Ingenieurmathematik IV / Numerik von Differentialgleichungen

Olaf Ippisch
Institut für Mathematik
TU Clausthal
Erzstr. 1
D-38678 Clausthal-Zellerfeld
E-mail: olaf.ippisch@tu-clausthal.de

Peter Bastian
Interdisziplinäres Zentrum für
Wissenschaftliches Rechnen
Universität Heidelberg
Im Neuenheimer Feld 368
D-69120 Heidelberg
E-mail: peter.bastian@iwr.uni-heidelberg.de

19. Januar 2017

Der erste Teil dieses Skripts basiert auf einer Vorlesungsmitschrift der Vorlesung "Numerik von gewöhnlichen Differentialgleichungen", die von Stefan Breuning angefertigt wurde.

Inhaltsverzeichnis

Teil I.
Gewöhnliche Differentialgleichungen

Organisatorisches

Übungen

- Übung
  - 3 Stunden Vorlesung/1 Stunde große Übung, variabel verteilt
  - Tutorien
  - Anmeldung über Stud.IP
  - Korrektur einer Aufgabe pro Übungsblatt für Bonus (70 % der Punkte notwendig)
  - Abgabe
    - in Zweier oder Dreiergruppen
    - Freitag zu Beginn der Vorlesung
  - Rückgabe in der Tutoriumsgruppe
- Programmieraufgaben in Python.
  - Wir werden für die Lösung partieller Differentialgleichungen im zweiten Teil der Vorlesung FEniCS verwenden. Installationsanleitung für Linux und Virtuelle Maschinen für Windows/MacOS werden in StudIP bereitgestellt.

Intensive Beteiligung an den Übungen ist essentiell für den Lernerfolg (und für das Bestehen der Klausur)

Klausur

- Montag, 18. Juli 2016, 9–11 Uhr
- (Nachklausur Ingenieurmathematik III: Freitag, 15. Juli 2016, tagsüber)
- Erlaubte Hilfsmittel: ein eigenhändig beschriebenes (nicht kopiertes) DIN A4 Blatt. Vorder- und Rückseite dürfen verwendet werden.
- Taschenrechner (nicht programmierbar)

Literatur

- Vorlesungsmitschrift
  - wird während der Vorlesung erstellt
  - enthält den Tafelanschrieb
  - erhältlich über Stud.IP

- Allgemeine Literatur:
  - Wikipedia
  - Bärwolf, G.: „Numerik für Ingenieure, Physiker und Informatiker: für Bachelor und Diplom“, Spektrum Akademischer Verlag, 2008.
  - Dahmen, W. und Reusken, A.: „Numerik für Ingenieure und Naturwissenschaftler“, Springer, 2. korr. Aufl. 2008
  - Hanke-Bourgeois, M.: „Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens“, Vieweg+Teubner Verlag, 3. akt. Aufl. 2009
  - Plato, R.: „Numerische Mathematik kompakt: Grundlagenwissen für Studium und Praxis“, Vieweg+Teubner Verlag, 4. Aufl. 2010
  - Schwarz, H. R.: „Numerische Mathematik“, Vieweg+Teubner Verlag, 8. akt. Aufl. 2011

- Spezielle Literatur:
  - Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Vorlesungsskript, `http://numerik.iwr.uni-heidelberg.de/~lehre/notes/num1/Numerik_1.pdf`
  - Rannacher, R.: „Numerik 2: Numerik partieller Differentialgleichungen“, Vorlesungsskript, `http://numerik.iwr.uni-heidelberg.de/~lehre/notes/num2/numerik2.pdf`
  - Großmann, C. und H.-G. Roos: „Numerische Behandlung partieller Differentialgleichungen“, Teubner Studienbücher Mathematik, 2005
  - Knabner, P. und Angermann, L.: „Numerik Partieller Differentialgleichungen: Eine anwendungsorientierte Einführung“, Springer, 2000.
  - Larsson, S.: „Partielle Differentialgleichungen und numerische Methoden“, Springer 2005.

Studentenbeteiligung

Zur Vorlesung

- Rückmeldeblatt in jeder Vorlesung, nur zwei Fragen:
  - Was ist das Wichtigste, das Sie heute gelernt haben?
  - Was haben Sie am wenigsten verstanden?

  Kann gerne für weitere Anmerkungen genutzt werden.
- Rückmeldung über Tutoren
- Mail an mich: `olaf.ippisch@tu-clausthal.de`

Zu den Übungen

- Rückmeldung über Tutoren
- Mail an mich: `olaf.ippisch@tu-clausthal.de`

1. Motivation

1.1. Wachstumsmodelle

Sei $y(t)\colon [a,b] \to \mathbb{R}$ die Zahl der Individuen einer Population von Tieren oder Pflanzen

Wir treffen dabei folgende vereinfachende Annahmen:

- Die Anzahl der Individuen ist eine kontinuierlich Größe
- Die räumliche Verteilung wird vernachlässigt
- Die Zunahme der Individuen im Zeitintervall Δt ist proportional zu Δt und der aktuellen Anzahl an Individuen

Damit ergibt sich das Modell:

$$\underbrace{y(t+\Delta t)}_{\substack{\#\text{ Anzahl Individ.}\\ \text{am Ende}}} = \underbrace{y(t)}_{\substack{\#\text{ Anzahl Individ.}\\ \text{am Anfang}}} + \underbrace{\lambda \cdot \Delta t \cdot y(t)}_{\text{Netto Zuwachs}}$$

Dabei ist $\lambda \in \mathbb{R}$ die sogenannte Wachstumsrate (Wachstum für $\lambda > 0$, Zerfallsgesetz für $\lambda < 0$). Umstellen liefert

$$\frac{y(t+\Delta t) - y(t)}{\Delta t} = \lambda \cdot y(t)$$

Im Grenzwert für $\Delta t \to 0$ ergibt sich daraus:

$$\boxed{\frac{dy(t)}{dt} = \lambda \cdot y(t)} \tag{1.1}$$

Dies ist eine Differentialgleichung (DGL), da $y(t)$ durch eine Bedingung an die Ableitung bestimmt wird. Es ist eine gewöhnliche DGL, da y eine Funktion nur einer Variablen ist.
Zur Lösung der DGL machen wir den Ansatz $y(t) = e^{\lambda t}$ mit $\lambda \in \mathbb{R}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \underbrace{e^{\lambda t}}_{y(t)} = \lambda \cdot \underbrace{e^{\lambda t}}_{y(t)}$$

Daher ist $e^{\lambda t}$ tatsächlich eine Lösung von (1.1).
Für eine *beliebige* Lösung $y(t)$ von Gleichung (1.1) gilt:

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

also muss $y(t)\cdot e^{-\lambda t} = C$ für ein festes $C \in \mathbb{R}$ gelten. Daher haben alle Lösungen von (1.1) die Form

$$y(t) = C \cdot e^{\lambda t}.$$

Die Konstante C muss durch eine zusätzliche Bedingung zu Gleichung (1.1) festgelegt werden.

Als Anfangswertaufgabe (AWA) bezeichnet man

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

Dabei ist Y eine vorgegebene Anfangspopulation.
Wegen $y(a) = Ce^{\lambda a} = Y$ folgt $C = Y \cdot e^{-\lambda a}$ und schließlich

$$y(t) = Ce^{\lambda t} = y_0 \cdot e^{\lambda(t-a)}$$

als Lösung der AWA (1.2).

(1.2) mit $\lambda \in \mathbb{C}$ wird häufig als Testproblem verwendet. Als Wachstumsmodell ist die Gleichung nicht sehr realistisch Wachstum erfordert Ressourcen (Energie, Nahrung) $\Rightarrow$ es gibt eine Obergrenze für Wachstum.

Sei $y(t) \in [0,1]$ festgelegt (1 = Maximalpopulation). Dies wird durch das logistische Wachstumsmodell erreicht:

$$y'(t) = \underbrace{\lambda \cdot (1 - y(t))}_{\substack{\text{Wachstumsrate} \\ \to 0 \text{ für } y \to 1}} \cdot y(t) \tag{1.3}$$

Dies ist eine nichtlineare gewöhnliche DGL. Ihre Lösungsmenge kann durch die Methode der Trennung der Variablen bestimmt werden.

Für kompliziertere DGL: Numerische Lösungsverfahren nötig!

Einfache numerische Lösungsverfahren für (1.2):
Wir teilen die Zeit in gleichmäßige Intervalle:

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

und bezeichnen mit y_i die Lösung zur $y(t_i)$ zur Zeit t_i.

Explizites Euler-Verfahren:

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

Herleitung des Expliziten Euler-Verfahrens:

- 1. Möglichkeit: Näherung der Ableitung durch einen Differenzenquotienten:

$$\left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t_{i-1}} \approx \frac{y_i^h - y_{i-1}^h}{h} = \lambda \cdot y_{i-1}^h$$

- 2. Möglichkeit: Integration der Gleichung auf beiden Seiten:

$$\begin{aligned}
&\int_{t_{i-1}}^{t_i} \frac{\mathrm{d}y}{\mathrm{d}t}\,\mathrm{d}t = \int_{t_{i-1}}^{t_i} \lambda y(t)\,\mathrm{d}t \\
&\iff [y(t)]_{t_{i-1}}^{t_i} = \lambda \int_{t_{i-1}}^{t_i} y(t)\,\mathrm{d}t \\
&\iff y(t_i) - y(t_{i-1}) = \lambda \int_{t_{i-1}}^{t_i} y(t)\,\mathrm{d}t \overset{\substack{\text{Mittel-}\\ \text{wert-}\\ \text{satz}}}{=} \lambda \underbrace{(t_i - t_{i-1})}_{h} y(\xi) \quad \text{für ein } \xi \in]t_{i-1}, t_i[
\end{aligned}$$

  Nähern wir $\xi \approx t_{i-1}$ erhalten wir das explizite Euler-Verfahren:

$$y(t_i) = y(t_{i-1}) + \lambda \cdot h \cdot y(t_{i-1})$$

Implizites Euler-Verfahren:
Nähern wir $\xi \approx t_i$ erhalten wir das implizite Euler-Verfahren:

$$\begin{aligned}
&y(t_i) - y(t_{i-1}) = \lambda \cdot h \cdot y(t_i) \\
&\iff (1 - h\lambda)\, y_i^h = y_{i-1}^h \\
&\iff y_i^h = \frac{1}{1 - h\lambda} y_{i-1}^h i = 1, \ldots, N
\end{aligned}$$

Im allgemeinen Fall, z.B. für Gleichung (1.3) erfordert das implizite Euler-Verfahren das Lösen einer (nichtlinearen) Gleichung im Gegensatz zum expliziten Euler-Verfahren.
Frage: Lohnt sich der deutlich höhere Aufwand?

1.2. Chemische Reaktionen

Zwei Stoffe A und B reagieren zu einem Stoff C, also

$$A + B \to C.$$

Sei $c_A(t)$ die Konzentration (z.B. Mol pro Volumen) von A zur Zeit t.
Analog: $c_B(t)$, $c_C(t)$ und $k > 0$ die Reaktionsgeschwindigkeit. Dann erhalten wir folgende Gleichungen für die Konzentration von A:

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

Bei analogem Ansatz für B und C erhalten wir für $\Delta t \to 0$ ein System gewöhnlicher DGL:

$$\begin{aligned}
\frac{\mathrm{d}c_A(t)}{\mathrm{d}t} &= -k \cdot c_A(t) \cdot c_B(t) \\
\frac{\mathrm{d}c_B(t)}{\mathrm{d}t} &= -k \cdot c_A(t) \cdot c_B(t) \\
\frac{\mathrm{d}c_C(t)}{\mathrm{d}t} &= +k \cdot c_A(t) \cdot c_B(t)
\end{aligned} \tag{1.4}$$

Auch hier sind Anfangsbedingungen für alle drei Substanzen erforderlich.

Für die komplexere Reaktion

$$a \cdot A + b \cdot B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} c \cdot C + d \cdot D$$

mit $a, b, c, d \in \mathbb{N}$ (Stöchiometrie) erhalten wir die AWA

$$\begin{aligned}
&c'_A(t) = a \cdot R(t) \\
&c'_B(t) = b \cdot R(t) \\
&c'_C(t) = -c \cdot R(t) \\
&c'_D(t) = -d \cdot R(t) \\
&\text{mit: } R(t) = -k_1 \cdot c_A(t)^a \cdot c_B(t)^b + k_2 \cdot c_C(t)^c \cdot c_D(t)^d \\
&\text{und: } c_i(t_0) = C_i \quad i \in \{A, B, C, D\}\,.
\end{aligned} \tag{1.5}$$

Im chemischen Gleichgewicht gilt:

$$c'_i(t) = 0 \iff R(t) = 0 \iff \boxed{\frac{c_A^a \cdot c_B^b}{c_C^c \cdot c_D^d} = \frac{k_2}{k_1} = K_{\text{eq}}}$$

das sogenannte *Massenwirkungsgesetz*.

1.3. Astrophysikalisches N-Körper-Problem

Gegeben sind N als punktförmig angenommene Körper mit den Massem m_i, $i = 1, \ldots, N$. Zu berechnen sind die Positionen $\mathbf{x}_i(t) \in \mathbb{R}^3$ und die Geschwindigkeit $\mathbf{v}_i(t) \in \mathbb{R}^3$ der Bewegung der Körper in ihrem eigenen Schwerefeld.

Das Gravitationsgesetz beschreibt die Anziehungskraft ($\|.\|$ ist die euklidische Norm) und $G = 6.674 \cdot 10^{-11} \frac{\text{m}^3}{\text{kg s}^2}$ die Gravitationskonstante:

$$\mathbf{F}_{ij}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \underbrace{G \frac{m_i \cdot m_j}{\|\mathbf{x}_j - \mathbf{x}_i\|^2}}_{\text{Betrag}} \underbrace{\frac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|}}_{\text{Richtung}} \tag{1.6}$$

Die Dynamik ergibt sich aus dem *2. Newtonschen Gesetz* ($\mathbf{F} = m \cdot \mathbf{a}$)

$$m_i \cdot \mathbf{a}_i(t) = \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbf{F}_{ij}\left(\mathbf{x}_i(t), \mathbf{x}_j(t)\right) = G\, m_i \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{m_j \left(\mathbf{x}_j(t) - \mathbf{x}_i(t)\right)}{\|\mathbf{x}_j(t) - \mathbf{x}_i(t)\|^3}$$

Mit $\mathbf{a}_i(t) = \frac{\mathrm{d}\mathbf{v}_i(t)}{\mathrm{d}t}$ und $\frac{\mathrm{d}\mathbf{x}_i(t)}{\mathrm{d}t} = \mathbf{v}_i(t)$ erhält man ein System von $6N$ gewöhnlichen DGL.

$$\frac{\mathrm{d}\mathbf{x}_i(t)}{\mathrm{d}t} = \mathbf{v}_i(t) \qquad \mathbf{x}_i(t_0) = \mathbf{x}_i^0 \tag{1.7a}$$

$$\frac{\mathrm{d}\mathbf{v}_i(t)}{\mathrm{d}t} = G \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{m_j \cdot \left(\mathbf{x}_j(t) - \mathbf{x}_i(t)\right)}{\|\mathbf{x}_j - \mathbf{x}_i\|^3} \qquad \mathbf{v}_i(t_0) = \mathbf{v}_i^0 \tag{1.7b}$$

Potential:
Sei

$$\varphi(\mathbf{x},\mathbf{y}) = -\frac{1}{\|\mathbf{y}-\mathbf{x}\|} = -\left(\sum_{k=1}^{3}(y_k - x_k)^2\right)^{-\frac{1}{2}} \tag{1.8}$$

dann rechne

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

$$\Rightarrow \boxed{\nabla_y \varphi(\mathbf{x},\mathbf{y}) = \frac{\mathbf{y}-\mathbf{x}}{\|\mathbf{y}-\mathbf{x}\|^3}}$$

Damit können wir (1.7b) auch schreiben als

$$\frac{\mathrm{d}\mathbf{v}_i(t)}{\mathrm{d}t} = G\sum_{\substack{j=1\\ j\neq i}}^{N} m_j \nabla_{\mathbf{y}}\varphi(\mathbf{x}_i(t),\mathbf{x}_j(t))$$

Dieses System ist:

- nicht dissipativ (es gibt keine Reibung)
- konservativ (Energie bleibt erhalten)

Die Energie des Systems besteht nur aus

- potentieller Energie

$$E_{\mathrm{pot}}(t) = G\sum_{i=1}^{N}\sum_{j<i}^{N} m_i\, m_j\, \varphi(\mathbf{x}_i(t),\mathbf{x}_j(t))$$

- und kinetischer Energie

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

Die Gesamtenergie bleibt erhalten solange keine Kräfte von außen auf das System einwirken.

$$E_{\mathrm{tot}} = E_{\mathrm{pot}}(t) + E_{\mathrm{kin}}(t) = \mathrm{const} \tag{1.9}$$

Es ist nicht trivial, dass die Energieerhaltung (1.9) auch für ein numerisches Lösungsverfahren *exakt* für alle Zeiten und Schrittweiten gilt. Die Erhaltung solch spezieller Größen wie Energie, Impuls und Masse ist aber oft wichtig und erfordert *speziell angepasste Verfahren.*

Anmerkung 1.1

- *Da keine Energie zugeführt oder durch Reibung „verloren“ geht, bewegt sich der Schwerpunkt:*

$$\mathbf{R}(t) = \frac{1}{M}\sum_{i=1}^{N} m_i \mathbf{x}_i \qquad \textit{mit} \quad M = \sum_{i=1}^{N} m_i$$

mit der Geschwindigkeit

$$\mathbf{V}(t) = \frac{1}{M}\sum_{i=1}^{N} m_i \mathbf{v}_i.$$

„Korrigiere“ Anfangswerte durch

$$\tilde{\mathbf{x}}_i(0) \coloneqq \mathbf{x}_i(0) - \mathbf{R}(0), \qquad \tilde{\mathbf{v}}_i(0) \coloneqq \mathbf{v}_i(0) - \mathbf{V}(0)$$

- *Für $\mathbf{x}_i \to \mathbf{x}_j$ gilt $\mathbf{F}_{ij} \to \infty$. Wird $\|\mathbf{x}_j - \mathbf{x}_i\|$ zu klein, ist das Modell nicht mehr brauchbar, da Kollisionen nicht modelliert werden. Um diese Schwierigkeit zu umgehen „regularisiert“ man das Potential*

$$\varphi(\mathbf{x}, \mathbf{y}) = -\frac{1}{(\|\mathbf{y} - \mathbf{x}\|^2 + \varepsilon^2)^{\frac{1}{2}}}$$

und setzt

$$\frac{\mathrm{d}\mathbf{v}_i(t)}{\mathrm{d}t} = G\sum_{\substack{j=1 \\ j\neq i}}^{N} m_j \nabla_{\mathbf{y}} \varphi_\varepsilon(\mathbf{x}_i(t), \mathbf{x}_j(t))$$

für ein kleines ε. Dies ist vor allem ein Problem bei sehr großen N, z.B. Galaxiensimulationen.

1.4. Raketengleichung

Die Bewegung eines Raumfahrzeuges mit Raketenmotor wird beschrieben durch

$m(t)$ Masse des Raumfahrzeugs zur Zeit t
(verringert sich durch ausgestoßenen Treibstoff)
$\mathbf{x}(t)$ Position des Raumfahrzeugs
$\mathbf{v}(t)$ Geschwindigkeit des Raumfahrzeugs

Das 2. Newton'sche Gesetz lässt sich mit dem Impuls $\mathbf{p}(t) = m(t) \cdot \mathbf{v}(t)$ der Rakete schreiben als

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = \underbrace{\mathbf{F}_G(t)}_{\substack{\text{Gravitationskraft} \\ \text{z.B. aus } N\text{-Körper-Problem}}} + \underbrace{\mathbf{F}_T(t)}_{\text{Treibstoffausstoss}}$$

$$\mathbf{F}_T(t) = \underbrace{-}_{\text{actio = reactio}} \underbrace{c(t)}_{\substack{\text{Treibstoffaus-} \\ \text{stossrate}\left[\frac{\text{kg}}{\text{s}}\right]}} \cdot \underbrace{\mathbf{w}(t)}_{\substack{\text{Aussstoßgeschwindigkeit} \in \mathbb{R}^3 \\ \text{Richtung steuert die Rakete}}}$$

Daraus ergibt sich das System:

$$\frac{\mathrm{d}\left(m(t)\cdot\mathbf{v}(t)\right)}{\mathrm{d}t} = \mathbf{F}_G(t) - c(t)\cdot\mathbf{w}(t) \tag{1.10a}$$

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{v}(t) \tag{1.10b}$$

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = -c(t) \tag{1.10c}$$

plus Anfangsbedingungen und der Nebenbedingung

$$c(t) \geq 0 \int_{t_0}^{t_1} c(t)\,\mathrm{d}t \leq m(t_0)$$

- Beachte die „implizite“ Form von (1.10a). Dies könnte man vermeiden durch Verwendung von $p(t)$ als Primärvariable.
- Da $c(t)$ eine bekannte Funktion ist, gilt

  $$m(t) = m(t_0) - \int_{t_0}^{t} c(t)\,\mathrm{d}t,$$

  d.h. (1.10c) kann explizit integriert werden.
- $c(t)$ ist typischerweise nur stückweise glatt → Diese „Schaltpunkte“ sollten bei der Wahl des „Gitters“ berücksichtigt werden.

Erweiterung: *Optimalsteuerungsproblem*

- Wähle $c(t), \mathbf{w}(t)$ so dass ein Punkt im Raum mit möglichst wenig Treibstoff zu einem bestimmten Zeitpunkt erreicht wird.
- $c(t), \mathbf{w}(t)$ unterliegen technischen Beschränkungen (äußert sich mathematisch als komplizierte Nebenbedingungen): Schaltpunkte, Winkel der Düsen, Minimal- und Maximalwerte von $\|\mathbf{w}\|$ usw.
- liefert ein differentiell-algebraisches System.

1.5. Lagrange-Formalismus in der Mechanik

Formulierung der klassischen Mechanik über ein *Extremalprinzip*. Das physikalische System wird beschrieben durch die skalare Lagrange-Funktion

$$\mathcal{L} :]t_0, t_1[\times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

Beim obigen N-Körper-Problem ist $\mathcal{L}$ die Gesamtenergie

$$\mathcal{L}(t, \mathbf{x}(t), \mathbf{v}(t)) = E_{\mathrm{tot}}(t)$$

Die Aufgabe lautet dann: Finde eine Funktion $\mathbf{x}(t)$, sodass das folgende Funktional[1] minimal wird

$$\underbrace{I[\mathbf{x}]}_{\text{„Funktional“}} = \int_{t_0}^{t_1} \mathcal{L}\left(t, \mathbf{x}(t), \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}(t)\right) \mathrm{d}t \to \min.$$

Die Variationsrechnung liefert $\mathbf{x}$ als Lösung der *Euler-Lagrange-Gleichung*

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \mathbf{v}}\left(t, \mathbf{x}(t), \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right) = \frac{\partial \mathcal{L}}{\partial \mathbf{x}}\left(t, \mathbf{x}(t), \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right) \tag{1.11}$$

was wir schreiben können als

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \mathbf{v} \\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \mathbf{v}}(t, \mathbf{x}, \mathbf{v}) &= \frac{\partial \mathcal{L}}{\partial \mathbf{x}}(t, \mathbf{x}(t), \mathbf{x}(t)) \end{aligned}$$

Beachte: Die linke Seite der DGL ist nicht in Standardform.

1.6. Zusammenfassung

- Gewöhnliche Differentialgleichungen beschreiben viele Vorgänge in den Natur- und Ingenieurwissenschaften.
- Einfache numerische Lösungsverfahren sind das explizite und das implizite Euler-Verfahren. Das implizite Euler-Verfahren erfordert im Gegensatz zum expliziten Euler-Verfahren das Lösen eines Gleichungssystems.
- Erhaltungsgleichungen werden nicht notwendigerweise auch für ein numerisches Lösungsverfahren exakt für alle Zeiten und Schrittweiten erfüllt. Dies erfordert spezielle Verfahren.

[1] Ein Funktional ist eine reelle oder komplexe Größe, die für die Elemente eines Vektorraums berechnet werden kann, also eine „Funktion“ deren Argument nicht eine Zahl sondern ein Vektor ist. Im obigen Fall ist der Vektor $\mathbf{x}(t)$ eine auf dem Intervall $[t_0, t_1]$ integrierbare Funktion.

2. Zur Theorie gewöhnlicher Differentialgleichungen

2.1. Notation

Im weiteren wird folgende Notation verwendet:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{d} x_i y_i \qquad \text{(Skalarprodukt)}$$

$$\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}} \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \qquad \text{(Euklidische Norm)}$$

$$\|\mathbf{A}\| = \sup_{\mathbf{x} \in \mathbb{R}^d,\ \|\mathbf{x}\| \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \qquad \text{(Spektralnorm)}$$

$$\mathbf{u}'(t) = \frac{\mathrm{d}\mathbf{u}(t)}{\mathrm{d}t} = \left(\frac{\mathrm{d}u_1(t)}{\mathrm{d}t}, \ldots, \frac{\mathrm{d}u_d(t)}{\mathrm{d}t} \right)^T \qquad \text{(Ableitung von } \mathbf{u} \text{ nach } t\text{)}$$

$$\mathbf{f}'_t(t, \mathbf{x}) = \frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial t} = \left(\frac{\partial f_1(t, \mathbf{x})}{\partial t}, \ldots, \frac{\partial f_d(t, \mathbf{x})}{\partial t} \right)^T \qquad \text{(Ableitung von } \mathbf{f} \text{ nach } t\text{)}$$

$$\partial_i \mathbf{f}(t, \mathbf{x}) = \frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial x_i} = \left(\frac{\partial f_1(t, \mathbf{x})}{\partial x_i}, \ldots, \frac{\partial f_d(t, \mathbf{x})}{\partial x_i} \right)^T \qquad \begin{pmatrix} \text{Ableitung von } \mathbf{f} \\ \text{nach der } i\text{-ten Komponente von } \mathbf{x} \end{pmatrix}$$

$$\mathbf{J}_{\mathbf{f},\mathbf{x}}(t, \mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial x_1} & \cdots & \frac{\partial f_d}{\partial x_d} \end{pmatrix} \qquad \begin{pmatrix} \text{Jacobi-Matrix für die Ableitung von } \mathbf{f} \\ \text{nach den Komponenten von } \mathbf{x} \end{pmatrix}$$

2.2. Gewöhnliche Differentialgleichungen

Definition 2.1
Eine Gleichung zur Bestimmung einer Funktion $u = u(t)$ in der neben der Funktion $u(t)$ und Ausdrücken in t auch die Ableitungen von u bis zur n-ten Ordnung vorkommen, heißt gewöhnliche Differentialgleichung (DGL) n-ter Ordnung. Allgemein kann man diese in der Form

$$F\left(t, u, u', \ldots, u^{(n)}\right) = 0$$

schreiben.

Kann man diese nach der höchsten Ableitung $u^{(n)}$ auflösen:

$$u^{(n)} = f\left(t, u, u', \ldots, u^{(n-1)}\right)$$

so spricht man von einer expliziten gewöhnlichen Differentialgleichung n-ter Ordnung, ansonsten von einer impliziten.

Eine n-mal differenzierbare Funktion $u : I \to \mathbb{R}$, $I \subseteq \mathbb{R}$ heißt Lösung der DGL, wenn sie die DGL in allen Punkten $t \in I$ erfüllt.

Lässt sich die DGL als

$$\sum_{i=0}^{n} \alpha_i(t) \cdot u^{(i)}(t) + g(t) = 0,$$

schreiben, wobei die Koeffizienten $\alpha_i(t)$ nur Funktionen von t aber nicht von $u^{(i)}$ für $0 \leq i \leq n$ sind, so sprechen wir von einer linearen Differentialgleichung.

Definition 2.2
Ein Anfangswertproblem besteht aus einer gewöhnlichen Differentialgleichung

$$F\left(t, u, u', \ldots, u^{(n)}\right) = 0$$

zusammen mit vorgegebenen Werten für die ersten $n-1$ Ableitungen:

$$u(t_0) = u_0, \;\; u'(t_0) = u_1, \ldots, \;\; u^{(n-1)}(t_0) = u_{n-1}$$

wobei $u_0, u_1, \ldots, u_{n-1}$ gegebene reelle Zahlen sind, die Anfangsbedingungen genannt werden.

Durch die Anfangsbedingungen wird die Lösung der DGL eindeutig festgelegt. Der Name kommt daher, dass DGL häufig die zeitliche Dynamik eines Systems beschreiben. Die Lösung $u(t)$ beschreibt dann z.B. den Ort u eines Körpers zur Zeit t.

Einige Verfahren für die *analytische* Lösung von gewöhnlichen Differentialgleichungen haben wir bereits in Ingenieurmathematik I behandelt. In dieser Vorlesung soll es vor allem um die Frage gehen, unter welchen Bedingungen überhaupt Lösungen für gewöhnliche DGL existieren und wie sich diese *numerisch* berechnen lassen.

2.3. Systeme von gewöhnlichen Differentialgleichungen

Definition 2.3 (System von Differentialgleichungen)
Man spricht von einem impliziten *System von Differentialgleichungen, wenn* $\mathbf{u}(t) = (u_1(t), \ldots, u_d(t))^T$ *eine vektorwertige Abbildung ist und ein Gleichungssystem*

$$\mathbf{F}\left(t, \mathbf{u}(t), \mathbf{u}'(t), \ldots, \mathbf{u}^{(n)}(t)\right) = 0$$

zu erfüllen ist.

Im ersten Teil dieser Vorlesung behandeln wir Systeme gewöhnlicher Differentialgleichungen der *expliziten* Form

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

mit

$$\mathbf{u}(t) = (u_1(t), \ldots, u_d(t))^T, \quad \mathbf{f}(t, \mathbf{x}) = (f_1(t, \mathbf{x}), \ldots, f_d(t, \mathbf{x}))^T.$$

Die Funktion $\mathbf{f}$ sei auf $D = I \times \Omega \subset \mathbb{R}^1 \times \mathbb{R}^d$ definiert und dort *stetig* (Wir erwarten, dass die Lösung $\mathbf{u}(t)$ stetig differenzierbar ist, daher ist die Forderung, dass $\mathbf{f}$ stetig ist, eine notwendige Voraussetzung).

(2.1) ist ein Spezialfall der allgemeinen *impliziten* DGL

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

Der Satz der impliziten Funktion sagt, dass sich eine Funktion $\mathbf{F}(t, \mathbf{x}, \mathbf{y})$ lokal nach $\mathbf{y}$ auflösen lässt, falls $\mathbf{J}_{\mathbf{F},\mathbf{y}} = \left(\frac{\partial F_i}{\partial y_j}\right)_{i,j=1}^{d}$ invertierbar ist. Ein implizites DGL-System 1. Ordnung lässt sich also in ein explizites DGL-System 1. Ordnung umformen, falls $\mathbf{J}_{\mathbf{F},\mathbf{u}'}$ regulär ist.

