Analysis und Numerik von Differentialgleichungen

Benedikt Wirth

Inhaltsverzeichnis

1 Einleitung

Ziel: Einführung in ausgewählte, grundlegende Konzepte der angewandten Mathematik

Angewandte Mathematik: Mathematische Disziplinen, deren Ergebnisse oft in anderen Wissenschaften (Physik, Biologie, Medizin, Meteorologie, Geistes- & Sozialwissenschaften) angewandt werden, z.B. Analysis von (partiellen) Differentialgleichungen, Dynamische Systeme, Numerische Analysis, Approxima- tions-Theorie, Wissenschaftliches Rechnen & Komplexitätstheorie, Wahrscheinlichkeitstheorie & Stochastik & Statistik, Maschinelles Lernen, Variationsrechnung & Optimierung, ...

Bemerkung 1 (Angewandte Mathematik)**.**

- *Begriff ist nicht klar definiert und umstritten*
- *einige Anwendungen (Relativitätstheorie, Quantenmechanik) zählt man eher nicht dazu (stattdessen zu Differentialgeometrie, Operatoralgebren)*
- *Modellierung, die Übersetzung von Anwendungsproblemen in mathematische Probleme, zählt man typischerweise nicht zur Mathematik, gehört aber zu den Kompetenzen von Mathematikern*

Definition 2 (Stetig differenzierbare Funktionen)**.** *$C^n(I)$ für $n \in \mathbb{N}$ und $I := [a, b] \subset \mathbb{R}$ bezeichnet die* Menge der n-mal stetig differenzierbaren Funktionen $I \to \mathbb{R}$.

Definition 3 (Gewöhnliche Differentialgleichung)**.** *Seien $n \in \mathbb{N}$, $I := [a, b] \subset \mathbb{R}$ und $F : I \times \mathbb{R}^{n+1} \to \mathbb{R}$. Eine (skalare)* gewöhnliche Differentialgleichung *(gDgl) n-ter Ordnung für $y \in C^n(I)$ ist eine Gleichung der Form*

$$F(x, y(x), y'(x), y''(x), \dots, y^{(n)}(x)) = 0 \quad \forall x \in I.$$

Falls $F(x, y_0, \dots, y_n) = f(x, y_0, \dots, y_{n-1}) - y_n$ für ein $f : I \times \mathbb{R}^n \to \mathbb{R}$, sodass die Gleichung folgender Form ist,

$$y^{(n)}(x) = f(x, y(x), y'(x), y''(x), \dots, y^{(n-1)}(x)) \quad \forall x \in I,$$

dann heißt die Differentialgleichung explizit *(sonst* implizit*). Hängen F oder f nicht vom ersten Argument ab, heißt die Differentialgleichung* autonom*. Hängt f nur vom ersten Argument ab, heißt sie* elementar*. Eine Funktion $y \in C^n(I)$, die die Differentialgleichung erfüllt, heißt* Lösung der gewöhnlichen Differentialgleichung.
Bei einer vektor- ($\mathbb{R}^m$-)wertigen gewöhnlichen Differentialgleichung *sind $F : I \times (\mathbb{R}^m)^{n+1} \to \mathbb{R}^m$, $f : I \times (\mathbb{R}^m)^n \to \mathbb{R}^m$ und $y \in C^n(I)^m$.*

Fragen der Vorlesung:

- Existenz einer Lösung?
- Eindeutigkeit?
- Regularität der Lösung?
- (Langzeit-)verhalten der Lösung?
- Numerische Approximationsverfahren? (I.A. kann Lösung nur numerisch gefunden werden)
- Approximationsfehler?
- Approximationsaufwand?

Gewöhnliche Differentialgleichungen sind die kleine Schwester von & Vorbereitung auf partielle Differentialgleichungen mit vielen ungelösten Forschungsfragen, z.B. eines der 7 Millenium Prize Problems:

> Beweise oder widerlege: In 3D + Zeit, gegeben ein glattes Geschwindigkeitsfeld, existiert eine glatte global definierte Lösung der Navier–Stokes-Gleichungen.

2 Beispiele für Differentialgleichungsmodelle

1. Populationsdynamik. (Explizite Differentialgleichung erster Ordnung mit Anfangsbedingung)
   $u(t)$ = Größe einer Population zu Zeit t, $u(0) = u_0 > 0$, sage $u(t)$ vorher!

$$u(t) - u(0) = \text{Geburten} - \text{Tode in } [0,t] = \int_0^t \text{Geburtsrate } G(s) - \text{Sterberate } S(s)\,\mathrm{d}s$$
$$\text{oder} \qquad u'(t) = G(t) - S(t).$$

   Beide Raten sind proportional zu $u(t)$, d.h.

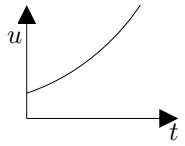
$$u'(t) = C(t,u(t))u(t), \quad u(0) = u_0,$$

   d.h. explizite Differentialgleichung erster Ordnung mit Anfangsbedingung.

   Der Proportionalitätsfaktor $C(t,u(t))$ kann von der Zeit abhängen (z.B. ist Geburtenrate im Sommer höher) und von $u(t)$ (etwa bei Überbevölkerung, ist $u(t)$ groß, wird C klein).

   (a) *Einfachstes Modell (exponentielles Wachstum):*
   $C(t,u(t)) = \lambda = konst.$ (passt z.B. gut für Bakterien in Nährlösung)

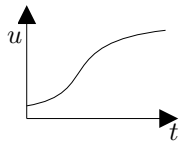
$$\begin{aligned} &\Longrightarrow \quad \lambda = u'(t)/u(t) = (\log u(t))' \\ &\Longrightarrow \quad \log u(t) - \log u_0 = \lambda t \\ &\Longrightarrow \quad u(t) = u_0 e^{\lambda t}. \end{aligned}$$



   Langzeitverhalten $t \mapsto \infty$:
   - für $\lambda < 0$ stirbt Population aus
   - für $\lambda > 0$ wächst Population exponentiell (unrealistisch wegen Nahrungsbegrenzung)

   (b) *Modell mit max. Umwelt-Kapazität für eine Populationsgröße M (logistisches Wachstum):*
   $C(t,u(t)) = \lambda\,(M - u(t))$ mit $\lambda > 0$, d.h. $C \geq 0 \Leftrightarrow$ ausreichend viel Platz, Nahrung, etc.

$$\begin{aligned} & u'(t) = \lambda(M - u(t))\,u(t) \\ \Longrightarrow \quad & \lambda = \frac{u'(t)}{u(t)\,(M-u(t))} = \left(\frac{1}{u(t)} + \frac{1}{M-u(t)}\right)\frac{u'(t)}{M} = \left(\frac{\log u - \log(M-u)}{M}\right)'(t) \\ \Longrightarrow \quad & \log\frac{u(t)}{M-u(t)} - \log\frac{u_0}{M-u_0} = M\lambda t \\ \Longrightarrow \quad & \frac{u(t)}{M-u(t)} = \frac{u_0}{M-u_0}e^{M\lambda t} \\ \Longrightarrow \quad & u(t) = \frac{M}{1 + \frac{M-u_0}{u_0}e^{-M\lambda t}}. \end{aligned}$$

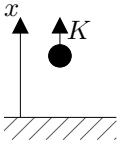


   Langzeitverhalten:

$$u(t) \xrightarrow{t\to\infty} M.$$

2. Freier Fall/Gravitation. (Explizite Differentialgleichung zweiter Ordnung mit Anfangsbedingung)

$$\begin{aligned} x(t) &= \text{Höhe eines Balls der Masse } m \\ v(t) &= x'(t) = \text{Geschwindigkeit} \\ a(t) &= v'(t) = \text{Beschleunigung} \\ K(t,x(t),x'(t)) &= \text{auf Ball wirkende Kraft} \end{aligned}$$



   Newtonsches Kraftgesetz:

$$ma(t) = K(t,x(t),x'(t)) \quad \Rightarrow \quad x''(t) = \frac{1}{m}K(t,x(t),x'(t)),$$

   d.h. explizite Differentialgleichung zweiter Ordnung; Anfangsbedingungen $x(0)$, $v(0)$

(a) *Ohne Luftwiderstand:*
Erbeschleunigung $g = 9.81 \frac{\mathrm{m}}{\mathrm{s}^2} \Longrightarrow K(t, x(t), x'(t)) = -mg.$

$$\begin{aligned} \Longrightarrow \quad & x''(t) = -\frac{1}{m} mg = -g \\ \Longrightarrow \quad & x'(t) = -gt + v(0) \\ \Longrightarrow \quad & x(t) = -\frac{1}{2} gt^2 + v(0)t + x(0) \end{aligned}$$

(b) *Mit Luftwiderstand:*
$K(t, x(t), x'(t)) = -\sigma x'(t) - mg$ mit $\sigma > 0$.

$$\begin{aligned} \Longrightarrow \quad & x''(t) = -\frac{\sigma}{m} x'(t) - g \\ \Longrightarrow \quad & v'(t) = -\frac{\sigma}{m} v(t) - g \\ \Longrightarrow \quad & -\frac{\sigma}{m} = \frac{v'(t)}{v(t) + \frac{mg}{\sigma}} = \log(v(t) + \tfrac{mg}{\sigma})' \\ \Longrightarrow \quad & \log(v(t) + \tfrac{mg}{\sigma}) - \log(v(0) + \tfrac{mg}{\sigma}) = -\frac{\sigma}{m} t \\ \Longrightarrow \quad & v(t) = -\tfrac{mg}{\sigma} + (v(0) + \tfrac{mg}{\sigma}) e^{-\frac{\sigma}{m} t} \\ \Longrightarrow \quad & x(t) = x(0) - \tfrac{mg}{\sigma} t + (v(0) + \tfrac{mg}{\sigma}) \left(e^{-\frac{\sigma}{m} t} - 1 \right) \end{aligned}$$

(c) *Aus Weltraum:*
$K(t, x(t), x'(t)) = -mg \frac{R^2}{x^2(t)}$ mit $R =$ Erdradius, $x =$ Entfernung von Erdmittelpunkt

$$\Longrightarrow x''(t) = -g \frac{R^2}{x^2(t)}$$

Mit dem Ansatz $x(t) = at^b$ erhalten wir $x'(t) = abt^{b-1}$ und $x''(t) = ab(b-1)t^{b-2}$. Einsetzen ergibt

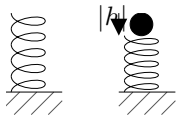
$$ab(b-1)t^{b-2} = -gR^2 \frac{1}{a^2 t^{2b}} = -\frac{gR^2}{a^2} t^{-2b}.$$

$$\Longrightarrow \begin{cases} \bullet \quad b - 2 = -2b \implies b = \frac{2}{3}, \\ \bullet \quad ab(b-1) = -\frac{gR^2}{a^2} \implies a = \left(\frac{9gR^2}{2} \right)^{1/3}. \end{cases}$$

$\Rightarrow x(t) = \left(\frac{9gR^2 t^2}{2} \right)^{\frac{1}{3}}$ ist eine Lösung (nicht einzige, aber für andere gibt es keine explizite Formel)

3. Harmonischer Oszillator. (Explizite Differentialgleichung zweiter Ordnung mit Anfangsbedingung)

Masse $m > 0$
Erdbeschleunigung g
Feder mit Konstante $c > 0$
Federauslenkung h



Newtonsches Kraftgesetz:

$$mh''(t) = -mg - ch(t)$$

d.h. explizite Differentialgleichung zweiter Ordnung; Anfangsbedingungen $h(0)$, $h'(0)$

In Ruhe, also für $h(t) = H =$ konst., gilt

$$0 = -m\,g - c\,H \Longleftrightarrow H = -\frac{m\,g}{c}.$$

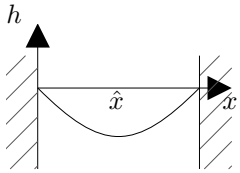
Vermutung: I.A. schwingt Masse um Ruhelage H. $\Longrightarrow$ substituiere $u(t) = h(t) - H$

$$\begin{aligned}
&\Longrightarrow \quad u''(t) = h''(t) = -g - \frac{c\,h(t)}{m} = -g - \frac{c\,(u(t)+H)}{m} = -\frac{c}{m}u(t) \\
&\Longrightarrow \quad \left(\frac{u'}{u}\right)'(t) = \frac{u''(t)}{u(t)} - \frac{u'(t)^2}{u(t)^2} = -\tfrac{c}{m} - \left(\frac{u'}{u}\right)^2(t) \\
&\overset{w=\frac{u'}{u}\sqrt{\frac{m}{c}}}{\Longrightarrow} \sqrt{\tfrac{c}{m}} = \frac{w'(t)}{-1-w(t)^2} = (\operatorname{arccot} w(t))' \\
&\Longrightarrow \quad \operatorname{arccot} w(t) = \sqrt{\tfrac{c}{m}}(t - c_1) \\
&\Longrightarrow \quad \cot\sqrt{\tfrac{c}{m}}(t-c_1) = w(t) = (\log u(t))' \\
&\Longrightarrow \quad \log\left(\sin\sqrt{\tfrac{c}{m}}(t-c_1)\right) = \log u(t) - \log c_2 \\
&\Longrightarrow \quad u(t) = c_2 \sin\sqrt{\tfrac{c}{m}}(t-c_1)
\end{aligned}$$

Langzeitverhalten: Oszillation mit Frequenz $\sqrt{\frac{c}{m}}$

4. Katenoide. (Explizite Differentialgleichung zweiter Ordnung mit Randbedingung)

Kabel aufgehangen in $(0,0)$ und $(1,H)$
Masse pro Länge ρ
Kabelhöhe $h(x)$
Erdbeschleunigung g
innere Zugkraft $T(x)$
tiefster Punkt $\hat{x}$



Einheitstangentialvektor in x: $v(x) = [1 \quad h'(x)]/\sqrt{1+h'(x)^2}$
Masse zwischen $\hat{x}$ und $x > \hat{x}$: $m = \rho\int_{\hat{x}}^{x}\sqrt{1+h'(s)^2}\,\mathrm{d}s$
Kräftegleichgewicht für Kabelstück zwischen $\hat{x}$ und x: $0 = T(x)v(x) + T(\hat{x})[-1 \quad 0] + mg[0 \quad -1]$

$$\begin{aligned}
&\Longrightarrow \quad \begin{cases} T(\hat{x}) = T(x)v_1(x) \\ mg = T(x)v_2(x) = T(\hat{x})v_2(x)/v_1(x) = T(\hat{x})h'(x) \end{cases} \\
&\Longrightarrow \quad \rho g\int_{\hat{x}}^{x}\sqrt{1+h'(x)^2}\,\mathrm{d}x = T(\hat{x})h'(x) \\
&\Longrightarrow \quad \rho g\sqrt{1+h'(x)^2} = T(\hat{x})h''(x)
\end{aligned}$$

d.h. explizite Differentialgleichung zweiter Ordnung; Randbedingungen $h(0) = 0$, $h(1) = 1$

$$\begin{aligned}
&\overset{u=h'}{\Longrightarrow} \quad \frac{\rho g}{T(\hat{x})} = \frac{u'(x)}{\sqrt{1+u(x)^2}} = (\operatorname{arsinh} u(x))' \\
&\overset{u(\hat{x})=0}{\Longrightarrow} \quad u(x) = \sinh\left(\frac{\rho g}{T(\hat{x})}(x-\hat{x})\right) \\
&\Longrightarrow \quad h(x) = \frac{T(\hat{x})}{\rho g}\cosh\left(\frac{\rho g}{T(\hat{x})}(x-\hat{x})\right) + c
\end{aligned}$$

c und $\hat{x}$ müssen nun bestimmt werden, um die Randbedingungen zu erfüllen
(und $T(\hat{x})$ muss aus dem Kräftegleichgewicht für das gesamte Kabel bestimmt werden)

3 Anfangswertprobleme

Definition 4 (Anfangswertproblem)**.** *Ein* Anfangswertproblem *(AWP) ist eine $\mathbb{R}^m$-wertige gewöhnliche Differentialgleichung n-ter Ordnung für $y \in C^n(I)^m$ auf $I = [a,b]$ zusammen mit* Anfangsbedingungen

$$y(a) = \bar{y}_0,\ y'(a) = \bar{y}_1,\ \ldots,\ y^{(n-1)}(a) = \bar{y}_{n-1}$$

für gegebene $\bar{y}_0, \ldots, \bar{y}_{n-1} \in \mathbb{R}^m$. $y \in C^n(I)^m$ heißt Lösung *des Anfangswertproblems, wenn es eine Lösung der Differentialgleichung ist und die Anfangsbedingungen erfüllt.*

Bemerkung 5 (Anfangswert im Innern von I)**.** *Statt in a kann die Anfangsbedingung auch an einem Punkt $c \in (a,b)$ gestellt werden, die Differentialgleichung wird also in einer Umgebung von c gelöst. Auch wenn c nicht am Intervallanfang liegt, ist das Problem grundsätzlich dasselbe (auf $[c,b]$ ist es ein normales Anfangswertproblem, auf $[a,c]$ wird es durch die Variablentransformation $x \rightsquigarrow -x$ zu einem Anfangswertproblem).*

3.1 Analytische Lösungsmethoden für skalare Differentialgleichungen

Einige Anfangswertprobleme sind explizit lösbar; wir behandeln einige typische Beispiele.

1. Elementare Differentialgleichung. Die Lösung von
$$y'(x) = f(x), \quad y(a) = y_0$$
ist offensichtlich gegeben durch
$$y(x) = y_0 + \int_a^x f(z)\,\mathrm{d}z.$$
Analog erhält man die Lösung elementarer Differentialgleichungen n-ter Ordnung mit Anfangsbedingungen,
$$y^{(n)}(x) = f(x), \quad y(a) = y_0, \ldots, y^{(n-1)}(a) = y_{n-1},$$
durch n-fache Integration,
$$y(x) = y_0 + \int_a^x y_1 + \int_a^{z_1} \cdots y_{n-1} + \int_a^{z_{n-1}} f(z_n)\,\mathrm{d}z_n \cdots \mathrm{d}z_2\,\mathrm{d}z_1.$$

2. Explizite autonome Differentialgleichung erster Ordnung. Betrachte das Anfangswertproblem
$$y'(x) = f(y(x)), \quad y(a) = y_0.$$
Theorem 6 (Autonome Differentialgleichung)**.** *Sei f stetig und ungleich 0 auf einer Umgebung von y_0, dann ist $y(x) = F^{-1}(x-a)$ eine Lösung des Anfangswertproblems auf einer Umgebung von a mit*
$$F(y) = \int_{y_0}^{y} \frac{1}{f(z)}\,\mathrm{d}z \qquad \text{auf einer Umgebung von } y_0.$$
Beweis. f stetig und ungleich 0 auf Intervall (c,d) um y_0
$\Longrightarrow 1/f$ stetig und größer oder kleiner 0 auf (c,d)
$\Longrightarrow F$ wohldefiniert, stetig diff.-bar & streng monoton auf (c,d) mit $F'(y) = 1/f(y)$, $F(y_0) = 0$
$\Longrightarrow$ nach Umkehrsatz ist F lokal invertierbar mit $F^{-1}(0) = y_0$, $(F^{-1})'(x) = \frac{1}{F'(F^{-1}(x))} = f(F^{-1}(x))$
$\Longrightarrow y$ ist Lösung □

Beispiel 7 ($y'(x) = y(x)^2$, $y(a) = y_0$)**.** *Setze $f(y) = y^2$ und*
$$F(y) = \int_{y_0}^{y} \frac{1}{f(z)}\,\mathrm{d}z = \int_{y_0}^{y} \frac{1}{z^2}\,\mathrm{d}z = \frac{1}{y_0} - \frac{1}{y}.$$
Die Umkehrfunktion F^{-1} ergibt sich durch Lösen von $F(z) = x$ nach z als $F^{-1}(x) = z = 1/(1/y_0 - x)$. Folglich
$$y(x) = F^{-1}(x-a) = \frac{1}{\frac{1}{y_0} + a - x}.$$
Dies löst tatsächlich das Anfangswertproblem für $x \neq \frac{1}{y_0} + a$.

