

Numerik 1 – Olaf Ippisch

1. Februar 2012

Inhaltsverzeichnis

1	Motivation	3
1.1	Wachstumsmodelle	3
1.2	Chemische Reaktionen	4
1.3	Neurowissenschaft	5
1.4	Astrophysikalisches N-Körper-Problem	6
1.5	Raketengleichung	8
2	Zur Theorie von gewöhnlichen Differentialgleichungen	8
2.1	Problemstellung	8
2.2	Existenzaussagen	10
2.3	Eindeutigkeit und Stabilität	13
2.4	Globale Stabilität	17
3	Einschrittverfahren	20
3.1	Das explizite Euler-Verfahren	20
3.2	Taylor und Runge-Kutta Verfahren	24
3.3	Konvergenz allgemeiner Einschrittverfahren	29
3.4	Schrittweitensteuerung	31
4	Numerik steifer Differentialgleichungen	37
4.1	Motivation	37
4.2	(Skalare, lineare) Modellprobleme	39
4.3	Lineare Stabilitätsanalyse	42
4.4	Implizite Runge-Kutta Verfahren	43
5	Mehrschrittverfahren	53
5.1	Verfahrenskonstruktion	53
5.2	Konsistenz von LMM	55
5.3	Nullstabilität von LMM	57
5.4	Konvergenz von LMM	58

6 Randwertprobleme	65
6.1 Schießverfahren	66
6.2 Differenzverfahren	68
7 Partielle Differentialgleichungen	70
8 Finite Differenzen Verfahren	76
9 Konvergenz des Finite Differenzen Verfahrens	84
9.1 M-Matrizen	84
10 Iterative Lösung von linearen Gleichungssystemen	90
10.1 Relaxationsverfahren	90
10.2 Abstiegsverfahren	96

Lizenz und Copyright

Die Originalversion dieses Mitschriebs stammt von Stefan Breuning. Der *Mitschrieb* steht unter Public Domain; Copyright des Inhalts liegt bei Peter Bastian und Olaf Ippisch. Diese Version ist vom 1. Februar 2012, die aktuellste Version findet sich auf

http://conan.iwr.uni-heidelberg.de/teaching/numerik1_ws2011/.

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

olaf.ippisch@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

$y(t): [a, b] \rightarrow \mathbb{R}$ „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:

$$\begin{aligned}
 y(t + \Delta t) &= y(t) + \lambda \Delta t y(t) \\
 \Leftrightarrow \frac{y(t + \Delta t) - y(t)}{\Delta t} &= \lambda y(t) \\
 \lim_{\Delta t \rightarrow 0} \frac{dy(t)}{dt} &= \lambda y(t) \tag{1.1}
 \end{aligned}$$

Gewöhnliche Differentialgleichung (GDGL).

Lösung:

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

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

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

$$\Rightarrow y(t)e^{-\lambda t} = C \in \mathbb{R}$$

$$\Rightarrow \boxed{y(t) = Ce^{\lambda t}} \text{ Form aller Lösungen von 1.1}$$

\Rightarrow Festlegen der Konstante C durch eine Anfangsbedingung

Anfangswertaufgabe (AWA)

$$\frac{dy(t)}{dt} = \lambda y(t) \quad t \in (a, b] \tag{1.2a}$$

$$y(a) = Y \tag{1.2b}$$

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

also

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

Logistisches Wachstumsmodell:

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

$$\frac{dy(t)}{dt} = \lambda \cdot (1 - y(t)) \cdot y(t)$$

→ nichtlineare GDGL

Einfache numerische Lösungsverfahren

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

Expliziter Euler:

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

$$\frac{dy}{dt}(t_{i-1}) \approx \frac{y_i^h - y_{i-1}^h}{h} = \lambda y_{i-1}^h$$

Impliziter Euler:

Ansatz:

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

$$\Leftrightarrow (1 - h\lambda)y_i^h = y_{i-1}^h$$

$$\Leftrightarrow y_i^h = \frac{1}{1-h\lambda} y_{i-1}^h \quad i = 1, \dots, N$$

⇒ im allgemeinen deutlich höherer Aufwand. Frage: Lohnt das?

1.2 Chemische Reaktionen

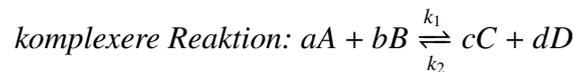
A und B reagieren zu C ($A + B \rightarrow C$).

Sei $c_A(t)$ die Konzentration (z.B. Mol pro Volumen) von A zur Zeit t . Analog: $c_B(t), c_C(t)$.

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

$$\Leftrightarrow \begin{aligned} \frac{dc_A(t)}{dt} &= -kc_A(t)c_B(t) \\ \frac{dc_B(t)}{dt} &= -kc_A(t)c_B(t) \\ \frac{dc_C(t)}{dt} &= +kc_A(t)c_B(t) \end{aligned}$$

System von GDGL + Anfangsbedingungen



$c_x(t)$ für $x \in \{A, B, C, D\}$

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

$$\begin{aligned} \frac{dc_A(t)}{dt} &= aR(t) & \frac{dc_B(t)}{dt} &= bR(t) \\ \frac{dc_C(t)}{dt} &= -cR(t) & \frac{dc_D(t)}{dt} &= -dR(t) \end{aligned}$$

chemisches Gleichgewicht:

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

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

$$\underbrace{C_M}_{\text{Membran- kapazität}[\frac{As}{Vm^2}]} \frac{dv}{dt} = \underbrace{I_{\text{ext}}(v, t)}_{\text{Strominjektion Synapse}[Am^{-2}]} - g_{\text{Na}} m^3 h (v - v_{\text{Na}}) - g_{\text{K}} n^4 (v - v_{\text{K}}) - g_{\text{L}} (v - v_{\text{L}})$$

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

$$\frac{dk}{dt} = \alpha_k(v)(1 - k) - \beta_k(v)k \quad k \in \{n, m, h\}$$

⇒ Nobelpreis 1963 Hodgkin/Huxley

Betrachtung der räumlichen Ausbreitung

$$\underbrace{c_M}_{\text{spez. Membran- kapazität}[\frac{As}{Vm}]} \frac{\partial v(x, t)}{\partial t} = \frac{\partial}{\partial x} \left(\underbrace{Q_m}_{\text{spez. Membran- leitfähigkeit}[\Omega^{-1}m]} \frac{\partial v}{\partial x} \right) + i_{\text{ext}}(x, t, v(x, t)) - i_{\text{Na}}(\dots) - i_{\text{K}}(\dots) - i_{\text{L}}(\dots)$$

→ Erstes Beispiel für partielle DGL. Zusätzlich sind hier *Randbedingungen* anzugeben, z.B. v oder $Q_m \frac{\partial v}{\partial x}$ am Anfang/Ende des Axons.

Zeichnung zu einer Nervenzelle. x_L und x_R bezeichnen die Ränder des Axons in „Stromrichtung“.

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

$$v(x, t_0) = v_0(x)$$

Anfangsbedingung

$$v(x_L, t) = v_0(t) \text{ und } v(x_R, t) = v_R(t)$$

„Randbedingung“

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

$$\frac{d}{dx} \left(\rho_m \frac{dv(x)}{dx} \right) + i(v) = 0$$

$$v(x_L) = v_L$$

$$v(x_R) = v_R$$

„Randwertproblem“ 2. Ordnung

→ erster Teil der Vorlesung *nur* AWA.

$$\begin{aligned} \rho_m \frac{dv(x)}{dx} &= u(x) \\ \frac{du(x)}{dx} &= -i(v(x)) \end{aligned}$$

⇒ 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 Schwerfeld.

$$\rightarrow x_i(t): [t_0, t_1] \rightarrow \mathbb{R}^3 \quad v_i(t): [t_0, t_1] \rightarrow \mathbb{R}^3$$

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

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

2. Newtonsches Gesetz ($F = \frac{dp}{dt} = ma$)

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

Mit $a_i(t) = \frac{dv_i(t)}{dt}$ und $\frac{dx_i(t)}{dt} = v_i(t)$ erhält man:

$$\frac{dx_i(t)}{dt} = v_i(t) \quad x_i(t_0) = x_i^0 \quad (1.12a)$$

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

$\Rightarrow 6N$ gew. DGL

Potential:

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

Setze

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

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

also:

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

Dieses System ist:

- nicht dissipativ
- konservativ

Es gibt:

- potentielle Energie:

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

- kinetische Energie:

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

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

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)

$$\frac{dp(t)}{dt} = \underbrace{F_G(t)}_{\text{Gewichtskraft}} + \underbrace{F_T(t)}_{\text{Antrieb}}$$

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

ergibt zusammen das System:

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

$$\frac{dx(t)}{dt} = v(t) \quad \text{NB: } c(t) \geq 0 \quad \int_{t_0}^{t_1} c(t) dt \leq m(t_0)$$

$$\frac{dm(t)}{dt} = -c(t) \quad \rightarrow m(t) = m(t_0) - \int_{t_0}^t \underbrace{c(s)}_{\text{(Ausgestossene Masse)}} ds$$

⇒ 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-algebraische Systeme.

2 Zur Theorie von gewöhnlichen Differentialgleichungen

2.1 Problemstellung

$$u'(t) = f(t, u(t)) \quad (2.1)$$

mit

$$u(t) = (u_1(t), \dots, u_d(t)) \quad f(t, x) = (f_1(t, x), \dots, f_d(t, x))$$

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

Standardnotation:

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

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

$$u'(t) = \frac{du(t)}{dt} \quad f'_t(t, x) = \frac{\partial f(t, x)}{\partial t} \quad \partial_i f(t, x) = \frac{\partial f(t, x)}{\partial x_i}$$

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

Implizite Formen

(2.1) ist Spezialfall der allgemeinen, impliziten Form

$$F(t, u(t), u'(t)) = 0 \quad (2.2)$$

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

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

Ist nicht nach $u(t)$ auflösbar, da es sich um ein DAE-System handelt.

In der Praxis kommt oft folgende Form vor:

$$\frac{dm(t, u(t))}{dt} = f(t, u(t)) \quad (2.3)$$

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

Reduktion von Gleichungen höherer Ordnung

Eine DGL höherer Ordnung der Form

$$F\left(t, u(t), \frac{du(t)}{dt}, \dots, \frac{d^m u(t)}{dt^m}\right) = 0$$

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

$$\left. \begin{aligned} \frac{du_0(t)}{dt} &= u_1(t) \\ \vdots &\vdots \\ \frac{du_{m-2}(t)}{dt} &= u_{m-1}(t) \end{aligned} \right\} m-1\text{-Gleichungen}$$

$$u_0(t) = u(t) \quad u_i(t) = \frac{d^i u(t)}{dt^i}$$

$$F\left(t, u_0(t), u_1(t), \dots, u_{m-1}(t), \frac{du_{m-1}(t)}{dt}\right) = 0$$

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

$$\begin{aligned}\bar{u}(t) &:= (u(t), t) \\ \bar{f}(\bar{u}(t)) &:= (f(t, u(t)), 1)^T\end{aligned}$$

Integralgleichungsform

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

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

Anfangswertaufgabe (AWA)

Definition 2.1

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

- 1) $\text{Graph}(u) := \{(t, u(t)) : t \in I\} \subset D$
- 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

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

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

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

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 $t \rightarrow t_0 + T^* = \infty$, als auch durch

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

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

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

mit einer festen stetigen Funktion $\beta: \mathbb{R} \rightarrow \mathbb{R}$.

Dann lässt sich u zu einer Lösung auf ganz \mathbb{R} fortsetzen.

Beweis .

[Rannacher Korollar 1.1](#) □

Beispiele:

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

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

$$\begin{aligned} |u(t)| &= \left| u(t_0) + \int_{t_0}^t \sin(u(s)) \, ds \right| \\ &\leq \underbrace{|u(t_0)|}_{=0} + \int_0^t \underbrace{|\sin(u(s))|}_{\leq 1} \, ds \leq t \end{aligned}$$

→ Lösung auf ganz \mathbb{R} fortsetzbar nach [Korollar 2.4](#)

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

Für beliebiges $C \geq 0$ ist jede Funktion

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

eine Lösung.

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

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

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

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

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

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

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

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

die auf ganz \mathbb{R} existierende Lösung

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

Kleine Änderung, große Wirkung.

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

$$\operatorname{Re} \lambda < 0 : \lim_{t \rightarrow \infty} |u(t)| = 0$$

$$\operatorname{Re} \lambda = 0 : \lim_{t \rightarrow \infty} |u(t)| = |u_0|$$

$$\operatorname{Re} \lambda > 0 : \lim_{t \rightarrow \infty} |u(t)| = \infty$$

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 $m \geq 1$, dann ist $u \in C^{m+1}(I)$.

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

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:

$$\|f(t, x) - f(t, x')\| \leq L(t)\|x - x'\| \quad (t, x), (t, x') \in D \quad (2.5)$$

- 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:

- $d = 1$. 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

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

lässt sich erweitern auf $d > 1$.

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

$$u'(t) = f(t, u(t)) \quad t \in I \quad u(t_0) = u_0 \quad (2.6a)$$

$$v'(t) = g(t, v(t)) \quad t \in I \quad v(t_0) = v_0 \quad (2.6b)$$

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).