Anmerkung 2.4 (Reduktion von Differentialgleichungssystemen höherer Ordnung)
Ein DGL-System höherer Ordnung der Form

$$\mathbf{F}\left(t, \mathbf{u}(t), \mathbf{u}'(t), \ldots, \mathbf{u}^{(n)}(t)\right) = 0$$

lässt sich mittels Einführung von Hilfsvariablen $\mathbf{u}_i(t) = \mathbf{u}'_{i-1}(t)$ *auf ein System erster Ordnung reduzieren:*

$$\left.\begin{array}{rcl} \mathbf{u}_1(t) & := & \frac{\mathrm{d}\mathbf{u}(t)}{\mathrm{d}t} \\ & \vdots & \\ \mathbf{u}_{n-1}(t) & := & \frac{\mathrm{d}\mathbf{u}_{n-2}(t)}{\mathrm{d}t} \end{array}\right\} (n-1) \text{ Gleichungen aus der Definition der Hilfsvariabeln}$$

$\mathbf{F}\left(t, \mathbf{u}(t), \ldots, \mathbf{u}_{n-1}(t), \mathbf{u}'_{n-1}(t)\right) = 0$ *Originalgleichung, jetzt DGL 1. Ordnung für Hilfsvariablen*

Beispiel 2.5

- $$\mathbf{u}''(t) = \mathbf{u} + \mathbf{a}$$

  Durch Einführung der Hilfsvariablen $\mathbf{u}_1(t) = \mathbf{u}'(t)$ *erhalten wir das System 1. Ordnung*

  $$\mathbf{u}'(t) = \mathbf{u}_1$$
  $$\mathbf{u}'_1(t) = \mathbf{u} + \mathbf{a}$$

- $$\frac{\mathrm{d}^3 u}{\mathrm{d}t^3} = \sin(t) \cdot u(t) + 12u'(t)$$

  Durch Einführung der Hilfsvariablen $u_1(t) = u'(t)$ *und* $u_2(t) = u'_1(t)$ *erhalten wir das System 1. Ordnung*

  $$u'(t) = u_1$$
  $$u'_1(t) = u_2$$
  $$u'_2(t) = \sin(t) \cdot u(t) + 12u_1(t)$$

- $$u(t) \cdot \frac{\mathrm{d}^3 u}{\mathrm{d}t^3} - \sin(t \cdot u(t)) \cdot u(t) + 12u'(t)^2 = 0$$

  Durch Einführung der Hilfsvariablen $u_1(t) = u'(t)$ *und* $u_2(t) = u'_1(t)$ *erhalten wir das (implizite) System 1. Ordnung*

  $$u'(t) = u_1$$
  $$u'_1(t) = u_2$$
  $$u(t) \cdot u'_2(t) - \sin(t \cdot u(t)) \cdot u(t) + 12u_1(t)^2 = 0$$

Definition 2.6 (Spezielle Formen der rechten Seite)
Ein System von DGL der Form (2.1) *heißt*

- autonom, *falls die Funkton* $\mathbf{f}$ *nicht von der Zeit abhängt, falls also* $\mathbf{f}(t, \mathbf{x}) = \mathbf{f}(\mathbf{x})$.
- separiert, *falls* $\mathbf{f}(t, \mathbf{x}) = a(t) \cdot \mathbf{g}(\mathbf{x})$, *d.h.* $\mathbf{f}$ *ist ein Produkt aus einer Funktion* a, *die nur von* t *abhängt und von einer Funktion* $\mathbf{g}$, *die nur von* $\mathbf{x}$ *abhängt.*
- linear, *falls* $\mathbf{f}(t, \mathbf{x}) = \mathbf{A}(t) \cdot \mathbf{x} + \mathbf{b}(t)$ *mit* $\mathbf{A}(t) \in \mathbb{R}^{d \times d}$ *und* $\mathbf{b}(t) \in \mathbb{R}^d$. *Das lineare DGL-System heißt homogen, wenn* $\mathbf{b} \equiv \mathbf{0}$.

Beispiel 2.7

- *Die Wachstumsmodelle (Abschnitt 1.1), die chemischen Reaktionsgesetze (Abschnitt 1.2) und das N-Körper-Problem (Abschnitt 1.3) sind autonom.*
- *Das einfache Wachstumsmodell ist außerdem linear.*
- $u'(t) = u(t)^2 + 2$ *ist autonom und nicht-linear.*
- $u'(t) = u(t) \cdot t$ *ist separiert und linear.*
- $u'(t) = u(t) \cdot t + 2$ *ist weder autonom noch separiert, aber linear.*
- *Ein lineares DGL-System erster Ordnung ist:*

$$\begin{pmatrix} 27t & -3 \\ 18 & \sin(4t) \end{pmatrix} \cdot \mathbf{u} + \begin{pmatrix} \frac{3}{t} \\ e^t \end{pmatrix} = \mathbf{0}$$

das zugehörige homogene lineare DGL-System ist:

$$\begin{pmatrix} 27t & -3 \\ 18 & \sin(4t) \end{pmatrix} \cdot \mathbf{u} = \mathbf{0}$$

Definition 2.8 (Anfangswertaufgabe(AWA))
Zu einem gegebenen Punkt $(t_0, \mathbf{u}_0) \in D$ *ist eine (stetig) differenzierbare Funktion* $\mathbf{u}\colon I \to \mathbb{R}^d$ *gesucht mit den Eigenschaften*

1) $\text{Graph}(\mathbf{u}) \coloneqq \{(t, \mathbf{u}(t)) : t \in I\} \subset D$ *(d.h.* $\mathbf{f}$ *ist für alle Punkte* $(t, \mathbf{u}(t))$ *wohldefiniert)*

2) $\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t)) \quad \forall\, t \in I$ *(d.h.* $\mathbf{u}(t)$ *erfüllt die DGL)*

3) $\mathbf{u}(t_0) = \mathbf{u}_0$ *(d.h.* $\mathbf{u}(t)$ *erfüllt die Anfangsbedingung)*

Dann heißt $\mathbf{u}(t)$ *Lösung der Anfangswertaufgabe.*

Nach dem Hauptsatz der Differential- und Integralrechnung ist eine stetige Funktion $\mathbf{u} : I \to \mathbb{R}^d$ genau dann Lösung der AWA, wenn $Graph(\mathbf{u}) \in D$ ist, und wenn $\mathbf{u}$ folgende „Integralgleichung“ erfüllt:

$$\mathbf{u}(t) = \mathbf{u}(t_0) + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{u}(s))\, \mathrm{d}s, \quad t \in I \tag{2.3}$$

2.4. Existenz von Lösungen einer AWA

Wir betrachten zunächst Aussagen darüber, unter welchen Bedingungen wir beweisen können, das Lösungen existieren.

Satz 2.9 (Existenzsatz von Peano)
Die Funktion $\mathbf{f}(t, \mathbf{x})$ *sei stetig auf dem* $(d+1)$*-dimensionalen Zylinder*

$$D \coloneqq \Big\{(t, \mathbf{x}) \in \mathbb{R}^1 \times \mathbb{R}^d : |t - t_0| \leq \alpha,\ \|\mathbf{x} - \mathbf{u}_0\| \leq \beta\Big\}$$

Dann existiert eine Lösung $\mathbf{u}(t)$ *der Anfangswertaufgabe auf dem Intervall* $I \coloneqq [t_0 - T, t_0 + T]$*, wobei*

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

Beweis:
Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Satz 1.1

Idee: Konstruktion stückweise linearer Funktionen $\mathbf{u}^h$ mit dem expliziten Euler-Verfahren abhängig von Schrittweite h. Nachweis, dass für $h \to 0$ eine Teilfolge der Funktionen gegen die Lösung konvergiert. □

Anmerkung 2.10
Nur solange die Ableitung von $\mathbf{u}$*, d.h. die Funktion* $\mathbf{f}(t, \mathbf{u})$ *stetig ist, kann eine Lösung mit dem expliziten Euler-Verfahren konstruiert werden, dies ist also eine* hinreichende *Bedingung. Das heißt nicht* notwendigerweise*, dass eine Lösung nicht existiert, wenn* $\mathbf{f}$ *unstetig ist. Dies ist wichtig, da viele relevante Probleme Schaltern enthalten, z.B. die Raketengleichung (Abschnitt 1.4).*

Satz 2.11 (Fortsetzungssatz)
Die Funktion $\mathbf{f}(t, \mathbf{x})$ *sei stetig auf einem abgeschlossenen Bereich* D *des* $\mathbb{R}^1 \times \mathbb{R}^d$*, welcher den Punkt* $(t_0, \mathbf{u}_0)$ *enthält und es sei* $\mathbf{u}$ *eine Lösung der Anfangswertaufgabe auf einem Intervall* $I = [t_0 - T, t_0 + T]$*.*

Dann ist die lokale Lösung $\mathbf{u}$ *nach rechts und links auf ein „maximales“ Existenzintervall* $I_{\max} =]t_0 - T_*, t_0 + T^*[$ *stetig differenzierbar fortsetzbar, solange der Graph von* $\mathbf{u}$ *nicht an den Rand von* D *stößt. Dabei kann*

$$\mathrm{Graph}(\mathbf{u}) \coloneqq \{(t, \mathbf{u}(t)) : t \in I_{\max}\}$$

unbeschränkt sein, sowohl durch $t \to t_0 + T^* = \infty$*, als auch durch* $\|\mathbf{u}(t)\| \to \infty$ *für* $t \to t_0 + T^*$ *(entsprechend für* $t_0 - T_*$*).*

Beweis:
Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Satz 1.2 □

Korrolar 2.12 (Globale Existenz)
$\mathbf{f}(t, \mathbf{x})$ sei auf ganz $\mathbb{R}^1 \times \mathbb{R}^d$ definiert und stetig. Für jede durch den Satz von Peano gelieferte lokale Lösung $\mathbf{u}(t)$, gelte

$$\|\mathbf{u}(t)\| \leq \beta(t) \qquad t \in [t_0 - T, t_0 + T] \tag{2.4}$$

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

Beweis:
Da die Norm von $\mathbf{u}(t)$ durch die Bedingung (2.4) immer beschränkt bleibt, ist es ausgeschlossen, dass die Funktion $\mathbf{u}$ so stark anwächst, dass $\|\mathbf{u}\| \to \infty$ für $|t| < \infty$ („blow-up in finite time"). □

Beispiel 2.13
1)

$$u'(t) = \sin(u(t)), \quad t \geq 0, \quad u(0) = 0$$

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

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

sind diese nach Korollar 2.12 auf ganz $\mathbb{R}$ fortsetzbar.

2)

$$u'(t) = u(t)^{1/3}, \quad t \geq 0, \quad u(0) = 0$$

Für beliebiges $c \geq 0$ ist folgende Funktion eine Lösung der AWA:

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

Wir überprüfen das:

- *Für $u_c(t) = 0$ ist die AWA trivialerweise erfüllt, da $0^{1/3} = 0$.*
- *Sonst:*

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

Für die Anfangsbedingung $u(0) := 0$ hat die Anfangswertaufgabe also unendlich viele Lösungen. Das explizite Euler-Verfahren liefert dabei die Lösung $u(t) = 0$.

Wählen wir stattdessen die Anfangsbedingung $u(0) = 1$, dann hat diese (andere) Anfangswertaufgabe die eindeutige Lösung

$$u(t) = \left(\frac{2}{3}t + 1\right)^{3/2}.$$

3)

$$u'(t) = u(t)^2, \quad 0 \le t < 1, \quad u(0) = 1$$

besitzt eine lokale Lösung der Form

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

d.h. $u(t) \to \infty$ *für* $t \to 1 \Rightarrow$ *„Blow-up in finite time“.*

Dagegen hat

$$u'(t) = -2\,t\,u(t)^2, \quad t \ge 0, \quad u(0) = 1$$

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

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

Kleine Änderungen können also eine große Wirkung haben.

4) Das skalare „Modellproblem“

$$u'(t) = \lambda u(t) \quad t \ge 0 \quad u(0) = u_0 \qquad (\lambda \in \mathbb{C})$$

hat die global eindeutige Lösung $u(t) = u_0 \cdot e^{\lambda \cdot t}$ *(das haben wir oben gezeigt) mit*

$$\lambda < 0 \Rightarrow \lim_{t\to\infty} |u(t)| = 0$$
$$\lambda = 0 \Rightarrow \lim_{t\to\infty} |u(t)| = |u_0|$$
$$\lambda > 0 \Rightarrow \lim_{t\to\infty} |u(t)| = \infty$$

2.5. Eindeutigkeit und Stabilität

Definition 2.14 (Lipschitz-Bedingung)

(i) Die Funktion $\mathbf{f}(t, \mathbf{x})$ *genügt auf ihrem Definitionsbereich* $D \subset \mathbb{R}\times\mathbb{R}^d$ *einer (gleichmäßigen) „Lipschitz-Bedingung“, wenn mit einer stetigen Funktion* $L(t) > 0$ *gilt:*

$$\|\mathbf{f}(t,\mathbf{x}) - \mathbf{f}(t,\mathbf{y})\| \le L(t)\|\mathbf{x}-\mathbf{y}\|, \quad (t,\mathbf{x}),(t,\mathbf{y}) \in D \quad \text{(ein } L \text{ für alle } \mathbf{x},\mathbf{y}). \tag{2.5}$$

(ii) Die Funktion $\mathbf{f}(t, \mathbf{x})$ *genügt einer „lokalen“ Lipschitz-Bedingung, wenn* $\mathbf{f}(t, \mathbf{x})$ *auf jeder beschränkten Teilmenge von* D *einer Lipschitz-Bedingung genügt (*L *darf von Teilmenge abhängen).*

Beispiel 2.15

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

$$\max \left|\frac{\partial f}{\partial x}(t,x)\right| \le K, \quad (t,x) \in D.$$

Dann gilt

$$|f(t,x) - f(t,y)| = \left|\int_x^y \frac{\partial f}{\partial s}(t,s)\,\mathrm{d}s\right| \le K|x-y|.$$

Lässt sich auf $d > 1$ *erweitern.*

2) $f(t,x) = x^{1/3}$ *aus dem Beispiel oben ist nicht Lipschitz-stetig in* $x = 0$*, aber in* $[\epsilon, \infty)$ *mit* $\epsilon > 0$*.*

Satz 2.16 (Lokaler Stabilitätssatz)
Wir betrachten die beiden Anfangswertaufgaben:

$$\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t)), \qquad t \in I, \qquad \mathbf{u}(t_0) = \mathbf{u}_0, \tag{2.6a}$$
$$\mathbf{v}'(t) = \mathbf{g}(t, \mathbf{v}(t)), \qquad t \in I, \qquad \mathbf{v}(t_0) = \mathbf{v}_0, \tag{2.6b}$$

mit $\mathbf{f}, \mathbf{g}$ *stetig. Die Funktion* $\mathbf{f}(t, \mathbf{x})$ *genüge der (gleichmäßigen) Lipschitz-Bedingung 2.14 auf* D *mit* $L := \sup_{t\in I} L(t) < \infty$*. Dann gilt für zwei beliebige Lösungen* $\mathbf{u}$ *von* (2.6a) *und* $\mathbf{v}$ *von* (2.6b)*:*

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq e^{L\cdot(t-t_0)} \left[\|\mathbf{u}_0 - \mathbf{v}_0\| + \int_{t_0}^{t} \varepsilon(s)\, \mathrm{d}s \right], \quad t \in I,$$

mit

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

Beweis: Nach Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Satz 1.4:

Mit der Integraldarstellung gilt:

$$\begin{aligned}\mathbf{u}(t) - \mathbf{v}(t) &= \mathbf{u}(t_0) + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{u}(s))\, \mathrm{d}s - \mathbf{v}(t_0) - \int_{t_0}^{t} \mathbf{g}(s, \mathbf{v}(s))\, \mathrm{d}s \\ &= \int_{t_0}^{t} \mathbf{f}(s, \mathbf{u}(s)) - \underbrace{\mathbf{f}(s, \mathbf{v}(s))}_{\text{eingeschoben}}\, \mathrm{d}s + \int_{t_0}^{t} \underbrace{\mathbf{f}(s, \mathbf{v}(s))}_{\text{eingeschoben}} - \mathbf{g}(s, \mathbf{v}(s))\, \mathrm{d}s + \mathbf{u}_0 - \mathbf{v}_0\end{aligned}$$

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

$$\left\| \underbrace{\int_{t_0}^{t} \mathbf{e}(s) ds}_{\substack{\text{Integral über} \\ \text{jede Komponente}}} \right\| = \lim_{N\to\infty} \left\| \sum_{i=1}^{N} \mathbf{e}(s_i)(s_i - s_{i-1}) \right\| \underset{\substack{\text{Norm-} \\ \text{eigensch.} \\ \text{(Dreiecksungl.)}}}{\leq} \lim_{N\to\infty} \sum_{i=1}^{N} \|\mathbf{e}(s_i)\| \, (s_i - s_{i-1}) = \int_{t_0}^{t} \|\mathbf{e}(s)\|\, ds.$$

Damit gilt für $\mathbf{e}(t) = \mathbf{u}(t) - \mathbf{v}(t)$ die lineare Integralungleichung:

$$\begin{aligned}\|\mathbf{e}(t)\| &\underset{\substack{\text{Hilfs-} \\ \text{resultat}}}{\leq} \int_{t_0}^{t} \|\mathbf{f}(s, \mathbf{u}(s)) - \mathbf{f}(s, \mathbf{v}(s))\|\, \mathrm{d}s + \int_{t_0}^{t} \|\mathbf{f}(s, \mathbf{v}(s)) - \mathbf{g}(s, \mathbf{v}(s))\|\, \mathrm{d}s + \|\mathbf{u}_0 - \mathbf{v}_0\| \\ &\underset{\substack{\text{Lipschitz-} \\ \text{stetigkeit} \\ \text{für } \mathbf{f}, \\ \text{Def. von } \varepsilon(s)}}{\leq} L \int_{t_0}^{t} \|\mathbf{e}(s)\|\, \mathrm{d}s + \int_{t_0}^{t} \varepsilon(s)\, \mathrm{d}s + \|\mathbf{u}_0 - \mathbf{v}_0\|.\end{aligned}$$

Zum Abschluss des Beweises brauchen wir das sogenannte Gronwall-Lemma. □

Hilfssatz 2.17 (Gronwallsches Lemma)
Die stückweise stetige Funktion $w(t) \geq 0$ genüge mit zwei Konstanten $a, b \geq 0$ der Integralungleichung

$$w(t) \leq a \int_{t_0}^{t} w(s)\,\mathrm{d}s + b, \quad t \geq t_0. \tag{2.7}$$

Dann gilt die Abschätzung

$$w(t) \leq e^{a(t-t_0)} b, \quad t \geq t_0. \tag{2.8}$$

Beweis:
Für die Funktion

$$\psi(t) := a \int_{t_0}^{t} w(s)\,\mathrm{d}s + b$$

gilt $\psi'(t) = aw(t)$ und somit gemäß Voraussetzung $\psi'(t) \leq a\psi(t)$. Damit gilt:

$$\left(e^{-at}\psi(t)\right)' = e^{-at}\left(\psi'(t) - a\psi(t)\right) \leq 0,$$

d.h. die Funktion $e^{-at}\psi(t)$ ist monoton fallend. Dies bedeutet, dass

$$e^{-at}w(t) \leq e^{-at}\psi(t) \leq \psi(t_0)e^{-at_0} = be^{-at_0}, \quad t \geq t_0,$$

woraus die behauptete Ungleichung folgt. □

Anmerkung 2.18
Das Gronwallsche Lemma lässt sich wie folgt verallgemeinern:
Falls

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

mit einer stetigen Funktion $a(t) \geq 0$ und einer monoton steigenden Funktion $b(t) \geq 0$, dann gilt

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

Abschluss des Beweises von Satz 2.16:
Es gilt mit den Bezeichnungen des Gronwall'schen-Lemmas und aus Satz 2.16

$$w(t) = \|\mathbf{e}(t)\|, \quad a(t) = L, \quad b(t) = \underbrace{\int_{t_0}^{t} \varepsilon(s)\,\mathrm{d}s + \|\mathbf{u}_0 - \mathbf{v}_0\|}_{\geq 0,\text{ nicht fallend.}}.$$

Also:

$$\|\mathbf{e}(t)\| \leq e^{L(t-t_0)} \left(\int_{t_0}^{t} \varepsilon(s)\,\mathrm{d}s + \|\mathbf{u}_0 - \mathbf{v}_0\| \right).$$

Korrolar 2.19 (Eindeutigkeitssatz)
Erfüllt $\mathbf{f}(t, \mathbf{x})$ eine Lipschitz-Bedingung, so ist die durch den Existenzsatz von Peano gelieferte Lösung eindeutig bestimmt.

Beweis:
Seien $\mathbf{u}(t), \mathbf{v}(t)$ zwei verschiedene Lösungen *derselben* AWA, also

$$\begin{aligned} \mathbf{u}'(t) &= \mathbf{f}(t, \mathbf{u}(t)), & \mathbf{u}(t_0) &= \mathbf{u}_0, \\ \mathbf{v}'(t) &= \mathbf{f}(t, \mathbf{v}(t)), & \mathbf{v}(t_0) &= \mathbf{v}_0 = \mathbf{u}_0. \end{aligned}$$

Dann gilt nach dem Stabilitätssatz

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq e^{L(t-t_0)} \cdot \left(\underbrace{\|\mathbf{u}_0 - \mathbf{v}_0\|}_{=0} + \int_{t_0}^{t} \sup_{x \in \Omega} \underbrace{\|\mathbf{f}(s, \mathbf{x}) - \mathbf{f}(s, \mathbf{x})\|}_{=0,\text{ da } g=f} \,\mathrm{d}s \right) = 0$$

□

Dies zeigt $\mathbf{u} = \mathbf{v}$ im Widerspruch zur Annahme.

Korrolar 2.20
Betrachte die DGL n-ter Ordnung

$$\mathbf{u}^{(n)}(t) = \mathbf{f}(t, \mathbf{u}(t), \ldots, \mathbf{u}^{(n-1)}(t)), \quad t \geq t_0.$$

$\mathbf{f}\colon I \times \mathbb{R}^d$ sei Lipschitz-stetig bezüglich der letzten n Argumente. Dann existiert für jeden Satz von n Werten $\mathbf{u}_0, \ldots, \mathbf{u}_{n-1} \in \mathbb{R}$ genau eine lokale Lösung, die den Anfangsbedingungen $\mathbf{u}^{(i)}(t_0) = \mathbf{u}_i, i = 0, \ldots, n-1$ genügt.

Beweis:
Umschreiben als System erster Ordnung. Die Systemfunktion $\mathbf{F}(t, \mathbf{x})$ ist dann Lipschitz-stetig. Nach dem Eindeutigkeitssatz folgt die Behauptung. □

Beispiel 2.21

1) $u'(t) = u^2$

d.h. zu prüfen: Ist $f(t,x) = x^2$ *„lokal“ Lipschitz-stetig?*

$$|x^2 - y^2| = |(x+y)(x-y)| \leq \underbrace{|x+y|}_{\leq L := \max_{x,y \in D} |x+y|} \cdot |x-y| \leq L|x-y|$$

Solange die Lösung existiert, ist sie eindeutig.

2) $u''(t) + ku(t) = 0$ *(lineare DGL 2. Ordnung, harmonischer Oszillator) hat die beiden Lösungen* $u_1(t) = \cos(\sqrt{k}t), \quad u_2(t) = \sin(\sqrt{k}t)$.
Jede Linearkombination $u'(t) = c_1 u_1(t) + c_2 u_2(t)$ *ist ebenfalls eine Lösung.*
Es ist $u(0) = c_1$ *und* $u'(0) = c_2\sqrt{k}$ *und damit*

$$u(t) = u(0)\cos(\sqrt{k}t) + \frac{u'(0)}{\sqrt{k}}\sin(\sqrt{k}t)$$

die Lösung zu diesen Anfangswerten. Nach Korollar 2.20 ist die Lösung eindeutig.

3) Wir betrachten nochmal die AWA

$$u'(t) = u(t)^{1/3}, \quad t \geq 0, \quad u(0) = 0$$

$x^{1/3}$ *ist nicht Lipschitz-stetig an der Stelle* $x = 0$. *Daher ist die Lösung für die Anfangsbedingung* $u(0) = 0$ *nicht eindeutig.*

Korrolar 2.22 (Globale Existenz)
$\mathbf{f}(t,\mathbf{x})$ *sei stetig auf* $D = \mathbb{R}^1 \times \mathbb{R}^d$ *und genüge der Wachstumsbedingung*

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

mit stetigem $\alpha(t), \beta(t) \geq 0$. *Dann besitzt die AWA eine „globale“ Lösung. Erfüllt* $\mathbf{f}(t,\mathbf{x})$ *eine Lipschitz-Bedingung, ist die Lösung eindeutig.*

Beweis: Siehe Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Korrolar 1.1
Sei $\mathbf{u}(t)$ eine lokale Lösung nach Peano auf $I = [t_0, t_0 + T]$ zum Startpunkt $(t_0, \mathbf{u}_0)$.
Dann gilt

$$\|\mathbf{u}(t)\| = \left\| \mathbf{u}_0 + \int_{t_0}^{t} \mathbf{f}\left(s, \mathbf{u}(s)\right) \, \mathrm{d}s \right\| \leq \|\mathbf{u}_0\| + \int_{t_0}^{t} \alpha(s)\, \|\mathbf{u}(s)\| + \beta(s) \, \mathrm{d}s, \quad t \in I.$$

Nach dem erweiterten Gronwall'schen Lemma (Anmerkung 2.18) gilt:

$$\|\mathbf{u}(t)\| \leq \underbrace{\exp\left(\int_{t_0}^{t} \alpha(s)\,\mathrm{d}s\right) \cdot \left[\|\mathbf{u}_0\| + \int_{t_0}^{t} \beta(s)\,\mathrm{d}s\right]}_{=: G(t,\alpha,\beta)}, \quad t \in I,$$

d.h.

$$\|\mathbf{u}(t)\| \leq G(T,\alpha,\beta) \quad \text{(kein Blow-up)}.$$

Fortsetzungssatz Satz 2.11: Die Lösung lässt sich bis auf den Rand von D, also ganz $\mathbb{R} \times \mathbb{R}^d$ fortsetzen, also existiert $\mathbf{u}$ für alle $t \geq t_0$.

Die Eindeutigkeit folgt aus Korollar 2.19. □

Beispiel 2.23

1) Sei $\mathbf{f}$ *global Lipschitz-stetig. Dann gilt*

$$\|\mathbf{f}(t,\mathbf{x})\| \leq \|\mathbf{f}(t,\mathbf{x}) - \mathbf{f}(t,\mathbf{0})\| + \|\mathbf{f}(t,\mathbf{0})\| \leq L\|\mathbf{x}\| + \|\mathbf{f}(t,\mathbf{0})\|\,.$$

Aufgrund von Korrolar 2.22 existiert dann eine globale und eindeutige Lösung.

2) Wir betrachten die lineare Anfangswertaufgabe

$$\mathbf{u}'(t) = \mathbf{A}(t)\,\mathbf{u}(t) + \mathbf{b}(t), \quad t \geq t_0, \quad \mathbf{u}(t_0) = \mathbf{u}_0,$$

mit stetigen Funktionen $\mathbf{A}(t) \in \mathbb{R}^{d\times d}$ *und* $\mathbf{b}(t) \in \mathbb{R}^d$ *für* $t \in [t_0, \infty]$.

Dann gilt:

$$\|\mathbf{f}(t,\mathbf{u})\| = \|\mathbf{A}(t)\,\mathbf{u} + \mathbf{b}(t)\| \leq \|\mathbf{A}(t)\,\mathbf{u}\| + \|\mathbf{b}(t)\| \leq \|\mathbf{A}(t)\|\,\|\mathbf{u}\| + \|\mathbf{b}(t)\|$$

Da $\|\mathbf{A}(t)\|$ *und* $\|\mathbf{b}(t)\|$ *als stetige Funktionen global beschränkt sind, existiert eine Lösung für alle* $t > t_0$. *Da* $\mathbf{f}$ *mit* $L = \|\mathbf{A}\|$ *Lipschitz-stetig ist, ist diese Lösung eindeutig.*

2.6. Globale Stabilität

Nach Satz 2.16 (lokaler Stabilitätssatz) wächst der Unterschied zwischen zwei Lösungen zu gestörten Anfangswerten oder bei Störung von $\mathbf{f}$ exponentiell mit der Länge des Zeitintervalls T. Für $T \to \infty$ wächst der Unterschied somit unbeschränkt (und sehr schnell). Bei bestimmten Annahmen an $\mathbf{f}$ lässt sich jedoch „globale Stabilität“ zeigen.

Definition 2.24 (Monotone Anfangswertaufgabe)

Die Funktion $\mathbf{f}(t,\mathbf{x})$ *genügt einer „Monotoniebedingung“, wenn mit* $\lambda(t) > 0$ *und* $\lambda \coloneqq \inf_{t\in I} \lambda(t) > 0$ *gilt:*

$$-\langle \mathbf{f}(t,\mathbf{x}) - \mathbf{f}(t,\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq \lambda(t)\|\mathbf{x}-\mathbf{y}\|^2, \quad (t,\mathbf{x}),(t,\mathbf{y}) \in D,\ \|\mathbf{x}-\mathbf{y}\| \neq 0.$$

Anmerkung 2.25

Die Monotoniebedingung verallgemeinert den Begriff „monoton fallend“ für vektorwertige Funktionen:

- *Für eine* montone skalare *Funktion* $f(x)$ *gilt:*

$$\begin{aligned} & -\langle f(x) - f(y), x-y\rangle = -\left(f(x) - f(y)\right)(x-y) \geq \lambda(t)\cdot(x-y)^2 \\ \iff & \underbrace{\frac{f(x)-f(y)}{x-y}}_{\text{Steigung}} \leq -\lambda < 0 \end{aligned}$$

bzw. $f' \leq -\lambda < 0$ *falls* f *differenzierbar.*

- *Für* lineares $\mathbf{f}(t, \mathbf{x}) = \mathbf{A}(t) \cdot \mathbf{x} + \mathbf{b}(t)$ *ist die Bedingung aus Definition 2.24*

$$\begin{aligned} & - \langle \mathbf{A}(t) \cdot \mathbf{x} - \mathbf{A}(t, \mathbf{y})\mathbf{y}, \mathbf{x} - \mathbf{y} \rangle = - \langle \mathbf{A}(t) \cdot (\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \\ \iff & - (\mathbf{x} - \mathbf{y})^T \cdot \mathbf{A}^T(t) \cdot (\mathbf{x} - \mathbf{y}) \geq \lambda(t) \|\mathbf{x} - \mathbf{y}\|^2 \end{aligned}$$

  $-\mathbf{A}^T(t)$ ist daher positiv definit, damit ist $\mathbf{A}^T(t)$ negativ definit. Da die Eigenwerte einer Matrix und ihrer Transponierten gleich ist, ist also auch $\mathbf{A}(t)$ negativ definit für alle t.