Beispiel 8 ($y'(x) = 1 + y(x)^2$, $y(a) = y_0$)**.** *Setze*
$$F(y) = \int_{y_0}^{y} \frac{1}{1+z^2}\,\mathrm{d}z = \arctan y - \arctan y_0.$$
Somit
$$y(x) = F^{-1}(x-a) = \tan(x - a + \arctan y_0).$$
Dies löst tatsächlich das Anfangswertproblem für $x \neq a - \arctan y_0 + \frac{\pi}{2} + \pi\mathbb{Z}$.

Bemerkung 9 (Gebiet der Lösung)**.** *Die berechneten Lösungen der Beispiele haben Pole, sind also nicht auf ganz $\mathbb{R}$ definiert. Der Satz garantiert nur eine Lösbarkeit in einer Umgebung von a, die Gleichung besitzt nicht notwendig globale Lösungen.*

3. Trennung der Variablen. Benennung durch Johann I Bernoulli in einem Brief an Leibniz 1694.

Theorem 10 (Trennung der Variablen)**.** *Seien $a, y_0 \in \mathbb{R}$ mit Umgebungen $U, V \subset \mathbb{R}$ sowie $h : U \to \mathbb{R}$ und $f : V \to \mathbb{R}$ stetig mit $0 \notin f(V)$. Eine Lösung des Anfangswertproblems*

$$y'(x) = f(y(x))h(x), \qquad y(a) = y_0$$

in einer Umgebung von a ist gegeben durch $F^{-1}(H(x))$ mit den Stammfunktionen

$$F(y) = \int_{y_0}^{y} \frac{1}{f(z)}\,\mathrm{d}z \quad \text{auf einer Umgebung von } y_0, \qquad H(x) = \int_a^x h(s)\,\mathrm{d}s.$$

Beweis. Hausaufgabe. □

Formal kann man die Trennung der Variablen wie folgt beschreiben:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(y)h(x) \quad \Rightarrow \quad \frac{\mathrm{d}y}{f(y)} = h(x)\mathrm{d}x \quad \Rightarrow \quad \int_{y_0}^{y} \frac{1}{f(y)}\mathrm{d}y = \int_a^x h(x)\mathrm{d}x$$

Bemerkung 11 (Trennung der Variablen für autonome Differentialgleichungen)**.** *Autonome und elementare Differentialgleichungen sind Spezialfälle für die Trennung der Variablen.*

4. Exakte Differentialgleichung.

Theorem 12 (Exakte Differentialgleichung)**.** *Sei $F : [a, b] \times \mathbb{R} \to \mathbb{R}$ zweimal stetig differenzierbar und*

$$g(x, y) = \tfrac{\partial F}{\partial x}(x, y), \quad h(x, y) = \tfrac{\partial F}{\partial y}(x, y).$$

Dann heißt die Differentialgleichung des Anfangswertproblems

$$g(x, y(x)) + h(x, y(x))y'(x) = 0, \qquad y(a) = y_0$$

exakt*. Sei weiter $y : [a, b] \to \mathbb{R}$ eine Funktion mit $F(x, y(x)) = F(a, y_0)$ und $h(x, y(x)) \neq 0$ für alle $x \in [a, b]$, dann löst y das Anfangswertproblem.*

Beweis. Hausaufgabe. □

Formal kann man die Lösung einer exakten Differentialgleichung wie folgt beschreiben:

$$0 = g(x, y) + h(x, y)\frac{\mathrm{d}y}{\mathrm{d}x} \quad \Rightarrow \quad 0 = h(x, y)\mathrm{d}y + g(x, y)\mathrm{d}x = \mathrm{d}F(x, y) \quad \Rightarrow \quad F(x, y) = \text{konst.}$$

Ob eine Differentialgleichung exakt ist, kann man durch Prüfen des Satzes von Schwarz ermitteln, also durch prüfen von

$$\tfrac{\partial h}{\partial x}(x, y) = \tfrac{\partial g}{\partial y}(x, y).$$

Ist dies der Fall, so erhält man ein F für feste $x_0, y_0 \in \mathbb{R}$ durch

$$F(x, y) = \int_{x_0}^{x} g(t, y)\,\mathrm{d}t + \int_{y_0}^{y} h(x_0, s)\,\mathrm{d}s.$$

Durch Erweitern einer nicht-exakten Differentialgleichung $y'(x) = -\frac{g(x,y(x))}{h(x,y(x))}$ mit einem sogenannten *integrierenden Faktor* kann sie manchmal in eine exakte umgewandelt werden.

Theorem 13 (Integrierender Faktor)**.** *Sei $M(x,y)$ eine Funktion mit*

$$M(x,y)\tfrac{\partial}{\partial y}g(x,y)+g(x,y)\tfrac{\partial}{\partial y}M(x,y)=M(x,y)\tfrac{\partial}{\partial x}h(x,y)+h(x,y)\tfrac{\partial}{\partial x}M(x,y).$$

Dann ist die Differentialgleichung

$$y'(x)=-\frac{M(x,y(x))g(x,y(x))}{M(x,y(x))h(x,y(x))}=-\frac{g(x,y(x))}{h(x,y(x))}$$

exakt. M heißt integrierender Faktor*.*

Beweis. Hausaufgabe. □

5. Lineare Differentialgleichung erster Ordnung.

Definition 14 (Lineare Differentialgleichung)**.** *Eine explizite Differentialgleichung der Form*

$$y^{(n)}(x)=\beta(x)+\alpha_0(x)y(x)+\ldots+\alpha_{n-1}(x)y^{(n-1)}(x)$$

heißt linear*. Für $\beta=0$ heißt sie* homogen*, sonst* inhomogen*.*

Betrachte das lineare Anfangswertproblem erster Ordnung

$$y'(x)=\beta(x)+\alpha(x)y(x),\qquad y(a)=y_0.$$

Im homogenen Fall ($\beta=0$) ist eine Lösung gegeben durch

$$y(x)=y_0Y(x)\qquad \text{für } Y(x)=\exp\left(\int_a^x\alpha(s)\,\mathrm{d}s\right).$$

Im inhomogenen Fall machen wir den Ansatz (sog. *Variation der Konstanten*)

$$y(x)=C(x)Y(x).$$

Damit y Lösung des Anfangswertproblems ist, muss gelten $C(a)=y_0$ sowie

$$C(x)Y'(x)+C'(x)Y(x)=y'(x)=\alpha(x)y(x)+\beta(x)=\alpha(x)C(x)Y(x)+\beta(x)\quad\Leftrightarrow\quad C'(x)Y(x)=\beta(x)$$

und somit

$$C(x)=\int_a^x\frac{\beta(s)}{Y(s)}\,\mathrm{d}s+y_0.$$

Theorem 15 (Variation der Konstanten)**.** *Das Anfangswertproblem*

$$y'(x)=\beta(x)+\alpha(x)y(x),\qquad y(a)=y_0$$

wird gelöst durch

$$y(x)=\left(\int_a^x\frac{\beta(s)}{Y(s)}\,\mathrm{d}s+y_0\right)Y(x)\qquad \textit{für } Y(x)=\exp\left(\int_a^x\alpha(s)\,\mathrm{d}s\right).$$

Beweis. Nachrechnen. □

Beispiel 16 ($y'(x)=y(x)+1$, $y(0)=1$)**.** *Es ist* $Y(x)=\exp(\int_0^x 1\,\mathrm{d}s)=e^x$*, somit* $C(x)=\int_0^x\frac{1}{e^s}\,\mathrm{d}s+1=-e^{-x}+2$

$$\Rightarrow\quad y(x)=\left(-e^{-x}+2\right)e^x=2e^x-1.$$

6. Bernoullische Differentialgleichung. Die Bernoullische Differentialgleichung ist

$$y'(x)=p(x)y(x)+r(x)y(x)^n,\qquad n\notin\{0,1\}.$$

Durch die Transformation $z(x)=y(x)^{1-n}$ erhält man die lineare Differentialgleichung

$$z'(x)=(1-n)y(x)^{-n}y'(x)=(1-n)p(x)z(x)+(1-n)r(x).$$

Beispiel 17 ($y'(x) = -y(x)/x + x^2 y(x)^2$, $x > 0$)**.** *Wähle* $z(x) = y(x)^{1-2} = 1/y(x)$

$$\begin{aligned}
&\Longrightarrow \quad z'(x) = z(x)/x - x^2 \\
&\Longrightarrow \quad z(x) = \left(\int \frac{-x^2}{Y(x)} \, \mathrm{d}x \right) Y(x) \qquad \text{für } Y(x) = \exp\left(\int \frac{1}{x} \, \mathrm{d}x \right) = \exp(\log x + C) = cx \\
&\Longrightarrow \quad z(x) = \left(\int \frac{-x^2}{cx} \, \mathrm{d}x \right) cx = \left(\frac{-x^2}{2} + K \right) x \\
&\Longrightarrow \quad y(x) = \frac{1}{z(x)} = \frac{2}{-x^3 + kx},
\end{aligned}$$

wobei die Konstanten $K, k \in \mathbb{R}$ *aus den Anfangsbedingungen bestimmt werden können.*

7. Riccatische Differentialgleichung. Die Riccatische Differentialgleichung ist

$$y'(x) = p(x)y(x) + r(x)y(x)^2 + q(x).$$

Sie besitzt i.A. keine explizite Lösungsformel, doch kennt man eine spezielle Lösung $u(x)$, so liefert der Ansatz $y(x) = u(x) + v(x)$

$$\begin{aligned}
u'(x) + v'(x) &= p(x)(u(x) + v(x)) + r(x)(u(x) + v(x))^2 + q(x) \\
&= [p(x)u(x) + r(x)u(x)^2 + q(x)] + [p(x) + 2r(x)u(x)]v(x) + r(x)v(x)^2
\end{aligned}$$

und somit für v die Bernoullische Differentialgleichung

$$v'(x) = [p(x) + 2r(x)u(x)]v(x) + r(x)v(x)^2.$$

Beispiel 18 ($y'(x) = y(x) - 3y(x)^2 + 10$)**.** $u(x) = 2$ *ist eine spezielle Lösung, somit* $y(x) = u(x) + v(x)$ *mit*

$$v'(x) = [1 - 6 \cdot 2]v(x) - 3v(x)^2 = -11v(x) - 3v(x)^2.$$

8. Potenzreihenanzatz. Sind die Koeffizienten $\beta, \alpha_0, \ldots, \alpha_{n-1}$ der linearen Differentialgleichung

$$y^{(n)}(x) = \beta(x) + \alpha_0(x)y(x) + \ldots + \alpha_{n-1}(x)y^{(n-1)}(x)$$

im Anfangspunkt $a \in \mathbb{R}$ in eine Potenzreihe mit positivem Konvergenzradius entwickelbar (also *lokal analytisch*), kann man für eine Lösung den Potenzreihenanzatz wählen,

$$y(x) = \sum_{k=0}^{\infty} c_k (x - a)^k.$$

Einsetzen (unter gliedweisem Differenzieren) ergibt Gleichungen für die Koeffizienten c_k, die konsekutiv gelöst werden können. Am Schluss muss noch überprüft werden, ob der Konvergenzradius der Potenzreihe positiv ist (dann war auch gliedweises Differenzieren erlaubt).

Beispiel 19 ($(x-1)y'(x) + y(x) = 0$, $y(0) = 1$)**.** *Wähle Ansatz*

$$\begin{aligned}
&y(x) = \sum_{k=0}^{\infty} c_k x^k \qquad \Longrightarrow \qquad y'(x) = \sum_{k=1}^{\infty} k c_k x^{k-1} \\
&\Longrightarrow \qquad 0 = \sum_{k=1}^{\infty} k c_k (x^k - x^{k-1}) + \sum_{k=0}^{\infty} c_k x^k \\
&\Longrightarrow \qquad 0 = \sum_{k=0}^{\infty} (k+1)(c_k - c_{k+1}) x^k \\
&\Longrightarrow \qquad 0 = c_1 - c_0 = c_2 - c_1 = \ldots \\
&\Longrightarrow \qquad 1 = c_0 = c_1 = c_2 = \ldots \\
&\Longrightarrow \qquad y(x) = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x},
\end{aligned}$$

wobei Konvergenz für $|x| < 1$ *gilt.*

3.2 Existenz, Eindeutigkeit und Stabilität

I.A. reicht die Betrachtung von (vektorwertigen) Differentialgleichungen erster Ordnung.

Bemerkung 20 (Reduktion von gewöhnlichen Differentialgleichungen).

- *Eine skalare explizite Differentialgleichung n-ter Ordnung in $y \in C^n([a,b])$,*

$$y^{(n)}(x) = f(x, y(x), \dots, y^{(n-1)}(x)),$$

kann in eine äquivalente $\mathbb{R}^n$-wertige explizite Differentialgleichung erster Ordnung in $u \in C^1([a,b])^n$ umgewandelt werden mittels $u = (u_0, \dots, u_{n-1})^T = (y^{(0)}, \dots, y^{(n-1)})^T$,

$$u'(x) = \begin{pmatrix} u_1(x) \\ \vdots \\ u_{n-1}(x) \\ f(x, u_0(x), \dots, u_{n-1}(x)) \end{pmatrix}.$$

- *Eine skalare explizite Differentialgleichung $y'(x) = f(x, y(x))$ kann in eine äquivalente $\mathbb{R}^2$-wertige autonome Differentialgleichung umgewandelt werden mittels $u(x) = \binom{u_1(x)}{u_2(x)} = \binom{x}{y(x)}$,*

$$u'(x) = \begin{pmatrix} 1 \\ f(u_1(x), u_2(x)) \end{pmatrix}.$$

- *Funktioniert analog für vektorwertige und für implizite Differentialgleichungen (wie?).*
- *Analog können zugehörige Anfangswertprobleme reduziert werden (wie?).*

Ein Anfangswertproblem erster Ordnung kann man auf eine äquivalente Integralgleichung zurückführen.

Theorem 21 (Integralgleichung). *Sei $m \in \mathbb{N}$, $I = [a,b] \subset \mathbb{R}$, $f : I \times \mathbb{R}^m \to \mathbb{R}^m$ stetig, $y \in C^0(I)^m$. y ist genau dann in $C^1(I)^m$ und Lösung des Anfangswertproblems*

$$y'(x) = f(x, y(x)), \qquad y(a) = y_0, \tag{1}$$

wenn es folgende Integralgleichung löst,

$$y(x) = y_0 + \int_a^x f(s, y(s))\,\mathrm{d}s \qquad \text{für alle } x \in I.$$

Beweis. Hausaufgabe. □

Diese Integralgleichung ist eine Fixpunkt-Gleichung, sodass Fixpunkt-Theorie angewandt werden kann.

Definition 22 (Lipschitz-stetig, Kontraktion, Fixpunkt). *Seien X, Y normierte Vektorräume, $D \subset X$. $T : D \to Y$ heißt* Lipschitz-stetig *mit* Lipschitz-Konstante $L > 0$, *wenn $\|T(x) - T(z)\| \le L\|x - z\|$ für alle $x, z \in D$ gilt. T heißt* Kontraktion, *wenn $Y = X$, $T : D \to D$, und $L < 1$. Ein $\bar{x} \in D$ mit $T(\bar{x}) = \bar{x}$ heißt* Fixpunkt *von T. Die iterative Vorschrift $x_k = T(x_{k-1})$ für $k \in \mathbb{N}$ und gegebenes $x_0 \in D$ heißt* Fixpunkt-Iteration.

Die Fixpunkt-Iteration für obige Integralgleichung nennt man auch Picard-Iteration.

Definition 23 (Picard-Iteration). *Setze $y_0(x) = y_0$, dann lautet die* Picard-Iteration *zum Anfangswertproblem* (1)

$$y_k(x) = y_0 + \int_a^x f(s, y_{k-1}(s))\,\mathrm{d}s \qquad \text{für alle } x \in I \text{ und } k \in \mathbb{N}.$$

Beispiel 24 (Exponentielles Populationswachstum). *Die Picard-Iteration für*

$$y'(t) = \lambda y(t), \quad y(0) = y_0 > 0$$

lautet

$$y_1(t) = y_0 + \int_0^t \lambda y_0 \,\mathrm{d}s = y_0(1+\lambda t),$$
$$y_2(t) = y_0 + \int_0^t \lambda y_0(1+\lambda s) \,\mathrm{d}s = y_0(1+\lambda t + \tfrac{(\lambda t)^2}{2}),$$
$$y_k(t) = y_0 \sum_{j=0}^{k} \frac{(\lambda t)^j}{j!},$$

sie konvergiert also (gleichmäßig auf jedem endlichen Intervall) gegen die Lösung $y(t) = y_0 \exp(\lambda t)$.

Wann existiert nun eine Lösung des Anfangswertproblems bzw. der Integralgleichung? Konvergiert die Picard-Iteration dagegen? Beides ist eine direkte Anwendung des Banachschen Fixpunkt-Satzes.

Theorem 25 (Banachscher Fixpunkt-Satz)**.** *Sei X ein Banachraum, $D \subset X$ abgeschlossen, $T : D \to D$ eine Kontraktion mit Konstante L, dann gilt:*

1. *T besitzt genau einen Fixpunkt $\bar{x} \in D$.*
2. *Die Fixpunkt-Iteration $x_k = T(x_{k-1})$ konvergiert gegen $\bar{x}$.*
3. *Der Fehler erfüllt* $\|x_k - \bar{x}\| \le \underbrace{\frac{L}{1-L}\|x_k - x_{k-1}\|}_{\text{„a posteriori-,}} \le \underbrace{\frac{L^k}{1-L}\|x_1 - x_0\|}_{\text{a priori-Abschätzung“}}$.

Beweis. 0)
$$\|x_l - x_{l-1}\| = \|T(x_{l-1}) - T(x_{l-2})\| \le L\|x_{l-1} - x_{l-2}\|$$
$$\le L^2\|x_{l-2} - x_{l-3}\| \le \ldots \le L^{l-k-1}\|x_{k+1} - x_k\| \tag{2}$$
$$\|x_l - x_k\| \le \|x_l - x_{l-1}\| + \|x_{l-1} - x_{l-2}\| + \ldots + \|x_{k+1} - x_k\|$$
$$\le L^{l-k-1}\|x_{k+1} - x_k\| + L^{l-k-2}\|x_{k+1} - x_k\| + \ldots + L^0\|x_{k+1} - x_k\|$$
$$= \underbrace{\sum_{j=0}^{l-k-1} L^j}_{\le \sum_{j=0}^{\infty} L^j = \frac{1}{1-L}} \|x_{k+1} - x_k\| \le \frac{1}{1-L}\|x_{k+1} - x_k\| \overset{(2)}{\le} \frac{L}{1-L}\|x_k - x_{k-1}\| \tag{3}$$

1) & 2)
- x_k ist eine Cauchy-Folge: Für $\epsilon > 0$ wähle $N \in \mathbb{N}$ mit $L^N < \epsilon(1-L)/\|x_1 - x_0\|$, dann ist

$$\|x_l - x_k\| \overset{(3)}{\le} \frac{L}{1-L}\|x_k - x_{k-1}\| \overset{(2)}{\le} \frac{L^k}{1-L}\|x_1 - x_0\| \le \epsilon \qquad \forall l \ge k \ge N$$

- somit $x_k \to \bar{x}$ für ein $\bar{x} \in D$, und $\bar{x}$ ist Fixpunkt, da

$$\|T(\bar{x}) - \bar{x}\| = \lim_{l\to\infty} \|T(\bar{x}) - x_{l+1}\| = \lim_{l\to\infty} \|T(\bar{x}) - T(x_l)\| \le \lim_{l\to\infty} L\|\bar{x} - x_l\| = 0$$

- $\bar{x}$ ist eindeutig, denn $y = T(y) \Rightarrow \|\bar{x} - y\| = \|T(\bar{x}) - T(y)\| \le L\|\bar{x} - y\| \Rightarrow \|\bar{x} - y\| = 0$

3) $\|\bar{x} - x_k\| = \lim_{l\to\infty} \|x_l - x_k\| \overset{(3)}{\le} \frac{L}{1-L}\|x_k - x_{k-1}\| \overset{(2)}{\le} \frac{L^k}{1-L}\|x_1 - x_0\|$

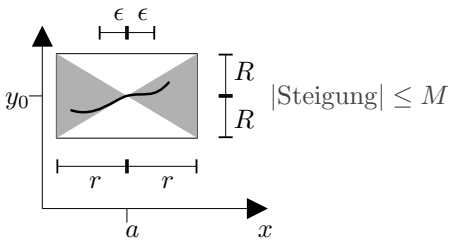
□

Die Anwendung auf unsere Integralgleichung liefert Existenz und Eindeutigkeit unter Bedingungen an f.