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

mit

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

Beweis. Rannacher □

Mit der Integraldarstellung gilt:

$$\begin{aligned} u(t) - v(t) &= u(t_0) + \int_{t_0}^t f(s, u(s)) \, ds - v(t_0) - \int_{t_0}^t g(s, v(s)) \, ds \\ &= \int_{t_0}^t f(s, u(s)) - f(s, v(s)) \, ds + \int_{t_0}^t f(s, v(s)) - g(s, v(s)) \, ds + u_0 - v_0 \end{aligned}$$

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

$$\begin{aligned} \left\| \int_{t_0}^t e(s) \, ds \right\| &= \left\| \lim_{N \rightarrow \infty} \sum_{i=1}^N e(S_i) \cdot (S_i - S_{i-1}) \right\| \stackrel{\text{stetig}}{=} \lim_{N \rightarrow \infty} \left\| \sum_{i=1}^N e(S_i)(S_i - S_{i-1}) \right\| \\ &= \lim_{N \rightarrow \infty} \sum_{i=1}^N \|e(S_i)\| (S_i - S_{i-1}) = \int_{t_0}^t \|e(s)\| \, ds \end{aligned}$$

Dann gilt für $e(t) = u(t) - v(t)$:

$$\begin{aligned} \|e(t)\| &\leq \int_{t_0}^t \|f(s, u(s)) - f(s, v(s))\| \, ds + \int_{t_0}^t \|f(s, v(s)) - g(s, v(s))\| \, ds + \|u_0 - v_0\| \\ &\leq L \int_{t_0}^t \|e(s)\| \, ds + \int_{t_0}^t \varepsilon(s) \, ds + \|u_0 - v_0\| \end{aligned}$$

Lemma 2.8 Gronwall

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

$$w(t) \leq \int_{t_0}^t a(s)w(s) \, ds + b(t) \quad t \geq t_0$$

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

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

Beweis. Setze

$$\varphi(t) := \int_{t_0}^t a(s)w(s) \, ds$$

$$\psi(t) := w(t) - \int_{t_0}^t a(s)w(s) \, ds \leq b(t)$$

$\varphi(t)$ erfüllt

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

Weiter ist

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

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

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

Diese AWA hat die (eindeutige) Lösung:

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

Einsetzen in die Integralgleichung

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

□

Damit weiter im Satz von 2.7:

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

also nach Gronwall:

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

Korollar 2.9 Eindeutigkeitsatz

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.

$$\begin{aligned} u'(t) &= f(t, u(t)) & u(t_0) &= u_0 \\ v'(t) &= f(t, v(t)) & v(t_0) &= v_0 = u_0 \end{aligned}$$

Dann gilt nach dem Stabilitätssatz:

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

□

Korollar 2.10

Betrachte die DGL d -ter Ordnung

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

$f: I \times \mathbb{R}^d$ sei Lipschitz-stetig bzgl. der letzten d Argumente. Dann existiert für jeden Satz von d Werten $u_0, \dots, u_{d-1} \in \mathbb{R}$ 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

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

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) .

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

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 $t \geq t_0$. \square

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

$$\text{Matrixfunktion: } A: [t_0, \infty) \rightarrow \mathbb{R}^{d \times d}$$

$$\text{Vektorfunktion: } b: [t_0, \infty) \rightarrow \mathbb{R}^d$$

Dann hat die lineare AWA:

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

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

Beweis.

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

\square

2.4 Globale Stabilität

- bisherige Abschätzung im Stabilitätssatz sehr pessimistisch
- für $T \rightarrow \infty$ 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:

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

Verallgemeinerung von „monoton fallend“ für vektorwertige f :

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

$$\begin{aligned} -(f(x) - f(y), x - y) &= -(f(x) - f(y))(x - y) \geq \lambda(x - y)^2 \\ &\Leftrightarrow -\frac{f(x) - f(y)}{x - y} \geq \lambda \\ &\stackrel{(-1)}{\Leftrightarrow} \underbrace{\frac{f(x) - f(y)}{x - y}}_{\text{Steigung}} \leq -\lambda < 0 \end{aligned}$$

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

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

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

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

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

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

$u(t) = u^e$ 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

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

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

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

Setze

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

also

$$\begin{aligned} w'(t) &= (f(t, v(t)) - f(t, u(t))) = 0 \mid \text{multipliziere Skalar mit } w(t) \\ \underbrace{(w'(t), w(t))}_{\text{Skalarprodukt}} &= (f(t, v(t)) - f(t, u(t)), w(t)) = 0 \end{aligned}$$

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

$$\Leftrightarrow 0 = \frac{1}{2} \frac{d}{dt} \|w(t)\|^2 - \underbrace{(f(t, v(t)) - f(t, u(t)), v(t) - u(t))}_{\text{Skalarprodukt}} = 0$$

Monotoniebedingung:

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

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

$$\begin{aligned} e^{2\lambda(t-t_*)} \frac{d}{dt} \|w(t)\|^2 + 2\lambda e^{2\lambda(t-t_*)} \|w(t)\|^2 &\leq 0 \\ \Leftrightarrow \frac{d}{dt} [e^{2\lambda(t-t_*)} \|w(t)\|^2] &\leq 0 \end{aligned}$$

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

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

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

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

eine eindeutige globale Lösung

$$u: [t_0, \infty) \rightarrow \mathbb{R}^d$$

die beschränkt und exponentiell stabil ist.

Beweis .

- (i) Lipschitzbedingung \rightarrow eindeutige, globale Lsg.
- (ii) $A(t)$ gleichmässig negativ definit \rightarrow AWA ist monoton (siehe oben)
- (iii)

$$\sup_{t \in [t_0, \infty)} \|f(t, 0)\| = \sup_{t \in [t_0, \infty)} \|b(t)\| \underbrace{\leq \infty}_{\text{nach Vor.}}$$

somit ist die Lösung nach Satz 2.15 beschränkt.

□

3 Einschrittverfahren

Wir betrachten die AWA

$$u'(t) = f(t, u(t)) \quad t \in [t_0, t_0 + T] = I \quad u(t_0) = u_0 \quad (3.1)$$

$$\text{Die Funktion } f \text{ sei stetig und erfülle eine } L\text{-Bedingung.} \quad (3.2)$$

\Rightarrow globale, eindeutige Lösung

Zeitgitter: Unterteile $I = [t_0, t_0 + T]$ in

$$t_0 < t_1 < \dots < t_N = t_0 + T \quad T < \infty \quad N \in \mathbb{N}$$

und setze:

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

„nicht-äquidistantes Gitter“

3.1 Das explizite Euler-Verfahren

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

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

und

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

Setze

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

und definiere den „Differenzenoperator“

$$L_h: \mathbb{R}^{Nd} \times \mathbb{R}^d \rightarrow \mathbb{R}^{Nd} \quad \left(L_h(y^h, y_0^h) \right)_n = h_n^{-1} (y_n^h - y_{n-1}^h) - f(t_{n-1}, y_{n-1}^h) \quad 1 \leq n \leq N \quad (3.4)$$

y^h ist Lösung der Gleichung:

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

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

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

$$u^h := (u_1^h)^T, \dots, (u_N^h)^T)^T$$

Lokaler Diskretisierungsfehler

Die folgende Größe

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

heißt „Abschneidefehler“ oder „lokaler Diskretisierungsfehler“.

Für den Abschneidefehler gilt:

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

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

Globaler Diskretisierungsfehler

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

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

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:

$$\|e_n^h\| \leq \underbrace{\|e_0^h\|}_{=y_0-u_0} + L \sum_{v=0}^{n-1} h_{v+1} \|e_v^h\| + \sum_{v=1}^n h_v \|\tau_v\| \tag{3.6}$$

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

$$w_0 \leq b_0 \text{ und } w_n \leq \sum_{v=0}^{n-1} a_v w_v + b_n \quad n \geq 1$$

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

$$w_n \leq \exp\left(\sum_{v=0}^{n-1} a_v\right) b_n \quad n \geq 1$$

Beweis .

Definiere

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

Zeige per Induktion

$$S_n \leq \exp\left(\sum_{v=0}^{n-1} a_v\right) b_n \quad n \geq 0$$

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

2) $n - 1 \rightarrow n$

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

also

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

□

damit weiter für den globalen Fehler:

$$\begin{aligned} \|e_n^h\| &\leq \sum_{v=0}^{n-1} \underbrace{L h_{v+1}}_{a_v} \underbrace{\|e_v^h\|}_{w_v} + \underbrace{\sum_{v=1}^n h_v \|\tau_n^h\| + \|e_0^h\|}_{b_n} \\ &\stackrel{\text{Lemma 3.1}}{\leq} \underbrace{\exp\left(\sum_{v=0}^{n-1} L h_{v+1}\right)}_{=L(t_n-t_0)=LT} \left\{ \|e_0^h\| + \underbrace{\sum_{v=1}^n h_v \|\tau_n^h\|}_{\leq T} \right\} \\ &\leq e^{LT} \left\{ \|e_0^h\| + \frac{T}{2} h \cdot \max_{t \in I} \|u''(t)\| \right\} \end{aligned}$$

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

Kurzer Ausflug: impliziter Euler

$$\begin{aligned} e_n^h &= y_n^h - u_n^h = \underbrace{y_{n-1}^h + h_n f(t_n, y_n^h)}_{\text{impliziter Euler}} - \underbrace{(u_n^h - u_{n-1}^h - h_n f(t_n, u_n^h))}_{\text{hinzugefügt}} - u_{n-1}^h - h_n f(t_n, u_n^h) \\ &= y_{n-1}^h - u_{n-1}^h + h_n (f(t_n, y_n^h) - f(t_n, u_n^h)) - h_n \tau_n^h \\ \|e_n^h\| &\leq \|e_{n-1}^h\| + h_n L \|e_n^h\| + h_n \|\tau_n^h\| \\ &\leq \sum_{v=1}^n h_v L \|e_v^h\| + \|e_0^h\| + \sum_{v=1}^n h_v \|\tau_v^h\| \end{aligned}$$

Man erhält also eine implizite Summengleichung:

$$\begin{aligned} w_n &\leq \sum_{v=0}^n a_v w_v + b_n & n \geq 1 \\ w_0 &\leq b_0 & n = 0 \end{aligned}$$

Sei $a_n < 1$, dann gilt

$$\begin{aligned} w_n &\leq a_n w_n + \sum_{v=0}^{n-1} a_v w_v + b_n \\ \underbrace{(1 - a_n) w_n}_{\tilde{w}_n} &\leq \sum_{v=0}^{n-1} \underbrace{\frac{a_v}{1 - a_v}}_{\tilde{a}_v} \underbrace{(1 - a_v) w_v}_{\tilde{w}_v} + b_n \end{aligned}$$

jetzt diskretes Gronwall-Lemma anwendbar:

$$\begin{aligned} \tilde{w}_n &\leq \exp\left(\sum_{v=0}^{n-1} \tilde{a}_v\right) b_n \\ \Leftrightarrow (1 - a_n) w_n &\leq \exp\left(\sum_{v=0}^{n-1} \frac{a_v}{1 - a_v}\right) b_n \\ w_n &\leq \underbrace{\frac{1}{1 - a_n}}_{\leq \exp\left(\frac{a_n}{1 - a_n}\right)} \exp\left(\sum_{v=0}^{n-1} \frac{a_v}{1 - a_v}\right) \\ \boxed{w_n} &\leq \exp\left(\sum_{v=0}^n \frac{a_v}{1 - a_v}\right) b_n \quad \text{wobei } \boxed{a_v < 1} \end{aligned}$$

zurück zum impliziten Euler: Bedingung $a_v < 1$ bedeutet Konvergenz falls $h, L < 1$ klein genug.

3.2 Taylor und Runge-Kutta Verfahren

Ziel: Konstruktion von Einschritt-Verfahren höherer Ordnung

Idee:

(a) Taylorentwicklung um $t - h$:

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

(b) Differenzieren der DGL:

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

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

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

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

⇒ 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

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

gilt

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

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

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

Runge-Kutta Verfahren

Idee: Ersetze Ableitungen (von f) durch numerische Approximation (Differenzenquotienten).

Beispiel: $R = 2$

$$\begin{aligned} f^{(1)}(t-h, u(t-h)) &= \frac{1}{h} (f(t, u(t)) - f(t-h, u(t-h))) + \mathcal{O}(h) \\ &= \frac{1}{h} \left(\underbrace{f(t, u(t-h) + hf(t-h, u(t-h)) + \mathcal{O}(h^2))}_{=:k_2} - f(t-h, u(t-h)) \right) + \mathcal{O}(h) \\ &= \frac{1}{h} \left(f(t, u(t-h) + hf(t-h, u(t-h))) + \mathcal{O}(h^2) - f(t-h, u(t-h)) \right) + \mathcal{O}(h) \\ &= \frac{1}{h} [f(t, u(t-h) + hf(t-h, u(t-h))) - f(t-h, u(t-h))] + \mathcal{O}(h) \end{aligned}$$

und damit für das 2-stufige Taylor-Verfahren:

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

somit erhält man das folgende Verfahren mit Konsistenzordnung 2:

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

Man nennt dies das Verfahren von Heun.

Das Beispiel motiviert den allgemeinen Ansatz für Runge-Kutta Verfahren:

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

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“:

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

A strikte unter Dreiecksmatrix \rightarrow explizit

A untere Dreiecksmatrix \rightarrow diagonal implizit

A voll \rightarrow voll implizit

$s = 1$

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

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

$s = 2$ (skalar)

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

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

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

vergleiche mit

$$\begin{aligned}u(t_n) &= u(t_{n-1}) + h_n f + h_n^2 \frac{1}{2} f_t + h_n^2 \frac{1}{2} f f_x + \mathcal{O}(h_n^3) \\ &\Rightarrow b_1 + b_2 = 1 \quad b_2 c_2 = \frac{1}{2} \quad b_2 a_{21} = \frac{1}{2}\end{aligned}$$

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:

$$\begin{array}{c|ccc} \text{Heun} & 0 & 0 & \\ \hline & 1 & 1 & 0 \\ \hline & & \frac{1}{2} & \frac{1}{2} \end{array} \quad \begin{array}{c|ccc} \text{Modifizierter Euler} & 0 & 0 & \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & & 0 & 1 \end{array}$$

Heun:

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

Modifizierter Euler:

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

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:

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

Runge-Kutta 3. Ordnung:

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

s=4

13 Parameter, 11 Gleichungen

Runge-Kutta 4. Ordnung:

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

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

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

ein Einschrittverfahren zur Lösung der AWA:

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

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

$$\|\tau_n^h\| \leq C \cdot h^m \quad (h \rightarrow 0)$$

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 !)