Eine spezielle (relativ starke) Form der Stabilität ist die

Definition 2.26 (Exponentielle Stabilität)
Eine globale Lösung $\mathbf{u}(t)$ einer AWA heißt „exponentiell stabil", wenn es $\delta, \alpha, \beta > 0$ gibt, so dass gilt:
Zu jedem Zeitpunkt $t_* \geq t_0$ und jedem $\mathbf{w}_* \in \mathbb{R}^d$ mit $\|\mathbf{w}_*\| < \delta$ hat die gestörte AWA

$$\mathbf{v}'(t) = \mathbf{f}(t, \mathbf{v}(t)), \quad t \geq t_*, \quad \mathbf{v}(t_*) = \mathbf{u}(t_*) + \mathbf{w}_*,$$

ebenfalls eine globale Lösung $\mathbf{v}(t)$, für die gilt

$$\|\mathbf{v}(t) - \mathbf{u}(t)\| \leq \beta e^{-\alpha(t - t_*)} \|\mathbf{w}_*\|, \quad t \geq t_*.$$

Bei einer Störung läuft man also wieder exponentiell auf die ungestörte Lösung zu.

Satz 2.27 (Globaler Stabilitätssatz)
Alle Lösungen einer L-stetigen und monotonen AWA sind global und exponentiell stabil mit δ beliebig, $\alpha = \lambda$ und $\beta = 1$. Im Fall $\sup_{t > t_0} \|\mathbf{f}(t, \mathbf{0})\| < \infty$ sind alle Lösungen gleichmäßig (d.h. unabhängig von t) beschränkt.

Beweis: Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen", Satz 1.7. □

Beispiel 2.28
Wir betrachten noch einmal die lineare Anfangswertaufgabe

$$\mathbf{u}'(t) = \mathbf{A}(t)\, \mathbf{u}(t) + \mathbf{b}(t), \quad t \geq t_0, \quad \mathbf{u}(t_0) = \mathbf{u}_0,$$

mit dabei sei $\mathbf{A}(t) \in \mathbb{R}^{d \times d}$ stetig und gleichmäßig (d.h. überall) negativ definit und $\mathbf{b}(t) \in \mathbb{R}^d$ stetig für $t \in [t_0, \infty]$.

(i) Wir haben bereits gezeigt, dass für eine entsprechende lineare Anfangswertaufgabe eine eindeutige globale Lösung existiert (Beispiel 2.23).

(ii) Da $\mathbf{A}(t)$ gleichmäßig negativ definit ist, ist die Anfangswertaufgabe monoton (siehe Anmerkung 2.25), daher ist die Lösung global und exponentiell stabil.

(iii) Da nach Voraussetzung

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

ist die Lösung nach Satz 2.27 gleichmäßig beschränkt.

2.7. Zusammenfassung

- Differentialgleichungen sind Gleichungen in denen neben einer Funktion $u(t)$ selbst auch deren Ableitungen vorkommen.
- Die höchste vorkommende Ableitung bestimmt die Ordnung der DGL.
- Ein Anfangswertproblem besteht aus einer DGL und vorgegebenen Werten für die ersten $n-1$ Ableitungen
- Falls n DGL für n Funktionen $u_i(t)$ $(1 \leq i \leq n)$ gegeben sind, die voneinander abhängen können, spricht man von einem System von DGL.
- In dieser Vorlesung behandeln wir vor allem Systeme gewöhnlicher Differentialgleichungen erster Ordnung in expliziter Form, d.h. man kann sie als $\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t))$ schreiben.
- Die Existenz einer Lösung ist bei DGL nicht selbstverständlich, ebensowenig dass es nur eine Lösung gibt (Eindeutigkeit).
- Es gibt eine lokale Lösung für $\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t))$, falls $\mathbf{f}(t, \mathbf{u}(t))$ in einem Bereich um $t, \mathbf{u}(t)$ stetig ist.
- Diese Lösung lässt sich fortsetzen, so lange $\mathbf{f}$ stetig bleibt und $\|\mathbf{u}(t)\|$ nicht gegen unendlich geht. Wenn dies für beliebiges $t \in \mathbb{R}$ möglich ist, gibt es eine „globale Lösung“.
- Zwei Lösungen mit verschiedenen Anfangswerten oder rechten Seiten können mit der Zeit exponentiell auseinander laufen (Lokaler Stabilitätssatz).
- Für die Existenz einer globalen Lösung reicht es zu zeigen, dass die Norm von $\mathbf{f}(t, \mathbf{x})$ nicht mehr als linear mit $\mathbf{x}$ wächst: $\mathbf{f}(t, \mathbf{x}) \leq \alpha(t)\|\mathbf{x}\| + \beta(t)$ mit beliebigen stetigen Funktionen $\alpha(t)$ und $\beta(t)$.
- Die Lösung ist eindeutig, wenn $\mathbf{f}(t, \mathbf{x})$ Liptschitz-stetig in $\mathbf{x}$ ist, wenn also $\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ mit einer sogenannten Lipschitz-Konstanten $L \in \mathbb{R}^+$. Für differenzierbare Funktionen ist das äquivalent dazu, dass die Ableitung nicht unendlich wird.
- Für sogenannte „monotone Anfangswertaufgabe“ (z.B. lineare DGL mit negativ definiter Matrix $\mathbf{A}$) ist die Lösung exponentiell stabil, d.h. Unterschiede in den Anfangswerten oder der rechten Seite werden mit der Zeit exponentiell gedämpft (statt exponentiell verstärkt im allgemeinen Fall).

3. Einschrittverfahren

Wir betrachten die Anfangswertaufgabe

$$\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t), \qquad t \in [t_0, t_0 + T] = I, \qquad \mathbf{u}(t_0) = \mathbf{u}_0. \tag{3.1}$$

Die Funktion $\mathbf{f}(t, \mathbf{x})$ sei stetig und erfülle eine globale Lipschitz-Bedingung

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|, \quad (t, \mathbf{x}), (t, \mathbf{y}) \in I \times \mathbb{R}^d. \tag{3.2}$$

Somit existiert eine eindeutige Lösung für alle $t \geq t_0$.

$T \to \infty$, also $I = [t_0, \infty[$ sei auch erlaubt.

Zeitgitter: Für $N \in \mathbb{N}$ wähle

$$t_0 < t_1 < \ldots < t_N = t_0 + T \qquad (T < \infty)$$

und setze:

$$I_n \coloneqq [t_{n-1}, t_n], \qquad h_n \coloneqq t_n - t_{n-1}, \qquad h \coloneqq \max_{1 \leq n \leq N} h_n.$$

3.1. Das explizite Eulerverfahren

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

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \, \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}^h), \qquad 1 \leq n \leq N \tag{3.3}$$

wobei der Startwert gegeben ist durch

$$\mathbf{y}_0^h = \mathbf{y}_0 \qquad \text{(muss nicht unbedingt } \mathbf{u}_0 \text{ sein, z.B. } \mathbf{y}_0 = \mathrm{rd}(\mathbf{u}_0)).$$

Es sei nun $\mathbf{u}(t)$ die Lösung der AWA (3.1). Die diskreten Werte der exakten Lösung für das gewählte Zeitgitter sind dann

$$\mathbf{u}_n^h \coloneqq \mathbf{u}(t_n) \quad 0 \leq n \leq N \qquad \text{(und damit } \mathbf{u}_0^h = \mathbf{u}(t_0) = \mathbf{u}_0)$$

Lokaler Diskretisierungsfehler

Für die Konvergenz der Einschrittverfahren erweist sich der sogenannte „Abschneidefehler“ als zentral.

Definition 3.1 (Abschneidefehler)
Als „lokalen Diskretisierungsfehler“ bzw. „Abschneidefehler“ $\boldsymbol{\tau}_n^h$ bezeichnen wir den Fehler, der innerhalb eines Verfahrensschritts von $\mathbf{y}_{n-1}$ nach $\mathbf{y}_n$ unter der Annahme auftritt, dass $\mathbf{y}_{n-1}$ exakte Daten sind.

$$\boldsymbol{\tau}_n^h \coloneqq h_n^{-1} \left[\mathbf{u}_n - \mathbf{y}_n(t_{n-1}, h_n, \mathbf{u}_{n-1})\right]$$

Für das explizite Euler-Verfahren erhält man:

$$\boldsymbol{\tau}_n^h = h_n^{-1}\left[\mathbf{u}_n^h - \left(\mathbf{u}_{n-1}^h + h_n\mathbf{f}(t_{n-1},\mathbf{u}_{n-1}^h)\right)\right] = h_n^{-1}\left[\mathbf{u}_n^h - \mathbf{u}_{n-1}^h\right] - \mathbf{f}(t_{n-1},\mathbf{u}_{n-1}^h).$$

Taylor-Entwicklung von $\mathbf{u}$ um den Punkt t_{n-1}:

$$\mathbf{u}(t_{n-1}+h) = \mathbf{u}(t_{n-1}) + h\mathbf{u}'(t_{n-1}) + \frac{h^2}{2}\mathbf{u}''(\xi), \quad \xi \in I_n$$

Wegen $\mathbf{u}'(t_{n-1}) = \mathbf{f}(t_{n-1},\mathbf{u}_{n-1}^h)$ gilt damit:

$$\boldsymbol{\tau}_n^h = h_n^{-1}\left[\mathbf{u}_{n-1}^h + h_n\mathbf{u}_{n-1}' + \frac{h_n^2}{2}\mathbf{u}''(\xi) - \mathbf{u}_{n-1}^h\right] - \mathbf{u}_{n-1}' = \frac{h_n}{2}\mathbf{u}''(\xi), \quad \xi \in I_n$$

und damit für die Norm

$$\|\boldsymbol{\tau}_n^h\| = \left\|\frac{h_n}{2}\mathbf{u}''(\xi)\right\|,\ \xi \in I_n \le \frac{1}{2}h_n \max_{t\in I_n}\|\mathbf{u}''(t)\| \tag{3.4}$$

- Diskretisierung „erster Ordnung“ in h.
- erfordert entsprechende „Regularität“ der Lösung, d.h. Beschränktheit der 2. Ableitung (siehe Polynominterpolation, Quadratur,...).

Definition 3.2 (Globaler Diskretisierungsfehler)
Sei $\mathbf{y}_n^h$ die mit einem Einschrittverfahren der Schrittweite h berechnete Näherung der Lösung $\mathbf{u}_n^h$ einer AWA unter der Annahme $\mathbf{y}_0^h = \mathbf{u}_0^h = \mathbf{u}_0$. Dann bezeichnen wir

$$\mathbf{e}_n^h = \mathbf{y}_n^h - \mathbf{u}_n^h$$

als globalen Diskretisierungsfehler.

Für das explizite Eulerverfahren erhalten wir dann

$$\begin{aligned}\mathbf{e}_n^h &= \underbrace{\mathbf{y}_{n-1}^h + h_n\mathbf{f}(t_{n-1},\mathbf{y}_{n-1}^h)}_{\text{Def. von } \mathbf{y}_n^h} - \underbrace{(\mathbf{u}_n^h - \underbrace{\mathbf{u}_{n-1}^h - h_n\mathbf{f}(t_{n-1},\mathbf{u}_{n-1}^h)}_{\text{fügen wir hinzu}})}_{=h_n\boldsymbol{\tau}_n^h} - \underbrace{\mathbf{u}_{n-1}^h - h_n\mathbf{f}(t_{n-1},\mathbf{u}_{n-1}^h)}_{\text{und ziehen wir wieder ab}} \\ &= \underbrace{\mathbf{y}_{n-1}^h - \mathbf{u}_{n-1}^h}_{\mathbf{e}_{n-1}^h} + h_n\left[\mathbf{f}(t_{n-1},\mathbf{y}_{n-1}^h) - \mathbf{f}(t_{n-1},\mathbf{u}_{n-1}^h)\right] - h_n\boldsymbol{\tau}_n^h\end{aligned}$$

Lipschitz-Bedingung und Dreiecksungleichung liefern:

$$\|\mathbf{e}_n^h\| \le \|\mathbf{e}_{n-1}^h\| + h_nL\,\|\mathbf{e}_{n-1}^h\| + h_n\|\boldsymbol{\tau}_n^h\|$$

Abspulen der Rekursion liefert:

$$\|\mathbf{e}_n^h\| \le \|\mathbf{e}_0^h\| + L\sum_{i=0}^{n-1} h_{i+1}\|\mathbf{e}_i^h\| + \sum_{i=1}^{n} h_i\|\boldsymbol{\tau}_i^h\| \tag{3.5}$$

Für eine weitergehende Abschätzung brauchen wir jetzt eine Aussage ähnlich dem Gronwall-Lemma, allerdings für den diskreten Fall.

Lemma 3.3 (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

$$w_0 \leq b_0 \text{ und } w_n \leq \sum_{i=0}^{n-1} a_i w_i + b_n, \qquad n \geq 1.$$

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

$$w_n \leq \exp\left(\sum_{i=0}^{n-1} a_i\right) b_n, \qquad n \geq 1.$$

Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Hilfssatz 2.1

Damit kann man nun den Konvergenzbeweis vollenden:

$$\begin{aligned}
\underbrace{\|\mathbf{e}_n^h\|}_{=w_n} &\underset{(3.5}{\leq} \sum_{i=0}^{n-1} \underbrace{L\, h_{i+1}}_{a_i} \underbrace{\|\mathbf{e}_i^h\|}_{w_i} + \underbrace{\sum_{i=1}^{n} h_i \|\boldsymbol{\tau}_n^h\| + \|\mathbf{e}_0^h\|}_{b_n} \\
&\underset{\text{Lemma 3.3}}{\leq} \exp\underbrace{\left(\sum_{i=0}^{n-1} L\, h_{i+1}\right)}_{\substack{L\,(t_n - t_0) \\ \text{(Teleskopsumme)}}} \cdot \left[\|\mathbf{e}_0^h\| + \sum_{i=1}^{n} \Big(h_i \underbrace{\|\boldsymbol{\tau}_n^h\|}_{\leq \max_{1\leq i\leq n} \|\boldsymbol{\tau}_n^h\|}\Big)\right] \\
&\leq e^{L\,(t_n - t_0)} \cdot \left[\|\mathbf{e}_0^h\| + \underbrace{\max_{1\leq i\leq n} \|\boldsymbol{\tau}_n^h\|}_{\leq \frac{h}{2} \max_{t\in I} \|\mathbf{u}''(t)\|} \underbrace{\sum_{i=1}^{n} h_i}_{\leq T}\right] \\
&\leq e^{L\,(t_n - t_0)} \cdot \left[\|\mathbf{e}_0^h\| + \frac{h}{2} \max_{t\in I} \|\mathbf{u}''(t)\| T\right]
\end{aligned}$$

und somit

$$\max_{1\leq n\leq N} \|\mathbf{e}_n^h\| \leq \underbrace{e^{LT}}_{\text{da } (t_n - t_0)\leq T} \left[\|\mathbf{e}_0^h\| + T \max_{1\leq i\leq n} \|\boldsymbol{\tau}_n^h\|\right] \leq e^{LT} \left[\|\mathbf{e}_0^h\| + \frac{T}{2} h \max_{t\in I} \|\mathbf{u}''(t)\|\right]. \quad (3.6)$$

- Die erste Abschätzung gilt unabhängig vom verwendeten (expliziten) Verfahren.
- Globale Konvergenzordnung ist gleich der lokalen Konvergenzordnung, beim expliziten Euler also $\mathcal{O}(h)$.
- Erfordert auch hier entsprechende Differenzierbarkeit der Lösung (Regularitätsannahme).
- e^{LT} ist in der Regel sehr pessimistisch (z.B. bei monotonen AWA).

Im Wesentlichen analog für das implizite Euler-Verfahren mit $\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n\, \mathbf{f}(t_n, \mathbf{y}_n^h)$.

3.2. Taylor-Verfahren

Ziel: Konstruktion von Einschritt-Verfahren höherer Ordnung.

Idee:

(a) Taylorentwicklung um $t-h$:

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

(b) Differenzieren der Differentialgleichung:

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

R-stufiges „Taylor-Verfahren“: Setze (b) in (a) ein.

$$\mathbf{u}(t) = \underbrace{\mathbf{u}(t-h)}_{r=0} + \sum_{r=1}^{R} \frac{h^r}{r!} \underbrace{\mathbf{f}^{(r-1)}(t-h, \mathbf{u}(t-h))}_{\text{(b) eingesetzt}} + \frac{h^{R+1}}{(R+1)!}\mathbf{u}^{(R+1)}(\xi), \qquad \xi \in [t-h, t].$$

Weglassen des Restgliedes ergibt:

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \cdot \underbrace{\sum_{r=1}^{R} \frac{h_n^{r-1}}{r!}\mathbf{f}^{(r-1)}(t_{n-1}, \mathbf{y}_{n-1}^h)}_{\substack{=:\mathbf{F}(h_n, t_{n-1}, \mathbf{y}_n^h, \mathbf{y}_{n-1}^h) \\ \text{„Verfahrensfunktion“}}} \tag{3.7}$$

Anmerkung 3.4

- *Auch „implizite“ Verfahren sind möglich.*

- *Die praktische Durchführung erfordert die manuelle Berechnung höherer Ableitungen von* $\mathbf{f}$. *Im* skalaren *Fall gilt:*

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

$$f^{(2)}(t, u(t)) = \frac{\mathrm{d}}{\mathrm{d}t} f^{(1)}(t, u(t)) = \Big[\underbrace{f_{tt} + f_{tx} f}_{\frac{\mathrm{d}}{\mathrm{d}t} f_t} + \underbrace{(f_{xt} + f_{xx} f) f + f_x (f_t + f_x f)}_{\frac{\mathrm{d}}{\mathrm{d}x} f_x f}\Big]\Big|_{(t,u(t))}$$

$$= \left[f_{tt} + 2 f_{tx} f + f_{xx} f^2 + f_x f_t + f_x^2 f\right]\Big|_{(t,u(t))}.$$

Im vektorwertigen Fall ($d > 1$) bzw. noch höherer Ordnung ist das noch sehr viel aufwändiger (kombinatorische Explosion).

3.3. Konvergenz allgemeiner Einschrittverfahren

Definition 3.5 (Konsistenz)
Ein Verfahren der Form

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \mathbf{F}(h_n, t_{n-1}, \mathbf{y}_n^h, \mathbf{y}_{n-1}^h) \tag{3.8}$$

heißt allgemeines „Einschrittverfahren". Dabei ist die Verfahrensfunktion $\mathbf{F}(h_n, t_{n-1}, \mathbf{y}_n^h, \mathbf{y}_{n-1}^h)$ *eine für das jeweilige Verfahren charakteristische Linearkombination von Auswertungen der rechten Seite* $\mathbf{f}$ *oder deren Ableitungen.*

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

$$\boldsymbol{\tau}_n^h := h_n^{-1}\left[\mathbf{u}_n^h - \left(\mathbf{u}_{n-1}^h + \mathbf{F}(h_n, t_{n-1}, \mathbf{u}_{n-1}^h, \mathbf{u}_n^h)\right)\right] = h_n^{-1}\left(\mathbf{u}_n^h - \mathbf{u}_{n-1}^h\right) - \mathbf{F}(h_n, t_{n-1}, \mathbf{u}_{n-1}^h, \mathbf{u}_n^h) \tag{3.9}$$

gilt

$$\lim_{h\to 0}\max_{t_n\in I}\|\boldsymbol{\tau}_n^h\| = 0 \qquad \textit{bzw.} \qquad \max_{t_n\in I}\|\boldsymbol{\tau}_n^h\| = \mathcal{O}(h^m) \textit{ für } h \to 0.$$

Wie man sich leicht überlegt, hat das R-stufige Taylorverfahren nach Konstruktion gerade die Konsistenzordnung $m = R$.

Satz 3.6
Sei

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \mathbf{F}(h_n, t_{n-1}, \mathbf{y}_{n-1}^h, \mathbf{y}_n^h), \qquad n \geq 0, \qquad \mathbf{y}_0^h = \mathbf{y}_0,$$

ein explizites oder implizites Einschrittverfahren zur Lösung der AWA

$$\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t)), \qquad t \geq t_0, \qquad \mathbf{u}(t_0) = \mathbf{u}_0.$$

$\mathbf{f}(t, \mathbf{x})$ *erfülle eine Lipschitzbedingung mit Konstante* L*, das Verfahren sei konsistent mit Ordnung* m*, d.h.* $\|\boldsymbol{\tau}_n^h\| \leq C h_n^m$*. Im impliziten Fall gelte für* $\mathbf{F}$ *die folgende Lipschitzbedingung*

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

sowie die Schrittweitenbedingung $h\,\tilde{L} \leq \frac{1}{2}$*.*
Dann gilt für den globalen Fehler

$$\max_{1\leq n\leq N}\|\mathbf{e}_n^h\| \leq e^{LT}\|\mathbf{u}_0 - \mathbf{y}_0\| + \frac{\alpha C h^m}{L}(e^{LT} - 1)$$

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

ohne Beweis.

Anmerkung 3.7 (Wichtige Bemerkungen zum Resultat)

- *Bei Einschrittverfahren folgt aus Konsistenz sofort Konvergenz.*
- *Die Ordnung bleibt dabei erhalten.*
- *Im expliziten Fall, genügt die Stabilität der Anfangswertaufgabe (Lipschitz-Bedingung an* $\mathbf{f}$*), im impliziten Fall gibt es zusätzlich eine Schrittweitenbedinung. Diese ist jedoch eher eine Folge der Beweistechnik und daher eine hinreichende aber keine notwendige Voraussetzung für Konvergenz.*

3.4. Runge-Kutta-Verfahren

Problem: Manuelle Ableitungsberechnung im Taylor-Verfahren ist viel zu aufwändig.

Idee: Ersetze Ableitungen durch numerische Approximation (Differenzenquotienten)

Beispiel 3.8
Ersetze erste Ableitung durch einseitigen Differentienquotienten:

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

Zwischenrechnung:

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

Wir setzen dies nun in das 2-stufige Taylor-Verfahren ein:

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

Somit ist das Verfahren

$$y_n^h = y_{n-1}^h + \frac{h}{2}f(t_{n-1},y_{n-1}^h) + \frac{h}{2}f(\underbrace{t_n}_{=},\underbrace{y_{n-1}^h + hf(t_{n-1},y_{n-1}^h)}_{\text{Näherung an der Stelle } t_n})$$

ein Verfahren der Konsistenzordnung 2 (Verfahren von Heun).

Definition 3.9 (Runge-Kutta-Verfahren)
Wir bezeichnen Verfahren der Form

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \sum_{i=1}^{s} b_i \cdot \mathbf{k}_i$$

mit

$$\mathbf{k}_i = \mathbf{f}(\tilde{t}_i,\tilde{\mathbf{y}}_i) \quad \textit{und} \quad \tilde{t}_i = t_{n-1} + c_i \cdot h_n$$

sowie

$$\tilde{\mathbf{y}}_i = \mathbf{y}_{n-1}^h + h_n \sum_{j=1}^{s} a_{ij}\, \mathbf{k}_j$$

als s-stufige Runge-Kutta-Verfahren.

Die Koeffizienten b_i, c_i und a_{ij} werden so gewählt, dass das Verfahren möglichst vorteilhafte Eigenschaften hat. Sie werden oft in einem sog. „Butcher-Tableau" dargestellt:

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array} \qquad \textit{bzw.} \qquad \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\mathsf{T} \end{array}$$

Falls $\mathbf{k}_1 = \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}^h)$ und $a_{ij} = 0$ für $i \leq j$ ($\mathbf{A}$ strikte untere Dreiecksmatrix) erhält man explizite Verfahren, bei denen $\mathbf{k}_i$ nur von bereits bekannten Zwischenwerten $\mathbf{k}_j$ mit $j < i$ abhängt.

Falls $a_{ij} = 0$ für $i < j$ ($\mathbf{A}$ untere Dreiecksmatrix) spricht man von diagonal-impliziten Runge-Kutta-Verfahren. Für diese ist die Lösung eines Gleichungssystems der Dimension d auf jeder Stufe notwendig.

Für a_{ij} beliebig ($\mathbf{A}$ voll besetzt) spricht man von voll-impliziten RK-Verfahren. Für diese Varianten ist die Lösung eines linearen Gleichungssystems für alle Stufen gleichzeitig (also der Dimension $s \cdot d$) notwendig.

3.4.1. Explizite Runge-Kutta-Verfahren

Die Konstruktion expliziter Verfahren erfolgt folgendermaßen:

Einstufiges-Verfahren ($\mathbf{s = 1}$):

$$\mathbf{k}_1 = \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}^h)$$
$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n b_1 \mathbf{k}_1 = \mathbf{y}_{n-1}^h + h_n b_1 \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}^h)$$

Zur Bestimmung von b_1 vergleichen wir mit der Taylorentwicklung von $\mathbf{u}$ um t_{n-1} (s.o.):

$$\mathbf{u}(t_n) = \mathbf{u}(t_{n-1}) + h_n \mathbf{f}(t_{n-1}, \mathbf{u}(t_{n-1})) + h_n^2 \frac{1}{2}(\mathbf{f}_t + \mathbf{f}_x\, \mathbf{f})(t_{n-1}, \mathbf{u}(t_{n-1})) + \mathcal{O}(h_n^3)$$

Bestimmung der Parameter durch Koeffizientenvergleich:

- Es ist maximal Konsistenzordnung $m = 1$ erreichbar (1 freier Parameter)
- Das einzige Verfahren ist $b_1 = 1$, also der explizite Euler.

Zweistufige Verfahren (s = 2): Wir betrachten nur den skalaren Fall.

$$k_2 = f(t_{n-1} + c_2 h_n, y^h_{n-1} + h_n a_{21} k_1)$$

Da f hier von h_n und k_1 abhängt, brauchen wir eine Taylorentwicklung einer Funktion mehrerer Argumente. Diese berechnet sich allgemein nach (vgl. IngMathe 2):

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{1}{2} \langle \mathbf{x} - \mathbf{x}_0, \mathbf{H}_f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) \rangle + \mathcal{O}(\|\mathbf{x} - \mathbf{x}_0\|^3)$$

Dabei ist $\langle ., . \rangle$ das Skalarprodukt zweier Vektoren, ∇f der Gradient und $\mathbf{H}_f$ die Hessenbergmatrix der Funktion f.

In unserem Fall reicht eine lineare Näherung von f, wir können den Term mit der Hessenbergmatrix also weglassen (bei Verfahren mit mehr als zwei Stufen kann man das nicht). $\mathbf{x} = (t_{n-1} + h_n c_2, y_{n-1^h} + h_n a_{21} k_1)^T$ und $\mathbf{x}_0 = (t_{n-1}, y_{n-1^h})^T$, die Änderung im Argument ist also $\mathbf{x} - \mathbf{x}_0 = (h_n c_2, h_n a_{21} k_1)^T$.

$$\begin{aligned} k_2 &= f(t_{n-1}, y^h_{n-1}) + \langle \begin{pmatrix} f_t(t_{n-1}, y^h_{n-1}) \\ f_x(t_{n-1}, y^h_{n-1}) \end{pmatrix}, \begin{pmatrix} h_n c_2 \\ h_n a_{21} k_1 \end{pmatrix} \rangle + \mathcal{O}(h_n^2) \\ &= f(t_{n-1}, y^h_{n-1}) + h_n c_2 f_t(t_{n-1}, y^h_{n-1}) + h_n a_{21} \underbrace{k_1}_{\searrow \text{einzige Änderung}} f_x(t_{n-1}, y^h_{n-1}) + \mathcal{O}(h_n^2) \\ &\overset{\substack{k_1 \text{von} \\ \text{oben} \\ \text{einsetzen}}}{=} f(t_{n-1}, y^h_{n-1}) + h_n c_2 f_t(t_{n-1}, y^h_{n-1}) + h_n a_{21} \overbrace{f(t_{n-1}, y^h_{n-1})} \cdot f_x(t_{n-1}, y^h_{n-1}) + \mathcal{O}(h_n^2) \end{aligned}$$

und hiermit (Argumente von $f, f_t, \ldots$ sind immer (t_{n-1}, y^h_{n-1}))

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

Vergleich mit

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

liefert die 3 Bedingungen

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

für die 4 Koeffizienten c_2, b_1, b_2, a_{21}.
$\Rightarrow$ Im Allgemeinen ist ein unterbestimmtes nichtlineares algebraisches System zu lösen.
Es gibt unendlich viele Lösungen, gebräuchliche Verfahren sind:

1) $b_1 = b_2 = \frac{1}{2}, \quad c_2 = 1, \quad a_{21} = 1 \rightarrow$ Verfahren von Heun

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

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

Dies entspricht der Anwendung einer genäherten Trapezregel zur Berechnung des Integrals in der Integralform.

2) $b_1 = 0, \quad b_2 = 1, \quad c_2 = \frac{1}{2} \quad a_{21} = \frac{1}{2} \rightarrow$ „modifizierter Euler“

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

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

Dies entstpricht einer genäherten Mittelpunktsregel, wird auch Halbschrittverfahren genannt.

Bemerkung: Ordnung $m = 3$ ist mit $s = 2$ Stufen nicht erreichbar.

Dreistufige Verfahren (s = 3):
Es gibt 8 Parameter und 6 Gleichungen. Übliche Verfahren sind:
Heun 3. Ordnung:

0	0		
$\frac{1}{3}$	$\frac{1}{3}$		
$\frac{2}{3}$	0	$\frac{2}{3}$	
	$\frac{1}{4}$	0	$\frac{3}{4}$

Kutta 3. Ordnung:

0	0		
$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Vierstufige Verfahren (s = 4):
13 Parameter, 11 Gleichungen
Klassisches Runge-Kutta Verfahren:

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

Anmerkung 3.10

- *Das Prinzip von Taylor-Verfahren lässt sich direkt auf Systeme übertragen, die Berechnung der Ableitungen ist jedoch noch aufwendiger.*

- *Runge-Kutta-Verfahren sind nicht unbedingt auf Systeme übertragbar, da hier die zu eliminierenden Ableitungen komplizierten sind und deshalb mehr Parameter benötigt werden, um die gleiche Ordnung zu erhalten.*

  Es gilt im Allgemeinen (Rannacher, R.: „Numerik 1: Numerik gewöhnlicher Differentialgleichungen“, Seite 51):

  $m \leq 4$ *Die Ordnung für den skalaren Fall überträgt sich auf Systeme.*

  $m > 4$ *Die Ordnung für Systeme ist in der Regel reduziert*

- *Die maximal erreichbare Ordnung für explizite, s-stufige RK-Verfahren (skalar) ist:*

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

  (Quelle: Simeon, Vorlesungsskript TUM).
  Dies erklärt die Beliebtheit des klassischen RK-Verfahrens 4. Ordnung.