Theorem 26 (Satz von Picard–Lindelöf)**.** *Sei $I = [a-r, a+r]$, $B = \{y \in \mathbb{R}^m \,|\, \|y - y_0\| \le R\}$, $f : I \times B \to \mathbb{R}^m$ stetig und Lipschitz-stetig im zweiten Argument, d.h.*

$$\exists L > 0 : \ \|f(x,y) - f(x,z)\| \le L\|y - z\| \ \forall x \in I, \ y, z \in B.$$

Sei $M = \max_{x\in I, y\in B} \|f(x,y)\|$ und $r \le R/M$, dann existiert eine eindeutige Lösung $y : I \to B$ von (1).



Beweis. Zunächst zeige Existenz und Eindeutigkeit auf $\tilde{I} = [a-\epsilon, a+\epsilon]$ für $\epsilon = \min\{r, \frac{1}{2L}\}$.

- Sei $C(\tilde{I}; B) = \{y \in C^0(\tilde{I})^m \mid y(\tilde{I}) \subset B\}$ und T die Picard-Iterationsabbildung, dann ist $T : C(\tilde{I}; B) \to C(\tilde{I}; B)$, denn $T(y)$ ist stetig mit $\|T(y)(x) - y_0\| \le \int_a^x \|f(s, y(s)\| \,\mathrm{d}s \le M\epsilon \le R$.
- T ist eine Kontraktion: Mit der Supremumnorm $\|g\|_\infty = \sup_{t\in\tilde{I}} \|g(t)\|$ gilt

$$\|T(y)(x) - T(z)(x)\| = \left\| \int_a^x f(s, y(s)) - f(s, z(s)) \,\mathrm{d}s \right\| \le \left| \int_a^x \|f(s, y(s)) - f(s, z(s))\| \,\mathrm{d}s \right|$$
$$\le |x-a| L \|y - z\|_\infty \le \tfrac{1}{2}\|y - z\|_\infty \quad \forall x \in \tilde{I}$$

  $\Rightarrow \|T(y) - T(z)\|_\infty \le \frac{1}{2}\|y - z\|_\infty$.
- $C(\tilde{I}; B)$ ist abgeschlossen im Banachraum $(C^0(\tilde{I})^m, \|\cdot\|_\infty)$
  $\overset{\text{Banachscher Fixpunkt-Satz}}{\Longrightarrow}$ $\exists!$ Fixpunkt $\bar{y} \in C(\tilde{I}; B)$
  $\overset{\text{Theorem 21}}{\Longrightarrow}$ $\bar{y}$ löst (1).

Wiederhole das Argument nun für $y'(x) = f(x, y(x))$, $y(a+\epsilon) = \bar{y}(a+\epsilon)$, um eine eindeutige Lösung auf $[a, a+2\epsilon]$ zu erhalten, dann analog für Anfangswerte bei $a+2\epsilon$, $a+3\epsilon$ usw., bis die eindeutige Lösung auf ganz I definiert ist. □

Bemerkung 27 (Voraussetzungen von Picard–Lindelöf).
- *Lipschitz-Bedingung ist nötig für Eindeutigkeit (Hausaufgabe). Existenz einer Lösung erhielte man auch unter schwächeren Bedingungen, z.B. f stetig (Satz von Peano) oder noch schwächer (Satz von Caratheodory).*
- *$r \le R/M$ ist nötig für Existenz der Lösung auf ganz I (Hausaufgabe).*

Zur Wohlgestelltheit eines mathematischen Problems gehört nach Hadamard nicht nur Existenz und Eindeutigkeit einer Lösung, sondern auch Stabilität, d.h. stetige Abhängigkeit von den Eingabedaten (in unserem Fall rechte Seite und Anfangsbedingungen von (1)), was wir als nächstes behandeln.

Lemma 28 (Gronwall). *Sei $I = [a, b]$, $\alpha, \beta, y : I \to \mathbb{R}$ stetig.*

1. $y'(x) \le \alpha(x) + \beta(x)y(x)\ \forall x \in I \Longrightarrow y(x) \le y(a)\exp\left(\int_a^x \beta(s)\,\mathrm{d}s\right) + \int_a^x \alpha(r)\exp\left(\int_r^x \beta(s)\,\mathrm{d}s\right)\,\mathrm{d}r\ \forall x \in I$,
2. $y(x) \le \alpha(x) + \int_a^x \beta(s)y(s)\,\mathrm{d}s\ \forall x \in I \Longrightarrow y(x) \le \alpha(x) + \int_a^x \alpha(r)\beta(r)\exp\left(\int_r^x \beta(s)\,\mathrm{d}s\right)\,\mathrm{d}r\ \forall x \in I$.

Beweis. Hausaufgabe. □

Theorem 29 (Stabilität). (1) *erfülle die Voraussetzungen des Satzes von Picard–Lindelöf, $\tilde{f}, \tilde{y}_0$ seien Näherungen mit $\|\tilde{y}_0 - y_0\| \le \epsilon$ und $\|\tilde{f} - f\|_\infty \le \delta$. Für eine Lösung $\tilde{y} : I \to \mathbb{R}^m$ des Anfangswertproblems*

$$\tilde{y}'(x) = \tilde{f}(x, \tilde{y}(x)), \qquad \tilde{y}(a) = \tilde{y}_0$$

gilt $\|\tilde{y}(x) - y(x)\| \le (\epsilon + \delta|x-a|)\exp(L|x-a|)$ für alle $x \in I$.

Beweis. Hausaufgabe. □

Man kann sogar genauer untersuchen, wie sich die Lösung bei Perturbationen von y_0 und f ändert, also die sogenannte *Sensitivität* ausrechnen. Damit kann man vorhersagen, welche Konsequenzen die Änderung von Modellparametern hat (z.B. „Wie wird sich der Verlauf der durch gewöhnliche Differentialgleichungen beschriebenen Klimaerwärmung in den nächsten 20 Jahren ändern, wenn wir die CO_2-Emission im nächsten Jahr um 20 % senken?“), und letztlich diese Modellparameter optimieren.

Theorem 30 (Sensitivität). *Seien $y_0 = y_0^p$ und $f = f^p$ stetig differenzierbar von einem Parameter $p \in \mathbb{R}$ abhängig sowie $f^p(x, y)$ stetig in (x, y, p) und stetig differenzierbar in y. Sei $y^p \in C^1(I)^m$ die von p abhängige Lösung des Anfangswertproblems*

$$(y^p)'(x) = f^p(x, y^p(x)), \qquad y^p(0) = y_0^p.$$

Dann ist $z^p := \frac{\partial y^p}{\partial p}$ die Lösung des linearen Anfangswertproblems

$$(z^p)'(x) = \tfrac{\partial f^p}{\partial p}(x, y^p(x)) + \tfrac{\partial f^p}{\partial y}(x, y^p(x))z^p(x), \qquad z^p(0) = \tfrac{\partial y_0^p}{\partial p}.$$

Beweis. Hausaufgabe (leite die äquivalente Integralgleichung ab). □

3.3 Phasenportrait-Analyse

Eine $\mathbb{R}^m$-wertige autonome gewöhnliche Differentialgleichung

$$y'(x) = f(y(x)) \tag{4}$$

mit Lipschitz-stetigem f kann als das Vektorfeld $\mathbb{R}^m \ni y \mapsto f(y) \in \mathbb{R}^m$ (sogenanntes *Phasenportrait*) aufgefasst werden. Die Trajektorie $x \mapsto y(x)$ einer Lösung y ist immer tangential an das Vektorfeld. Im $\mathbb{R}^2$ lässt sich das Phasenportrait leicht visualisieren.

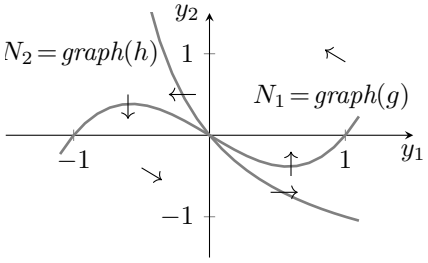
Definition 31 (Nullkline)**.** *Die y_j-*Nullkline *von* (4) *ist die Menge*

$$N_j = \{y \in \mathbb{R}^m \mid f_j(y) = 0\}.$$

Auf der y_j-Nullkline ist $y_j' = 0$. Die Nullklinen zerlegen den $\mathbb{R}^m$ in Gebiete, in denen die Lösungstrajektorien in unterschiedliche Richtungen zeigen, z.B. den $\mathbb{R}^2$ in die Gebiete

$$\{y_1' > 0, y_2' > 0\},\ \{y_1' > 0, y_2' < 0\},\ \{y_1' < 0, y_2' > 0\},\ \{y_1' < 0, y_2' < 0\}.$$

Beispiel 32 ($y_1'(x) = g(y_1(x)) - y_2(x)$, $y_2'(x) = y_2(x) - h(y_1(x))$ mit $g(s) = s^3 - s$, $h(s) = -2s/(s+1)$)**.**



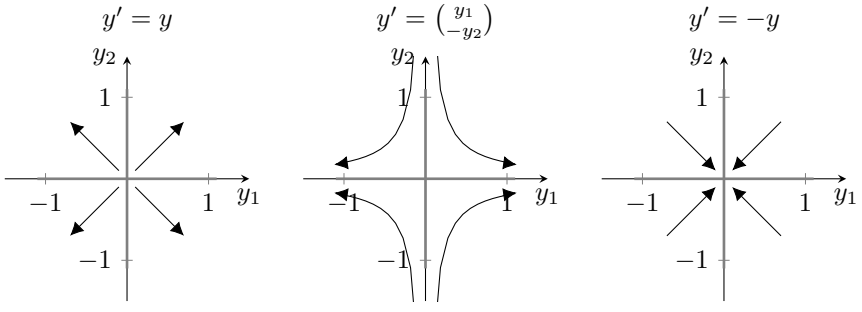
Die Schnittpunkte aller Nullklinen sind Gleichgewichtspunkte oder stationäre Punkte.

Definition 33 (Gleichgewichtspunkt, Stabilität)**.** *Gegeben sei eine explizite autonome gewöhnliche Differentialgleichung $y'(x) = f(y(x))$ in $\mathbb{R}^m$.*

1. *Ein* Gleichgewichtspunkt *ist ein $\bar{y} \in \mathbb{R}^m$ mit $f(\bar{y}) = 0$ ($y(x) = \bar{y}$ is Lösung).*
2. *$\bar{y}$ heißt* stabil*, wenn für jede Umgebung U von $\bar{y}$ eine Umgebung $V \subset U$ von $\bar{y}$ existiert, sodass jede Lösung $y(x)$ mit $y(0) \in V$ für alle $x > 0$ definiert ist und $y(x) \in U$ erfüllt.*
3. *$\bar{y}$ heißt* asymptotisch stabil*, wenn eine Umgebung V von $\bar{y}$ existiert mit $\lim_{x\to\infty} y(x) = \bar{y}$ für alle Lösungen y mit $y(0) \in V$.*
4. *$\bar{y}$ heißt* instabil*, wenn es nicht stabil ist.*

(In-)stabilität sagt etwas über das Langzeitverhalten der gewöhnlichen Differentialgleichung aus.

Beispiel 34 (Asymptotisch stabile und instabile Gleichgewichtspunkte in Phasenportraits)**.**



Asymptotische Stabilität kann man auf verschiedene Arten prüfen:

- Prüfe, ob der Gleichgewichtspunkt nach Linearisierung der Differentialgleichung asymptotisch stabil ist (machen wir später).
- Finde eine Lyapunov-Funktion.

Theorem 35 (Lyapunov-Funktion)**.** *Sei $\bar{y} \in \mathbb{R}^m$ mit $f(\bar{y}) = 0$, U eine Umgebung von $\bar{y}$ und $\mathcal{L} : U \to \mathbb{R}$ stetig differenzierbar mit $\mathcal{L}(y) > \mathcal{L}(\bar{y})\ \forall y \neq \bar{y}$. Dann gilt:*

1. *$\nabla\mathcal{L}(y) \cdot f(y) \leq 0\ \forall y \in U \implies \bar{y}$ ist stabil.*
2. *$\nabla\mathcal{L}(y) \cdot f(y) < 0\ \forall y \in U \setminus \{\bar{y}\} \implies \bar{y}$ ist asymptotisch stabil.*

Ein $\mathcal{L}$ mit diesen Eigenschaften heißt Lyapunov-Funktion.

Beweis. 1. Sei B ein Ball in U um $\bar{y}$ und $\alpha = \min_{y \in \partial B} \mathcal{L}(y)$, $V = \{y \in B \mid \mathcal{L}(y) < \alpha\}$.
Sei y Lösung der Differentialgleichung mit $y(0) \in V$, dann ist

$$\tfrac{\mathrm{d}}{\mathrm{d}x}\mathcal{L}(y(x)) = \nabla\mathcal{L}(y(x)) \cdot y'(x) = \nabla\mathcal{L}(y(x)) \cdot f(y(x)) \leq 0,$$

somit $\mathcal{L}(y(x)) < \alpha\ \forall x > 0$. Somit bleibt $y(x)$ in V.

2. Sei $R > 0$ beliebig, $B_R(\bar{y})$ offener Ball um $\bar{y}$ mit Radius R. Z.z.: $y(x) \in \overline{B_R(\bar{y})}$ für x groß genug.
Sei $\beta = \min\{\mathcal{L}(y) \mid y \in \bar{B} \setminus B_R(\bar{y})\} > \mathcal{L}(\bar{y})$
(Minimum einer stetigen Funktion auf kompaktem Gebiet existiert nach Satz von Weierstraß).
Sei $r \in (0, R)$ so dass $\mathcal{L} < \beta$ auf $B_r(\bar{y})$.
Setze $\delta = \min\{|\nabla\mathcal{L}(y) \cdot f(y)| \mid y \in \bar{B} \setminus B_r(\bar{y})\} > 0$.
Fall 1: y erreicht irgendwann $B_r(\bar{y}) \implies$ ab dann $\mathcal{L} \circ y < \beta \implies$ ab dann $y \in B_R(\bar{y})$.
Fall 2: y erreicht nie $B_r(\bar{y}) \implies \tfrac{\mathrm{d}}{\mathrm{d}x}\mathcal{L}(y(x)) = \nabla\mathcal{L}(y(x)) \cdot y'(x) < -\delta\ \forall x \implies \lim_{x\to\infty} \mathcal{L}(y(x)) = -\infty$. ↯

□

3.4 Lineare Systeme von gewöhnlichen Differentialgleichungen

Betrachte das $\mathbb{R}^m$-wertige lineare Anfangswertproblem

$$y'(x) = A(x)y(x) + B(x), \quad y(a) = y_0. \tag{5}$$

Theorem 36 (Existenz und Eindeutigkeit)**.** *Sei $I = [a, b]$, $A : I \to \mathbb{R}^{m\times m}$ und $B : I \to \mathbb{R}^m$ stetig, $y_0 \in \mathbb{R}^m$. Dann existiert genau eine Lösung von (5) auf ganz I.*

Beweis. $f(x, y) = A(x)y + B(x)$ ist stetig in x und Lipschitz-stetig in y mit Konstante $L = \max_{x\in I} \|A(x)\|$ (Operatornorm), da

$$\|f(x, y) - f(x, \tilde{y})\| = \|A(x)(y - \tilde{y})\| \leq \|A(x)\|\,\|y - \tilde{y}\| \leq L\|y - \tilde{y}\|.$$

Nutze nun Picard–Lindelöf (wie?). □

Wie im skalaren Fall kann man Lösungen von (5) auf ähnliche Weise explizit angeben. Dies erfordert etwas Vorarbeit.

Theorem 37 (Lösungsisomorphismus)**.** *Sei $I = [a, b]$, $A : I \to \mathbb{R}^{m\times m}$ stetig.*

1. *Die Lösungen der $\mathbb{R}^m$-wertigen homogenen gewöhnlichen Differentialgleichung*

$$y'(x) = A(x)y(x) \tag{6}$$

bilden einen m-dimensionalen Unterraum von $C^1(I)^m$.

2. *Die Abbildung*

$$\Phi : \mathbb{R}^m \to C^1(I)^m, \quad \Phi(y_0) = \text{eindeutige Lösung von (6) mit } y(a) = y_0$$

ist eine Vektorraumisomorphismus von $\mathbb{R}^m$ in den Lösungsraum.

Beweis. Erster folgt aus zweitem Teil.
Φ ist linear: $\Psi = \alpha\Phi(y_0) + \beta\Phi(y_1)$ erfüllt

$$\Psi'(x) = \alpha\Phi(y_0)'(x) + \beta\Phi(y_1)'(x) = \alpha A(x)\Phi(y_0)(x) + \beta A(x)\Phi(y_1)(x) = A(x)\Psi(x)$$

$\Longrightarrow \Psi$ ist Lösung für Anfangsbedingung $\Psi(a) = \alpha y_0 + \beta y_1$, d.h. $\Phi(\alpha y_0 + \beta y_1) = \Psi = \alpha\Phi(y_0) + \beta\Phi(y_1)$.
Φ ist invertierbar: Sei y eine beliebige Lösung, dann ist $y = \Phi(y(0))$. □

Korollar 38 (Fundamentalsystem)**.** *Seien $y_1, \dots, y_m$ Lösungen von (6), $x \in I$ beliebig. Folgende Aussagen sind äquvalent.*

1. *$y_1, \dots, y_m$ sind Basis des Lösungsraums.*
2. *$y_1, \dots, y_m$ sind linear unabhängig in $C^1(I)^m$.*
3. *$y_1(a), \dots, y_m(a)$ sind linear unabhängig in $\mathbb{R}^m$.*
4. *$y_1(x), \dots, y_m(x)$ sind linear unabhängig in $\mathbb{R}^m$.*

Gilt eine, so ist jede Lösung y eine eindeutige Linearkombination der $y_1, \dots, y_m$.

Definition 39 (Fundamentalmatrix, Wronski-Determinante)**.** *Ein* Fundamentalsystem *zu (6) sind m linear unabhängige Lösungen $y_1, \dots, y_m$. $Y : I \to \mathbb{R}^{m\times m}$, $Y(x) = (y_1(x), \dots, y_m(x))$ heißt* Fundamentalmatrix*. $W : I \to \mathbb{R}$, $W(x) = \det Y(x)$ heißt* Wronski-Determinante*.*

Bemerkung 40 (Erinnerung: Adjunkte/Kofaktormatrix)**.** *Aus $M \in \mathbb{R}^{m\times m}$ erhalte man $M^{ij} \in \mathbb{R}^{(m-1)\times(m-1)}$ durch Streichen der i-ten Zeile und j-ten Spalte. Die Einträge der Kofaktormatrix $\operatorname{cof} M \in \mathbb{R}^{m\times m}$ sind dann definiert durch $(\operatorname{cof} M)_{ij} = (-1)^{i+j} \det M^{ij}$, und es gilt $\operatorname{cof} M = \det M\, M^{-T}$ (Cramersche Regel). Sei $E \in \mathbb{R}^{m\times m}$ die Matrix mit einzigem Eintrag $E_{ij} = 1$ ungleich 0 und $\tilde{M} \in \mathbb{R}^{m\times m}$ entstehe aus M durch Tauschen der j-ten Spalte mit E. Die Ableitung von $\det M$ nach dem (i,j)-Eintrag ist dann (mithilfe Laplacescher Entwicklung nach der j-ten Spalte) gegeben durch*

$$\frac{\partial}{\partial M_{ij}} \det M = \lim_{t\to 0} \frac{\det(M + tE) - \det M}{t} = \lim_{t\to 0} \frac{\det M + t \det \tilde{M} - \det M}{t} = \det \tilde{M} = (\operatorname{cof} M)_{ij}.$$

Weiter sei $M : N = \sum_{ij} M_{ij} N_{ij} = \operatorname{tr}(NM^T)$ das Frobenius-Skalarprodukt zweier Matrizen.

