}^h + h_n F(h_n, t_{n-1}, y_{n-1}^h, y_n^h) \qquad n \geq 0 \qquad y_0^h = y_0$$

ein Einschrittverfahren zur Lösung der AWA:

$$u'(t) = f(t, u(t)) \qquad t \geq 0 \qquad u(t_0) = u_0$$

$f(t, x)$ erfülle die Lipschitzbedingung mit Konstante L, das Verfahren sei konsistent mit Ordnung m:

$$\|\tau_n^h\| \leq C \cdot h^m \qquad (h \to 0)$$

Im *impliziten* Fall gelte für F die Lipschitzbedingung

$$\|F(h, t, x, \tilde{x}) - F(h, t, y, \tilde{y})\| \leq \tilde{L}(\|x - y\| + \|\tilde{x} - \tilde{y}\|)$$

sowie die Schrittweitenbedingung

$$h\tilde{L} \leq \frac{1}{2}$$

Dann gilt für den globalen Fehler

$$\max_{1 \leq n \leq N} \|e_n^h\| \leq e^{LT} \left\{ \|u_0 - y_0\| + \frac{\alpha C}{K} h^m \right\}$$

mit $\alpha = 1$ für explizite und $\alpha = 2$ für implizite Verfahren.

Beweis .
Nach [Bernd Simeon, Vorlesungsskript, TU München]:

(i) Für $0 \leq n \leq N$ sei $u_n(t)$ die Lösung der AWA (nicht zu verwechseln mit u_n^h!)

$$u_n'(t) = f(t, u_n(t)) \qquad t \geq t_n \qquad u(t_n) = y_n^h$$

für den Fehler $e_n^h = u(t_n) - y_n^h$ gilt:

$$\begin{aligned} e_n^h &= u(t_n) - y_n^h \\ &= \underbrace{u(t_n) - u_0(t_n)} + \underbrace{u_0(t_n) - u_1(t_n)} + \underbrace{u_1(t_n) - u_2(t_n)} + \ldots + \underbrace{u_{n-2}(t_n) - u_{n-1}(t_n)} + \underbrace{u_{n-1}(t_n) - y_n^h} \\ &= u(t_n) - u_0(t_n) + \sum_{i=0}^{n-2} (u_i(t_n) - u_{i+1}(t_n)) + u_{n-1}(t_n) - y_n^h \end{aligned}$$

Normbilden, Dreiecksungleichung

$$\begin{aligned} \|e_n^h\| &\leq \|u(t_n) - u_0(t_n)\| + \sum_{i=0}^{n-2} \|u_i(t_n) - u_{i+1}(t_n)\| + \|u_{n-1}(t_n) - y_n^h\| \\ &\underset{\text{Stabilitätssatz}}{\leq} e^{L(t_n - t_0)} \|u_0 - y_0\| + \sum_{i=0}^{n-2} e^{L(t_n - t_{i+1})} \|u_i(t_{i+1}) - y_{i+1}^h\| + \|u_{n-1}(t_n) - y_n^h\| \end{aligned}$$

(Graph zur Visualisierung des Fehlers)

(ii)

$$h_{i+1}\Big\{u_i(t_{i+1}) - \underbrace{u_i(t_i)}_{y_i^h}\Big\} - F(h_{i+1}, t_i, \underbrace{u_i(t_i)}_{y_i^h}, u_i(t_{i+1})) = \tau_{i+1}^h$$

$$\Leftrightarrow u_i(t_{i+1}) - y_{i+1}^h + \underbrace{\cancel{y_i^h} + h_{i+1}F(h_{i+1}, t_i, y_i^h, y_{i+1}^h)}_{y_{i+1}^h} - \cancel{y_i^h} - h_{i+1}F(h_{i+1}, t_i, u_i(t_i), u_i(t_{i+1})) = h_{i+1}\tau_{i+1}^h$$

$$\Leftrightarrow u_i(t_{i+1}) - y_{i+1}^h = h_{i+1}\tau_{i+1}^h + h_{i+1}\left(F(h_{i+1}, t_i, y_i^h, u_i(t_{i+1})) - F(h_{i+1}, t_i, y_i^h, y_{i+1}^h)\right)$$

und damit

$$\|u_i(t_{i+1}) - y_{i+1}^h\| \le h_{i+1}\|\tau_{i+1}^h\| + \begin{cases} 0 & \text{explizit} \\ h_{i+1}\tilde{L}\|u_i(t_{i+1}) - y_{i+1}^h\| & \text{implizit} \end{cases}$$

Im impliziten Fall gilt mit $h_{i+1}\tilde{L} \le \frac{1}{2}$

$$\|u_i(t_{i+1}) - y_{i+1}^h\| \le \frac{h_{i+1}}{1 - h_{i+1}\tilde{L}}\|\tau_{i+1}^h\| \le 2h_{i+1}\|\tau_{i+1}^h\|$$

(iii) weiter i): und $\alpha = \begin{cases} 1 & \text{explizit} \\ 2 & \text{implizit} \end{cases}$:

$$\begin{aligned}
\|e_n^h\| &\le e^{L(t_n - t_0)}\|u_0 - y_0\| + \sum_{i=0}^{n-2}\left(e^{L(t_n - t_{i+1})}\alpha h_{i+1}\|\tau_{i+1}^h\|\right) + \alpha h_n\|\tau_n^h\| \\
&= e^{L(t_n - t_0)}\|u_0 - y_0\| + \alpha\sum_{i=0}^{n-1} e^{L(t_n - t_{i+1})}h_{i+1}\|\tau_{i+1}^h\| \\
&\underset{\text{Konsistenz } \|\tau_{i+1}^h\| \le Ch^m}{\le} e^{L(t_n - t_0)}\|u_0 - y_0\| + \alpha \cdot C \cdot h^m \cdot \underbrace{\sum_{i=0}^{n-1} h_{i+1}e^{L(t_n - t_{i+1})}}_{\text{nicht vorhande Zeichnung}} \\
&\le e^{L(t_n - t_0)}\|u_0 - y_0\| + \alpha \cdot C \cdot h^m \cdot \int_{t_0}^{t_n} e^{L(t_n - t)}\,\mathrm{d}t \\
&= e^{L(t_n - t_0)}\|u_0 - y_0\| + \alpha \cdot (h^m\frac{1}{L}(e^{L(t_n - t_0)} - 1) \\
&= e^{L(t_n - t_0)}\left\{\|u_0 - y_0\| + \frac{\alpha \cdot C}{L} \cdot h^m\right\}
\end{aligned}$$

Maximum bilden → fertig

□

Bemerkungen

- Bei Einschrittverfahren folgt aus Konsistenz sofort Konvergenz allgemeiner Einschrittverfahren
- Ordnung bleibt erhalten
- Im expliziten Fall genügt Stabilität der AWA

Bemerkung zum impliziten Fall

$$h\tilde{L} \leq \frac{1}{2} \text{ notwendig?}$$

$$u_i(t_{i+1}) - y^h_{i+1} = h_{i+1}\tau^h_{i+1} + h_{i+1}\left(F(h_{i+1}, t_i, y^h_i, u_i(t_{i+1})) - F(h_{i+1}, t_i, y^h_i, y^h_{i+1})\right) \qquad (\star)$$

- **Lösbarkeit im impliziten Verfahren**

  $$y^h_n = y^h_{n-1} + h_n F(h_n, t_{n-1}, y^h_{n-1}, y^h_n) =: g_n(y^h_n)$$

  Lösung der nichtlinearen Gleichung mit Fixpunktiteration

  $$\begin{aligned}\|g_n(x) - g_n(y)\| &= \|h_n F(\ldots, x) - h_n F(\ldots, y)\| \\ &\leq h_n\tilde{L}\|x - y\|\end{aligned}$$

  d.h. wenn $h_n\tilde{L} < 1 \Rightarrow g_n$ ist Kontraktion $\Rightarrow$ Konvergenz der Fixpunktiteration
- spezielle Wahl
  - impliziter Euler
  - $f(t, x) = Ax + b$ mit A negativ definit

  Einsetzen in ($\star$):

  $$\begin{aligned}u_i(t_{i+1}) - y^h_{i+1} &= h_{i+1}\tau^h_{i+1} + h_{i+1}\left(\underbrace{Au_i(t_{i+1}) + b}_{\text{impliziter Euler}} - \underbrace{Ay^h_{i+1} - b}\right) \\ &= h_{i+1}\tau^h_{i+1} + h_{i+1}A(u_i(t_{i+1}) - y^h_{i+1}) \\ \|u_i(t_{i+1}) - y^h_{i+1}\| &\leq \left\|(I - h_{i+1}A)^{-1}\right\| h_{i+1}\|\tau^h_{i+1}\|\end{aligned}$$

  A negativ definit $\Rightarrow$ alle EW von $I - hA$ größer 1 $\Rightarrow$ alle EW von $(I - hA)^{-1}$ kleiner 1 $\Rightarrow$ $\|(I - h_{i+1}A)^{-1}\| \leq 1$ (Spektralnorm).
  $\Rightarrow$ Schrittweitenbedingung ist hinreichend, aber evtl. nicht notwendig

3.4 Schrittweitensteuerung

Bisher: h_n fest (äquidistant), nun h_n variabel
Ziel:

1) Erreiche vorgegebene Genauigkeit $\max_{1\leq n\leq N} \|u(t_n) - y^h_n\| \leq \varepsilon$ wobei man sich das ε selbst vorgibt

2) mit möglichst wenig Rechenaufwand

Ideen:

- A priori Fehlerabschätzung

  - Stabilitätsfaktor e^{LT} (möglicherweise sehr pessimistisch, was ein sehr kleines h und damit viel Rechenaufwand zur Folge hätte)
  - höhere Ableitung $u^{(n+1)}(\xi)$
    $\rightsquigarrow$ höhere Ableitung von f

  $\Rightarrow$ dieser Weg funktioniert nicht bzw. ist zu aufwendig
- Praxis: Steuerung über „lokalen Fehler“

$$\max_{t_n \in I} \|e_n^h\| \leq e^{LT} \sum_{n-1}^{N} h_n \underbrace{\|\tau_n^h\|}$$

$\|\tau_n^h\|$ wird „**a posteriori**“ aus der numerischen Lösung geschätzt.

Zurück zum lokalen Fehler (explizite Verfahren):

$$\begin{aligned}
h_n\tau_n^h &= u(t_n) - u(t_{n-1}) - h_n F\left(h_n, t, u(t_{n-1})\right) \\
&= \left[\cancel{u(t_{n-1})} + h_n \cancel{\sum_{i=1}^{m} \frac{h_n^{i-1}}{i!} f^{(i-1)}(t_{n-1}, u(t_{n-1}))} + c(t_{n-1})h_n^{m+1} + O(h_n^{m+2})\right] \\
&\quad - \cancel{u(t_{n-1})} - h_n\Bigg[\underbrace{\cancel{\sum_{i=1}^{m} \frac{h_n^{i-1}}{i!} f^{(i-1)}(t_{n-1}, u(t_{n-1}))}}_{\text{Taylor-Verfahren } m \text{ Stufen}} + \underbrace{\tilde{c}(t_{n-1})h_n^m + O(h_m^{m+1})}_{\substack{\text{Zusätzlich bei RK} \\ \text{Verfahren der Ordnung } m}}\Bigg] \\
&= \underbrace{C(t_{n-1})}_{\substack{\text{Funktion ist unabhängig von } h}} h_n^{m+1} + O(h_n^{m+2})
\end{aligned}$$

Also:

$$\begin{aligned}
u(t_n) &= u(t_{n-1}) + h_n F(h_n, t_{n-1}, u(t_{n-1})) + h_n\tau_n^h \\
y_n^h &= y_{n-1}^h + h_n F(h_n, t_{n-1}, y_{n-1}^h)
\end{aligned}$$

Sei $\tilde{u}(t)$ die Lösung zum Startwert $\tilde{u}(t_{n-1}) = y_{n-1}^h$, so gilt nach einem Schritt:

$$\tilde{u}(t_n) - y_n^h = h_n\tau_n^h = C(t_{n-1})h_n^{m+1} + O(h_n^{m+2})$$

Fehlerabschätzung mit Verfahren unterschiedlicher Ordnung

y_n^h werde erzeugt mit Verfahren der Ordnung m
$\hat{y}_n^h$ werde erzeugt mit Verfahren der Ordnung $m+1$
Dann gilt nach *einem* Schritt mit Startwert $\hat{y}_n^h$

$$\begin{aligned}
\hat{y}_n^h - y_n^h &= \tilde{u}(t_n) - \hat{C}(t_{n-1})h_n^{m+2} + O(h_n^{m+3}) - (\cancel{\tilde{u}(t_n)} - C(t_{n-1})h_n^{m+1} + O(h_n^{m+2})) \\
&= C(t_{n-1})h_n^{m+1} + O(h_n^{m+2}) \doteq C(t_{n-1})h_n^{m+1}
\end{aligned}$$

Für die „optimale“ Schrittweite sollte gelten:

$$\|\tilde{u}(t_n) - y_n^h\| \doteq \|C(t_{n-1})\|h_{\text{opt}}^{m+1} = \mathsf{TOL} \text{ (=vorgegebene Toleranz)}$$

Für $C(t_{n-1})$ gilt:

$$\|C(t_{n-1})\| \doteq \frac{\|\hat{y}_n^h - y_n^h\|}{h_n^{m+1}}$$

und damit

$$\boxed{h_{\text{opt}} = h_n \left(\frac{\mathsf{TOL}}{\|\hat{y}_n^h - y_n^h\|} \right)^{\frac{1}{m+1}}}$$

Algorithmus:
Eingabe: y_{n-1}, Vorschlag für h_n
Ausgabe: y_n, so dass geschätzter Fehler $\leq \mathsf{TOL}$, Vorschlag für h_{n+1}

1) Berechne mit h_n die Näherungen $y_n^h, \hat{y}_n^h$ (Konsistenzordnung $m, m+1$)

2) Berechne

$$h_{\text{opt}} = h_n \cdot \left(\frac{\varrho \cdot \mathsf{TOL}}{\|\hat{y}_n^h - y_n^h\|} \right)^{\frac{1}{m+1}}$$

$$h_{\text{opt}} = \min(\beta h_n, \max(\alpha h_n, h_{\text{opt}})) \qquad 0 < \varrho \leq 1 \text{ (Sicherheitsfaktor)}$$

$$\alpha < 1 < \beta \qquad \text{Schranken für Änderung in einem Schritt, z.B. } \alpha = \frac{1}{4}\ \beta = 4$$

3) Falls $\|\hat{y}_n^h - y_n^h\| \leq \mathsf{TOL}$. Setze $h_{n+1} = h_{\text{opt}}$, $y_n = \hat{y}_n^h$. Sonst $h_n = h_{\text{opt}}$ und gehe nach 1)

Frage: welche Verfahren sind zu wählen, um den lokalen Fehler bzw. die Schrittweite möglichst effektiv zu bestimmen?

Eingebettete Runge-Kutta Verfahren

C	A	
	$b \rightarrow y_n^h = y_{n-1}^h + h_n \sum_{j=1}^{s-1} b_j k_j$	Ordnung m
	$\hat{b} \rightarrow \hat{y}_n^h = y_{n-1}^h + h_n \sum_{j=1}^{s} \hat{b}_j k_j$	Ordnung $m+1$

Runge-Kutta-Fehlberg Methode 4(5)

$s = 6 \qquad m = 4 \qquad m + 1 = 5$

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{72}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{24}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{4}$	
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Beispiel 2-Körper-Problem

$$G = 1 \qquad m_1 = 1 \qquad m_2 = 0.01$$
$$x_1 = (-1, 0)^T \qquad v_1 = 0$$
$$x_2 = (1, 0)^T \qquad v_2 = (0, 1/5)$$

Heun 2 (äquidistant)	$\lvert e_0 - e_N\rvert / \lvert e_0\rvert$	$\#f$ Ausw.
	$5.7 \cdot 10^{-1}$	51200
	$2.3 \cdot 10^{-3}$	409600
$RK4$	$2.7 \cdot 10^{-3}$	102400
	$2.7 \cdot 10^{-6}$	409600
$RKF45$	$4.4 \cdot 10^{-2}$	3282
	$2.8 \cdot 10^{-6}$	16542

Einbeziehung des globalen Fehlers

Ziel:

1) Fehlerkontrolle: $\max_{t_n \in I} \|e_n^h\| \leq \mathsf{TOL}$

2) Effizienz: Wähle h_n so dass 1) mit möglichst geringem Aufwand erreicht wird

Ausgangspunkt:

$$\max_{t_n \in I} \|e_n^h\| \leq e^{LT} \sum_{n=1}^{N} h_n \|\tau_n^h\|$$

Oben wurde auch gezeigt:

$$\tau_n^h = \underbrace{\tau^m(t_n)}_{\text{hieß oben } C(t) \to \text{Taylormethode: } \tau^m(t_n) = \frac{1}{(m+1)!} u^{(m+1)}(t_{n-1})} h_n^m + O(h_n^{m+1})$$

Stabilitätsfaktor $K := e^{LT}$ sehr pessimistisch. Im folgenden sei K als fest und bekannt vorausgesetzt.

Strategie 1

Idee: Verteile Fehler gleichmäßig auf das Intervall $[t_0, t_0 + T]$ Wähle

$$h_n \approx \left(\frac{\mathsf{TOL}}{KT \|\tau^m(t_n)\|} \right)^{1/m}$$

$$\max_{t_n \in I} \|e_n^h\| \leq K \sum_{n=1}^{N} h_n \|\tau_n^h\| \doteq K \sum_{n=1}^{N} h_n h_n^m \|\tau^m(t_n)\|$$
$$\approx \cancel{K} \sum_{n=1}^{N} h_n \frac{\mathsf{TOL}}{\cancel{KT \|\tau^m(t_n)\|}} \cancel{\|\tau^m(t_n)\|} = \frac{\mathsf{TOL}}{\cancel{T}} \sum_{n=1}^{N} \cancel{h_n} = \mathsf{TOL}$$

$$N = \sum_{n=1}^{N} h_n h_n^{-1} = \sum_{n=1}^{N} h_n \left(\frac{KT \|\tau^m(t_n)\|}{\mathsf{TOL}} \right)^{1/m}$$
$$= \left(\frac{KT}{\mathsf{TOL}} \right)^{1/m} \underbrace{\sum_{n=1}^{N} h_n \|\tau^m(t_n)\|^{1/m}} \overset{\text{Taylor-Methode}}{\approx} \left(\frac{KT}{\mathsf{TOL}\,(m+1)!} \right)^{1/m} \int_I \|u^{(m+1)}\|^{1/m} \, \mathrm{d}t$$

Strategie 2

Verteile Fehler gleichmäßig auf die Schritte $1 \dots N$. Wähle

$$h_n \approx \left(\frac{\mathsf{TOL}}{KN \|\tau^m(t_n)\|} \right)^{1/(m+1)}$$

N ist a priori nicht bekannt und muss iterativ bestimmt werden.

$$\max_{t_n \in I} \|e_n^h\| \leq K \sum_{n=1}^{N} h_n^m \|\tau(t_n)\| \doteq K \sum_{n=1}^{N} h_n^{m+1} \|\tau^m(t_n)\| = \cancel{K} \sum_{n=1}^{N} \frac{\mathsf{TOL}}{\cancel{KN \|\tau^m(t_n)\|}} \cancel{\|\tau^m(t_n)\|} = \mathsf{TOL}$$

$$N \approx \sum_{n=1}^{N} h_n \left(\frac{KN \|\tau^m(t_n)\|}{\mathsf{TOL}} \right)^{\frac{1}{m+1}} = \left(\frac{KN}{\mathsf{TOL}} \right)^{\frac{1}{m+1}} \sum_{n=1}^{N} h_n \|\tau^m(t_n)\|^{\frac{1}{m+1}}$$

$$N^{\frac{m}{m+1}} = \left(\frac{K}{\mathsf{TOL}\,(m+1)!} \right)^{\frac{1}{m+1}} \underbrace{\sum_{n=1}^{N} h_n \|u^{(m+1)}(t_{n-1})\|^{\frac{1}{m+1}}}_{} \approx \left(\frac{K}{\mathsf{TOL}\,(m+1)!} \right)^{\frac{1}{m+1}} \int_I \|u^{(m+1)}(t)\|^{\frac{1}{m+1}} \, \mathrm{d}t$$

$$N \approx \left(\frac{K}{\mathsf{TOL}\,(m+1)!} \right)^{\frac{1}{m}} \left(\int_I \|u^{(m+1)}\|^{\frac{1}{m+1}} \, \mathrm{d}t \right)^{\frac{m+1}{m}}$$

Vergleich der Strategien

- Beide liefern $N \sim \mathsf{TOL}^{-1/m}$
- Strategie 2 aufwändiger
- Betrachte $m = 1$ (expliziter Euler)

$$\text{Strategie 1: } N \approx \frac{KT}{2\mathsf{TOL}} \int_I \|u''(t)\| \, \mathrm{d}t$$

$$\text{Strategie 2: } N \approx \frac{K}{2\mathsf{TOL}} \left(\int_I \|u''(t)\|^{1/2} \, \mathrm{d}t \right)^2$$

Wende Cauchy-Schwarz an:

$$\int_I \|u''(t)\|^{\frac{1}{2}} \cdot 1 \, \mathrm{d}t \leq \left(\int_I \|u''(t)\| \, \mathrm{d}t \right)^{\frac{1}{2}} \cdot \left(\int_I 1 \, \mathrm{d}t \right)^{\frac{1}{2}}$$

$$\Rightarrow \quad \left(\int_I \|u''(t)\|^{\frac{1}{2}} \, \mathrm{d}t \right)^2 \leq \int_I \|u''(t)\| \, \mathrm{d}t \cdot T$$

$$\Rightarrow \quad N_2 \leq N_1$$

Strategie 2 effizienter, falls

$$\left(\int_I \|u''(t)\|^{\frac{1}{2}} \, \mathrm{d}t \right)^2 \ll \int \|u''(t)\| \, \mathrm{d}t \cdot T,$$

z.B. bei Singularität.

Richardson-Extrapolation zur Schätzung von $\tau^m(t_n)$

Idee: (Graph und so)

- Berechne y_n^H mit einem Schritt eines Verfahrens (der Konsistenzordnung m) der Länge H
- Berechne $y_n^{H/2}$ mit zwei Schritten desselben Verfahrens der Länge $\frac{H}{2}$
- Schätze Fehler aus $y_n^H - y_n^{H/2}$

$$\begin{aligned} u(t_n) &= u(t_{n-1}) + h_n F(h_n, t_{n-1}, u(t_{n-1})) + h_n \tau_n^h \\ y_n^h &= y_{n-1}^h + h_n F(h_n, t_{n-1}, y_{n-1}^h) \end{aligned}$$

für Schrittweite H folgt:

$$\begin{aligned} y_n^H - u(t_n) &= e_{n-1} + H \cdot (F(H, t_{n-1}, \underbrace{y_{n-1}}_{=u(t_{n-1})+\underbrace{y_{n-1}-u(t_{n-1})}}) - F(H, t_{n-1}, u(t_{n-1})) - H\tau_n^H \\ &= e_{n-1} + H \cdot (\cancel{F(H, t_{n-1}, u(t_{n-1}))} + e_{n-1} F_x(H, t_{n-1}, \xi) - \cancel{F(H, t_{n-1}, u(t_{n-1}))}) - H\tau_n^H \\ &= (1 + O(H)) e_{n-1} - H\tau^m(t_n) + O(H^{m+2}) \end{aligned}$$

Schrittweite $\frac{H}{2}$

$$\begin{aligned} e^{\frac{H}{2}}_{n-\frac{1}{2}} &= (1 + O(H)) e_{n-1} - \left(\frac{H}{2}\right)^{m+1} \tau^m\left(t_n - \frac{H}{2}\right) + O(H^{m+2}) \\ e_n^{\frac{H}{2}} &= y_n^{\frac{H}{2}} - u(t_n) = e^{\frac{H}{2}}_{n-\frac{1}{2}} + \frac{H}{2}\left(F\left(\frac{H}{2}, t_{n-\frac{1}{2}}, y^{\frac{H}{2}}_{n-\frac{1}{2}}\right) - F\left(\frac{H}{2}, t_{n-\frac{1}{2}}, u\left(t_{n-1} + \frac{H}{2}\right)\right)\right) - \frac{H}{2}\tau_n^{\frac{H}{2}} \\ &= (1 + O(H)) e^{\frac{H}{2}}_{n-\frac{1}{2}} - \left(\frac{H}{2}\right)^{m+1} \tau^m(t_n) + O(H^{m+2}) \\ &= (1 + O(H)) e_{n-1} - 2\left(\frac{H}{2}\right)^{m+1} \tau^m(t_n) + O(H^{m+2}) \end{aligned}$$

Differenz bilden

$$\begin{aligned} y_n^{\frac{H}{2}} - y_n^H &= \underbrace{y_n^{\frac{H}{2}} - u(t_n)} - \underbrace{y_n^H - u(t_n))} = e_n^{\frac{H}{2}} - e_n^H \\ &= O(H) e_{n-1} - \left(2\left(\frac{H}{2}\right)^{m+1} - H^{m+1}\right)\tau^m(t_n) + O(H^{m+2}) \\ \|\tau^m(t_n)\| &= \frac{\|y_n^{\frac{H}{2}} - y_n^H\|}{H^{m+1}(1 - 2^{-m})} + \|e_{n-1}\| O(H^{-m}) + O(H) \end{aligned}$$

- H sei so klein, dass erster Summand den Dritten dominiert
- eine Möglichkeit: Nehme an, dass $\|e_{n-1}\| = 0$ (τ^m ist lokaler Fehler nach einem Schritt)
- Alternativ: $\exists y_n^H - u(t_n) = a^m(t_n) H^m + O(H^{m+1})$ (Asymptotische Entwicklung des *globalen* Fehlers) dann ist
$$\tilde{y}_n := \frac{2^m y_n^{H/2} - y_n^H}{2^m - 1}$$
eine Näherung mit Fehler $O(H^{m+1})$
- $\exists$ sogenannte „Extrapolationsverfahren“, die dieses Prinzip weiter nutzen

Adaptiver Algorithmus

Eingabe: y_{n-1}, letzte Schrittweite h_{n-1}, t_{n-1}
Ausgabe: y_n, berechnet mit „optimalen“ h_n, t_n

1) $h_n = h_{n-1}$

2) $H = 2h_n$, berechne y_n^H und $y_n^{H/2}$ (3 Schritte insgesamt)

3) $\alpha < 1 < \beta$

$$\tau_n^m = \frac{\|y_n^{H/2} - y_n^H\|}{H^{m+1}(1-2^{-m})} \qquad h_{\text{opt}} = \left(\frac{\delta \cdot TOL}{KT\tau_n^m}\right)^{\frac{1}{m}} \qquad h_{\text{opt}} = \min(\beta h_n, \max(\alpha h_n, h_{\text{opt}}))$$

4) Falls $h_n \leq h_{\text{opt}}$ (akzeptiere Schritt)

$$t_n = t_{n-1} + H \qquad y_n = \frac{(2^m y_n^{\frac{H}{2}} - y_n^H)}{2^m - 1} \qquad h_n = h_{\text{opt}}$$

sonst:

$$h_n = h_{\text{opt}} \qquad \text{und gehe nach 2)}$$

Vorteil: Sofort auf alle Verfahren anwendbar (auch implizite)
Nachteil: 50% mehr Aufwand im Vergleich zu ohne Fehlerabschätzung. Dafür aber eine Ordnung besser.
Vergleich mit RK45: 12 f-Auswertungen bei Ordnung 5
Extrapolation + RK4: 12 f-Auswertungen bei Ordnung 5
Anwendung auf 2-Körper-Problem:

	$\|e_0-e_N\|/\|e_N\|$	f-Auswertungen
$RK4$	$2.7 \cdot 10^{-6}$	409600
$RK45$	$2.8 \cdot 10^{-6}$	16542
$RK4$-Extrapolation	$2.7 \cdot 10^{-6}$	47316

4 Numerik steifer Differentialgleichungen

4.1 Motivation

Beispiel 4.1 Van der Pol Oszillator

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -u_2 \\ \frac{1}{\varepsilon}(u_1 - \frac{u_2^3}{3} + u_2) \end{pmatrix} \qquad t \geq 0 \qquad u(0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

RK45 kann keine großen Schrittweiten wählen, auch wenn TOL sehr groß gewählt wird, obwohl Lösung sehr glatt für $t \in [1, 1.7]$.