- *Konstruktion von Verfahren höherer Ordnung mit*
  - *Butcher-Bäumen (systematische Darstellung der Ableitungen)*
  - *Computeralgebrasystemen.*

3.5. Schrittweitensteuerung

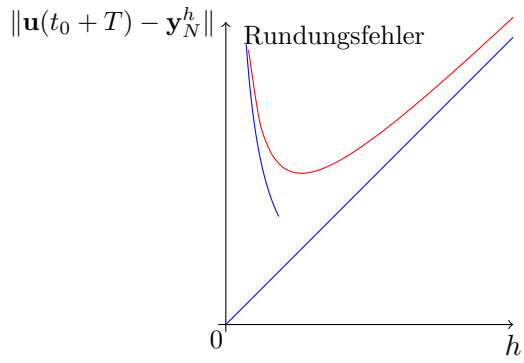
Bisher: Zeitpunkte t_n fest gewählt, z.B. äquidistant.
Dies ist ineffizient, wenn Lösung mit t stark variiert.

Ziel:

1) Erreiche vorgegebene Genauigkeit $\max\limits_{1\leq n\leq N} \|\mathbf{u}(t_n) - \mathbf{y}_n^h\| \leq \epsilon$

2) mit möglichst kleinem Aufwand.

Qualitatives Verhalten von $\|\mathbf{u}(t_0+T) - \mathbf{y}_N^h\|$:



$\Rightarrow$ h sollte für gegebene Fließkommazahlen eine untere Schranke nicht unterschreiten.

Die a priori Fehlerabschätzung ist zur Gittersteuerung nicht brauchbar, da e^{LT} viel zu pessimistisch und $\boldsymbol{\tau}_n^h$ höhere Ableitungen enthält.
Steuerung über „lokalen Fehler“:

- globaler Fehler ergibt sich durch Aufsummation der lokalen Abschneidefehler (gewichtet mit Stabilitätsfaktor)
- lokaler Fehler klein $\overset{\text{Hoffnung}}{\Rightarrow}$ globaler Fehler klein.

Sei $\tilde{\mathbf{u}}(t)$ die Lösung zum Startwert $\tilde{\mathbf{u}}(t_{n-1}) = \mathbf{y}^h_{n-1}$. Wegen

$$\mathbf{u}(t_n) = \mathbf{u}(t_{n-1}) + h_n \mathbf{F}(h_n, t_{n-1}, \mathbf{u}(t_{n-1})) + h_n \boldsymbol{\tau}^h_n$$

und

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

gilt für den lokalen Fehler:

$$\tilde{\mathbf{u}}(t_n) - \mathbf{y}^h_n = h_n \boldsymbol{\tau}^h_n \tag{3.10}$$

Für ein explizites Verfahren mit Konsistenzordnung m gilt:

$$\begin{aligned}
h_n \boldsymbol{\tau}^h_n &\overset{\text{(Def.)}}{=} \mathbf{u}(t_n) - \mathbf{u}(t_{n-1}) - h_n \mathbf{F}\left(h_n, t_{n-1}, \mathbf{u}(t_{n-1})\right) \\
&\overset{\downarrow}{=} \Bigg[\underbrace{\mathbf{u}(t_{n-1}) + h_n \sum_{i=1}^{m} \frac{h_n^{i-1}}{i!} \mathbf{f}^{(i-1)}\left(t_{n-1}, \mathbf{u}(t_{n-1})\right) + \mathbf{c}(t_{n-1}) h_n^{m+1} + \mathcal{O}(h_n^{m+2})}_{\text{Taylor-Entwicklung von } \mathbf{u}(t_n)} \Bigg] \\
&\quad - \mathbf{u}(t_{n-1}) - h_n \underbrace{\Bigg[\sum_{i=1}^{m} \frac{h_n^{i-1}}{i!} \mathbf{f}^{(i-1)}\left(t_{n-1}, \mathbf{u}(t_{n-1})\right) + \underbrace{\tilde{\mathbf{c}}(t_{n-1}) h_n^m + \mathcal{O}(h_m^{m+1})}_{\substack{=0 \text{ für Taylor, aber} \\ \neq 0 \text{ für RK}}} \Bigg]}_{\mathbf{F}(h_n, t_{n-1}, \mathbf{u}(t_{n-1})} \\
&= \underbrace{\mathbf{C}(t_{n-1})}_{=:\mathbf{c}(t_{n-1}) - \tilde{\mathbf{c}}(t_{n-1})} h_n^{m+1} + \mathcal{O}(h_n^{m+2})
\end{aligned} \tag{3.11}$$

$\mathbf{C}(t_{n-1})$ hängt von höheren Ableitungen von $\mathbf{f}$ ab, ist aber unabhängig von h_n. $\mathbf{C}(t_{n-1})$ ist unbekannt (bzw. viel zu kompliziert zu berechnen).

3.5.1. Schrittweitenschätzung mit Verfahren unterschiedlicher Ordnung

$\mathbf{y}^h_n$ werde erzeugt von Verfahren der Ordnung m
$\hat{\mathbf{y}}^h_n$ werde erzeugt von Verfahren der Ordnung $m+1$.
Dann ist nach einem Schritt mit gleichen Startwerten $\mathbf{y}^h_n$ wegen 3.11:

$$\begin{aligned}
\hat{\mathbf{y}}^h_n - \mathbf{y}^h_n &= \tilde{\mathbf{u}}(t_n) - \hat{\mathbf{C}}(t_{n-1}) h_n^{m+2} + \mathcal{O}(h_n^{m+3}) - \left(\tilde{\mathbf{u}}(t_n) - \mathbf{C}(t_{n-1}) h_n^{m+1} + \mathcal{O}(h_n^{m+2})\right) \\
&= \mathbf{C}(t_{n-1}) h_n^{m+1} + \mathcal{O}(h_n^{m+2})
\end{aligned}$$

Für h_n klein genug dominiert der führende Term und wir haben

$$\hat{\mathbf{y}}^h_n - \mathbf{y}^h_n \doteq \mathbf{C}(t_{n-1}) h_n^{m+1} \tag{3.12}$$

$$\Rightarrow \|\mathbf{C}(t_{n-1})\| \doteq \frac{\|\hat{\mathbf{y}}^h_n - \mathbf{y}^h_n\|}{h_n^{m+1}} \tag{3.13}$$

Wollen wir den Fehler in aktuellen Schritt beschränken, dann erhalten wir für die optimale Schrittweite h_{opt}

$$\|\tilde{\mathbf{u}}(t_n) - \mathbf{y}_n^h\| \doteq \|\mathbf{C}(t_{n-1})\| h_{\text{opt}}^{m+1} = \text{TOL}, \tag{3.14}$$

wobei die vorgegebene Toleranz TOL, deutlich größer als die Maschinengenauigkeit sein sollte, und nach Einsetzen von (3.12) und Auflösen nach h_{opt}:

$$\boxed{h_{\text{opt}} = h_n \cdot \left(\frac{\text{TOL}}{\|\hat{\mathbf{y}}_n^h - \mathbf{y}_n^h\|} \right)^{\frac{1}{m+1}}} \tag{3.15}$$

Eingebettete Runge-Kutta Verfahren

Die Fehlerschätzung wird effizienter, wenn die Berechnung von $\mathbf{y}_n^h$ keinen Zusatzaufwand bedeutet. Dies erreicht man durch Verwendung sogenannter „ eingebetteter“ Runge-Kutta Verfahren. Beide Verfahren nutzen dabei unterschiedliche Linearkombinationen derselben $\mathbf{k}_i$.

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \\ \hline \mathbf{b}^T & \\ \hline \hat{\mathbf{b}}^T & \end{array} \quad \begin{array}{l} \\ \\ \rightarrow \mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \sum_{j=1}^{s-1} b_j \mathbf{k}_j \\ \rightarrow \hat{\mathbf{y}}_n^h = \mathbf{y}_{n-1}^h + h_n \sum_{j=1}^{s} \hat{b}_j \mathbf{k}_j \end{array}$$

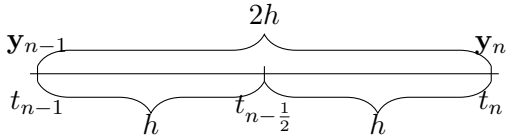
Runge-Kutta-Fehlberg Methode (1969):

$s = 6, \qquad m = 4, \qquad m+1 = 5, \quad \Rightarrow$ RKF45 Methode.

$$\begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{4} & \frac{1}{4} & & & & & \\ \frac{3}{8} & \frac{3}{32} & \frac{9}{32} & & & & \\ \frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & & & \\ 1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & & \\ \frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & \\ \hline & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\ \hline & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55} \end{array}$$

3.5.2. Schrittweitenschätzung mit Richardson-Extrapolation

Alternative Idee:



- Berechne $\mathbf{y}_n^h = \mathbf{y}(t_{n-1} + 2h)$ ausgehend von $\mathbf{y}_{n-1}^h$ mit zwei Schritten eines Verfahrens der Konsistenzordnung m zur Schrittweite h.
- Berechne $\mathbf{y}_n^{2h} = \mathbf{y}(t_{n-1} + 2h)$ mit einem Schritt desselben Verfahrens.
- Schätze den lokalen Fehler aus $\mathbf{y}_n^{2h} - \mathbf{y}_n^h$.

Mit Hilfe der Richardson-Extrapolation aus zwei Werten erhält man:

$$\tilde{\mathbf{u}}(t_n) - \mathbf{y}_n^h \approx \mathbf{y}_n^h - \mathbf{y}_n^{2h} = \|\mathbf{e}_{n-1}\| \, \mathcal{O}(2h) + \mathbf{C} \cdot \left(2h^{m+1} - (2h)^{m+1}\right) + \mathcal{O}((2h)^{m+2}) \quad (3.16)$$

Nun nimmt man an, dass:

- h so klein ist, dass der zweite gegen den dritten Term dominiert.
- $\|\mathbf{e}_{n-1}\| = 0$ ist (betrachte Fehler nach einem Schritt) oder alternativ $\|\mathbf{e}_{n-1}\| = \mathcal{O}(h^{m+1})$ und somit auch der zweite Term vom ersten dominiert wird.

Wir erhalten damit:

$$\|\mathbf{C}\| \doteq \frac{\|\mathbf{y}_n^h - \mathbf{y}_n^{2h}\|}{h_n^{m+1} \cdot 2 \cdot (2^m - 1)}$$

und durch Einsetzen in (3.14):

$$\boxed{h_{\text{opt}} = h_n \cdot \left(\frac{\text{TOL} \cdot 2 \cdot (2^m - 1)}{\|\mathbf{y}_n^h - \mathbf{y}_n^{2h}\|}\right)^{\frac{1}{m+1}}} \quad (3.17)$$

- Vorteile:
  - Funktioniert für jedes Verfahren (auch implizite) der Konsistenzordnung m.
  - Verwendung mit minimalen Änderungen am Code möglich.
  - Verbesserte Lösung der Ordnung $m + 1$ (bei entsprechender Regularität) aus:

$$\mathbf{y}_n = \frac{2^m \mathbf{y}_n^h - \mathbf{y}_n^{2h}}{2^m - 1}$$

- Nachteile:
  - Verfahren macht „2 Schritte der Schrittweite h auf einmal“.
  - Durch zusätzliche Berechnung eines Schritts der Schrittweite $2h$ alle zwei Schritte ist der Aufwand pro Schritt um 50% höher im Vergleich zur Berechnung ohne Fehlerschätzung.

3.5.3. Adaptiver Algorithmus zur Beschränkung des Gesamtfehlers

Wir möchten am liebsten nicht nur den Fehler in einem Zeitschritt kontrollieren, sondern den Gesamtfehler. Unter der Annahme exakter Arithmetik und fehlerfreier Startwerte lässt sich für das Zeitintervall $[t_0, T]$ die *a priori* Fehlerabschätzung

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

mit $K = e^{LT}$ zeigen. Oben haben wir außerdem berechnet, dass

$$h_n \tau_n^h \doteq \mathbf{C} h_n^{m+1} \iff \tau_n^h \doteq \mathbf{C} h_n^m$$

Wir wohlen den Fehler gleichmäßig über die Zeit verteilen. Dazu wählen wir die Schrittweite als

$$h_n \approx \left(\frac{\mathrm{TOL}}{KT\|\mathbf{C}\|} \right)^{1/m}$$

und erhalten damit

$$\max_{t_n \in I} \|e_n^h\| \leq K \sum_{n=1}^{N} h_n \|\tau_n^h\| \doteq K \sum_{n=1}^{N} h_n \|\mathbf{C}\| h_n^m = K \sum_{n=1}^{N} h_n \|\mathbf{C}\| \frac{\mathrm{TOL}}{KT\|\mathbf{C}\|} = \frac{\mathrm{TOL}}{T} \underbrace{\sum_{n=1}^{N} h_n}_{=T} = \mathrm{TOL}.$$

Der Gesamtfehler ist also unter den gegebenen Annahmen unterhalb der Toleranz.

Algorithmus 3.11
Eingabe: $\mathbf{y}_{n-1}^h$, *letzte Schrittweite* h_{n-1}, t_{n-1}
Ausgabe: $\mathbf{y}_n^h$, *berechnet mit neuer Schrittweite* h_n, t_n

1) *Setze* $h_n = h_{n-1}$ *(Anfangsschätzung der Schrittweite)*

2) *Berechne Näherung der Lösung und des optimalen Zeitschritts:*

   a) *Eingebettetes RK-Verfahren:*
   Berechne $\mathbf{y}_n^h$ *(mit Konsistenzordnung* m*) und* $\hat{\mathbf{y}}_n^h$ *(mit Konsistenzordnung* $m+1$*)*

   $$\|\mathbf{C}\| = \frac{\|\hat{\mathbf{y}}_n^h - \mathbf{y}_n^h\|}{h_n^{m+1}}$$

   b) *Richardson Extrapolation:*
   Berechne mit Verfahren der Konsistenzordnung m *die Werte* $\mathbf{y}^{2h}(t_{n+1})$ *und* $\mathbf{y}^h(t_n)$ *ausgehend von* $\mathbf{y}_{n-1}^h$.

   $$\|\mathbf{C}\| = \frac{\|\mathbf{y}_{n+1}^h - \mathbf{y}_{n+1}^{2h}\|}{h^{m+1} \cdot 2 \cdot (2^m - 1)}.$$

3) *Berechne Zeitschritt*

   $$h_{opt} = \left(\frac{TOL}{KT\|\mathbf{C}\|} \right)^{1/m}$$

   mit geforderter Toleranze TOL und geschätztem K.

4) *Falls* $h_{opt} < \alpha h_n$ *(z.B.* $\alpha = \frac{1}{4}$*) dann ist die Rechnung zu ungenau. Setze* $h_n = \alpha h_n$ *und gehe zu 2).*

5) Sonst akzeptiere Schritt:

$$t_n = t_{n-1} + h_n$$

$$\mathbf{y}_n = \begin{cases} \hat{\mathbf{y}}_n^h & \textit{Eingebettetes RK-Verfahren} \\ \frac{2^m \mathbf{y}_n^h - \mathbf{y}_n^{2h}}{2^m - 1} & \textit{Richardson-Extrapolation} \end{cases}$$

6) Beschränke Zeitschrittvergrößerung für nächsten Schritt:

$$h_{n+1} = \min(\beta \cdot h_n, h_{opt})$$

mit $\beta > 1$ (z.B. $\beta = 4$) und gehe zu 2)

Beispiel 3.12 (2-Körper-Problem)
$G = 1, m_1 = 1, m_2 = 0.01, x_1 = (-1, 0)^T, v_1 = 0, x_2 = (1, 0)^T, v_2 = (0, 1/5), I = [0, 100]$.
Mit konstantem Zeitschritt:

Methode	Δt	$\|e_0 - e_N\|/\|e_0\|$	$\#f$ *eval.*
Heun 2	$1/256$	$5.7 \cdot 10^{-1}$	51200
	$1/512$	$1.3 \cdot 10^{-1}$	102400
	$1/1024$	$1.8 \cdot 10^{-2}$	204800
	$1/2048$	$2.3 \cdot 10^{-3}$	409600
Runge-Kutta 4	$1/128$	$8.8 \cdot 10^{-2}$	51200
	$1/256$	$2.7 \cdot 10^{-3}$	102400
	$1/512$	$8.6 \cdot 10^{-5}$	204800
	$1/1024$	$2.7 \cdot 10^{-6}$	409600
	$1/2048$	$8.4 \cdot 10^{-8}$	819200

Mit adaptiver Zeitschrittsteuerung:

Methode	*TOL*	$\|e_0 - e_N\|/\|e_0\|$	$\#f$ *eval.*
RKF 45	10^{-3}	$4.4 \cdot 10^{-2}$	3282
	10^{-5}	$1.1 \cdot 10^{-3}$	5910
	$3 \cdot 10^{-8}$	$2.8 \cdot 10^{-6}$	16542
	10^{-10}	$2.0 \cdot 10^{-9}$	35694
RungeKutta4 extrapoliert	10^{-3}	$7.6 \cdot 10^{-2}$	10644
	10^{-5}	$1.3 \cdot 10^{-4}$	24348
	10^{-6}	$6.1 \cdot 10^{-6}$	40416
	$5 \cdot 10^{-7}$	$2.7 \cdot 10^{-6}$	47316
	$1.3 \cdot 10^{-9}$	$2.1 \cdot 10^{-9}$	217152
Heun2 extrapoliert	$5 \cdot 10^{-3}$	$5.9 \cdot 10^{-2}$	30318
	$5 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$	96798
	10^{-4}	$2.3 \cdot 10^{-4}$	215256
	$5 \cdot 10^{-6}$	$2.7 \cdot 10^{-6}$	1081812
	10^{-6}	$2.5 \cdot 10^{-7}$	2562834

3.6. Zusammenfassung

- Einschrittverfahren berechnen eine Näherungslösung $\mathbf{y}_n^h$ einer DGL zum Zeitpunkt $t_n = t_{n-1} + h_n$ ausgehend von einer Lösung $\mathbf{y}_{n-1}^h$ zum Zeitpunkt t_{n-1}.
- Ein allgemeines Einschrittverfahren hat die Form

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \mathbf{F}(h_n, t_{n-1}, \mathbf{y}_n^h, \mathbf{y}_{n-1}^h) \tag{3.18}$$

- Die lokale Genauigkeit eines Zeitschrittverfahrens wird durch den Abschneidefehler

$$\boldsymbol{\tau}_n^h \coloneqq h_n^{-1} \left(\mathbf{u}_n^h - \mathbf{u}_{n-1}^h \right) - \mathbf{F}(h_n, t_{n-1}, \mathbf{u}_{n-1}^h, \mathbf{u}_n^h)$$

  charakterisiert, d.h. den Fehler, der in einem Zeitschritt pro Zeit gemacht wird, wenn man mit einem exakten Wert startet.
- Ein Verfahren für das der lokale Abschneidefehler für $h \to 0$ gegen Null geht, heißt konsistent. Ein Verfahren hat die Konsistenzordnung m, wenn die Norm des Abschneidefehlers schneller als $\mathcal{O}(h^m)$ gegen Null geht.
- Das einfachste explizite Einschrittverfahren ist das explizite Eulerverfahren mit

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \, \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}^h)$$

  und Konsistenzordnung 1, das einfachste implizite Einschrittverfahren das implizite Eulerverfahren

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \, \mathbf{f}(t_n, \mathbf{y}_n^h)$$

  ebenfalls mit Konsistenzordnung 1.
- Der globale Fehler lässt sich für explizite Verfahren durch

$$\max_{1 \le n \le N} \|\mathbf{e}_n^h\| \le e^{LT} \left(\|\mathbf{e}_0^h\| + T \max_{1 \le i \le n} \|\boldsymbol{\tau}_n^h\| \right) \tag{3.19}$$

  abschätzen. Dabei ist L die Konstante aus der Lipschitz-Bedingung für $\mathbf{f}$. Der Verstärkungsfaktor e^{LT} ist jedoch im Allgemeinen sehr pessimistisch.
- Bei Einschrittverfahren folgt aus der Konsistenz sofort globale Konvergenz. Der globale Fehler lässt sich durch

$$\max_{1 \le n \le N} \|\mathbf{e}_n^h\| \le e^{LT} \|\mathbf{u}_0 - \mathbf{y}_0\| + \frac{\alpha C h^m}{L} (e^{LT} - 1)$$

  abschätzen (mit $\alpha = 1$ für explizite und $\alpha = 2$ für implizite Verfahren). Die Konvergenzordnung ist dabei gleich der Konsistenzordnung.
- Bei Taylor-Verfahren wird ein Verfahren der Ordnung R durch Taylorentwicklung der Funktion $\mathbf{u}(t)$ bis zur Ordnung R und Weglassen des Restgliedes konstruiert. n-te Ableitungen von $\mathbf{u}(t)$ können dabei als $(n-1)$-te Ableitungen der rechten Seite $\mathbf{f}$ berechnet werden. Das Problem bei Taylor-Verfahren ist die schwierige manuelle Berechnung höherer Ableitungen von $\mathbf{f}$, insbesondere bei Systemen von DGL (kombinatorische Explosion).

- Runge-Kutta-Verfahren kann man entweder als Näherung der Ableitungen durch numerische Approximation interpretieren. Ein s-stufiges Runge-Kutta-Verfahren hat die allgemeine Form

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \sum_{i=1}^{s} b_i \cdot \mathbf{k}_i$$

$$\mathbf{k}_i = \mathbf{f}(t_{n-1} + c_i \cdot h_n, \mathbf{y}_{n-1}^h + h_n \sum_{j=1}^{s} a_{ij}\, \mathbf{k}_j) \quad 1 \leq i \leq s.$$

  Dabei werden die Koeffizienten b_i, c_i und a_{ij} für verschiedene Verfahren so gewählt, dass die Verfahren möglichst vorteilhafte Eigenschaften haben (z.B. hohe Konsistenzordnung, besondere Stabilitätseigenschaften).

- Die Koeffizienten eines Runge-Kutta-Verfahren werden häufig in sogenannten Butcher-Tableaus angegeben:

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array} \qquad \text{bzw.} \qquad \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\mathsf{T} \end{array}$$

  Dabei erhält man explizite Verfahren, wenn $\mathbf{A}$ eine strikt untere Dreiecksmatrix ist (auf der Diagonalen und darüber stehen Nullen), diagonal-implizite Verfahren, wenn $\mathbf{A}$ eine untere Dreiecksmatrix ist (über der Diagonalen stehen Nullen) und voll-implizite Verfahren sonst.

- Verbreitete explizite Verfahren: Expliziter Euler (1. Ordnung, $s = 1$), Verfahren von Heun, modifiziertes Euler-Verfahren (beide 2. Ordnung, $s = 2$), Heun 3. Ordnung, Kutta 3. Ordnung (beide $s = 3$), klassisches Runge-Kutta-Verfahren 4. Ordnung ($s = 4$).

- Die maximal erreichbare Ordnung ist für $s \leq 4$ gleich der Stufenzahl. Danach liegt sie darunter.

- Die Ordnung von Runge-Kutta-Verfahren gilt für $m \leq 4$ auch für Systeme. Für $s > m > 4$ ist die Ordnung in der Regel für Systeme niedriger als für skalare Gleichungen.

- Die letzten beiden Beobachtungen erklären die Popularität des klassischen RK4-Verfahrens.

- Eine Kontrolle der Genauigkeit und eine Optimierung der Schrittweite ist durch Schätzung des lokalen Fehlers möglich. Dafür gibt es zwei Möglichkeiten:
  - Kombination von zwei Verfahren unterschiedlicher Ordnung:
    Dies geht am effizientesten mit sogenannten eingebetteten Runge-Kutta-Verfahren, z.B. Runge-Kutta-Fehlberg-Verfahren. Vorteil: Effiziente Berechnung. Nachteil: Funktioniert nur mit speziellen Verfahren.
  - Berechnung mit unterschiedlicher Schrittweite mit gleichem Verfahren und Fehlerschätzung durch Richardson-Extrapolation:
    Vorteil: Funktioniert mit beliebigen Einschrittverfahren (möglichst mit $m \geq 2$), auch impliziten. Nachteil: Mindestens 50 Prozent höherer Rechenaufwand.

4. Numerik steifer Differentialgleichungen

4.1. Motivation

Beispiel 4.1 (Modellproblem)

$$u' = \lambda u, \qquad \lambda \in \mathbb{R}, \quad \lambda < 0 \tag{4.1}$$

mit exakter Lösung $u(t) = u_0 e^{\lambda t}$.

	$\lambda = -10$	$\lambda = -100$
expliziter Euler	$\Delta t = 0.1 \to 1, 0, 0, \ldots$ $\Delta t = 0.2 \to$ *Oszillation* $\Delta t > 0.2$ (*z.B.* $\Delta t = 0.3$) $\to$ *explodiert*	$\Delta t = 0.01 \to 1, 0, 0,$ $\Delta t = 0.02 \to$ *Oszillation* $\Delta t > 0.02 \to$ *explodiert*
impliziter Euler	$y_n^h \to 0$ *für* $t \to \infty$ *für alle* Δt	*dito*

Für das Modellproblem (4.1) und die beiden Euler-Verfahren können wir das Verhalten auch theoretisch verstehen:

Expliziter Euler:

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

also

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

und

$$|y_n^h| = |1 + h\lambda|^n \, |u_0|$$

Mit der Voraussetzung $\lambda < 0$ folgt

$$\begin{aligned} |1 + h\lambda|^n \leq 1 &\iff |1 + h\lambda| \leq 1 \\ &\iff -1 \leq \underbrace{1 + h\lambda \leq 1}_{\text{immer erfüllt, da } \lambda<0} \\ &\Rightarrow -2 \leq h\lambda \\ &\iff \boxed{-\frac{2}{\lambda} \geq h} \end{aligned}$$

Somit ist für $\lambda = -10$ der Wert $h = 0.2$ genau die Stabilitätsgrenze.
Wegen $|f(t,x) - f(t,y)| = |\lambda\,(x - y)| \leq |\lambda| \, |x - y|$ ist $L = |\lambda|$ die Lipschitzkonstante und wir können die Bedingung auch als $hL \leq 2$ schreiben.

Impliziter Euler

$$\begin{aligned} \frac{y_n^h - y_{n-1}^h}{h} = \lambda y_n^h &\iff (1 - h\lambda) y_n^h = y_{n-1}^h \\ &\iff y_n^h = \frac{1}{1 - h\lambda} y_{n-1}^h \end{aligned}$$

also

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

und

$$|y_n^h| = \frac{1}{|1-h\lambda|^n}\, u_0.$$

Wir erhalten

$$\frac{1}{|1-h\lambda|^n} \leq 1 \iff \frac{1}{|1-h\lambda|} \leq 1 \iff 1 \leq |1-h\lambda| \text{ was wegen } \lambda < 0 \text{ für alle } h \geq 0 \text{ erfüllt ist.}$$

$\Rightarrow$ Für den impliziten Euler ist keine Schrittweitenbedingung erforderlich.

Was versteht man nun unter einer „steifen“ AWA?
Hier gibt es keine eindeutige Definition. In der Literatur findet man:

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

2) Ein lineares System $\mathbf{u}' = \mathbf{A}\mathbf{u}$ mit $\mathrm{Re}(\lambda_i) < 0$ für alle Eigenwerte λ_i heißt steif, wenn

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

   Im nichtlinearen Fall betrachtet man die Eigenwerte der Jacobimatrix $\mathbf{f}_{\mathbf{x}}(t, \mathbf{u}(t))$.

   Anschaulich betrachtet heißt das, dass es Prozesse gibt, die auf sehr unterschiedlichen Zeitskalen ablaufen (oder auf sehr unterschiedlichen Entfernungen, wenn die DGL einen Prozess im Ort beschreibt).

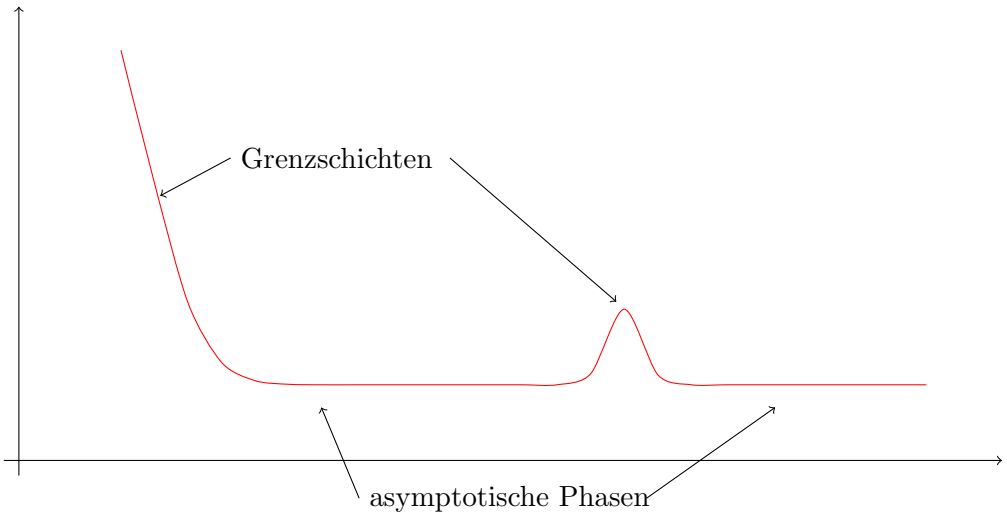
3) Differentialgleichungen der Form

$$\begin{aligned} \mathbf{u}' &= \mathbf{f}(t, \mathbf{u}, \mathbf{z}) \\ \varepsilon\, \mathbf{z}' &= \mathbf{g}(t, \mathbf{u}, \mathbf{z}) \end{aligned}$$

   mit $\varepsilon \in \mathbb{R}, \varepsilon \ll 1$ heißen „singulär gestört“ und führen für $\epsilon \to 0$ auf steife Anfangswertaufgaben.

   Es gibt also einen (schnelleren) Prozess, der sehr wenig zur Lösung beiträgt, aber den Lösungsprozess stört.

Qualitatives Lösungsverhalten:



In den asymptotischen Phasen möchte man große Zeitschritte zulassen. Bei steifen Problemen bestimmt aber der schnellste Prozess den notwendigen Zeitschritt, auch wenn dieser gerade keinen Einfluss auf die Lösung hat.