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

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

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

Normbilden, Dreiecksungleichung

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

(Graph zur Visualisierung des Fehlers)

(ii)

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

und damit

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

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

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

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

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

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

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

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

- **Lösbarkeit im impliziten Verfahren**

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

Lösung der nichtlinearen Gleichung mit Fixpunktiteration

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

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):

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

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 $\max_{1 \leq n \leq N} \|u(t_n) - y_n^h\| \leq \varepsilon$ 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):

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

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

$$\begin{aligned} \hat{y}_n^h - y_n^h &= \tilde{u}(t_n) - \hat{C}(t_{n-1})h_n^{m+2} + O(h_n^{m+3}) - (\tilde{u}(t_n) - \underbrace{C(t_{n-1})h_n^{m+1}}_{\text{Zusätzlich bei RK Verfahren der Ordnung } m} + O(h_n^{m+2})) \\ &= C(t_{n-1})h_n^{m+1} + O(h_n^{m+2}) \doteq C(t_{n-1})h_n^{m+1} \end{aligned}$$

Für die „optimale“ Schrittweite sollte gelten:

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

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

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

Algorithmus:

Eingabe: y_{n-1} , Vorschlag für h_n

Ausgabe: y_n , so dass geschätzter Fehler \leq 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

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

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

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

- 3) Falls $\|\hat{y}_n^h - y_n^h\| \leq \text{TOL}$. 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)

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

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	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$
	$\frac{16}{135}$		0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$
						$\frac{2}{55}$

Beispiel 2-Körper-Problem

$$\begin{aligned}
 G &= 1 & m_1 &= 1 & m_2 &= 0.01 \\
 x_1 &= (-1, 0)^T & v_1 &= 0 \\
 x_2 &= (1, 0)^T & v_2 &= (0, 1/5)
 \end{aligned}$$

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

Einbeziehung des globalen Fehlers

Ziel:

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

$$\tau_n^h = \underbrace{\tau^m(t_n)} h_n^m + O(h_n^{m+1})$$

hie oben $C(t) \rightarrow$ Taylormethode: $\tau^m(t_n) = \frac{1}{(m+1)!} u^{(m+1)}(t_{n-1})$

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

Strategie 1

Idee: Verteile Fehler gleichmäig auf das Intervall $[t_0, t_0 + T]$ Wähle

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

$$\begin{aligned}
 \max_{t_n \in I} \|e_n^h\| &\leq K \sum_{n=1}^N h_n \|\tau_n^h\| \doteq K \sum_{n=1}^N h_n h_n^m \|\tau^m(t_n)\| \\
 &\approx K \sum_{n=1}^N h_n \frac{\text{TOL}}{KT \|\tau^m(t_n)\|} \|\tau^m(t_n)\| = \frac{\text{TOL}}{K} \sum_{n=1}^N 1 = \text{TOL}
 \end{aligned}$$

$$\begin{aligned}
 N &= \sum_{n=1}^N h_n h_n^{-1} = \sum_{n=1}^N h_n \left(\frac{KT \|\tau^m(t_n)\|}{\text{TOL}} \right)^{1/m} \\
 &= \left(\frac{KT}{\text{TOL}} \right)^{1/m} \underbrace{\sum_{n=1}^N h_n \|\tau^m(t_n)\|^{1/m}}_{\text{Taylor-Methode}} \approx \left(\frac{KT}{\text{TOL} (m+1)!} \right)^{1/m} \int_I \|u^{(m+1)}\|^{1/m} dt
 \end{aligned}$$

Strategie 2

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

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

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

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

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

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

$$N \approx \left(\frac{K}{\text{TOL}(m+1)!} \right)^{\frac{1}{m}} \left(\int_I \|u^{(m+1)}\|^{\frac{1}{m+1}} dt \right)^{\frac{m+1}{m}}$$

Vergleich der Strategien

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

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

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

Wende Cauchy-Schwarz an:

$$\begin{aligned} \int_I \|u''(t)\|^{1/2} \cdot 1 dt &\leq \left(\int_I \|u''(t)\| dt \right)^{1/2} \cdot \left(\int_I 1 dt \right)^{1/2} \\ \Rightarrow \left(\int_I \|u''(t)\|^{1/2} dt \right)^2 &\leq \int_I \|u''(t)\| dt \cdot T \\ \Rightarrow N_2 &\leq N_1 \end{aligned}$$

Strategie 2 effizienter, falls

$$\left(\int_I \|u''(t)\|^{1/2} dt \right)^2 \ll \int_I \|u''(t)\| dt \cdot T,$$

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 $y_n^H - y_n^{H/2}$

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

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

für Schrittweite H folgt:

$$y_n^H - u(t_n) = e_{n-1} + H \cdot \underbrace{(F(H, t_{n-1}, y_{n-1}^H) - F(H, t_{n-1}, u(t_{n-1})))}_{=u(t_{n-1})+y_{n-1}^H-u(t_{n-1})} - H\tau_n^H$$

$$= e_{n-1} + H \cdot (\underbrace{F(H, t_{n-1}, u(t_{n-1}))}_{=u(t_{n-1})+y_{n-1}^H-u(t_{n-1})} + e_{n-1} F_x(H, t_{n-1}, \xi) - \underbrace{F(H, t_{n-1}, u(t_{n-1}))}_{=u(t_{n-1})+y_{n-1}^H-u(t_{n-1})}) - H\tau_n^H$$

$$= (1 + O(H))e_{n-1} - H\tau^m(t_n) + O(H^{m+2})$$

Schrittweite $\frac{H}{2}$

$$e_{n-\frac{1}{2}}^{\frac{H}{2}} = (1 + O(H))e_{n-1} - \left(\frac{H}{2}\right)^{m+1} \tau^m\left(t_n - \frac{H}{2}\right) + O(H^{m+2})$$

$$e_n^{\frac{H}{2}} = y_n^{\frac{H}{2}} - u(t_n) = e_{n-\frac{1}{2}}^{\frac{H}{2}} + \frac{H}{2} \left(F\left(\frac{H}{2}, t_{n-\frac{1}{2}}, y_{n-\frac{1}{2}}^{\frac{H}{2}}\right) - F\left(\frac{H}{2}, t_{n-\frac{1}{2}}, u\left(t_{n-1} + \frac{H}{2}\right)\right) \right) - \frac{H}{2} \tau_n^{\frac{H}{2}}$$

$$= (1 + O(H))e_{n-\frac{1}{2}}^{\frac{H}{2}} - \left(\frac{H}{2}\right)^{m+1} \tau^m(t_n) + O(H^{m+2})$$

$$= (1 + O(H))e_{n-1} - 2\left(\frac{H}{2}\right)^{m+1} \tau^m(t_n) + O(H^{m+2})$$

Differenz bilden

$$y_n^{\frac{H}{2}} - y_n^H = \underbrace{y_n^{\frac{H}{2}} - u(t_n)}_{=e_{n-\frac{1}{2}}^{\frac{H}{2}}} - \underbrace{y_n^H - u(t_n)}_{=e_n^H} = e_{n-\frac{1}{2}}^{\frac{H}{2}} - e_n^H$$

$$= O(H)e_{n-1} - \left(2\left(\frac{H}{2}\right)^{m+1} - H^{m+1}\right) \tau^m(t_n) + O(H^{m+2})$$

$$\|\tau^m(t_n)\| = \frac{\|y_n^{\frac{H}{2}} - y_n^H\|}{H^{m+1}(1 - 2^{-m})} + \|e_{n-1}\| O(H^{-m}) + O(H)$$

- 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 AlgorithmusEingabe: 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) $\alpha < 1 < \beta$

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

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

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

sonst:

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

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 5Extrapolation + 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

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

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

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

	$\lambda = -10$	$\lambda = -100$
	$h = 0.1 \rightarrow 1, 0, 0, \dots$	$h = 0.01 \rightarrow 1, 0, 0, \dots$
expliziter Euler	$h = 0.2 \rightarrow 1, -1, 1, -1, \dots$	$h = 0.02 \rightarrow 1, -1, 1, -1, \dots$
	$h > 0.2 \rightarrow$ Divergenz	$h > 0.02 \rightarrow$ 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

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

Beispiel 4.3 Simeon

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

hat die Lösung

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

(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 $\text{Re}(\lambda_i) < 0$ für alle EW λ_i heißt steif, wenn

$$\frac{\max_i |\text{Re}(\lambda_i)|}{\min_i |\text{Re}(\lambda_i)|} \gg 1$$

(im nichtlinearen Fall nehme $f_x(t, u(t))$)

- 3) Differentialgleichungen der Form

$$\begin{aligned} u' &= f(t, u, z) && \text{mit } \varepsilon \in \mathbb{R}. \varepsilon \ll 1 \text{ heißen } \textit{singular gestört} \\ \varepsilon' &= f(t, u, z) && \text{und führen für } \varepsilon \rightarrow 0 \text{ auf steife AWAn} \end{aligned}$$

4.2 (Skalare, lineare) Modellprobleme

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

dient nun als Modellproblem zur Bewertung unserer Verfahren.

Definition 4.4 absolute Stabilität

Eine Einzschrittmethode 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

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

heißt *Stabilitätsgebiet* einer Einschrittformel.

SG des expliziten Euler (EE): $\text{SG} = \{z \in \mathbb{C} \mid |1 + z| \leq 1\}$

(Zeichnung des Kreises in der komplexen Zahlenebene)

Für das Taylor-Verfahren der Stufe R gilt:

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

- (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:

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

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

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

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

$\Rightarrow k_i$ 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)

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

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

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 expliziten 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:

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

Also:

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

impliziter Euler

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

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

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

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 $\lim_{z \rightarrow -\infty} \omega(z) = \pm\infty \Rightarrow \nexists$ A-stabilen Taylor und *expliziten* RK-Verfahren.

4.3 Lineare Stabilitätsanalyse

\rightarrow Erweitern der Modellproblemanalyse auf nichtlineare Systeme

Stabilität beim skalaren MP: Näherungen für $\operatorname{Re}(\lambda) < 0$ bleiben beschränkt.

Definition 4.6

Die Lösung u der L-stetigen AWA

$$u'(t) = f(t, u(t)) \quad t \geq t_0 \quad u(t_0) = u_0 \quad (4.2)$$

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

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

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

$$\|v(t) - u(t)\| \rightarrow 0 \quad (t \rightarrow \infty)$$

Übertragen auf den diskreten Fall:

Definition 4.7

Die AWA (4.2) sei mit dem Einschrittverfahren

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

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

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

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

$$\|z_n - y_n\| \rightarrow 0 \quad (n \rightarrow \infty)$$

Übertragen der Erkenntnisse vom skalaren Modellproblem führt auf

Hypothese 4.8

Ein Differenzenverfahren mit Stabilitätsgebiet $\text{SG} \subset \mathbb{C}$ 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 $\text{Re}(\lambda(t)) \leq 0$ gilt:

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

Bemerkungen:

- Hypothese und kein Satz, also nicht rigoros beweisbar.
- Man sollte hier

$$\text{SG} = \left\{ z = \lambda h \in \mathbb{C} \mid |\omega(z)| \underbrace{\leq}_{\text{wichtig}} 1 \right\}$$

verwenden

- Sei AWA linear: $u' = Au$ und A diagonalisierbar. Dann reduziert sich die AWA auf entkoppelte, skalare Probleme und mit $h_n \lambda_i \in \text{SG}$ 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 $\rightarrow 0$ geht obwohl $h_n \lambda_i(t) \in \text{SG}$.

4.4 Implizite Runge-Kutta Verfahren

Nichtlineares System für die k_i :

$$k_i = f \left(t_{n-1} + c_i h_n, y_{n-1}^h + h_n \sum_{j=1}^i a_{ij} k_j \right) \quad i = 1, \dots, s \quad (4.3)$$

Anschließend kombiniere

$$y_n^h = y_{n-1}^h + h_n \sum_{i=1}^s b_i k_i \quad (4.4)$$

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

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

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

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

Zentrale Frage: Lösbarkeit und Lösung von (4.4) \rightarrow 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) Einschnitt Θ -Verfahren, $\Theta \in [0, 1]$

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

also

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

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

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

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

(c) Mittelpunkregel

$$\begin{aligned}
 y_n^h &= y_{n-1}^h + h_n \underbrace{f\left(t_{n-1} + \frac{h_n}{2}, \frac{1}{2}(y_{n-1}^h + y_n^h)\right)}_{=:k_1} \\
 k_1 &= f\left(t_{n-1} + \frac{h_n}{2}, \frac{1}{2}y_{n-1}^h + \frac{1}{2}(y_{n-1}^h + h_n k_1)\right) \\
 &= f\left(t_{n-1} + \frac{h_n}{2}, y_{n-1}^h + \frac{h_n}{2}k_1\right) \\
 &\quad \frac{\frac{1}{2} \mid \frac{1}{2}}{1} \quad \text{Konsistenzordnung 2}
 \end{aligned}$$

Lemma 4.9

Trapez- und Mittelpunkregel sind A-stabil. Beweis siehe Übungen

Lemma 4.10

Die Stabilitätsfunktion eines allgemeinen IRK-Verfahrens hat die Darstellungen:

(i)

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

(ii)

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

(iii)

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

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

(i)

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

$$\Leftrightarrow (I - h_n \lambda A) k \stackrel{(*)}{=} \lambda y_{n-1}^h \mathbb{1}$$

$$\Leftrightarrow k = (I - h_n \lambda A)^{-1} \lambda y_{n-1}^h \mathbb{1}$$

und dann

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

(ii) Fasse (★) und (★★) 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 $t_0 = 0$

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

\Rightarrow für $h \rightarrow 0$ 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

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

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

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

Padé-Tafel für $\exp(z)$:

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

(Die Einträge je rechts sind die Fehlerordnungen der Padé-Approximation, 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 $z \rightarrow \infty$ genauer:

(i)

$$\lim_{y \rightarrow \infty} |\exp(-y)| = 0$$

reelle Achse

$$\lim_{y \rightarrow \infty} |\exp(\pm iy)| = 1$$

imaginäre Achse

(ii)

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

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

- Verfahren mit $\lim_{z \rightarrow \infty} |\omega(z)| = 1$ (Hauptdiagonale der Padé-Tafel) sind für Schwingungsprobleme geeignet
- Verfahren mit $\lim_{z \rightarrow \infty} |\omega(z)| = 0$ 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

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

Kollokationsverfahren

Das Polynom w von Grad s welches den Bedingungen

$$\begin{aligned} w(t_{n-1}) &= y_{n-1}^h \\ w'(t_{n-1} + c_i h_n) &= f(t_{n-1} + c_i h_n, w(t_{n-1} + c_i h_n)) \quad i = 1, \dots, s \\ y_n^h &:= w(t_n) \end{aligned} \tag{4.5}$$

genügt heißt „Kollokationspolynom“.

- 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, \dots, c_s entspricht das Kollokationsverfahren (4.5) einem IRK mit:

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

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 $s - 1$.