Beispiel 4.2 Modellproblem

$$u' = \lambda u \qquad \mathbb{R} \ni \lambda < 0$$

	$\lambda = -10$	$\lambda = -100$
expliziter Euler	$h = 0.1 \rightarrow 1, 0, 0, \cdots$ $h = 0.2 \rightarrow 1, -1, 1, -1, \cdots$ $h > 0.2 \rightarrow$ Divergenz	$h = 0.01 \rightarrow 1, 0, 0, \cdots$ $h = 0.02 \rightarrow 1, -1, 1, -1, \cdots$ $h > 0.02 \rightarrow$ Divergenz
impliziter Euler	„stabil“ für alle $h > 0$	ebenfalls

Analyse von EE/IE für das simple Modellproblem.

Expliziter Euler:

$$y_n^h = y_{n-1}^h + h\lambda y_{n-1}^h = (1 + h\lambda) y_{n-1}^h$$

also

$$y_n^h = (1 + h\lambda)^n u_0$$

also

$$|y_n^h| = \underbrace{|1 + h\lambda|^n}_{\text{beschränkt } |1+h\lambda| \leq 1} |u_0|$$

für $\lambda < 0$ gilt demnach

$$-1 \leq 1 + h\lambda \leq 1 \text{ trivial}$$
$$\Leftrightarrow \boxed{-\frac{2}{\lambda} \geq h}$$

wegen

$$|f(t,x)| - f(t,y)| = |\lambda(x-y)| = \underbrace{|\lambda|}_{=:L} |x-y|$$

d.h. unsere Bedingung können wir schreiben als

$$\boxed{hL \leq 2}$$

Dies ist eine *notwendige* Bedingung für die Beschränktheit des EE-Verfahrens!

Impliziter Euler

$$\frac{y_n^h - y_{n-1}^h}{h} = \lambda y_n^h \Leftrightarrow (1 - h\lambda) y_n^h = y_{n-1}^h \Leftrightarrow y_n^h = \frac{1}{1 - h\lambda} y_{n-1}^h$$

also

$$y_n^h = \left(\frac{1}{1 - h\lambda}\right)^n u_0$$

d.h. Lösung beschränkt, falls

$$\left|\frac{1}{1 - h\lambda}\right| \leq 1 \Leftarrow |1 - h\lambda| \text{ ist für alle } h > 0 \text{ und } \lambda < 0 \text{ erfüllt}$$

Beispiel 4.3 Simeon

$$u'(t) = \lambda(u(t) - \underbrace{\varphi(t)}_{\text{gegebene stetige Funktion}}) + \varphi'(t) \qquad t \in [t_0, t_0 + T] \qquad u(t_0) = u_0$$

hat die Lösung

$$u(t) = (u_0 - \varphi(t)) e^{\lambda(t - t_0)} + \varphi(t)$$

(Graph zum Beispiel)

Wann ist eine AWA „steif“? $\exists$ verschiedene Definitionen:

1) Wenn explizite Verfahren kleine Zeitschritte einsetzen müssen, *obwohl* sich die Lösung kaum ändert, (ausgewählte) implizite Verfahren jedoch große Schritte einsetzen können

2) (auch im Rannacher) Ein lineares System $u' = Au$ mit $\mathrm{Re}(\lambda_i) < 0$ für alle EW λ_i heißt steif, wenn
$$\frac{\max_i |\mathrm{Re}(\lambda_i)|}{\min_i |\mathrm{Re}(\lambda_i)|} \gg 1$$
(im nichtlinearen Fall nehme $f_x(t, u(t))$)

3) Differentialgleichungen der Form

$u' = f(t, u, z)$	mit $\varepsilon \in \mathbb{R}$. $\varepsilon \ll 1$ heißen *singulär gestört*
$\varepsilon' = f(t, u, z)$	und führen für $\varepsilon \to 0$ auf steife AWAn

4.2 (Skalare, lineare) Modellprobleme

$$u'(t) = \lambda u(t) \qquad t \geq 0 \qquad u(0) = u_0 \qquad \lambda \in \mathbb{C} \qquad \mathrm{Re}(\lambda) \leq 0 \tag{4.1}$$

dient nun als Modellproblem zur Bewertung unserer Verfahren.

Definition 4.4 absolute Stabilität
Eine Einschrittmethode heißt „absolut stabil“ für ein $h\lambda \neq 0$, wenn sie angewandt auf das Modellproblem (4.1) beschränkte Näherungen

$$\sup_{n \geq 0} |y_n^h| < \infty$$

erzeugt.

expliziter Euler

Für den expliziten Euler gilt:

$$y_n^h = \underbrace{(1 + h\lambda)}_{=:\omega(h\lambda)} y_{n-1}^h = \omega(h\lambda) y_{n-1}^h$$

Expliziter Euler ist absolut stabil für

$$|\omega(\underbrace{h\lambda}_{z})| \leq 1.$$

Die Menge

$$\mathsf{SG} = \{z = h\lambda \in \mathbb{C} \mid |\omega(z)| \leq 1\}$$

heißt *Stabilitätsgebiet* einer Einschrittformel.
SG des expliziten Euler (EE): $\mathsf{SG} = \{z \in \mathbb{C} \mid |1 + z| \leq 1\}$
(Zeichnung des Kreises in der komplexenen Zahlenebene)
Für das Taylor-Verfahren der Stufe R gilt:

$$\begin{aligned}
y_n^h &= y_{n-1}^h + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t_{n-1}, y_{n-1}^h) \\
&= y_{n-1}^h + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} \lambda^r y_{n-1}^h && \begin{matrix} f(t,u(t))=\lambda u(t) \\ f^{(1)}(t,u(t))=\lambda u'(t)=\lambda^2 u(t) \end{matrix} \\
&= y_{n-1}^h + \sum_{r=1}^{R} \frac{(h\lambda)^r}{r!} y_{n-1}^h \\
&= \underbrace{\left(\sum_{r=0}^{R} \frac{(h\lambda)^r}{r!} \right)}_{\omega(h\lambda)} y_{n-1}^h \\
\Rightarrow \mathsf{SG}_R &= \left\{ z \in \mathbb{C} \;\middle|\; \left| \sum_{r=0}^{R} \frac{z^r}{r!} \right| \leq 1 \right\}
\end{aligned}$$

- (Zeichnung der Stabilitätsgebiete für das Taylor-Verfahren), siehe z.B. in diesem PDF aus Wien
- größeres R, größeres SG.
- ab $R = 2$ sind Teil der imaginären Achse dabei

Stabilitätsintervall:

$$SI = \{z \in \mathbb{R} \mid |\omega(z)| \leq 1\}$$

$$SI_{\text{Expliziter Euler}} = \begin{cases} [-2, 0] & R = 1 \\ [-2, 0] & R = 2 \\ [-2.51\ldots, 0] & R = 3 \\ [-2.78\ldots, 0] & R = 4 \end{cases}$$

Runge-Kutta Verfahren

$$y_n^h = y_{n-1}^h + h \underbrace{\left[\sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t_{n-1}, y_{n-1}^h) + O(h^R) \right]}_{=:F(h,t_{n-1},y_{n-1}^h)}$$

Andererseits gilt

$$k_1 = f(t,u) = \lambda u \qquad k_i = f\left(t - h + c_i h, u + h \sum_{j=1}^{i-1} a_{ij} k_j \right) = \lambda u + \lambda h \sum_{j=1}^{i-1} a_{ij} k_j$$

$\Rightarrow k_i$ ist ein *Polynom in h vom Grad* $i-1$.

Damit ist $F(h,t,u) = \sum_{i=1}^{S} b_i k_i$ ein Polynom vom Grad $s-1$.

Für $s = R$ gilt damit

1) $F(h,t,u)$ ist Polynom vom Grad $R-1$ in h

2)
$$F(h,t,u) = \underbrace{\sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t,u)}_{\text{Polynom in } h \text{ Grad } R-1} + \underbrace{O(h^R)}_{C_R h^R + C_{R+1} h^{R+1} + \ldots}$$

$\Rightarrow C_R = C_{R+1} = \ldots = 0!$

Bis $R = 4$ stimmen SG von Taylor-Verfahren und Runge-Kutta Verfahren (maximaler Konsistenzordnung) überein.

3) Für $s > 4$ können Freiheiten bei der Definition der Koeffizienten des RK-Verfahrens zur Optimierung des SG verwendet werden.

Für die Taylor- und explizten RK-Verfahren ist bei Anwendung auf das Testproblem (4.1) eine Schrittweitenbedingung einzuhalten.

Definition 4.5 A-Stabilität
Ein Einschrittverfahren heißt „A-stabil“, wenn das zugehörige Stabilitätsgebiet die ganze linke Halbebene umfasst:

$$\mathbb{C}^- = \{ z \in \mathbb{C} \mid \mathrm{Re}(z) \leq 0 \} \subset \mathsf{SG}$$

Also:

$$\text{A-stabil} \Leftrightarrow \mathbb{C}^- \subset \mathsf{SG} \Leftrightarrow |\omega(z)| \leq 1 \ \ \forall z \in \mathbb{C}^-$$

impliziter Euler

$$y_n^h = \underbrace{\frac{1}{1-h\lambda}}_{\omega(h\lambda)} y_{n-1}^h$$

$\omega(z) = \frac{1}{1-z}$ rational und $\lim_{z\to-\infty} \frac{1}{1-z} = 0$

$$\left|\frac{1}{1-(a+ib)}\right| \leq 1 \Leftrightarrow (1-a)^2 + b^2 \geq 1$$

Graph siehe wieder im PDF aus Wien.
$\Rightarrow$ Impliziter Euler ist A-stabil!

Taylor-Verfahren

Für das Taylor-Verfahren ist $\omega(z) = \sum_{r=0}^{R} \frac{z^r}{r!}$ Polynom in $\mathbb{Z}$ vom Grad $R > 0$ und damit $\lim_{z\to-\infty} \omega(z) = \pm\infty \Rightarrow \nexists$ A-stabilen Taylor und *expliziten* RK-Verfahren.

4.3 Lineare Stabilitätsanalyse

$\rightarrow$ Erweitern der Modellproblemeanalyse auf nichtlineare Systeme
Stabilität beim skalaren MP: Näherungen für $\mathrm{Re}(\lambda) < 0$ bleiben beschränkt.

Definition 4.6
Die Lösung u der L-stetigen AWA

$$u'(t) = f(t, u(t)) \qquad t \geq t_0 \qquad u(t_0) = u_0 \tag{4.2}$$

heißt „(asymptotisch) stabil“, wenn für jede Lösung v der gestörten AWA

$$v'(t) = f(t, v(t)) \qquad t \geq t_* \qquad v(t_*) = u(t_*) + w_*$$

mit $t_* \geq t_0$ und $\|w_*\| \leq \delta$ hinreichend klein gilt:

$$\|v(t) - u(t)\| \to 0 \qquad (t \to \infty)$$

Übertragen auf den diskreten Fall:

Definition 4.7
Die AWA (4.2) sei mit dem Einschrittverfahren

$$y_n = y_{n-1} + h_n F(h_n, t_{n-1}, y_{n-1}, y_n) \qquad n \geq 0 \qquad y_0 = u_0$$

und L-stetiger Verfahrensfunktion F diskretisiert. Die Lösung $(y_n)_{n\geq 0}$ heißt „*(numerisch) stabil*“ wenn für die gestörte Folge

$$(z_n)_{n\geq n_*} \qquad z_n = z_{n-1} + h_n F(h_n, t_{n-1}, z_{n-1}, z_n) \qquad n \geq n_* \qquad z_{n_*} = y_{n_*} + w_*$$

mit $n_* \geq 0$ und $\|w_*\| \leq \delta$ hinreichend klein gilt

$$\|z_n - y_n\| \to 0 \qquad (n \to \infty)$$

Übertragen der Erkenntnisse vom skalaren Modellproblem führt auf

Hypothese 4.8
Ein Differenzenverfahren mit Stabilitätsgebiet $\mathsf{SG} \subset \mathbb{C}$ ist „numerisch stabil“ für eine allgemeine AWA, wenn die Schrittweiten h_n so gewählt werden, dass für alle EW $\lambda(t)$ der Jacobimatrix $f_x(t, u(t))$ mit $\mathrm{Re}(\lambda(t)) \leq 0$ gilt:

$$\boxed{h_n \lambda(t) \in \mathsf{SG}, \quad n \geq 0}.$$

Bemerkungen:

- Hypothese und kein Satz, also nicht rigoros beweisbar.
- Man sollte hier
$$\mathsf{SG} = \left\{ z = \lambda h \in \mathbb{C} \;\middle|\; |\omega(z)| \underbrace{<}_{\text{wichtig}} 1 \right\}$$
verwenden
- Sei AWA linear: $u' = Au$ und A diagonalisierbar. Dann reduziert sich die AWA auf entkoppelte, skalare Probleme und mit $h_n \lambda_i \in \mathsf{SG}$ ist numerische Stabilität gegeben.
- Reduktion des allgemeinen, nicht-linearen Falles durch Linearisierung

Eine genaue Analyse findet sich im Rannacher Skript, S 78ff. Die Lektüre ist jedem empfohlen.

Gegenbeispiel im Rannacher

Diagonalisierbarkeit ist entscheidend:
Die implizite, A-stabile Trapezregel liefert exponentiell anwachsende Näherungen für ein Problem dessen Lösung $\to 0$ geht obwohl $h_n \lambda_i(t) \in \mathsf{SG}$.

4.4 Implizite Runge-Kutta Verfahren

Nichtlineares System für die k_i:

$$k_i = f\left(t_{n-1} + c_i h_n, y^h_{n-1} + h_n \sum_{j=1}^{i} a_{ij} k_j \right) \qquad i = 1, \cdots, s \tag{4.3}$$

Anschließend kombiniere

$$y^h_n = y^h_{n-1} + h_n \sum_{i=1}^{s} b_i k_i \tag{4.4}$$

Bestimmung der $s^2 + 2s$ Koeffizienten über Taylorreihenentwicklung… oder Quadraturformeln.
Gehe aus von

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

Ersetze Integral durch Quadraturformel

$$u(t_n) = u(t_{n-1}) + h_n \sum_{i=1}^{s} b_i f(t_{n-1} + c_i h_n, \underbrace{u(t_{n-1} + c_i h_n)}_{\text{zu bestimmen}}) + O(h_n^{m+1})$$

mit geeigneten b_i, c_i. Bestimme die *Unbekannten* $u(t_{n-1} + c_i h_n)$ ebenfalls durch Quadraturformeln:

$$u(t_{n-1} + c_i h_n) = u(t_{n-1}) + h_n \sum_{j=1}^{s} a_{ij} f(t_{n-1} + c_j h_n, u(t_{n-1} + c_j h_n)) + O(h_n^m)$$

Darstellung der Koeffizienten im Butcher-Tableau

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \qquad A \text{ voll besetzt, } c_1 \neq 0 \text{ möglich}$$

Zentrale Frage: Lösbarkeit und Lösung von (4.4) → später

Beispiele

(a) Impliziter Euler als IRK

$$y_n^h = y_{n-1}^h + h_n \underbrace{f(t_n, y_n^h)}_{=:k_1}$$

also

$$k_1 = f(t_n, y_n^h) = f(t_{n-1} + h_n, y_{n-1}^h + h_n k_1) \rightarrow \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

(b) Einschritt Θ-Verfahren, $\Theta \in [0, 1]$

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

$$\underbrace{h_n((1-\Theta)\overbrace{f(t_{n-1}, u(t_{n-1}))}^{=:k_1} + \Theta \overbrace{f(t_n, u(t_n))}^{=:k_2}}_{\text{Quadratur}} + \begin{cases} O(h_n^2) & \Theta \neq \frac{1}{2} \\ O(h_n^3) & \Theta = \frac{1}{2} \text{ (Trapezregel)} \end{cases}$$

also

$$\begin{aligned} k_1 &= f(t_{n-1}, y_{n-1}^h) \\ k_2 &= f(t_{n-1} + h_n, y_{n-1}^h + h_n((1-\Theta)k_1 + \Theta k_2)) \\ y_n^h &= y_{n-1}^h + h_n((1-\Theta)k_1 + \Theta k_2)) \end{aligned}$$

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

Konsistenzordnung 2 für $\Theta = \frac{1}{2}$ (Trapezregel). Konsistenzordnung 1 sonst.

($\Theta = 0$: EE, $\Theta = 1$: IE)

(c) Mittelpunktregel

$$y_n^h = y_{n-1}^h + h_n \underbrace{f(t_{n-1} + \frac{h_n}{2}, \frac{1}{2}(y_{n-1}^h + y_n^h))}_{=:k_1}$$

$$\begin{aligned} k_1 &= f(t_{n-1} + \frac{h_n}{2}, \frac{1}{2}y_{n-1}^h + \frac{1}{2}(y_{n-1}^h + h_n k_1)) \\ &= f(t_{n-1} + \frac{h_n}{2}, y_{n-1}^h + \frac{h_n}{2}k_1) \end{aligned}$$

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

Lemma 4.9
Trapez- und Mittelpunktregel sind A-stabil. Beweis siehe Übungen

Lemma 4.10
Die Stabilitätsfunktion eines allgemeinen IRK-Verfahrens hat die Darstellungen:

(i)
$$\omega(z) = 1 + zb^T(I - zA)^{-1}\mathbb{1}$$

(ii)
$$\omega(z) = \frac{\det(I - zA + z\mathbb{1}b^T)}{\det(I - zA)}$$

(iii)
$$\omega(z) = \frac{P(z)}{Q(z)}$$

mit $\mathbb{1} = (\underbrace{1, \ldots, 1}_{s\text{-mal}})^T$, P, Q Polynome vom Grad $\leq s$

Beweis .

(i)
$$\mathrm{k} = \begin{pmatrix} k_1 = & \lambda y_{n-1}^h + \lambda h_n \sum_{j=1}^s a_{1j}k_j \\ k_s = & \lambda y_{n-1} + \lambda h_n \sum_{j=1}^s a_{sj}k_j \end{pmatrix} = \lambda y_{n-1}^h \mathbb{1} + h_n \lambda A k$$

$$\begin{aligned} \Leftrightarrow \quad & (I - h_n\lambda A)k \overset{(\star)}{=} \lambda y_{n-1}^h \mathbb{1} \\ \Leftrightarrow \quad & k = (I - h_n\lambda A)^{-1}\lambda y_{n-1}^h \mathbb{1} \end{aligned}$$

und dann

$$\begin{aligned} y_n^h &\overset{(\star\star)}{=} y_{n-1}^h + h_n b^T k = y_{n-1}^h + h_n \lambda b^T (I - h_n\lambda A)^{-1}\mathbb{1} y_{n-1}^h \\ &= (1 + h_n\lambda b^T(I - h_n\lambda A)^{-1}\mathbb{1})y_{n-1}^h \end{aligned}$$

(ii) Fasse ($\star$) und ($\star\star$) in einem System zusammen:

$$\begin{pmatrix} I - h_n\lambda A & 0 \\ -h_n b^T & 1 \end{pmatrix}\begin{pmatrix} k \\ y_n^h \end{pmatrix} = \begin{pmatrix} \lambda y_{n-1}^h \mathbb{1} \\ y_{n-1}^h \end{pmatrix}$$

Auflösen mit Cramerscher Regel

$$\begin{aligned} y_n^h &= \frac{\det\begin{pmatrix} I - h_n\lambda A & \lambda y_{n-1}^h \mathbb{1} \\ -h_n b^T & y_{n-1}^h \end{pmatrix}}{\det\begin{pmatrix} I - h_n\lambda A & 0 \\ -h_n b^T & 1 \end{pmatrix}} \\ &= \frac{\det\begin{pmatrix} I - h_n\lambda + h_n\lambda \mathbb{1} b^T & 0 \\ -h_n b^T & y_{n-1}^h \end{pmatrix}}{\det(I - h_n\lambda A)} \\ &= \frac{\det(I + h_n\lambda(\mathbb{1}b^T - A))}{\det(I - h_n\lambda A)} y_{n-1}^h \end{aligned}$$

(iii) Determinante in Zähler/Nenner sind Polynome in $h_n\lambda$ vom Grad $\leq s$

□

Ziel: Genauere Untersuchung der Eigenschaften allgemeiner $\omega(z)$. Lösung des Modellproblems $u' = \lambda u$ nach einem Schritt und $t_0 = 0$

$$\begin{aligned} u(t_1) &= e^{h_1\lambda} u_0 && \text{(ex. Lösung)} \\ y_1^h &= \omega(h_1\lambda) u_0 && \text{(für die numerische Lösung)} \end{aligned}$$

$$\Rightarrow u(t_1) - y_1^h = (e^{h_1\lambda} - \omega(h_1\lambda)) u_0$$

$\Rightarrow$ für $h \to 0$ muss $\omega(z)$ die Exponentialfunktion möglichst gut approximieren.
Optimale Approximation mit rationalen Funktionen führt auf Padé-Approximation. .

Definition 4.11
Sei $g(z)$ analytisch in $z = 0$. Die rationale Funktion $R(z) = \frac{P(z)}{Q(z)}$ mit P Polynom vom Grad k und Q Polynom vom Grad j heißt die „Padé-Approximation“ vom Grad (j, k) falls

$$R^{(l)}(0) = g^{(l)}(0) \text{ für } l = 0, \ldots, j + k$$

Anwendung auf $g = \exp(z)$:

$$\omega(z) = \frac{P(z)}{Q(z)} = \exp(z) \Leftrightarrow \sum_{r=0}^{k} a_r z^r = \left(\sum_{n=0}^{\infty} \frac{z^n}{n!}\right)\sum_{s=0}^{j} b_s z^s \overset{n+s=m}{=} \sum_{m=0}^{\infty}\left(\sum_{s=0}^{\min(j,m)} \frac{b_s}{(m-s)!}\right) z^m$$

Padé-Tafel für $\exp(z)$:

$\downarrow j \quad \vec{k}$	0		1		2	
0	$\frac{1}{1}$	1	$\frac{1+z}{1}$	2	$\frac{1+z+\frac{z^2}{2}}{1}$	3
1	$\frac{1}{1-z}$	2	$\frac{1+\frac{z}{2}}{1-\frac{z}{2}}$	3	$\frac{1+\frac{2z}{3}+\frac{z^2}{6}}{1-\frac{z}{3}}$	4
2	$\frac{1}{1-z+\frac{z^2}{3}}$	3	$\frac{1+\frac{z}{3}}{1-\frac{2z}{3}+\frac{z^2}{6}}$	4	$\frac{1+\frac{z}{2}+\frac{z^2}{12}}{1-\frac{z}{2}+\frac{z^2}{12}}$	5

(Die Einträge je rechts sind die Fehlerordnungen der Padé-Aprroximation, die Konsistenzordnung eines entsprechenden Verfahrens ist eins niedriger.)

Bemerkungen

- erste Zeile Taylor-Verfahren
- $(1,0)$: Impliziter Euler, $(1,1)$ Trapez/Mittelpunkt
- Diagonale: Beste Ordnung $2j$ für max. Polynomgrad
- Hauptdiagonale und erste beiden *unteren* Nebendiagonalen liefern A-stabile Verfahren

Verhalten für $z \to \infty$ genauer:

(i)

$$\lim_{y\to\infty} |\exp(-y)| = 0 \qquad \text{reelle Achse}$$
$$\lim_{y\to\infty} |\exp(\pm iy)| = 1 \qquad \text{imaginäre Achse}$$

(ii)

$$\lim_{y\to\infty} |\omega(-y)| = \lim_{y\to\infty} |\omega(\pm iy)| \text{ da } \omega(z) \text{ rational}$$

$\Rightarrow$ kein numerisches Verfahren kann beide Bedingungen aus (i) erfüllen.