4.2. Modellproblemanalyse (skalar, linear)

Zur Bewertung verschiedener Verfahren betrachten wir das Modellproblem

$$u'(t) = \lambda u(t), \quad t \geq 0, \quad u(0) = u_0, \quad \lambda \in \mathbb{C}, \quad \mathrm{Re}(\lambda) \leq 0. \tag{4.2}$$

Wir wir in der großen Übung gesehen haben, lässt sich die Lösung für alle linearen DGL-Systeme erster Ordnung, als Linearkombination von Funktionen dieser Form zusammensetzen. DGL höherer Ordnung lassen sich durch Transformation auf Systeme erster Ordnung bringen und nicht-lineare DGL lokal durch Linearisierung (Jacobi-Matrix).. Die Analyse des Modellproblems untersucht also ein Lösungsverhalten, das prinzipiell bei allen Arten von steifen DGL auftreten kann.

Definition 4.2 (Absolute Stabilität)
Eine Einschrittmethode heißt „absolut stabil“ für ein $h\lambda \neq 0$, wenn durch ihre Anwendung auf das skalare Modellproblem (4.2) für $Re(\lambda) \leq 0$ beschränkte Näherungen erzeugt werden: $\sup_{n\geq 0} \|\mathbf{y}_n^h\| < \infty$.

Für das explizite Euler-Verfahren gilt

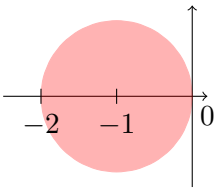
$$y_n^h = \underbrace{(1 + h\lambda)}_{=:\omega(h\lambda)} y_{n-1}^h = \omega(h\lambda)\, y_{n-1}^h$$

Die Methode ist absolut stabil für $|\omega(h\lambda)| \leq 1$.
Die Menge

$$\mathrm{SG} := \{z = \lambda h \in \mathbb{C} \colon |\omega(z)| \leq 1\}$$

heißt „Stabilitätsgebiet“ einer Einschrittformel. Das Stabilitätsgebiet beinhaltet die Menge aller Punkte $z = h\lambda$ für die das Verfahren stabil ist. Ein großes Stabilitätsgebiet bedeutet

also, dass bei gleichem λ größere Zeitschritte möglich sind. Da wir nur an Fällen interessiert sind, bei denen $Re(\lambda) \leq 0$, sind nur die Werte auf der linken Halbebene der komplexen Zahlen interessant.
Für das explizite Euler-Verfahren gilt $\mathrm{SG}_{EE} = \{z \in \mathbb{C} \colon |1+z| \leq 1\}$



$|z| \leq 1$: Einheitskreisscheibe.
$|1+z| \leq 1$: um 1 nach links verschobene Einheitskreisscheibe.
Formel:
$|1+a+ib| = \sqrt{(1+a)^2+b^2} \leq 1 \iff (1+a)^2+b^2 \leq 1$

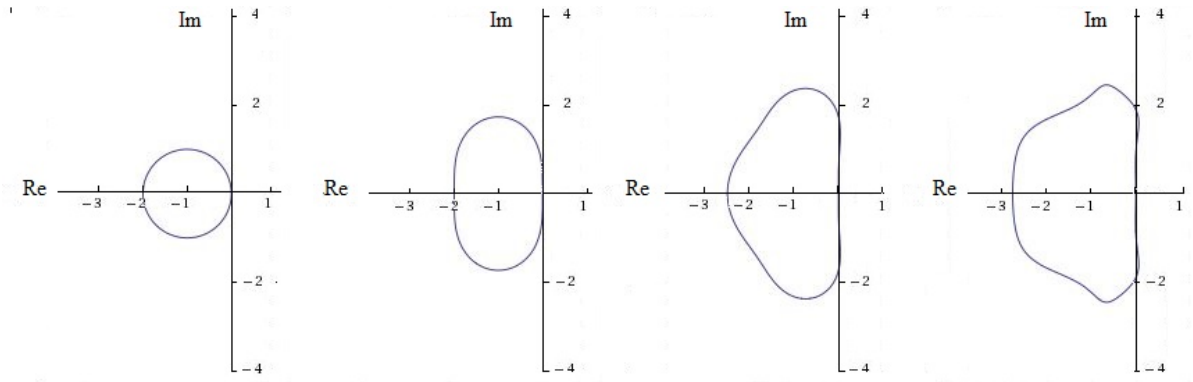
Für das Taylor-Verfahren der Stufe R gilt

$$
\begin{aligned}
y_n^h &= y_{n-1}^h + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t_{n-1}, y_{n-1}^h) \\
&= y_{n-1}^h + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} \lambda^r y_{n-1}^h = \left(\underbrace{1}_{=\frac{(h\lambda)^0}{0!}} + \sum_{r=1}^{R} \frac{(h\lambda)^r}{r!} \right) y_{n-1}^h \\
&= \underbrace{\sum_{r=0}^{R} \frac{(h\lambda)^r}{r!}}_{=:\omega(h\lambda)} y_{n-1}^h
\end{aligned}
$$

Wobei oben die folgenden Gleichungen benutzt werden:

$$
\begin{aligned}
f(t,u(t)) &= \lambda u(t) \\
f^{(1)}(t,u(t)) &= \lambda u'(t) = \lambda f(t,u(t)) = \lambda^2 u(t) \\
f^{(2)}(t,u(t)) &= \lambda f^{(1)}(t,u(t)) = \lambda^3 u(t) \\
\text{und damit } f^{(j)} &= \lambda^{j+1} u(t)
\end{aligned}
$$

Die Bestimmung der Stabilitätsgebiete ist schwierig, sie sehen so aus (für $R = 1, \ldots, 4$):



Einfacher ist das „Stabilitätsintervall“

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

$$\text{SI} = \begin{cases} [-2,0], & R = 1 \\ [-2,0], & R = 2 \\ [-2,51\ldots,0], & R = 3 \\ [-2,78\ldots,0], & R = 4 \end{cases}$$

RK-Verfahren mit $s = R \leq 4$ und optimaler Konsistenzordnung haben den gleichen Verstärkungsfaktor $\omega(z)$. Für die expliziten RK-Verfahren ist deshalb bei Anwendung auf das Testproblem (4.2) ebenfalls eine Schrittweitenbeschränkung erforderlich. Diese ist umso weniger restriktiv, je weiter das Stabilitätsgebiet in der komplexen Ebene nach links reicht.
Optimal wäre

Definition 4.3 (A-Stabilität)
Ein Einschrittverfahren heißt „A-stabil“, wenn das zugehörige Stabilitätsgebiet die ganze linke Halbebene der komplexen Zahlen umfasst:

$$\mathbb{C}^- := \{z \in \mathbb{C} \colon Re(z) \leq 0\} \subset SG.$$

Wir haben also

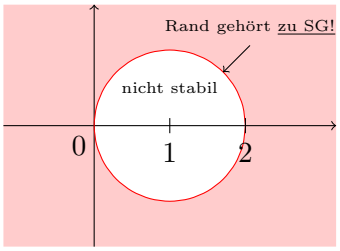
$$\text{A-stabil} \iff |\omega(z)| \leq 1 \ \forall\, z \in \mathbb{C}^- \iff \mathbb{C}^- \subset SG$$

A-stabile Verfahren lassen bei steifen Differentialgleichungen beliebig große Schrittweiten zu, ohne dass die Lösung gegen $\pm\infty$ geht.

Gehen wir zurück zum impliziten Euler-Verfahren.
Dort gilt

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

$\omega(z) = \frac{1}{1-z}$ ist rational und es gilt $\lim_{z\to-\infty} \frac{1}{1-z} = 0$.
Für das Stabilitätsgebiet ergibt sich



denn
$\left|\frac{1}{1-a+ib}\right| = \frac{1}{\sqrt{(1-a)^2+b^2}} \leq 1 \iff (1-a)^2 + b^2 \geq 1$
$x^2 + y^2 \geq 1$ sind alle Punkte nicht
im Inneren des Einheitskreises.
$(1-x)^2 + y^2$ ist eine Verschiebung um Eins nach rechts.

Damit ist der implizite Euler A-stabil (Achtung: die gesamte imaginäre Achse gehört zum SG!).
Für Taylor- und Runge-Kutta-Verfahren ist $\omega(z)$ ein Polynom in z (Taylor: $\omega(z) = \sum_{r=0}^{R} \frac{z^r}{r!}$).
Ein Polynom vom Grad > 0 hat immer die Eingenschaft $\lim\limits_{z\to-\infty} \omega(z) = \pm\infty$ und somit kann nicht $\mathbb{C}^- \subset$ SG gelten.

$\Rightarrow$ Es gibt kein A-stabiles Taylor bzw. explizites RK-Verfahren, d.h. diese Verfahren benötigen immer eine Schrittweitenbeschränkung!

Für unser Modellproblem (4.2) und ähnliche Probleme sollte die Lösung nicht nur beschränkt bleiben, sondern gegen Null gehen für $h \to \infty$. Wir definieren deshalb

Definition 4.4 (L-Stabil)
Ein Einschrittverfahren heißt L-stabil (auch stark A-stabil), falls es A-stabil ist und $\lim_{z\to-\infty} |\omega(z)| = 0$ *gilt.*

4.3. Implizite Runge-Kutta-Verfahren

Implizite Runge-Kutta-Verfahren (IRK) sind definiert über die Lösung des nichtlinearen Systems:

$$\mathbf{k}_i = \mathbf{f}\left(t_{n-1} + c_i h_n, \mathbf{y}_{n-1}^h + h_n \sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right) \qquad i = 1, \ldots, s \tag{4.3}$$

und anschließende Berechnung der „neuen“ Lösung

$$\mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \sum_{i=1}^{s} b_i \mathbf{k}_i. \tag{4.4}$$

Neben der Herleitung über eine Taylorentwicklung kann man RK-Verfahren auch über Quadraturformeln entwickeln. Ausgehend von

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

berechne das Integral durch eine Quadraturformel:

$$\mathbf{u}(t_n) = \mathbf{u}(t_{n-1}) + h_n \sum_{i=1}^{s} b_i \mathbf{f}(t_{n-1} + c_i h_n, \mathbf{u}(t_{n-1} + c_i h_n)) + \mathcal{O}(h_n^{m+1})$$

mit geeigneten Gewichten b_i und Zwischenstellen c_i.

Typische Implizite Runge-Kutta-Verfahren

(a) Impliziter Euler als IRK

$$\frac{\mathbf{y}_n^h - \mathbf{y}_{n-1}^h}{h_n} = \mathbf{f}(t_n, \mathbf{y}_n^h) \iff \mathbf{y}_n^h = \mathbf{y}_{n-1}^h + h_n \underbrace{\mathbf{f}(t_n, \mathbf{y}_n^h)}_{=:\mathbf{k}_1}$$

also $\mathbf{k}_1 = \mathbf{f}(t_n, \mathbf{y}_n^h) = \mathbf{f}(t_{n-1} + h_n, \mathbf{y}_{n-1}^h + h_n\mathbf{k}_1)$

Butcher-Tableau:

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

Das Implizite-Euler-Verfahren ist L-stabil.

(b) Einschritt-θ-Verfahren. Für $\theta \in [0,1]$

$$\begin{aligned}\mathbf{u}(t_n) &= \mathbf{u}(t_{n-1}) + \int_{t_{n-1}}^{t_n} \mathbf{f}(t, \mathbf{u}(t))\,\mathrm{d}t \\ &= \mathbf{u}(t_{n-1}) + h_n\big[(1-\theta)\underbrace{\mathbf{f}(t_{n-1}, \mathbf{u}(t_{n-1}))}_{\mathbf{k}_1} + \theta\underbrace{\mathbf{f}(t_n, \mathbf{u}(t_n))}_{\mathbf{k}_2}\big] + \begin{cases}\mathcal{O}(h_n^2) & \theta \neq \frac{1}{2} \\ \mathcal{O}(h_n^3) & \theta = \frac{1}{2}\end{cases}\end{aligned}$$

also

$$\begin{aligned}\mathbf{k}_1 &= \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}^h) \\ \mathbf{k}_2 &= \mathbf{f}\left(t_{n-1} + h_n, \mathbf{y}_{n-1}^h + h_n\left[(1-\theta)\mathbf{k}_1 + \theta\mathbf{k}_2\right]\right) \\ \mathbf{y}_n^h &= \mathbf{y}_{n-1}^h + h_n\left[(1-\theta)\mathbf{k}_1 + \theta\mathbf{k}_2\right]\end{aligned}$$

Butcher-Tableau:

0	0	0
1	$1-\theta$	θ
	$1-\theta$	θ

Konsistenzordnung:

2 für $\theta = \frac{1}{2}$(„Trapezregel“)
1 sonst ($\theta = 0$: Expliziter Euler, $\theta = 1$: Impliziter Euler) .

A-stabil falls $\theta \geq 0.5$, L-stabil falls $\theta = 1$.

(c) Mittelpunktsregel

$$\begin{aligned}\mathbf{y}_n^h &= \mathbf{y}_{n-1}^h + h_n \underbrace{\mathbf{f}(t_{n-1} + \frac{h_n}{2}, \mathbf{y}_{n-1}^h + \frac{h_n}{2}\mathbf{k}_1)}_{\mathbf{k}_1} \\ \mathbf{k}_1 &= \mathbf{f}\big(t_{n-1} + \frac{h_n}{2}, \mathbf{y}_{n-1}^h + \frac{h_n}{2}\mathbf{k}_1\big)\end{aligned}$$

Butcher-Tableau:

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

Konsistenzordnung 2. A-stabil.

(d) Gauß-Verfahren

Wähle c_i als Nullstellen der Legendre-Polynome (Gauß-Quadratur ist normalerweise auf dem Intervall $[-1,1]$ definiert, transformiere auf $[0,1]$).

Die Konvergenzordnung $m = 2s$ ist optimal. Die Verfahren sind A-stabil, aber nicht L-stabil $\lim_{t\to\infty} \omega(z) = 1$, gut geeignet für Schwingungsprobleme.
$s = 1$: Implizite Mittelpunktregel
$s = 2$ (Konsistenzordnung 4):

$1/2 - \sqrt{3}/6$	$1/4$	$1/4 - \sqrt{3}/6$
$1/2 + \sqrt{3}/6$	$1/4 + \sqrt{3}/6$	$1/4$
	$1/2$	$1/2$

(e) Radau-Verfahren

Es gibt es zwei Klassen IA und IIA von Radau-Verfahren:

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

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

Konsistenzordnung $2s-1$, L-stabil

$s = 1$: impliziter Euler

$s = 2$ (Konsistenzordnung 3):

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

(f) Lobatto-Regeln

Hier fordert man $c_1 = 0$, $c_s = 1$. Die Konsistenzordnung ist $2s-2$.
$s = 2$ (Konsistenzordnung 2):

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

Verfahren von Heun, explizit, nicht stabil.

$s = 3$ (Konsistenzordnung 4):

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

Erste Zeile, letzte Spalte immer Null $\Rightarrow$ 2 explizite Schritte. Stabilität?

(g) DIRK-Verfahren (DIRK = diagonally implicit Runge-Kutta method)

Die Matrix $\mathbf{A}$ im Butcher-Tableau ist eine untere Dreiecksmatrix $\Rightarrow$ In jeder Stufe ist ein nichtlineares System mit Dimension d (statt $s \cdot d$) zu lösen.

SDIRK = singly DIRK

$$a_{11} = a_{22} = \cdots = a_{ss}$$

Im linearen Fall besitzt das System in jeder Stufe dieselbe Matrix $\Rightarrow$ nur eine LR-Zerlegung nötig.

$s = 2$ (Konsistenzordnung 2): Alexander-Verfahren

$$
\begin{array}{c|cc|l}
\alpha & \alpha & 0 & \alpha \quad = 1 \pm \frac{\sqrt{2}}{2} \\
1 & 1-\alpha & \alpha & \\
\hline
 & 1-\alpha & \alpha & \leftarrow b^T \quad = \text{letzte Zeile von } \mathbf{A} \Rightarrow \mathbf{y}_n^h = \mathbf{k}_2 \text{ (weitere Einsparung)}
\end{array}
$$

L-stabil.

$s = 2$ (Konsistenzordnung 3): Crouzieux-Verfahren

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

A-stabil.

4.4. Zusammenfassung

- Man spricht von „steifen“ Anfangswertaufgaben, wenn explizite Verfahren sehr kleine Zeitschritte einsetzen müssen, obwohl dies aus Genauigkeitsgründen nicht notwendig wäre, während bestimmte implizite Verfahren große Zeitschritte einsetzen können. Dies ist insbesondere dann der Fall, wenn Prozesse auf sehr unterschiedlichen (Raum- oder Zeit-) Skalen stattfinden.
- Die Stabilität von verschiedenen Verfahren untersucht man anhand des sogenannten Modellproblems $u'(t) = \lambda u(t)$ mit $t \geq 0$, $u(0) = u_0$ und $\lambda \in \mathbb{C}$, $\mathrm{Re}(\lambda) \leq 0$.
- Wir bezeichnen ein Einschrittverfahren als „A-stabil“, wenn ihre Anwendung auf das Modellproblem für beliebig große Zeitschritte beschränkte Näherungen erzeugt und als „L-stabil“, wenn das Verfahren A-stabil ist und die Lösung für immer größere Zeitschritte sogar gegen Null geht.
- Es gibt keine A-stabilen expliziten Runge-Kutta-Verfahren.
- Implizite Runge-Kutta-Verfahren leitet man am leichtesten durch Anwendung einer Quadraturformel zur Berechnung von $\mathbf{u}(t_n) = \mathbf{u}(t_{n-1}) + \int_{t_{n-1}}^{t_n} \mathbf{f}(t, \mathbf{u}(t))$ her.
- Die maximale Konsistenzordnung von A-stabilen impliziten Runge-Kutta-Verfahren ist $2s$ (Gauß-Verfahren). Die maximale Ordnung von L-stabilen impliziten Runge-Kutta-Verfahren ist $2s - 1$ (Radau-Verfahren).
- Typische implizite Runge-Kutta-Verfahren sind:
  - Impliziter Euler ($s = 1$, 1. Ordnung), L-stabil
  - Einschritt-θ-Verfahren ($s = 2$, 1. oder 2. Ordnung, abhängig von θ). A-stabil falls $\theta \geq 0.5$.
  - Mittelpunktsregel ($s = 2$, 2. Ordnung)
  - Gauß-Verfahren (Ordnung $2s$): A-stabil
  - Radau IA oder IIA-Verfahren (Ordnung $2s - 1$): L-stabil

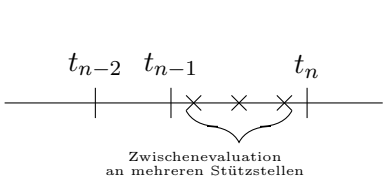
  - Lobatto-Regeln (Ordnung $2s - 2$), teilweise explizit, Stabilität unklar.
  - Diagonal-Implizite-Runge-Kutta-Verfahren (DIRK): z.B. Alexander-Verfahren ($s = 2$, 2. Ordnung, L-stabil), Crouzieux-Verfahren ($s = 2$, 3. Ordnung, A-stabil).

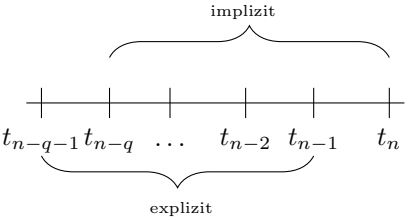
- Implizite Verfahren erfordern die Lösung einer linearen oder nichtlinearen Gleichung. Nichtlineare Gleichungen oder Gleichungssysteme löst man mit dem Newton-Verfahren oder einer Fixpunktiteration.

- Die höhere Stabilität impliziter Verfahren zahlt sich vor allem bei steifen DGL aus. Bei DGL, die nicht zu stabilen Lösungen im Sinne von Satz 2.27 führen, lohnt sich der Mehraufwand in der Regel nicht.

5. Mehrschrittverfahren

Einschrittverfahren (ESV): Mehrschrittverfahren (MSV):





nutze q letzte Zeitschritte
Ziel: niedrigerer Aufwand pro Genauigkeit
wegen nur 1 f-Auswertung pro h_n-Schritt
und trotzdem hoher Genauigkeit.
(ob besser → Fehlerkonstante)

- Wir beschränken uns auf äquidistante Stützstellen $t_n = t_0 + nh$
- Variable Schrittweite möglich, aber wesentlich aufwändiger als bei ESV.

5.1. Integrationsbasierte Verfahren

Wir wollen $\mathbf{u}(t_n)$ ausgehend von $\mathbf{u}(t_{n-\sigma})$ mit einem gewählten $\sigma \geq 1$ durch Integration berechnen:

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

Um die Quadratur durchführen zu können interpolieren wir $\mathbf{f}$ an äquidistanten Stellen t_{k-q} bis t_k mit einem Polynom vom Grad q und integrieren das Polynom (Newton-Cotes Quadratur). t_k ist entweder gleich t_{n-1} (explizite Verfahren) oder t_n (implizite Verfahren).

$$\mathbf{p}_q(t) = \sum_{\mu=0}^{q} \mathbf{f}(t_{k-\mu}, \mathbf{u}(t_{k-\mu}))L_\mu^{(q)}(t), \qquad L_\mu^{(q)}(t) = \prod_{\substack{\ell=0 \\ \ell\neq\mu}}^{q} \frac{t - t_{k-\ell}}{t_{k-\mu} - t_{k-\ell}}$$

Es gilt für den Interpolationsfehler

$$\mathbf{f}(t, \mathbf{u}(t)) - \mathbf{p}_q(t) = \frac{\mathbf{f}^{(q+1)}(\xi_t, \mathbf{u}(\xi_t))}{(q+1)!} \underbrace{\prod_{\ell=0}^{q}(t - t_{k-\ell})}_{=:L(t)} = \frac{L(t)}{(q+1)!}\mathbf{u}^{(q+2)}(\xi_t), \qquad \xi_t \in [t_{k-q}, t_k]$$

und damit

$$\mathbf{u}(t_n) = \mathbf{u}(t_{n-\sigma}) + \sum_{\mu=0}^{q} \mathbf{f}(t_{k-\mu}, \mathbf{u}(t_{k-\mu})) \underbrace{\int_{t_{n-\sigma}}^{t_n} L_\mu^{(q)}(s)\,\mathrm{d}s}_{\text{Gewichte} \to \text{ausrechnen}} + \int_{t_{n-\sigma}}^{t_n} \frac{L(t)}{(q+1)!}\mathbf{u}^{(q+2)}(\xi_t)\mathrm{d}t, \qquad \xi_t \in [t_{k-q}, t_k]$$

Abschätzung des Fehlerterms:

$$|L(t)| = \prod_{\ell=0}^{q} |t - t_{k-\ell}| = \prod_{\ell=0}^{q} \big|t - \underbrace{(\underbrace{t_0 + kh}_{=t_k} - \ell h)}_{=t_0+(k-\ell)h}\big| = \prod_{\ell=0}^{q} |(t - t_k) + \ell h|$$

$$= h^{q+1} \prod_{\ell=0}^{q} \left| \frac{t - t_k}{h} + \ell \right| \leq h^{q+1} \prod_{\ell=0}^{q} \Bigg(\underbrace{\left| \frac{t - t_k}{h} \right|}_{\substack{t_{n-\sigma} \leq t \leq t_n \\ \text{und} \\ t_{n-1} \leq t_k \leq t_n \\ \Rightarrow |t - t_k| \leq \sigma h}} + \ell \Bigg)$$

$$\leq h^{q+1} \prod_{\ell=0}^{q} (\sigma + \ell) = \mathcal{O}(h^{q+1})$$

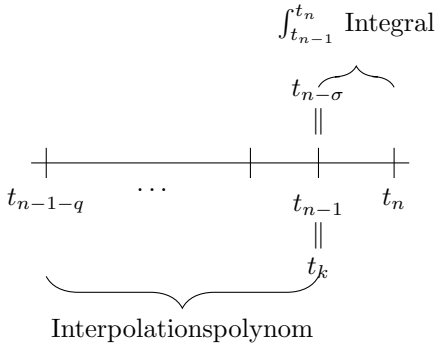
Da die Intervallbreite, über die integriert wird, auch proportional zu h ist, gilt:

$$\left\| \int_{t_{n-\sigma}}^{t_n} \frac{L(t)}{(q+1)!} \mathbf{u}^{(q+2)}(\xi_t)\, \mathrm{d}t \right\| = \mathcal{O}(h^{q+2}), \qquad \xi_t \in [t_{k-q}, t_k]$$

falls $\mathbf{u}^{(q+2)}$ beschränkt ist.

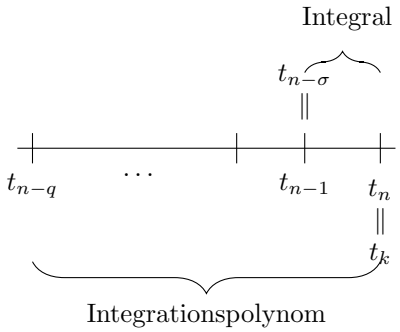
Verfahrensklassen

(a) Adams-Bashfort-Formeln: $\sigma = 1$, $k = n - 1$ (explizit)
d.h.



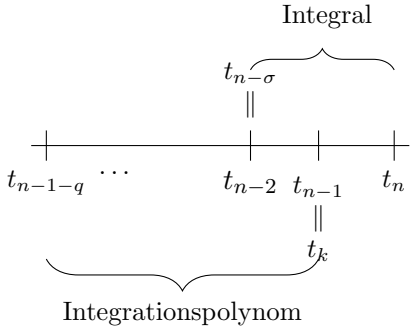
$q = 0$	$\mathbf{y}_n = \mathbf{y}_{n-1} + h\mathbf{f}_{n-1}$	(Expliziter Euler)	$(m = 1)$
$q = 1$	$\mathbf{y}_n = \mathbf{y}_{n-1} + \frac{1}{2}h(3\mathbf{f}_{n-1} - \mathbf{f}_{n-2})$		$(m = 2)$
$q = 2$	$\mathbf{y}_n = \mathbf{y}_{n-1} + \frac{1}{12}h(23\mathbf{f}_{n-1} - 16\mathbf{f}_{n-2} + 5\mathbf{f}_{n-3})$		$(m = 3)$

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



$q=0$	$\mathbf{y}_n = \mathbf{y}_{n-1} + h\mathbf{f}_n$	(Impliziter Euler)	$(m=1)$
$q=1$	$\mathbf{y}_n = \mathbf{y}_{n-1} + \frac{h}{2}(\mathbf{f}_n + \mathbf{f}_{n-1})$	(Trapezregel)	$(m=2)$
$q=2$	$\mathbf{y}_n = \mathbf{y}_{n-1} + \frac{1}{12}h(5\mathbf{f}_n + 8\mathbf{f}_{n-1} - \mathbf{f}_{n-2})$		$(m=3)$

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



$q=0 \quad \mathbf{y}_n = \mathbf{y}_{n-2} + 2h\mathbf{f}_{n-1}$ (explizite Mittelpunktregel) $(m=1)$

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

$$q=2 \quad \mathbf{y}_n = \mathbf{y}_{n-2} + \frac{1}{3}h(\mathbf{f}_n + 4\mathbf{f}_{n-1} + \mathbf{f}_{n-2}) \qquad (m=3)$$

Startwerte

- Die Verfahren erfordern $n \geq q+1$ (explizite) und $n \geq q$ (implizite Formeln) Startwerte.
- Zwei Möglichkeiten
  (a) Berechne Startwerte mittels Einschrittverfahren entsprechender Ordnung.
  (b) Starte mit geringer Ordnung und erhöhe diese schrittweise.
- Ähnliche Probleme treten bei einer Änderung der Schrittweite auf.

5.2. Differentiationsbasierte Verfahren

Idee: Lege Polynom vom Grad q durch die Funktionswerte $\mathbf{u}$ (statt durch die Werte von $\mathbf{f}$ wie oben):

$$\mathbf{p}_q(t) = \sum_{\mu=o}^{q} \mathbf{u}(t_{k-\mu}) L_\mu^{(q)}(t), \qquad L_\mu^{(q)}(t) \text{ wie oben.}$$

Analog lässt sich zeigen, dass

$$\mathbf{u}(t) - \mathbf{p}_q(t) = \frac{L(t)}{(q+1)!} \mathbf{u}^{(q+1)}(\xi_t), \qquad \xi_t \in [t_{k-q}, t_k]$$

Einsetzen in die Differentialgleichung (Ableiten von $\mathbf{p}_q(t)$) ergibt

$$\mathbf{p}_q'(t_n) = \sum_{\mu=0}^{q} \left(L_\mu^{(q)} \right)' (t_n)\ \mathbf{u}(t_{k-\mu}) = \mathbf{f}(t_n, \mathbf{u}(t_n)) + \underbrace{\mathcal{O}(h^q)}_{\substack{\text{wegen Differenzieren} \\ \text{von } L(T) \text{ nach } t}}$$

$t_n = t_k$ (Auswertepunkt = „vorderster“ Punkt) ergibt „Rückwärtsdifferenzenformel“ (BDF - backward difference formula)

$$\begin{array}{lll} q=1: & \mathbf{y}_n - \mathbf{y}_{n-1} = h\mathbf{f}_n & \text{(Impliziter Euler)} \quad (m=1) \\ q=2: & \mathbf{y}_n - \frac{4}{3}\mathbf{y}_{n-1} + \frac{1}{3}\mathbf{y}_{n-2} = \frac{2}{3}h\mathbf{f}_n & (m=2) \\ q=3: & \mathbf{y}_n - \frac{18}{11}\mathbf{y}_{n-1} + \frac{9}{11}\mathbf{y}_{n-2} - \frac{2}{11}\mathbf{y}_{n-3} = \frac{6}{11}h\mathbf{f}_n & (m=3) \end{array}$$

Anmerkung 5.1
Beachte:

- *Adams-Moulton: $q = 0$ entspricht Implizitem Euler*
- *BDF: $q = 0$ entspricht Implizitem Euler*

5.3. Konsistenz und Stabilität von linearen Mehrschrittmethoden

Definition 5.2
Eine allgemeine s-stufige lineare Mehrschrittmethode (LMM) hat die Form:

$$\sum_{r=0}^{s} \alpha_{s-r} \mathbf{y}_{n-r} = h \sum_{r=0}^{s} \beta_{s-r} \mathbf{f}_{n-r}, \qquad \textit{mit } \mathbf{f}_i = \mathbf{f}(t_i, \mathbf{y}_i), \qquad n \geq s \tag{5.1}$$

Man normiert $\alpha_s = 1$, d.h. es gibt $2s-1$ zu bestimmende Koeffizienten. $\beta_s = 0$ liefert explizite Verfahren, $\beta_s \neq 0$ implizite. Weiter fordern wir $|\alpha_0| + |\beta_0| \neq 0$ für ein s-stufiges Verfahren.