Korollar 41 (Wronski-Determinante)**.** *$W(x) \neq 0\ \forall x \in I$. Außerdem $W'(x) = \operatorname{tr}(A(x))W(x)$.*

Beweis. $W'(x) = \sum_{i,j} \frac{\partial \det Y(x)}{\partial Y_{ij}(x)} Y'_{ij}(x) = \operatorname{cof} Y(x) : Y'(x) = \det Y(x) Y(x)^{-T} : (A(x)Y(x)) = W(x)\operatorname{tr}(A(x))$. □

Für eine Fundamentalmatrix Y und beliebige $b \in \mathbb{R}^m$ ist $Y(x)b$ eine Lösung von (6). Für das inhomogene Anfangswertproblem (5) machen wir wieder den Ansatz $y(x) = Y(x)b(x)$ der Variation der Konstanten. Damit dies eine Lösung ist, muss gelten $y_0 = Y(a)b(a)$ und

$$Y'(x)b(x) + Y(x)b'(x) = y'(x) = A(x)y(x) + B(x) = A(x)Y(x)b(x) + B(x) \quad \Leftrightarrow \quad Y(x)b'(x) = B(x),$$

somit $b(a) = Y(a)^{-1}y_0$ und $b'(x) = Y(x)^{-1}B(x)$ ($\det Y(x) \neq 0$!), also $b(x) = Y(a)^{-1}y_0 + \int_a^x Y(s)^{-1}B(s)\,\mathrm{d}s$.

Theorem 42 (Variation der Konstanten)**.** (5) *wird gelöst durch*

$$y(x) = Y(x)\left(\int_a^x Y(s)^{-1}B(s)\,\mathrm{d}s + Y(a)^{-1}y_0\right) \qquad \text{für } Y \text{ eine Fundamentalmatrix.}$$

Beweis. Nachrechnen. □

Für $A(x) = A \in \mathbb{R}^{m\times m}$ unabhängig von x kann man ein Fundamentalsystem sogar explizit angeben. Hierzu ist das Matrix-Exponential hilfreich,

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

Theorem 43 (Konstante Koeffizienten). *Sei $A \in \mathbb{R}^{m\times m}$. Betrachte $y'(x) = Ay(x)$.*

1. *Seien $\lambda \in \mathbb{C}$, $v \in \mathbb{C}^m$ Eigenwert und -vektor von A, dann sind Real- und Imaginärteil von $y(x) = e^{\lambda x}v$ Lösungen.*
2. *Die Spalten von $Y(x) = \exp(xA)$ sind Lösungen.*

Beweis. Nachrechnen. □

Korollar 44 (Fundamentalmatrix). 1. *Ist $A = P\mathrm{diag}(\lambda_1, \dots, \lambda_m)P^{-1}$ diagonalisierbar, enthalten die Real- und Imaginärteile der Spalten von $P\mathrm{diag}(e^{\lambda_1 x}, \dots, e^{\lambda_m x})$ ein Fundamentalsystem.*
2. *$Y(x) = \exp(xA)$ ist eine Fundamentalmatrix.*

Beweis. 1. Da P invertierbar ist, spannen sie für $x = 0$ den $\mathbb{R}^m$ auf.

2. Wronski-Determinante in $x = 0$ ist $\det Y(0) = \det\exp(0) = \det I = 1 \neq 0$. □

Beispiel 45 ($y'(x) = Ay(x)$ mit $A = \left(\begin{smallmatrix} 3 & 5 \\ -5 & -3 \end{smallmatrix}\right)$).
Charakteristisches Polynom $\chi_A(\lambda) = \det(A - \lambda I) = \lambda^2 + 16 \implies$ Eigenwerte $\lambda_{1/2} = \pm 4i$.
Eigenvektoren liegen im Kern von $A - \lambda I \implies$ Eigenvektoren $v_{1/2} = \binom{-5}{3\mp 4i}$.

$$\mathfrak{Re}(e^{\lambda_1 x}v_1) = \mathfrak{Re}\left((\cos(4x) + i\sin(4x))\left(\begin{pmatrix}-5\\3\end{pmatrix} - i\begin{pmatrix}0\\4\end{pmatrix}\right)\right) = \begin{pmatrix}-5\cos(4x)\\3\cos(4x) + 4\sin(4x)\end{pmatrix},$$
$$\mathfrak{Im}(e^{\lambda_1 x}v_1) = \begin{pmatrix}-5\sin(4x)\\3\sin(4x) - 4\cos(4x)\end{pmatrix}$$

sind linear unabhängig in $x = 0$ und bilden somit ein Fundamentalsystem.

Wie berechnet man jedoch $\exp(xA)$ für ein $A \in \mathbb{R}^{m\times m}$?

Bemerkung 46 (Berechnung $\exp(xA)$). *Sei die (komplexe) Jordan-Normalform von A gegeben durch*

$$J = P^{-1}AP = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_n \end{pmatrix} \quad \text{mit Blöcken } J_i = \Lambda_i + N_i, \quad \Lambda_i = \begin{pmatrix} \lambda_i & & \\ & \ddots & \\ & & \lambda_i \end{pmatrix}, N_i = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}, \lambda_i \in \mathbb{C},$$

dann gilt (siehe Übung; beachte $\Lambda_i N_i = N_i \Lambda_i \in \mathbb{R}^{m_i \times m_i}$)

$$\exp(xA) = \exp(xPJP^{-1}) = P\exp(xJ)P^{-1} = P\begin{pmatrix} \exp(xJ_1) & & \\ & \ddots & \\ & & \exp(xJ_n) \end{pmatrix}P^{-1},$$
$$\exp(xJ_i) = \exp(x\Lambda_i + xN_i) = \exp(x\Lambda_i)\exp(xN_i) = \exp(x\lambda_i)\exp(xN_i),$$
$$\exp(xN_i) = \sum_{k=0}^{m_i-1}\frac{x^k}{k!}N_i^k = \begin{pmatrix} 1 & x & \frac{x^2}{2} & \cdots & \frac{x^{m_i-1}}{(m_i-1)!} \\ & 1 & x & \cdots & \frac{x^{m_i-2}}{(m_i-2)!} \\ & & \ddots & \ddots & \vdots \\ & & & 1 & x \\ & & & & 1 \end{pmatrix}.$$

Korollar 47 (Stabilität von Gleichgewichtspunkten). *Besitzt A einen Eigenwert mit positivem Realteil, ist $\bar{y} = 0$ ein instabiler Gleichgewichtspunkt von*

$$y'(x) = Ay(x). \tag{7}$$

Haben alle Eigenwerte von A negativen Realteil, ist $\bar{y} = 0$ asymptotisch stabil.

Beweis. Hausaufgabe. □

Zu asymptotisch stabilen Gleichgewichtspunkten kann man eine Lyapunov-Funktion konstruieren.

Theorem 48 (Lyapunov-Funktion)**.** *Haben alle Eigenwerte von $A \in \mathbb{R}^{m\times m}$ negativen Realteil, dann existiert für jede symmetrische positiv definite Matrix $Q \in \mathbb{R}^{m\times m}$ eine symmetrische positiv definite Matrix $P \in \mathbb{R}^{m\times m}$ mit*

$$A^T P + PA = -Q.$$

$\mathcal{L}\colon \mathbb{R}^m \to \mathbb{R}$, $\mathcal{L}(y) = y^T Py$ ist dann Lyapunov-Funktion für die asymptotische Stabilität von $\bar{y} = 0$ in (7).

Beweis. $P = \int_0^\infty \exp(tA^T)Q\exp(tA)\,\mathrm{d}t$ erfüllt

$$\begin{aligned} A^T P + PA &= \int_0^\infty A^T \exp(tA^T)Q\exp(tA) + \exp(tA^T)Q\exp(tA)A\,\mathrm{d}t \\ &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}[\exp(tA^T)Q\exp(tA)]\,\mathrm{d}t = [\exp(tA^T)Q\exp(tA)]_{t=0}^{\infty} = 0 - IQI = -Q. \end{aligned}$$

Damit ist $\nabla\mathcal{L}(y)\cdot Ay = 2y^T PAy = y^T(A^T P + PA)y = -y^T Qy < 0$. □

Für nichtlineare autonome gewöhnliche Differentialgleichungen kann man die asymptotische Stabilität nun auch über Linearisierung prüfen.

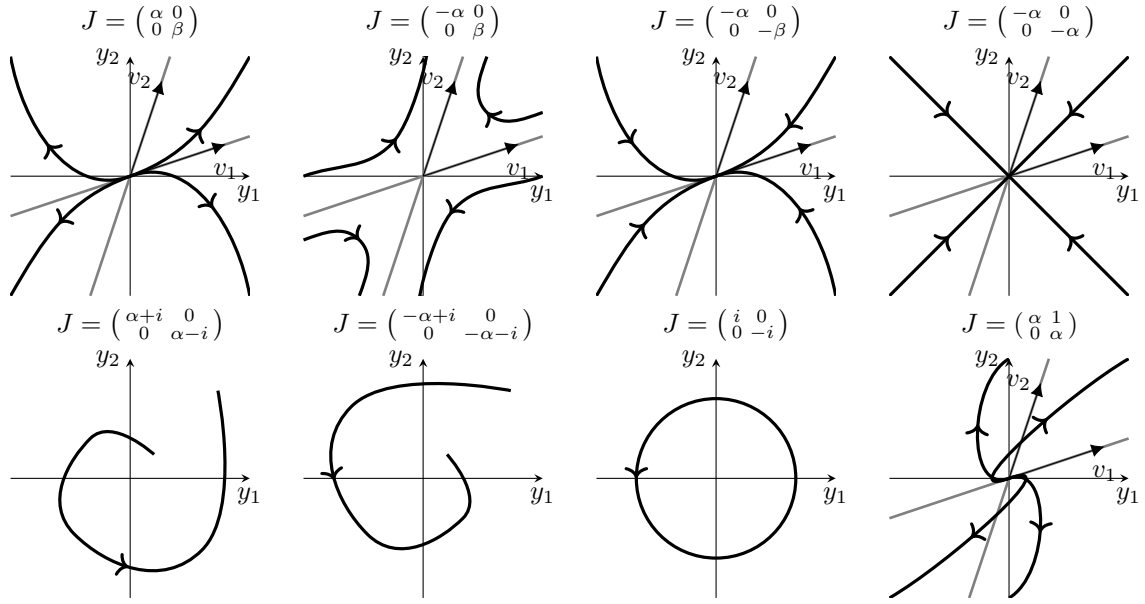
Korollar 49 (Linearisierte Stabilität)**.** *Sei $f : \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz-stetig und stetig differenzierbar in $\bar{y} \in \mathbb{R}^m$ mit $f(\bar{y}) = 0$. Die autonome gewöhnliche Differentialgleichung $y'(x) = f(y(x))$ hat den asymptotisch stabilen (instabilen) Gleichgewichtspunkt $\bar{y}$, wenn dies für ihre Linearisierung $y'(x) = A(y(x) - \bar{y})$ mit $A = Df(\bar{y})$ gilt.*

Beweis für Stabilität. O.B.d.A. sei $\bar{y} = 0$. Seien $\mathcal{L}$ und Q, P positiv definit gewählt wie in Theorem 48. Sei $\lambda > 0$ der kleinste Eigenwert von Q und $r > 0$ klein genug so dass $2\|P\|\|f(y) - Ay\| < \lambda\|y\|$ für $\|y\| < r$. Für $0 \neq \|y\| < r$ gilt dann

$$\nabla\mathcal{L}(y)\cdot f(y) = 2y^T P(Ay + (f(y) - Ay)) < -y^T Qy + \lambda\|y\|^2 \leq 0,$$

somit ist $\mathcal{L}$ auch eine Lyapunov-Funktion für $y'(x) = f(y(x))$. □

In 2D können nur wenige qualitativ unterschiedliche Gleichgewichtspunkte auftreten. Sei $J = V^{-1}AV$ die Jordan-Normalform von $A \in \mathbb{R}^{2\times 2}$ mit $V = (v_1\ v_2) \in \mathbb{C}^{2\times 2}$. Sei $0 < \alpha < \beta$, dann sind die wichtigsten Fälle:



Zum Schluss betrachten wir noch einmal lineare Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten,

$$0 = a_0 y(x) + a_1 y'(x) + \ldots + a_n y^{(n)}(x), \quad a_n = 1.$$

Die Reduktion ($u_i = y^{(i)}$) auf erste Ordnung liefert

$$u'(x) = \begin{pmatrix} 0 & 1 & 0 & \cdots & & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ & & \cdots & & & \\ 0 & & \cdots & & 0 & 1 \\ -a_0 & -a_1 & & \cdots & & -a_{n-1} \end{pmatrix} u(x).$$

Die Eigenwerte der Matrix sind genau die Nullstellen λ_j (mit Vielfachheit m_j) des Polynoms

$$\chi(\lambda) = a_0 + a_1\lambda + \ldots + a_n\lambda^n,$$

wobei der zu λ_j gehörige Jordanblock der Größe $m_j \times m_j$ ist (Hausaufgabe). Somit sind die Lösungen der Form

$$y(x) = \sum_j \exp(\lambda_j x) p_j(x)$$

für Polynome (mit komplexen Koeffizienten) p_j der Ordnung $m_j - 1$ (Hausaufgabe).

4 Numerische Methoden für Anfangswertprobleme

Sei $I = [a, b]$, $f : I \times \mathbb{R}^m \to \mathbb{R}^m$;

$$y'(x) = f(x, y(x)), \qquad y(a) = y_0 \tag{8}$$

habe eine eindeutige Lösung auf I.

Definition 50 (Numerisches Verfahren, Konvergenz)**.** 1. *Ein* Gitter *auf I ist ein Tupel $I_h = (x_0, \ldots, x_n) \in I^{n+1}$ mit* Stützstellen *(*Gitterpunkten*,* Knoten*)* $x_0 = a < x_1 < \ldots < x_n = b$.

2. *Die* Schritt- *oder* Gitterweite *ist $h = \max_{i=1,\ldots,n} h_i$ mit $h_i = x_i - x_{i-1}$.*

3. *Eine zugehörige* numerische Approximation *der Lösung y von* (8) *ist eine Abbildung $y_h : I_h \to \mathbb{R}^m$ bzw. das Tupel $(y_0, \ldots, y_n) = (y_h(x_0), \ldots, y_h(x_n)) \in (\mathbb{R}^m)^n$.*

4. *Der (globale)* Diskretisierungsfehler *der numerischen Approximation ist $e_h = \max_{i=0,\ldots,n} \|e_i\|$ mit $e_i = y_i - y(x_i)$.*

5. *Ein* numerisches Verfahren *für* (8) *ist eine Vorschrift, die für gegebenes h ein Gitter I_h und eine numerische Approximation y_h bestimmt.*

6. *Das Verfahren heißt* konvergent *(von Ordnung p), falls $e_h \to_{h\to 0} 0$ ($e_h = O(h^p)$).*

Konvergenzraten lassen sich effizient mit der O-Notation schreiben.

Definition 51 (Landau-Symbole)**.** *Seien $g, h : D \to \mathbb{R}$ für einen normierten Raum D (z.B. $D = \mathbb{R}$ oder $D = \mathbb{N}$).*

- $g \in O(h)$ *($x \to a$)* $\quad :\Leftrightarrow \quad h \in \Omega(g)$ *($x \to a$)* $\quad :\Leftrightarrow \quad \lim_{x\to a} \frac{\|g(x)\|}{\|h(x)\|} < \infty$
- $g \in o(h)$ *($x \to a$)* $\quad :\Leftrightarrow \quad h \in \omega(g)$ *($x \to a$)* $\quad :\Leftrightarrow \quad \lim_{x\to a} \frac{\|g(x)\|}{\|h(x)\|} < 0$

Bemerkung 52 (Landau-Symbole)**.**

- *a kann $\pm\infty$ sein*
- *a ist häufig aus Kontext klar $\Rightarrow$ „$(x \to a)$“ wird weggelassen*
- $O(h) = \{g : D \to \mathbb{R} \mid \lim_{x\to a} \frac{\|g(x)\|}{\|h(x)\|} < \infty\}$
- *man schreibt auch $g(x) = O(h(x))$ statt $g \in O(h)$*

Definition 53 (Verfahrenstypen)**.** *Ein Verfahren der Form*

- $y_{k+1} = y_k + h_{k+1}\varphi(x_k, y_k, h_{k+1})$ *heißt* explizites Einschrittverfahren.
- $y_{k+1} = y_k + h_{k+1}\varphi(x_k, y_k, y_{k+1}, h_{k+1})$ *heißt* implizites Einschrittverfahren.

- $y_{k+1} = y_k + h_{k+1}\varphi(x_0, \ldots, x_k, y_0, \ldots, y_k, h_{k+1})$ *heißt* explizites Mehrschrittverfahren.
- $y_{k+1} = y_k + h_{k+1}\varphi(x_0, \ldots, x_k, y_0, \ldots, y_k, y_{k+1}, h_{k+1})$ *heißt* implizites Mehrschrittverfahren.