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

(ii)

$$\int_0^{c_i} w'(t_{n-1} + th_n) dt = \left[\frac{1}{h_n} w(t_{n-1} + th_n) \right]_0^{c_i} = \frac{1}{h_n} (w(t_{n-1} + c_i h_n) - y_{n-1}^h)$$

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

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

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

(iii)

$$y_n^h = w(t_{n-1} + h_n) = y_{n-1}^h + \sum_{j=0}^s \left(\int_0^1 L_j(t) dt \right) k_j$$

□

Wähle Stützstellen $c_1, \dots, c_s \Rightarrow$ 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:

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

(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

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

$$\begin{aligned} [w(t) - u(t)]_{t_{n-1}}^{t_n} &= \int_{t_{n-1}}^{t_n} w'(t) - u'(t) dt = \int_{t_{n-1}}^{t_n} r(t) dt \\ &\stackrel{\text{Quadratur}}{=} h \cdot \sum_{i=1}^s b_i \underbrace{(w'(t_{n-1} + c_i h_n) - f(t_{n-1} + c_i h_n, w(t_{n-1} + c_i h_n)))}_{=0 \text{ wegen (4.5)}} + O(h^{p+1}) \\ e(t_n) = h_n \tau_n^h &\Rightarrow \tau_n^h = O(h^p) \end{aligned}$$

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 \rightarrow A-stabil
 $s = 1$: Mittelpunkregel

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

(ii) Radau-Verfahren

$$\begin{array}{lll} \text{IA: } c_i \text{ Nullstelle von } \frac{d^{s-1}}{dx^{s-1}} [x^s (x-1)^{s-1}] & c_1 = 0 & c_s < 1 \\ \text{IIA: } c_i \text{ Nullstelle von } \frac{d^{s-1}}{dx^{s-1}} [x^{s-1} (x-1)^s] & c_1 > 0 & c_s = 1 \end{array}$$

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

$$s = 1: \text{impliziter Euler} \quad s = 2: \begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}$$

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

$$s = 2, p = 2: \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array} \rightarrow \text{Heun} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/4 & 1/4 & 0 \\ 1 & 0 & 1 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

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

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

Beispiele

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

$$\begin{array}{c|cc} \alpha & \alpha & 0 \\ 1 & 1 - \alpha & \alpha \\ \hline & 1 - \alpha & \alpha \end{array} \leftarrow b^T \quad \alpha = 1 \pm \frac{\sqrt{2}}{2} = \text{letzte Zeile von A spart weitere Berechnungen}$$

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

$$\begin{array}{c|cc} 1/2 + 1/2 \cdot \sqrt{3} & 1/2 + 1/2 \cdot \sqrt{3} & 0 \\ 1/2 - 1/2 \cdot \sqrt{3} & -1/\sqrt{3} & 1/2 + 1/2 \cdot \sqrt{3} \\ \hline & 1/2 & 1/2 \end{array}$$

Lösung der nichtlinearen Systeme

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

ist in Fixpunktform

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

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

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

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

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

$$x^{(k+1)} = x^{(k)} + \underbrace{z^{(k)}} = x^{(k)} - (F_x(x^{(k)}))^{-1} F(x^{(k)})$$

Lösung von $F(x^{(k)}) + F_x(x^{(k)})z^{(k)} = 0$

- 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

$$\|F(x^{(k)} + \lambda^l z^{(k)})\| \leq (1 - \frac{1}{4} \lambda^l) \|F(x^{(k)})\| \quad \text{„line search“}$$

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:

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

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

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

Kontraktion falls

$$\mathbb{Z}_Z : \|g(x) - g(\tilde{x})\| \leq \underbrace{\dots}_{\text{nicht groß!}} \|x - \tilde{x}\|$$

$$h_n \underbrace{\|(I - h_n A)^{-1}\|}_{(I - h_n A)^{-1} \text{ hat EW } \frac{1}{1 - h_n \lambda_i}} \underbrace{(\|L_b\| + \|b_x(y_n^{(k)})\|)}_{\lambda_j \text{ EW von } A} < 1$$

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 $z = (z_1, \dots, z_s)^T$, $k = (k_1, \dots, k_s)$ gilt

$$\boxed{z = h_n A k} \rightarrow k = \frac{1}{h_n} A^{-1} z \text{ f\u00fcr } \det(A) \neq 0$$

F\u00fcr die z_i gilt:

$$\begin{aligned} z_i &= h_n \sum_{j=1}^s a_{ij} k_j = h_n \sum_{j=1}^s a_{ij} f(t_{n-1} + c_j h_n, y_{n-1}^h + \underbrace{h_n \sum_{l=1}^s a_{jl} k_l}_{z_j}) \\ &= h_n \sum_{j=1}^s a_{ij} f(t_{n-1} + c_j h_n, y_{n-1}^h + z_j) \quad \text{weniger anf\u00e4llig gegen Ausl\u00f6schung} \end{aligned}$$

Au\u00dferdem: falls $b^T =$ letzte Zeile von A (z.B. Alexander), dann gilt

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

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

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

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

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

und damit

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

(Beweis siehe Übungen)

Verfahrensklassen:

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

(Grafik zur Erläuterung)

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

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

(Grafik)

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

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

(Grafik)

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

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

(Grafik)

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

Differentiationsbasierte Verfahren

Idee: Lege Polynom vom Grad m durch die Funktionswerte:

$$p_m(t) = \sum_{\mu=0}^m \underbrace{u(t_{k-\mu})}_{\text{zu w\u00e4hlen}} \cdot L_\mu(t)$$

Fehler der Ableitung:

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

$k = n$ ergibt „R\u00fcckw\u00e4rtdifferenzenformeln“ (engl. BDF: „backward difference formula“)

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

- Geht analog f\u00fcr Systeme von DGL

Definition 5.1

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

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

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

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

f\u00fcr 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 $y_{n-r} = u_{n-r}$, $r = 1, \dots, R$ gilt für jede LMM die Beziehung

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

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, $\alpha_R = 1$:

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

□

Definition 5.3

Eine LMM heißt konsistent mit der AWA, wenn

$$\max_{t_n \in I} \|\tau_n^h\| \rightarrow 0 \quad (h \rightarrow 0)$$

bzw. konsistent, von der Ordnung p falls

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

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

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

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. \square

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 $C_0 = C_1 = \dots = C_p = 0, C_{p+1} \neq 0$.

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 $C_0 = \dots = C_p = 0$ gilt

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

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:

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

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:

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

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

Untersuchung der homogenen Differenzgleichung

$$\alpha_R y_n + \alpha_{R-1} y_{n-1} + \dots + \alpha_0 y_{n-R} = 0 \quad n \geq R \quad (5.2)$$

und $\alpha_R = 1, \alpha_0 \neq 0$.

Neben der trivialen Lösung $(y_n)_{n \in \mathbb{N}} = (0, \dots)$ gibt es noch weitere, nicht triviale Lösungen.

Ansatz: $y_n = \lambda^n$ führt auf:

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

\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“ $\left(1 \quad \lambda_i \quad \lambda_i^2 \quad \dots \quad \lambda_i^{R-1} \right)$

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

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

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

⇒ 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:

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

5.4 Konvergenz von LMM

Definition 5.9 Konvergenz

Eine LMM heißt konvergent, wenn für jede AWA gilt:

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

vorausgesetzt die Startwerte konvergieren

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

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 $\|\cdot\|_{A,\varepsilon}$ sodass für den Spektralradius $\varrho(A) := \max\{|\lambda| \mid \lambda \text{ ist Eigenwert von } A\}$ 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 $\|\cdot\|_{A,0}$ mit $\|A\|_{A,0} = \varrho(A)$

Beweis .

Konstruktiv:

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

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

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

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

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

Dann gilt:

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

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

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

Also

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

also

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

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

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

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

mit $\max_{s+1 \leq j \leq m} |r_{jj}| < \varrho(A)$

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:

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

□

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.

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

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

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

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

$$\|\tilde{y} - y_n\| \leq K \cdot e^{\Gamma(t_n - t_0)} \left\{ \max_{0 \leq v < R} \|\varrho_v\| + \sum_{r=R}^n \|\varrho_v\| \right\} \quad (n \geq R)$$

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

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

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

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

In Matrixform lauten die R Gleichungen dann:

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

Indextransformation:

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

d.h.

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

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:

$$E_i = D_i^{-1}\{C_i A E_{i-1} + B_i\} \quad (i \geq 1) \quad (\text{V})$$

$$E_0 = (\varrho_{R-1}, \dots, \varrho_0)^T \quad (i = 0) \quad (\text{V})$$

(ii) Die Matrix A hat das charakteristische Polynom

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

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 Determinante nach erster Zeile liefert $\chi_A(\lambda)$.

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

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

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

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

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

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

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)

$$\begin{aligned}
\|E_i\|_0 &\leq \|D_i^{-1}\|_0 \left\{ \|C_i\|_0 \underbrace{\|A\|_0}_{=\varrho(A) \leq 1} \|E_{i-1}\|_0 + \|B_i\|_0 \right\} \\
&\leq \left(1 + \frac{\gamma^2 h|\beta_R|L}{1 - h|\beta_R|L} \right) \left\{ (1 + \gamma^2 h|\beta_R|L) \|E_{i-1}\|_0 + \gamma^2 h\beta L \|E_{i-1}\|_0 + \gamma |Q_{i+R-1}| \right\} \\
&= \|E_{i-1}\|_0 + h\Gamma \|E_{i-1}\|_0 + \Lambda |Q_{i+R-1}| \\
&\leq \|E_{i-2}\|_0 + h\Gamma \|E_{i-2}\|_0 + h\Gamma \|E_{i-1}\|_0 + \Lambda |Q_{i+R-1}| + \Lambda |Q_{i+R-2}| \\
&\leq h\Gamma \sum_{v=0}^{i-1} \|E_v\|_0 + \underbrace{\|E_0\|_0 + \Lambda \sum_{v=1}^i |Q_{v+R-1}|}_{=: b_i} \\
\|E_i\|_0 &\leq \exp \left(\underbrace{\Gamma \sum_{v=0}^{i-1} h}_{t_i - t_0} \right) \left(\|E_0\|_0 + \Lambda \sum_{v=1}^i |Q_{v+R-1}| \right)
\end{aligned}$$

\Rightarrow Behauptung

□

Satz 5.12 Konvergenz LMM

LMM sei Null-Stabil, AWA L-stetig

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

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

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

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

Beweis .

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

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. \square

Anwendung auf Adams- \star Verfahren

$$u(t_n) - u(t_{n-\sigma}) = \sum_{\mu=0}^R \int_{t_{n-\sigma}}^{t_n} L_{\mu}^{(m)}(s) \, ds \cdot f(t_{h-\mu}, u(t_{h-\mu})) + h \underbrace{\tau_n^h}_{O(h^{R+1})}$$

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

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

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

$$\begin{aligned} \Pi(\lambda, qh) &:= \sum_{r=0}^R (\alpha_r - hq\beta_r) \lambda^r \\ \bar{h} &= qh \end{aligned}$$

Stabilitätsgebiet:

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

- 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 $f_n^{(k)} = f(t_n, y_n^{(k)})$

⇒ 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:

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

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

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

Umschreibe als System erster Ordnung:

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

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 Parameter
- Löse Gleichung für „das Treffen des Endwertes“

Demonstration am Beispiel 6.1:

Finde $y(t, s)$ so dass:

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

s ist gegeben durch die Bedingung:

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

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

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

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:

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

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

Beweis .

$$\begin{aligned}
 \text{erste Komponente:} \quad & \frac{\partial}{\partial s} \frac{\partial y_1}{\partial t}(t, s) = \frac{\partial}{\partial s} y_2(t, s) \\
 & \Leftrightarrow \frac{\partial}{\partial t} \underbrace{\frac{\partial y_1}{\partial s}(t, s)}_{=: v_1(t, s)} = v_2(t, s) \\
 \\
 \text{zweite Komponente:} \quad & \frac{\partial}{\partial s} \frac{\partial y_2}{\partial t}(t, s) = \frac{\partial}{\partial s} f(t, y_1(t, s), y_2(t, s)) \\
 & \frac{\partial}{\partial t} v_2(t, s) = \frac{\partial f}{\partial u}(t, y_1(t, s), y_2(t, s)) v_1(t, s) \\
 & \quad + \frac{\partial f}{\partial u'}(t, y_1(t, s), y_2(t, s)) v_2(t, s) \\
 \\
 \text{Anfangswerte:} \quad & v_1(a, s) = \frac{\partial y_1}{\partial s}(a, s) = \frac{\partial}{\partial s} u_a = 0 \\
 & v_2(a, s) = \frac{\partial y_2}{\partial s}(a, s) = \frac{\partial}{\partial s} 1 = 1
 \end{aligned}$$

□

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 $\frac{\partial y_1^{(i)}}{\partial s}$ durch Lösen der AWA (6.6) numerisch
- 5) Newton-Schritt:

$$s^{(i+1)} = s^{(i)} + \left(\frac{\partial y_1}{\partial s}(b, s^{(i)}) \right)^{-1} (u_b - y_1^{(i)}(b, s^{(i)}))$$

- 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 Verfahren in Schritt 2/4).

Problem des Schießverfahrens:

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

\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, \dots, s_M sowie interne Startwerte w_2, \dots, w_m sodass:

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

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

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

→ Löse (6.7) mit dem Newton-Verfahren.

6.2 Differenzverfahren

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

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

Nichtlineares GLS für die $N + 1$ „Unbekannten“ y_0, \dots, 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)

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

(Skizze zum Verfahren/Beispiel)

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

Eigenschaften:

- 1) Invariant gegenüber Zeitumkehr
- 2) Drehimpuls erhaltend, nicht Energie erhaltend, aber Energiefehler bleibt beschränkt \Rightarrow Langzeitstabilität
- 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 $u \in U = \{v \in C^1(I)^d : v(t_0) = u_0\}$ so dass

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

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.

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

„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

$$Q_u(t) = \int_i c(x)\varrho(x)T(x, t) dx$$

$$[c] = \frac{j}{Kkg}$$

Spez. Wärme

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

Dichte

$$[T] = K$$

Temperatur

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

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

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

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

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

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

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)

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

- Konvektiver Fluss
(hübsche Zeichnung mit Fläche von $\omega(t) \rightarrow (\omega(t + \tau))$ 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 $T(x, t_0) = T_0(x) \quad \forall x \in \Omega$
- Randbedingungen: $T(x, t) = g(x, t) \quad \forall x \in \partial\Omega, t \in [t_0, t_{\text{end}}]$

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:

$$\begin{aligned} \partial_t T &= 0 \text{ (stationär)} \\ v &= 0 \text{ (keine Konv.)} \\ \lambda &= 1 \\ -\nabla \cdot (\nabla T) &= q \quad (-\Delta T = q) \\ T(x) &= g(x) \quad \forall x \in \partial\Omega \end{aligned}$$

Grundwassergleichung:

$$\begin{aligned} \partial_t \varphi(x, t) + \nabla \cdot j_w(x, t) &= f(x, t) \quad (\forall (x, t) \in \Omega \times \Sigma) \quad (\Sigma: \text{Zeitintervall}) \\ j_w(x, t) &= -K(x) [\nabla p(x, t) - \varrho(x, t)g e_z] \end{aligned}$$

$\varphi(x, t)$: 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:

$$F\left(\frac{\partial^m u}{\partial^m x_1}, \dots, \frac{\partial^m u}{\partial^m x_n}, \frac{\partial^{m-1} u}{\partial^{m-1} x_1}, \dots, \frac{\partial u}{\partial x_n}, u\right) = 0 \quad \forall x \in \Omega$$

Zur Eindeutigkeit fehlen noch zusätzliche Bedingungen.