Hilfssatz 5.3
Eine LMM ist genau dann konsistent mit jeder AWA, wenn gilt:

$$\sum_{r=0}^{s} \alpha_r = 0 \textit{ und } \sum_{r=0}^{s} (r\alpha_{s-r} + \beta_{s-r}) = 0$$

ohne Beweis.

Konvergenz:

- Für die Konvergenz einer lineare Mehrschrittmethode gibt es eine zusätzliche Anforderung, die sogenannte „Nullstabilität“

- Nullstabilität ist eine Bedingung, die es nur bei LMM gibt.

- Adams-Bashfort, Adams-Moulton, Nyström, Milne-Simpson sind nullstabil.

- BDF ist nur bis $s = 6$ nullstabil!

- Ordnungsbarriere von Dahlquist: Eine nullstabile LMM hat maximal die Ordnung $s+2$ für s gerade und $s+1$ für s ungerade. Bei den Einschrittverfahren war dagegen die maximale Ordnung $2s$.

Stabilität:

Wie bei Einschrittverfahren wendet man die LMM auf das Modellproblem (4.2) an. Dadurch erhält man:

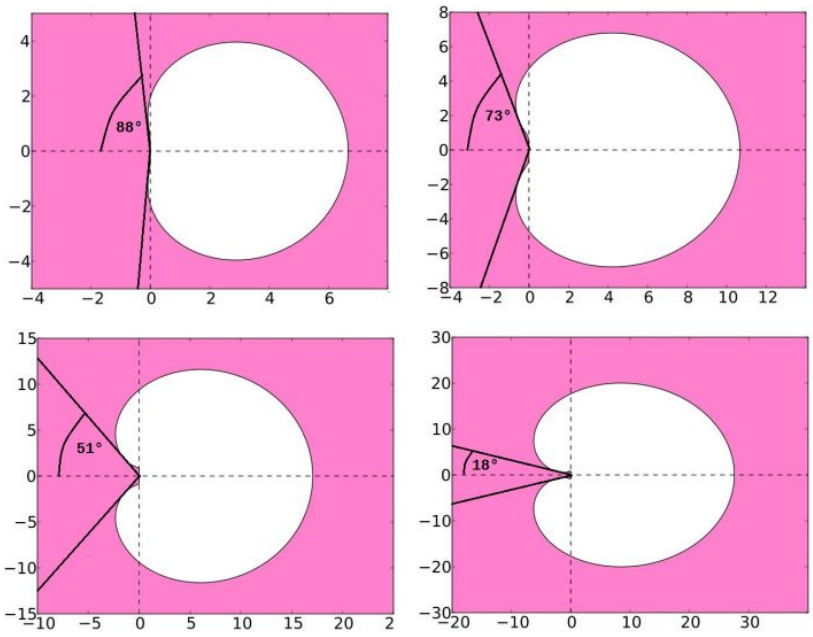
- „optimale“ Verfahren mit Ordnung $s+2$ bei s gerade (z.B. Simpson) haben $\mathrm{SG} = \{0\}$.

- explizite Verfahren haben immer beschränkte Stabilitätsgebiete (wie bei den Einschrittverfahren).

- die maximale Ordnung einer (impliziten) A-stabilen LMM ist $m = 2$.

- BDF-Verfahren bis 2. Ordnung A-stabil, darüber $A(\alpha)$-stabil, d.h. es gibt einen Bereich der negativen Halbebene der komplexen Zahlen für den Sie nicht A-stabil sind (siehe folgende Abbildung). Dieser wird durch Geraden begrenzt, die die reelle Achse in einem Winkel α schneiden

m	3	4	5	6
α	88°	73°	51°	18°



Quelle: http://de.wikipedia.org/wiki/BDF-Verfahren

5.4. Prädiktor-Korrektor-Methoden

Idee:

- Um ein stabiles Verfahren zu erhalten verwendet man eine implizite LMM (Korrektor, C).
- Man nutzt eine Fixpunktiteration zur Lösung des nichtlinearen Gleichungssystems (welches als nicht steif betrachtet wird).
- Ein geeigneter Startwert für die Fixpunktieration wird mit einer expliziten LMM (Prädiktor, P) bestimmt.
- Jeder Schritt der Fixpunktiteration kostet <u>eine</u> Auswertung von $\mathbf{f}$ (zur Berechnung von $\mathbf{f}_n^{(k)} = \mathbf{f}(t_n, \mathbf{y}_n^{(k)})) \rightsquigarrow$ „Evaluate“

Schema: $P \overset{\mathbf{y}_n^{(0)}}{\downarrow} E \quad C \overset{\mathbf{y}_n^{(1)}}{\downarrow} EC \ldots = P(EC)^k \qquad (k \geq 1)$

$\uparrow$ $\mathbf{f}_n^{(0)}$ — $\uparrow$ hier hat man $\mathbf{y}_n^k$, aber nur $\mathbf{f}_n^{(k-1)}$ berechnet

bzw. $P(EC)^k E$

$\uparrow$das $\mathbf{f}_n^{(k)}$ wird im Schritt $n{+}1$ verwendet

Man zeigt für die Ordnung

$$m^{(PC)} = \min\{ \underbrace{m^{(C)}}_{\substack{\text{Ordnung} \\ \text{Korrektor}}}, \underbrace{m^{(P)}}_{\substack{\text{Ordnung} \\ \text{Prädiktor}}} + \underbrace{k}_{\substack{\#\text{ Iterationen} \\ \text{das heißt es} \\ \text{genügt ein Prädiktor} \\ \text{„niedrigerer“ Ordnung}}} \}$$

5.5. Zusammenfassung

- Mit Mehrschrittverfahren (LMM=Lineare Mehrschrittmethode) versucht man Rechenzeit durch Wiederverwendung alter Zeitschritte zu sparen. Dies resultiert evtl. in einem höheren Speicherbedarf.
- Herleitung entweder durch Integration: Adams-Bashfort (explizit), Adams-Moulton (implizit), Nyström (explizit), Milne-Simpson (implizit) oder durch Differentiation: BDF-Verfahren (implizit).
- Für den Nachweis der Konvergenz reicht die Konsistenz nicht mehr aus, es gibt eine zusätzliche Bedingung, die sogenannte „Nullstabilität“. BDF-Verfahren sind nur bis sechster Ordnung nullstabil und damit konvergent.
- Nach Dahlquist gib es keine konvergente LMM mit höherer Ordnung als $s+2$. Verfahren mit Ordnung $s+2$ sind jedoch vollständig instabil, also unbrauchbar.
- Im wesentlichen erreicht man mit einem LMM eine Ordnung von s durch einmalige Auswertung der rechten Seite und zusätzliche Verwendung von Werten aus vorherigen Zeitschritten.
- Explizite Verfahren haben beschränkte Stabilitätsgebiete.
- Es gibt keine (impliziten) A-stabilen LMM mit mehr als 2. Ordnung. BDF-Verfahren sind A(α)-stabil, für reelle λ darf der Zeitschritt beliebig groß werden.
- Es gibt keine L-stabilen LMM.
- Damit eignen sich nur BDF-Verfahren gut für steife Differentialgleichungen. Für nichtsteife DGL können LMM aber zu merklichen Zeiteinsparungen führen.
- Nachteil der LMM ist vor allem die geringere Flexibilität. Bei einem Wechsel der Schrittweite sind wieder entsprechend viele Startwerte im Abstand des neuen Zeitschritts notwendig. Diese müssen entweder mit einer Einschrittmethoden berechnet werden, oder man nimmt einen Genauigkeitsverlust beim Wechsel der Schrittweite in Kauf.
- Eine effiziente Variante von LMM ist die Kombination einer impliziten LMM, deren Lösung mit einer Fixpunktiteration berechnet wird, mit einer expliziten LMM zur Berechnung des Startwerts der Fixpunktiteration. Dies nennt man Prädiktor-Korrektor-Verfahren.

6. Randwertprobleme

Definition 6.1
Unter einer Randwertaufgabe (RWA) verstehen wir eine Aufgabe der Form

$$\mathbf{u}'(t) = \mathbf{f}(t, \mathbf{u}(t)) \qquad t \in I = [a, b] \qquad\qquad \mathbf{r}(\mathbf{u}(a), \mathbf{u}(b)) = 0$$

mit zwei mindestens zweimal stetig-differenzierbaren vektorwertigen Funktionen $\mathbf{f} : I \times \mathbb{R}^d \to \mathbb{R}^d$ *(rechte Seite der DGL) und* $\mathbf{r} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ *(allgemeine Randbedingung für* $t = a$ *und* $t = b$*).*

Anmerkung 6.2

- *Bei einer Randwertaufgabe sind statt mehrerer Anfangsbedingungen Randbedingungen am Anfang und am Ende gegeben.*
- *Damit es mehr als eine Bedingung geben kann, muss die DGL in der Regel mindestens zweiter Ordnung sein.*

Beispiel 6.3
Betrachte die skalare DGL zweiter Ordnung (z.B. für die Krümmung eines eingespannten Balkens):

$$\begin{aligned} u''(t) &= f(t, u(t), u'(t)) && t \in I = [a, b] \\ u(a) &= u_a, \qquad u(b) = u_b \end{aligned} \tag{6.1}$$

Verwandle die DGL in ein System erster Ordnung:

$$\tilde{\mathbf{u}}(t) = \begin{pmatrix} \tilde{u}_1(t) \\ \tilde{u}_2(t) \end{pmatrix} = \begin{pmatrix} u(t) \\ u'(t) \end{pmatrix}$$

$$\tilde{\mathbf{u}}'(t) = \tilde{\mathbf{f}}(t, \tilde{\mathbf{u}}(t)) = \begin{pmatrix} \tilde{u}_2(t) \\ f(t, \tilde{u}_1(t), \tilde{u}_2(t)) \end{pmatrix} = \begin{pmatrix} u'(t) \\ f(t, u(t), u'(t)) \end{pmatrix}$$

Randbedingungen:

$$\mathbf{r}(\tilde{\mathbf{u}}(a), \tilde{\mathbf{u}}(b)) = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} \tilde{u}_1(a) \\ \tilde{u}_2(a) \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}}_{\mathbf{B}} \begin{pmatrix} \tilde{u}_1(b) \\ \tilde{u}_2(b) \end{pmatrix} = \underbrace{\begin{pmatrix} u_a \\ u_b \end{pmatrix}}_{\mathbf{c}} \iff \mathbf{A}\,\tilde{\mathbf{u}}(a) + \mathbf{B}\,\tilde{\mathbf{u}}(b) = \mathbf{c} \tag{6.2}$$

Die Existenz und Eindeutigkeit der Lösungen von RWA ist nur unter sehr eingeschränkten (und in der Regel nicht realistischen) Bedingungen zu beweisen, wir nehmen im weiteren jedoch an, dass es für die behandelten Probleme eine Lösung gibt.

Beispiel 6.4
Folgendes Beispiel ergibt eine Eindruck davon, was möglich ist:

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

(b) $u'' + u = 0$ *auf* $[0, \pi]$, $u(0) = 0$, $u(\pi) = 0 \to u(t) = c\sin(t)$, $c \in \mathbb{R}$ *d.h.* ∞*-viele Lösungen.*

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

Wir beschränken uns auf die Betrachtung zweier numerischer Verfahren.

6.1. Schießverfahren

Idee:

- Umformulieren als Randwertproblem mit „unbekannter“ Anfangsbedingung.
- Lösen eines Gleichungssystems für die Anfangsbedingung

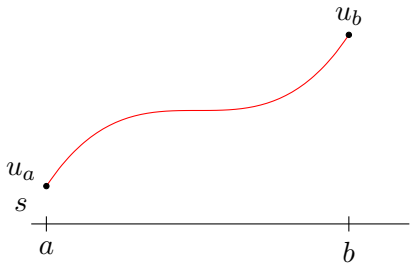
Wir demonstrieren das Schießverfahren anhand von Beispiel 6.3:

Finde $\tilde{\mathbf{u}}(t,s)$ mit

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

mit unbekanntem $s \in \mathbb{R}$. Betrachte s als zusätzlichen Parameter, der so gewählt werden muss, dass gilt:

$$\tilde{u}_1(b,s) \overset{!}{=} u_b. \tag{6.3}$$



Finde s durch Lösen von (6.3) mittels Newton-Verfahren. Dazu brauchen wir $\frac{\partial}{\partial s}\tilde{u}_1(b,s)$.

Lemma 6.5
Sei $\tilde{\mathbf{u}}(t,s)$ für festes $s \in \mathbb{R}$ die Lösung der „AWA“

$$\frac{\partial \tilde{\mathbf{u}}}{\partial t}(t,s) = \tilde{\mathbf{f}}(t,\tilde{\mathbf{u}}(t,s)), \qquad t \in [a,b], \qquad \tilde{\mathbf{u}}(a) = \begin{pmatrix} u_a \\ s \end{pmatrix} \tag{6.4}$$

mit $\tilde{\mathbf{f}}$ aus Beispiel 6.3. Dann ist die Funktion $\mathbf{v}(t,s) = \frac{\partial}{\partial s}\tilde{\mathbf{u}}(t,s)$ gegeben als Lösung der AWA

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{v}(t,s) &= \begin{pmatrix} v_2(t,s) \\ \left[\frac{\partial}{\partial \tilde{u}_1}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s))\right] \cdot v_1(t,s) + \left[\frac{\partial}{\partial \tilde{u}_2}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s))\right] \cdot v_2(t,s) \end{pmatrix} \\ &= \begin{pmatrix} v_2(t,s) \\ \left[\frac{\partial}{\partial u} f(t,u(t,s),u'(t,s))\right] \cdot v_1(t,s) + \left[\frac{\partial}{\partial u'} f(t,u(t,s),u'(t,s))\right] \cdot v_2(t,s) \end{pmatrix} \end{aligned} \tag{6.5}$$

mit den Anfangswerten

$$\mathbf{v}(a,s) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Beweis: *Differenziere die DGL nach dem Parameter s*

erste Gleichung:

$$\frac{\partial}{\partial s}\frac{\partial}{\partial t}\tilde{u}_1(t,s) = \frac{\partial}{\partial s}\tilde{u}_2(t,s)$$

zweite Gleichung:

$$\begin{aligned}\frac{\partial}{\partial s}\frac{\partial}{\partial t}\tilde{u}_2(t,s) =& \frac{\partial}{\partial s}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s)) \\ \overset{\text{Ketten-}}{\underset{\text{regel}}{=}}& \frac{\partial}{\partial \tilde{u}_1}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s))\cdot\frac{\partial}{\partial s}\tilde{u}_1(t,s) + \frac{\partial}{\partial \tilde{u}_2}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s))\cdot\frac{\partial}{\partial s}\tilde{u}_2(t,s)\end{aligned}$$

Vertauschen der Ableitungen auf der linken Seite und Abkürzung $v_i(t,s) \coloneqq \frac{\partial}{\partial s}\tilde{u}_i(t,s)$ liefert:

$$\begin{aligned}&\frac{\partial v_1}{\partial t}(t,s) = \frac{\partial}{\partial s}\tilde{u}_2(t,s) = v_2(t,s) \\ &\frac{\partial v_2}{\partial t} = \left[\frac{\partial}{\partial \tilde{u}_1}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s))\right]\cdot v_1(t,s) + \left[\frac{\partial}{\partial \tilde{u}_2}\tilde{f}_2(t,\tilde{\mathbf{u}}(t,s))\right]\cdot v_2(t,s).\end{aligned}$$

Anfangswerte:

$$\begin{aligned}v_1(a,s) &= \frac{\partial}{\partial s}\tilde{u}_1(a,s) = \frac{\partial}{\partial s}u_a = 0 \\ v_2(a,s) &= \frac{\partial}{\partial s}\tilde{u}_2(a,s) = \frac{\partial}{\partial s}s = 1\end{aligned}$$

□

Damit erhält man das sogenannte „einfache“ Schießverfahren:

Algorithmus 6.6 (Schießverfahren)
1) Wähle Startwert $s^{(0)}$

2) Berechne numerische Lösung $\mathbf{y}^{(i)}$ von (6.4) zum Startwert $s^{(i)}$

2b) Falls $|u_b - y_1(b,s^{(i)})| < \epsilon \rightarrow$ FERTIG!

3) Berechne $\frac{\partial \mathbf{y}}{\partial s}$ durch Lösung von(6.5) (numerisch)

4) Newton-Update: $y_1(b,s^{(i)} + \Delta s) \approx y_1(b,s^{(i)}) + \frac{\partial}{\partial s}y_1(b,s^{(i)})\cdot\Delta s = u_b$

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

5) $i = i+1$; Gehe nach 2)

Man zeigt: Hat das im Schritt 3) und 4) verwendete Lösungsverfahren die Ordnung m, konvergiert das Schießverfahren für genügend kleines h ebenfalls mit der Ordnung m.

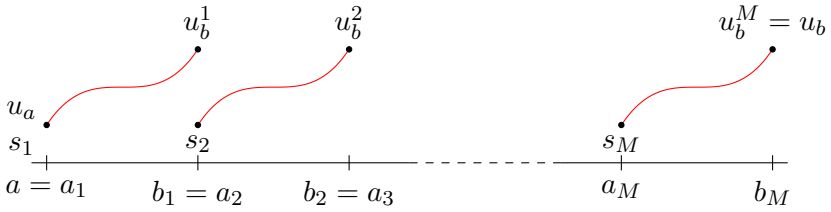
Problem des Schießverfahrens:
Der Stabilitätssatz liefert

$$\|\mathbf{y}(b,s) - \mathbf{y}(b,s+\epsilon)\| \leq c \cdot e^{L\cdot(b-a)} \cdot \epsilon$$

Kleine Änderungen in s bewirken also bei großer Lipschitzkonstante L entsprechend große Abweichungen. Dies erfordert möglicherweise sehr kleine Schrittweiten bzw. es genügt eventuell nicht mehr mit doppelter Genauigkeit zu rechnen.

Lösung: „Multiple Shooting“ Verfahren

Zerlege $[a,b]$ in M Teilintervalle $[a_i, b_i]$, $a_1 = a$, $a_{i+1} = b_i$, $1 \leq i \leq M$, $b_M = b$, so dass $e^{L(b_i - a_i)}$ klein genug.
z.B. für $L = 10, b - a = 10 \Rightarrow e^{L(b-a)} = e^{100}$ zu groß.
Mit $M = 10$ ist $e^{L(b_i-a_i)} \leq e^{10}$, das ist o.k.



Zu bestimmen sind nun $s_1, \ldots, s_M$ sowie interne Startwerte $w_2, \ldots, w_M$ so dass auf jedem Intervall $1 \leq i \leq M$ gilt:

$$\tilde{\mathbf{u}}^i(\underset{\substack{\text{rechtes}\\ \text{Ende}}}{b_i}, s_i) = \begin{pmatrix} w_{i+1} \\ s_{i+1} \end{pmatrix} \qquad 1 \leq i < M \tag{6.6a}$$

$$\tilde{u}_1^i(b_M, s_M) = u_b \tag{6.6b}$$

s_i ist dabei jeweils der Startwert für $\tilde{u}_2^i$. Diese Bedingung stellt die stetige Differenzierbarkeit der Lösung sicher, da $\tilde{u}_2 = u'$.

Wir erhalten insgesamt $2M - 1$ Bedingungen für die $2M - 1$ Unbekannten, das Problem ist also eindeutig lösbar, wenn die RWA eindeutig lösbar ist.

Die Startwerte sind:

$$\tilde{\mathbf{u}}^1(a_1, s_1) = \begin{pmatrix} u_a \\ s_1 \end{pmatrix}$$

$$\tilde{\mathbf{u}}^i(a_i, s_i) = \begin{pmatrix} w_i \\ s_i \end{pmatrix} \qquad 2 \leq iM$$

Löse (6.6a) (6.6b) mit dem Newton-Verfahren.

6.2. Differenzenverfahren

Wir betrachten wieder die RWA 6.1.
Diskretisiere nun die Ableitung von $\mathbf{u}$ mit einem Differenzenquotienten:

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

Damit erhalten wir ein nichtlineares System für die $\mathbf{y}_n \approx \mathbf{u}(t_n)$, $0 \le n \le N$

$$\begin{aligned} &\mathbf{y}_n - \mathbf{y}_{n-1} = h_n \mathbf{f}(t_n, \mathbf{y}_n) \qquad 1 \le n \le N \\ &\mathbf{r}(\mathbf{y}_0, \mathbf{y}_N) = 0 \end{aligned}$$

Insgesamt sind dies $(N+1)$ (gegebenenfalls vektorielle) Gleichungen für $N+1$ (vektorielle) Unbekannte.
Zu zeigen wäre nun, dass das Verfahren konvergiert, dass also

$$\max_{1 \le n \le N} \|\mathbf{u}(t_n) - \mathbf{y}_n\| \to 0 \text{ für } h \to 0.$$

Dies werden wir bei den partiellen DGL zeigen, die bei der Methode der finiten Differenzen auf die gleiche Weise behandelt werden.

Beachte: Für $h \to 0$ sind immer größere (nicht-)lineare Gleichungssysteme zu lösen!

6.3. Zusammenfassung

- Bei Randwertaufgaben (RWA) werden statt mehrerer Anfangsbedingungen Randbedingungen an beiden Seiten des Gebiets angegeben.
- Die Frage nach der Existenz und Eindeutigkeit von Lösungen ist für Randwertaufgaben deutlich schwieriger zu beantworten.
- Zwei Lösungsverfahren wurden anhand einer Beispiel-DGL zweiter Ordnung vorgestellt:

  Schießverfahren Bei Schießverfahren wird die RWA in eine AWA mit einem zusätzlichen unbekannten Parameter umgewandelt. Dieser Parameter wird in einem iterativen Verfahren zusammen mit der Lösung der RWA bestimmt. Dabei ist in jedem Schritt eine AWA für die Lösung und eine AWA für die Ableitung der Lösung nach dem Parameter (Sensitivität der Lösung auf den Parameter) zu berechnen.

  Zur Erhöhung der Stabilität kann bei „Multiple Shooting“-Verfahren das Zeitintervall in Teilintervalle aufgeteilt werden. Für jedes Teilintervall wird ein Schießverfahren berechnet. Die Kontinuität der Lösung wird durch zusätzliche Gleichungen gefordert und sichergestellt.

  Differenzenverfahren Die Ableitung von $\mathbf{u}$ wird durch ein Differenzenverfahren genähert. Dies führt auf ein System (nicht-) linearer Gleichungen. Dies ist im wesentlichen dasselbe wie ein Finite-Differenzen-Verfahren zur Lösung partieller Differenzialgleichungen.

7. Ausblick zu gewöhnlichen Differentialgleichungen

Unter anderem haben wir uns bei der Behandlung gewöhnlicher Differentialgleichungen nicht um strukturerhaltende Verfahren gekümmert.

Strukturerhaltende Lösungsverfahren:
Physikalische Systeme erfüllen oft Erhaltungsgleichungen, z.B. Massen-, Impuls-, Drehimpuls-, Energieerhaltung.

Beispiele:

- Energieerhaltung beim Pendel,
- Impuls-, Drehimpuls- und Energieerhaltung beim N-Körper-Problem.

Die bisher behandelten Verfahren erhalten diese Größen in der Regel nicht exakt (nur im Grenzwert für $h \to 0$). Es gibt jedoch Verfahren, die dies auch im diskreten Fall sicherstellen.

Beispiel 7.1 (Leapfrog-Verfahren)
N-Körperproblem

$$(*)\ \mathbf{x}_i'(t) = \mathbf{v}_i(t)$$

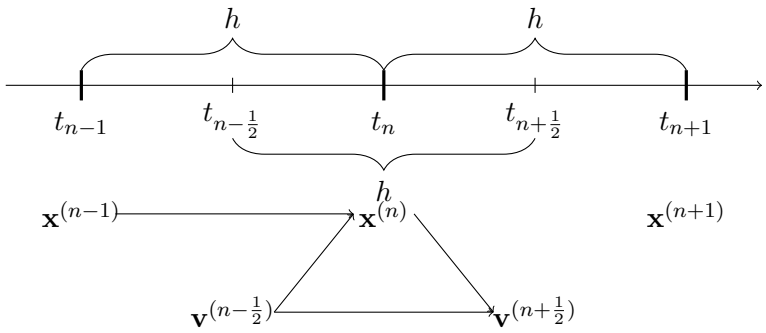
$$(**)\ \mathbf{v}_i'(t) = G \sum_{\substack{j=1 \\ j\neq i}}^{N} m_j \frac{\mathbf{x}_j(t)-\mathbf{x}_i(t)}{\|\mathbf{x}_j(t)-\mathbf{x}_i(t)\|^3} =: \underbrace{\mathbf{a}_i(\mathbf{x}(t))}_{\text{Beschleunigung}}$$

Nähere nun (*) durch

$$\overbrace{\frac{\mathbf{x}_i^{(n)} - \mathbf{x}_i^{(n-1)}}{h}}^{\text{Approximation 2. Ordnung für } \mathbf{x}_i'^{(n-\frac{1}{2})}} = \mathbf{v}_i^{(n-1/2)} \iff \mathbf{x}_i^{(n)} = \mathbf{x}_i^{(n-1)} + h\,\mathbf{v}_i^{(n-1/2)}$$

und nähere (**) durch

$$\overbrace{\frac{\mathbf{v}_i^{(n+1/2)} - \mathbf{v}_i^{(n-1/2)}}{h}}^{\text{Approximation 2. Ordnung für } \mathbf{v}_i'^{(n)}} = \mathbf{a}_i(\mathbf{x}^{(n)}) \iff \mathbf{v}_i^{(n+1/2)} = \mathbf{v}_i^{(n-1/2)} + h\,\mathbf{a}_i(\mathbf{x}^{(n)})$$



Die Behandlung der beiden Gleichungen auf zwei zueinander versetzten Zeitgittern ist zulässig, weil die rechte Seite der Differentialgleichung in (*) nur von $\mathbf{v}$ *und in (**) nur von* $\mathbf{x}$ *abhängt.*

Vorteile dieses Verfahrens:

1) *Das N-Körper-Problem ist invariant gegen Zeitumkehr, d.h. ersetzt man* $\mathbf{v}$ *in der DGL überall durch* $-\mathbf{v}$*, kehrt man exakt an den Ausgangsort zurück, d.h. falls* $\mathbf{x}$ *und* $\mathbf{v}$ *eine Lösung der AWA*

$$\begin{aligned}\mathbf{x}'(t) &= \mathbf{v}(t), \quad \mathbf{x}(0) = \mathbf{x}_0\\ \mathbf{v}'(t) &= \mathbf{a}(\mathbf{x}(t)), \quad \mathbf{v}(0) = \mathbf{v}_0\end{aligned}$$

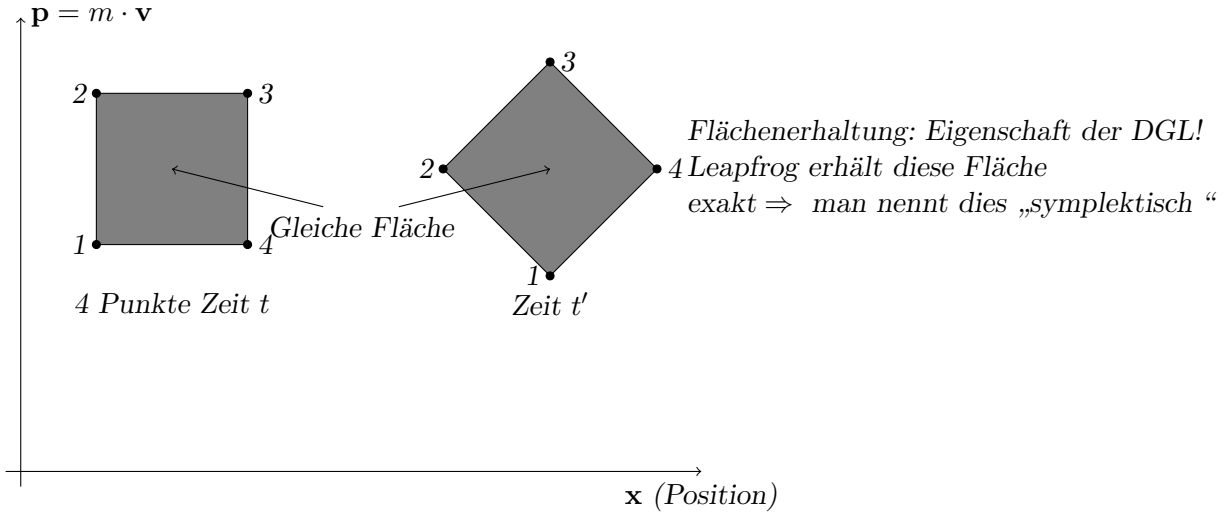
im Intervall $[0, T]$ *sind, dann sind* $\tilde{\mathbf{x}}(\tilde{t}) \coloneqq \mathbf{x}(T - \tilde{t})$ *und* $\tilde{\mathbf{v}}(\tilde{t}) = \mathbf{v}(T - \tilde{t})$ *Lösungen der AWA*

$$\begin{aligned}\tilde{\mathbf{x}}'(\tilde{t}) &= -\tilde{\mathbf{v}}(\tilde{t}), \quad \tilde{\mathbf{x}}(0) = \mathbf{x}(T)\\ \tilde{\mathbf{v}}'(\tilde{t}) &= -\mathbf{a}(\tilde{\mathbf{x}}(\tilde{t})), \quad \tilde{\mathbf{v}}(0) = \mathbf{v}(T).\end{aligned}$$

Das Leapfrog-Verfahren erfüllt dies exakt.

2) *Der Drehimpuls bleibt erhalten, die Energie (leider) nicht.*
Aber: Der Energiefehler bleibt beschränkt $\Rightarrow$ *Langzeitstabilität.*

3) *Symplektisch: (z.B. Pendel)*



Flächenerhaltung: Eigenschaft der DGL!
Leapfrog erhält diese Fläche exakt $\Rightarrow$ *man nennt dies „symplektisch “*

Ein anderer nicht behandelter Aspekt wären strukturausnutzende Verfahren: Nutze Struktur von $\mathbf{f}(t, \mathbf{x})$ zur Beschleunigung der Berechnungen.

Teil II.
Partielle Differentialgleichungen

8. Partielle Differentialgleichungen

Partielle Differentialgleichungen tauchen in allen naturwissenschaftlichen Disziplinen auf.

8.1. Erhaltungsgleichungen und Wärmetransport

Partielle Differentialgleichungen tauchen z.B. auf, wenn die Erhaltung von Größen wie Masse, Energie, Impuls oder Drehimpuls betrachtet wird.