Bemerkung 54 (Verfahrenstypen).
- *Man findet nacheinander $y_0, y_1, \ldots$. In einem impliziten Verfahren muss hierzu in jedem Schritt eine Gleichung gelöst werden.*
- *Ein Mehrschrittverfahren muss nicht alle Argumente $x_0, \ldots, x_k, y_0, \ldots, y_k$ nutzen.*

Theorem 55 (Wohldefiniertheit implizites Verfahren). *Sei $y_{k+1} = y_k + h_{k+1}\varphi(x_0, \ldots, x_k, y_0, \ldots, y_{k+1}, h_{k+1})$ ein implizites (Ein- oder Mehrschritt-)Verfahren für (8). Sei (analog zum Satz von Picard–Lindelöf) $B = \{y \in \mathbb{R}^m \mid \|y - y_0\| \le R\}$, φ Lipschitz-stetig in y_{k+1} mit Konstante L, $\|\varphi(\ldots)\| \le M$ für alle $x_0, \ldots, x_k \in I$ und $y_0, \ldots, y_{k+1} \in B$ (unabhängig von $k \in \mathbb{N}$) und $b - a \le \frac{R}{M}$. Dann existiert ein h, sodass bei Wahl von $h_k \le h$ für alle k das Verfahren in jedem Schritt k eine eindeutige Lösung $y_{k+1} = \Psi(x_0, \ldots, x_k, y_0, \ldots, y_k, h_{k+1}) \in B$ besitzt. Ist φ Lipschitz-stetig in $y_0, \ldots, y_k$, so auch Ψ.*

Bemerkung 56 (Implizite Verfahren). *Somit können wir ein implizites Verfahren auch als explizites interpretieren (ggfs. mit gleichen Lipschitz-Stetigkeits-Eigenschaften).*

Beweis. Sei $h < \frac{1}{L}$ und $h < \frac{R}{M}$ (h entspricht ϵ in Picard–Lindelöf), $h_k \le h \ \forall k$.
Zeige mit vollständiger Induktion $y_k \in B_k = \{y \in \mathbb{R}^m \mid \|y - y_0\| \le M(x_k - a)\}$ und die Satz-Aussagen.
Induktionsanfang $k = 0$: trivial.
Induktionsschritt $k \rightsquigarrow k + 1$:

- Setze $g(y) = y_k + h_{k+1}\varphi(x_0, \ldots, x_k, y_0, \ldots, y_k, y, h_{k+1})$.
- g ist Kontraktion: $g : B \to B_{k+1} \subset B$, denn für $y \in B$ ist
  $\|g(y) - y_0\| \le \|g(y) - y_k\| - \|y_k - y_0\| \le h\|\varphi(\ldots)\| + M(x_k - a) \le M(x_{k+1} - a)$.
  Außerdem $\|g(y) - g(z)\| = h_{k+1}\|\varphi(\ldots, y, h_{k+1}) - \varphi(\ldots, z, h_{k+1})\| \le hL\|y - z\|$.
- $\Longrightarrow \exists!$ Fixpunkt $y_{k+1} = \Psi(x_0, \ldots, x_k, y_0, \ldots, y_k, h_{k+1}) \in B_{k+1}$ von g, also eine eindeutige Lösung.
- Lipschitz-stetigkeit von Ψ: Lipschitz-Konstante von φ in $(y_0, \ldots, y_k)$ sei K.
  Sei $z = \Psi(x_0, \ldots, x_k, z_0, \ldots, z_k, h_{k+1})$, $u = \Psi(x_0, \ldots, x_k, u_0, \ldots, u_k, h_{k+1})$, dann ist

$$\begin{aligned}
\|z - u\| &= \|z_k + h_{k+1}\varphi(\ldots, z_0, \ldots, z_k, z, h_{k+1}) - u_k - h_{k+1}\varphi(\ldots, u_0, \ldots, u_k, u, h_{k+1})\| \\
&\le \|z_k - u_k\| + h\|\varphi(\ldots, z_0, \ldots, z_k, z, h_{k+1}) - \varphi(\ldots, u_0, \ldots, u_k, z, h_{k+1})\| \\
&\qquad + h\|\varphi(\ldots, u_0, \ldots, u_k, z, h_{k+1}) - \varphi(\ldots, u_0, \ldots, u_k, u, h_{k+1})\| \\
&\le \|z_k - u_k\| + hK\|(z_0, \ldots, z_k) - (u_0, \ldots, u_k)\| + hL\|z - u\|
\end{aligned}$$

$\Longrightarrow \|z - u\| \le \frac{1+hK}{1-hL}\|(z_0, \ldots, z_k) - (u_0, \ldots, u_k)\| \Longrightarrow \Psi$ hat Lipschitz-Konstante $\frac{1+hK}{1-hL}$. □

Bemerkung 57 (Lösung mit Fixpunktiteration). *Die Fixpunkt-Iteration $y_{k+1}^{i+1} = g(y_{k+1}^i)$ aus dem Beweis kann mit $y_{k+1}^0 = y_k$ genutzt werden, um y_{k+1} im impliziten Verfahren zu berechnen.*

Wie kann man sinnvolle Verfahren herleiten? Mittels Integralapproximation:

$$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x))\,\mathrm{d}t \approx y(x_k) + h_{k+1}\begin{cases} f(x_k, y(x_k)) & \text{Riemann links} \\ f(x_{k+1}, y(x_{k+1})) & \text{Riemann rechts} \\ (f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1})))/2 & \text{Trapez-Regel} \end{cases}$$

Beispiel 58 (Euler-Verfahren). *Zu (8) heißt*

$$y_{k+1} = y_k + h_{k+1}\begin{cases} f(x_k, y_k) & \text{explizites Euler-Verfahren,} \\ f(x_{k+1}, y_{k+1}) & \text{implizites Euler-Verfahren,} \\ (f(x_k, y_k) + f(x_{k+1}, y_{k+1}))/2 & \text{Trapez-Verfahren,} \\ (1-\theta)f(x_k, y_k) + \theta f(x_{k+1}, y_{k+1}) & \theta\text{-Verfahren,} \\ (f(x_k, y_k) + f(x_{k+1}, y_k + h_{k+1}f(x_k, y_k)))/2 & \text{verbessertes Euler-Verfahren.} \end{cases}$$

```
% löse y'(x) = x - y(x)^2, y(0) = 0
f = @(x,y) x-y^2;
h = .1; % Gitterweite
x = 0:h:1; % Gitter
n = length(x)-1;
y = zeros(3,n+1);
for k = 1:n
    % explizites Eulerverfahren
    y(1,k+1) = y(1,k) + h * f( x(k), y(1,k) );
    % implizites Eulerverfahren
    y(2,k+1) = ( -.5 + sqrt(.25+h*(h*x(k+1)+y(2,k))) ) / h;
    % verbessertes Eulerverfahren
    y(3,k+1) = y(3,k) + h * ( f( x(k), y(3,k) ) ...
                            + f( x(k+1), y(3,k)+h*f(x(k),y(3,k)) ) ) / 2;
end
plot(x',y','.-','Linewidth',3,'Markersize',20);
hold on;
x = linspace(0,1,100);
y = lsode(@(y,x)f(x,y),0,x);
plot(x,y,'k','Linewidth',3);
```

Das verbesserte Euler-Verfahren ist viel genauer – woran liegt das?

4.1 Konvergenz- und Stabilitätsanalyse von Einschrittverfahren

Wir beginnen mit einem Stabilitätsresultat für numerische Verfahren. Wie im Kontinuierlichen basiert es auf einem Gronwall-Lemma, diesmal werden jedoch Ableitungen und Integrale durch Differenzen und Summen ersetzt.

Lemma 59 (Gronwall diskret)**.** *Seien $\alpha_k, \beta_k, y_k \geq 0$ reelle Folgen mit $y_{k+1} - y_k \leq \alpha_k + \beta_k y_k \ \forall k$. Dann gilt $y_k \leq y_0 \exp(\sum_{j=0}^{k-1} \beta_j) + \sum_{i=0}^{k-1} \alpha_i \exp(\sum_{j=i+1}^{k-1} \beta_j)$.*

Beweis. Vollständige Induktion (Hausaufgabe). □

Der Einfachheit halber sei im Folgenden $h_k = h$ für alle k (die Verallgemeinerung zu h_k ist einfach).

Theorem 60 (Stabilität)**.** *Seien $\tilde{y}_0, \tilde{\varphi}$ Näherungen für y_0, φ mit $\|\tilde{y}_0 - y_0\| \leq \epsilon$, $\|\tilde{\varphi} - \varphi\|_\infty \leq \delta$ für ein (o.B.d.A. explizites) Einschrittverfahren $y_{k+1} = y_k + h\varphi(x_k, y_k, h)$. φ sei Lipschitz-stetig mit Konstante L. Die Lösung $\tilde{y}_k$ des gestörten Verfahrens erfüllt*

$$\|\tilde{y}_k - y_k\| \leq (\epsilon + \delta(x_k - a)) \exp(L(x_k - a)).$$

Man sagt, das Verfahren ist stabil.

Beweis. Sei $u_k = \tilde{y}_k - y_k$.

$$\begin{aligned}\|u_{k+1}\| &= \|\tilde{y}_k + h\tilde{\varphi}(x_k, \tilde{y}_k, h) - y_k - h\varphi(x_k, y_k, h)\| \\ &\leq \|u_k\| + h\left[\|\tilde{\varphi}(x_k, \tilde{y}_k, h) - \varphi(x_k, \tilde{y}_k, h)\| + \|\varphi(x_k, \tilde{y}_k, h) - \varphi(x_k, y_k, h)\|\right] \\ &\leq \|u_k\| + h\delta + hL\|u_k\|,\end{aligned}$$

mit diskretem Gronwall-Lemma folgt also

$$\|u_k\| \leq \|u_0\| \exp(khL) + \sum_{i=0}^{k-1} h\delta \exp((k-i-1)hL) \leq \epsilon \exp(L(x_k - a)) + (x_k - a)\delta \exp(L(x_k - a)). \quad \square$$

Bemerkung 61 (Schwächere Bedingung)**.** *Wir brauchen in Theorem 60 eigentlich nur $\max_k \|\tilde{\varphi}(x_k, \tilde{y}_k, h) - \varphi(x_k, \tilde{y}_k, h)\| \leq \delta$ statt $\|\tilde{\varphi} - \varphi\|_\infty \leq \delta$, also nur eine Abweichung $\leq \delta$ entlang der gestörten Lösung!*

Definition 62 (Abschneidefehler). *Für ein Verfahren* $y_{k+1} = y_k + h_{k+1}\varphi(x_0, \ldots, x_k, y_0, \ldots, y_k, y_{k+1}, h_{k+1})$ *und die Lösung* y *von* (8) *heißt*

$$\tau_k = \tfrac{1}{h_{k+1}}[y(x_{k+1}) - (y(x_k) + h_{k+1}\varphi(x_0, \ldots, x_k, y(x_0), \ldots, y(x_k), y(x_{k+1}), h_{k+1}))]$$

lokaler Diskretisierungsfehler *oder* Abschneidefehler.
Das Verfahren heißt konsistent *(von Ordnung* p*), falls* $\max_k \|\tau_k\| \to_{h\to 0} 0$ *(*$\max_k \|\tau_k\| = O(h^p)$*).*

Für den Abschneidefehler setzen wir quasi die kontinuierliche Lösung in das numerische Verfahren

$$0 = \frac{y_{k+1} - y_k}{h_{k+1}} - \varphi(\ldots)$$

ein und messen, wie stark sie die Gleichung verletzt.

Beispiel 63 (Abschneidefehler Euler-Verfahren).
- *Explizites Euler-Verfahren:*
  Taylor: $y(x_{k+1}) = y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(\xi_k)$ *für ein* $\xi_k \in [x_k, x_{k+1}]$
  $\tau_k = \frac{y(x_{k+1})-y(x_k)}{h} - f(x_k, y(x_k)) = y'(x_k) + \frac{h}{2}y''(\xi_k) - f(x_k, y(x_k)) = \frac{h}{2}y''(\xi_k)$
  $\Longrightarrow$ *konsistent von Ordnung* 1 *(sofern* f *ausreichend differenzierbar)*
- *Implizites Euler-Verfahren: konsistent von Ordnung* 1 *(Hausaufgabe)*
- *Verbessertes Euler-Verfahren: konsistent von Ordnung* 2 *(Hausaufgabe)*

Man kann sich merken: „Stabilität+Konsistenz=Konvergenz“

Theorem 64 (Konvergenz). *Das Einschrittverfahren* $y_{k+1} = y_k + h\varphi(x_k, y_k, h)$ *sei konsistent (von Ordnung* p*) und stabil. Dann ist das Verfahren konvergent (von Ordnung* p*).*

Beweis.
- Definiere $\tilde{\varphi}(\hat{x}, \hat{y}, h) = \frac{\tilde{y}(\hat{x}+h)-\tilde{y}(\hat{x})}{h}$ für die Lösung $\tilde{y}$ von $\tilde{y}'(x) = f(x, \tilde{y}(x))$ mit $\tilde{y}(\hat{x}) = \hat{y}$.
- Das Verfahren mit $\tilde{\varphi}$ liefert die exakte Lösung $y(x_k)$, das mit φ unsere numerische Approximation.
- $\tilde{\varphi}(\hat{x}, \hat{y}, h) - \varphi(\hat{x}, \hat{y}, h) = \frac{\tilde{y}(\hat{x}+h)-\tilde{y}(\hat{x})}{h} - \varphi(\hat{x}, \tilde{y}(\hat{x}), h) \overset{\hat{x}=x_k, \tilde{y}=y}{=} \tau_k$.
- Der Stabilitätssatz bzw. Bemerkung 61 liefert nun

$$\|e_k\| = \|y_k - y(x_k)\| \leq \max_j \|\tau_j\|(x_k - a)\exp(L(x_k - a)) \leq (b-a)\exp(L(b-a)) \cdot \max_j \|\tau_j\|. \quad \square$$

4.2 Typische Einschrittverfahren

Es gibt Verfahren beliebiger Konsistenzordnung, z.B. das folgende.

Definition 65 (Taylorverfahren). *Sei* $f \in C^p(\mathbb{R} \times \mathbb{R}^m)$ *und* $\tilde{y}$ *die Lösung zu* $\tilde{y}'(x) = f(x, \tilde{y}(x))$, $\tilde{y}(x_k) = y_k$*. Das Verfahren mit* $h\varphi(x_k, y_k, h) = \sum_{j=1}^p \frac{h^j}{j!}\tilde{y}^{(j)}(x_k)$*, der Taylorentwicklung von* $\tilde{y}(x_{k+1}) - y_k$ *und* x_k *von Ordnung* p*, heißt* Taylorverfahren.

Bemerkung 66 (Taylorverfahren). *Die Ableitungen* $\tilde{y}^{(j)}(x_k)$ *können, ohne* $\tilde{y}$ *zu kennen, mithilfe der Ableitungen von* f *und des Werts* y_k *bestimmt werden:*

$$\tilde{y}(x_k) = y_k,\ \tilde{y}'(x_k) = f(x_k, y_k),\ \tilde{y}''(x_k) = \tfrac{\mathrm{d}}{\mathrm{d}x}f(x, \tilde{y}(x))|_{x_k} = (\tfrac{\partial f}{\partial x} + \tfrac{\partial f}{\partial y}\tilde{y}')|_{\substack{x=x_k \\ y=y_k}} = (\tfrac{\partial f}{\partial x} + \tfrac{\partial f}{\partial y}f)|_{\substack{x=x_k \\ y=y_k}},\ \ldots \quad (9)$$

Theorem 67 (Taylorverfahren). *Das Taylorverfahren ist stabil und konsistent (somit konvergent) von Ordnung* p*.*

Beweis. Stabilität: φ ist Lipschitz-stetig in y_k (da $\tilde{y}^{(j)}(x_k)$ eine Summe von Produkten von bis zu $(j-1)$ten Ableitungen von f in (x_k, y_k) ist und $f \in C^p$).
Konsistenz: $\tau_k = \frac{\tilde{y}(x_{k+1})-\tilde{y}(x_k)}{h} - \varphi(x_k, y_k, h) = \sum_{j=1}^p \frac{h^{j-1}}{j!}\tilde{y}^{(j)}(x_k) + O(h^p) - \sum_{j=1}^p \frac{h^{j-1}}{j!}\tilde{y}^{(j)}(x_k) = O(h^p)$. $\square$

Im Taylorverfahren müssen hohe Ableitungen von f ausgewertet werden – oft sind die Formeln zu aufwändig oder gar nicht bekannt. Stattdessen kann man versuchen, höhere Ordnung nur mit Auswertungen von f zu erhalten. Hierzu approximiert man

$$y(x_{k+1}) - y(x_k) = \int_{x_k}^{x_{k+1}} f(x, y(x))\,\mathrm{d}x \approx h \sum_{i=0}^{p} c_i f(x_k + a_i h, y(x_k + a_i h))$$

mit einer Riemann-Summe (dies ist eine sog. *Quadraturformel*, siehe später). Darin muss $y(x_k + a_i h) = y(x_k) + \int_{x_k}^{x_k + a_i h} f(x, y(x))\,\mathrm{d}x$ wiederum mittels Quadratur approximiert werden.

Definition 68 (Runge–Kutta-Verfahren)**.** *Seien $c_i, a_i, b_{ij} \in \mathbb{R}$ für $i, j = 1, 2, \ldots$.*

- *Ein Verfahren mit*

$$\varphi(x, y, h) = \sum_{r=1}^{R} c_r k_r, \qquad k_r = f\left(x + a_r h, y + h \sum_{s=1}^{r-1} b_{rs} k_s\right), \qquad r = 1, \ldots, R$$

  heißt R-stufiges explizites Runge–Kutta-Verfahren*.*

- *Ein Verfahren mit*

$$\varphi(x, y, h) = \sum_{r=1}^{R} c_r k_r, \qquad k_r = f\left(x + a_r h, y + h \sum_{s=1}^{R} b_{rs} k_s\right), \qquad r = 1, \ldots, R$$

  heißt R-stufiges implizites Runge–Kutta-Verfahren*.*

- *Runge–Kutta-Verfahren werden im* Butcher-Diagramm *dargestellt,*

$$\begin{array}{c|ccc} a_1 & b_{11} & \cdots & b_{1R} \\ \vdots & \vdots & & \vdots \\ a_R & b_{R1} & \cdots & b_{RR} \\ \hline & c_1 & \cdots & c_R \end{array}$$

  (für explizite Verfahren ist $b_{rs} = 0$ für $s \geq r$).

Beispiel 69 (Runge–Kutta-Verfahren)**.**

- *Euler-Verfahren:* *explizit* $\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$ *implizit* $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ *modifiziert* $\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$ *verbessert* $\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$
- *4.-Ordnung Runge–Kutta-Verfahren* $\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & & \frac{1}{2} & & \\ 1 & & & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$

Runge–Kutta-Verfahren sind der Standard unter den Einschrittverfahren. Geeignete Parameter c_i, a_i, b_{ij} erhält man durch Ermitteln des Abschneidefehlers.

Beispiel 70 (2-stufige explizite Runge–Kutta-Verfahren)**.** *2-stufige explizite Runge–Kutta-Verfahren haben die Form*

$$y_{k+1} = y_k + h(c_1 k_1 + c_2 k_2), \qquad k_1 = f(x_k + a_1 h, y_k), \qquad k_2 = f(x_k + a_2 h, y_k + b_{21} h k_1).$$

$$\begin{aligned} \Longrightarrow \quad \tau_k &= \tfrac{y(x_{k+1}) - y(x_k)}{h} - c_1 f(x_k + a_1 h, y(x_k)) - c_2 f(x_k + a_2 h, y(x_k) + b_{21} h f(x_k + a_1 h, y(x_k))) \\ &= y'(x_k) + \tfrac{h}{2} y''(x_k) + O(h^2) - c_1 y'(x_k) - c_1 h a_1 \partial_x f - c_2 [f(x_k, y(x_k)) + h(a_2 \partial_x f + b_{21} f \partial_y f) + O(h^2))] \\ &\overset{(9)}{=} y'(x_k)(1 - c_1 - c_2) + \partial_x f(x_k, y(x_k)) h(\tfrac{1}{2} - c_1 a_1 - c_2 a_2) + f(x_k, y(x_k)) \partial_y f(x_k, y(x_k)) h(\tfrac{1}{2} - c_2 b_{21}) + O(h^2) \end{aligned}$$

$\Longrightarrow$ *wähle* $c_1 + c_2 = 1$, $c_2 b_{21} = \frac{1}{2}$, $c_1 a_1 + c_2 a_2 = \frac{1}{2}$ $\Longrightarrow$ *Verfahren ist konsistent von Ordnung 2.*

- $a_2 = \frac{1}{2},\ a_1 = 0 \Rightarrow y_{k+1} = y_k + hf(x_k + \frac{h}{2}, y_k + \frac{h}{2}f(x_k, y_k))$ *„modifiziertes Euler-Verfahren“*
- $a_2 = 1,\ a_1 = 0 \Rightarrow y_{k+1} = y_k + h\frac{f(x_k,y_k)+f(x_k+h,y_k+hf(x_k,y_k))}{2}$ *„verbessertes Euler-Verfahren“*

Theorem 71 (Runge–Kutta-Verfahren). *1. f Lipschitz in $y \Rightarrow$ Runge–Kutta-Verfahren ist stabil.*

2. Ist $f \in C^1$ und $\sum_{r=1}^R c_r = 1$, so ist das Verfahren konsistent von Ordnung ≥ 1.

3. Ist zusätzlich $f \in C^2$, $\sum_{r=1}^R c_r \sum_{s=1}^R b_{rs} = \frac{1}{2}$, $\sum_{r=1}^R a_r c_r = \frac{1}{2}$, ist es konsistent von Ordnung ≥ 2.