Eine spezielle Klasse sind die linearen PDGL. Für $n = 2$ und $m = 2$:

$$\begin{aligned} a\partial_{xx}^2 u + 2b\partial_{xy}^2 u + c\partial_{yy}^2 u & \quad \leftarrow \text{Hauptteil} \\ + d\partial_x u + e\partial_y u + fu + g & = 0 \end{aligned}$$

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 $x^T D x = 0$, wobei D die Diagonalisierte von A ist. ($\alpha, \beta \geq 0$)

$$\begin{aligned} \text{Elliptisch: } x^T D x &= \alpha x_1^2 + \beta x_2^2 = 0 \\ \text{Hyperbolisch: } x^T D x &= \alpha x_1^2 - \beta x_2^2 = 0 \\ \text{Parabolisch: } x^T D x &= \alpha x_1^2 = 0 \end{aligned}$$

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 $(A(x))_{ij} = a_{ij}(x)$, $a(x) = (a_1(x), \dots, a_n(x))^T$

- 1) Elliptisch in $x \in \Omega$, falls alle EW von A gleiches Vorzeichen, kein EW Null
- 2) Hyperbolisch in $x \in \Omega$, falls kein EW Null und $n - 1$ gleiches Vorzeichen, 1 EW entgegenges. 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“.

$$\Gamma_D \cup \Gamma_N = \partial\Omega, \text{ meas}(\Gamma_D) \neq \emptyset$$

allgemeine Diffusionsgleichung:

$$-\nabla \cdot \{K(x)\nabla u\} = f \text{ in } \Omega \quad K(x) \text{ Tensor!}$$

$$u = g \text{ auf } \Gamma_D \subseteq \partial\Omega$$

$$-(K(x)\nabla u) \cdot \nu = j \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D$$

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) $\alpha(x) := \inf \{ \xi^T K(x)\xi \mid \|\xi\| = 1 \} \geq \alpha_0 > 0$ (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:

$$\begin{array}{l} y = 0: u(x, 0) = u_0(x), \frac{\partial u}{\partial y}(x, 0) = u_1(x). \\ x = 0: u(0, y) = g_0(y) \\ x = 1: u(1, y) = g_1(y) \end{array} \left. \vphantom{\begin{array}{l} y = 0: u(x, 0) = u_0(x), \frac{\partial u}{\partial y}(x, 0) = u_1(x). \\ x = 0: u(0, y) = g_0(y) \\ x = 1: u(1, y) = g_1(y) \end{array}} \right\} \begin{array}{l} \text{„Anfangbedingungen“} \\ \text{„Randbedingungen“} \end{array}$$

Wärmeleitungsgleichung

Vorzeichen ist aus Typdefinition nicht klar!

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial y} = f \text{ in } \Omega$$

Prototyp einer parabolischen PDGL.

Beispiele für Randbedingungen:

$$\begin{array}{l} y = 0: u(x, 0) = u_0(x) \text{ „Anfangbedingung“} \\ x = 0: u(0, y) = g_0(y) \\ x = 1: u(1, y) = g_1(y) \end{array} \left. \vphantom{\begin{array}{l} y = 0: u(x, 0) = u_0(x) \text{ „Anfangbedingung“} \\ x = 0: u(0, y) = g_0(y) \\ x = 1: u(1, y) = g_1(y) \end{array}} \right\} \text{„Randbedingungen“}$$

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: $L: X \rightarrow Y$, 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

$$\nabla \cdot \{\vec{v}(x)u(x)\} = f \text{ in } \Omega \quad \vec{v}(x) \text{ gegebenes Vektorfeld } \vec{v}: \mathbb{R}^n \rightarrow \mathbb{R}^n$$

(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 $u \in C^2(\Omega) \wedge C^0(\bar{\Omega})$ welche (7.1) löst heißt „klassische Lösung“.

Polarkoordinaten:

$$\begin{aligned}x(r, \varphi) &= r \cos \varphi & y(r, \varphi) &= r \sin \varphi \\ \hat{u}(r, \varphi) &= u(x(r, \varphi), y(r, \varphi)) \text{ wobei } \Delta u = f\end{aligned}$$

Gesucht: Gleichung für $\hat{u}(r, \varphi)$

$$\frac{\partial^2 \hat{u}}{\partial r^2} + \frac{1}{r} \frac{\partial \hat{u}}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \hat{u}}{\partial \varphi^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f\tag{7.2}$$

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

$$\Omega_i = \{x \in \mathbb{R}^3 \mid \|x\| < R\}$$

$$\Omega_a = \{x \in \mathbb{R}^3 \mid \|x\| > R\}$$

$$\Omega = \mathbb{R}^3$$

in Ω_i :

$$\Delta \Phi_i = 4\pi\gamma\rho$$

in Ω_a :

$$\Delta \Phi_a = 0 \text{ in } \Omega_a$$

$$\Phi_a(x) = 0 \text{ für } \|x\| \rightarrow \infty$$

Übergangsbedingungen auf $\Gamma = \partial\Omega_i \cap \partial\Omega_a$:

$$\Omega_a(x) = \Omega_i(x)$$

$$x \in \Gamma$$

$$\nabla \Omega_a(x) \cdot \nu_i = \nabla \Omega_i(x) \cdot \nu_i$$

Gesamtproblem:

$$\begin{aligned} \nabla\Phi &= f \text{ in } \Omega & f(x) &= \begin{cases} 4\pi\gamma\varrho & \|x\| < R \\ 0 & \text{sonst} \end{cases} \\ \Phi(x) &= 0 & \|x\| &\rightarrow \infty \end{aligned}$$

⇒ unstetige rechte Seite. Wird durch klassische Lösung nicht erfasst.

8 Finite Differenzen Verfahren

Der eindimensionale Fall

Betrachte die eindimensionale RWA:

$$\begin{aligned} -u''(x) &= f(x) \quad x \in (0, 1) \\ u(0) &= \varphi_0 \quad u(1) = \varphi_1 \end{aligned} \tag{8.1}$$

Taylorreihenentwicklung:

$$\begin{aligned} u(x+h) &= u(x) + \underbrace{hu'(x)} + \frac{h^2}{2}u''(x+\delta^+h) \quad \delta^+ \in [0, 1] \\ \Leftrightarrow u'(x) &= \frac{u(x+h) - u(x)}{h} - \frac{h}{2}u''(x+\delta^+h) \end{aligned} \tag{8.2}$$

$$\begin{aligned} u(x-h) &= u(x) - \underbrace{hu'(x)} + \frac{h^2}{2}u''(x-\delta^-h) \quad \delta^- \in [0, 1] \\ \Leftrightarrow u'(x) &= \frac{u(x) - u(x-h)}{h} + \frac{h}{2}u''(x-\delta^-h) \end{aligned} \tag{8.3}$$

Man setzt:

$$\begin{aligned} (\partial^+u)(x) &:= [u(x+h) - u(x)]/h && \text{„Vorwärtsdifferenz“} \\ (\partial^-u)(x) &:= [u(x) - u(x-h)]/h && \text{„Rückwärtsdifferenz“} \end{aligned}$$

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

$$\begin{aligned}
 (\partial^+ u)(x) &= u'(x) + \frac{h}{2}u''(x) + \frac{h^2}{6}u'''(x) + \frac{h^3}{24}u^{iv}(x) + \dots \\
 (\partial^- u)(x) &= u'(x) - \frac{h}{2}u''(x) + \frac{h^2}{6}u'''(x) - \frac{h^3}{24}u^{iv}(x) + \dots \\
 (\partial^- \partial^+ u)(x) &= \partial^- \left(u'(x) + \frac{h}{2}u''(x) + \frac{h^2}{6}u'''(x) + \frac{h^3}{24}u^{iv}(x) \right) \\
 &= \left(\cancel{u''(x)} - \frac{h}{2}\cancel{u'''(x)} + \frac{h^2}{6}u^{iv}(x) - \dots \right) + \frac{h}{2} \left(\cancel{u'''(x)} - \frac{h}{2}u^{iv}(x) + \dots \right) + \frac{h^2}{6} (u^{iv}(x) - \dots) \\
 &= u''(x) + h^2 \left(\frac{1}{6} - \frac{1}{4} + \frac{1}{6} \right) u^{iv}(x) + \dots \\
 &= u''(x) + \frac{h^2}{12} u^{iv}(x) + \mathcal{O}(h^4)
 \end{aligned}$$

Formel zweiter Ordnung für die erste Ableitung:

$$\begin{aligned}
 u(x+h) - u(x-h) &= 2hu'(x) + \mathcal{O}(h^3) \\
 \Leftrightarrow u'(x) &= \frac{u(x+h) - u(x-h)}{2h} + \mathcal{O}(h^2) \quad \text{„zentraler Differenzenquotient“}
 \end{aligned}$$

Lemma 8.1 Differenzenformel

$$(\partial^+ u)(x) = \frac{u(x+h) - u(x)}{h} = u'(x) + hR \text{ mit } |R| \leq \frac{1}{2} \|u\|_{C^2(\bar{\Omega})} \text{ falls } u \in C^2(\bar{\Omega}) \quad (8.4a)$$

$$(\partial^- u)(x) = \frac{u(x) - u(x-h)}{h} = u'(x) + hR \text{ mit } |R| \leq \frac{1}{2} \|u\|_{C^2(\bar{\Omega})} \text{ falls } u \in C^2(\bar{\Omega}) \quad (8.4b)$$

$$\|u\|_{C^k(\bar{\Omega})} = \sup_{x \in \bar{\Omega}} \max_{i=0, \dots, k} |u^{(i)}(x)|$$

$$\frac{u(x+h) - u(x-h)}{2h} = u'(x) + hR \text{ mit } |R| \leq \frac{1}{6} \|u\|_{C^3(\bar{\Omega})} \quad (8.4c)$$

$$\frac{u(x-h) - 2u(x) + u(x+h)}{h^2} = u''(x) + h^2 R \text{ mit } |R| \leq \frac{1}{12} \|u\|_{C^4(\bar{\Omega})} \quad (8.4d)$$

Bemerkung:

- Für die Formeln zweiter Ordnung sind äquidistante Gitter notwendig!
- $u \in C^4(\bar{\Omega})$ ist sehr starke Forderung, die in der Regel nicht erfüllt ist!

Zurück zum RWP (8.1):

Unterteile $\Omega = (0, 1)$ in N Teilintervalle ($N \in \mathbb{N}$)

$$[x_i, x_{i+1}] \quad i = 0, \dots, N-1 \quad x_i = ih \quad h = \frac{1}{N}$$

„äquidistantes Gitter“

$$\begin{aligned}\Omega_h &= \{ih \mid i \in \mathbb{N}_0 \wedge 0 < i < N\} \\ \overline{\Omega}_h &= \{ih \mid i \in \mathbb{N}_0 \wedge 0 \leq i \leq N\}\end{aligned}$$

Für $u \in C^4(\overline{\Omega})$ gilt:

$$-\frac{1}{h^2}[u(x-h) - 2u(x) + u(x+h)] = f(x) + O(h^2) \quad \forall x \in \Omega_h \quad (8.5)$$

(analog zu lokalem Abschneidefehler)

Streichen des Fehlerterms liefert $|\Omega_h| = N - 1$ lineare Gleichungen:

$$-\frac{1}{h^2}[u_h(x-h) - 2u_h(x) + u_h(x+h)] = f(x) \quad \forall x \in \Omega_h \quad (8.6a)$$

Zusätzlich gelten die Randbedingungen

$$u_h(0) = \varphi_0 \quad u_h(1) = \varphi_1 \quad (8.6b)$$

$u_h : \overline{\Omega}_h \rightarrow \mathbb{R}$ heißt „Gitterfunktion“.

Alternativ fassen wir u_h als Vektor im \mathbb{R}^{N-1} auf:

$$u_h = (u_h(h), u_h(2h), \dots, u_h(1-h))^T$$

Elimination der Randwerte aus (8.6a) mittels (8.6b) ergibt ein LGS für u_h (den Vektor!)

$$\underbrace{\frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & \end{bmatrix}}_{L_h} \underbrace{\begin{bmatrix} u_h(h) \\ u_h(2h) \\ u_h(3h) \\ \vdots \\ u_h(1-h) \end{bmatrix}}_{u_h} = \underbrace{\begin{bmatrix} f(h) + \frac{\varphi_0}{h^2} \\ f(2h) \\ \vdots \\ f(1-2h) \\ f(1-h) + \frac{\varphi_1}{h^2} \end{bmatrix}}_{q_h}$$

- 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

$$-\Delta u(x) = - \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}(x) = f(x) \quad x \in \Omega = (0, 1)^n \quad (8.7a)$$

$$u(x) = g(x) \quad x \in \partial\Omega \quad (8.7b)$$

In unserem Beispiel:

$$\frac{1}{h^2} \begin{bmatrix} & & 1 \\ 1 & -4 & 1 \\ & & 1 \end{bmatrix} \text{ „Fünfpunktstern“}$$

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:

$$a_{xx}(x, y) \frac{\partial^2 u}{\partial x^2} + 2a_{xy}(x, y) \frac{\partial^2 u}{\partial x \partial y} + a_{yy}(x, y) \frac{\partial^2 u}{\partial y^2} + a_x(x, y) \frac{\partial u}{\partial x} + a_y(x, y) \frac{\partial u}{\partial y} + a(x, y)u = f(x, y) \quad (8.9)$$