- Verfahren mit $\lim_{z\to\infty} |\omega(z)| = 1$ (Hauptdiagonale der Padé-Tafel) sind für Schwingungsprobleme geeignet
- Verfahren mit $\lim_{z\to\infty} |\omega(z)| = 0$ sind für Probleme mit $Re(\lambda) < 0$ (Dämpfungseigenschaft) besser.

Definition 4.12 L-Stabil

Ein Einschrittverfahren heißt L-stabil (stark A-stabil) , falls es A-stabil ist und zusätzlich

$$\lim_{z\to\infty} |\omega(z)| = 0$$

Kollokationsverfahren

Das Polynom w von Grad s welches den Bedingungen

$$\begin{aligned} w(t_{n-1}) &= y^h_{n-1} \\ w'(t_{n-1} + c_i h_n) &= f(t_{n-1} + c_i h_n, w(t_{n-1} + c_i h_n)) \qquad i = 1, \cdots, s \\ y^h_n &:= w(t_n) \end{aligned} \tag{4.5}$$

genügt heißt „Kollaktionspolynom“.

- w erfüllt DGL in den s Punkten
- $s+1$ Bedingungen für $s+1$ Koeffizienten
- das ist ein nichtlineares System
- Lösbarkeit folgt aus dem Untenstehenden

Satz 4.13
Für gegebene Stützstellen $c_1, \ldots, c_s$ entspricht das Kollokationsverfahren (4.5) einem IRK mit:

$$a_{ij} = \int_0^{c_i} L_j(t)\,\mathrm{d}t \qquad b_i = \int_0^1 L_i(t)\,\mathrm{d}t \quad i,j = 1,\ldots,s$$

und $L_j(t) = \prod_{l \neq j} \frac{t - c_l}{c_j - c_l}$ den Lagrange-Polynomen zu den s Stützstellen c_i

Beweis .

(i) w Grad s, w' Polynom vom Grad $s-1$.

$$t \in [0,1] \qquad w'(t_{n-1} + th_n) = \sum_{j=1}^{s} \underbrace{w'(t_{n-1} + c_j h_n)}_{=f\ldots} \cdot L_j(t)$$

(ii)

$$\int_0^{c_i} w'(t_{n-1} + th_n)\,\mathrm{d}t = \left[\frac{1}{h_n} w(t_{n-1} + th_n)\right]_0^{c_i} = \frac{1}{h_n}(w(w_{n-1} + c_i h_n) - y^h_{n-1})$$

$$\begin{aligned} w(t_{n-1} + c_i h_n) &\overset{(ii)}{=} y^h_{n-1} + h_n \int_0^{c_i} w'(t_{n-1} + th_n)\,\mathrm{d}t \\ &\overset{(i)}{=} y^h_{n-1} + h_n \int_0^{c_i} \sum_{j=1}^{s} w'(t_{n-1} + c_j h_n) L_j(t)\,\mathrm{d}t \\ &\overset{(4.5)}{=} y^h_{n-1} + h_n \cdot \sum_{j=1}^{s} \underbrace{\left(\int_0^{c_i} L_j(t)\,\mathrm{d}t\right)}_{a_{ij}} \underbrace{f(t_{n-1} + c_j h_n, w(t_{n-1} + c_j h_n)}_{k_j} \end{aligned}$$

$$k_i = f(t_{n-1} + c_j h_n, w(t_{n-1} + c_i h_n))$$

$$\boxed{k_i = f(t_{n-1} + c_j h_n, y^h_{n-1} + h_n \sum_{j=1}^{s} a_{ij} k_j) \qquad i = 1,\ldots,s}$$

(iii)

$$y^h_n = w(t_{n-1} + h_n) = y^h_{n-1} + \sum_{j=0}^{s} \left(\int_0^1 L_j(t)\,\mathrm{d}t\right) k_j$$

□

Wähle Stützstellen $c_1, \ldots, c_s \Rightarrow$ IRK-Verfahren.

Satz 4.14
Ein durch Kollokation erzeugtes RK-Verfahren hat die Konsistenzordnung p, wenn die durch die Stützstellen c_i und Gewichte b_i definierte Quadraturformel die Ordnung p besitzt.

Beweis Skizze.

(i) Quadraturformel hat Ordnung p d.h. für $g \in C^p$ gilt:

$$\int_{t_{n-1}}^{t_n} g(t)\,\mathrm{d}t = h \cdot \int_0^1 g(t_{n-1} + xh)\,\mathrm{d}x = h \cdot \sum_{i=1}^{s} b_i g(t_{n-1} + c_i h) + O(h^{p+1})$$

(ii) Für w' gilt:

$$w'(t) = f(t, w(t)) + \underbrace{w'(t) - f(t, w(t))}_{r(t)}$$

d.h. w wird als Lösung einer „gestörten“ DGL interpretiert. Mit $u(t)$ Lösung des AWP

$$u' = f(t, u(t)) \qquad u(t_{n-1}) = y_{n-1}^h$$

$$[w(t) - u(t)]_{t_{n-1}}^{t_n} = \int_{t_{n-1}}^{t_n} w'(t) - u'(t)\,\mathrm{d}t = \int_{t_{n-1}}^{t_n} r(t)\,\mathrm{d}t$$

$$\overset{\text{Quadratur}}{=} h \cdot \sum_{i=1}^{s} b_i \underbrace{\left(w'(t_{n-1} + c_i h_n) - f(t_{n-1} + c_i h_n, w(t_{n-1} + c_i h_n))\right)}_{=0 \text{ wegen } (4.5)} + O(h^{p+1})$$

$$e(t_n) = h_n \tau_n^h \Rightarrow \tau_n^h = O(h^p)$$

Es fehlt zu zeigen: Beschränktheit von $\|r^{(p)}\|$ □

Zur Konstruktion von Verfahren sind geeignete Quadraturen zu wählen:

(i) Gauß-Verfahren: c_i sind Nullstellen (auf $[0, 1]$ verschobener) Legendre-Polynome
$p = 2s$ ist optimal. Stabilitätsfunktion ist Index (s, s) in Padé-Tafel → A-stabil
$s = 1$: Mittelpunktregel

$$s = 2: \quad \begin{array}{c|cc} 1/2 - \sqrt{3}/6 & 1/4 & 1/4 - \sqrt{3}/6 \\ 1/2 + \sqrt{3}/6 & 1/4 + \sqrt{3}/6 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}$$

(ii) Radau-Verfahren

$$\text{IA: } c_i \text{ Nullstelle von } \frac{\mathrm{d}^{s-1}}{\mathrm{d}x^{s-1}}[x^s(x-1)^{s-1}] \qquad c_1 = 0 \qquad c_s < 1$$

$$\text{IIA: } c_i \text{ Nullstelle von } \frac{\mathrm{d}^{s-1}}{\mathrm{d}x^{s-1}}[x^{s-1}(x-1)^s] \qquad c_1 > 0 \qquad c_s = 1$$

Konsistenzordnung $p = 2s - 1$, $w(z)$ ist Index $(s, s-1) \rightarrow$ L-stabil

$s = 1$: impliziter Euler

$s = 2$:

$1/3$	$5/12$	$-1/12$
1	$3/4$	$1/4$
	$3/4$	$1/4$

(iii) Lobatto-Regeln $c_1 = 0$, $c_s = 1$ Konsistenzordnung $p = 2s - 2$.

$s = 2$, $p = 2$:

0	0	0
1	1	0
	$1/2$	$1/2$

$\rightarrow$ Heun

0	0	0	0
$1/2$	$1/4$	$1/4$	0
1	0	1	0
	$1/6$	$2/3$	$1/6$

DIRK-Verfahren

DIRK = diagonally implicit Runge-Kutta method
A ist untere Dreiecksmatrix
voll implizit $\rightarrow$ 1 System der Größe sd
A untere Dreiecksmatrix $\rightarrow$ s Systeme der Größe d
$\Rightarrow$ Vorteil
SDIRK = singly diagonally implicit Runge-Kutta method

$$a_{11} = a_{22} = \ldots = a_{ss} \Rightarrow \text{im linearen Fall nur ein LR-Verfahren}$$

Beispiele

Alexander-Verfahren:
$s = 2$, $p = 2$, L-stabil (sogar stark L-stabil)

α	α	0
1	$1-\alpha$	α
	$1-\alpha$	α

$\alpha = 1 \pm \frac{\sqrt{2}}{2}$

$\leftarrow b^T$ = letzte Zeile von A spart weitere Berechnungen

Crouzieux-Verfahren
$s = 2$, $p - 3$, A-stabil

$1/2 + 1/2 \cdot \sqrt{3}$	$1/2 + 1/2 \cdot \sqrt{3}$	0
$1/2 - 1/2 \cdot \sqrt{3}$	$-1/\sqrt{3}$	$1/2 + 1/2 \cdot \sqrt{3}$
	$1/2$	$1/2$

Lösung der nichtlinearen Systeme

$$k_i = f(t_{n-1} + c_i h_n, y_{n-1}^h + h_n \sum_{j=1}^{s} a_{ij} k_j) \qquad i = 1, \ldots, s$$

ist in Fixpunktform

$$k = F(k) \qquad F_i(k) = f(t_{n-1} + c_i h_n, \sum a_{ij} k_j)$$

$$
\begin{aligned}
\|F_i(k) - F_i(\tilde{k})\| &= \left\| f\left(\ldots, y_{n-1}^h + h_n \cdot \sum a_{ij} k_j\right) - f\left(\ldots, y_{n-1}^h + h_n \cdot \sum a_{ij} \tilde{k}_j\right) \right\| \\
&\underset{\text{f L-stetig}}{\leq} L \cdot \|h_n \sum_{j=1}^{s} a_{ij}(k_j - \tilde{k}_j)\| \\
&\leq L \cdot h_n \cdot \sum |a_{ij}| \|k_j - \tilde{k}_j\| \\
&\leq L \cdot h_n \cdot \max_j \|k_j - \tilde{k}_j\| \underbrace{\sum_{j=1}^{s} |a_{ij}|}_{\leq \|A\|_\infty} \\
&\leq L \cdot h_n \cdot \max_j \|k_j - \tilde{k}_j\| \cdot \|A\|_\infty \\
\max_i \|F_i(k) - F_i(\tilde{k})\| &\leq hL\|A\|_\infty \cdot \max_i \|k_i - \tilde{k}_i\|
\end{aligned}
$$

Kontraktion für $h\|A\|_\infty L < 1$ d.h.

$$\boxed{h < \frac{1}{L \cdot \|A\|_\infty}}$$

Bedingung ist notwendig für die Konvergenz der Fixpunktiteration, aber nur hinreichend für die Lösung der nichtlinearen Systeme.
⇒ Lösung der nichtlinearen Systeme mit dem Newton-Verfahren.

Lösung von $F(x) = 0$

Idee: $F(x + z) = F(x) + F_x(x) \cdot z + O(\|z\|^2)$

$$x^{(k+1)} = x^{(k)} + \underbrace{z^{(k)}}_{\text{Lösung von } F(x^{(k)}) + F_x(x^{(k)}) z^{(k)} = 0} = x^{(k)} - (F_x(x^{(k)}))^{-1} F(x^{(k)})$$

- Iteration konvergiert quadratisch für $x^{(k)}$ hinreichend nahe an der Lösung x (und f glatt genug).
- Verbesserung der „globalen“ Konvergenz durch „Dämpfung“ Wähle Startwert $x^{(0)}$ (Lösung zum letzten Zeitpunkt)
  while ($\|F(x^{(k)}\| > \varepsilon \|F(x^{(0)}\|$) {
  1) Löse $F_x(x^{(k)}) z^{(k)} = -F(x^{(k)})$
  2) Für $\lambda \in (0, 1)$ finde kleinstes $l \in \mathbb{N}_0$ sodass
     $$\|F(x^{(k)} + \lambda^l z^{(k)}\| \leq (1 - \frac{1}{4}\lambda^l)\|F(x^{(k)})\| \qquad \text{„line search“}$$
  3) Setze $x^{(k+1)} = x^{(k)} + \lambda^l z^{(k)}$

  }
- Löst man in (1) $Az^{(k)} = -F(x^{(k)})$ mit $A \neq F_x(x^{(k)})$ spricht man von *inexakten* Newton-Verfahren
  - Nehme immer $A = F_x(x^{(0)})$
  - inexakte Lösung mit iterativen Verfahren
  - „Rang A“ updates
  - Für $F(x) = F^1(x) + F^2(x)$ setze $A = F_x^1(x^{(k)})$

Warum ist Newton besser?

Motivation am Beispiel des impliziten Euler:

$$y_n = y_{n-1} + hf(t_n, y_n) \Leftrightarrow \boxed{F(y_n) := hf(t_n, y_n) - y_n + y_{n-1} = 0}$$

Newton:

$$\begin{aligned}
y_n^{(k+1)} &= y_n^{(k)} - F_x^{-1}\left(y_n^{(k)}\right) F\left(y_n^{(k)}\right) \\
&= y_n^{(k)} - \underbrace{\left(h_n f_x\left(t_n, y_n^{(k)}\right) - I\right)}_{=:J^{(k)}} \left(h_n f\left(t_n, y_n^{(k)}\right) - y_n^{(k)} + y_{n-1}\right) \\
&= -\left(J^{(k)}\right)^{-1} y_{n-1} - \left(J^{(k)}\right)^{-1} \left(h_n f\left(t_n, y_n^{(k)}\right) - y_n^{(k)} - \left(h_n f_x\left(t_n, y_n^{(k)}\right) - I\right) y_n^{(k)}\right) \\
&= \underbrace{\left(J^{(k)}\right)^{-1} y_{n-1} - \left(J^{(k)}\right)^{-1} \left(h_n f\left(t_n, y_n^{(k)}\right) - h_n f_x\left(t_n, y_n^{(k)}\right) y_n^{(k)}\right)}_{g\left(y_n^{(k)}\right)}
\end{aligned}$$

Nun sei f von der Gestalt $f(t, x) = Ax + b(x)$ mit *steifen* Anteil A und *nicht-steifen* Anteilt b. Verwende inexaktes Newton-Verfahren mit $J^{(k)} := h_n A - I$, d.h. $f_x \approx A$

$$\begin{aligned}
g(y_n^{(k)}) &= -(h_n A - I)^{-1} y_{n-1} - (h_n A - I)^{-1} \left[h_n A y_n^{(k)} + h_n b y_n^{(k)} - h_n (A + b_x(y_n^{(k)}) y_n^{(k)})\right] \\
&= (I - h_n A)^{-1} y_{n-1} + h_n (I - h_n A)^{-1} \left[b(y_n^{(k)}) - b_x(y_n^{(k)}) y_n^{(k)}\right]
\end{aligned}$$

Kontraktion falls $\exists L : \|g(x) - g(\tilde{x})\| \leq \underbrace{\ldots}_{L} \|x - \tilde{x}\|$

$$h_n \underbrace{\|(I - h_n A)^{-1}\|}_{(I-h_nA)^{-1} \text{ hat EW } \frac{1}{1-h\lambda_i}} (\overbrace{L_b}^{\text{nicht groß!}} + \|b_x(y_n^{(k)})\|) < 1$$

λ_i EW von A

In diesem Fall große Schrittweiten möglich.

Alternative Formulierung von IRK-Verfahren

Setze $z_i := h_n \sum_{j=1}^{s} a_{ij} k_j$ mit $z = (z_1, \ldots, z_s)^T$, $\quad k = (k_1, \ldots, k_s)$ gilt

$$\boxed{z = h_n A k} \rightarrow k = \frac{1}{h_n} A^{-1} z \text{ für } \det(A) \neq 0$$

Für die z_i gilt:

$$\begin{aligned}
z_i &= h_n \sum_{j=1}^{s} a_{ij} k_j = h_n \sum_{j=1}^{s} a_{ij} f(t_{n-1} + c_j h_n, y_{n-1}^h + \underbrace{h_n \sum_{l=1}^{s} a_{jl} k_l}_{z_j}) \\
&= h_n \sum_{j=1}^{s} a_{ij} f(t_{n-1} + c_j h_n, y_{n-1}^h + z_j) \qquad \text{weniger anfällig gegen Auslöschung}
\end{aligned}$$

Außerdem: falls b^T = letzte Zeile von A (z.B. Alexander), dann gilt

$$y_n = y_{n-1} + h_n b^T k \overset{a_{sj}=b_j}{=} y_{n-1} + z_s$$

5 Mehrschrittverfahren

(Graph zur Veranschaulichung, wie die MSV im Vergleich zu den ESV Verfahren funktionieren werden)
Ziel: Man möchte die selbe Fehlergenauigkeit bei gleichem Aufwand erreichen.

5.1 Verfahrenskonstruktion

- Beschränkung auf äquidistante Schrittweite
- Variable Schrittweite prinzipiell möglich, aber aufwändiger als bei Einschrittverfahren

Integrationsbasierte Verfahren

$$u(t_n) = u(t_{n-\sigma}) + \int_{t_{n-\sigma}}^{t_n} f(s, u(s))\,\mathrm{d}s \qquad \sigma \in \mathbb{N}$$

Newton-Cotes Quadratur für das Integral: Interpoliere Integrand durch Polynom vom Grad m

$$p_m(t) = \sum_{\mu=0}^{m} f(\underbrace{t_{k-\mu}}_{\text{noch zu wählen}}, u(t_{k-\mu}))L_\mu(t) \qquad L_\mu(t) = \prod_{\substack{\ell=0 \\ \ell\neq\mu}}^{m} \frac{t - t_{k-\ell}}{t_{k-\mu} - t_{k-\ell}}$$

und damit

$$u(t_n) = u(t_{n-\sigma}) + \sum_{\mu=0}^{m} f(t_{k-\mu}, u(t_{k-\mu})) \underbrace{\int_{t_{n-\sigma}}^{t_n} L_\mu(s)\,\mathrm{d}s}_{\text{Gewichte} \to \text{ausrechnen}} + \underbrace{O(h^{m+2})}_{\text{falls } u^{(m+2)} \text{ beschränkt}}$$

(Beweis siehe Übungen)
Verfahrensklassen:

1) Adams-Bashforth-Formeln: $\sigma = 1$, $k = n - 1$ (explizit)
   (Grafik zur Erläuterung)

$m = 0$	$y_n = y_{n-1} + hf_{n-1}$	(EE) $f_{n-i} := f(t_{n-i}, u(t_{n-i}))$
$m = 1$	$y_n = y_{n-1} + \frac{1}{2}h(3f_{n-1} - f_{n-2})$	
$m = 2$	$y_n = y_{n-1} + \frac{1}{12}h(23f_{n-1} - 16f_{n-2} + 5f_{n-3})$	

2) Adams-Moulton: $\sigma = 1$, $k = n$ (implizit)
   (Grafik)

$m = 0$	$y_n = y_{n-1} + h \cdot f_n$	(IE)
$m = 1$	$y_n = y_{n-1} + \frac{1}{2}h \cdot (f_n + f_{n-1})$	(Trapezregel)
$m = 2$	$y_n = y_{n-1} + \frac{1}{12}h \cdot (5f_n + 8f_{n-1} - f_{n-2})$	

3) Nyström-Verfahren : $\sigma = 2$, $k = n - 1$ (explizit)
(Grafik)

$$m = 0 \quad y_n = y_{n-2} + 2hf_{n-1} \qquad \text{(explizite Mittelpunktregel)}$$

4) Milne-Simpson : $\sigma = 2$, $k = n$ (implizit)
(Grafik)

$$m = 2 \quad y_n = y_{n-2} + \frac{1}{3}h(f_n + 4f_{n-1} + f_{n-2}) \qquad \text{(Simpson-Regel)}$$

Differentiationsbasierte Verfahren

Idee: Lege Polynom vom Grad m durch die Funktionswerte:

$$p_m(t) = \sum_{\mu=o}^{m} u(\underbrace{t_{k-\mu}}_{\text{zu wählen}}) \cdot L_\mu(t)$$

Fehler der Ableitung:

$$p'_m(t_n) = \sum_{\mu=0}^{m} \left(L_\mu\right)'(t_n) u(t_{k-\mu}) = f(t_n, \overbrace{u(t_n)}^{\text{implizit}}) + \overbrace{O(h^m)}^{\text{wegen Ableitung}}$$

$k = n$ ergibt „Rückwärtsdifferenzenformeln“ (engl. BDF: „backward difference formula“)

$$\begin{array}{lrcrcrcrcl} m=1: & y_n & - & y_{n-1} & & & & & = & hf_n \quad \text{(IE)} \\ m=2: & y_n & - & \frac{4}{3}y_{n-1} & + & \frac{1}{3}y_{n-2} & & & = & \frac{2}{3}hf_n \\ m=3: & y_n & - & \frac{18}{11}y_{n-1} & + & \frac{9}{11}y_{n-2} & - & \frac{2}{11}y_{n-3} & = & \frac{6}{11}hf_n \end{array}$$

- Geht analog für Systeme von DGL

Definition 5.1
Eine allgemeine *lineare* Mehrschrittmethode (LMM) hat die Form:

$$\sum_{r=0}^{R} \alpha_{R-r} y_{n-r} = h \sum_{r=0}^{R} \beta_{R-r} f_{n-r} \qquad f_m = f(t_m, y_m) \qquad n \geq R \tag{5.1}$$

Man normiert $\alpha_R = 1$, d.h. es gibt insgesamt $2R - 1$ freie Parameter. $\beta_R = 0$ liefert explizite Verfahren, sonst implizite. Weiter fordern wir, dass

$$|\alpha_0| + |\beta_0| \neq 0$$

für ein „echtes“-R-stufiges Verfahren.

5.2 Konsistenz von LMM

Wie bei ESV definiert man „lokalen Diskretisierungsfehler“ durch Einsetzen der exakten Lösung in die Differenzenformel:

$$\tau_n^h = \tau^h(t_n) := h^{-1} \sum_{r=0}^{R} \alpha_{R-r} \underbrace{u_{n-r}}_{u(t_{n-r})} - \sum_{r=0}^{R} \beta_{R-r} f(t_{n-r}, u_{n-r})$$

Anderer Name: lokaler Abschneidefehler

Hilfssatz 5.2
Bei *exakten* Startwerten $y_{n-r} = u_{n-r}$, $r = 1, \ldots, R$ gilt für jede LMM die Beziehung

$$u_n - y_n = h\tau_n^h(1 + O(h))$$

Beweis .
Definition Abschneidefunktion:

$$\sum_{r=0}^{R} \alpha_{R-r} u_{n-r} = h \sum_{r=0}^{R} \beta_{R-r} f(t_{n-r}, u_{n-r}) + h\tau_n^h$$

Definition LMM:

$$\sum_{r=0}^{R} \alpha_{R-r} y_{n-r} = h \sum_{r=0}^{R} \beta_{R-r} f(t_{n-r}, y_{n-r})$$

Abziehen der Gleichungen voneinander, $\alpha_R = 1$:

$$\begin{aligned} u_n - y_n &= h\beta_R(f(t_n, u_n) - f(t_n, y_n)) + h\tau_m^h \\ \|u_n - y_n\| &\leq h|\beta_R|L\|u_n - y_n\| + h\|\tau_n^h\| \\ \|u_n - y_n\| &\leq \frac{h\|\tau_n^h\|}{1 - h|\beta_R|L} = h\|\tau_n^h\|(1 + O(h)) \end{aligned}$$

□

Definition 5.3
Eine LMM heißt konsistent mit der AWA, wenn

$$\max_{t_n \in I} \|\tau_n^h\| \to 0 \qquad (h \to 0)$$

bzw. konsistent, von der Ordnung p falls

$$\max_{t_n \in I} \|\tau_n^h\| = O(h^p)$$

Hilfssatz 5.4
Der lokale Diskretisierungsfehler einer LMM besitzt für eine analytische Lösung $u(t)$ die Darstellung:

$$\tau^h(t) = h^{-1} \sum_{i=0}^{\infty} c_i h^i u^{(i)}(t)$$

mit

$$C_0 = \sum_{r=0}^{R} \alpha_{R-r} \qquad C_i = (-1)^i \left\{ \frac{1}{i!} \sum_{r=0}^{R} r^i \alpha_{R-r} + \frac{1}{(i-1)!} \sum_{r=0}^{R} r^{i-1} \beta_{R-r} \right\}$$

Beweis .
Entwickle $u(t_n - rh)$ und $f(t_n - rh, u(t_n - rh))$ in Taylorreihe um t_n. Anschließend einsetzen in $\tau^h(t_n)$, nach h-Potenzen sortieren. □

Hilfssatz 5.5
Eine LMM ist genau dann konsistent mit jeder AWA, wenn gilt:

$$\sum_{r=0}^{R} \alpha_{R-r} = 0 \text{ und } \sum_{r=0}^{R} (r\alpha_{R-r} + \beta_{R-r}) = 0$$

bzw. konsistent von der Ordnung genau p, wenn gilt $C_0 = C_1 = \cdots = C_p = 0, C_{P+1} \neq 0$.
Beweis:
Nach Hilfssatz 5.4

$$\tau_n^h = \sum_{i=0}^{\infty} C_i h^{i-1} u^{(i)}(t_n)$$

Konsistenz $\Leftrightarrow C_0 = C_1 = 0$. Konsistenzordnung genau p wegen $C_0 = \ldots = C_p = 0$ gilt

$$\tau_n^h = \underbrace{C_{p+1}}_{\text{Fehlerkonstante}} h^p u^{(p+1)}(t_n) + O(h^{p+1})$$

Beispiel 5.6
Allgemeine 2-Schritt LMM. Fünf freie Koeffizienten $\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2$.
$p = 3$ benötigt vier Bedingungen: $C_0 = C_1 = C_2 = C_3 = 0$
$\alpha := \alpha_0$ sei noch frei und es ergibt sich:

$$y_n - (1+\alpha)y_{n-1} + \alpha y_{n-2} = \frac{h}{12}[(5+\alpha)f_n + 8(1-\alpha)f_{n-1} - (1+5\alpha)f_{n-2}]$$

weiter gilt

$$C_4 = -\frac{1}{4!}(1+\alpha)$$

(i) $\alpha = -1 \Rightarrow C_4 = 0, C_5 \neq 0$, also $p = 4$: Simpson-Regel

(ii) $\alpha \neq -1 \Rightarrow C_4 \neq 0$, also $p = 3$. $\alpha = 0$: Adams-Moulton (z.B.)