Wir wollen als Beispiel den Transport von Wärme betrachten in einem mit einer Festsubstanz oder einem Fluid gefüllten begrenzten Gebiet $\Omega \subseteq \mathbb{R}^3$ betrachten. Die erhaltene Größe ist die thermische Energie E. Wir nehmen an, dass die Energiedichte e proportional zur Temperatur ist

$$e = \rho c T$$

wobei c die spezifische Wärmekapazität (in J kg^{-1} K^{-1}), ρ die Dichte des Materials (in kg m^{-3}) und T die absolute Temperatur (in K) ist.

In Feststoffen und Fluiden kann Wärmeleitung durch die Beziehung

$$\mathbf{j}_{\text{cond}} = -\lambda \nabla T.$$

modelliert werden. Die Wärmeflussdichte $\mathbf{j}_{\text{cond}}$ (also die transportierte Energie pro Fläche und Zeit in J m^{-2} s^{-1}) ist proportional zum Temperaturgradienten mit Proportionalitätskonstante λ (in J s^{-1} m^{-1} K^{-1}) und fließt in Richtung von der höheren zur niedrigeren Temperatur.

In Fluiden kann Wärme außerdem mit der mit einer Geschwindigkeit $\mathbf{v}$ (in m s^{-1}) fließenden Flüssigkeit transportiert werden. Dies konvektive Wärmeflussdichte $\mathbf{j}_{\text{conv}}$ lässt sich durch

$$\mathbf{j}_{\text{conv}} = e\mathbf{v} = \rho c T \mathbf{v}$$

beschreiben.

Außerdem kann es noch Quellen oder Senken im Gebiet geben. Quellen- oder Senken ohne Massenaustausch können durch einen Quell-/Senkenterm q_h (in J m^{-3} s^{-1}) beschrieben werden (positiv für Quellen, negativ für Senken). Wird hingegen ein Fluid zu- oder abgeführt, dann wird damit auch Energie transportiert, die sich als $\rho\, c\, T_l\, q_l$ beschreiben lässt, wobei q_l die Entnahme-/Zuflussrate ist (in kg m^{-3} s^{-1}). Bei einem Zufluss ist die Temperatur bekannt und die Quelle kann in den Quell-/Senkenterm q_h integriert werden. Bei einem Abfluss ist die Temperatur vom System gegeben, wir erhalten daher einen zusätzlichen Term in der Gleichung.

Die Energieerhaltung wird dann beschrieben durch

$$\begin{aligned} &\frac{\partial e}{\partial t} + \nabla \cdot \mathbf{j}_{\text{tot}} + \rho c T\, q_{l,\text{out}} = \frac{\partial(\rho c T)}{\partial t} + \nabla \cdot (\mathbf{j}_{\text{cond}} + \mathbf{j}_{\text{conv}}) + \rho c T\, q_{l,\text{out}} = q_h \qquad \text{in } \Omega \\ \iff &\frac{\partial(\rho c T)}{\partial t} + \nabla \cdot (\rho c T\, \mathbf{v} - \lambda \nabla T) + \rho c T\, q_{l,\text{out}} = q_h \qquad \text{in } \Omega \end{aligned} \tag{8.1}$$

Dies ist eine lineare partielle Differentialgleichung 2. Ordnung für die Temperatur T.

Für den Fall stationären Wärmetransports ($\frac{\partial T}{\partial t} = 0$) und reiner Wärmeleitung ohne Zufluss von Flüssigkeit erhalten wir:

$$\nabla \cdot (-\lambda \nabla T) = q_h$$

Ist die Wärmeleitfähigkeit konstant, können wir durch $-\lambda$ teilen und erhalten die Poisson-Gleichung

$$\nabla \cdot \nabla T = \Delta T = -\frac{q_h}{\lambda} = f \iff \Delta T = f$$

wobei Δ der Laplace-Operator ist mit $\Delta T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}$. Wenn $f = 0$ dann nennt man die Gleichung auch Laplace-Gleichung.

8.2. Definition

Definition 8.1

Eine Gleichung zur Bestimmung einer Funktion $u : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, $u \in \mathscr{C}^m$ in der neben der Funktion $u(\mathbf{x})$ und Ausdrücken in $\mathbf{x}$ auch partielle Ableitungen von u bis zur m-ten Ordnung vorkommen, heißt partielle Differentialgleichung (PDGL oder PDE) m-ter Ordnung. Allgemein kann man diese in der Form

$$F\left(\frac{\partial^m u}{\partial x_1^m}(\mathbf{x}), \frac{\partial^m u}{\partial x_1^{m-1} \partial x_2}(\mathbf{x}), \ldots, \frac{\partial^m u}{\partial x_n^m}(\mathbf{x}), \ldots, \frac{\partial^{m-1} u}{\partial x_1^{m-1}}(\mathbf{x}), \ldots, u(\mathbf{x})\right) = 0 \qquad \forall \mathbf{x} \in \Omega \quad (8.2)$$

schreiben.

Eine m-mal partiell differenzierbare Funktion $u : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$ heißt Lösung der PDGL, wenn sie die PDGL in allen Punkten $\mathbf{x} \in \Omega$ erfüllt. Zur eindeutigen Festlegung von u sind noch „zusätzliche Bedingungen“ erforderlich, sogenannte Randbedingungen (bei Randbedingungen in der Zeit spricht man von Anfangsbedingungen).

Ein Anfangs- oder Randwertproblem mit partiellen Differentialgleichungen heißt sachgemäß gestellt (oder wohlgestellt), falls eine Lösung existiert, eindeutig ist (mit angemessenen Randbedingungen) und stetig von den Eingangsdaten abhängt.

Lässt sich die PDGL als Linearkombination partieller Ableitungen schreiben, bei der die Koeffizienten vor den partiellen Ableitungen nicht von u abhängen, sprechen wir von einer linearen partiellen Differentialgleichung.

Ist $\mathbf{u}$ eine vektorwertige Funktion, bei der eine partielle Differentialgleichung für jede Komponente des Vektors angegeben ist, dann sprechen wir von einem System partieller Differentialgleichungen.

Anmerkung 8.2

- *Eine PDGL ist ein funktioneller Zusammenhang zwischen partiellen Ableitungen.*
- *$\mathbf{x}$ kann als Komponenten nicht nur Ortskoordinaten, sondern auch z.B. die Zeit t enthalten.*
- *Eine PDGL wird auf einem Gebiet definiert.*

Definition 8.3

Eine Teilmenge $\Omega \subseteq \mathbb{R}^n$ heißt Gebiet, wenn gilt:

1. *Ω ist offen*

2. *Ω ist zusammenhängend, d.h. zu je zwei Punkten $\mathbf{x}_0, \mathbf{y}_0 \in \Omega$ gibt es eine reguläre Kurve $\mathbf{x} : [a, b] \to \Omega$ mit $\mathbf{x}(a) = \mathbf{x}_0$, $\mathbf{x}(b) = \mathbf{y}_0$, die komplett in Ω verläuft.*

Mit $\overline{\Omega}$ bezeichnet man den Abschluss des Gebiets, d.h. Ω zusammen mit dem Grenzwert aller Folgen, die sich aus Elementen von Ω bilden lassen. Der Rand des Gebiets ist dann $\overline{\Omega} \setminus \Omega =: \partial\Omega$. Mit $\boldsymbol{\nu}(\mathbf{x})$ bezeichnet man den äußeren Normalenvektor am Punkt $\mathbf{x} \in \partial\Omega$.

Häufig sind zusätzliche Bedingungen an die Glattheit des Randes notwendig (z.B. dass der Rand aus stückweise regulären Kurven zusammengesetzt ist).

8.3. Klassifikation von partiellen Differentialgleichungen 1. und 2. Ordnung

Definition 8.4
Wir betrachten eine allgemeine lineare partielle Differentialgleichung 2. Ordnung:

$$\underbrace{\sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\partial_{x_i}\partial_{x_j}u}_{\text{Hauptteils}} + \sum_{i=1}^{n} b_i(\mathbf{x})\partial_{x_i}u + a_0(\mathbf{x})u = f(\mathbf{x}) \quad \text{in } \Omega.$$

Eine partielle Differentialgleichung zweiter Ordnung wird aufgrund der Koeffizientenmatrix $\mathbf{A}$ der zweiten Ableitungen (des sogenannten Hauptteils klassifiziert). Die PDGL heißt

elliptisch *wenn $\mathbf{A}$ positiv oder negativ definit ist.*

hyperbolisch *wenn $\mathbf{A}$ indefinit mit genau einem negativen Eigenwert ist (d.h. ein EW hat anderes Vorzeichen als alle anderen EW und kein EW ist Null).*

parabolisch *wenn $\mathbf{A}$ positiv oder negativ semidefinit ist (d.h. alle EW≥ 0), wobei der Eigenwert Null genau einmal vorkommt und wenn* $\operatorname{Rang}\left(\mathbf{A}\,\mathbf{b}\right) = n$.

Für eine PDGL in zwei Dimensionen

$$\underbrace{a(x,y)\frac{\partial^2 u}{\partial x^2}(x,y) + 2b(x,y)\frac{\partial^2 u}{\partial x \partial y}(x,y) + c(x,y)\frac{\partial^2 u}{\partial y^2}(x,y)}_{\text{Hauptteil}}$$
$$+ d(x,y)\frac{\partial u}{\partial x}(x,y) + e(x,y)\frac{\partial u}{\partial y}(x,y) + f(x,y)u(x,y) + g(x,y) = 0 \qquad \text{in } \Omega \quad (8.3)$$

lässt sich das etwas anschaulicher darstellen. Die PDGL heißt hier im Punkt (x, y)

elliptisch falls $\det \begin{pmatrix} a & b \\ b & c \end{pmatrix} = a(x,y)c(x,y) - b^2(x,y) > 0$

hyperbolisch falls $\det \begin{pmatrix} a & b \\ b & c \end{pmatrix} = a(x,y)c(x,y) - b^2(x,y) < 0$

parabolisch falls $\det \begin{pmatrix} a & b \\ b & c \end{pmatrix} = a(x,y)c(x,y) - b^2(x,y) = 0$ und $\operatorname{Rang} \begin{pmatrix} a & b & d \\ b & c & e \end{pmatrix} = 2$.

Definition 8.5
Eine lineare PDGL zweiter Ordnung heißt elliptisch (hyperbolisch, parabolisch) im Gebiet Ω, wenn sie an allen Punkten $\mathbf{x} \in \Omega$ elliptisch (hyperbolisch, parabolisch) ist.

Anmerkung 8.6

- *Der Typ der PDGL hat eine entscheidende Bedeutung für die Existenz und Eindeutigkeit von Lösungen, die Art von notwendigen Randbedingungen und die Auswahl des numerischen Lösungsverfahrens.*
- *Die Klassifikation ist für lineare PDGL mit $n = m = 2$ vollständig. In höheren (Raum-) dimensionen ist sie nicht mehr vollständig.*
- *Der Typ der PDGL ist invariant unter Koordinatentransformation.*
- *Der Typ kann an verschiedenen Punkt in Ω unterschiedlich sein.*
- *Der Typ hängt nur vom Hauptteil der PDGL ab (außer für parabolische Gleichungen)*
- *Patologische Fälle wie $\frac{\partial^2 u}{\partial u^2} + \frac{\partial u}{\partial x} = 0$ mit der einzigen Lösung $u(x, y) = 0$ werden ausgeschlossen.*

Definition 8.7
Eine Gleichung der Form

$$d(x,y)\frac{\partial u}{\partial x}(x,y) + e(x,y)\frac{\partial u}{\partial y}(x,y) + f(x,y)u(x,y) + g(x,y) = 0 \qquad \text{in } \Omega. \quad (8.4)$$

heißt hyperbolische Gleichung erster Ordnung, falls $|d(x, y)| \cdot |e(x, y)| > 0$ für alle $(x, y) \in \Omega$. Für $n \geq 2$ heißt die Gleichung

$$\mathbf{v}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) + f(\mathbf{x}) \cdot u(\mathbf{x}) + g(\mathbf{x}) = 0$$

hyperbolisch.

Anmerkung 8.8

- *Für nicht-lineare PDGL erster und zweiter Ordnung (d.h. die Koeffizienten hängen von der Lösung u ab) kann der Typ der PDGL räumlich und zeitlich variieren.*
- *In dieser Vorlesung behandeln wir nur skalare PDGL. Auch für Systeme von PDGL gibt es Klassifikationssysteme, die hier aber nicht behandelt werden können.*

8.3.1. Beispiele für verschiedene Typen

Beispiel 8.9 (Poisson-Gleichung)

$$-\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y) \qquad \forall (x,y) \in \Omega \quad (8.5)$$

heißt Poisson-Gleichung.

Diese ist der Prototyp für eine elliptische *PDGL. (8.5) bestimmt die Lösung nicht eindeutig. Mit $u(x, y)$ ist z. B. auch $u(x, y) + c_1 + c_2x + c_3y$ für beliebige c_1, c_2, c_3 eine Lösung. Um u eindeutig festzulegen sind Randbedingungen erforderlich.*

Hierbei gibt es zwei gebräuchliche Bedingungen:

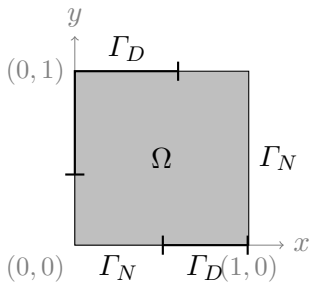
1. *$u(x, y) = \varphi(x, y)$ für $(x, y) \in \Gamma_D \subseteq \partial\Omega$ (Dirichlet[2]-Randbedingung),*

[2]Peter Gustav Lejeune Dirichlet, 1805-1859, dt. Mathematiker.

2. $\frac{\partial u}{\partial \boldsymbol{\nu}}(x,y) = \phi(x,y)$ *für* $(x,y) \in \Gamma_N \subset \partial\Omega$ *(Neumann*[3]*-Randbedingung, Fluss-Randbedingung),*

und $\Gamma_D \cup \Gamma_N = \partial\Omega$*. Wichtig ist auch* $\Gamma_N \neq \partial\Omega$*, da sonst die Lösung nur bis auf eine Konstante bestimmt ist.*

Die vollständige Poisson-Gleichung lautet also



$$\boxed{\begin{aligned} -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} &= f \text{ in } \Omega \\ u &= \varphi \text{ auf } \Gamma_D \subseteq \partial\Omega \\ \frac{\partial u}{\partial \boldsymbol{\nu}} &= \phi \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D \neq \partial\Omega \end{aligned}}$$

Verallgemeinerung auf d *Raumdimensionen:*

$$\begin{aligned} -\sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} =: -\Delta u &= f \text{ in } \Omega \\ u &= \varphi \text{ auf } \Gamma_D \subseteq \partial\Omega \\ \nabla u \cdot \boldsymbol{\nu} &= \phi \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D \end{aligned}$$

Auch diese Gleichung ist elliptisch. Ist $f \equiv 0$ *so spricht man auch von der Laplace-Gleichung.*

Beispiel 8.10 (Allgemeine Diffusionsgleichung)
Sei $\Omega \subset \mathbb{R}^d$ *ein Gebiet und* $K : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ *eine Abbildung, die jedem Punkt* $\mathbf{x} \in \Omega$ *eine* $d \times d$ *Matrix* $\mathbf{K}(\mathbf{x})$ *zuordnet.*

Für $\mathbf{K}(\mathbf{x})$ *fordern wir zusätzlich (für alle* $\mathbf{x} \in \Omega$*)*

1. $\mathbf{K}(\mathbf{x})$ *ist symmetrisch positiv definit*
2. $\min\left\{\boldsymbol{\xi}^{\mathrm{T}}\mathbf{K}(\mathbf{x})\boldsymbol{\xi} \,:\, \|\xi\| = 1\right\} \geq c_0 > 0$ *(uniforme Elliptizität).*

Dann ist

$$\boxed{\begin{aligned} \nabla \cdot (-\mathbf{K}(\mathbf{x})\nabla u(\mathbf{x})) &= f \text{ in } \Omega \\ u &= \varphi \text{ auf } \Gamma_D \subseteq \partial\Omega \\ \Big(-\mathbf{K}(\mathbf{x})\nabla u(\mathbf{x})\Big) \cdot \boldsymbol{\nu}(\mathbf{x}) &= \phi \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D \neq \partial\Omega \end{aligned}} \tag{8.6}$$

die ebenfalls elliptische allgemeine Diffusionsgleichung. In der Praxis ist (8.6) *für sehr variables* $\mathbf{K}$ *schwierig zu lösen.*

Beispiel 8.11 (Wellengleichung)
Der Prototyp einer hyperbolischen Gleichung zweiter Ordnung ist die Wellengleichung:

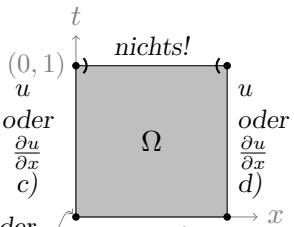
$$-\frac{\partial^2 u}{\partial x^2}(x,t) + \frac{\partial^2 u}{\partial t^2}(x,t) = 0 \qquad \text{in } \Omega \quad . \tag{8.7}$$

[3]Carl Gottfried Neumann, 1832-1925, deutscher Mathematiker.

Als Randwertvorgabe kommt für $\Omega = (0,1)^2$ etwa in Frage:
$x \in [0,1]$:

a) $u(x,0) = u_0(x)$

b) $\dfrac{\partial u}{\partial t}(x,0) = u_1(x)$



Kompatibilität der Randbedingungen für u, $\frac{\partial u}{\partial t}$!

Zwei Anfangsbedingungen wegen $\frac{\partial^2 u}{\partial t^2}$!

$t \in [0,1]$:

c) $u(0,t) = \varphi_0(t) \quad$ *oder* $\dfrac{\partial u}{\partial x}(0,t) = \phi_0(t)$

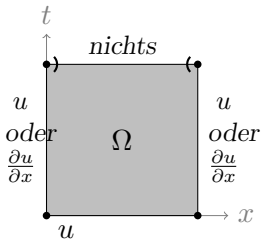
d) $u(1,t) = \varphi_1(t) \quad$ *oder* $\dfrac{\partial u}{\partial x}(1,t) = \phi_1(t)$

Beachte die ausgezeichnete (Zeit-)Richtung t! a) + b) heißen deshalb Anfangswerte und c) + d) Randwerte. Vorgaben auf dem ganze Rand sind nicht möglich!

Beispiel 8.12 (Wärmeleitungsgleichung)
Der Prototyp einer parabolischen Gleichung ist die Wärmeleitungsgleichung:

$$-\frac{\partial^2 u}{\partial x^2}(x,t) + \frac{\partial u}{\partial t}(x,t) = 0 \qquad \text{in } \Omega.$$

Bem.: Das + ist nicht klar, auch − wäre nach Def. 8.4 parabolisch
⇒ zusätzliche Forderung nach Stabilität bzw. sachgemäß gestelltem Problem.



nur eine Randbedingung da PDGL erster Ordnung in t

Als Randwertvorgabe in $\Omega = (0,1)^2$ wählt man für $x \in [0,1], t \in [0,1]$ z.B.

$$u(x,0) = u_0(x)$$
$$u(0,t) = \varphi_0(t)$$
$$u(1,t) = \varphi_1(t)$$

Beispiel 8.13 (Transportgleichung)
Sei $\Omega \subset \mathbb{R}^d, \mathbf{v} : \Omega \to \mathbb{R}^d$ ein gegebenes Vektorfeld. Die Gleichung

$$\nabla \cdot (\mathbf{v}(\mathbf{x}) \cdot u(\mathbf{x})) = f(\mathbf{x}) \qquad \text{in } \Omega$$

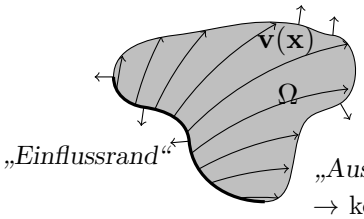
heißt stationäre Transportgleichung und ist hyperbolisch erster Ordnung.

Als Randwertvorgabe kommt in Betracht

$$u(\mathbf{x}) = \varphi(\mathbf{x})$$

für $\mathbf{x} \in \partial\Omega$ *so dass* $\mathbf{v}(\mathbf{x}) \cdot \boldsymbol{\nu}(\mathbf{x}) < 0$ *(Randvorgabe abhängig von den Daten)*

Auch $\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}(\mathbf{x}) \cdot u(\mathbf{x})) = f(\mathbf{x}, t)$ *ist hyperbolisch 1. Ordnung.*



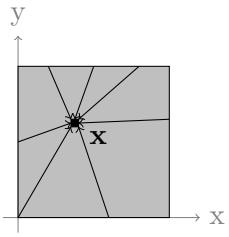
„Einflussrand“ *„Ausflussrand“* → keine *Randbedingung*

8.3.2. Einflussbereich

Der Typ einer partiellen Differentialgleichung wird auch bei folgender Frage deutlich:

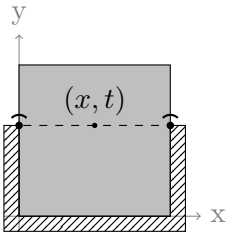
> Gegeben $\mathbf{x} \in \Omega$. Welche Randwerte/Anfangswerte beeinflussen die Lösung u am Punkt $\mathbf{x}$?

Elliptisch $-u_{xx} - u_{yy} = 0$



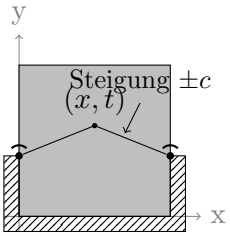
alle Randwerte beeinflussen $u(\mathbf{x})$, d. h. eine Änderung in $u(\mathbf{y})$ für beliebiges $\mathbf{y} \in \partial\Omega$ bewirkt eine Änderung in $u(\mathbf{x})$.

Parabolisch $-u_{xx} + u_t = 0$



für (x, t) beeinflussen alle (x', t') mit $t' \leq y$ den Wert in $\mathbf{x}$.
„unendliche Ausbreitungsgeschwindigkeit“

Hyperbolisch (2. Ordnung) $-u_{xx} + u_{tt} = 0$

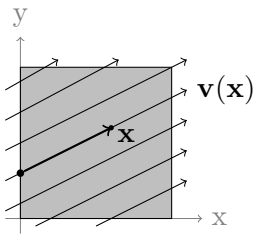


Lösung in (x, t) wird beeinflusst von allen Randpunkten unterhalb des Kegels

$$\{(x', t') : t' \leq (x' - x) \cdot c + t \ \wedge \ t' \leq (x - x') \cdot c + \} \cap \partial\Omega$$

„endliche Ausbreitungsgeschwindigkeit“

Hyperbolisch (1. Ordnung) $u_x + u_t = 0$



Genau ein Randpunkt beeinflusst den Wert.

9. Numerische Lösung elliptischer PDGL

Partielle Differentialgleichungen können nur für einige Sonderfälle (bei einer speziellen Wahl der Form des Gebiets, der Randbedingungen und der Parameterfelder) analytisch gelöst werden. Näherungen der Lösung werden daher mit numerischen Methoden berechnet. Dabei erhält man Näherungen

- der Lösung an diskreten Punkten im Raum (z.B. Finite-Differenzen-Verfahren)
- der Lösung mit einer parametrisierten Funktion (z.B. Finite-Elemente-Verfahren, Diskontinuierliche-Galerkin-Verfahren...)
- bestimmter mathematischer Eigenschaften (Massenerhaltung, Kontinuität von Flüssen) der Gleichung (z.B. Finite-Volumen-Verfahren, Mimetische-Finite-Differenzen-Verfahren, Diskontinuierliche-Galerkin-Verfahren)

9.1. Gitter

Bei all diesen Verfahren wird zunächst das Gebiet Ω in Teilgebiete (Elemente e) mit einer einfachen Geometrie zerlegt (Triangulierung).

- Typische Elementgeometrien sind

  1D Geradenstücke

  2D Dreiecke, Vierecke

  3D Tetraeder, Hexaeder, Pyramiden, Prismen

- Alle Elemente zusammen bezeichnet man als Gitter
- Es ist nicht immer möglich das gesamte Gebiet mit solchen Elementen einfacher Geometrie zu füllen (z.B. bei einer Kugel). Das Gitter sollte aber keine Löcher enthalten und $\lim_{n\to\infty} \bigcup_{i=1}^{n} e_i = \bar{\Omega}$

Abhängig von der Zielsetzung und dem Diskretisierungsverfahren gibt es unterschiedliche Arten von Gittern. Ein Gitter ist

strukturiert wenn es aus gleichförmigen Elementen nach einem einfachen Konstruktionsprinzip zusammengesetzt ist. Typische Beispiele sind Gitter mit rechteckigen Gitterelementen. Man spricht dann von

äquidistanten Gittern wenn die Elementgröße h_i in einer Raumdimension $i \in \{x, y, z\}$ konstant ist.

Tensorprodukt-Gittern wenn die Elementgröße $h_i = f(x_i)$ in einer Raumdimension $i \in \{x, y, z\}$ jeweils nur eine Funktion dieser Raumdimension ist (die Gitterweite in x-Richtung ändert sich also nicht entlang der y- oder z-Koordinate)

unstrukturiert wenn es aus Elementen unterschiedlicher Geometrie und/oder Form zusammengesetzt ist.

konform wenn es keine hängenden Knoten gibt, d.h. wenn die Schnittmenge $e_i \cap e_j$ zwischen zwei Elementen e_i und e_j

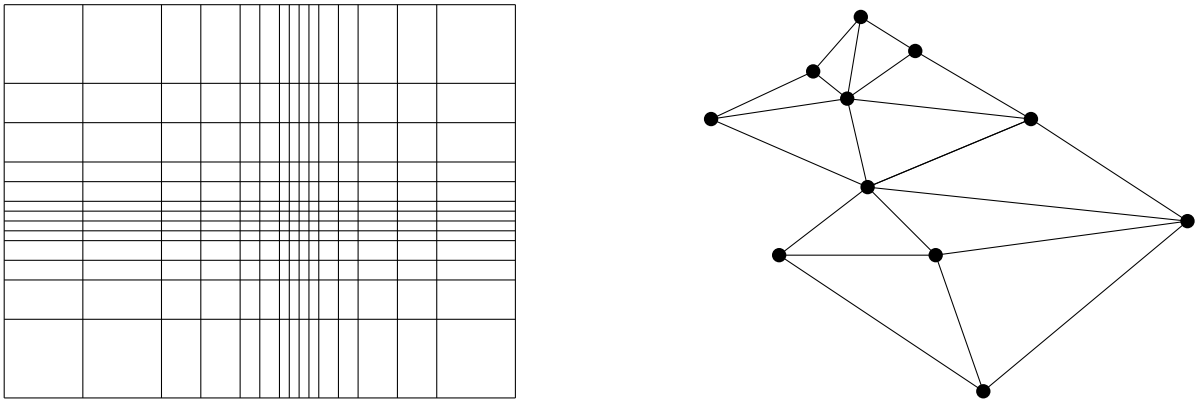


Abbildung 1: Strukturiertes (Tensorprodukt-) Gitter (links) und unstrukturiertes Gitter (rechts).

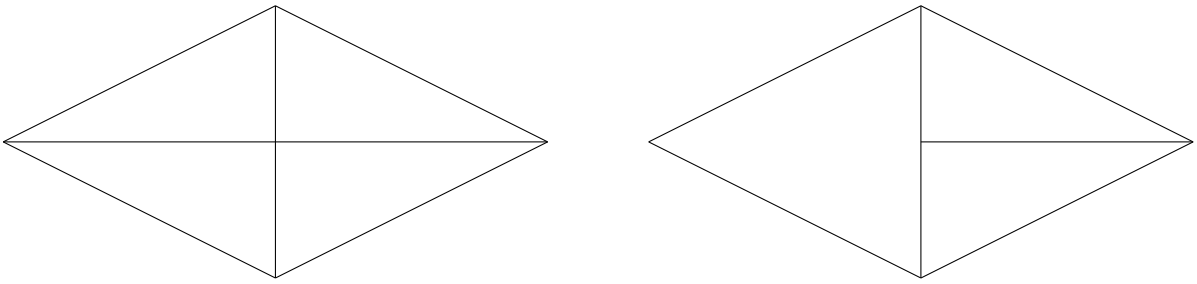


Abbildung 2: Konformes (links) und nicht-konformes Gitter (rechts).

- ein Punkt ist, dann haben die Elemente einen gemeinsam Knoten
- eine Linie ist, dann haben Sie eine gemeinsame Kante
- eine Fläche ist, dann haben Sie eine gemeinsame Seitenfläche

nicht konform wenn es hängende Knoten gibt, d.h. Knoten eines Elements, die nicht gleichzeitig Knoten eines anderen Elements sind, mit dem eine Schnittmenge existiert.

9.2. Finite-Differenzen-Verfahren

Grundidee: Partielle Ableitungen werden durch Differenzenquotienten ersetzt.