Diskretisierung von $\frac{\partial^2 u}{\partial x \partial y}$

$$\begin{aligned} u(x + h, y + h) &= u(x, y + h) + hu_x(x, y + h) + \frac{h^2}{2} u_{xx}(x, y + h) + \frac{h^3}{6} u_{xxx}(x, y + h) + \frac{h^4}{24} u_{xxxx}(\xi_1) \\ &= u(x, y) + hu_y(x, y) + \frac{h^2}{2} u_{yy}(x, y) + \frac{h^3}{6} u_{yyy}(x, y) + \frac{h^4}{24} u_{yyyy}(\xi_2) \\ &\quad + hu_x(x, y) + h^2 u_{xy}(x, y) + \frac{h^3}{6} u_{xyy}(x, y) + \frac{h^4}{6} u_{xyyy}(\xi_3) \\ &\quad + \frac{h^2}{2} u_{xx}(x, y) + \frac{h^3}{2} u_{xxy}(x, y) + \frac{h^4}{4} u_{xxyy}(\xi_4) \\ &\quad + \frac{h^3}{6} u_{xxx}(x, y) + \frac{h^4}{6} u_{xxxxy}(\xi_5) \end{aligned}$$

Analog für $u(x - h, y - h)$: Bei allen ungeraden h -Potenzen entsteht ein Minuszeichen.
Addition liefert:

$$\begin{aligned} &u(x + h, y + h) + u(x - h, y - h) \\ &= 2u(x, y) + h^2 \left[u_{yy}(x, y) + 2u_{xy}(x, y) + u_{xx}(x, y) \right] + O(h^4) \\ &= 2u(x, y) + h^2 \left[\frac{u(x, y - h) - 2u(x, y) + u(x, y + h)}{h^2} \right. \\ &\quad \left. + 2u_{xy}(x, y) + \frac{u(x - h, y) - 2u(x, y) + u(x + h, y)}{h^2} \right] + O(h^4) \end{aligned}$$

$$\Leftrightarrow u_{xy}(x, y) = \frac{1}{2h^2} [-u(x, y+h) + u(x+h, y+h) - u(x-h, y) + 2u(x, y) - u(x+h, y) + u(x-h, y-h) - u(x, y-h)] + \mathcal{O}(h^2) = \frac{1}{2h^2} \begin{bmatrix} & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & \end{bmatrix} \quad (8.10)$$

Mit $u(x+h, y-h)$ und $u(x-h, y+h)$ folgt analog:

$$u_{xy}(x, y) = \frac{1}{2h^2} [-u(x-h, y+h) + u(x, y+h) + u(x-h, y) - 2u(x, y) + u(x+h, y) + u(x, y-h) - u(x+h, y-h)] + \mathcal{O}(h^2) = \frac{1}{2h^2} \begin{bmatrix} -1 & 1 \\ 1 & -2 & 1 \\ & 1 & -1 \end{bmatrix} \quad (8.11)$$

Welche wählt man? → hängt von den Koeffizienten a_{xx}, a_{xy}, a_{yy} ab!

(8.9) sei als *elliptisch* vorausgesetzt, d.h.

$$a_{xx}(x, y) \cdot a_{yy}(x, y) > a_{xy}^2(x, y)$$

→ a_{xx}, a_{yy} haben *gleiches* Vorzeichen, o.B.d.A. sei dies *positiv*, $a_{xx} > 0, a_{yy} > 0$. *Zusätzlich* sei

$$|a_{xy}| \leq \min \{a_{xx} \cdot a_{yy}\}$$

(aus Elliptizität folgt nur $|a_{xy}| \leq \sqrt{a_{xx} \cdot a_{yy}}$)

Dann wähle Formel wie folgt:

(a) (8.10) „rechts oben“-Formel, falls $a_{xy} \geq 0$

(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:

$$a_{xx} \frac{\partial^2 u}{\partial x^2} + 2a_{xy} \frac{\partial^2 u}{\partial x \partial y} + a_{yy} \frac{\partial^2 u}{\partial y^2} = \frac{1}{h^2} \begin{bmatrix} -a_{xy}^- & a_{yy} - |a_{xy}| & a_{xy}^+ \\ a_{xx} - |a_{xy}| & 2(|a_{xy}| - a_{xx} - a_{yy}) & a_{xx} - |a_{xy}| \\ a_{xy}^+ & a_{yy} - |a_{xy}| & -a_{xy}^- \end{bmatrix} u(x, y) + \mathcal{O}(h^2)$$

Vorzeichen

- Nicht-Diagonalelemente

- $a_{xy}^+ \leq 0, -a_{xy}^- \geq 0$

- $|a_{xy}| \leq \underbrace{\min\{a_{xx}, a_{yy}\}}_{>0} \Rightarrow \min\{a_{xx}, a_{yy}\} - |a_{xy}| \geq 0$

⇒ alle Nicht-Diagonalelemente nicht negativ.

- Diagonalelement

$$2(|a_{xy}| - a_{xx} - a_{yy}) = 2(\underbrace{|a_{xy}|}_{\leq 0} - \underbrace{\min\{a_{xx}, a_{yy}\}}_{>0} - \underbrace{\max\{a_{xx}, a_{yy}\}}_{>0}) < 0$$

- Betragssumme der Nicht-Diagonalelemente

$$\begin{aligned} & 2|a_{xx} - |a_{xy}|| + 2|a_{yy} - |a_{xy}|| + 2|a_{xy}| \\ &= 2|a_{xx} - |a_{xy}|| + 2|a_{yy}| \end{aligned}$$

Für das Diagonalelement gilt:

$$\begin{aligned} & 2||a_{xy} - a_{xx} - a_{yy}|| = 2\left| \underbrace{a_{yy}}_{>0} + \underbrace{a_{xx} - |a_{xy}|}_{\geq 0} \right| \\ &= 2|a_{yy}| + 2|a_{xx} - |a_{xy}|| \end{aligned}$$

⇒ 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) $\sum_{j \neq i} |a_{ij}| = |a_{ii}| \quad \forall i$

Dann gilt diskretes Maximumsprinzip:

Zeile i :

$$a_{ii}x_i + \sum_{j \neq i} a_{ij}x_j = 0$$

$$\Leftrightarrow x_i = \sum_{j \neq i} \underbrace{-\frac{a_{ij}}{a_{ii}}}_{=: \alpha_{ij} \geq 0} x_j = \sum_{j \neq i} \alpha_{ij} x_j \quad \alpha_{ij} \geq 0$$

$$\sum_{j \neq i} \alpha_{ij} = \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} = \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| = \frac{|a_{ii}|}{|a_{ii}|} = 1$$

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

$$x_i = \sum_{j \neq i} \alpha_{ij} x_j \geq \min_{k \neq i} x_k$$

⇒ Das Maximum/Minimum wird auf dem Rand angenommen.

Gleichungen in Erhaltungssform

$$-\nabla \cdot \{k(x, y) \nabla u\} = f \text{ in } \Omega$$

$$u = g \text{ auf } \partial\Omega$$

Hat nicht die Form (8.9) → Produktregel anwenden. Setzt Glattheit von $k(x, y)$ voraus. (Zeichnung)

Idee

Hier für $n = 1$:

$$q = -k(x) \frac{\partial u}{\partial x} \qquad \frac{\partial q}{\partial x} = f$$

$$(q^+ u)_x = k(x + h/2) \frac{u(x + h) - u(x)}{h} \quad (\text{Zahlenstrahlzeichnung})$$

$$(\partial^- q^+ u)(x) = \frac{1}{h} \left[\underbrace{k(x + h/2) \frac{u(x + h) - u(x)}{h}}_{(q^+ u)(x)} - \underbrace{k(x - h/2) \frac{u(x) - u(x - h)}{h}}_{(q^+ u)(x-h)} \right]$$

$$= \frac{\partial}{\partial x} \left(k(x) \frac{\partial u}{\partial x} \right) (x) + \mathcal{O}(h^2)$$

Direkte Diskretisierung:

$$= \frac{1}{h^2} \left[k\left(x + \frac{h}{2}\right) u(x + h) - \left(k\left(x + \frac{h}{2}\right) + k\left(x - \frac{h}{2}\right) \right) u(x) + k\left(x - \frac{h}{2}\right) u(x - h) \right]$$

$$- \frac{1}{2} \left(k\left(x + \frac{h}{2}\right) + k\left(x - \frac{h}{2}\right) \right) \frac{u(x + h) - 2u(x) + u(x - h)}{h^2}$$

$$+ \frac{1}{2} \left(k\left(x + \frac{h}{2}\right) + k\left(x - \frac{h}{2}\right) \right) \frac{u(x + h) - 2u(x) + u(x - h)}{h^2}$$

$$= \frac{2k\left(x + \frac{h}{2}\right) u(x + h) + 2k\left(x - \frac{h}{2}\right) u(x - h) - \left(k\left(x + \frac{h}{2}\right) + k\left(x - \frac{h}{2}\right) \right) (u(x + h) + u(x - h))}{2h^2}$$

$$\underbrace{\frac{k\left(x + \frac{h}{2}\right) - k\left(x - \frac{h}{2}\right)}{h} \cdot \frac{(u(x+h) - u(x-h))}{2h} = (k'(x) + \mathcal{O}(h^2)) \cdot (u'(x) + \mathcal{O}(h^2))}_{\text{Taylor expansion}}}$$

$$+ \frac{1}{2} \left(k\left(x + \frac{h}{2}\right) + k\left(x - \frac{h}{2}\right) \right) \frac{u(x + h) - 2u(x) + u(x - h)}{h^2}$$

$$\underbrace{k(x) + \mathcal{O}(h^2)}_{\text{Taylor expansion}} \underbrace{\frac{\partial^2 u}{\partial x^2} + \mathcal{O}(h^2)}_{\text{Taylor expansion}}$$

$$k(x) \frac{\partial^2 u}{\partial x^2} + \mathcal{O}(h^2)$$

⇒ Konsistenzordnung 2 (für k glatt)

Vorteil der diskreten Formel: ist auch für nicht glattes $k(x)$ sinnvoll: (Grafik)

$$\begin{aligned}
& \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \frac{\partial}{\partial x} \left(k(x) \frac{\partial u}{\partial x} \right) dx = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(x) dx \\
& \Leftrightarrow \left(k(x) \frac{\partial u}{\partial x} \right) \Big|_{x+\frac{h}{2}} - \left(k(x) \frac{\partial u}{\partial x} \right) \Big|_{x-\frac{h}{2}} = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(x) dx \\
& \Leftrightarrow k \left(x + \frac{h}{2} \right) \left(\frac{u(x+h) - u(x)}{h} + O(h^2) \right) - k \left(x - \frac{h}{2} \right) \left(\frac{u(x) - u(x-h)}{h} + O(h^2) \right) = f(x)h + O(h^3) \\
& \Leftrightarrow \frac{1}{h^2} \left[k \left(x + \frac{h}{2} \right) u(x+h) - \left(k \left(x + \frac{h}{2} \right) + k \left(x - \frac{h}{2} \right) \right) u(x) + k \left(x - \frac{h}{2} \right) u(x-h) \right] = f(x) + O(h)
\end{aligned}$$

(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 $u \notin C^4(\bar{\Omega})$?

9 Konvergenz des Finite Differenzen Verfahrens

9.1 M-Matrizen

Bezeichnungen:

$A \in \mathbb{R}^{n \times n}$ mit Elementen $a_{\alpha\beta}$ mit $\alpha, \beta \in I = \{1, \dots, n\}$

I heißt „Indexmenge“

$A \geq B$, falls $a_{\alpha\beta} \geq b_{\alpha\beta} \quad \forall \alpha, \beta \in I$. ($A > B, A < B, A \leq B$ 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) $a_{\alpha\beta} \leq 0 \quad \forall \alpha, \beta \in I, \alpha \neq \beta$
- (iii) A nicht singular 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 $\alpha = \alpha_0, \alpha_1, \dots, \alpha_k = \beta$ existiert mit $(\alpha_{i-1}, \alpha_i) \in E \quad \forall i = 1, \dots, k$

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* Permutation 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

$$B(z, r) = \{\xi \in \mathbb{C} \mid |z - \xi| < r\} \quad \text{und} \quad \overline{B(z, r)} = \{\xi \in \mathbb{C} \mid |z - \xi| \leq r\}$$

- (a) Alle Eigenwerte von A liegen in

$$\bigcup_{\alpha \in I} \overline{B(a_{\alpha\alpha}, r_\alpha)} \quad \text{mit} \quad r_\alpha = \sum_{\substack{\beta \neq \alpha \\ \beta \in I}} |a_{\alpha\beta}|$$

- (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

$$\sum_{\beta \in I} a_{\gamma\beta} u_\beta = \lambda u_\gamma \Leftrightarrow (\lambda - a_{\gamma\gamma}) u_\gamma = \sum_{\beta \neq \gamma} a_{\gamma\beta} u_\beta$$

$$|\lambda - a_{\gamma\gamma}| |u_\gamma| = \left| \sum_{\beta \neq \gamma} a_{\gamma\beta} u_\beta \right| \leq \sum_{\beta \neq \gamma} |a_{\gamma\beta}| \underbrace{|u_\beta|}_{\leq 1} \leq \sum_{\beta \neq \gamma} |a_{\gamma\beta}| =: r_\gamma \quad (9.1)$$

$$\lambda \in \overline{B(a_{\gamma\gamma}, r_\gamma)} \subseteq \bigcup_{\alpha \in I} \overline{B(a_{\alpha\alpha}, r_\alpha)} \quad \text{für jedes } \lambda \in \sigma(A)$$

(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):

$$|\lambda - a_{\gamma\gamma}| = \sum_{\beta \neq \gamma} |a_{\gamma\beta}| |u_\beta| = \sum_{\beta \neq \gamma} |a_{\gamma\beta}| = r_\gamma$$

$\Rightarrow |u_\beta| = 1$ für alle $\beta \in I$ mit $a_{\gamma\beta} \neq 0$

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:

$$|u_\alpha| = 1, \quad |\lambda - a_{\alpha\alpha}| = r_\alpha \quad \forall \alpha \in I \Leftrightarrow \lambda \in \bigcap_{\alpha \in I} \partial B(a_{\alpha\alpha}, r_\alpha)$$

□

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 \rightarrow 9.4(b) anwendbar:

innere	Randknoten \neq Ecke	Eckknoten
$\frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$	$\frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & 0 & \end{bmatrix}$	$\frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & 0 \\ & 0 & \end{bmatrix}$

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

$$(i) \quad \sum_{\substack{\beta \neq \alpha \\ \beta \in I}} |a_{\alpha\beta}| < |a_{\alpha\alpha}| \quad \forall \alpha \in I$$

und irreduzibel diagonaldominant, falls

(ii) A irreduzibel

$$(iii) \quad \sum_{\substack{\beta \neq \alpha \\ \beta \in I}} |a_{\alpha\beta}| \leq |a_{\alpha\alpha}| \quad \forall \alpha \in I$$

(iv) Bedingung (i) für mindestens ein $\alpha \in I_0$ gilt.