(iii) $\alpha = -5 \Rightarrow$ explizites Verfahren (d.h. mit 2-f Auswertungen $p = 3$. Vergleiche mit expliziten RK-Verfahren!

Anwendung des Verfahrens mit $\alpha = -5$ führt zu:

$$|y_N| \to \infty \text{ für } h \to 0 \text{ und } N = \frac{T}{h}!$$

Verfahren ist *nicht* konvergent obwohl

- Konsistenzordnung 3
- AWA L-stabil

Warum?

5.3 Nullstabilität von LMM

Beispiel:

Löse triviale AWA $u'(t) = 0, t \geq 0, u(0) = u_0$ mit der trivialen Lösung $u(t) = u_0$. Führt auf:

$$\sum_{r=0}^{R} \alpha_{R-r} y_{n-r} = 0 \qquad n \geq R$$

Beobachtung: Winzige Störung in den Startwerten können zu großen, anwachsenden Fehlern führen.

Untersuchung der homogenen Differenzengleichung

$$\alpha_R y_n + \alpha_{R-1} y_{n-1} + \ldots + \alpha_0 y_{n-R} = 0 \qquad n \geq R \tag{5.2}$$

und $\alpha_R = 1$, $\alpha_0 \neq 0$.
Neben der trivialen Lösung $(y_n)_{n\in\mathbb{N}} = (0, \ldots)$ gibt es noch weitere, nicht triviale Lösungen.
Ansatz: $y_n = \lambda^n$ führt auf:

$$\sum_{r=0}^{R} \alpha_{R-r} y_{n-r} = \sum_{r=0}^{R} \alpha_{R-r} \lambda^{n-r} = \sum_{r=0}^{R} \alpha_{R-r} \lambda^{\overbrace{n-R}} {}^{+R-r} = \underbrace{\lambda^{n-R}}_{} \underbrace{\sum_{i=0}^{R} \alpha_i \lambda^i}_{=:\varrho(\lambda)} \overset{!}{=} 0$$

$\Rightarrow$ Neben $\lambda = 0$ sind Nullstellen von $\varrho(\lambda)$ nicht triviale Lösungen.

Definition 5.7

(i) $\varrho(\lambda) = \sum_{i=0}^{R} \alpha_i \lambda^i$ heißt erstes charakteristisches Polynom

(ii) $\sigma(\lambda) = \sum_{i=0}^{R} \beta_i \lambda^i$ heißt zweites charakteristisches Polynom

einer LMM.

Seien λ_i die Nullstellen von $\varrho(\lambda)$. Im Falle *einfacher* Nullstellen haben *alle* Lösungen von (5.2) die Form $\sum_{i=1}^{R} c_i \lambda_i^n$, d.h. sie bleiben beschränkt für $|\lambda_i| \leq 1$. Skizze:

(i) Die Lösungen von (5.2) bilden einen R-dimensionalen Vektorraum

(ii) Jedes λ_i liefert „Startvektor“ $\begin{pmatrix} 1 & \lambda_i & \lambda_i^2 & \ldots & \lambda_i^{R-1} \end{pmatrix}$

$$\rightarrow \text{ Vandermondematrix } \begin{pmatrix} 1 & \lambda_1 & \ldots & \lambda_1^{R-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_R & \ldots & \lambda_R^{R-1} \end{pmatrix} \text{ ist regulär}$$

Fall mehrfache Nullstelle λ_i: Dann ist neben λ_i^n auch $n\lambda_i^n$ eine Lösung:

$$\begin{aligned}
\sum_{r=0}^{R} \alpha_{R-r}(n-r)\lambda_i^{n-r} &= \sum_{r=0}^{R} \alpha_{R-r}(\underbrace{n-R}+\underbrace{R-r})\lambda_i^{\overbrace{n-R}+\overbrace{R-r}} \\
&= (n-R)\lambda^{n-R}\underbrace{\sum_{r=0}^{R}\alpha_{R-r}\lambda_i^{R-r}}_{\varrho(\lambda_i)} + \sum_{r=0}^{R}\alpha_{R-r}(R-r)\lambda_i^{\overbrace{n+1-R}+\overbrace{R-r-1}} \\
&= (n-R)\lambda^{n-R}\varrho(\lambda_i) + \lambda_i^{n+1-R}\underbrace{\sum_{r=0}^{R}\alpha_{R-r}(R-r)\lambda_i^{R-r-1}}_{\varrho'(\lambda_i)} \\
&= 0
\end{aligned}$$

$\Rightarrow$ Für Beschränktheit der Lösung ist nun $|\lambda_i| < 1$ notwendig.

Definition 5.8
Eine LMM erfüllt die *Wurzelbedingung* bzw. heißt *null-stabil*, wenn für die Nullstellen λ_i des ersten charakteristischen Polynoms gilt:

$$|\lambda_i| \leq 1, \text{ falls } \lambda_i \text{ einfache Nullstelle} \qquad |\lambda_i| < 1, \text{ falls } \lambda_i \text{ mehrfache Nullstelle}$$

5.4 Konvergenz von LMM

Definition 5.9 Konvergenz
Eine LMM heißt konvergent, wenn für jede AWA gilt:

$$\max_{R\leq n\leq N} \|y_n - u(t_n)\| \to 0 \qquad t_n \in I \qquad h \to 0$$

vorausgesetzt die Startwerte konvergieren

$$\max_{0\leq n<R} \|y_n - u(t_n)\| \to 0 \qquad h \to 0$$

Lemma 5.10 Spektralradius und Matrixnorm
Zu jeder Matrix $A \in \mathbb{R}^{m\times m}$ gibt es für jedes $\varepsilon > 0$ eine natürliche Matrixnorm $\|\quad\|_{A,\varepsilon}$ sodass für den Spektralradius $\varrho(A) := \max\{|\lambda| \;\big|\; \lambda \text{ ist Eigenwert von } A\}$ gilt:

$$\varrho(A) \leq \|A\|_{A,\varepsilon} \leq \varrho(A) + \varepsilon$$

Ist jeder Eigenwert von A mit $|\lambda| = \varrho(A)$ nur einfach, so existiert sogar eine natürliche Matrixnorm $\|\quad\|_{A,0}$ mit $\|A\|_{A,0} = \varrho(A)$

Beweis .
Konstruktiv:

(a) Schur-Normalform: A ist ähnlich zu einer Dreiecksmatrix:

$$A = T^{-1}RT \qquad R = \begin{bmatrix} r_{11} & & \star \\ & \ddots & \\ 0 & & r_{mm} \end{bmatrix} \qquad r_{ii} \text{ Eigenwert von } A$$

d.h. $\varrho(A) = \max\limits_{1\le i\le n} |r_{ii}|$ (T ist unitär, d.h. alle EW von T haben den Betrag eins).

Für $\delta \in (0, 1]$ definiere Matrizen:

$$S_\delta = \begin{bmatrix} 1 & & & & 0 \\ & \delta & & & \\ & & \delta^2 & & \\ & & & \ddots & \\ 0 & & & & \delta^{m-1} \end{bmatrix} \qquad D = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{mm} \end{bmatrix} \qquad Q_\delta = \begin{bmatrix} 0 & r_{12} & \delta \cdot r_{13} & \dots & \delta^{m-2} \cdot r_{1m} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta \cdot r_{m-2,m} \\ & & & \ddots & r_{m-1,m} \\ 0 & & & & 0 \end{bmatrix}$$

Dann gilt:

$$R_\delta := S_\delta^{-1} R S_\delta = S_\delta^{-1}(D + \underbrace{N}_{R-D})S_\delta = \underbrace{S_\delta^{-1} D S_\delta}_{D} + \underbrace{S_\delta^{-1} N S_\delta}_{(r_{ij}\cdot\delta^{j-i})_{i,j=0,j-i>0}^{n}} = D + \delta Q_\delta$$

$S_\delta^{-1}T$ ist regulär, $\|x\|_\delta := \|S_\delta^{-1}Tx\|_2$, $x \in \mathbb{R}$ erklärt eine Vektornorm

Weiter gilt: $A = T^{-1}RT \overset{\text{Def. von } R_\delta}{=} T^{-1}S_\delta R_\delta S_\delta^{-1} T$.

Also

$$\begin{aligned}
\|Ax\|_\delta &= \|\underbrace{T^{-1}S_\delta R_\delta S_\delta^{-1}T}_{=A} x\|_\delta \overset{\text{Def. von } \|\cdot\|_\delta}{=} \|\underbrace{S_\delta^{-1}TT^{-1}S_\delta}_{I} R_\delta \underbrace{S_\delta^{-1}Tx}_{=:y}\|_2 \\
&= \|R_\delta y\|_2 = \|(D + \delta Q_\delta)y\|_2 \\
&\le \|Dy\|_2 + \delta\|Q_\delta y\|_2 \\
&\le \underbrace{\max_{1\le i\le m}(r_{ii})}_{=\varrho(A)} \|y\|_2 + \delta \underbrace{\|Q_\delta\|_F}_{=:\mu} \|y\|_2 \\
&= (\varrho(A) + \delta\mu)\|y\|_2 \\
&= (\varrho(A) + \delta\mu)\|\underbrace{S_\delta^{-1}Tx}_{\text{Def. von } y}\|_2 \\
&= (\varrho(A) + \delta\mu)\|x\|_\delta \qquad\qquad \delta := \frac{\varepsilon}{\mu}
\end{aligned}$$

also

$$\|A\|_\delta = \sup_{x\neq 0} \frac{\|Ax\|_\delta}{\|x\|_\delta} \le \varrho(A) + \varepsilon.$$

(b) Alle EW mit $|\lambda| = \varrho(A)$ seien einfach. Zu jedem dieser EW existieren daher l.u. EV. Dann gilt für Schurform:

$$A = T^{-1}RT \qquad R = \begin{bmatrix} R^{11} & 0 \\ 0 & R^{22} \end{bmatrix} \qquad R^{11} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_S \end{bmatrix} \qquad |\lambda_i| = \varrho(A)$$

$$R^{22} = \begin{bmatrix} r_{s+1,s+1} & \star \\ 0 & r_{m,m} \end{bmatrix}$$

mit $\max\limits_{s+1\le j\le m} |r_{jj}| < \varrho(A)$

Definiere S_δ, Q_δ wie oben:

$$Q_\delta = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & Q_\delta^{22} \end{array}\right]$$

$$Q_\delta^{22} = \begin{bmatrix} 0 & & \star \\ & \ddots & \\ & & 0 \end{bmatrix}$$

Also:

$$\begin{aligned} \|Ax\|_\delta^2 &= \|T^{-1}S_\delta R_\delta \underbrace{S_\delta^{-1}Tx}_{=y}\|_\delta^2 = \|R_\delta y\|_2^2 \\ &= \|R_\delta^{11}y_1\|_2^2 + \|R_\delta^{22}y_2\|_2^2 && y = (y_1, y_2)^T \\ &\le \varrho^2(A)\|y_1\|_2^2 + (\underbrace{\varrho^2(R^{22})}_{<\varrho(A)} + \underbrace{\delta^2}_{\substack{\text{wähle } \delta \text{ so dass} \\ \varrho^2(R_{22})+\delta^2\mu^2=\varrho^2(A)}} \|Q_\delta^{22}\|_F^2)\|y_2\|_2^2 \\ &\le \varrho^2(A)(\|y_1\|_2^2 + \|y_2\|_2^2) \\ &= \varrho^2(A)\|x\|_\delta^2 \end{aligned}$$

□

Satz 5.11 Stabilität von LMM
Die AWA genüge der Lipschitz-Bedingung und die LMM sei null-stabil. Ist $\{y_n\}_{n=0}^\infty$ Lösung der ungestörten LMM.

$$\sum_{r=0}^{R} \alpha_{R-r} y_{n-r} = h \cdot \sum_{r=0}^{R} \beta_{R-r} f_{n-r} \qquad (n \ge R),\ y_n = u_n (n < R) \qquad \text{(I)}$$

und $\{\tilde{y}_n\}_{n=0}^\infty$ Lösung der „gestörten“ LMM

$$\sum_{r=0}^{R} \alpha_{R-r} \tilde{y}_{n-r} = h \cdot \sum_{r=0}^{R} \beta_{R-r} \underbrace{\tilde{f}_{n-r}}_{=f(t_{n-r},\tilde{y}_{n-r})} + \varrho_n \qquad (n \ge R), \tilde{y}_n = y_n + \varrho_n, (n < R) \qquad \text{(II)}$$

dann gibt es unter der Voraussetzung $h < \frac{1}{L|\beta_R|}$ (nur für $\beta_R \neq 0$) die Abschätzung:

$$\|\tilde{y} - y_n\| \le K \cdot e^{\Gamma(t_n - t_0)} \left\{ \max_{0\le\nu<R} \|\varrho_\nu\| + \sum_{r=R}^{n} \|\varrho_\nu\| \right\} \qquad (n \ge R)$$

Beweis .

(i) Rekursionsgleichung: (II) - (I) liefert

$$\sum_{r=0}^{R} \alpha_{R-r} \underbrace{(\tilde{y}_{n-r} - y_{n-r})}_{=:e_{n-r}} = h \cdot \sum_{r=0}^{R} \beta_{R-r}(\tilde{f}_{n-r} - f_{n-r}) + \varrho_n$$

Umstellen, $\alpha_R = 1$, $e_{n-r} := \tilde{y}_{n-r} - y_{n-r}$

$$e_n - h\beta_R(\tilde{f}_n - f_n) = -\sum_{r=1}^{R} \alpha_{R-r} e_{n-r} + \underbrace{h \cdot \sum_{r=1}^{R} \beta_{R-r}(\tilde{f}_{n-r} - f_{n-r}) + \varrho_n}_{:=b_n}$$

Definiere Abkürzung:

$$\sigma_n := \begin{cases} \frac{\tilde{f}_n - f_n}{e_n} & e_n \neq 0 \\ 0 & e_n = 0 \end{cases}$$

$$(1 - h\beta_R\sigma_n)e_n = -\sum_{r=1}^{R} \alpha_{R-r} e_{n-r} + b_n \tag{III}$$

Füge triviale Gleichungen hinzu

$$(1 - h\beta_R\sigma_{n-r})e_{n-r} = (1 - h\beta_R\sigma_{n-r})e_{n-r} \qquad r = 1, \cdots, R-1 \tag{IV}$$

In Matrixform lauten die R Gleichungen dann:

$$\underbrace{\begin{bmatrix} 1 - h \cdot \beta_R\sigma_n & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & 1 - h \cdot \beta_R\sigma_{n-R+1} \end{bmatrix}}_{=:D_{n-R+1}} \underbrace{\begin{bmatrix} e_n \\ e_{n-1} \\ \vdots \\ e_{n-R+1} \end{bmatrix}}_{=:E_{n-R+1}}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ & 1 - h\beta_R\sigma_{n-1} & & \vdots \\ & & \ddots & \vdots \\ 0 & & & 1 - h\beta_R\sigma_{n-R+1} \end{bmatrix}}_{=:C_{n-R+1}} \underbrace{\begin{bmatrix} -\alpha_{R-1} & \cdots & \cdots & -\alpha_0 \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{bmatrix}}_{=:A} \underbrace{\begin{bmatrix} e_{n-1} \\ e_{n-2} \\ \vdots \\ \vdots \\ e_{n-R} \end{bmatrix}}_{=:E_{n-R}} + \underbrace{\begin{bmatrix} b_n \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}}_{=:B_{n-R+1}}$$

Indextransformation:

$$i := n - R + 1 \Leftrightarrow n = i + R - 1$$

d.h.

$$D_i E_i = C_i A E_{i-1} + B_i,\ i \geq 1,\ E_0 = (\varrho_{R-r}, \cdots, \varrho_0)^T$$

wegen

$$|\sigma_n| = \left| \frac{\tilde{f}_n - f_n}{e_n} \right| \leq L \frac{|e_n|}{|e_n|} = L \text{ ist } |h\beta_R\sigma_n| \leq h \cdot |\beta_R| L < 1 \text{ nach Voraussetzung}$$

und somit D_i immer invertierbar und wir erhalten:

$$E_i = D_i^{-1}\{C_i A E_{i-1} + B_i\} \qquad (i \geq 1) \qquad \text{(V)}$$
$$E_0 = (\varrho_{R-1}, \cdots, \varrho_0)^T \qquad (i = 0) \qquad \text{(V)}$$

(ii) Die Matrix A hat das charakteristische Polynom

$$\chi_A(\lambda) = (-1)^R(\alpha_0 + \alpha_1\lambda + \cdots + \alpha_{R-1}\lambda^{R-1} + \lambda^R) = (-1)^R\varrho(\lambda)$$

Wegen Nullstabilität der LMM ($|\lambda_i| < 1$ oder $\lambda_i = 1$ und $|\lambda_i|$ einfach) gibt es nach Lemma 5.10 eine natürliche Norm, so dass $\|A\|_0 = \varrho(A)$.

Zum charakteristischen Polynom:

$$\underbrace{\begin{bmatrix} -(\alpha_R+\lambda) & -\alpha_{R-2} & \cdots & -\alpha_0 \\ 1 & -\lambda & & \\ & \ddots & \ddots & \\ & & 1 & -\lambda \end{bmatrix}}_{A-\lambda I}$$

Entwickeln der Determinate nach erster Zeile liefert $\chi_A(\lambda)$.

(iii) Abschätzen von $\|D_i^{-1}\|_0$, $\|C_i\|_0$, $\|B_i\|_0$.

$B_i = (b_{i+R-1}, 0, \cdots, 0)^T$.

$$\text{Äquivalenz aller Normen: } \frac{1}{\gamma}\|z\|_0 \leq \underbrace{\sum_{r=1}^{R}|z_r|}_{=\|z\|_1} \leq \gamma\|z\|_0$$

$$\begin{aligned}
\|B_i\|_0 &\leq \gamma\|B_i\|_0 = \gamma|b_{i+R-1}| \\
&= \gamma\left|h\cdot\sum_{r=1}^{R}\beta_{R-r}(\tilde{f}_{i+R-1-r} - f_{i+R-1-r}) + \varrho_{i+R-1}\right| \\
&\leq \gamma\left(h\sum_{r-1}^{R}|\beta_{R-r}|L|e_{i+R-1-r}| + |\varrho_{i+R-1}|\right) \\
&\overset{\beta:=\max_{r=1,\ldots,R}|\beta_i|}{\leq} \gamma\left(h\beta L\sum_{r=1}^{R}|e_{i+R-1-r}| + |\varrho_{i+R-1}|\right) \\
&\leq hL\beta\gamma^2\|E_{i-1}\|_0 + \gamma|\varrho_{i+R-1}|
\end{aligned}$$

$$\begin{aligned}
D_i^{-1} &= \operatorname{diag}((1-h\beta_R\sigma_{i+R-1})^{-1}, \ldots, (1-h\beta_R\sigma_i)^{-1}) \\
&\overset{\frac{1}{1-a}=1+\frac{a}{1-a}}{=} I + \operatorname{diag}\left(\frac{h\beta_R\sigma_{i+R-1}}{1-h\beta_R\sigma_{i+R-1}}, \ldots, \frac{h\beta_R\sigma_i}{1-h\beta_R\sigma_i}\right) \\
&= I + \tilde{D}
\end{aligned}$$

$$\begin{aligned}\frac{||D_i^{-1}x||_0}{||x||_0} &= \frac{||(I+\tilde{D})x||_0}{||x||_0} \\ &\leq 1 + \frac{\gamma \sum |\tilde{d}_{ii}x_i|}{\frac{1}{\gamma}\sum |x_i|} \\ &\leq 1 + \gamma^2 \max_{r=0,\ldots,R-1} \tilde{d}_{jj}\end{aligned}$$

nun gilt:

$$\begin{aligned}|h\beta_R\sigma_n| &< h|\beta_R|L < 1 \\ 1 - h|\beta_R\sigma_n| &\geq 1 - h|\beta_R|L > 0 \\ \Rightarrow ||D_i^{-1}||_0 &= \sup \frac{||D_i^{-1}||_0}{||x||_0} \leq 1 + \gamma^2 \frac{h|\beta_R|L}{1-h|\beta_R|L}\end{aligned}$$

Analog: $||C_i||_0 \leq 1 + \gamma^2 h|\beta_R|L$

(iv)

$$\begin{aligned}||E_i||_0 &\leq ||D_i^{-1}|| \left\{ ||C_i||_0 \underbrace{||A||_0}_{=\varrho(A)\leq 1} ||E_{i-1}|| + ||B_i||_0 \right\} \\ &\leq \left(1 + \frac{\gamma^2 h|\beta_R|L}{1-h|\beta_R|L}\right)\left\{\left(1+\gamma^2 h|\beta_R|L\right)||E_{i-1}||_0 + \gamma^2 h\beta L||E_{i-1}||_0 + \gamma|\varrho_{i+R-1}|\right\} \\ &= ||E_{i-1}||_0 + h\Gamma||E_{i-1}||_0 + \Lambda|\varrho_{i+R-1}| \\ &\leq ||E_{i-2}||_0 + h\Gamma||E_{i-2}||_0 + h\Gamma||E_{i-1}||_0 + \Lambda|\varrho_{i+R-1}| + \Lambda|\varrho_{i+R-2}| \\ &\leq h\Gamma \sum_{\nu=0}^{i-1} ||E_\nu||_0 + \underbrace{||E_0||_0 + \Lambda \sum_{\nu=1}^{i} |\varrho_{\nu+R-1}|}_{=:b_i}\end{aligned}$$

$$||E_i||_0 \leq \exp\left(\Gamma \underbrace{\sum_{\nu=0}^{i-1} h}_{t_i - t_0}\right)\left(||E_0||_0 + \Lambda \sum_{\nu=1}^{i} |\varrho_{\nu+R-1}|\right)$$

$\Rightarrow$ Behauptung

□

Satz 5.12 Konvergenz LMM
LMM sei Null-Stabil, AWA L-stetig

$$h < (|\beta_R|L)^{-1} \qquad \text{falls} \qquad \beta_R \neq 0$$

$$\delta_h := \max_{0\leq n\leq R-1} ||y_n - u(t_n)|| \to 0 \qquad (h \to 0)$$

folgt aus Konvergenz auch Konvergenz und es gilt die a-priori Fehlerabschätzung

$$||y_n - u(t_n)|| \leq \kappa e^{\Gamma(t_n - t_0)} \left\{ \delta_h + (t_n - t_0) \max_{R\leq\nu\leq n} ||\tau_\nu^h|| \right\}$$

Beweis .

$$\sum_{r=0}^{R} \alpha_{R-\nu} u(t_{n-r}) = h \sum_{r=0}^{R} \beta_{R-r} f(t_{n-r}, u(t_{n-r})) + h\tau_n^h$$

Behauptung folgt aus dem Stabilitätssatz für LMM 5.11 mit $\tilde{y}_n = u(t_n)$ und $\varrho_n = h\tau_n^h$, ziehe $\max \|\tau_n^h\|$ raus. □

Anwendung auf Adams-⋆ Verfahren

$$u(t_n) - u(t_{n-\sigma}) = \sum_{\mu_0}^{R} \int_{t_{n-\sigma}}^{t_n} L_\mu^{(m)}(s)\,\mathrm{d}s \cdot f(t_{h-\mu}, u(t_{h-\mu})) + h \underbrace{\tau_n^h}_{O(h^{R+1})}$$

$$(\alpha_R = 1, \alpha_{R-1} = -1)$$

$$\begin{aligned} \varrho(\lambda) &= \lambda^R - \lambda^{R-1} \\ &= \lambda^{R-1}(\lambda - 1) \end{aligned}$$

Bemerkungen:

- Nyström, Milne-Simpson LMM sind nullstabil
- BDF nur bis $R = 6$ nullstabil
- Dahlquist: Eine nullstabile LMM hat maximal Ordnung $R + 2$ (R ungerade), $R + 1$ (R gerade)

Stabilität für LMM

Betrachte: $u' = qu$

$$\sum_{r=0}^{R} (\alpha_{R-r} - hq\beta_{R-r}) y_{n-r} = 0$$

Stabilitätspolynom

$$\Pi(\lambda, qh) := \sum_{r=0}^{R} (\alpha_r - hq\beta_r)\lambda^r$$

$$\overline{h} = qh$$

Stabilitätsgebiet:

$$SG := \left\{ \overline{h} \in \mathbb{C}^- : \Pi(\lambda, \overline{h}) \text{ absolut stabil} \right\}$$

- optimale Verfahren mit Ordnung $R + 2$ bei R gerade (z.B. Simson) haben $SG = \{0\}$
- Explizite LMM haben immer beschränkte SG
- Maximale Ordnung für implizite und A-stabile LMM ist zwei.

BDF Verfahren sind $A(\alpha)$ stabil.
Zeichnung zu den Stabilitätsgebieten und Erläuterung, wo dort der Winkel α zu finden ist, der zur Charakterisierung der Verfahren verwendet wird.