9.2.1. 1D-Poisson-Gleichung

$$-\frac{\partial^2 u}{\partial x^2} = f(x) \qquad x \in]0,1[$$
$$u(0) = \varphi_0, \quad u(1) = \varphi_1.$$

Unterteile $\Omega =]0,1[$ in N Teilintervalle ($N \in \mathbb{N}$), ein sogenanntes „äquidistantes Gitter“

$$e_i = [x_i, x_{i+1}], \qquad x_i = i \cdot h, \quad i = 0, \ldots, N, \quad h = \tfrac{1}{N}.$$

$$\Omega_h = \{ih : i \in \mathbb{N} \wedge 0 < i < N\}, \qquad \overline{\Omega}_h = \{ih : i \in \mathbb{N}_0 \wedge 0 \leq i \leq N\}$$

Einseitiger Differenzenquotient (erster Ordnung):

$$u'(x) = \frac{u(x+h) - u(x)}{h} - \underbrace{\frac{h}{2}u''(\xi)}_{\mathcal{O}(h)} \qquad \xi \in]x, x+h[$$

Zentraler Differenzenquotient (zweiter Ordnung) für den Gradienten von u:

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} \quad - \underbrace{\frac{h^2}{12}\left(u'''(\xi^+) + u'''(\xi-)\right)}_{\mathcal{O}(h^2)}$$

Zentraler Differenzenquotient (zweiter Ordnung) für die zweite Ableitung von u:

$$u''(x) = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2} \quad - \underbrace{\frac{h^2}{24}\left(u''''(\xi^+) + u''''(\xi^-)\right)}_{\mathcal{O}(h^2)}.$$

Ersetzt man in der partielle Differentialgleichung die partielle Ableitung durch einen zentralen Differenzenquotienten, erhält man für jeden inneren Gitterpunkt x_i eine Gleichung

$$-\frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} = f(x_i).$$

Dirichlet-Randbedingungen lassen sich dadurch berücksichtigen, dass man $u_0 = \varphi_0$ bzw. $u_n = \varphi_n$ setzt und die entsprechenden Terme auf die rechte Seite bringt. Neumann-Randbedingungen $\frac{\partial u}{\partial x} = \phi$ können z.B. durch einseitige Differenzenquotienten eingebaut werden. Man erhält dann z.B. am rechten Rand die Gleichung

$$\frac{\partial u}{\partial x} = \frac{u(x_1) - u(x_0)}{h} + \mathcal{O}(h) \approx \frac{u(x_1) - u(x_0)}{h} = \phi_0$$

Nehmen wir nun an, dass Dirichlet-Randbedingungen $u_0 = \varphi_0$ und $u_N = \varphi_N$ gegeben sind und fassen die Lösung an den Gitterpunkten in einem Vektor $\mathbf{u}_h$ im $\mathbb{R}^{N-1}$ zusammen:

$$\mathbf{u}_h = (u_h(h), u_h(2h), \ldots, u_h(1-h))^T$$

Bringen wir die (gegebenen) Randwerte auf die rechte Seite, erhalten wir das lineare Gleichungssystem für $\mathbf{u}_h$

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}}_{\mathbf{L}_h} \underbrace{\begin{pmatrix} u_h(h) \\ u_h(2h) \\ u_h(3h) \\ \vdots \\ \vdots \\ u_h(1-h) \end{pmatrix}}_{\mathbf{u}_h} = \underbrace{\begin{pmatrix} f(h) + \frac{\varphi_0}{h^2} \\ f(2h) \\ \vdots \\ \vdots \\ f(1-2h) \\ f(1-h) + \frac{\varphi_N}{h^2} \end{pmatrix}}_{\mathbf{q}_h}$$

Das lineare Gleichungssystem hat folgende Eigenschaften:

- $\mathbf{L}_h$ ist eine Triagonalmatrix
- $\mathbf{L}_h$ ist dünn besetzt (nur $\mathcal{O}(N)$ von N^2 Einträgen sind ungleich Null, maximal 3 Einträge pro Zeile)
- $\mathbf{L}_h$ ist symmetrisch und positiv definit
- Dieses LGS ist in linearer Laufzeit mit dem Gauß-Algorithmus (dieser heißt für Tridiagonalmatrizen auch Thomas-Verfahren) lösbar.

9.2.2. 2D-Poisson-Gleichung

In 2D lautet die Poisson-Gleichung

$$-\Delta u(x,y) = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x,y)$$

Überziehen wir Ω mit einem Gitter $x_i = i \cdot h$ und $y_j = j \cdot h$ für $1 \leq i,j \leq N$ und ersetzen an jedem Gitterpunkt $\frac{\partial^2 u}{\partial x^2}$ und $\frac{\partial^2 u}{\partial y^2}$ mit einem zentralen Differenzenquotienten

$$\Delta u(x_i,y_j) \approx \frac{u(x_{i-1},y_j) - 2u(x_i,y_j) + u(x_{i+1},y_j)}{h^2} + \frac{u(x_i,y_{j-1}) - 2u(x_i,y_j) + u(x_i,y_{j+1})}{h^2},$$

erhalten wir für jeden Gitterpunkt eine lineare Gleichung

$$\frac{4u(x_i,y_j) - u(x_{i-1},y_j) - u(x_{i+1},y_j) - u(x_i,y_{j-1}) - u(x_i,y_{j+1})}{h^2} = f(x_i,y_j).$$

Wenn die Gitterpunkte Zeile für Zeile nummerieren sind (lexikographische Anordnung) und mit $4N$-Dirichlet-Randbedingungen erhält man das lineare Gleichungssystem

$$\underbrace{\frac{1}{h^2}\begin{pmatrix}
4 & -1 & & -1 & & & & & & & & \\
-1 & \ddots & -1 & & \ddots & & & & & & & \\
 & -1 & 4 & & & -1 & & & & & & \\
\hline
-1 & & & 4 & -1 & & -1 & & & & & \\
 & \ddots & & -1 & \ddots & -1 & & \ddots & & & & \\
 & & -1 & & -1 & 4 & & & -1 & & & \\
\hline
 & & & -1 & & & 4 & -1 & & -1 & & \\
 & & & & \ddots & & -1 & \ddots & -1 & & \ddots & \\
 & & & & & -1 & & -1 & 4 & & & -1 \\
\hline
 & & & & & & -1 & & & 4 & -1 & \\
 & & & & & & & \ddots & & -1 & \ddots & -1 \\
 & & & & & & & & -1 & & -1 & 4
\end{pmatrix}}_{\mathbf{L}_h} \cdot \underbrace{\begin{pmatrix} u(h,h) \\ u(2h,h) \\ \vdots \\ \hline u(h,2h) \\ \hline \\ \hline \\ \hline \\ \end{pmatrix}}_{\mathbf{u}_h} = \underbrace{\begin{pmatrix} f(h,h) + {}^{[\varphi(0,h)+\varphi(h,0)]}/_{h^2} \\ f(2h,h) + {}^{\varphi(h,0)}/_{h^2} \\ \hline \\ \hline \\ \hline \\ \end{pmatrix}}_{\mathbf{q}_h}$$

Jeder Gitterpunkt trägt zu dieser Matrix fünf Punkte bei, die sich aus fünf Gitterpunkten ergeben, die in einem sogenannten Differenzenstern angeordnet sind:

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

Im dreidimensionalen erhält man entsprechend einen „Siebenpunktstern“. Schon im zweidimensionalen, noch mehr aber in 3D, können die Gleichungssysteme sehr groß werden. Eine wesentliche Herausforderung ist dann die effiziente Lösung der LGS. Man verwendet dafür i.d.R. spezielle iterative Lösungsverfahren.

9.2.3. Konvergenz

Bei einem zentralen Differenzenquotienten erhalten wir auf äquidistanten Gittern eine Näherung der Lösung an den Gitterpunkten, deren Fehler mit $\mathcal{O}(h^2)$ gegen Null geht für $h \to 0$. Bei einem einseitigen Differenzenquotienten (z.B. am Rand) oder auf nicht-äquidistanten Gittern geht der Fehler nur mit $\mathcal{O}(h)$ gegen Null für $h \to 0$. Für dieses Konvergenzabschätzung muss u jedoch viermal-stetig differenzierbar sein, was nicht immer erfüllt ist.

9.2.4. Wesentliche Eigenschaften des Finite-Differenzen-Verfahrens

Vorteile:

- Die Gleichungen sind einfach aufzustellen.
- Das Verfahren ist einfach zu implementieren.
- Es ist gut geeignet für strukturierte Gitter.

Nachteile:

- Das FD-Verfahren hat nur lineare Konvergenzrate für nicht-äquidistante Gitter.
- Die Beschreibung komplizierterer Gebiete schwierig (z.B. bei gekrümmten Rändern).
- Was ist der Wert von u zwischen zwei Gitterpunkten?
- In der Regel werden Erhaltungsgleichung (z.B. Massenerhaltung) nicht diskret erfüllt.

9.3. Finite-Elemente-Verfahren

Die sogenannten „Finite-Elemente-Verfahren“ eignen sich auch für unstrukturierte Gitter und damit für Gebiete mit komplexer Geometrie.

9.3.1. Schwache Formulierung

Zentraler Baustein von Finite-Elemente-Verfahren ist die sogenannte „schwache Formulierung“. Dazu wählt man eine Testfunktion v, z.B. aus dem Raum der einmal stetig-differenzierbaren Funktionen, die auf dem Rand von Ω Null sind:

$$v \in \left\{ v \in \mathcal{C}^1(\overline{\Omega}) : v(\mathbf{x}) = 0 \; \forall \, \mathbf{x} \in \partial\Omega \right\}.$$

Wir verlangen nun nicht mehr, dass die PDGL

$$F\left(\frac{\partial^m u}{\partial x_1^m}(\mathbf{x}), \frac{\partial^{m-1} u}{\partial x_1^{m-1}}(\mathbf{x}), \ldots, \frac{\partial^m u}{\partial x_1^{m-1} \partial x_2}(\mathbf{x}), \ldots, \frac{\partial^m u}{\partial x_n^m}(\mathbf{x}), \frac{\partial^{m-1} u}{\partial x_n^{m-1}}(\mathbf{x}), \ldots, u(\mathbf{x}), \mathbf{x} \right) = 0$$

punktweise erfüllt ist, sondern nur im Integral über Ω multipliziert mit der Testfunktion v:

$$\int_\Omega F\left(\frac{\partial^m u}{\partial x_1^m}(\mathbf{x}), \frac{\partial^{m-1} u}{\partial x_1^{m-1}}(\mathbf{x}), \ldots, \frac{\partial^m u}{\partial x_1^{m-1}\partial x_2}(\mathbf{x}), \ldots, \frac{\partial^m u}{\partial x_n^m}(\mathbf{x}), \frac{\partial^{m-1} u}{\partial x_n^{m-1}}(\mathbf{x}), \ldots, u(\mathbf{x}), \mathbf{x}\right) \cdot v(\mathbf{x})\ \mathrm{d}\mathbf{x} = 0.$$

Dies bezeichnet man als „schwache Formulierung“ der PDGL. Für eine klassische Lösung der PDGL, ist diese Integralgleichung für jede beliebige Funktion v erfüllt. Eine Lösung der „schwachen Formulierung“ erfüllt die PDGL jedoch nur in einer Art gewichtetem Mittel.

Beispiel 9.1
Für die Poisson-Gleichung ist die schwache Formulierung:

$$-\int_\Omega \Delta u \cdot v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x.$$

Durch partielle Integration erhalten wir (weil v stetig differenzierbar und auf dem Rand Null ist)

$$-\int_\Omega \Delta u \cdot v \,\mathrm{d}x = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_{\partial\Omega} \frac{\partial u}{\partial \boldsymbol{\nu}} \cdot v \,\mathrm{d}s = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x$$

und damit

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x.$$

Eine Lösung u dieser Gleichung muss nur einmal stetig differenzierbar sein (im Gegensatz zu $u \in \mathcal{C}^2(\overline{\Omega})$ für eine klassische Lösung).

9.3.2. Test- und Ansatzfunktionen

Um eine Näherung $y(\mathbf{x})$ der Lösung zu berechnen, wählt man eine Ansatzfunktion als Linearkombination von N linear unabhängigen Basisfunktionen (i.d.R. Polynome)

$$y(x) = \sum_{i=1}^{N} c_i \cdot \psi_i(\mathbf{x}).$$

Diese Ansatzfunktion setzt man in die schwache Formulierung ein. Zusammen mit einer Wahl von N geeigneten (ebenfalls linear unabhängigen) Testfunktionen v führt dies dann auf ein Gleichungssystem zur Bestimmung der Koeffizienten c_i.

Wegen der Linearität der Integration erhält man eine Summe über Integrale für jede Basisfunktion v_k:

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega \nabla \left(\sum_{i=1}^{N} c_i \cdot \psi_i(\mathbf{x})\right) \cdot \nabla v \,\mathrm{d}x = \sum_{i=1}^{N} c_i \int_\Omega \nabla \psi_i \cdot \nabla v_k \,\mathrm{d}x$$

Die Berechnung jedes dieser Integrale über das gesamte Gebiet wäre sehr teuer. Außerdem hätte man dann jeweils Polynome mit sehr vielen Stützstellen. Daher diskretisiert man das

Gebiet und erhält ein Gitter aus Elementen. Auf jedem Element wählt man einen Satz von Basisfunktionen, die nur auf diesem Element ungleich Null sind, aber auf allen anderen Elementen gleich Null. Die Ansatzfunktion wird damit

$$y(\mathbf{x}) = \sum_{e_i} \sum_{j=1}^{N_{e_i}} c_{e_i,j} \cdot \psi_{e_i,j}(\mathbf{x}),$$

wobei N_{e_i} die Anzahl der Basisfunktionen ist, die auf dem Element e_i ungleich Null sind.

Auch die Testfunktionen wählt man Element-weise. Die Integrale über das Gesamtgebiet können jetzt durch eine Summe von Integralen über die einzelnen Elemente ersetzt werden.

$$\sum_{i=1}^{N} c_i \int_{\Omega} \nabla\psi_i \cdot \nabla v \,\mathrm{d}x = \sum_{e_i} \sum_{j=1}^{N_{e_i}} c_{e_i,j} \int_{e_i} \nabla\psi_{e_i,j} \cdot \nabla v_{e_i,k} \,\mathrm{d}x$$

Die Koeffizienten in der Zeile des linearen Gleichungssystems, die man für die Basisfunktion $v_{e_i,k}$ erhält, die Integrale

$$\int_{e_i} \nabla\psi_{e_i,j} \cdot \nabla v_{e_i,k} \,\mathrm{d}x,$$

sind dabei für eine bestimmte schwache Formulierung nur vom Gitter, sowie von der Wahl der Basis- und Testfunktionen abhängig

Verschiedene Finite-Elemente-Verfahren unterschieden sich in der Wahl der Basis- und der Testfunktionen. Als Test- und Basisfunktionen können die gleichen Funktionen verwendet werden (sogenannte „Galerkin-Verfahren“).

Eine beliebte Wahl von Basisfunktionen ist eine Lagrange-Basis, bei der die Parameter c_i gerade die Werte der gesuchten Funktion an bestimmten Knoten in jedem Element sind (zunächst den Werten an den Ecken, dann an zusätzlichen Stellen an den Kanten oder Seiten). Jede Basisfunktion ist dann an genau einem Knoten Eins und an allen anderen Null (wie bei der Langrage-Interpolation). Für ein konformes Gitter erhält man dadurch eine global stetige Lösung. Die Basisfunktionen definiert man auf einem sogenannten Referenzelement (z.B. dem Einheitsquadrat oder dem Einheitswürfel) und transformiert sie auf die reale Geometrie des Elements. Abbildung 3 und 4 zeigen Basisfunktionen erster und zweiter Ordnung auf dem Referenzdreieck (ein rechtwinkliges Dreieck mit Kantenlänge 1).

9.3.3. Eindimensionale Poisson-Gleichung

Wir wollen das Finite-Elemente-Verfahren auf die eindimensionale Poisson-Gleichung

$$-\frac{\partial^2 u(x)}{\partial x^2} = f(x) \qquad \text{in }]0,1[$$

mit den Randbedingungen $u(0) = 0$ und $u(1) = 0$ anwenden. Wir verwenden dabei wieder ein äquidistantes Gitter. Als Ansatzfunktionen verwenden wir die sogenannten Hutfunktionen $\psi_i, \quad i = 1, ..., n-1$

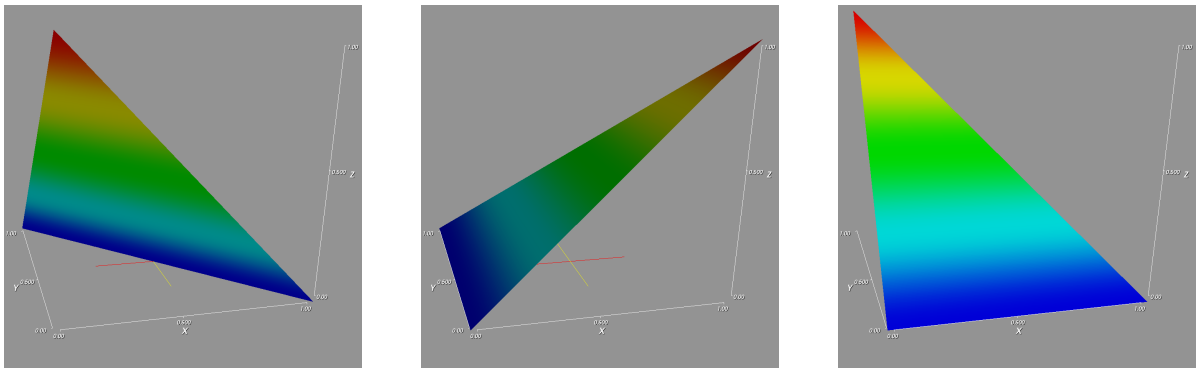


Abbildung 3: Basisfunktionen erster Ordnung auf dem Referenzdreieck (Rechteckiges Dreieck mit Kantenlänge 1). Die Knoten liegen hier auf den Ecken.

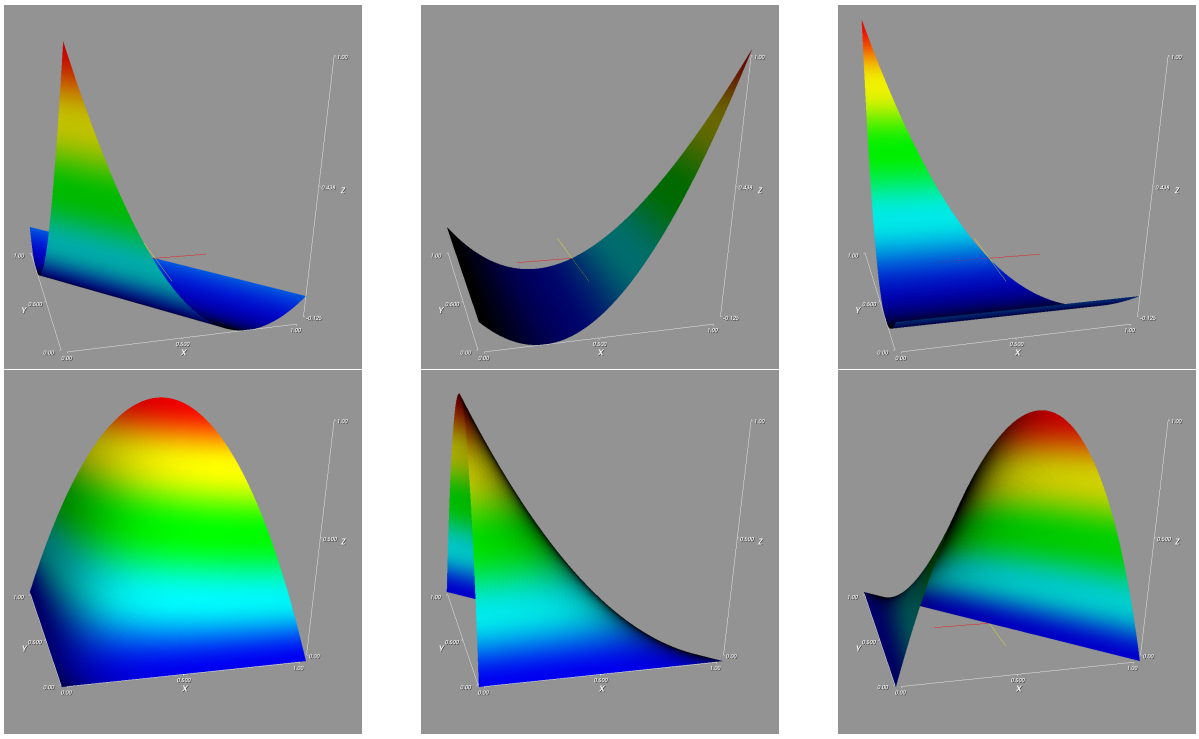
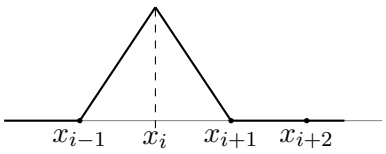


Abbildung 4: Basisfunktionen zweiter Ordnung auf dem Referenzdreieck. Hier gibt es zusätzliche Knoten an den Kantenmitten.



$$\psi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}} = \frac{x-x_{i-1}}{h} & x \in]x_{i-1}, x_i[\\ \frac{x-x_{i+1}}{x_i-x_{i+1}} = -\frac{x-x_{i+1}}{h} & x \in]x_i, x_{i+1}[\\ 0 & \text{sonst} \end{cases}$$

mit der Eigenschaft

$$\psi_i(x_j) = \begin{cases} 1 & i = j \\ 0 & \text{sonst} \end{cases}.$$

Dies entspricht einer Wahl von linearen Basisfunktionen auf jedem Element mit Freiheitsgraden an den Punkten x_i. Für jede Testfunktion v_j bekommen wir eine Gleichung

$$-\int_0^1 \frac{\partial^2 y}{\partial x^2} \cdot v_j \,\mathrm{d}x = \int_0^1 f \cdot v_j \,\mathrm{d}x$$

und durch partielle Integration

$$-\left[-\int_0^1 \frac{\partial y}{\partial x} \cdot \frac{\partial v_j}{\partial x} \,\mathrm{d}x + \frac{\partial y}{\partial x}(1) \cdot v_j(1) - \frac{\partial y}{\partial x}(0) \cdot v_j(0)\right] = \int_0^1 f \cdot v_j \,\mathrm{d}x$$

da $v_j(x)$ auf dem Rand Null ist folgt

$$\int_0^1 \frac{\partial y}{\partial x} \cdot \frac{\partial v_j}{\partial x}\, dx = \int_0^1 f \cdot v_j \, dx$$

Wir verwenden die Hutfunktionen sowohl als Ansatz- als auch als Testfunktionen (Galerkin-Verfahren). Mit $y(x) = \sum_i^N y_i \cdot \psi_i(x)$ erhalten wir für jede Testfunktion eine Zeile eines linearen Gleichungssystems:

$$y_{i-1} \int_{x_{i-1}}^{x_i} \frac{\partial \psi_{i-1}}{\partial x} \cdot \frac{\partial \psi_i}{\partial x} dx + y_i \int_{x_{i-1}}^{x_{i+1}} \frac{\partial \psi_i}{\partial x} \cdot \frac{\partial \psi_i}{\partial x} dx + y_{i+1} \int_{x_i}^{x_{i+1}} \frac{\partial \psi_{i+1}}{\partial x} \cdot \frac{\partial \psi_i}{\partial x} dx = \int_{x_{i-1}}^{x_{i+1}} f \cdot \psi_i dx$$

Alle anderen Terme sind Null, da die Testfunktion $\psi_i(x)$ außerhalb des Intervalls $]x_{i-1}, x_{i+1}[$ Null ist. Die Integrale sind:

$$\int \frac{\partial \psi_i}{\partial x} \cdot \frac{\partial \psi_k}{\partial x} dx = \begin{cases} \int_{x_i-h}^{x_i} \frac{1}{h} \cdot \frac{1}{h} dx + \int_{x_i}^{x_i+h} (-\frac{1}{h}) \cdot (-\frac{1}{h}) dx = \frac{1}{h} + \frac{1}{h} = \frac{2}{h} & k = i \\ \int_{x_i}^{x_i+h} (-\frac{1}{h}) \cdot \frac{1}{h} dx = -\frac{1}{h} & k = i \pm 1 \\ & \quad x_i \; x_{i+1} \\ 0 & \text{else} \end{cases}$$

Integration der rechten Seite mit der Trapezregel liefert

$$\int_{x_{i-1}}^{x_{i+1}} f \cdot \psi_i dx \approx \frac{h}{2} \cdot (0 \cdot f_{i-1} + f_i + f_i + 0 \cdot f_{i+1}) = h \cdot f_i$$

und damit erhalten wir die Gleichung

$$\frac{1}{h}(-y_{i-1} + 2y_i - y_{i+1}) = h \cdot f_i$$

Für eindimensionale äquidistante Gitter erhalten wir mit dem Finite-Differenzen und Finite-Elemente-Verfahren also die exakt gleiche Lösung (in etwas unterschiedlicher Schreibweise):

$$-\frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} = f(x_i)$$
$$\frac{1}{h}(-y_{i-1} + 2y_i - y_{i+1}) = h \cdot f_i$$

Das Finite-Elemente-Verfahren lässt sich aber wesentlich einfacher auf unstrukturierte Gitter übertragen.

9.3.4. Randbedingungen und Konvergenzrate

- Dirichlet-Randbedingungen können direkt in die Ansatzfunktion eingebaut werden.
- Neumann-Randbedingungen werden bei der Berechnung der Integrale berücksichtigt und ergeben zusätzliche Terme auf der rechten Seite.
- Die Konvergenzordnung hängt von der konkreten Wahl der Test- und Ansatzfunktionen ab.
- Die Integrale werden häufig mit numerischer Integration berechnet. Wenn eine Quadraturformel mit genügend hoher Ordnung verwendet wird, wird die gleiche Konvergenzordnung erzielt, wie bei analytischer Berechnung der Integrale.

9.3.5. Wesentliche Eigenschaften des Finite-Elemente-Verfahrens

Vorteile:

- Das Finite-Elemente-Verfahren lässt sich für Gebiete mit beliebiger Form verwenden
- Es liefert Näherungen der Lösung an jedem Punkt des Gebiets.
- Es ist gut geeignet für unstrukturierte Gitter.
- Lokale Gitterverfeinerung zur Erzielung hoher Genauigkeit bei überschaubarem Rechenaufwand ist möglich.

Nachteile:

- Die Gittererzeugung kann schwierig sein, da das Gitter oft gewisse Eigenschaften haben muss (z.B. keine Elemente mit sehr spitzen Winkeln).
- Für einfache Probleme ist es rechenaufwändiger als Finite-Differenzen.
- In der Regel werden Erhaltungsgleichung (z.B. Massenerhaltung) nicht diskret erfüllt.

9.4. Finite-Volumen und Discontinuous-Galerkin-Verfahren

Bei Finite-Volumen-Verfahren und Discontinuous-Galerkin-Verfahren (die als eine Verallgemeinerung von Finite-Volumen-Verfahren angesehen werden können) wird eine diskrete Erfüllung von Erhaltungsgleichungen angestrebt.

Betrachten wir dazu die Poisson-Gleichung als Erhaltungsgleichung mit einer Flussdichte $\mathbf{j} = -\nabla u$ und einem Quell-/Senkenterm f:

$$-\Delta u = \nabla \cdot (-\nabla u) = \nabla \cdot \mathbf{j} = f$$

Mit einer schwachen Formulierung und einer Element-weisen Wahl der Basis- und Testfunktionen erhalten wir

$$\int_\Omega v \cdot \nabla \cdot \mathbf{j} \, \mathrm{d}\mathbf{x} \approx \sum_{e_i \in \Omega} \int_{e_i} v \cdot \nabla \cdot \mathbf{j} \, \mathrm{d}\mathbf{x} = \sum_{e_i \in \Omega} \int_{e_i} v \, f \, \mathrm{d}\mathbf{x}$$

Jetzt wenden wir auf jedem Element den Gauß'schen Satz an:

$$\sum_{e_i \in \Omega} \int_{e_i} v \cdot \nabla \cdot \mathbf{j} \, \mathrm{d}\mathbf{x} = \sum_{e_i \in \Omega} \left(- \int_{e_i} \mathbf{j} \cdot \nabla v \, \mathrm{d}\mathbf{x} + \int_{\partial e_i} v \cdot \mathbf{j} \cdot \boldsymbol{\nu} \, \mathrm{d}s \right) = \sum_{e_i \in \Omega} \int_{e_i} v \, f \, \mathrm{d}\mathbf{x}$$

Jetzt fordert man, dass die Gleichung Element-weise erfüllt ist:

$$- \int_{e_i} \mathbf{j} \cdot \nabla v \, \mathrm{d}\mathbf{x} + \int_{\partial e_i} v \cdot \mathbf{j} \cdot \boldsymbol{\nu} \, \mathrm{d}s = \int_{e_i} v \, f \, \mathrm{d}\mathbf{x} \qquad \forall e_i \in \Omega.$$

Diese Formulierung führt dazu, dass Änderungen der Flüsse genau dem Integral über Quellen im Element entsprechen, man erhält eine lokale Massenerhaltung.

Bei den Finite-Volumen-Verfahren verwendet man nun Testfunktionen, die auf jedem Element konstant sind, dadurch fällt der erste Term auf der linken Seite weg. Bei Diskontinuierlichen-Galerkin-Verfahren macht man dies nicht, sondern wählt auf jedem Element beliebige Basisfunktionen. Dies führt zu einer Lösung, die nicht mehr stetig ist. Man bestraft jedoch Sprünge in der Lösung durch zusätzliche Terme, so dass die Lösung für beliebig feine Gitter $h \to 0$ wieder stetig wird.

10. Numerische Lösung parabolischer PDGL

Der Prototyp einer parabolischen Gleichung ist die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} - \Delta u = 0 \qquad \text{für } (x,t) \in [t_0, T] \times \Omega$$

Zur Lösung verwendet man i.d.R. einen sogenannten „Method of Lines“-Ansatz. Dazu wird die PDGL zunächst im Raum diskretisiert, was auf ein System von gewöhnlichen Differentialgleichungen führt. Dieses wird dann z.B. mit einem Runge-Kutta-Verfahren gelöst.

Wir wollen das Vorgehen am Beispiel der eindimensionale Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad \text{für } (x,t) \in [t_0, T] \times]0:1[$$

zeigen. Die räumliche Diskretisierung mit dem Finite-Differenzen-Verfahren und zentralen Differenzenquotienten ergibt für den Gitterpunkt x_i die gewöhnliche Differentialgleichung

$$\frac{\partial u(x_i,t)}{\partial t} = \frac{u(x_{i-1},t) - 2u(x_i,t) + u(x_{i+1},t)}{h^2}$$

Wir erhalten ein System von Anzahl Gitterpunkten gewöhnlicher Differentialgleichungen. Diese können wir z.B. mit dem expliziten Euler-Verfahren auf einem Zeitgitter $t_i = t_0 + i\tau$ lösen und erhalten für x_i:

$$\begin{aligned} u(x_i,t_i) &= u(x_i,t_{i-1}) + \tau \cdot \frac{u(x_{i-1},t_{i-1}) - 2u(x_i,t_{i-1}) + u(x_{i+1},t_{i-1})}{h^2} \\ &= u(x_i,t_{i-1}) + \frac{\tau}{h^2} \cdot \Big(u(x_{i-1},t_{i-1}) - 2u(x_i,t_{i-1}) + u(x_{i+1},t_{i-1}) \Big) \end{aligned}$$

Mit gegebenen Anfangsbedingungen $u(x,t_0)$ lässt sich die Lösung direkt ausrechnen. Da das explizite Euler-Verfahren nicht A-stabil ist, gibt es eine Beschränkung für den maximalen Zeitschritt τ. Es lässt sich zeigen, dass aus Stabilitätsgründen

$$\tau \leq \frac{1}{2} h^2.$$

sein muss. Bei feinen Gittern wird der Zeitschritt somit sehr klein.

Da Wärmeleitung als diffusiver Prozess zu einer immer glatteren Lösung und damit immer langsameren Änderungen führt, möchte man den Zeitschritt gerne beliebig wählen können. Dies ist durch die Auswahl eines (impliziten) A- oder L-stabilen Verfahrens für die Zeitdiskretisierung möglich. Allerdings muss dann in jedem Zeitschritt ein sehr großes Gleichungssystem gelöst werden. Bei 100 Elementen in jede Raumrichtung erhält man in 2D 10.000 und in 3D bereits eine Million Gleichungen. Auch hier werden meist iterative Lösungsverfahren verwendet.