4. Es gibt kein zweistufiges explizites Verfahren der Konsistenzordnung ≥ 3.

Beweis. 1. $\varphi(x, y, h) =$ Summe und Komposition von f, welches Lipschitz-stetig in y ist $\Rightarrow$ auch φ

2. $\tau_k = \frac{y(x_{k+1})-y(x_k)}{h} - \sum_{r=1}^R c_r f(x_k + arh, y(x_k) + h\sum_{s=1}^R b_{rs}k_s)$
$= y'(x_k) + \frac{h}{2}y''(x_k) + O(h^2) - \sum_{r=1}^R c_r[f(x_k, y(x_k)) + h\{a_r\partial_x f + \partial_y f \sum_{s=1}^R b_{rs}k_s\}_{x_k,y(x_k)} + O(h^2)]$
$= h[\frac{y''(x_k)}{2} - \sum_{r=1}^R c_r\{a_r\partial_x f + \partial_y f \sum_{s=1}^R b_{rs}k_s\}_{x_k,y(x_k)}] + O(h^2)$

3. $\tau_k = h[\frac{y''(x_k)}{2} - \sum_{r=1}^R c_r\{a_r\partial_x f + \partial_y f \sum_{s=1}^R b_{rs}(f + O(h))\}_{x_k,y(x_k)}] + O(h^2)$
$= h[\frac{y''(x_k)}{2} - \frac{1}{2}\{\partial_x f + f\partial_y f\}_{x_k,y(x_k)} + O(h)] + O(h^2) \overset{(9)}{=} O(h^2)$

4. O.B.d.A. sei (8) autonom.

$\tau_k = \frac{y(x_{k+1})-y(x_k)}{h} - c_1 f(y(x_k)) - c_2 f(y(x_k) + b_{21}hf(y(x_k)))$
$= y'(x_k) + \frac{h}{2}y''(x_k) + \frac{h^2}{6}y'''(x_k) + O(h^3) - c_1 y'(x_k) - c_2[f + b_{21}hff' + \frac{b_{21}^2h^2}{2}f^2f'' + O(h^3)]_{y=y(x_k)}$
$= h^2[\frac{y'''}{6} - \frac{c_2}{2}b_{21}^2 f^2(y)f''(y)]_{x=x_k} + O(h^3),$

jedoch ist $y'''(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}f(y(x)) = \frac{\mathrm{d}}{\mathrm{d}x}[f'(y(x))f(y(x))] = f'(y(x))^2 f(y(x)) + f(y(x))^2 f''(y(x))$. □

Bemerkung 72 (Runge–Kutta-Verfahren). • *Die Konsistenzordnung eines expliziten R-stufigen Runge–Kutta-Verfahrens ist $\leq R$, z.B. ist die maximal mögliche Konsistenzordnung eines 5-stufigen Verfahrens 4.*

- *Für speziell gewählte Stützstellen a_i und Gewichte c_i (der Gauß-Quadratur, siehe später) gibt es b_{ij}, sodass das implizite R-stufige Verfahren die maximal mögliche Konsistenzordnung $2R$ hat.*
- *Matlab & Octave impliementieren mit* `ode45` *das* Dormand–Prince-Verfahren *(siehe Wikipedia: Dormand–Prince-method). Dies sind zwei 7-stufige Runge–Kutta-Verfahren mit Konsistenzordnung 4 bzw. 5, die sich nur in den c_i unterscheiden und somit die gleichen Funktionsauswertungen brauchen. Die Differenz beider Methoden liefert eine Fehlerschätzung für das 4.-Ordnung-Verfahren und kann somit genutzt werden, um die Schrittweite h adaptiv (d.h. so, dass eine vorgegebene Fehlerschranke nicht überschritten wird) anzupassen.*

4.3 Lineare Mehrschrittverfahren

Einschrittverfahren haben den Nachteil, dass sie alle vorherigen Funktionsauswertungen wieder vergessen und y_{k+1} nur aus y_k erhalten. Für hohe Konsistenzordnungen muss f in jedem Schritt häufig ausgewertet werden. Mehrschrittverfahren hingegen benutzen die alten Auswertungen weiter.

Definition 73 (Lineares Mehrschrittverfahren). *Sei die Schrittweite konstant, $h = h_k\ \forall k$. Ein Verfahren der Form*

$$\sum_{i=0}^{n} \alpha_i y_{k+i} = h\sum_{i=0}^{n} \beta_i f(x_{k+i}, y_{k+i}), \quad \alpha_n \neq 0,$$

*zu lösen für y_{k+n}, heißt lineares n-*Schritt-Verfahren *(explizit für $\beta_n = 0$, implizit für $\beta_n \neq 0$).*

Bemerkung 74 (Mehrschrittverfahren durch Polynominterpolation oder Quadratur). *Lineare Mehrschrittverfahren sind durch Polynominterpolation oder Quadratur motiviert (kommt später):*

- *Sei $p = P(x_k, \dots x_{k+n}, y_k, \dots, y_{k+n})$ das Polynom von Grad n mit $p(x_i) = y_i$, dann ist Gleichung $P(x_k, \dots x_{k+n}, y_k, \dots, y_{k+n})'(x_{k+s}) = f(x_{k+s}, y_{k+s})$ ein Mehrschrittverfahren, zu lösen für y_{k+n}.*
- *Die Quadraturformel $y(x_{k+n}) - y(x_k) = \int_{x_k}^{x_{k+n}} f(x, y(x))\,\mathrm{d}x \approx \sum_{i=0}^{n} \beta_i f(x_{k+i}, y(x_{k+i}))$ liefert das Verfahren $y_{k+n} - y_k = \sum_{i=0}^{n} \beta_i f(x_{k+i}, y_{k+i})$, zu lösen für y_{k+n}.*

Beispiel 75 (Mehrschrittverfahren)**.** *Kürze ab $f_k := f(x_k, y_k)$.*

- *Simpson-Regel: $y_{k+2} - y_k = \frac{h}{3}[f_k + 4f_{k+1} + f_{k+2}]$ $\left(\approx \int_{x_k}^{x_{k+2}} f(x, y(x))\,\mathrm{d}x\right)$*
- *Adams–Bashforth-Verfahren: $y_{k+4} = y_{k+3} + \frac{h}{24}[55f_{k+3} - 59f_{k+2} + 37f_{k+1} - 9f_k]$*
- *Adams–Moulton-Verfahren: $y_{k+3} = y_{k+2} + \frac{h}{24}[9f_{k+3} + 19f_{k+2} - 5f_{k+1} + f_k]$*

Bemerkung 76 (Lineare Mehrschrittverfahren)**.**

- *$y_0, \dots, y_{n-1}$ sind Eingabedaten für ein n-Schritt-Verfahren (sie müssen mit einem anderen Verfahren berechnet werden).*
- *Oft werden lineare Mehrschrittverfahren in Paaren als Prädiktor-Korrektor-Verfahren genutzt: Der Prädiktor, ein explizites Verfahren, liefert einen einfach zu berechnenden Startwert für eine Fixpunkt-Iteration für den Korrektor, ein implizites Verfahren.*
- *Das Verfahren kann geschrieben werden als $y_{k+1} = y_k + h\varphi(x_0, \dots, x_{k+1}, y_0, \dots, y_{k+1}, h)$ für $\varphi(\dots) = \sum_{i=0}^{n} \beta_i f(x_{k+i+1-n}, y_{k+i+1-n}) + \frac{1}{h}[y_{k+1} - y_k - \sum_{i=0}^{n} \alpha_i y_{k+i+1-n}]$, der* Abschneidefehler *wird daher oft eingeführt als*

$$\tau_{k+n-1} = \frac{1}{h}\sum_{i=0}^{n} \alpha_i y(x_{k+i}) - \sum_{i=0}^{n} \beta_i \underbrace{f(x_{k+i}, y(x_{k+i}))}_{y'(x_{k+i})}.$$

Definition 77 (Charakteristisches Polynom)**.** *Das* erste und zweite charakteristische Polynom *eines Mehrschrittverfahrens $\sum_{i=0}^{n} \alpha_i y_{k+i} = h\sum_{i=0}^{n} \beta_i f(x_{k+i}, y_{k+i})$ sind gegeben durch*

$$\rho(x) = \sum_{j=0}^{n} \alpha_j x^j \qquad \& \qquad \sigma(x) = \sum_{j=0}^{n} \beta_j x^j.$$

Theorem 78 (Konsistenz)**.** *Sei $C_0 = \rho(1)$, $C_1 = \rho'(1) - \sigma(1)$, $C_q = \sum_{j=1}^{n} \frac{j^q}{q!}\alpha_j - \sum_{j=1}^{n} \frac{j^{q-1}}{(q-1)!}\beta_j$ $\forall q > 1$. Das n-Schritt-Verfahren ist konsistent von Ordnung p, wenn $C_0 = \dots = C_p = 0$.*

Beweis. Sei $y \in C^{p+1}$, Taylor liefert $y(x_{k+i}) = \sum_{j=0}^{p} \frac{y^{(j)}(x_k)}{j!}(ih)^j + O(h^{p+1})$.

$$\begin{aligned}
\implies \tau_{k+n-1} &= \sum_{i=0}^{n} \frac{\alpha_i y(x_{k+i}) - h\beta_i y'(x_{k+i})}{h} = \frac{1}{h}\sum_{i=0}^{n}\left(\alpha_i \sum_{j=0}^{p} \frac{y^{(j)}(x_k)}{j!}(ih)^j - h\beta_i \sum_{j=0}^{p} \frac{y^{(j+1)}(x_k)}{j!}(ih)^j\right) + O(h^p) \\
&= \frac{y(x_k)}{h}\left[\sum_{i=0}^{n}\alpha_i\right] + y'(x_k)\left[\sum_{i=0}^{n} i\alpha_i - \sum_{i=0}^{n}\beta_i\right] + \sum_{j=2}^{p} y^{(j)}(x_k)\left[\sum_{i=0}^{n}\alpha_i \frac{i^j}{j!} - \sum_{i=0}^{n}\beta_i \frac{i^{j-1}}{(j-1)!}\right]h^{j-1} + O(h^p) \\
&= \sum_{j=0}^{p} C_j y^{(j)}(x_k) h^{j-1} + O(h^p)
\end{aligned}$$

□

Gegeben Eingabedaten $y_0, \dots, y_{n-1}$ definiert man Stabilität für Mehrschrittverfahren wie folgt (sog. „Nullstabilität“, da es ausreichen wird, die Differentialgleichung $y'(x) = 0$ zu betrachten).

Definition 79 (Nullstabilität)**.** *Ein lineares n-Schritt-Verfahren $y_{k+1} = y_k + h\varphi(x_k, y_{k-n+1}, \dots, y_k, y_{k+1}, h)$ mit Schrittweite h auf $I = [a, b]$ heißt* (asymptotisch) stabil *oder* nullstabil*, wenn $K, h_0 > 0$ existieren, sodass Folgendes gilt: Sind $\tilde{y}_0, \dots, \tilde{y}_{n-1}, \tilde{\varphi}$ Näherungen für $y_0, \dots, y_{n-1}, \varphi$ mit $\|\tilde{y}_i - y_i\| \le \epsilon$, $\|\tilde{\varphi} - \varphi\|_\infty \le \delta$, dann erfüllen die Lösungen y_k und $\tilde{y}_k$ des ursprünglichen und des gestörten Verfahrens mit $h \le h_0$*

$$\max_k \|\tilde{y}_k - y_k\| \le K(\epsilon + \delta).$$

Beispiel 80 (Instabiles Mehrschrittverfahren). *Das 2-Schritt-Verfahren*

$$y_{k+2} + 4y_{k+1} - 5y_k = h(4f(x_{k+1}, y_{k+1}) + 2f(x_k, y_k))$$

hat Konsistenzordnung 3 (Hausaufgabe).

```
f = @(x,y) 0;   % oder 1
for h = logspace(-1,-1.5,3)
    x = 0:h:1;
    n = length(x);
    y = zeros(1,n);
    y(2) = eps; % oder x(2)
    for k = 3:n
        y(k) = 5*y(k-2) - 4*y(k-1) + h*( 2*f(x(k-2),y(k-2)) + 4*f(x(k-1),y(k-1)) );
    end;
    plot(x,y,'Linewidth',3); hold on; pause;
end;
```

(Rundungs-)Fehler in Anfangswerten oder Rechenschritten werden verstärkt (umso mehr je kleiner h) $\Longrightarrow$ *Verfahren ist weder stabil noch konvergent!*

Es gibt ein einfaches Kriterium, Nullstabilität von Mehrschrittverfahren zu prüfen.

Theorem 81 (Wurzelbedingung von Dahlquist). *Ein lineares Mehrschrittverfahren ist nullstabil für* (8) *mit (beliebigem) Lipschitz-stetigem f genau dann, wenn alle Nullstellen des ersten charakteristischen Polynoms ρ im komplexen Einheitskreis liegen, wobei Nullstellen auf dem Rand einfach sein müssen.*

Beispiel 82 (Instabiles Mehrschrittverfahren II). *Obiges 2-Schritt-Verfahren hat $\rho(x) = x^2 + 4x - 5$ mit Nullstellen $\{-5, 1\}$* $\Longrightarrow$ *instabil.*

Der Beweis erfordert eine Einführung in Differenzengleichungen.

Definition 83 (Polynome vom Grad n). *Mit $\mathcal{P}_n = \{p : \mathbb{C} \to \mathbb{C} \,|\, p(x) = a_0x^0 + \ldots + a_nx^n$ für $a_0, \ldots, a_n \in \mathbb{C}\}$ bezeichnen wir die komplexen* Polynome vom Grad $\leq n$.

Definition 84 (Differenzengleichung). *Die* lineare Differenzengleichung *zu den Koeffizienten $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$, $\alpha_n \neq 0$, und der Folge $b_0, b_1, \ldots$ in $\mathbb{R}$ ist*

$$\sum_{i=0}^{n} \alpha_i y_{k+i} = b_k. \tag{10}$$

Für $0 = b_0 = b_1 = \ldots$ heißt sie homogen, *sonst* inhomogen.

Theorem 85 (Differenzengleichungen).

1. *Die Menge aller Lösungen $y_0, y_1, \ldots$ von* (10) *mit $0 = b_0 = b_1 = \ldots$ ist ein n-dimensionaler Vektorraum.*
2. *y_k löst die homogene Gleichung $\Leftrightarrow$ $y_k = \sum_{j=1}^{r} q_j(k)\lambda_j^k$ für $\lambda_1, \ldots, \lambda_r$ die Nullstellen des charakteristischen Polynoms ρ mit Vielfachheiten $n_1, \ldots, n_r$ und $q_j \in \mathcal{P}_{n_j-1}$, $j = 1, \ldots, r$.*
3. *Alle Lösungen von* (10) *haben die Form $y = \sum_{k=0}^{\infty} b_k v^k + \sum_{k=0}^{n-1} y_k w^k$, wobei*
   - *w^k die Lösung der homogenen Gleichung mit Anfangswerten $w_i^k = \delta_{ik}$, $i = 0, \ldots, n-1$, ist,*
   - *$v^k = (\underbrace{0, \ldots, 0}_{(k+1)-mal}, \frac{1}{\alpha_n} w^{n-1})$.*

Beweis. 1.
- Vektorraumeigenschaften sind klar.
- Wegen $\alpha_n \neq 0$ ist eine Lösung eindeutig durch ihre Anfangswerte $y_0, \ldots, y_{n-1}$ bestimmt.
- Der Raum der Anfangswerte $(y_0, \ldots, y_{n-1})$ ist n-dimensional.

2. (Idee: Sei $y_k = \lambda_j^k$, dann ist $\sum_{i=0}^n \alpha_i y_{k+i} = \lambda_j^k \sum_{i=0}^n \alpha_i \lambda_j^i = 0$, also y_k eine Lösung.)
„$\Rightarrow$“: Sei y_k eine Lösung, dann erfüllt $Y_k := (y_k, y_{k+1}, \dots, y_{k+n-1})^T$

$$Y_{k+1} = AY_k \qquad \text{für } A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\frac{\alpha_0}{\alpha_n} & -\frac{\alpha_1}{\alpha_n} & \cdots & \cdots & -\frac{\alpha_{n-1}}{\alpha_n} \end{pmatrix}.$$

Sei $J = S^{-1}AS$ die Jordannormalform; wir wissen (alte Hausaufgabe):

- Eigenwerte λ_i (mit Vielfachheit n_i) sind die (komplexen) Nullstellen des charakteristischen Polynoms χ_A von A und somit von $\rho = \alpha_n \chi_A$.
- Zugehörige Jordanblöcke J_i haben Größe $n_i \times n_i$.
- $\Longrightarrow A^k = SJ^kS^{-1} = S \begin{pmatrix} J_1^k & & \\ & \ddots & \\ & & J_r^k \end{pmatrix} S^{-1}$ mit $J_i^k = (\lambda_i I + N_i)^k = \sum_{j=0}^{n_i-1} \binom{k}{j} N_i^j \lambda_i^{k-j} = \sum_{j=0}^{n_i-1} p_j^i(k) N_i^j \lambda_i^k$ für $p_j^i(k) = \lambda_i^{-j}\binom{k}{j} = \lambda_i^{-j} \frac{k(k-1)\cdots(k-j+1)}{j!}$ in $\mathcal{P}_j \subset \mathcal{P}_{n_i-1}$.
- $\Longrightarrow$ Einträge von J^k und somit auch A^k sind Linearkombinationen von Termen $p^i(k)\lambda_i^k$, $p^i \in \mathcal{P}_{n_i-1}$, $i = 1, \dots, r$
- $\Longrightarrow$ Einträge von $Y_k = A^k Y_0$ und somit von y_k sind auch solche Linearkombinationen

„$\Leftarrow$“: Sei V der Vektorraum aller Folgen der Form $y_k = \sum_{i=1}^r q_i(k)\lambda_i^k$; er hat Dimension $n_1 + \ldots + n_i = n$. Der Lösungsraum W ist ein n-dimensionaler Teilraum, also $V = W$.

3. - 1. $\Rightarrow w^0, \dots, w^{n-1}$ bilden eine Basis der Lösungen der homogenen Gleichung, d.h. jede solche Lösung lässt sich schreiben als $\sum_{k=0}^{n-1} a_k w^k$, $a_k \in \mathbb{R}$.
   - Seien z_1, z_2 zwei Lösungen von (10) $\Rightarrow z_1 - z_2$ löst homogene Gleichung $\Longrightarrow$ jede Lösung von (10) ist von der Form $z_1 + \sum_{k=0}^{n-1} a_k w^k$ für eine spezielle Lösung z_1 und $a_0, \dots, a_{n-1} \in \mathbb{R}$.
   - Noch zu zeigen: $z_1 = \sum_{k=0}^{\infty} b_k v^k$ ist spezielle Lösung (dann folgt $a_i = y_i$ wegen $v_i^k = 0 \; \forall i < n$).
   - z_1 ist wohldefiniert, da $(z_1)_j = \sum_{k=0}^{\infty} b_k v_j^k = \sum_{k=0}^{j-n} b_k v_j^k$ für alle j nur endlich viele Summanden enthält. Außerdem ist $\sum_{j=0}^n \alpha_j (z_1)_{l+j} = \sum_{k=0}^{\infty} b_k \sum_{j=0}^n \alpha_j v_{l+j}^k = b_l$, da

$$\sum_{j=0}^n \alpha_j v_{l+j}^k = \begin{cases} 0 & \text{falls } l < k, \\ \frac{1}{\alpha_n} \sum_{j=0}^n \alpha_j w_{l-k-1+j}^{n-1} & \text{sonst,} \end{cases} \qquad \sum_{j=0}^n \alpha_j w_{l-k-1+j}^{n-1} = \begin{cases} 0 & \text{falls } l > k, \\ \alpha_n & \text{falls } l = k. \end{cases}$$ □

Beweis Theorem 81. Notwendigkeit: Betrachte $y'(x) = 0$, $y(0) = 0$ auf $[0, 1]$. Das n-Schritt-Verfahren lautet $\sum_{i=0}^n \alpha_i y_{k+i} = 0$.