Prädiktor-Korrektor Verfahren

- Idee: Fixpunktiteration LMM (implizit) (Korrektor).
- Startwert aus explizitem LMM
- Jeder Schritt kostet eine f-Auswertung (Evaluate) zur Berechnung von $f_n^{(k)} = f(t_n, y_n^{(k)})$

$\Rightarrow$ Schreibweise $P(EC)^k E$ für $k \geq 1$
Die Ordnung ist

$$P_m = \min(\underbrace{m^{(c)}}_{\text{Korrektor}}, \underbrace{m^{(p)}}_{\text{Prädiktor}} + \underbrace{k}_{\text{\# Iterationen}})$$

Milnes-Device

Die Fehlerkonstanten von Prädiktor (p) und Korrektor (c) lassen sich berechnen. Unter Annahme exakter Startwerte und Iteration bis zur Konvergenz erhält man damit:

$$\begin{aligned}
y_n^{(p)} &= u(t_n) + C_5^{(p)} h^5 u^{(5)}(t_n) + O(h^6) \\
y_n^{(c)} &= u(t_n) + C_5^{(c)} h^5 u^{(5)}(t_n) + O(h^6) \\
\Rightarrow u^{(5)}(t_n) &= \frac{y_n^{(c)} - y_n^{(p)}}{h^5 (C_5^{(p)} - C_5^{(c)})} + O(h) \\
\tau_5^{(c),h} &= C_5^{(c)} h^4 u^{(5)}(t_n) + O(h) \\
&= \frac{C_5^{(c)}}{C_5^{(p)} - C_5^{(c)}} \frac{y_n^{(c)} - y_n^{(p)}}{h} + O(h^5)
\end{aligned}$$

Die Schätzung für den lokalen Abschneidefehler $\tau_5^{(c),h}$ kann zur Schrittweitenkontrolle verwendet werden.

6 Randwertprobleme

nach [J. M. Melenk, Vorlesungsnotizen zu "Numerik von Differentialgleichungen", TU Wien]
Betrachte folgende Randwertaufgabe (RWA)

$$\begin{aligned}
u'(t) &= f(t, u(t)) \quad t \in I = [a, b] \\
Au(a) + Bu(b) &= c
\end{aligned} \tag{6.1}$$

Beispiel 6.1
DGL zweiter Ordnung

$$\begin{aligned}
u''(t) &= f(t, u(t), u'(t)) && t \in I = [a, b] \\
u(a) &= u_a \qquad u(b) = u_b
\end{aligned} \tag{6.2}$$

Umschreibe als System erster Ordnung:

$$\begin{aligned}
y(t) &= (y_1(t), y_2(t)) = (u(t), u'(t)) \\
y'(t) &= F(t, y(t)) = \begin{pmatrix} y_2(t) \\ f(t, y_1(t), y_2(t)) \end{pmatrix} \\
&\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1(a) \\ y_2(a) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} u_a \\ u_b \end{pmatrix}
\end{aligned} \tag{6.3}$$

Existenz und Eindeutigkeit: Schwieriger als bei AWA!

Beispiel 6.2

(a) $y'' + y = 0$ auf $[0, \frac{\pi}{2}]$, $y(0) = 0$, $y(\frac{\pi}{2}) = 1 \rightarrow y(t) = \sin(t)$

(b) $y'' + y = 0$ auf $[0, \pi]$, $y(0) = 0$, $y(\pi) = 0 \rightarrow y(t) = c\sin(t)$, $c \in \mathbb{R}$ löst diese RWA!

(c) $y'' + y = 0$ auf $[0, \pi]$, $y(0) = 0$, $y(\pi) = 1 \rightarrow$ hat keine Lösung

6.1 Schießverfahren

Idee

- Umformulieren in AWA mit Paramater
- Löse Gleichung für „das Treffen des Endwertes“

Demonstration am Beispiel 6.1:
Finde $y(t, s)$ so dass:

$$\frac{\partial y}{\partial t}(t, s) = F(t, y(t, s)) \qquad t \in [a, b] \qquad y(a) = \underbrace{\begin{pmatrix} u_a \\ s \end{pmatrix}}_{\text{nur hier taucht } s \text{ auf}}$$

s ist gegeben durch die Bedingung:

$$y_1(b, s) \stackrel{!}{=} u_b \tag{6.4}$$

Idee: Verwende Newton-Verfahren zur Lösung von (6.4). Dazu brauchen wir $\frac{\partial y_1}{\partial s}(b, s)$.

Lemma 6.3
Sei $y(t, s)$ für gegebenes $s \in \mathbb{R}$ die Lösung der AWA

$$\frac{\partial y}{\partial t}(t, s) = F(t, y(t, s)) \qquad t \in [a, b] \qquad y(a) = \begin{pmatrix} u_a \\ s \end{pmatrix} \tag{6.5}$$

mit F aus Beispiel 6.1. Dann ist die Funktion $v(t, s) = \frac{\partial y}{\partial s}(t, s)$ gegeben als Lösung der AWA:

$$\frac{\partial v}{\partial t}(t, s) = \begin{pmatrix} v_2(t, s) \\ \frac{\partial f}{\partial u}(t, y_1(t, s), y_2(t, s))v_1(t, s) + \frac{\partial f}{\partial u'}(t, y_1(t, s), y_2(t, s))v_2(t, s) \end{pmatrix} \tag{6.6}$$

mit den Anfangswerten $v(a, s) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Dies ist eine lineare AWA für v!

Beweis .

erste Komponente:

$$\frac{\partial}{\partial s}\frac{\partial y_1}{\partial t}(t,s) = \frac{\partial}{\partial s}y_2(t,s)$$

$$\Leftrightarrow \frac{\partial}{\partial t}\underbrace{\frac{\partial y_1}{\partial s}(t,s)}_{=:v_1(t,s)} = v_2(t,s)$$

zweite Komponente:

$$\frac{\partial}{\partial s}\frac{\partial y_2}{\partial t}(t,s) = \frac{\partial}{\partial s}f(t,y_1(t,s),y_2(t,s))$$

$$\frac{\partial}{\partial t}v_2(t,s) = \frac{\partial f}{\partial u}(t,y_1(t,s),y_2(t,s)v_1(t,s) + \frac{\partial f}{\partial u'}(t,y_1(t,s),y_2)t,s))v_2(t,s)$$

Anfangswerte:

$$v_1(a,s) = \frac{\partial y_1}{\partial s}(a,s) = \frac{\partial}{\partial s}u_a = 0$$

$$v_2(a,s) = \frac{\partial y_2}{\partial s}(a,s) = \frac{\partial}{\partial s} = 1$$

□

Einfaches Schießverfahren:

1) Wähle Startwert $s^{(0)}$
2) Berechne Lösung $y^{(i)}$ von (6.5) zum Startwert $s^{(i)}$ numerisch
3) Falls $|u_b - y_1^{(i)}(b,s^{(i)})| < \varepsilon \rightarrow$ FERTIG!
4) Berechne $\frac{\partial y^{(i)}}{\partial s}$ durch Lösen der AWA (6.6) numerisch
5) Newton-Schritt:
$$s^{(i+1)} = s^{(i)} + \left(\frac{\partial y_1}{\partial s}(b,s^{(i)})\right)^{-1}\left(u_b - y_1^{(i)}(b,s^{(i)})\right)$$
6) $i = i+1$, gehe nach 2)

Man zeigt: Für genügend kleines h konvergiert das Schießverfahren mit der Ordnung $O(h^m)$ (m Ordnung der Verfaren in Schritt 2/4).
Problem des Schießverfahrens:

$$\text{Stabilitätssatz } \|y(b,s) - y(b,s+\varepsilon)\| \leq ce^{L\overbrace{(b-a)}^{T}}\varepsilon$$

$\rightarrow$ Funktioniert nur gut, falls e^{LT} hinreichend klein. Beispiel: $L = 10$, $T = 10$, dann ist $e^{LT} = e^{100}$!
Abhilfe: „multiple shooting“ $M = 10 \rightarrow T = 1$, $e^{LT} = e^{10\cdot 1}$
Zerlege $[a,b]$ in M Teilintervalle $[a_i,b_i]$, $a_1 = a$, $a_{i+1} = b_i$, $1 \leq i \leq M$, $b_M = b$
(Skizze zur Erläuterung. Da wir mittlerweile drei Indizes haben sind die oben rechts *ohne* Klammern die für Intervalle)

Zu bestimmen sind nun $s_1, \cdots, s_M$ sowie interne Startwerte $w_2, \cdots, w_m$ sodass:

$$\begin{aligned} y_1^i(b_i, s_i) &= \begin{cases} w_{i+1} & 1 \le i < M \\ u_b & i = M \end{cases} \\ y_2^i(b_i, s_i) &= s_{i+1} \qquad 1 \le i < M \end{aligned} \tag{6.7}$$

(insgesamt also $2M - 1$ Bedingungen) mit den Startwerten

$$\begin{aligned} y_1^i(a_i, s_i) &= \begin{cases} u_a & i = 1 \\ w_i & 2 \le i \le M \end{cases} \\ y_2^i(a_i, s_i) &= s_i \qquad 1 \le i \le M \end{aligned} \tag{6.8}$$

→ Löse (6.7) mit dem Newton-Verfahren.

6.2 Differenzverfahren

$$u'(t_n) = \frac{u(t_n) - u(t_{n-1})}{h_n} + \mathcal{O}(h_n) = f(t_n, u(t_n))$$

$$\begin{aligned} \Rightarrow y_n - y_{n-1} - h_n f(t_n, y_n) &= 0 \qquad\qquad 1 \le n \le N \\ Ay_0 + By_n - c &= 0 \end{aligned}$$

Nichtlineares GLS für die $N + 1$ „Unbekannten“ $y_0, \cdots, y_N$
→ Löse mit Newton/Fixpunktiteration

Weitere Probleme bei der Lösung gewöhnlicher DGL

„Strukturerhaltende“ Integratoren

Physikalische Systeme erfüllen Massen, Energie, Impuls, . . . -Erhaltung
Leisten dies auch numerische Verfahren?

Beispiel

Leapfrog-Verfahren für N-Körper-Problem (Newtonsche Bewegungsgleichung)

$$\begin{aligned} x_i'(t) &= v_i(t) \\ v_i'(t) &= G \sum_{\substack{j=1 \\ j \ne i}}^{N} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3} =: a_i(x) \end{aligned}$$

(Skizze zum Verfahren/Beispiel)

$$\begin{aligned} \frac{x_i^{(n)} - x_i^{(n-1)}}{h} &= v_i(t) \qquad \Rightarrow \qquad x_i^{(n)} = x_i^{(n+1)} + h v_i^{n-1/2} \\ \frac{v_i^{(n+1/2)} - v_i^{(n-1/2)}}{h} &= a_i(x^{(n)}) \quad \Rightarrow \quad v_i^{(n+1/2)} = v_i^{(n-1/2)} + h a_i(x^{(n)}) \end{aligned}$$

Eigenschaften:

1) Invariant gegenüber Zeitumkehr

2) Drehimpuls erhaltend, nicht Energier erhaltend, aber Energiefehler bleibt beschränkt ⇒ Langszeitstabilität

3) „symplektisch“: Fläche im Phasenraum bleibt erhalten (Eigenschaft der DGL aus der klassischen Mechanik) wird exakt erfüllt.

Galerkinverfahren:

Grundlage ist die “Variationelle Formulierung” der AWA:
Finde $u \in U = \left\{ v \in C^1(I)^d : v(t_0) = u_0 \right\}$ so dass

$$\int_I (u' - f(t,u), \varphi)\, dt = 0 \qquad \forall \varphi \in V = C^0(I)^d$$

Dies ist eine äquivalente Formulierung der AWA. (siehe Ranacher).
Numerische Näherung: Ersetze U, V durch endlich dimensionale Räume, etwa Polynome.
Problem: ergibt gekoppeltes Problem wie bei RWA:
Idee: Stückweise Anwendung auf Teilintervallen mit geeigneten Kopplungsbedingungen, die stetige Differenzierbarkeit sicherstellen.

$$\sum_{n=1}^{N} \left\{ \int_{I_n} (u' - f(t,u), \varphi)\, dt + ([u]_{n-1}, \varphi^+_{n-1}) \right\} = 0 \qquad \forall \varphi \in V = C^0(I)^d$$

“unstetiges Galerkinverfahren”.
Vorteile:

- verläßliche Schrittweitensteuerung
- a posteriori Fehlerschätzung möglich
- Monotonie-Eigenschaften der AWA bleiben erhalten

Kriterien für die Auswahl von Lösungsverfahren

- Genauigkeit (Konvergenzordnung)
- Rechenaufwand
- Implementierungsaufwand
- Numerische Stabilität (besonders für steife DGL)
  - A-stabil aber nicht L-stabil für Schwingungsprobleme
  - L-stabil für gedämpfte (monotone) AWA
- Schrittweitensteuerung

- Fehlerkontrolle
- Konservierung von Erhaltungsgrößen
- Regularität der Lösung

7 Partielle Differentialgleichungen

Einleitung:

PDGL entstehen in diversen Disziplinen der Naturwissenschaften.

Beispiel: Wärmeleitung

(Zeichnung)
offen, beschr., Lipschitzrand
Sei $\omega \subseteq \Omega$ ein Testvolumen

$$Q_u(t) = \int_i c(x)\varrho(x)T(x,t)\,\mathrm{d}x$$

$$[c] = \frac{j}{Kkg} \qquad \text{Spez. Wärme}$$

$$[\varrho] = \frac{kg}{m^3} \qquad \text{Dichte}$$

$$[T] = K \qquad \text{Temperatur}$$

$$Q_u(t+\Delta t) - Q_u(t) = \int_t^{t+\Delta t} \left\{ \int_{\partial\omega} j(x,t)\cdot\nu(x)\,\mathrm{d}s + \int_\omega q(x,t)\,\mathrm{d}x \right\} \mathrm{d}t$$

$$\Rightarrow \qquad \int_\omega c(x)\varrho(x)\frac{T(x,t+\Delta t)-T(x,t)}{\Delta t}\,\mathrm{d}x = \int_\omega -\nabla\cdot j(x,t) + q(x,t)\,\mathrm{d}x + O(\Delta t)$$

$$\Rightarrow \qquad \int_\Omega c(x)\varrho(x)\partial_t T(x,t)\chi_\omega(x)\,\mathrm{d}x = \int_\Omega (-\nabla\cdot j + q)\chi_\omega(x)\,\mathrm{d}x$$

$$\chi_\omega(x) = \begin{cases} 1 & x \in \omega \\ 0 & \text{sonst} \end{cases}$$

Da $\omega \subseteq \Omega$ beliebig gewählt:

$$c(x)\varrho(x)\partial_t T(x,t) = -\nabla j(x,t) + q(x,t) \quad \forall\, x \in \Omega$$

Wenn $q(x) \equiv 0$, dann nennt man Vektorfeld j divergenzfrei.
PDGL dieser Form treten auch für andere physikalische Erhaltungsgrößen auf, z.B. Masse oder Impuls.

Wärmefluss:

- Konduktiver Fluss: (Wärmeübertragung durch thermische Bewegung der Atome)

$$j_d(x,t) = -\lambda \nabla T(x,t) \qquad \lambda \text{ Wärmeleitfähigkeit}$$

- Konvektiver Fluss
(hübsche Zeichnung mit Fläche von $\omega(t) \to (\omega(t+\tau)$ im Gebiet Ω)

$$j_c(x,t) = c(x)\varrho(x,t)v(x,t)T(x,t)$$

Gesamtfluss:

$$j = j_c + j_d$$

Um $T(x,t)$ eindeutig festzulegen brauchen wir:

  - Anfangsbedingungen $T(x,t_0) = T_0(x) \;\; \forall\, x \in \Omega$
  - Randbedingungen: $T(x,t) = g(x,t) \;\; \forall\, x \in \partial\Omega,\, t \in [t_0, t_{\text{end}}]$

Allgemein bezeichnet man

$$\partial_t C(x,t) + \nabla \cdot [v(x,t)C(x,t) - D(x,t)\nabla C(x,t)] = q(x,t)$$

als Konvektions-Dispersionsgleichung.

Weitere Beispiele

Poisson Gleichung:

$$\begin{aligned} \partial_t T &= 0 \text{ (stationär)} \\ v &= 0 \text{ (keine Konv.)} \\ \lambda &= 1 \\ -\nabla \cdot (\nabla T) &= q \qquad (-\Delta T = q) \\ T(x) &= g(x) \qquad \forall x \in \partial\Omega \end{aligned}$$

Grundwassergleichung:

$$\begin{aligned} \partial_t \varphi(x,t) + \nabla \cdot j_w(x,t) &= f(x,t) \quad (\forall (x,t) \in \Omega \times \Sigma) \qquad (\Sigma\text{: Zeitintervall}) \\ j_w(x,t) &= -K(x)\left[\nabla p(x,t) - \varrho(x,t) g e_z\right] \end{aligned}$$

$\varphi(x,t)$: Porosität, $j_w(x,t)$: Wasserflussdichte, $f(x,t)$: Quell-/Senkenterm,
$K(x)$: hydraulischer Leitfähigkeitstensor, p: Wasserdruck, g: Erdbeschleunigung,
e_z: Einheitsvektor in z-Richtung
Mit $\partial_t \varphi(x,t) = 0$:

$$\nabla \cdot [K(x)\nabla p(x,t)] = f - \nabla \cdot [K(x)\varrho(x,t) g e_z]$$

Typeneinteilung PDGL

Allgemeine Definition: PDGL determiniert eine Funktion $u \in C^m(\mathbb{R}^n)$ durch:

$$F\left(\frac{\partial^m u}{\partial^m x_1}, \ldots, \frac{\partial^m u}{\partial^m x_n}, \frac{\partial^{m-1} u}{\partial^{m-1} x_1}, \ldots \frac{\partial u}{\partial x_n}, u\right) = 0 \qquad \forall x \in \Omega$$

Zur Eindeutigkeit fehlen noch zusätzliche Bediungen.
Eine spezielle Klasse sind die linearen PDGL. Für $n = 2$ und $m = 2$:

$$\begin{aligned} & a\partial^2_{xx} u + 2b\partial^2_{xy} u + c\partial^2_{yy} u && \leftarrow \text{Hauptteil} \\ & \quad + d\partial_x u + e\partial_y u + fu + g = 0 \end{aligned}$$

Hauptteil schreiben wir als:

$$\begin{pmatrix} \partial_x u \\ \partial_y u \end{pmatrix}^T \underbrace{\begin{pmatrix} a & b \\ b & c \end{pmatrix}}_{=:A} \begin{pmatrix} \partial_x u \\ \partial_y u \end{pmatrix}$$

Es gilt: $\det(A) = ac - b^2$.

Definition 7.1 Typeneinteilung
Wir nennen eine lineare PDGL zweiter Ordnung...

1) elliptisch, wenn $ac - b^2 > 0$
2) hyperbolisch, wenn $ac - b^2 < 0$
3) parabolisch, wenn $\det(A) = 0$

Die Namensgebung folgt aus dem Typ der Gleichung $x^T Dx = 0$, wobei D die Diagonalisierte von A ist. $(\alpha, \beta \geq 0)$

$$\begin{aligned} \text{Elliptisch:} \quad & x^T Dx = \alpha x_1^2 + \beta x_2^2 && = 0 \\ \text{Hyperbolisch:} \quad & x^T Dx = \alpha x_1^2 - \beta x_2^2 && = 0 \\ \text{Parabolisch:} \quad & x^T Dx = \alpha x_1^2 && = 0 \end{aligned}$$

Definition 7.2 Typeinteilung in höheren Raumdimensionen

$$\underbrace{\sum_{i,j=1}^{n} a_{ij}(x)\partial_{x_i}\partial_{x_j} u}_{\text{Hauptteil}} + \sum_{i=1}^{n} a_i(x)\partial_{x_i} u + a_0(x) = 0 \text{ in } \Omega$$

o.B.d.A $a_{ij} = a_{ji}(x)$, Setze $(A(x))_{ij} = a_{ij}(x)$, $a(x) = (a_1(x), \ldots, a_n(x))^T$

1) Elliptisch in $x \in \Omega$, falls alle EW von A gleiches Verzeichen, kein EW Null
2) Hyperbolisch in $x \in \Omega$, falls kein EW Null und $n - 1$ gleiches Vorzeichen, 1 EW entgegenes VZ.
3) Parabolisch in $x \in \Omega$, falls EW Null, restl. $n - 1$ gleiches VZ, $\text{Rang}[A(x), a(x)] = n$

Bemerkungen:

- Für $n = 2$ ist Typeinteilung vollständig, für $n > 2$ nicht
- Warum diese Einteilung?
  - Theorie für Existenz und Eindeutigkeit hängt vom Typ ab.
  - Entsprechend hängen numerische Methoden vom Typ ab!
  - Notwendige Rand- und Anfangsbedingungen hängt ebenfalls von Typ ab!
- Invarianz gegenüber Koordinatentransformation:
  $n = 2$: $u(x, y)$ sei Lösung der PDGL. $\xi = \xi(x, y)$, $\eta = \eta(x, y)$

  $$\hat{u}(\xi(x, y), \eta(x, y)) = u(x, y)$$

  $\Rightarrow$ PDGL für $\hat{u}$: $\hat{L}\hat{u} = \hat{f}$ mit anderen Koeffizienten $\hat{a}_{ij}$, $\hat{a}_i$, $\hat{a}_0$. Es gilt:
  Hat L am Punkt (x, y) den Typ T, so
  hat $\hat{L}$ am Punkt $(\xi(x, y), \eta(x, y))$ ebenfalls den Typ T.
- Typ hängt nur vom Hauptteil ab

Beispiele für verschiedene Typen

Poisson-Gleichung

$$\frac{\partial u}{\partial x^2} + \frac{\partial u}{\partial y^2} = f \text{ in } \Omega$$

Prototyp für *elliptische* PDGL. Typische Randbedingungen:

1) $u(x, y) = g(x, y)$ für $(x, y) \in \partial\Omega$ „Dirichlet“ (Γ_D)

2) $\frac{\partial u}{\partial \nu}(x, y) = j(x, y)$ für $(x, y) \in \partial\Omega$ „Neumann“ (Γ_N)

$\Rightarrow$ genau eine Bedingung 1) oder 2) muss an jedem Punkt des Randes vorgegeben werden $\Rightarrow$ „Randwertproblem“.
$\Gamma_D \cup \Gamma_N = \partial\Omega$, $\text{meas}(\Gamma_D) \neq \varnothing$
allgemeine Diffusionsgleichung:

$$\begin{aligned} -\nabla \cdot \{K(x)\nabla u\} &= f \text{ in } \Omega \qquad K(x) \text{ Tensor!} \\ u &= g \text{ auf } \Gamma_D \subseteq \partial\Omega \\ -(K(x)\nabla u) \cdot \nu &= j \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D \end{aligned}$$

mit

1) $K(x) = K^T(x)$ und $\xi^T K(x)\xi > 0 \forall \xi \in \mathbb{R}^n$, $\xi \neq 0$, $x \in \Omega$

2) $\alpha(x) := \inf \left\{ \xi^T K(x)\xi \;\middle|\; \|\xi\| = 1 \right\} \geq \alpha_0 > 0$ (gleichmäßige Elliptizität)

$\Rightarrow$ wir betrachten hauptsächlich elliptische PDGL!

Wellengleichung

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \text{ in } \Omega$$

Prototyp für hyperbolische PDGL (2. Ordnung)
Beispiele für Randbedingungen:
$y = 0$: $u(x,0) = u_0(x)$, $\frac{\partial u}{\partial y}(x,0) = u_1(x)$. } „Anfangbedingungen“

$$\left.\begin{array}{l} x = 0\text{: } u(0,y) = g_0(y) \\ x = 1\text{: } u(1,y) = g_1(y) \end{array}\right\} \text{„Randbedingungen“}$$

Wärmeleitungsgleichung

$$\frac{\partial^2 u}{\partial x^2} \overbrace{-}^{\text{Vorzeichen ist aus Typdefinition nicht klar!}} \frac{\partial u}{\partial y} = f \text{ in } \Omega$$

Prototyp einer parabolischen PDGL.
Beispiele für Randbedingungen:
$y = 0$: $u(x,0) = u_0(x)$ „Anfangbedingung“

$$\left.\begin{array}{l} x = 0\text{: } u(0,y) = g_0(y) \\ x = 1\text{: } u(1,y) = g_1(y) \end{array}\right\} \text{„Randbedingungen“}$$

Definition 7.4 Sachgemäß gestellt
Eine partielle DGL heißt „sachgemäß gestellt“, falls

1) Eine eindeutige Lösung existiert und

2) diese *stetig* von den Daten (Rand- und Anfangsbedingung, Koeffizienten) abhängt.

Formal: $L\colon X \to Y$, x, y normierte Funktionenräume.

1) zu jedem $f \in Y\, \exists\, u \in X$ sodass $L(u) = f$

2) Es gibt $C > 0$, unabhängig f sodass: $\|u\|_X \leq C\|f\|_Y$

Transportgleichung

$$\nabla \cdot \{\vec{v}(x)u(x)\} = f \text{ in } \Omega \qquad \vec{v}(x) \text{ gegebenes Vektorfeld } \vec{v}\colon \mathbb{R}^n \to \mathbb{R}^n$$

(modelliert rein konvektiven, stationären Wärmetransport)
Mögliche Randwerte: $u(x) = g(x)$ für $x \in \partial\Omega$ mit $\vec{v}(x) \cdot \nu(x) < 0$
„hyperbolische Gleichung“ erster Ordnung (Einziger Typ für Gleichungen erster Ordnung)

Grenzen des klassischen Lösungsbegriffs

$$\begin{aligned} \Delta u &= f \text{ in } \Omega \\ u &= g \text{ auf } \partial\Omega \end{aligned} \tag{7.1}$$

Definition 7.5 klassische Lösung
Eine Funktion $u \in C^2(\Omega) \wedge C^0(\overline{\Omega})$ welche (7.1) löst heißt „klassische Lösung“.