Bemerkung: L_h aus Fünfpunktstern ist irreduzibel diagonaldominant.

Satz 9.7

Es sei A diagonaldominant oder irreduzibel diagonaldominant und $A = D - B$, wobei $D = \text{diag}(A)$. Dann ist

$$\rho(D^{-1}B) < 1$$

Beweis.

$$C := D^{-1}B$$

$$c_{\alpha\beta} = \begin{cases} 0 & \text{falls } \alpha = \beta \\ -\frac{a_{\alpha\beta}}{a_{\alpha\alpha}} & \text{sonst} \end{cases}$$

(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: $r_\beta \leq 1 \quad \forall \beta \in I$ und

$r_\alpha < 1$ für mindestens ein α

$$\lambda \in \sigma(A): \lambda \in \left[\bigcup_{\beta \in I} B(0, r_\beta) \right] \cup \underbrace{\left[\bigcap_{\beta \in I} \partial B(0, r_\beta) \right]}_{=:\Gamma}$$

Fallunterscheidung:

(i) Alle r_β gleich: $r_\beta = r_\alpha < 1 \Rightarrow \Gamma \subset B(0, 1)$

(ii) nicht alle r_β gleich $\Rightarrow \Gamma = \emptyset$

□

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.

„ \Leftarrow “: 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 $C^n \rightarrow 0$ wegen $\varrho(C) < 1 \Rightarrow S := \sum_{\nu=0}^{\infty} C^\nu$. Da $SD^{-1} = A^{-1}$ und da $S \geq 0 \Rightarrow A^{-1} \geq 0$.

„ \Rightarrow “: Sei $D^{-1}Bu = \lambda u$ für ein $u \in \mathbb{R}^n$ mit $u \neq 0$. Setze

$$|u| = (|u_1|, \dots, |u_n|)^T \quad |\lambda||u| = |\lambda u| = |D^{-1}Bu| \leq D^{-1}B|u|$$

$$\begin{aligned} |u| &= A^{-1}A|u| = A^{-1}(D - B)|u| = A^{-1}D(I - D^{-1}B)|u| \\ &= A^{-1}D|u| - A^{-1}DD^{-1}B|u| \leq A^{-1}D|u| - A^{-1}D|\lambda||u| \\ &= (1 - |\lambda|)A^{-1}D|u| \end{aligned}$$

Angenommen $|\lambda| \geq 1 \Rightarrow |u| \leq 0 \nabla$

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. $\stackrel{9.7}{\Rightarrow} \varrho(D^{-1}B) < 1 \stackrel{9.8}{\Rightarrow} A$ ist M -Matrix.

Es gilt auch: Ist A M -Matrix, irreduzibel, dann ist $A^{-1} > 0$.

□

Definition 9.10 Restriktion

Sei $U_L \{f_h \mid f_h: \bar{\Omega}_h \rightarrow \mathbb{R}\}$ der VR der Gitterfunktionen, dann ist die Restriktion

$$R_h: C^0(\bar{\Omega}) \rightarrow U_h$$

definiert durch

$$(R_h u)(x) := u(x) \quad \forall x \in \bar{\Omega}_h$$

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:

$$|u(x) - u_h(x)| \rightarrow 0 \quad \forall x \in \Omega_h \text{ wenn } h \rightarrow 0$$

und wir nennen das Verfahren konvergent.

Beweis.

Sei $e_h := R_h u - u_h = L_h^{-1} \eta_h$ der Fehler, dann gilt:

$$\|e_h\|_\infty \|L_h^{-1} \eta_h\|_\infty \leq \|L_h^{-1}\|_\infty \|\eta_h\|_\infty \leq \kappa Ch^p \rightarrow 0 \text{ für } h \rightarrow 0$$

□

Bemerkung: Für 2D-Fall (Fünfpunktstern) haben wir bereits Konsistenz 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.

$$\forall u \in \mathbb{R}^n: |u| \leq \|u\|_\infty \cdot \mathbb{1} \leq \|u\|_\infty Aw$$

wegen $A^{-1} \geq 0$ gilt

$$|A^{-1}u| \leq A^{-1}|u| \leq A^{-1}\|u\|_\infty Aw = \|u\|_\infty w$$

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 Fünfpunktstern. 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

$$Ax = b \quad A \in \mathbb{R}^{N \times N} \quad x, b \in \mathbb{R}^N$$

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:

$$A(N) = \sum_{i=1}^N n \cdot 2n = \sum_{i=1}^N 2n^2 = 2Nn^2 = 2NN^{2\alpha} = 2N^{2\alpha+1}$$

$$d = 1 \quad \alpha = 0 \quad \Rightarrow O(N)$$

$$d = 2 \quad \alpha = \frac{1}{2} \quad \Rightarrow O(N^2)$$

$$d = 3 \quad \alpha = \frac{1}{3} \quad \Rightarrow O(N^{\frac{7}{3}})$$

10.1 Relaxationsverfahren

Idee: Sukzessives Auflösen

$$\text{i-te Gl.: } \sum_{j=1}^N a_{ij}x_j = b_i \Leftrightarrow_{a_{ii} \neq 0} \boxed{x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}x_j \right)}$$

Konstruierte Folge $x^{(k)} \rightarrow x$

Jacobi-Verfahren (Gesamtschrittverfahren)

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}x_j^{(k)} \right) \quad i = 1, \dots, N$$

gedämpftes Jacobi-Verfahren:

$$x_i^{(k+1)} = (1 - \omega) \cdot x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right) \quad \begin{array}{l} i = 1, \dots, N \\ \omega \in (0, 1] \end{array}$$

Gauß-Seidel (Einzelschritt) Verfahren:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j < i} a_{ij} x_j^{(k+1)} - \sum_{j > i} a_{ij} x_j^{(k)} \right) \quad i = 1, \dots, N$$

gedämpftes Gauß-Seidel Verfahren:

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j < i} a_{ij} x_j^{(k+1)} - \sum_{j > i} a_{ij} x_j^{(k)} \right) \quad i = 1, \dots, N$$

$$x_i^{(k+1)} = (1 - \omega) x_i^{(k)} + \omega \tilde{x}_i^{(k+1)} \quad \begin{array}{l} i = 1, \dots, N \\ \omega \in (0, 1] \end{array}$$

SOR (successive over relaxation)-Verfahren

$$x_i^{(k+1)} = (1 - \omega) x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j < i} a_{ij} x_j^{(k+1)} - \sum_{j > i} a_{ij} x_j^{(k)} \right) \quad \begin{array}{l} i = 1, \dots, N \\ \omega \in (0, 2) \end{array}$$

gedämpfte Richardson-Iteration

$$x_i^{(k+1)} = (1 - \omega) x_i^{(k)} + \omega \left(b - \sum_{j \neq i} a_{ij} x_j^{(k)} \right) \quad i = 1, \dots, N$$

Matrixschreibweise der Relaxationsverfahren

$$A = \underbrace{\quad}_{L} + \underbrace{\quad}_{D} + \underbrace{\quad}_{U}$$

$$l_{ij} = \begin{cases} a_{ij} & i > j \\ 0 & \text{sonst} \end{cases} \quad d_{ij} = \begin{cases} a_{ij} & i = j \\ 0 & \text{sonst} \end{cases} \quad u_{ij} = \begin{cases} a_{ij} & i < j \\ 0 & \text{sonst} \end{cases}$$

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

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} \right) \quad i = 1, \dots, N$$

$$\Leftrightarrow \omega \sum_{j<i} a_{ij}x_j^{(k+1)} + a_{ii}x_i^{(k+1)} = a_{ii}(1 - \omega)x_i^{(k)} + \omega \left(b_i - \sum_{j>i} a_{ij}x_j^{(k)} \right) \quad i = 1, \dots, N$$

$$\Leftrightarrow \omega Lx^{(k+1)} + Dx^{(k+1)} = (1 - \omega)Dx^{(k)} + \omega(b - Ux^{(k)})$$

$$x^{(k+1)} = (1 - \omega)(\omega L + D)^{-1}Dx^{(k)} + \omega(\omega L + D)^{-1}(b - Ux^{(k)}) + \underbrace{\omega(\omega L + D)^{-1}(L + D)x^{(k)} - \omega(\omega L + D)^{-1}(L + D)x^{(k)}}_{=0}$$

$$= (\omega L + D)^{-1}[(1 - \omega)D + \omega(L + D)]x^{(k)} +$$

$$\underbrace{\omega(\omega L + D)^{-1}}_{(L + \frac{1}{\omega}D)^{-1}}(b - \underbrace{(L + D + U)}_{=A})x^{(k)}$$

$$x^{(k+1)} = x^{(k)} + (L + \frac{1}{\omega}D)^{-1}(b - Ax^{(k)})$$

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 $x^{(k+1)} = x^{(k)} + v^{(k)} = x^{(k)} + M^{-1}(b - Ax^{(k)})$

$$M = D \quad \text{Jacobi}$$

$$M = L + D \quad \text{Gauß-Seidel}$$

$$M = L + \frac{1}{\omega}D \quad \text{SOR-Verfahren}$$

$$M = \frac{1}{\omega}(L + D) \quad \text{ged. Gauß-Seidel}$$

$$M = \frac{1}{\omega}I \quad \text{ged. Richardson-Iteration}$$

Konvergenz linearer Iterationsverfahren

Fehlerfortpflanzung:

$$\begin{aligned} e^{(k+1)} &= x - x^{(k+1)} = \underbrace{x - x^{(k)}}_{e^{(k)}} - M^{-1}(b - Ax^{(k)}) \\ &= e^{(k)} - M^{-1}Ae^{(k)} \\ &= \underbrace{(I - M^{-1}A)}_{=:S \text{ „Iterationsmatrix“}} e^{(k)} \\ &=:S \text{ „Iterationsmatrix“} \end{aligned}$$

$$e^{(k)} = S^k e^{(0)}$$

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.

„ \Leftarrow “ 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 $\|\cdot\|_\varepsilon$ mit $\|S\|_\varepsilon \leq \varrho(S) + \varepsilon < 1$ und damit

$$\lim_{k \rightarrow \infty} \|S\|_\varepsilon^k \rightarrow 0 \text{ also } \|e^{(k)}\|_\varepsilon \rightarrow 0$$

„ \Rightarrow “ 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 $\Leftrightarrow x^T A x > 0 \forall x \neq 0$) und $\omega < \frac{2}{\lambda_{\max}(A)}$. Dann konvergiert die Richardson-Iteration.

Beweis.

$$\lambda_i \in \sigma(A) \text{ mit } 0 < \lambda_{\min}(A) = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N = \lambda_{\max}(A)$$

$$S_{\text{Rich}} = I - \omega A \Rightarrow \sigma(S_{\text{Rich}}) = \{\mu_i \mid \mu_i = 1 - \omega \lambda_i; \lambda_i \in \sigma(A)\}$$

(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: ($\omega = \frac{1}{\lambda_{\max}}$):

$$\varrho(S_{\text{Rich}}) = 1 - \frac{\lambda_{\min}}{\lambda_{\max}} = 1 - \frac{1}{\kappa(A)} \quad \kappa(A) = \frac{\lambda_{\min}}{\lambda_{\max}}$$

$$k(\varepsilon) = \frac{\log \varepsilon}{\log\left(1 - \frac{1}{\kappa}\right)} \approx \frac{\log \varepsilon}{-\frac{1}{\kappa}} = \kappa(A) |\log \varepsilon|$$

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: $O(N)$.

⇒ Gesamtaufwand (zur Reduktion des Fehlers um Faktor ε): $O(|\log \varepsilon| N^{1+\frac{2}{d}})$

	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):

$$\sup_{x \neq 0} \frac{\langle x, Cx \rangle}{\langle x, x \rangle} = \lambda_{\max}(C) \quad \inf_{x \neq 0} \frac{\langle x, Cx \rangle}{\langle x, x \rangle} = \lambda_{\min}(C)$$

$S_{\text{Jac}} = I - D^{-1}A$ ist nicht symmetrisch, aber mit „ $D^{\frac{1}{2}}$ “ = $\text{diag}(\sqrt{a_{ii}})$ 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 λ_{\min}

$$\begin{aligned}\lambda_{\min}(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}) &= \inf_{x \neq 0} \frac{\langle x, (I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}})x \rangle}{\langle x, x \rangle} \\ &= 1 - \sup_{x \neq 0} \frac{\langle x, D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x \rangle}{\langle x, x \rangle} \\ &= 1 - \sup_{y \neq 0} \underbrace{\frac{\langle y, Ay \rangle}{\langle y, Dy \rangle}}_{< 2 \text{ nach Vor.}} > -1\end{aligned}$$

$$\Rightarrow \varrho(S_{\text{Jac}}) < 1 \quad \square$$

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:

$$(Sx)_i = (-D^{-1}(U + LS)x)_i = -\frac{1}{a_{ii}} \left(\sum_{j>i} a_{ij}x_j + \sum_{j<i} a_{ij}(Sx)_j \right)$$

Betrag bilden liefert

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j>i} |a_{ij}||x_j| + \sum_{j<i} |a_{ij}||x_j| \right)$$

(ii) diagonalominanter Fall, d.h. $\sum_{j \neq i} |a_{ij}| < |a_{ii}| \quad \forall i$. Per Induktion zeige: $|(Sx)_i| < \|x\|_{\infty}$:

$i = 1$:

$$\begin{aligned}|(Sx)_1| &\leq \frac{1}{|a_{11}|} \left(\sum_{j>1} |a_{1j}||x_j| \right) \\ &\leq \|x\|_{\infty} \underbrace{\frac{1}{|a_{11}|} \sum_{j>1} |a_{1j}|}_{< 1} < \|x\|_{\infty}\end{aligned}$$

$i - 1 \rightarrow i$, d.h. $|(Sx)_j| < \|x\|_\infty \quad \forall j < i$

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j>i} |a_{ij}| \underbrace{|x_j|}_{\leq \|x\|_\infty} + \sum_{j<i} |a_{ij}| \underbrace{|(Sx)_j|}_{< \|x\|_\infty} \right) \leq \|x\|_\infty \underbrace{\frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}|}_{< 1} < \|x\|_\infty$$

Damit ist $\|S\|_\infty = \sup_{x \neq 0} \frac{\|Sx\|_\infty}{\|x\|_\infty} < \frac{\|x\|_\infty}{\|x\|_\infty} < 1$

(iii) irreduzibel diagonaldom. Fall: jetzt gilt $\sum_{j \neq i} |a_{ij}| \leq |a_{ii}|$, 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

$$\begin{aligned} 1 = |(Sx)_i| &\leq \frac{1}{|a_{ii}|} \left(\sum_{j>i} |a_{ij}| |x_j| + \sum_{j<i} |a_{ij}| |(Sx)_j| \right) \leq \|x\|_\infty = 1 \\ \Rightarrow \sum_{j>i} |a_{ij}| |x_j| + \sum_{j<i} |a_{ij}| |(Sx)_j| &= |a_{ii}| \end{aligned}$$

Dies impliziert:

$$\begin{aligned} \sum_{j \neq i} |a_{ij}| &= |a_{ii}| \\ |x_j| = 1, j > i \text{ und } |(Sx)_j| = 1, j < i &\text{ für } a_{ij} \neq 0 \end{aligned}$$

\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 $j = 1, \dots, N$ 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

$$F := \frac{1}{2} x^T A x - b^T x = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle$$

sein eindeutiges Minimum in $x^* = A^{-1}b$ an. D.h. die Lösung des Minimierungsproblems $\min_{x \in \mathbb{R}^N} F(x)$ stimmt mit der Lösung des Gleichungssystems $Ax = b$ überein.