- Sei λ Nullstelle von ρ mit $|\lambda| > 1$. Betrachte Lösung $y_k = 0$ und gestörte Lösung $u_k = \epsilon\lambda^{k-n}$ mit Anfangswertfehler $|u_k - y_k| \le \epsilon$, $k = 0, \dots, n-1$. Für $N \in \mathbb{N}$ sei Schrittweite $h = \frac{1}{N}$ und y_N bzw. u_N die Approximation an $y(1)$, dann gilt $|u_N - y_N| = \epsilon|\lambda|^{1/h-n} \to_{h\to 0} \infty \Rightarrow$ Verfahren ist instabil.
- Ist λ Nullstelle mit $|\lambda| = 1$ und Vielfachheit ≥ 2, wiederhole Argument für $u_k = \epsilon\frac{k}{n}\lambda^k$.

Hinlänglichkeit: Das n-Schritt-Verfahren sei $\sum_{i=0}^n \alpha_i y_{k+i} = h\sum_{i=0}^n \beta_i f(x_{k+i}, y_{k+i}) =: h\Psi(x_k, y_k, \dots, y_{k+n}, h)$.
Sei $\tilde{L}$ die Lipschitz-Konstante von f, dann ist Ψ Lipschitz in $(y_k, \dots, y_{k+n})$ mit Konstante $L = \tilde{L}\sum_{i=0}^n |\beta_i|$.
Sei y_k eine Lösung und u_k eine gestörte Lösung, setze

$$e_k = u_k - y_k, \quad \epsilon = \max\{e_0, \dots, e_{n-1}\}, \quad F_k = (\tilde{\Psi} - \Psi)(x_k, u_k, \dots, u_{k+n}, h), \quad \delta = \max_k \|F_k\|.$$

- Es gilt $\sum_{i=0}^n \alpha_i e_{k+i} = h\left[\Psi(x_k, u_k, \dots, u_{k+n}, h) + F_k - \Psi(x_k, y_k, \dots, y_{k+n}, h)\right] =: hb_k \; \forall k$.
- $\Rightarrow e = \sum_{k=0}^{\infty} hb_k v^k + \sum_{k=0}^{n-1} e_k w^k$ (Notation aus Theorem 85)
- Die w^k und somit auch v^k sind wegen der Wurzelbedingung beschränkt, $|v_j^k|, |w_j^k| \le M \; \forall j, k$ (denn w^k hat die Form $w_l^k = \sum_{i=1}^r q_i(l)\lambda_i^l$ für q_i Polynome und λ_i Nullstellen von ρ mit $|\lambda_i| < 1$ oder $|\lambda_i| = 1$ und $q_i \in \mathcal{P}_0$; dies ist beschränkt).

- $\|b_k\| \leq \|F_k\| + L\delta_k$ für $\delta_k = \max_{j=0,\ldots,n} \|e_{k+j}\|$

$$\Rightarrow \|e_l\| = \left\| \sum_{k=0}^{l-n} h b_k v_l^k + \sum_{k=0}^{n-1} e_k w_l^k \right\| \leq hM \sum_{k=0}^{l-n} (\|F_k\| + L\delta_k) + Mn\epsilon \leq M[n\epsilon + h(l-n+1)\delta] + hML \sum_{k=0}^{l-n} \delta_k$$

$$\Rightarrow \delta_l \leq \max_{j=0,\ldots,n} \left[M[n\epsilon + h(l+j-n+1)\delta] + hML \sum_{k=0}^{l-n+j} \delta_k \right] = M[n\epsilon + h(l+1)\delta] + hML \sum_{k=0}^{l} \delta_k$$

$$\Leftrightarrow \delta_l(1 - hML) \leq M[n\epsilon + h(l+1)\delta] + hML \sum_{k=0}^{l-1} \delta_k \leq M[n\epsilon + (b-a)\delta] + hML \sum_{k=0}^{l-1} \delta_k$$

Wähle nun $h_0 = \frac{1}{2ML}$, somit $\delta_l \leq 2M[n\epsilon + (b-a)\delta] + h2ML \sum_{k=0}^{l-1} \delta_k$. Hieraus folgt mittels vollständiger Induktion (analog zum diskreten Gronwall-Lemma)

$$\delta_l \leq 2M[n\epsilon + (b-a)\delta] e^{lh2ML}$$

und somit $\|e_k\| \leq \delta_k \leq 2M[n\epsilon + (b-a)\delta] e^{2ML(b-a)}$ $\forall k$. □

Bemerkung 86 (Spezialfälle der Stabilität).

- *Einschrittverfahren haben $\rho(x) = x - 1 \Rightarrow$ stabil.*
- *Aus Quadratur hergeleitete n-Schritt-Verfahren haben $\rho(x) = x^n - 1 \Rightarrow$ stabil.*

Wieder gilt „Stabilität+Konsistenz=Konvergenz“.

Theorem 87 (Konvergenz). *Ein stabiles und Ordnung-p-konsistentes lineares Mehrschrittverfahren ist konvergent von Ordnung p, sofern die Anfangswerte $y_0, \ldots, y_{n-1}$ konsistent von Ordnung p sind (d.h. $\|y_i - y(x_i)\| = O(h^p)$, $i = 0, \ldots, n-1$).*

Beweis. Analog zu Einschrittverfahren (Hausaufgabe). □

Bemerkung 88 (Sinnfreie konsistente Verfahren). *Es gibt konsistente Mehrschrittverfahren, die unabhängig von der Differentialgleichung sind (Hausaufgabe). Diese sind natürlich nicht konvergent und somit nicht stabil.*

Welche Konvergenzordnung (=Konsistenzordnung) kann ein lineares n-Schritt-Verfahren haben? Für $p$te Ordnung erhalten wir die $p+1$ Bedingungen $C_0 = \ldots = C_p = 0$ für die C_i aus Theorem 78. Dies sind $p+1$ lineare Bedingungen für die $2n+1$ zu wählenden Koeffizienten $\alpha_0, \ldots, \alpha_{n-1}, \beta_0, \ldots, \beta_n$ (α_n kann zu 1 normiert werden) für implizite bzw. $2n$ Koeffizienten $\alpha_0, \ldots, \alpha_{n-1}, \beta_0, \ldots, \beta_{n-1}$ für explizite Verfahren. Somit ist die maximal erreichbare Konsistenzordnung $p = 2n$ für implizite und $p = 2n-1$ für explizite Verfahren – wir werden jedoch sehen, dass stabile Verfahren nur eine geringere Ordnung erreichen.

Korollar 89 (Konsistenzordnung Mehrschrittverfahren). *Ein lineares n-Schritt-Verfahren hat Konsistenzordnung p genau dann, wenn $\Phi : \mathbb{C} \to \mathbb{C}$, $\Phi(z) = \frac{\rho(z)}{\log(z)} - \sigma(z)$ in $z = 1$ eine p-fache Nullstelle hat.*

Beweis. Φ hat p-fache Nullstelle in $z = 1$
$\Leftrightarrow \Phi(e^h) = \frac{1}{h}[\rho(e^h) - h\sigma(e^h)]$ hat in $h = 0$ eine p-fache Nullstelle
$\Leftrightarrow g(h) = h\Phi(e^h) = \sum_{i=0}^{n} (\alpha_i - h\beta_i) e^{ih}$ hat in $h = 0$ eine $(p+1)$-fache Nullstelle
$\Leftrightarrow \underbrace{g(0)}_{=C_0} = \underbrace{g'(0)}_{=C_1} = \ldots = \underbrace{g^{(p)}(0)}_{=p!C_p} = 0$ □

Das folgende Korollar zeigt, wie für vorgegebene $\alpha_0, \ldots, \alpha_n$ (bzw. ρ) die bzgl. der Konsistenz optimalen β_i (bzw. σ) berechnet werden (natürlich können wir dies bereits mittels Taylorentwicklung).

Korollar 90 (Optimales charakteristisches Polynom). *Sei $\rho \in \mathcal{P}_n$ mit $\rho(1) = 0$ und $0 \leq l \leq n$. Es gibt ein eindeutiges Polynom $\sigma \in \mathcal{P}_l$, für das das zugehörige n-Schritt-Verfahren Konsistenzordnung $\geq l+1$ hat. Es sind $\sigma(z) = \sum_{i=0}^{l} c_i (z-1)^i$ die ersten $l+1$ Terme der Taylorentwicklung von $\rho(z)/\log z$.*

Beweis. $\frac{\rho(z)}{\log z}$ ist holomorph um $z = 1$ und hat somit eine Taylorentwicklung $\sum_{i=0}^{l} c_i (z-1)^i + O((z-1)^{l+1})$
$\Rightarrow \Phi(z) = \frac{\rho(z)}{\log z} - \sigma(z) = \sum_{i=0}^{l} c_i (z-1)^i - \sigma(z) + O((z-1)^{l+1})$ hat genau dann eine $(l+1)$-fache Nullstelle in $z = 1$, wenn $\sigma(z) = \sum_{i=0}^{l} c_i (z-1)^i$. □

Beispiel 91 (Maximale Ordnung 2-Schritt-Verfahren)**.** *Ein 2-Schritt-Verfahren mit* $\rho(z) = (z-1)(z-\lambda)$ *ist stabil für* $\lambda \in [-1, 1)$*, und es ist*

$$\frac{\rho(z)}{\log z} = 1 - \lambda + \frac{3-\lambda}{2}(z-1) + \frac{5+\lambda}{12}(z-1)^2 + \frac{1+\lambda}{24}(z-1)^3 + O((z-1)^4).$$

$\Rightarrow$ *maximale Konsistenzordnung wird erreicht für* $\sigma(z) = 1 - \lambda + \frac{3-\lambda}{2}(z-1) + \frac{5+\lambda}{12}(z-1)^2 = -\frac{1+5\lambda}{12} + \frac{2-2\lambda}{3}z + \frac{5+\lambda}{12}z^2$*. Für* $\lambda \in (-1, 1)$ *ergibt sich 3. Ordnung, für* $\lambda = -1$ *das* Simpson-Regel-Verfahren *mit 4. Ordnung,*

$$y_{k+2} - y_k = \frac{h}{3}(f_{k+2} + 4f_{k+1} + f_k).$$

Theorem 92 (1. Dahlquistbarriere)**.** *Ein stabiles lineares n-Schritt-Verfahren hat höchstens Konsistenzordnung*

- $n + 2$*, falls* n *gerade,*
- $n + 1$*, falls* n *ungerade ist.*

Beweis.

- $T : \mathbb{C} \setminus \{-1\} \to \mathbb{C} \setminus \{1\}$, $T(\zeta) = \frac{\zeta-1}{\zeta+1}$ hat Inverse $T^{-1}(z) = \frac{1+z}{1-z}$ und bildet den komplexen Einheitsball $B = \{\zeta \in \mathbb{C} \,|\, |\zeta| \leq (<)1\}$ auf die linke Halbebene $H = \{z \in \mathbb{C} \,|\, \mathfrak{Re} z \leq (<)0\}$ ab.
- $r(z) = (\frac{1-z}{2})^n \rho(\frac{1+z}{1-z})$, $s(z) = (\frac{1-z}{2})^n \sigma(\frac{1+z}{1-z})$ erfüllen $r, s \in \mathcal{P}_n$
- Verfahren ist stabil $\overset{\rho(1)=0 \text{ für Konsistenz}}{\Rightarrow}$ $\zeta = 1$ ist einfache Nullstelle von ρ
  $\Rightarrow z = 0$ ist einfache Nullstelle von $r \Rightarrow r(z) = a_1 z + a_2 z^2 + \ldots + a_n z^n$, $a_1 \neq 0$; o.B.d.A. $a_1 > 0$
- Verfahren ist stabil $\Rightarrow$ Nullstellen von ρ liegen in $B \Rightarrow$ Nullstellen z_i von r liegen in H
  (sollte ρ eine Nullstelle in $\zeta = -1$ haben, geht diese in r verloren, d.h. eigentlich $r \in \mathcal{P}_{n-1}$)
  $\Rightarrow r(z) = Cz(z - z_1) \cdots (z - z_n) = a_1 z + a_2 z^2 + \ldots + a_n z^n$ hat $a_1, \ldots, a_n \geq 0$
  (Faktoren sind der Form $(z + p)$ oder $(z + p + qi)(z + p - qi) = (z^2 + 2pz + p^2 + q^2)$ für $p > 0$)
- Verfahren ist konsistent von Ordnung $p \geq n \Rightarrow \Phi$ hat p-fache Nullstelle in $\zeta = 1$
  $\Rightarrow q(z) = (\frac{1-z}{2})^n \Phi(\frac{1+z}{1-z}) = \frac{1}{\log \frac{1+z}{1-z}} r(z) - s(z)$ hat p-fache Nullstelle in $z = 0$
  $\Rightarrow$ erste p Terme der Taylorentwicklung $\frac{z}{\log \frac{1+z}{1-z}} \frac{r(z)}{z} = b_0 + b_1 z + b_2 z^2 + \ldots$ stimmen mit s überein
  $\Rightarrow$ da $p \geq n$ aber $s \in \mathcal{P}_n$ muss gelten $b_i = 0$ für $i = n + 1, \ldots, p - 1$
- Taylorentwicklung von $\frac{z}{\log \frac{1+z}{1-z}}$ ist von der Form $c_0 + c_2 z^2 + c_4 z^4 + \ldots$ mit $c_2, c_4, \ldots < 0$
  (Hausaufgabe; die Funktion ist gerade, somit sind $c_1 = c_3 = \ldots = 0$)
- Wir setzen $a_i = 0$ für $i > n$, dann gilt

$$\begin{aligned}
b_0 &= c_0 a_1 \\
b_1 &= c_0 a_2 \\
&\vdots \\
b_{2i} &= c_0 a_{2i+1} + c_2 a_{2i-1} + \ldots + c_{2i} a_1 \\
b_{2i+1} &= c_0 a_{2i+2} + c_2 a_{2i} + \ldots + c_{2i} a_2, \qquad i = 1, 2, \ldots
\end{aligned}$$

- Für $k > n$ gerade gilt $b_k = c_0 \underbrace{a_{k+1}}_{=0} + c_2 a_{k-1} + \ldots + c_k a_1 \leq c_k a_1 < 0$, insbesondere ist $b_k \neq 0$ $\lightning$
  $\Rightarrow \{n + 1, \ldots, p - 1\}$ darf keine gerade Zahl enthalten $\Rightarrow p \leq \begin{cases} n + 2, & n \text{ gerade}, \\ n + 1, & n \text{ ungerade}. \end{cases}$ $\square$

Beispiel 93 (Simpson-Regel-Verfahren)**.** *Das Simpson-Regel-Verfahren hat höchstmögliche Konsistenzordnung.*

4.4 Absolute Stabilität & steife Probleme

Auch wenn ein Verfahren stabil und konvergent ist, kann die numerische Lösung manchmal unbefriedigend sein, insbesondere wenn qualitative Eigenschaften der exakten Lösung erst für unverhältnismäßig kleine Schrittweiten h korrekt reproduziert werden. Ein besonders wichtiges Beispiel für eine solche Eigenschaft ist die Tatsache (A), dass die Lösung von $y'(x) = \lambda y(x)$, $\mathfrak{Re}\lambda < 0$, betragsmäßig mit der Zeit abnimmt. Das Konzept der absoluten Stabilität untersucht, wann ein Verfahren diese Eigenschaft korrekt reproduziert.

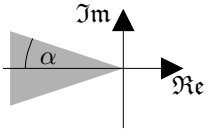
```
% löse y'(x)=-20y(x) mittels Euler-Verfahren
for h = [.01 .05 .1 0.2]
  x = 0:h:1;
  y = ones(size(x));
  for k = 2:length(x)
    y(k) = y(k-1) - h*20*y(k-1);
  end
  plot(x,y,'Linewidth',3); hold on; pause;
end
```

Ein Verfahren $y_{k+1} = y_k + h\varphi(x_k, y_0, \ldots, y_{k+1}, h)$ für $y'(x) = \lambda y(x)$ mit Schrittweite h kann auch aufgefasst werden als Verfahren für $y'(x) = \tilde{\lambda} y(x)$, $\tilde{\lambda} = \frac{\lambda}{\gamma}$, mit Schrittweite $\tilde{h} = h\gamma$; für das Eulerverfahren beispielsweise ist $y_{k+1} = y_k + h\lambda y_k$ äquivalent zu $y_{k+1} = y_k + \tilde{h}\tilde{\lambda} y_k$. Daher spielt bei der Analyse nur das Produkt $h\lambda$ eine Rolle.

Definition 94 (Absolute Stabilität)**.** *Das* Gebiet der absoluten Stabilität *eines numerischen Verfahrens ist*

$R_A = \{h\lambda \in \mathbb{C} \,|\, es\ gilt\ |y_{k+1}| < |y_k| \,\forall k\ für\ die\ Lösung\ y_k\ des\ Verfahrens\ zu\ y'(x) = \lambda y(x)\ mit\ Schrittweite\ h\}$.

Das Verfahren heißt A-stabil*, wenn* $R_A \supset \{z \in \mathbb{C} \,|\, \mathfrak{Re} z < 0\}$ *(d.h. (A) gilt für alle* h *&* $\mathfrak{Re}\lambda < 0$*) und* A(α)-stabil*, wenn* $R_A \supset \{z \in \mathbb{C} \,|\, \mathfrak{Re} z < 0, |\frac{\mathfrak{Im} z}{\mathfrak{Re} z}| \leq \tan\alpha\}$.



Warum ist A-Stabilität, also im Grunde eine Art Stabilität für alle h, wichtig? Sollte man nicht einfach h klein genug wählen, dass das Verfahren stabil ist? Dies verursacht bei steifen Problemen allerdings unverhältnismäßig großen Aufwand (genau dies soll *Steifigkeit* bedeuten, allerdings ist eine genaue mathematische Definition schwierig – wir geben eine beispielhafte).

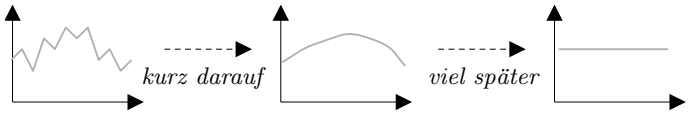
Definition 95 (Steifes Problem)**.** *Eine vektorwertige Differentialgleichung* $y'(x) = f(x, y(x))$ *heißt* steif*, wenn die Jakobi-Matrix* $J = D_y f$ *an einer Stelle* $(\hat{x}, \hat{y})$ *diagonalisierbar ist, ihre Eigenwerte* $\lambda_1, \ldots, \lambda_r$ $\mathfrak{Re}\lambda_i < 0$ *erfüllen und* $\mathfrak{Re}\lambda_j \ll \mathfrak{Re}\lambda_l$ *für mindestens zwei Indizes* j, l*.*

Bemerkung 96 (Steife Probleme)**.**

- *Nahe* $(\hat{x}, \hat{y})$ *verhält sich die Differentialgleichung etwa wie ihre Linearisierung* $y'(x) = f(\hat{x}, \hat{y}) + J(y(x) - \hat{y}) + \partial_x f(\hat{x}, \hat{y})(x - \hat{x})$ *bzw. nach der Variablentransformation* $\tilde{y}(x) = X^{-1}(y(\hat{x} + x) - \hat{y})$ *(wobei* $J = X\Lambda X^{-1}$ *die Diagonalisierung ist) wie* $\tilde{y}'(x) = b + cx + \Lambda\tilde{y}(x)$ *für* $b = X^{-1} f(\hat{x}, \hat{y})$*,* $c = X^{-1}\partial_x f(\hat{x}, \hat{y})$*. Daher reicht es, die Eigenschaften steifer Systeme an linearen diagonalen Differentialgleichungen zu untersuchen.*

- *Man betrachtet nur* $\mathfrak{Re}\lambda_i < 0$*, da für* $\mathfrak{Re}\lambda_i > 0$ *exponentielles Wachstum auftritt und die Schrittweite* h *für ausreichende Genauigkeit ohnehin sehr klein sein muss.*

- $\mathfrak{Re}\lambda_j \ll \mathfrak{Re}\lambda_l$ *bedeutet starke unterschiedliche Abklingzeiten. Dies taucht in vielen Systemen auf, z.B. werden bei der Wärmeleitung räumlich schnell oszillierende Inhomogenitäten schneller geglättet als langsam oszillierende.*



Beispiel 97 (Steife Probleme).