Polarkoordinaten:

$$\begin{aligned} x(r,\varphi) &= r\cos\varphi \qquad y(r,\varphi) = r\sin\varphi \\ \hat{u}(r,\varphi) &= u(x(r,\varphi), y(r,\varphi)) \text{ wobei } \Delta u = f \end{aligned}$$

Gesucht: Gleichung für $\hat{u}(r,\varphi)$

$$\frac{\partial^2\hat{u}}{\partial r^2} + \frac{1}{r}\frac{\partial\hat{u}}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial\varphi^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f \tag{7.2}$$

Einspringende Ecke

(Schaubild für das Beispiel)

$$\Omega = \{(r,\varphi) \mid 0 < r < 1, 0 < \varphi < \Phi\}$$

Die Funktion $\hat{u}(r,\varphi) = r^k \sin(k\varphi)$ löst die Laplace-Gleichung $\Delta u = 0$ in Polarkoordinaten.
Speziell $k = \frac{\pi}{\Phi}$: $s(r,\varphi) = r^{\frac{\pi}{\Phi}} \sin(\frac{\pi}{\Phi}\varphi)$.

$$\frac{\partial s}{\partial r}(r,\varphi) = \frac{\pi}{\Phi} r^{\frac{\pi}{\Phi}-1} \sin(\frac{\pi}{\Phi}\varphi)$$

$\Rightarrow \frac{\partial s}{\partial r}(0,0)$ wird ∞, falls $\Phi > \pi$ (nicht konvexes Gebiet)

Gravitationsfeld einer Kugel

$$\begin{aligned} \Omega_i &= \left\{ x \in \mathbb{R}^3 \;\middle|\; \|x\| < R \right\} \\ \Omega_a &= \left\{ x \in \mathbb{R}^3 \;\middle|\; \|x\| > R \right\} \\ \Omega &= \mathbb{R}^3 \end{aligned}$$

in Ω_i:
$\Delta\Phi_i = 4\pi\gamma\varrho$

in Ω_a:
$\Delta\Phi_a = 0$ in Ω_a
$\Phi_a(x) = 0$ für $\|x\| \to \infty$

Übergangsbedingungen auf $\Gamma = \partial\Omega_i \cap \partial\Omega_a$:

$$\begin{aligned} \Omega_a(x) &= \Omega_i(x) && x \in \Gamma \\ \nabla\Omega_a(x)\cdot\nu_i &= \nabla\Omega_i(x)\cdot\nu_i \end{aligned}$$

Gesamtproblem:

$$\nabla\Phi = f \text{ in } \Omega \qquad f(x) = \begin{cases} 4\pi\gamma\varrho & \|x\| < R \\ 0 & \text{sonst} \end{cases}$$
$$\Phi(x) = 0 \qquad \|x\| \to \infty$$

$\Rightarrow$ unstetige rechte Seite. Wird durch klassische Lösung nicht erfasst.

8 Finite Differenzen Verfahren

Der eindimensionale Fall

Betrachte die eindimensionale RWA:

$$\begin{aligned} -u''(x) &= f(x) \quad x \in (0,1) \\ u(0) &= \varphi_0 \qquad u(1) = \varphi_1 \end{aligned} \tag{8.1}$$

Taylorreihenentwicklung:

$$\begin{aligned} u(x+h) &= u(x) + h\underset{\sim}{u'(x)} + \frac{h^2}{2}u''(x+\delta^+ h) \qquad \delta^+ \in [0,1] \\ \Leftrightarrow u'(x) &= \frac{u(x+h)-u(x)}{h} - \frac{h}{2}u''(x+\delta^+ h) \end{aligned} \tag{8.2}$$

$$\begin{aligned} u(x-h) &= u(x) - h\underset{\sim}{u'(x)} + \frac{h^2}{2}u''(x-\delta^- h) \qquad \delta^- \in [0,1] \\ \Leftrightarrow u'(x) &= \frac{u(x)-u(x-h)}{h} + \frac{h}{2}u''(x-\delta^- h) \end{aligned} \tag{8.3}$$

Man setzt:

$$(\partial^+ u)(x) \coloneqq [u(x+h) - u(x)]/h \qquad \text{„Vorwärtsdifferenz“}$$
$$(\partial^- u)(x) \coloneqq [u(x) - u(x-h)]/h \qquad \text{„Rückwärtsdifferenz“}$$

Zweite Ableitung:

$$\begin{aligned} (\partial^-\partial^+ u)(x) &= \partial^-\left(\frac{u(x+h)-u(x)}{h}\right) = \left[\frac{u(x+h)-u(x)}{h} - \frac{u(x)-u(x-h)}{h}\right]/h \\ &= \frac{u(x+h)-2u(x)+u(x-h)}{h^2} = \frac{\partial^2 u}{\partial x^2}(x) + \mathcal{O}(h^2) \end{aligned}$$

Dies rechnen wir nach

$$
\begin{aligned}
(\partial^+ u)(x) &= u'(x) + \frac{h}{2}u''(x) + \frac{h^2}{6}u'''(x) + \frac{h^3}{24}u^{iv}(x) + \dots \\
(\partial^- u)(x) &= u'(x) - \frac{h}{2}u''(x) + \frac{h^2}{6}u'''(x) - \frac{h^3}{24}u^{iv}(x) + \dots \\
(\partial^-\partial^+ u)(x) &= \partial^- \left(u'(x) + \frac{h}{2}u''(x) + \frac{h^2}{6}u'''(x) + \frac{h^3}{24}u^{iv}(x) \right) \\
&= \left(u''(x) - \cancel{\frac{h}{2}u'''(x)} + \frac{h^2}{6}u^{iv}(x) - \dots \right) + \frac{h}{2}\left(\cancel{u'''(x)} - \frac{h}{2}u^{iv}(x) + \dots \right) + \frac{h^2}{6}\left(u^{iv}(x) - \dots \right) \\
&= u''(x) + h^2\left(\frac{1}{6} - \frac{1}{4} + \frac{1}{6} \right) u^{iv}(x) + \dots \\
&= u''(x) + \frac{h^2}{12}u^{iv}(x) + O(h^4)
\end{aligned}
$$

Formel zweiter Ordnung für die erste Ableitung:

$$
\begin{aligned}
u(x+h) - u(x-h) &= 2hu'(x)\, O(h^3) \\
\Leftrightarrow u'(x) &= \frac{u(x+h) - u(x-h)}{2h} + O(h^2) \qquad \text{„zentraler Differenzenquotient“}
\end{aligned}
$$

Lemma 8.1 Differenzenformel

$$(\partial^+ u)(x) = \frac{u(x+h) - u(x)}{h} = u'(x) + hR \text{ mit } |R| \le \frac{1}{2}\|u\|_{C^2(\overline{\Omega})} \text{ falls } u \in C^2(\overline{\Omega}) \tag{8.4a}$$

$$(\partial^- u)(x) = \frac{u(x) - u(x-h)}{h} = u'(x) + hR \text{ mit } |R| \le \frac{1}{2}\|u\|_{C^2(\overline{\Omega})} \text{ falls } u \in C^2(\overline{\Omega}) \tag{8.4b}$$

$$\|u\|_{C^k(\overline{\Omega})} = \sup_{x\in\Omega} \max_{i=0,\dots,k} |u^{(i)}(x)|$$

$$\frac{u(x+h) - u(x-h)}{2h} = u'(x) + hR \text{ mit } |R| \le \frac{1}{6}\|u\|_{C^3(\overline{\Omega})} \tag{8.4c}$$

$$\frac{u(x-h) - 2u(x) + u(x+h)}{h^2} = u''(x) + h^2 R \text{ mit } |R| \le \frac{1}{12}\|u\|_{C^4(\overline{\Omega})} \tag{8.4d}$$

Bemerkung:

- Für die Formeln zweiter Ordnung sind äquidistante Gitter notwendig!
- $u \in C^4(\overline{\Omega})$ ist sehr starke Forderung, die in der Regel nicht erfüllt ist!

Zurück zum RWP (8.1):
Unterteile $\Omega = (0,1)$ in N Teilintervalle ($N \in \mathbb{N}$)

$$[x_i, x_{i+1}] \qquad i = 0, \dots, N-1 \qquad x_i = ih \qquad h = \frac{1}{N}$$

„äquidistantes Gitter“

$$\Omega_h = \{ih \mid i \in \mathbb{N}_0 \wedge 0 < i < N\}$$
$$\overline{\Omega}_h = \{ih \mid i \in \mathbb{N}_0 \wedge 0 \leq i \leq N\}$$

Für $u \in C^4(\overline{\Omega})$ gilt:

$$-\frac{1}{h^2}[u(x-h) - 2u(x) + u(x+h)] = f(x) + O(h^2) \qquad \forall\, x \in \Omega_h \tag{8.5}$$

(analog zu lokalem Abschneidefehler)
Streichen des Fehlerterms liefert $|\Omega_h| = N - 1$ lineare Gleichungen:

$$-\frac{1}{h^2}[u_h(x-h) - 2u_h(x) + u_h(x+h)] = f(x) \qquad \forall\, x \in \Omega_h \tag{8.6a}$$

Zusätzlich gelten die Randbedingungen

$$u_h(0) = \varphi_0 \qquad u_h(1) = \varphi_1 \tag{8.6b}$$

$u_h : \overline{\Omega}_h \to \mathbb{R}$ heißt „Gitterfunktion“.
Alternativ fassen wir u_h als Vektor im $\mathbb{R}^{N-1}$ auf:

$$u_h = (u_h(h), u_h(2h), \ldots, u_h(1-h))^T$$

Elimination der Randwerte aus (8.6a) mittels (8.6b) ergibt ein LGS für u_h (den Vektor!)

$$\underbrace{\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}}_{L_h} \underbrace{\begin{bmatrix} u_h(h) \\ u_h(2h) \\ u_h(3h) \\ \vdots \\ \vdots \\ u_h(1-h) \end{bmatrix}}_{u_h} = \underbrace{\begin{bmatrix} f(h) + \frac{\varphi_0}{h^2} \\ f(2h) \\ \vdots \\ \vdots \\ f(1-2h) \\ f(1-h) + \frac{\varphi_1}{h^2} \end{bmatrix}}_{q_h}$$

- L_h ist Triagonalmatrix (Spezialfall einer Bandmatrix)
- L_h ist dünn besetzt ($O(N)$ Einträge statt N^2), maximal 3 Einträge pro Zeile
- L_h symmetrisch und positiv definit
- Ist in linearer Laufzeit lösbar mit Gauß-Algorithmus (Heißt für Tridiagonalmatrizen auch Thomas-Verfahren).

Der mehrdimensionale Fall

$$-\Delta u(x) = -\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(x) = f(x) \qquad x \in \Omega = (0,1)^n \tag{8.7a}$$
$$u(x) = g(x) \qquad x \in \partial\Omega \tag{8.7b}$$

Mit e_i kartesischer Einheitsvektor in Richtung i:

$$\Delta u(x) = \frac{1}{h^2}\left\{\sum_{i=1}^{n}\left[u(x-he_i) - 2u(x) + u(x+he_i)\right]\right\} + h^2 R \qquad \text{mit } |R| \leq \frac{n}{12}\|u\|_{C^4(\overline{\Omega})}$$

Wir nennen

$$\Delta_h u(x) := \frac{1}{h^2}\left(-2nu(x) + \sum_{i=1}^{n}\left[u(x-h) + u(x+h)\right]\right)$$

den "diskreten Laplace-Operator"
$n = 2$: (Zeichnung: ein quadratisches Gitter mit eingezeichneten inneren Punkten an den Schnittstellen der Striche und Randpunkten)

$$\Omega_h = \{(x,y) \in \Omega | x/h, y/h \in \mathbb{Z}\}$$
$$\overline{\Omega}_h = \left\{(x,y) \in \overline{\Omega} | x/h, y/h \in \mathbb{Z}\right\}$$

mit $h = 1/N$ und setze $\Gamma_h := \overline{\Omega}_h \setminus \Omega_h$ (Randpunkte)
$(N-1)^2$ Gleichungen:

$$-\Delta_h u_h(x) = f(x) \qquad x \in \Omega_h \tag{8.8a}$$

$4N$ Randbedingungen:

$$u(x) = g(x) \qquad x \in \Gamma_h \tag{8.8b}$$

LGS bei zeilenweise lexikographischer Anordnung:

$$\underbrace{\frac{1}{h^2}\left[\begin{array}{ccc|ccc|ccc|ccc}
4 & -1 & & -1 & & & & & & & & \\
-1 & \ddots & -1 & & \ddots & & & & & & & \\
 & -1 & 4 & & & -1 & & & & & & \\
\hline
-1 & & & 4 & -1 & & -1 & & & & & \\
 & \ddots & & -1 & \ddots & -1 & & \ddots & & & & \\
 & & -1 & & -1 & 4 & & & -1 & & & \\
\hline
 & & & -1 & & & 4 & -1 & & -1 & & \\
 & & & & \ddots & & -1 & \ddots & -1 & & \ddots & \\
 & & & & & -1 & & -1 & 4 & & & -1 \\
\hline
 & & & & & & -1 & & & 4 & -1 & \\
 & & & & & & & \ddots & & -1 & \ddots & -1 \\
 & & & & & & & & -1 & & -1 & 4
\end{array}\right]}_{L_h}
\underbrace{\left[\begin{array}{c} u(h,h) \\ u(2h,h) \\ \vdots \\ \hline u(h,2h) \\ \\ \hline \\ \\ \hline \\ \\ \\ \end{array}\right]}_{u_h}
=
\underbrace{\left[\begin{array}{c} f(h,h) + {}^{[g(0,h)+g(h,0)]}/_{h^2} \\ f(2h,h) + {}^{g(h,0)}/_{h^2} \\ \hline \\ \\ \hline \\ \\ \hline \\ \\ \\ \end{array}\right]}_{q_h}$$

„Differenzenstern" (Andere Formulierung für das obige)

$$\begin{bmatrix} c_{-1,1}(x,y) & c_{0,1}(x,y) & c_{1,1}(x,y) \\ c_{-1,0}(x,y) & c_{0,0}(x,y) & c_{1,0}(x,y) \\ c_{-1,-1}(x,y) & c_{0,-1}(x,y) & c_{1,-1}(x,y) \end{bmatrix} u_h(x,y) := \sum_{i=-1}^{1}\sum_{j=-1}^{1} c_{i,j}(x,y) \cdot u_h(x+ih, y+jh)$$

entspricht einer Zeile des LGS.

In unserem Beispiel:

$$\frac{1}{h^2}\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}$$ „Fünfpunktestern“

Neumann-Randbedingungen

(Grafik)

$$\frac{\partial u}{\partial \nu} = \pm\frac{\partial u}{\partial x},\ \pm\frac{\partial u}{\partial y}$$

Diskretisierung mit entsprechenden Vorwärts- oder Rückwärtsdifferenzen
(Zeichnung des Gitters)

$$\begin{aligned}
\frac{\partial u}{\partial \nu} &= -\frac{\partial u}{\partial x} = \frac{u(x,y)-u(x+h,y)}{h} + O(h) \text{ (rechts)}\\
\frac{\partial u}{\partial \nu} &= \ \ \frac{\partial u}{\partial x} = \frac{u(x,y)-u(x-h,y)}{h} + O(h) \text{ (links)}\\
\frac{\partial u}{\partial \nu} &= -\frac{\partial u}{\partial y} = \frac{u(x,y)-u(x,y+h)}{h} + O(h) \text{ (unten)}\\
\frac{\partial u}{\partial \nu} &= \ \ \frac{\partial u}{\partial y} = \frac{u(x,y)-u(x,y-h)}{h} + O(h) \text{ (oben)}
\end{aligned}$$

Allgemeine lineare PDE 2. Ordnung

Die allgemeine lineare PDGL zweiter Ordnung für $n = 2$ lautet:

$$a_{xx}(x,y)\frac{\partial^2 u}{\partial x^2} + 2a_{xy}(x,y)\frac{\partial^2 u}{\partial x \partial y} + a_{yy}(x,y)\frac{\partial^2 u}{\partial y^2} + a_x(x,y)\frac{\partial u}{\partial x} + a_y(x,y)\frac{\partial u}{\partial y} + a(x,y)u = f(x,y) \qquad (8.9)$$

Diskretisierung von $\frac{\partial^2 u}{\partial x \partial y}$

$$\begin{aligned}
u(x+h,y+h) &= u(x,y+h) &&+hu_x(x,y+h) &&+\tfrac{h^2}{2}u_{xx}(x,y+h) &&+\tfrac{h^3}{6}u_{xxx}(x,y+h) &&+\tfrac{h^4}{24}u_{xxxx}(\xi_1)\\
&= u(x,y) &&+hu_y(x,y) &&+\tfrac{h^2}{2}u_{yy}(x,y) &&+\tfrac{h^3}{6}u_{yyy}(x,y) &&+\tfrac{h^4}{24}u_{yyyy}(\xi_2)\\
& &&+hu_x(x,y) &&+h^2 u_{xy}(x,y) &&+\tfrac{h^3}{2}u_{xyy}(x,y) &&+\tfrac{h^4}{6}u_{xyyy}(\xi_3)\\
& && &&+\tfrac{h^2}{2}u_{xx}(x,y) &&+\tfrac{h^3}{2}u_{xxy}(x,y) &&+\tfrac{h^4}{4}u_{xxyy}(\xi_4)\\
& && && &&+\tfrac{h^3}{6}u_{xxx}(x,y) &&+\tfrac{h^4}{6}u_{xxxy}(\xi_5)
\end{aligned}$$

Analog für $u(x-h,y-h)$: Bei allen ungeraden h-Potenzen entsteht ein Minuszeichen.
Addition liefert:

$$\begin{aligned}
&u(x+h,y+h) + u(x-h,y-h)\\
&\quad = 2u(x,y) + h^2\left[u_{yy}(x,y) + 2u_{xy}(x,y) + u_{xx}(x,y)\right] + O(h^4)\\
&\quad = 2u(x,y) + h^2\left[\frac{u(x,y-h)-2u(x,y)+u(x,y+h)}{h^2}\right.\\
&\qquad \left. +2u_{xy}(x,y) + \frac{u(x-h,y)-2u(x,y)+u(x+h,y)}{h^2}\right] + O(h^4)
\end{aligned}$$

$$\begin{aligned}\Leftrightarrow u_{xy}(x,y) =& \frac{1}{2h^2}[-u(x,y+h)+u(x+h,y+h)-u(x-h,y)+2u(x,y)-u(x+h,y)\\ &+u(x-h,y-h)-u(x,y-h)]+O(h^2) = \frac{1}{2h^2}\begin{bmatrix} & -1 & 1\\ -1 & 2 & -1\\ 1 & -1 & \end{bmatrix}\end{aligned} \tag{8.10}$$

Mit $u(x+h,y-h)$ und $u(x-h,y+h)$ folgt analog:

$$\begin{aligned}u_{xy}(x,y) =& \frac{1}{2h^2}[-u(x-h,y+h)+u(x,y+h)+u(x-h,y)-2u(x,y)+u(x+h,y)\\ &+u(x,y-h)-u(x+h,y-h)]+O(h^2) = \frac{1}{2h^2}\begin{bmatrix} -1 & 1 & \\ 1 & -2 & 1\\ & 1 & -1\end{bmatrix}\end{aligned} \tag{8.11}$$

Welche wählt man? → hängt von den Koeffizienten a_{xx}, a_{xy}, a_{yy} ab!
(8.9) sei als *elliptisch* vorausgesetzt, d.h.

$$\boxed{a_{xx}(x,y)\cdot a_{yy}(x,y) > a_{xy}^2(x,y)}$$

→ a_{xx}, a_{yy} haben *gleiches* Vorzeichen, o.B.d.A. sei dies *positiv*, $a_{xx} > 0$, $a_{yy} > 0$. *Zusätzlich* sei

$$|a_{xy}| \le \min\left\{a_{xx}\cdot a_{yy}\right\}$$

(aus Elliptizität folgt nur $|a_{xy}| \le \sqrt{a_{xx}\cdot a_{yy}}$)
Dann wähle Formel wie folgt:

(a) (8.10) „rechts oben"-Formel, falls $a_{xy} \ge 0$

(b) (8.11) „links oben"-Formel, falls $a_{xy} < 0$

Setze $a_{xy}^+ = \max\{a_{xy}, 0\}$, $a_{xy}^- = \min\{a_{xy}, 0\}$. Dann ergibt die Fallunterscheidung:

$$a_{xx}\frac{\partial^2 u}{\partial x^2} + 2a_{xy}\frac{\partial^2 u}{\partial x\partial y} + a_{yy}\frac{\partial^2 u}{\partial y^2} = \frac{1}{h^2}\begin{bmatrix} -a_{xy}^- & a_{yy}-|a_{xy}| & a_{xy}^+\\ a_{xx}-|a_{xy}| & 2(|a_{xy}|-a_{xx}-a_{yy}) & a_{xx}-|a_{xy}|\\ a_{xy}^+ & a_{yy}-|a_{xy}| & -a_{xy}^-\end{bmatrix} u(x,y) + O(h^2)$$

Vorzeichen

- Nicht-Diagonalelemente
  - $a_{xy}^+ \le 0$, $-a_{xy}^- \ge 0$
  - $|a_{xy}| \le \min\{\underbrace{a_{xx}, a_{yy}}_{>0}\} \Rightarrow \min\{a_{xx}, a_{yy}\} - |a_{xy}| \ge 0$

  ⇒ alle Nicht-Diagonalelemente nicht negativ.
- Diagonalelement

$$2(|a_{xy}| - a_{xx} - a_{yy}) = \underbrace{2(|a_{xy}| - \min\{a_{xx}, a_{yy}\}}_{\le 0} \underbrace{- \max\{\overbrace{a_{xx}, a_{yy}}^{>0}\}}_{<0} < 0$$

- Betragssumme der Nicht-Diagonalelemente

$$\begin{aligned} &2\left|a_{xx} - |a_{xy}|\right| + 2\left|a_{yy} - |a_{xy}|\right| + 2|a_{xy}| \\ =&2\left|a_{xx} - |a_{xy}|\right| + 2|a_{yy}| \end{aligned}$$

  Für das Diagonalelement gilt:

$$\begin{aligned} &2\left||a_{xy}| - a_{xx} - a_{yy}\right| = 2\Big|\underbrace{a_{yy}}_{>0} + \underbrace{a_{xx} - |a_{xy}|}_{\geq 0}\Big| \\ =&2|a_{yy}| + 2\left|a_{xx} - |a_{xy}|\right| \end{aligned}$$

  $\Rightarrow$ Betragssumme Nebendiagonalelemente = Betrag Hauptdiagonale!

Somit erfüllt die Diskretisierungsmatrix L_h die Bedingungen:

(i) $L_{ii} < 0$

(ii) $L_{ij} \geq 0$ für $j \neq i$

(iii) $\sum\limits_{j\neq i} |a_{ij}| = |a_{ii}| \quad \forall\, i$

Dann gilt diskretes Maximumsprinzip:
Zeile i:

$$a_{ii}x_i + \sum_{j\neq i} a_{ij}x_j = 0$$

$$\Leftrightarrow x_i = \sum_{j\neq i} \underbrace{-\frac{a_{ij}}{a_{ii}}}_{=:\alpha_{ih}\geq 0} x_j = \sum_{j\neq i} \alpha_{ij}x_j \qquad \alpha_{ij} \geq 0$$

$$\sum_{j\neq i} \alpha_{ij} = \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} = \frac{1}{|a_{ii}|} \underbrace{\sum_{j\neq i} |a_{ij}|}_{} = \frac{|a_{ii}|}{|a_{ii}|} = 1$$

und damit

$$x_i = \sum_{j\neq i} \alpha_{ij}x_j \leq \sum_{j\neq i} \alpha_{ij} \max_{k\neq i} x_k = \max_{k\neq i} x_k \underbrace{\sum_{j\neq i} \alpha_{ij}}_{=1} = \max_{k\neq i} x_k$$

ebenso

$$x_i = \sum_{j\neq i} \alpha_{ij}x_j \geq \min_{k\neq i} x_k$$

$\Rightarrow$ Das Maximum/Minimum wird auf dem Rand angenommen.

Gleichungen in Erhaltungsform

$$-\nabla \cdot \{k(x,y)\nabla u\} = f \text{ in } \Omega$$
$$u = g \text{ auf } \partial\Omega$$

Hat nicht die Form (8.9) → Produktregel anwenden. Setzt Glattheit von $k(x,y)$ voraus. (Zeichnung)

Idee

Hier für $n = 1$:

$$q = -k(x)\frac{\partial u}{\partial x} \qquad\qquad \frac{\partial q}{\partial x} = f$$

$$(q^+u)x = k\left(x + {}^h/_2\right)\frac{u(x+h) - u(x)}{h} \qquad \text{(Zahlenstrahlzeichnung)}$$

$$(\partial^- q^+ u)(x) = \frac{1}{h}\left[\underbrace{k(x + {}^h/_2)\frac{u(x+h) - u(x)}{h}}_{(q^+u)(x)} - \underbrace{k(x - {}^h/_2)\frac{u(x) - u(x-h)}{h}}_{(q^+u)(x-h)}\right]$$
$$= \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right)(x) + O(h^2)$$

Direkte Diskretisierung:

$$= \frac{1}{h^2}\left[k\left(x+\frac{h}{2}\right)u(x+h) - \left(k\left(x+\frac{h}{2}\right) + k\left(x-\frac{h}{2}\right)\right)u(x) + k\left(x-\frac{h}{2}\right)u(x-h)\right]$$
$$- \frac{1}{2}\left(k\left(x+\frac{h}{2}\right) + k\left(x-\frac{h}{2}\right)\right)\frac{u(x+h) - \cancel{2u(x)} + u(x-h)}{h^2}$$
$$+ \frac{1}{2}\left(k\left(x+\frac{h}{2}\right) + k\left(x-\frac{h}{2}\right)\right)\frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$
$$= \underbrace{\frac{2k\left(x+\frac{h}{2}\right)u(x+h) + 2k\left(x-\frac{h}{2}\right)u(x-h) - \left(k\left(x+\frac{h}{2}\right) + k\left(x-\frac{h}{2}\right)\right)(u(x+h) + u(x-h))}{2h^2}}_{\frac{k\left(x+\frac{h}{2}\right)-k\left(x-\frac{h}{2}\right)}{h}\cdot\frac{(u(x+h)-u(x-h))}{2h} = (k'(x)+O(h^2))\cdot(u'(x)+O(h^2))}$$
$$+ \underbrace{\underbrace{\frac{1}{2}\left(k\left(x+\frac{h}{2}\right) + k\left(x-\frac{h}{2}\right)\right)}_{k(x)+O(h^2)}\underbrace{\frac{u(x+h) - 2u(x) + u(x-h)}{h^2}}_{\frac{\partial^2 u}{\partial x^2}+O(h^2)}}_{k(x)\frac{\partial^2 u}{\partial x^2}+O(h^2)}$$

⇒ Konsistenzordnung 2 (für k glatt)
Vorteil der diskreten Formel: ist auch für nicht glattes $k(x)$ sinnvoll: (Grafik)

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) \mathrm{d}x = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(x)\,\mathrm{d}x$$

$$\Leftrightarrow \left(k(x)\frac{\partial u}{\partial x}\right)\Big|_{x+\frac{h}{2}} - \left(k(x)\frac{\partial u}{\partial x}\right)\Big|_{x-\frac{h}{2}} = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(x)\,\mathrm{d}x$$

$$\Leftrightarrow k\left(x+\frac{h}{2}\right)\left(\frac{u(x+h)-u(x)}{h}+O(h^2)\right) - k\left(x-\frac{h}{2}\right)\left(\frac{u(x)-u(x-h)}{h}+O(h^2)\right) = f(x)h + O(h^3)$$

$$\Leftrightarrow \frac{1}{h^2}\left[k\left(x+\frac{h}{2}\right)u(x+h) - \left(k\left(x+\frac{h}{2}\right)+k\left(x-\frac{h}{2}\right)\right)u(x) + k\left(x-\frac{h}{2}\right)u(x-h)\right] = f(x)+O(h)$$

(Selbe Formel wie oben) Nennt man Methode der integrierten Finiten-Differenzen oder „Finite Volumen Verfahren“

Fragen

- Methoden höherer Ordnung?
  - $\Rightarrow$mehr als nächste Nachbarn
  - $\Rightarrow$schwierig in Randnähe
- komplex berandete Gebiete?
  - angepasste Formeln (Shortley-Weller)
  - „Finite Elemente Verfahren“ $\Rightarrow$ nächstes Semester
- Effiziente Lösung der linearen Gleichungssysteme
- Konsistenzordnung > 1 an Neumann-Rand
- Globale Konvergenz des FD-Verfahrens?
  - $\max_{x\in\Omega_h} |u(x) - u_h(x)| \leq ?$
  - was ist, wenn $u \notin C^4(\overline{\Omega})$?