Beweis.

$x \in \mathbb{R}^N$. Wähle v so, dass $x = x^* + v$

$$\begin{aligned} F(x) &= \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle = \frac{1}{2} \langle A(x^* + v), x^* + v \rangle - \langle b, x^* + v \rangle \\ &= \frac{1}{2} [\langle Ax^*, x^* \rangle + 2\langle Ax^*, v \rangle + \langle Av, v \rangle] - \langle b, x^* \rangle - \langle b, v \rangle \\ &= \underbrace{\frac{1}{2} \langle Ax^*, x^* \rangle - \langle b, x^* \rangle}_{=F(x^*)} + \underbrace{\langle Ax^* - b, v \rangle}_{=0 \text{ für } x^*=A^{-1}b} + \frac{1}{2} \underbrace{\langle Av, v \rangle}_{>0 \text{ für } v \neq 0} \end{aligned}$$

$\Rightarrow F(x) > F(x^*)$ für alle $x \neq x^*$

Eindeutigkeit: Ang. \exists zwei Minima, $x' \neq x$

$\rightarrow x' = x^* + v$ d.h. $v \neq 0$

$\Rightarrow F(x') = F(x^*) + \frac{1}{2} \underbrace{\langle Av, v \rangle}_{>0}$ Widerspruch zu x' Minimum □

Eindimensionale Minimierung:

Idee: $x^{(k)}$, Suchrichtung $p^{(k)} \neq 0$ finde $\alpha^{(k)} \in \mathbb{R}$ so dass

$$F(x^{(k)} + \alpha^{(k)} p^{(k)}) \rightarrow \min$$

$$\begin{aligned} f(\alpha^{(k)}) &= F(x^{(k)} + \alpha^{(k)} p^{(k)}) = F(x^{(k)}) + \alpha^{(k)} \langle Ax^{(k)} - b, p^{(k)} \rangle + \frac{1}{2} (\alpha^{(k)})^2 \langle Ap^{(k)}, p^{(k)} \rangle \\ \frac{df}{d\alpha^{(k)}} &= \langle Ax^{(k)} - b, p^{(k)} \rangle + \alpha^{(k)} \langle Ap^{(k)}, p^{(k)} \rangle \stackrel{!}{=} 0 \Rightarrow \alpha^{(k)} = \frac{\overbrace{\langle b - Ax^{(k)}, p^{(k)} \rangle}^{\text{Defekt}}}{\langle Ap^{(k)}, p^{(k)} \rangle} \end{aligned}$$

Gradientenverfahren: steilster Abstieg

$N = 2$: $F: \mathbb{R}^2 \rightarrow \mathbb{R}$ (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: $\Delta, x = x^{(0)}, b, \varepsilon$

$d = b - Ax$

$\text{norm0} = \|d\|$

$\text{norm} = \text{norm0}$

while ($\text{norm} > \varepsilon \text{norm0}$) **do**

$q = Ad$

$\alpha = \frac{\langle d, d \rangle}{\langle q, d \rangle}$

$x = x + \alpha d$

$d = d - \alpha q$

$\text{norm} = \|d\|$

end while

Satz 10.8 Konvergenz des Gradientenverfahrens

Sei $A = A^T > 0$ und $Ax = b$. Definiere das Fehlerfunktional

$$E(y) := \|x - y\|_A^2 = \langle A(x - y), x - y \rangle \quad \forall y \in \mathbb{R}^N$$

$(\langle x, y \rangle)_A = \langle Ax, y \rangle$ heißt Energieskalarprodukt und $\|x\|_A = \sqrt{\langle Ax, x \rangle}$ Energienorm).

Dann gilt:

$$E(x^{(t+1)}) \leq \left(1 - \frac{1}{\kappa(A)}\right) E(x^{(t)})$$

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

$$\begin{aligned} \frac{E(x^{(t)}) - E(x^{(t+1)})}{E(x^{(t)})} &= \dots \geq \frac{\lambda_{\min}}{\lambda_{\max}} = \frac{1}{\kappa(A)} \\ \Leftrightarrow E(x^{(t+1)}) &\leq \left(1 - \frac{1}{\kappa(A)}\right) E(x^{(t)}) \end{aligned}$$

□

Konjugierte Gradienten VerfahrenGradientenverfahren $N = 2$: (Zeichnung)

Idee: Minimierung im Unterraum

$$B_t := \text{span} \{p^{(0)}, p^{(1)}, \dots, p^{(t-1)}\} \quad p^{(i)} \text{ linear unabhängig}$$

Gegeben: $x^{(0)}$ bestimme

$$x^{(t)} = x^{(0)} + \sum_{j=0}^{t-1} \alpha_j^{(t-1)} p^{(j)} \in x^{(0)} + B_t$$

sodass

$$F(x^{(t)}) = \min_{y \in x^{(0)} + B_t} F(y)$$

Also:

$$\begin{aligned} F(x^{(t)}) &= \frac{1}{2} \langle Ax^{(t)}, x^{(t)} \rangle - \langle b, x^{(t)} \rangle \\ &= \frac{1}{2} \left\langle A \left(x^{(0)} + \sum_{j=0}^{t-1} \alpha_j^{(t-1)} p^{(j)} \right), x^{(0)} + \sum_{k=0}^{t-1} \alpha_k^{(t-1)} p^{(k)} \right\rangle - \left\langle b, x^{(0)} + \sum_{k=0}^{t-1} \alpha_k^{(t-1)} p^{(k)} \right\rangle \\ &= \frac{1}{2} \langle Ax^{(0)} - b, x^{(0)} \rangle + \sum_{k=0}^{t-1} \alpha_k^{(t-1)} \langle Ax^{(0)} - b, p^{(k)} \rangle + \sum_{j=0}^{t-1} \frac{(\alpha_j^{(t-1)})^2}{2} \langle Ap^{(j)}, p^{(j)} \rangle \\ &\quad + \sum_{j=0}^{t-1} \sum_{k \neq j} \frac{\alpha_j^{(t-1)} \alpha_k^{(t-1)}}{2} \langle Ap^{(j)}, p^{(k)} \rangle \end{aligned}$$

$$\begin{aligned} \frac{\partial F}{\partial \alpha_i^{(t-1)}}(x^{(t)}) &= \langle Ax^{(0)} - b, p^{(i)} \rangle + \alpha_i^{(t-1)} \langle Ap^{(i)}, p^{(i)} \rangle + \sum_{k \neq i} \frac{\alpha_k^{(t-1)}}{2} \langle Ap^{(i)}, p^{(k)} \rangle + \sum_{j \neq i} \frac{\alpha_j^{(t-1)}}{2} \langle Ap^{(j)}, p^{(i)} \rangle \\ &= \langle Ax^{(0)} - b, p^{(i)} \rangle + \sum_{k=0}^{t-1} \alpha_k^{(t-1)} \underbrace{\langle Ap^{(k)}, p^{(i)} \rangle}_{c_{ik}} \stackrel{!}{=} 0 \quad i = 0, \dots, t-1 \end{aligned} \quad (10.1)$$

 \Rightarrow LGS für die Koeffizienten α_k :

$$C\alpha^{(t+1)} = r \quad c_{ik} = \langle Ap^{(k)}, p^{(i)} \rangle \quad r_i = \langle b - Ax^{(0)}, p^{(i)} \rangle$$

 C ist s.p.d.

$$\langle Cy, y \rangle = \sum_{i=0}^{t-1} \sum_{k=0}^{t-1} y_i y_k c_{ik} = \sum_i \sum_k y_i y_k \langle Ap^{(k)}, p^{(i)} \rangle = \left\langle A \overbrace{\left(\sum_{k=0}^{t-1} y_k p^{(k)} \right)}^{=: y'}, \sum_{j=0}^{t-1} y_j p^{(j)} \right\rangle = \langle Ay', y' \rangle > 0$$

(10.1) können wir schreiben als:

$$\langle Ax^{(t)} - b, p^{(i)} \rangle = 0 \quad \forall i = 0, \dots, t-1 \quad (10.2)$$

ment man „Galerkin-Gleichungen“

Aufwand steigt mit der Anzahl der Suchrichtungen stark an.

Idee: Wähle A -orthogonale Suchrichtungen: („konjugierte“ Suchrichtungen)

$$\langle Ap^{(j)}, p^{(i)} \rangle = 0 \quad \forall 0 \leq i < j \leq t-1$$

Dann gilt:

$$\begin{aligned} & \langle Ax^{(0)} - b, p^{(i)} \rangle + \sum_{k=0}^{t-1} \alpha_k^{(t-1)} \langle Ap^{(k)}, p^{(i)} \rangle \\ &= \langle Ax^{(0)} - b, p^{(i)} \rangle + \alpha_i^{(t-1)} \langle Ap^{(i)}, p^{(i)} \rangle = 0 \quad i = 0, \dots, t-1 \end{aligned}$$

also

$$\alpha_i := \alpha_i^{(t-1)} = \frac{\langle b - Ax^{(0)}, p^{(i)} \rangle}{\langle Ap^{(i)}, p^{(i)} \rangle} \quad i = 0, \dots, t-1 \quad (10.3)$$

Orthogonalisierung der Suchrichtungen mittels Gram-Schmidt:

Suchrichtungen $p^{(k)}$ werden sukzessive aufgebaut:

$$\begin{aligned} p^{(0)} &= -\nabla F(x^{(0)}) = b - Ax^{(0)} =: d^{(0)} \\ p^{(t)} &= -\nabla F(x^{(t)}) + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)} \text{ sodass } \langle Ap^{(t)}, p^{(i)} \rangle = 0 \quad \forall i < t \end{aligned} \quad (10.4)$$

Zwischenresultat:

$$\text{span}\{p^{(0)}, \dots, p^{(t)}\} = \text{span}\{d^{(0)}, Ad^{(0)}, \dots, A^t d^{(0)}\} =: K_t(d^{(0)}, A) \quad \text{„Krylovraum“}$$

Induktion über t :

$t = 0$:

$$p^{(0)} = d^{(0)} \quad \text{span}\{p^{(0)}\} = \text{span}\{d^{(0)}\} \quad \checkmark$$

$t-1 \rightarrow t$, d.h.

$$\begin{aligned} \text{span}\{p^{(0)}, \dots, p^{(t-1)}\} &= \text{span}\{d^{(0)}, \dots, A^{t-1} d^{(0)}\} = K_{t-1}(d^{(0)}, A) \\ p^{(t)} &= b - A \left(\underbrace{x^{(0)} + \sum_{i=0}^{t-1} \alpha_i p^{(i)}}_{x^{(t)}} \right) + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)} \\ &= \underbrace{b - Ax^{(0)}}_{d^{(0)}} - \underbrace{\sum_{i=0}^{t-2} \alpha_i Ap^{(i)} + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)}}_{\in K_{t-1}(d^{(0)}, A)} - \underbrace{\alpha_{t-1} Ap^{(t-1)}}_{\in K_t(d^{(0)}, A)} \end{aligned}$$

Bestimmung der $\beta_j^{(t-1)}$, $i < t$:

$$\begin{aligned} 0 &\stackrel{!}{=} \langle Ap^{(t)}, p^{(i)} \rangle = \langle b - Ax^{(t)} + \sum_{j=0}^{t-1} \beta_j^{(t-1)} p^{(j)}, Ap^{(i)} \rangle \\ &= \langle b - Ax^{(t)}, Ap^{(i)} \rangle + \sum_{j=0}^{t-1} \beta_j^{(t-1)} \underbrace{\langle p^{(j)}, Ap^{(i)} \rangle}_{=0 \text{ für } i \neq j} \quad \forall i \leq t-1 \\ &= \langle b - Ax^{(t)}, Ap^{(i)} \rangle + \beta_i^{(t-1)} \langle p^{(i)}, Ap^{(i)} \rangle \end{aligned}$$

Fall 1: $i \leq t-2$

$$Ap^{(i)} \leq \text{span}\{d^{(0)}, \dots, AA^i d^{(0)}\} = \text{span}\{p^{(0)}, \dots, p^{(i+1)}\}$$

$$\text{Galerkin: } \langle b - Ax^{(t)}, p^{(k)} \rangle = 0 \text{ für } k \leq t-1 \Rightarrow \beta^{(t-1)} = 0$$

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:

$$p^{(t)} = b - Ax^{(t)} + \beta_{t-1}^{(t-1)} p^{(t-1)} \quad (10.5)$$

Konvergenz:

$$\|x^{(t)} - x\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right) \|x^{(0)} - x\|_A \quad \varrho \approx 1 - \frac{1}{\sqrt{\kappa(A)}}$$

Zur Reduktion des Fehlers um den Faktor ε benötigt man $t(\varepsilon) = \mathcal{O}(\sqrt{\kappa(A)} |\log \varepsilon|)$ Iterationen. Dies führt für die Diskretisierung der Poissongleichung mit dem Fünfpunktstern auf die Komplexitäten:

	CG	Relaxation	Bandmatrix	Nested Diss
$d = 2$	$\mathcal{O}(N^{\frac{3}{2}})$	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2)$	$\mathcal{O}(N^{\frac{3}{2}})$
$d = 3$	$\mathcal{O}(N^{\frac{4}{3}})$	$\mathcal{O}(N^{\frac{5}{3}})$	$\mathcal{O}(N^{\frac{7}{3}})$	$\mathcal{O}(N^2)$