9 Konvergenz des Finite Differenzen Verfahrens

9.1 M-Matrizen

Bezeichnungen:
$A \in \mathbb{R}^{n\times n}$ mit Elementen $a_{\alpha\beta}$ mit $\alpha,\beta \in I = \{1,\ldots,n\}$
I heißt „Indexmenge“
$A \geq B$, falls $a_{\alpha\beta} \geq b_{\alpha\beta}\ \ \forall\, \alpha,\beta \in I$. ($A > B, A < B, A \leq B$ entsprechend)
0 steht für Nullmatrix.

Definition 9.1 M-Matrix
A heißt M-Matrix, genau dann wenn

(i) $a_{\alpha\alpha} > 0, \forall \alpha \in I$

(ii) $a_{\alpha\beta} \leq 0 \;\; \forall\, \alpha, \beta \in I, \alpha \neq \beta$

(iii) A nicht singulär und $A^{-1} \geq 0$.

Definition 9.2 Graph einer Matrix
$G(A) = (I, E)$, $E \subseteq I \times I$
$(\alpha, \beta) \in E \Leftrightarrow a_{\alpha\beta} \neq 0$
$(\alpha, \beta) \in E \rightarrow \alpha$ heißt direkt verbunden mit β.
α heißt verbunden mit β, falls Kette $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_k = \beta$ existiert mit $(\alpha_{i-1}, \alpha_i) \in E \;\; \forall\, i = 1, \ldots, k$

Definition 9.3 Irreduzible Matrizen
Eine Matrix A heißt irreduzibel, falls jedes $\alpha \in I$ mit jedem $\beta \in I$ verbunden ist.

Folgerung: A irreduzibel $\Rightarrow$ Es gibt *keine* Permuation der Indizes, so dass

$$P^T A P = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

Beispiel: L_h aus $-\Delta_h u_h = f_h$ (mit Grafik) $\rightsquigarrow$ L_h ist irreduzibel

Satz 9.4 Gerschgorin Kreise
Sei

$$B(z, r) = \{\xi \in \mathbb{C} \mid |z - \xi| < r\} \text{ und } \overline{B(z, r)} = \{\xi \in \mathbb{C} \mid |z - \xi| \leq r\}$$

(a) Alle Eigenwerte von A liegen in

$$\bigcup_{\alpha \in I} \overline{B(a_{\alpha\alpha}, r_\alpha)} \text{ mit } r_\alpha = \sum_{\substack{\beta \neq \alpha \\ \beta \in I}} |a_{\alpha\beta}|$$

(b) Ist A irreduzibel, so liegen alle EW von A in

$$\bigcup_{\alpha \in I} B(a_{\alpha\alpha}, r_\alpha) \cup \left(\bigcap_{\alpha \in I} \partial B(a_{\alpha\alpha}, r_\alpha) \right)$$

Beweis .

(a) (λ, a) Eigenpaar: $Au = \lambda u$. o.B.d.A.: $\|u\|_\infty = 1$.
Dann $\exists$ mindestens ein $\gamma \in I$: $|u_\gamma| = 1$ und damit gilt

$$\sum_{\beta \in I} a_{\gamma\beta} u_\beta = \lambda u_\gamma \Leftrightarrow (\lambda - a_{\gamma\gamma}) u_\gamma = \sum_{\beta \neq \gamma} a_{\gamma\beta} u_\beta$$

$$|\lambda - a_{\gamma\gamma}||u_\gamma| = \left| \sum_{\beta \neq \gamma} a_{\gamma\beta} u_\beta \right| \leq \sum_{\beta \neq \gamma} |a_{\gamma\beta}| \underbrace{|u_\beta|}_{\leq 1} \leq \sum_{\beta \neq \gamma} |a_{\gamma\beta}| =: r_\gamma \tag{9.1}$$

$$\lambda \in \overline{B(a_{\gamma\gamma}, r_\gamma)} \subseteq \bigcup_{\alpha \in I} \overline{B(a_{\alpha\alpha}, r_\alpha)} \text{ für jedes } \lambda \in \sigma(A)$$

(b) A ist zusätzlich irreduzibel: (a) gilt weiter.

$$\lambda \in \bigcup_{\alpha\in I} \overline{B(a_{\alpha\alpha}, r_\alpha)} = \underbrace{\bigcup_{\alpha\in I} B(a_{\alpha\alpha}, r_\alpha)}_{\Lambda_<} \cup \underbrace{\bigcup_{\alpha\in I} \partial B(a_{\alpha\alpha}, r_\alpha)}_{\Lambda_=}$$

Zu zeigen: Ist $\lambda \notin \Lambda_<$ und damit $\lambda \in \Lambda_=$ so gilt sogar

$$\lambda \in \bigcap_{\alpha\in I} \partial B(a_{\alpha\alpha}, r_\alpha)$$

(λ, u) Eigenpaar, $\|u\|_\infty = 1$, $\gamma \in I$, $|u_\gamma| = 1$ wie oben.

Zusätzlich: $\lambda \notin \Lambda_<$, d.h. $|\lambda - a_{\gamma,\gamma}| = r_\gamma$

Also (aus 9.1):

$$|\lambda - a_{\gamma\gamma}| = \underbrace{\sum_{\beta\neq\gamma} |a_{\gamma\beta}||u_\beta| = \sum_{\beta\neq\gamma} |a_{\gamma\beta}|}_{\Rightarrow |u_\beta|=1 \text{ für alle } \beta\in I \text{ mit } a_{\gamma\beta}\neq 0} = r_\gamma$$

Wegen A irreduzibel gibt es mindestens ein $\beta \neq \gamma$ mit $a_{\gamma\beta} \neq 0$ und damit $|u_\beta| = 1$.

Da A irreduzibel können auf diese Weise *alle* Indizes $\alpha \in I$ erreicht werden und es gilt:

$$|u_\alpha| = 1, \quad |\lambda - a_{\alpha\alpha}| = r_\alpha \ \ \forall\, \alpha \in I \Leftrightarrow \lambda \in \bigcap_{\alpha\in I} \partial B(a_{\alpha\alpha}, r_\alpha)$$

□

Folgerung 9.5
L_h sei Matrix aus Diskretisierung von $-\Delta u = f$ in Ω mit 5-Punkte-Sern. Dann ist L_h regulär.

Beweis .
L_h irreduzibel → 9.4(b) anwendbar:

$$\overbrace{\frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}}^{\text{innere}} \qquad \overbrace{\frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & 0 & \end{bmatrix}}^{\text{Randknoten} \neq \text{Ecke}} \qquad \overbrace{\frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & 0 \\ & 0 & \end{bmatrix}}^{\text{Eckknoten}}$$

also

$$\lambda \in \left[B\left(\frac{4}{h^2}, \frac{4}{h^2}\right) \cup B\left(\frac{4}{h^2}, \frac{3}{h^2}\right) \cup B\left(\frac{4}{h^2}, \frac{2}{h^2}\right)\right] \cup \underbrace{\left[\partial B\left(\frac{4}{h^2}, \frac{4}{h^2}\right) \cap \partial B\left(\frac{4}{h^2}, \frac{3}{h^2}\right) \cap \partial B\left(\frac{4}{h^2}, \frac{2}{h^2}\right)\right]}_{=\emptyset}$$

$\Rightarrow \lambda \in B\left(\frac{4}{h^2}, \frac{4}{h^2}\right)$
$\Rightarrow 0 < \lambda < \frac{8}{h^2}$

□

Diagonaldominante Matrizen

Definition 9.6
$A \in \mathbb{R}^{n\times n}$ heißt diagonal dominant, falls

(i)
$$\sum_{\substack{\beta\neq\alpha\\ \beta\in I}} |a_{\alpha\beta}| < |a_{\alpha\alpha}| \qquad \forall\, \alpha \in I$$

und irreduzibel diagonaldominant, falls

(ii) A irreduzibel

(iii)
$$\sum_{\substack{\beta\neq\alpha\\ \beta\in I}} |a_{\alpha\beta}| \leq |a_{\alpha\alpha}| \qquad \forall\, \alpha \in I$$

(iv) Bedingung (i) für mindestens ein $\alpha \in I_0$ gilt.

Bemerkung: L_h aus Fünfpunktestern ist irreduzibel diagonaldominant.

Satz 9.7
Es sei A diagonaldominant oder irreduzibel diagonaldominant und $A = D - B$, wobei $D = \operatorname{diag}(A)$. Dann ist
$$\varrho(D^{-1}B) < 1$$

Beweis .

$$C \coloneqq D^{-1}B$$
$$c_{\alpha\beta} = \begin{cases} 0 & \text{falls } \alpha = \beta \\ -\frac{a_{\alpha\beta}}{a_{\alpha\alpha}} & \text{sonst} \end{cases}$$

(a) A sei diagonaldominant
$$r_\alpha = \sum_{\beta\neq\alpha} |c_{\alpha\beta}| = \sum_{\beta\neq\alpha} \frac{|a_{\alpha\beta}|}{|a_{\alpha\alpha}|} < \frac{|a_{\alpha\alpha}|}{|a_{\alpha\alpha}|} = 1$$

(b) A sei irreduzibel diagonaldominant.

analog: $r_\beta \leq 1 \;\; \forall\, \beta \in I$ und
$r_\alpha < 1$ für mindestens ein α

$$\lambda \in \sigma(A)\colon \lambda \in \left[\bigcup_{\beta\in I} B(0, r_\beta)\right] \cup \underbrace{\left[\bigcap_{\beta\in I} \partial B(0, r_\beta)\right]}_{=:\Gamma}$$

Fallunterscheidung:

(i) Alle r_β gleich: $r_\beta = r_\alpha < 1 \Rightarrow \Gamma \subset B(0,1)$

(ii) nicht alle r_β gleich $\Rightarrow \Gamma = \varnothing$

□

Lemma 9.8
Sei $A = D - B$ und A erfülle die Vorzeichenbedingung der M-Matrix, dann gilt:

$$A \text{ ist } M\text{-Matrix} \Leftrightarrow \varrho(D^{-1}B) < 1$$

Beweis .
„⇐“: Es gilt $D \geq 0$, $B \geq 0$ wegen Vorzeichenbedingung $\Rightarrow C \geq 0$.

$$(I - C)^{-1} := S = (I - D^{-1}B)^{-1} = (I - D^{-1}(D - A))^{-1} = A^{-1}D$$

Zur Neumannreihe:

$$(I - C)\sum_{\nu=0}^{n} C^\nu = \sum_{\nu=0}^{n}(C^\nu - C^{\nu+1}) = I - C^{n+1}$$

Es gilt $C^n \to 0$ wegen $\varrho(C) < 1 \Rightarrow S := \sum_{\nu=0}^{\infty} C^\nu$. Da $SD^{-1} = A^{-1}$ und da $S \geq 0 \Rightarrow A^{-1} \geq 0$.
„⇒“: Sei $D^{-1}Bu = \lambda u$ für ein $u \in \mathbb{R}^n$ mit $u \neq 0$. Setze

$$|u| = (|u_1|, \ldots, |u_n|)^T \qquad |\lambda||u| = |\lambda u| = |D^{-1}Bu| \leq D^{-1}B|u|$$

$$\begin{aligned} |u| &= A^{-1}A|u| = A^{-1}(D - B)|u| = A^{-1}D(I - D^{-1}B)|u| \\ &= A^{-1}D|u| - A^{-1}DD^{-1}B|u| \leq A^{-1}D|u| - A^{-1}D|\lambda||u| \\ &= (1 - |\lambda|)A^{-1}D|u| \end{aligned}$$

Angenommen $|\lambda| \geq 1 \Rightarrow |u| \leq 0$ ↯
Also $\lambda < 1 \Rightarrow \varrho(D^{-1}B) < 1$ □

Satz 9.9
Ist A diagonaldominant oder irreduzibel diagonaldominant und erfüllt die Vorzeichenbedingung, dann ist A eine M-Matrix.

Beweis .
A d.d./ irr. d.d. $\overset{9.7}{\Rightarrow} \varrho(D^{-1}B) < 1 \overset{9.8}{\Rightarrow} A$ ist M-Matrix.
Es gilt auch: Ist A M-Matrix, irreduzibel, dann ist $A^{-1} > 0$. □

Definition 9.10 Restriktion
Sei $U_L \left\{ f_h \;\middle|\; f_h \colon \overline{\Omega}_h \to \mathbb{R} \right\}$ der VR der Gitterfunktionen, dann ist die Restriktion

$$R_h \colon C^0(\overline{\Omega}) \to U_h$$

definiert durch

$$(R_h u)(x) := u(x) \quad \forall\, x \in \overline{\Omega}_h$$

Bemerkung:
Jedes $f_h(x) \in U_h$ entspricht einem Vektor $f_h \in \mathbb{R}^{|\Omega_h|}$ welcher durch die Wahl einer Anordnung und das Weglassen der Dirichlet Randwerte entsteht.

Satz 9.11 Konvergenz
Sei u eine klassische Lösung der Poissongleichung mit reinem Dirichlet-Rand und $L_h u_h = q_h$ das diskrete System, das aus einer FD-Diskretisierung entsteht. L_h sei invertierbar und es gelte $\|L_h^{-1}\|_\infty \leq \kappa$ wobei κ unabhängig von h ist (Stabilität). Für $\eta_h := L_h(R_h u - u_h)$ gelte $\|\eta_h\|_\infty \leq Ch^p$ mit C unabhängig von h (Konsistenz). Dann gilt:

$$|u(x) - u_h(x)| \to 0 \ \ \forall\, x \in \Omega_h \text{ wenn } h \to 0$$

und wir nennen das Verfahren konvergent.

Beweis .
Sei $e_h := R_h u - u_h = L_h^{-1}\eta_h$ der Fehler, dann gilt:

$$\|e_h\|_\infty \|L_h^{-1}\eta_h\|_\infty \leq \|L_h^{-1}\|_\infty \|\eta_h\|_\infty \leq \kappa C h^P \to 0 \text{ für } h \to 0$$

□

Bemerkung: Für 2D-Fall (Fünfpunktestern) haben wir bereits Konsitenz für $p = 2$ bewiesen. Aus Satz 9.9 folgt, dass das zugehörige L_h eine M-Matrix ist.

Satz 9.12 Stabilität für M-Matrizen
Sei $A \in \mathbb{R}^{n\times n}$ eine M-Matrix und es sei $w \in \mathbb{R}^n$ mit $Aw \geq \mathbb{1}$, dann gilt: $\|A^{-1}\|_\infty \leq \|w\|_\infty$

Beweis .

$$\forall u \in \mathbb{R}^n \colon |u| \leq \|u\|_\infty \cdot \mathbb{1} \leq \|u\|_\infty A w$$

wegen $A^{-1} \geq 0$ gilt

$$|A^{-1}u| \leq A^{-1}|u| \leq A^{-1}\|u\|_\infty A w = \|u\|_\infty w$$

Normdefinition:

$$\|A^{-1}\|_\infty = \sup_u \frac{\|A^{-1}u\|_\infty}{\|u\|_\infty} = \sup_u \frac{\||A^{-1}u|\|_\infty}{\|u\|_\infty} \leq \sup_u \frac{\|u\|_\infty \|w\|_\infty}{\|u\|_\infty} = \|w\|_\infty$$

□

Satz 9.13 Stabilität des 5-Punktsterns
Sei L_h die Matrix aus der Diskretisierung mit dem Fünfpunktestern. Dann gilt:

$$\|L_h^{-1}\|_\infty \leq \frac{1}{8}$$

Beweis Idee.
Satz 9.12 angewandt mit

$$w(x,y) = x\frac{1-x}{2} \text{ und } w_h = R_h w(x,y)$$

□

10 Iterative Lösung von linearen Gleichungssystemen

FD führt auf

$$Ax = b \qquad A \in \mathbb{R}^{N\times N} \qquad x, b \in \mathbb{R}^N$$

A ist

- dünn besetzt
- oft sym. pos. definit
- N sehr groß

$$\|R_h u - u_h\|_\infty \leq Ch^2$$

$2d$: $h = \frac{1}{\sqrt{N}} \Rightarrow \|R_h u - u(h)\|_\infty \leq \frac{C}{N}$

Aufwand direkter Lösungsverfahren (LR-Zerlegung)

- A voll besetzt: $O(N^3)$
- A Bandmatrix (mit Grafik zur Erläuterung)
  $n = N^\alpha$ Bandbreite
  $\alpha = \frac{d-1}{d}$ für strukturiertes Gitter in d Raumdimensionen
  Aufwand:

$$A(N) = \sum_{i=1}^{N} n \cdot 2n = \sum_{i=1}^{N} 2n^2 = 2Nn^2 = 2NN^{2\alpha} = 2N^{2\alpha+1}$$

$$\begin{aligned} d = 1 \quad \alpha = 0 \quad &\Rightarrow O(N) \\ d = 2 \quad \alpha = \frac{1}{2} \quad &\Rightarrow O(N^2) \\ d = 3 \quad \alpha = \frac{1}{3} \quad &\Rightarrow O(N^{\frac{7}{3}}) \end{aligned}$$

10.1 Relaxationsverfahren

Idee: Sukzessives Auflösen

$$\text{i-te Gl.: } \sum_{j=1}^{N} a_{ij}x_j = b_i \underset{a_{ii}\neq 0}{\Leftrightarrow} \boxed{x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j\right)}$$

Konstruierte Folge $x^{(k)} \to x$

Jacobi-Verfahren (Gesamtschrittverfahren)

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\right) \qquad i = 1, \cdots, N$$

gedämpftes Jacobi-Verfahren:

$$x_i^{(k+1)} = (1-\omega)\cdot x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\right) \qquad \begin{array}{l} i = 1,\cdots,N \\ \omega \in (0,1] \end{array}$$

Gauß-Seidel (Einzelschritt) Verfahren:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right) \qquad i = 1,\cdots,N$$

gedämpftes Gauß-Seidel Verfahren:

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right) \qquad i = 1,\cdots .N$$

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega\tilde{x}_i^{(k+1)} \qquad \begin{array}{l} i = 1,\cdots,N \\ \omega \in (0,1] \end{array}$$

SOR (successive over relaxation)-Verfahren

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right) \qquad \begin{array}{l} i = 1,\cdots,N \\ \omega \in (0,2) \end{array}$$

gedämpfte Richardson-Iteration

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega\left(b - \sum_{j\neq i} a_{ij}x_j^{(k)}\right) \qquad i = 1,\cdots,N$$

Matrixschreibweise der Relaxionsverfahren

$$A = \underbrace{L}_{l_{ij}=\begin{cases} a_{ij} & i>j \\ 0 \end{cases}} + \underbrace{D}_{d_{ij}=\begin{cases} a_{ij} & i=j \\ 0 \end{cases}} + \underbrace{U}_{u_{ij}=\begin{cases} a_{ij} & i<j \\ 0 \end{cases}}$$

gedämpftes Jacobiverfahren:

$$\begin{aligned} x^{(k+1)} &= (1-\omega)x^{(k)} + \omega D^{-1}(b - (L+U)x^{(k)}) \\ &= x^{(k)} - \omega D^{-1}\underbrace{(b - Ax^{(k)})}_{=:d^{(k)}\text{ Defekt}} \end{aligned}$$

SOR-Iteration

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right) \qquad i = 1, \cdots, N$$

$$\Leftrightarrow \quad \omega \sum_{j<i} a_{ij}x_j^{(k+1)} + a_{ii}x_i^{(k+1)} = a_{ii}(1-\omega)x_i^{(k)} + \omega\left(b_i - \sum_{j>i} a_{ij}x_j^{(k)}\right) \qquad i = 1, \cdots, N$$

$$\Leftrightarrow \qquad \omega L x^{(k+1)} + Dx^{(k+1)} = (1-\omega)Dx^{(k)} + \omega(b - Ux^{(k)})$$

$$\begin{aligned} x^{(k+1)} &= (1-\omega)(\omega L + D)^{-1}Dx^{(k)} + \omega(\omega L + D)^{-1}(b - Ux^{(k)}) + \\ &\quad \underbrace{\omega(\omega L + D)^{-1}(L+D)x^{(k)} - \omega(\omega L + D)^{-1}(L+D)x^{(k)}}_{=0} \\ &= (\omega L + D)^{-1}[(1-\omega)D + \omega(L+D)]x^{(k)} + \\ &\quad \underbrace{\omega(\omega L + D)^{-1}}_{(L+\frac{1}{\omega}D)^{-1}}(b - \underbrace{(L+D+U)}_{=A}x^{(k)}) \end{aligned}$$

$$x^{(k+1)} = x^{(k)} + (L + \frac{1}{\omega}D)^{-1}(b - Ax^{(k)})$$

Allgemeiner Ansatz für Iterationsverfahren

Fehler: $e^{(k)} = x - x^{(k)}$
dann gilt: $Ae^{(k)} = A(x - x^{(k)}) = Ax - Ax^{(k)} = b - Ax^{(k)} = d^{(k)}$
Idee:
(1) Löse $Ae^{(k)} = d^{(k)}$
(2) Berechne $x = x^{(k)} + e^{(k)} = x^{(k)} + A^{-1}(b - Ax^{(k)})$
aber löse Defektgleichung inexakt:
(1') Löse $Mv^{(k)} = d^{(k)}$ mit $M \approx A$
(2) Setze $x^{(k+1)} = x^{(k)} + v^{(k)} = x^{(k)} + M^{-1}(b - Ax^{(k)})$

$M = D$	Jacobi
$M = L + D$	Gauß-Seidel
$M = L + \frac{1}{\omega}D$	SOR-Verfahren
$M = \frac{1}{\omega}(L + D)$	ged. Gauß-Seidel
$M = \frac{1}{\omega}I$	ged. Richardson-Iteration

Konvergenz linearer Iterationsverfahren

Fehlerfortpflanzung:

$$\begin{aligned} e^{(k+1)} &= x - x^{(k+1)} = \underbrace{x - x^{(k)}} - M^{-1}(b - Ax^{(k)}) \\ &= e^{(k)} - M^{-1}Ae^{(k)} \\ &= \underbrace{(I - M^{-1}A)}_{=:S\ \text{„Iterationsmatrix“}} e^{(k)} \end{aligned}$$

$e^{(k)} = S^k e^{(0)}$

Satz 10.1
Ein Verfahren der Form $x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)})$ konvergiert unabhängig vom Startwert $x^{(0)}$ genau dann wenn $\varrho(S) < 1$.

Beweis .
„⇐“ für eine zugeordnete Matrixnorm gilt:

$$e^{(k)} = S^k e^{(0)} \Rightarrow \|e^{(k)}\| \leq \|S\|^k \|e^{(0)}\|$$

Lemma 5.10: $\varrho(S) < 1$ dann $\exists$ Norm $\|\ \|_\varepsilon$ mit $\|S\|_\varepsilon \leq \varrho(S) + \varepsilon < 1$ und damit

$$\lim_{k \to 0} \|S\|_\varepsilon^k \to 0 \text{ also } \|e^{(k)}\|_\varepsilon \to 0$$

„⇒“ Sei $w \neq 0$ ein EV von S zum betragsgrößten EW λ, d.h. $|\lambda| = \varrho(S)$.
Wähle Startwert $x^{(0)} = x - w \Rightarrow e^{(0)} = x - x^{(0)} = w$
Also $e^{(k)} = S^k w = \lambda^k w$ wegen Konvergenz folgt $|\lambda| < 1$ □

Satz 10.2

Sei $A = A^T > 0$ (sym. pos. definit $\Leftrightarrow x^T Ax > 0 \forall x \neq 0$) und $\omega < \frac{2}{\lambda_{\max}(A)}$. Dann konvergiert die Richardson-Iteration.

Beweis .

$$\lambda_i \in \sigma(A) \text{ mit } 0 < \lambda_{\min}(A) = \lambda_1 \leq \lambda_2 \leq \cdots \lambda_N = \lambda_{\max}(A)$$

$$S_{\text{Rich}} = I - \omega A \Rightarrow \sigma(S_{\text{Rich}}) = \{\mu_i \mid \mu_i = 1 - \omega\lambda_i;\ \lambda_i \in \sigma(A)\}$$

(Grafik zur Erläuterung, wie man das praktisch „grafisch löst“)

$$1 - \omega\lambda_{\max}(A) > -1 \qquad \omega < \frac{2}{\lambda_{\max}(A)}$$

Optimal Konvergenzrate:

$$\frac{2}{\lambda_{\min} + \lambda_{\max}}$$

□

Bemerkung 10.3
Es sei $\varrho = \varrho(S)$. Dann benötigt man $\frac{\log \varepsilon}{\log \varrho}$ Schritte um den Fehler um den Faktor ε zu reduzieren.
$w \in \mathbb{R}^N$ zu EW λ, $|\lambda| = \varrho(S)$

$$e^{(k)} = S^k w = |\lambda|^k w$$
$$\text{also } \|e^{(k)}\| = |\lambda|^k \|w\|$$
$$|\lambda|^k = \varepsilon \Rightarrow k = \frac{\log \varepsilon}{\log |\lambda|}$$

Richardson: ($\omega = \frac{1}{\lambda_{\max}}$):

$$\varrho(S_{\text{Rich}}) = 1 - \frac{\lambda_{\min}}{\lambda_{\max}} = 1 - \frac{1}{\kappa(A)} \qquad \kappa(A) = \frac{\lambda_{\min}}{\lambda_{\max}}$$

$$k(\varepsilon) = \frac{\log \varepsilon}{\log\left(1 - \frac{1}{\kappa}\right)} \approx \frac{\log \varepsilon}{-\frac{1}{\kappa}} = \kappa(A)|\log \varepsilon|$$

Bemerkung 10.4 Aufwand zur Lösung der Poisson-Gleichung
L_h 5-Punkte-Stern: Es gilt $\kappa(L_h) = O(h^{-2})$
d Raumdimensionen, strukturiertes, äquidistantes Gitter: $h = N^{-\frac{1}{d}}$. Aufwand für eine Iteration: $O(N)$.
$\Rightarrow$ Gesamtaufwand (zur Reduktion des Fehlers um Faktor ε): $O\left(|\log \varepsilon| N^{1+\frac{2}{d}}\right)$

	Relaxation	Bandmatrix	Nested Diss
$d = 1$	$O(N^3)$	$O(N)$	$O(N)$
$d = 2$	$O(N^2)$	$O(N^2)$	$O(N^{\frac{3}{2}})$
$d = 3$	$O(N^{\frac{5}{3}})$	$O(N^{\frac{7}{3}})$	$O(N^2)$

Satz 10.5 Jacobi-Verfahren im s.p.d. Fall
$A = A^T > 0$ und $2D - A > 0$ (s.p.d.) Dann konvergiert die Jacobi-Iteration.

Beweis .

- $\langle x, y\rangle = \sum_{i=1}^{N} x_i y_i$
- $C = C^T$ dann gilt (Raleigh-Quotient):

$$\sup_{x\neq 0} \frac{\langle x, Cx\rangle}{\langle x, x\rangle} = \lambda_{\max}(C) \qquad \inf_{x\neq 0} \frac{\langle x, Cx\rangle}{\langle x, x\rangle} = \lambda_{\min}(C)$$

$S_{\text{Jac}} = I - D^{-1}A$ ist nicht symmetrisch, aber mit „$D^{\frac{1}{2}}$“ = diag($\sqrt{a_{ii}}$) gilt:

$$\sigma(S_{\text{Jac}}) = \sigma\left(D^{\frac{1}{2}} S_{\text{Jac}} D^{-\frac{1}{2}}\right) = \sigma\left(\underbrace{I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}}_{\text{symmetrisch}}\right)$$

- Voraussetzung $2D - A$ positiv definit bedeutet:

$$\begin{aligned}\langle x, (2D - A)x\rangle > 0 &\Leftrightarrow 2\langle x, Dx\rangle - \langle x, Ax\rangle > 0\\ &\Leftrightarrow \frac{\langle x, Ax\rangle}{\langle x, Dx\rangle} < 2\end{aligned}$$

Damit:

$$\begin{aligned}\lambda_{\max}(I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) &= \sup_{x\neq 0} \frac{\langle x, (I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}})x\rangle}{\langle x, x\rangle}\\ &= 1 - \inf_{x\neq 0} \frac{\langle x, D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x\rangle}{\langle x, x\rangle}\\ &= 1 - \inf_{x = D^{\frac{1}{2}} y \neq 0} \underbrace{\frac{\langle y, Ay\rangle}{\langle y, Dy\rangle}}_{>0 \text{ wg. } A, D \text{ pos. def.}} < 1\end{aligned}$$

Analog für $\lambda_{\min}$

$$\begin{aligned}\lambda_{\min}(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}) &= \inf_{x\neq 0} \frac{\langle x, (I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}})x\rangle}{\langle x, x\rangle} \\ &= 1 - \sup_{x\neq 0} \frac{\langle x, D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x\rangle}{\langle x, x\rangle} \\ &= 1 - \sup_{y\neq 0} \underbrace{\frac{\langle y, Ay\rangle}{\langle y, Dy\rangle}}_{<2 \text{ nach Vor.}} > -1\end{aligned}$$

$\Rightarrow \varrho(S_{\text{Jac}}) < 1$ □

Resultat für diagonaldominante Matrizen.

Satz 10.6
Sei A diagonaldominant oder irreduzibel diagonal dominant. Dann konvergiert das Gauß-Seidel Verfahren.

Beweis .

(i)

$$\begin{aligned}S = I - (L + D)^{-1}A &\Leftrightarrow (L + D)S = L + D - A = -U \\ &\Leftrightarrow DS = -(U + LS) \\ &\Leftrightarrow S = -D^{-1}(U + LS)\end{aligned}$$

Komponentenweise:

$$(Sx)_i = \left(-D^{-1}(U + LS)x\right)_i = -\frac{1}{a_{ii}}\left(\sum_{j>i} a_{ij}x_j + \sum_{j<i} a_{ij}(Sx)_j\right)$$

Betrag bilden liefert

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|}\left(\sum_{j>i} |a_{ij}||x_j| + \sum_{j<i} |a_{ij}||(Sx)_j|\right)$$

(ii) diagonalominanter Fall, d.h. $\sum_{j\neq i} |a_{ij}| < |a_{ii}| \ \ \forall\, i$. Per Induktion zeige: $|(Sx)_i| < \|x\|_\infty$:

$i = 1$:

$$\begin{aligned}|(Sx)_i| &\leq \frac{1}{|a_{ii}|}\left(\sum_{j>i} |a_{ij}||x_j|\right) \\ &\leq \|x\|_\infty \underbrace{\frac{1}{|a_{ii}|}\sum_{j>i} |a_{ij}|}_{<1} < \|x\|_\infty\end{aligned}$$

$i-1 \to i$, d.h. $|(Sx)_j| < \|x\|_\infty \;\; \forall\, j < i$

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|}\left(\sum_{j>i}|a_{ij}|\underbrace{|x_j|}_{\leq\|x\|_\infty} + \sum_{j<i}|a_{ij}|\underbrace{|(Sx)_j|}_{<\|x\|_\infty}\right) \leq \|x\|_\infty \underbrace{\frac{1}{|a_{ii}|}\sum_{j\neq i}|a_{ij}|}_{<1} < \|x\|_\infty$$

Damit ist $\|S\|_\infty = \sup\limits_{x\neq 0}\frac{\|Sx\|_\infty}{\|x\|_\infty} < \frac{\|x\|_\infty}{\|x\|_\infty} < 1$

(iii) irreduzibel diagonaldom. Fall: jetzt gilt $\sum\limits_{j\neq i}|a_{ij}| \leq |a_{ii}|$, für *mind. ein i* gilt $<$ statt $\leq$. Obiger Beweis liefert sofort $\|S\|_\infty \leq 1$. Führe $\|S\|_\infty = 1$ zum Widerspruch.

Sei x ein Vektor mit $\|Sx\|_\infty = 1$ und $\|x\|_\infty = 1$. Dann gibt es mindestens einen Index i sodass

$$1 = |(Sx)_i| \leq \frac{1}{|a_{ii}|}\left(\sum_{j>i}|a_{ij}||x_j| + \sum_{j<i}|a_{ij}||(Sx)_j|\right) \leq \|x\|_\infty = 1$$
$$\Rightarrow \sum_{j>i}|a_{ij}||x_j| + \sum_{j<i}|a_{ij}||(Sx)_j| = |a_{ii}|$$

Dies impliziert:

$$\sum_{j\neq i}|a_{ij}| = |a_{ii}|$$
$$|x_j| = 1,\ j > i \text{ und } |(Sx)_j| = 1,\ j < i \text{ für } a_{ij} \neq 0$$

$\Rightarrow$ Es ergeben sich weitere Indizes $j \neq i$ mit $|(Sx)_j| = 1$. Wende das Argument rekursiv an.

$\Rightarrow$ Da A irreduzibel werden, so alle $j = 1, \ldots, N$ erreicht und es gilt für alle Zeilen i:

$$\sum_{j\neq i}|a_{ij}| = |a_{ii}|$$

$\Rightarrow$ Widerspruch zu A-irreduzibel diagonaldominant.

□

10.2 Abstiegsverfahren

Satz 10.7
Es sei $A = A^T > 0$. Dann nimmt das Funktional

$$F := \frac{1}{2}x^TAx - b^Tx = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle)$$

sein eindeutiges Minimum in $x^\star = A^{-1}b$ an. D.h. die Lösung des Minimierungsproblems $\min\limits_{x\in\mathbb{R}^N} F(x)$ stimmt mit der Lösung des Gleichungssystems $Ax = b$ überein.

Beweis .
$x \in \mathbb{R}^N$. Wähle v so, dass $x = x^\star + v$

$$\begin{aligned} F(x) &= \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle = \frac{1}{2}\langle A(x^\star + v), x^\star + v\rangle - \langle b, x^\star + v\rangle \\ &= \frac{1}{2}\left[\langle Ax^\star, x^\star\rangle + 2\langle Ax^\star, v\rangle + \langle Av, v\rangle\right] - \langle b, x^\star\rangle - \langle b, v\rangle \\ &= \underbrace{\frac{1}{2}\langle Ax^\star, x^\star\rangle - \langle b, x^\star\rangle}_{=F(x^\star)} + \underbrace{\langle Ax^\star - b, v\rangle}_{=0 \text{ für } x^\star = A^{-1}b} + \frac{1}{2}\underbrace{\langle Av, v\rangle}_{>0 \text{ für } v\neq 0} \end{aligned}$$

$\Rightarrow F(x) > F(x^\star)$ für alle $x \neq x^\star$
Eindeutigkeit: Ang. $\exists$ zwei Minima, $x' \neq x$
$\rightarrow x' = x^\star + v$ d.h. $v \neq 0$
$\Rightarrow F(x') = F(x^\star) + \frac{1}{2}\underbrace{\langle Av, v\rangle}_{>0}$ Widerspruch zu x' Minimum □

Eindimensionale Minimierung:

Idee: $x^{(k)}$, Suchrichtung $p^{(k)} \neq 0$ finde $\alpha^{(k)} \in \mathbb{R}$ so dass

$$F(x^{(k)} + \alpha^{(k)} p^{(k)}) \rightarrow \min$$

$$f(\alpha^{(k)}) = F(x^{(k)} + \alpha^{(k)} p^{(k)}) = F(x^{(k)}) + \alpha^{(k)}\langle Ax^{(k)} - b, p^{(k)}\rangle + \frac{1}{2}(\alpha^{(k)})^2\langle Ap^{(k)}, p^{(k)}\rangle$$

$$\frac{\mathrm{d}f}{\mathrm{d}\alpha^{(k)}} = \langle Ax^{(k)} - b, p^{(k)}\rangle + \alpha^{(k)}\langle Ap^{(k)}, p^{(k)}\rangle \stackrel{!}{=} 0 \Rightarrow \alpha^{(k)} = \frac{\overbrace{\langle b - Ax^{(k)}}^{\text{Defekt}}, p^{(k)}\rangle}{\langle Ap^{(k)}, p^{(k)}\rangle}$$

Gradientenverfahren: steilster Abstieg

$N = 2$: $F\colon \mathbb{R}^2 \rightarrow \mathbb{R}$ (Zeichnung)
Idee: $p^{(k)} = -\nabla F(x^{(k)})$
$\nabla F(x^{(k)}) = Ax^{(k)} - b$
d.h. $p^{(k)} = -\nabla F(x^{(k)}) = b - Ax^{(k)} = d^{(k)}$

Gradientenverfahren

Require: Δ, $x = x^{(0)}$, b, ε
  $d = b - Ax$
  norm0 = $\|d\|$
  norm = norm0
  while (norm> ε norm0) **do**
    $q = Ad$
    $\alpha = \frac{\langle d,d\rangle}{\langle q,d\rangle}$
    $x = x + \alpha d$
    $d = d - \alpha q$
    norm = $\|d\|$

end while

Satz 10.8 Konvergenz des Gradientenverfahrens
Sei $A = A^T > 0$ und $Ax = b$. Definiere das Fehlerfunktional

$$E(y) := \|x - y\|_A^2 = \langle A(x-y), x-y\rangle \qquad \forall\, y \in \mathbb{R}^N$$

($\langle x, y\rangle_A = \langle Ax, y\rangle$ heißt Energieskalarprodukt und $\|x\|_A = \sqrt{\langle Ax, x\rangle}$ Energienorm).
Dann gilt:

$$E(x^{(t+1)}) \leq \left(1 - \frac{1}{\kappa(A)}\right) E(x^{(t)})$$

$$\text{oder: } \|x - x^{(t+1)}\|_A \leq \left(1 - \frac{1}{\kappa(A)}\right)^{\frac{1}{2}} \|x - x^{(t)}\|_A$$

Beweis .

$$\begin{aligned}
\frac{E(x^{(t)}) - E(x^{(t+1)})}{E(x^{(t)})} &= \frac{\langle e^{(t)}, Ae^{(t)}\rangle - \langle e^{(t+1)}, Ae^{(t+1)}\rangle}{\langle e^{(t)}, Ae^{(t)}\rangle} \\
&= \frac{\langle e^{(t)}, Ae^{(t)}\rangle - \langle e^{(t)} - \alpha^{(t)} d^{(t)}, A(e^{(t)} - \alpha^{(t)} d^{(t)})\rangle}{\langle e^{(t)}, Ae^{(t)}\rangle} \\
&= \frac{2\alpha^{(t)}\langle e^{(t)}, Ad^{(t)}\rangle - (\alpha^{(t)})^2 \langle d^{(t)}, Ad^{(t)}\rangle}{\langle e^{(t)}, Ae^{(t)}\rangle} \\
&= \frac{2\frac{\langle d^{(t)}, d^{(t)}\rangle}{\langle Ad^{(t)}, d^{(t)}\rangle}\langle d^{(t)}, d^{(t)}\rangle - \left(\frac{\langle d^{(t)}, d^{(t)}\rangle}{\langle Ad^{(t)}, d^{(t)}\rangle}\right)^2 \langle d^{(t)}, Ad^{(t)}\rangle}{\langle d^{(t)}, A^{-1} d^{(t)}\rangle} \\
&= \frac{\langle d^{(t)}, d^{(t)}\rangle^2}{\langle Ad^{(t)}, d^{(t)}\rangle \langle A^{-1} d^{(t)}, d^{(t)}\rangle} \\
&= \frac{\|d^{(t)}\|^4}{\langle Ad^{(t)}, d^{(t)}\rangle \langle A^{-1} d^{(t)}, d^{(t)}\rangle}
\end{aligned}$$

Raleigh Quotient:

$$\begin{aligned}
\lambda_{\min}\|y\|^2 \leq \langle Ay, y\rangle \leq \lambda_{\max}\|y\|^2 &\Rightarrow \frac{\|y\|^2}{\langle Ay, y\rangle} \geq \frac{1}{\lambda_{\max}} \\
\frac{1}{\lambda_{\max}}\|y\|^2 \leq \langle A^{-1}y, y\rangle \leq \frac{1}{\lambda_{\min}}\|y\|^2 &\leq \frac{\|y\|^2}{\langle A^{-1}y, y\rangle} \geq \lambda_{\min}
\end{aligned}$$

und damit

$$\begin{aligned}
\frac{E(x^{(t)}) - E(x^{(t+1)})}{E(x^{(t)})} &= \cdots \geq \frac{\lambda_{min}}{\lambda_{\max}} = \frac{1}{\kappa(A)} \\
\Leftrightarrow \qquad E(x^{(t+1)}) &\leq \left(1 - \frac{1}{\kappa(A)}\right) E(x^{(t)})
\end{aligned}$$

□

Konjugierte Gradienten Verfahren

Gradientenverfahren $N = 2$: (Zeichnung)
Idee: Minimierung im Unterraum

$$B_t := \text{span}\left\{p^{(0)}, p^{(1)}, \cdots, p^{(t-1)}\right\} \qquad p^{(i)} \text{linear unabhängig}$$

Gegeben: $x^{(0)}$ bestimme

$$x^{(t)} = x^{(0)} + \sum_{j=0}^{t-1} \underbrace{\alpha_j^{(t-1)}}_{\sim\sim} p^{(j)} \in x^{(0)} + B_t$$

sodass

$$F(x^{(t)}) = \min_{y \in x^{(0)} + B_t} F(y)$$

Also:

$$\begin{aligned}
F(x^{(t)}) &= \frac{1}{2}\langle Ax^{(t)}, x^{(t)}\rangle - \langle b, x^{(t)}\rangle \\
&= \frac{1}{2}\left\langle A\left(x^{(0)} + \sum_{j=0}^{t-1}\alpha_j^{(t-1)}p^{(j)}\right), x^{(0)} + \sum_{k=0}^{t-1}\alpha_k^{(t+1)}p^{(k)}\right\rangle - \left\langle b, x^{(0)} + \sum_{k=0}^{t-1}\alpha_k^{(t-1)}p^{(k)}\right\rangle \\
&= \frac{1}{2}\langle Ax^{(0)} - b, x^{(0)}\rangle + \sum_{k=0}^{t-1}\alpha_k^{(t-1)}\langle Ax^{(0)} - b, p^{(k)}\rangle + \sum_{j=0}^{t-1}\frac{(\alpha_j^{(t-1)})^2}{2}\langle Ap^{(j)}, p^{(j)}\rangle \\
&\qquad + \sum_{j=0}^{t-1}\sum_{k\neq j}\frac{\alpha_j^{(t-1)}\alpha_k^{(t-1)}}{2}\langle Ap^{(j)}, p^{(k)}\rangle
\end{aligned}$$

$$\begin{aligned}
\frac{\partial F}{\partial \alpha_i^{(t-1)}}(x^{(t)}) &= \langle Ax^{(0)} - b, p^{(i)}\rangle + \alpha_i^{(t-1)}\langle Ap^{(i)}, p^{(i)}\rangle + \sum_{k\neq i}\frac{\alpha_k^{(t-1)}}{2}\langle Ap^{(i)}, p^{(k)}\rangle + \sum_{j\neq i}\frac{\alpha_j^{(t-1)}}{2}\langle Ap^{(j)}, p^{(i)}\rangle \\
&= \langle Ax^{(0)} - b, p^{(i)}\rangle + \sum_{k=0}^{t-1}\alpha_k^{(t-1)}\underbrace{\langle Ap^{(k)}, p^{(i)}\rangle}_{c_{ik}} \stackrel{!}{=} 0 \qquad i = 0, \cdots, t-1 \qquad (10.1)
\end{aligned}$$

$\Rightarrow$ LGS für die Koeffizienten α_k:

$$C\alpha^{(t+1)} = r \qquad c_{ik} = \langle Ap^{(k)}, p^{(i)}\rangle \qquad r_i = \langle b - Ax^{(0)}, p^{(i)}\rangle$$

C ist s.p.d.

$$\langle Cy, y\rangle = \sum_{i=0}^{t-1}\sum_{k=0}^{t-1} y_i y_k c_{ik} = \sum_i\sum_k y_i y_k \langle Ap^{(k)}, p^{(i)}\rangle = \left\langle A\overbrace{\left(\sum_{k=0}^{t-1} y_k p^{(k)}\right)}^{=:y'}, \sum_{j=0}^{t-1} y_i p^{(i)}\right\rangle = \langle Ay', y'\rangle > 0$$

(10.1) können wir schreiben als:

$$\langle Ax^{(t)} - b, p^{(i)}\rangle = 0 \qquad \forall\, i = 0, \cdots, t-1 \qquad (10.2)$$

nent man „Galerkin-Gleichungen“
Aufwand steigt mit der Anzahl der Suchrichtungen stark an.
Idee: Wähle A-orthogonale Suchrichtungen: („konjugierte“ Suchrichtungen)

$$\left\langle Ap^{(j)}, p^{(i)} \right\rangle = 0 \qquad \forall\, 0 \le i < j \le t-1$$

Dann gilt:

$$\begin{aligned} &\langle Ax^{(0)} - b, p^{(i)} \rangle + \sum_{k=0}^{t-1} \alpha_k^{(t-1)} \langle Ap^{(k)}, p^{(i)} \rangle \\ &= \langle Ax^{(0)} - b, p^{(i)} \rangle + \alpha_i^{(t-1)} \langle Ap^{(i)}, p^{(i)} \rangle = 0 \qquad\qquad i = 0, \cdots, t-1 \end{aligned}$$

also

$$\alpha_i := \alpha_i^{(t-1)} = \frac{\langle b - Ax^{(0)}, p^{(i)} \rangle}{\langle Ap^{(i)}, p^{(i)} \rangle} \quad i = 0, \cdots, t-1 \tag{10.3}$$

Orthogonalisierung der Suchrichtungen mittels Gram-Schmidt:
Suchrichtungen $p^{(k)}$ werden sukzessive aufgebaut:

$$\begin{aligned} p^{(0)} &= -\nabla F(x^{(0)}) = b - Ax^{(0)} =: d^{(0)} \\ p^{(t)} &= -\nabla F(x^{(t)}) + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)} \text{ sodass } \langle Ap^{(t)}, p^{(i)} \rangle = 0 \ \ \forall\, i < t \end{aligned} \tag{10.4}$$

Zwischenresultat:

$$\operatorname{span}\left\{p^{(0)}, \cdots, p^{(t)}\right\} = \operatorname{span}\left\{d^{(0)}, Ad^{(0)}, \cdots, A^t d^{(0)}\right\} =: K_t(d^{(0)}, A) \qquad \text{„Krylovraum“}$$

Induktion über t:
$t = 0$:

$$p^{(0)} = d^{(0)} \qquad \operatorname{span}\left\{p^{(0)}\right\} = \operatorname{span}\left\{d^{(0)}\right\} \qquad \checkmark$$

$t - 1 \to t$, d.h.

$$\operatorname{span}\{p^{(0)}, \cdots, p^{(t-1)}\} = \operatorname{span}\{d^{(0)}, \cdots A^{t-1} d^{(0)}\} = K_{t-1}(d^{(0)}, A)$$

$$p^{(t)} = b - A \underbrace{\left(x^{(0)} + \sum_{i=0}^{t-1} \alpha_i p^{(i)} \right)}_{x^{(t)}} + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)}$$

$$= \underbrace{b - Ax^{(0)}}_{d^{(0)}} - \underbrace{\sum_{i=0}^{t-2} \alpha_i Ap^{(i)} + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)}}_{\in K_{t-1}(d^{(0)}, A)} - \underbrace{\alpha_{t-1} Ap^{(t-1)}}_{\in K_t(d^{(0)}, A)}$$

Bestimmung der $\beta_j^{(t-1)}$, $i < t$:

$$\begin{aligned} 0 \stackrel{!}{=} \langle Ap^{(t)}, p^{(i)} \rangle &= \langle b - Ax^{(t)} + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)}, Ap^{(i)} \rangle \\ &= \langle b - Ax^{(t)}, Ap^{(i)} \rangle + \sum_{j=0}^{t-1} \beta_j^{(t-1)} \underbrace{\langle p^{(j)}, Ap^{(i)} \rangle}_{=0 \text{ für } i \neq j} \qquad \forall\, i \leq t-1 \\ &= \langle b - Ax^{(t)}, Ap^{(i)} \rangle + \beta_i^{(t+1)} \langle p^{(i)}, Ap^{(i)} \rangle \end{aligned}$$

Fall 1: $i \leq t-2$

$$\begin{aligned} Ap^{(i)} &\leq \text{span}\{d^{(0)}, \cdots, AA^i d^{(0)}\} = \text{span}\{p^{(0)}, \cdots, p^{(i+1)}\} \\ \text{Galerkin: } \langle b - Ax^{(t)}, p^{(k)} \rangle &= 0 \text{ für } k \leq t-1 \Rightarrow \beta^{(t-1)} = 0 \end{aligned}$$

Fall 2: $i = t-1$ liefert

$$\beta_{t-1}^{(t-1)} = \frac{\langle b - Ax^{(t)}, Ap^{(t-1)} \rangle}{\langle p^{(t-1)}, Ap^{(t-1)} \rangle}$$

also:

$$p^{(t)} = b - Ax^{(t)} + \beta_{t-1}^{(t-1)} p^{(t-1)} \tag{10.5}$$

Konvergenz:

$$\|x^{(t)} - x\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right) \|x^{(0)} - x\|_A \qquad \varrho \approx 1 - \frac{1}{\sqrt{\kappa(A)}}$$

Zur Reduktion des Fehlers um den Faktor ε benötigt man $t(\varepsilon) = O(\sqrt{\kappa(A)} |log \varepsilon|)$ Iterationen. Dies führt für die Diskretisierung der Poisongleichung mit dem Fünfpunktstern auf die Komplexitäten:

	CG	Relaxation	Bandmatrix	Nested Diss
$d = 2$	$O(N^{\frac{3}{2}})$	$O(N^2)$	$O(N^2)$	$O(N^{\frac{3}{2}})$
$d = 3$	$O(N^{\frac{4}{3}})$	$O(N^{\frac{5}{3}})$	$O(N^{\frac{7}{3}})$	$O(N^2)$