- $\binom{y_1}{y_2}'(x) = \left(\begin{smallmatrix} -100 & 0 \\ 1 & -2 \end{smallmatrix}\right)\binom{y_1}{y_2}(x)$, $\binom{y_1}{y_2}(0) = \binom{\epsilon}{1}$, *hat die Lösung* $\binom{y_1}{y_2}(x) = \begin{pmatrix} \epsilon e^{-100x} \\ e^{-2x} + \frac{\epsilon}{98}(e^{-2x} - e^{-100x}) \end{pmatrix}$, y_1 *klingt daher sehr viel schneller ab als* y_2. *Ist* $\epsilon = 10^{-6}$, *ist* y_1 *praktisch vernachlässigbar, und die Differentialgleichung für* y_2 *wird* $y_2' = -2y_2$.

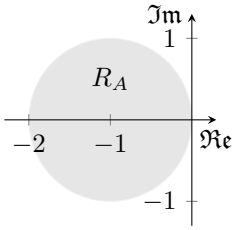
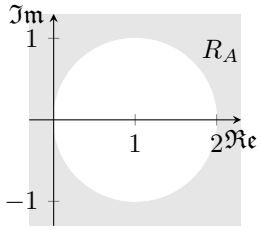
```
for h = [1e-4 1e-1] % für verschiedene Schrittweiten
  x = 0:h:1;
  n = length(x);
  y = zeros(2,n);   % Lösung des Systems
  y2 = zeros(1,n);  % Lösung für y2 unter Vernachlässigung von y1
  y(:,1) = [1e-6;1];
  y2(1) = 1;
  for k = 2:n       % Eulerverfahren
    y(:,k) = y(:,k-1) + h*[-100 0;1 -2]*y(:,k-1);
    y2(k) = y2(k-1) - h*2*y2(k-1);
  end
  plot(x,y2,'b','Linewidth',3); hold on; pause;
  plot(x,y,'r','Linewidth',3); hold off; pause;
end
```

- *Van der Pol-Oszillator* $y'' + \mu(y^2 - 1)y' + y = 0$, $y(0) = 2$, $y'(0) = 1 \Leftrightarrow \binom{y_1}{y_2}' = \begin{pmatrix} y_2 \\ \mu(1-y_1^2)y_2 - y_1 \end{pmatrix}$, $\binom{y_1}{y_2}(0) = \binom{2}{1}$

```
h = .0201021698315;
x = 0:h:60;
n = length(x);
y = zeros(2,n);
y(:,1) = [2;1];
for mu = [0 1 2 5 10 20 50]
  mu
  for k = 2:n
    y(:,k) = y(:,k-1) + h*[y(2,k-1);mu*(1-y(1,k-1)^2)*y(2,k-1)-y(1,k-1)];
  end
  plot(x,y,'Linewidth',3); pause;
  if mu == 20
    axis([50 52 2 3]); pause;
  end
end
```

Die Beispiele zeigen: In Bereichen, wo eine Komponente fast verschwindet, könnte man auch mit sehr großem h eine gute Genauigkeit erzielen – die Schritte in der verschwindenden Komponente führen dann jedoch zu Instabilität! Bei A-stabilen Verfahren gibt es dieses Problem nicht – das Verfahren ist stabil für alle h!

Beispiel 98 (Euler-Verfahren). *Für* $y' = \lambda y$ *liefert das Euler-Verfahren:*

explizit	*implizit*
$y_{k+1} = (1+h\lambda)y_k = (1+h\lambda)^{k+1}y_0$	$y_{k+1} = \frac{1}{1-h\lambda}y_k = \frac{1}{(1-h\lambda)^{k+1}}y_0$
$\Rightarrow R_A = \{z \in \mathbb{C} \,\vert\, \vert z+1\vert < 1\}$, *nicht A-stabil*	$\Rightarrow R_A = \{z \in \mathbb{C} \,\vert\, \vert 1-z\vert > 1\}$, *A-stabil*
	

Theorem 99 (Absolute Stabilität von Runge–Kutta-Verfahren)**.** *Betrachte ein n-stufiges Runge–Kutta-Verfahren angewandt auf $y' = \lambda y$.*

1. *Es ist $y_{k+1} = R(\lambda h)y_k$ für ein $R \in \mathcal{P}_n$ (explizites Verfahren) bzw. für ein $R = P/Q$ mit $P, Q \in \mathcal{P}_n$.*
2. *Ist das Verfahren konsistent der Ordnung p, gilt $\vert R(z) - e^z\vert = \vert R(z) - \sum_{i=0}^{\infty} \frac{z^i}{i!}\vert = O(z^{p+1})$.*
3. *$R_A = \{z \in \mathbb{C} \,\vert\, \vert R(z)\vert < 1\}$.*
4. *Kein explizites Runge–Kutta-Verfahren ist A-stabil oder A(0)-stabil.*

Beweis. 1. Hausaufgabe

2. Betrachte $y' = y$, $y(0) = 1$. Sei $y_{k+1} = y_k + h\varphi(y_k, y_{k+1}, h)$, φ Lipschitz in y_{k+1} mit Konstante L. Konsistenz von Ordnung $p \Rightarrow$

$$O(h^{p+1}) = h\tau_0 = y(h) - y(0) - h\varphi(x_0, y(0), y(h), h) - \underbrace{[y_1 - y_0 - h\varphi(x_0, y_0, y_1, h)]}_{=0}$$
$$= (y(h) - y_1)(1 + O(hL)) = (e^h - R(h))(1 + O(hL))$$

   Da R und exp um 0 herum holomorph sind, folgt die Behauptung für $z \in \mathbb{C}$.

3. trivial

4. Für jedes $R \in \mathcal{P}_n$ mit $n > 0$ gilt $\lim_{z\to-\infty} \vert R(z)\vert = \infty \Rightarrow \{z \in \mathbb{C} \,\vert\, \mathfrak{Re}z < 0,\ \mathfrak{Im}z = 0\} \not\subset R_A$. □

Für Mehrschrittverfahren angewandt auf $y' = \lambda y$ hängt die Monotonie der y_k von allen vorzugebenden Anfangswerten ab – daher wird die Definition der absoluten Stabilität hier oft angepasst. Wir wollen hier die strikte Monotonie $\vert y_{k+1}\vert < \vert y_k\vert$ erst für $k \geq K$ mit einem $K > 0$ fordern.

Theorem 100 (Absolute Stabilität linearer Mehrschrittverfahren)**.** *$\gamma \in \mathbb{C}$ liegt genau dann im absoluten Stabilitätsgebiet R_A eines linearen Mehrschrittverfahrens, wenn sein Stabilitätspolynom $\pi(z;\gamma) = \rho(z) - \gamma\sigma(z)$ nur Nullstellen $z \in \mathbb{C}$ mit $\vert z\vert < 1$ besitzt.*

Beweis.
- Anwendung auf $y' = \lambda y \Longrightarrow \sum_{i=0}^{n} \alpha_i y_{k+i} = h\sum_{i=0}^{n} \beta_i \lambda y_{k+i} \overset{\gamma=h\lambda}{\Longrightarrow} \sum_{i=0}^{n}(\alpha_i - \gamma\beta_i)y_{k+i} = 0$.
- Lösungen haben Form $y_k = \sum_{j=1}^{r} p_j(k)\lambda_j^k$ für Nullstellen $\lambda_1, \ldots, \lambda_r$ (mit Vielfachheiten $n_1, \ldots, n_r$) des charakteristischen Polynoms $\sum_{i=0}^{n}(\alpha_i - \gamma\beta_i)z^i = \rho(z) - \gamma\sigma(z)$ und Polynome $p_j \in \mathcal{P}_{n_j - 1}$.
- $\Rightarrow \vert y_k\vert$ strikt monoton fallend für k groß genug genau dann, wenn $\vert\lambda_j\vert < 1\ \forall j$. □

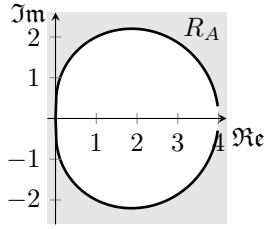
Folgendes ernüchterndes Resultat geben wir ohne Beweis an.

Theorem 101 (2. Dahlquistbarriere)**.**
- *Kein esplizites lineres Mehrschrittverfahren ist A(0)-stabil.*
- *A-stabile lineare Mehrschrittverfahren haben Konsistenzordnung ≤ 2.*
- *Das einzige A(0)-stabile lineare n-Schritt-Verfahren mit Konsistenzordnung $> n$ ist die Trapezregel $y_{k+1} - y_k = h\frac{f(x_k,y_k)+f(x_{k+1},y_{k+1})}{2}$.*

Beispiel 102 (2. Ordnung-Verfahren). *Die 2. Ordnung Rückwärts-Differenzen-Formel ist*

$$y_{k+2} - \frac{4}{3}y_{k+1} + \frac{1}{3}y_k = \frac{2}{3}hf(x_{k+2}, y_{k+2}).$$

$\pi(z;\gamma) = (1 - \frac{2}{3}\gamma)z^2 - \frac{4}{3}z + \frac{1}{3}$
$\Rightarrow$ *Nullstellen* $z_{1/2} = \frac{2 \pm \sqrt{1+2\gamma}}{3-2\gamma} = \frac{1}{2 \mp \sqrt{1+2\gamma}}$
$\Rightarrow R_A = \{\gamma \in \mathbb{C} \,|\, |2 \mp \sqrt{1+2\gamma}| > 1\}$
$\Rightarrow$ *A-stabil*



Wir wollen noch eine Methode angeben, wie man das Gebiet der absoluten Stabilität eines Mehrschrittverfahrens ermitteln kann.

Definition 103 (Schur-Polynom). $\Phi(z) = c_n z^n + \ldots + c_1 z + c_0$, $c_0 \neq 0$, $c_n \neq 0$, *heißt* Schur-Polynom, *wenn alle Nullstellen betragsmäßig kleiner* 1 *sind.*

Theorem 104 (Schur–Cohn-Test). $\Phi(z) = c_n z^n + \ldots + c_1 z + c_0$, $c_0 \neq 0$, $c_n \neq 0$ *ist ein Schurpolynom genau dann, wenn* $|\hat{\Phi}(0)| > |\Phi(0)|$ *und* Φ_1 *ein Schurpolynom ist für* $\hat{\Phi}(z) = \bar{c}_0 z^n + \ldots + \bar{c}_{n-1} z + \bar{c}_n$ *und* $\Phi_1(z) = \frac{1}{z}[\hat{\Phi}(0)\Phi(z) - \Phi(0)\hat{\Phi}(z)] \in \mathcal{P}_{n-1}$

Dieses Kriterium reduziert den Polynomgrad $\rightsquigarrow$ leichtere Nullstellenberechnung.

Beispiel 105 (Absolute Stabilität mittels Schur–Cohn-Test: $y_{k+2} - y_k = \frac{h}{2}(f(x_{k+1}, y_{k+1}) + 3f(x_k, y_k))$).

$$\pi(z;\gamma) = z^2 - \tfrac{\gamma}{2}z - (1 + \tfrac{3}{2}\gamma) \Rightarrow \begin{cases} \hat{\pi}(z;\gamma) & = -(1 + \frac{3}{2}\bar{\gamma})z^2 - \frac{\bar{\gamma}}{2}z + 1, \\ \pi_1(z;\gamma) & = \frac{1}{z}[z^2 - \frac{\gamma}{2}z - (1 + \frac{3}{2}\gamma) + (1 + \frac{3}{2}\gamma)(-(1 + \frac{3}{2}\bar{\gamma})z^2 - \frac{\bar{\gamma}}{2}z + 1)] \\ & = -(\frac{\gamma+\bar{\gamma}}{2} + 3|\frac{\gamma}{2}|^2)(3z + 1) \end{cases}$$

$\pi_1(z;\gamma)$ *ist Schur &* $|\hat{\pi}(0;\gamma)| = 1 > |1 + \frac{3}{2}\gamma| = |\pi(0;\gamma)| \Leftrightarrow \gamma \in B = \{\gamma \in \mathbb{C} \,|\, |\gamma - (-\frac{2}{3})| < \frac{2}{3}\}$, *d.h.* $R_A = B$

5 Interpolation und Quadratur

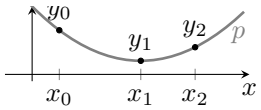
Den numerischen Verfahren für Anfangswertprobleme lagen numerische Interpolations-und Quadraturverfahren zugrunde, die wir nun vorstellen.

5.1 Grundlagen der Polynominterpolation

Interpolation ist nützlich, um Funktionen wie sin, cos, exp, ... anhand von vorberechneten Werten an bestimmten Stellen anzunähern (früher mithilfe von Wertetabellen, heute macht dies der Rechner intern) oder um eine unbekannte Funktion, von der nur einige Messwerte gegeben sind, zu approximieren. Wir beschreiben alles in $\mathbb{C}$, dies kann jedoch ohne Weiteres durch $\mathbb{R}$ ersetzt werden.

Definition 106 (Polynominterpolationsaufgabe). *Gegeben* Stützstellen $x_0, \ldots, x_n \in \mathbb{C}$ *und* Stützwerte $y_0, \ldots, y_n \in \mathbb{C}$ *ist das Interpolationsproblem:*

$$\textit{Finde } p \in \mathcal{P}_n \textit{ mit } p(x_k) = y_k,\ k = 0, \ldots, n. \tag{11}$$



Theorem 107 (Wohlgestelltheit). *Sind die Stützstellen paarweise verschieden, besitzt* (11) *eine eindeutige Lösung.*

Beweis.
- Existenz folgt aus Lagrange-Darstellung weiter unten.
- Eindeutigkeit: Seien $p, q \in \mathcal{P}_n$ Lösungen von (11)
  $\Rightarrow r := p - q \in \mathcal{P}_n$ mit $r(x_0) = \ldots = r(x_n) = 0$
  $\Rightarrow r \in \mathcal{P}_n$ hat mehr als n Nullstellen $\Rightarrow r = 0$. $\square$

Theorem 108 (Neville-Schema). *Sei p_{ik} das Interpolationspolynom zu x_j, y_j, $j = i, \dots, i+k$. Dann ist*

$$p_{i,k+1}(x) = \frac{1}{x_{i+k+1} - x_i}[(x_{i+k+1} - x)p_{ik}(x) + (x - x_i)p_{i+1,k}(x)], \quad i = 0, \dots, n-1, \quad k = 0, \dots, n-i-1,$$

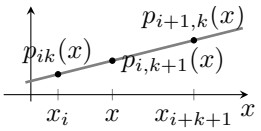
d.h. die Lösung p_{0n} von (11) kann rekursiv berechnet werden.

Beweis. $p_{i,k+1} \in \mathcal{P}_{k+1}$ & $p_{i,k+1}(x_j) = y_j$ für $j = i, \dots, i+k+1$ (durch Einsetzen) □

Die Berechnung nach Neville (oder Neville–Aitken) erfolgt nach dem Dreiecksschema

$$\begin{array}{llllll} x_0 & y_0{=}p_{00} & p_{01} & p_{02} & \ddots & \\ x_1 & y_1{=}p_{10} & p_{11} & \vdots & & p_{0n}. \\ x_2 & y_2{=}p_{20} & \vdots & & \ddots & \\ \vdots & \vdots & \vdots & p_{n-2,2} & & \\ x_n & y_n{=}p_{n0} & p_{n-1,1} & & & \end{array}$$

Bemerkung 109 (Polynomauswertung mit Neville–Aitken). *Man kann das gleiche Schema nutzen, um p_{0n} an einer Stelle x auszuwerten (ohne das Polynom explizit hinzuschreiben). Hierzu muss man in jedem Schritt nur eine Linearkombination mit Gewichten $w = \frac{x_{i+k+1}-x}{x_{i+k+1}-x_i}$ und $\tilde{w} = \frac{x-x_i}{x_{i+k+1}-x_i}$ bilden. Da $w + \tilde{w} = 1$, ist dies nichts anderes als eine lineare Inter-/Extrapolation:*



⇒ *Schema klappt nicht nur in $\mathbb{C}$, sondern in allen Räumen, wo „lineare Inter-/Extrapolation" definiert werden kann, z.B. auf Bildern (mit spezieller Inter-/Extrapolation):*

Beispiel 110 (3 Stützstellen, $x = 3$).

$$\begin{array}{ll} x_0 = 1 & y_0 = 0 \\ x_1 = 2 & y_1 = 0 \\ x_2 = 4 & y_2 = 5 \end{array} \quad \begin{array}{l} p_{01}(x) = \frac{-1}{1}0 + \frac{2}{1}0 = 0 \\ p_{11}(x) = \frac{1}{2}0 + \frac{1}{2}5 = \frac{5}{2} \end{array} \quad p_{02}(x) = \tfrac{1}{3}0 + \tfrac{2}{3}\tfrac{5}{2} = \tfrac{5}{3}$$

Sei $v_0, \dots, v_n \in \mathcal{P}_n$ eine Basis von $\mathcal{P}_n$. Die Lösung $p = \sum_{i=0}^n a_i v_i$ von (11) kann dann berechnet werden durch Lösen des linearen Gleichungssystems

$$\underbrace{\begin{pmatrix} v_0(x_0) & \cdots & v_n(x_0) \\ \vdots & & \vdots \\ v_0(x_n) & \cdots & v_n(x_n) \end{pmatrix}}_{=:V} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}. \tag{12}$$

Aus der Wohlgestelltheit von (11) folgt, dass (12) für alle rechten Seiten eine eindeutige Lösung hat ⇒ $V \in \mathbb{R}^{(n+1)\times(n+1)}$ ist regulär, unabhängig von der gewählten Basis.

Definition 111 (Darstellungsformen). *Für unterschiedliche Polynombasen $v_0, \dots, v_n$ erhalten wir unterschiedliche Darstellungen $p = \sum_{i=0}^n a_i v_i$ des Interpolationspolynoms:*

	Lagrange-Darstellung	Newton-Darstellung	Monom-Darstellung
Basis $v_i(x)$	$l_i(x) := \prod_{j=0, j\neq i}^n \frac{x-x_j}{x_i-x_j}$	$u_i(x) := \prod_{j=0}^{i-1}(x - x_j)$	$m_i(x) := x^i$
Systemmatrix V	I	$\begin{pmatrix} 1 & & & \\ 1 & x_1{-}x_0 & & \\ \vdots & \vdots & \ddots & \\ 1 & x_n{-}x_0 & \cdots & \prod_{j=0}^{n-1}(x_n - x_j) \end{pmatrix}$	$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}$ = *Vandermonde-Matrix*

Die Lagrange-Darstellung eignet sich gut für theoretische Zwecke, allerdings ändern sich alle Basispolynome, sobald eine neue Stützstelle hinzukommt. Die Newton-Darstellung ist so gewählt, dass nur neue Basispolynome hinzukommen und gleichzeitig V eine untere Dreiecksmatrix ist – somit ändern neue Stützstellen die alten Koeffizienten nicht. Um die Koeffizienten näher zu untersuchen, führen wir Folgendes ein.

Definition 112 (Dividierte Differenzen)**.** *Die* dividierten Differenzen *zu Stützstellen $x_0, \dots, x_n$ und -werten $y_0, \dots, y_n$ sind rekursiv definiert durch*

$$\begin{aligned} [y_i] &= y_i, && i = 0, \dots, n, \\ [y_i, \dots, y_{i+j}] &= \tfrac{[y_{i+1}, \dots, y_{i+j}] - [y_i, \dots, y_{i+j-1}]}{x_{i+j} - x_i}, && j = 1, \dots, n,\ i = 0, \dots, n-j. \end{aligned}$$

Sind $y_i = f(x_i)$ für ein $f : \mathbb{C} \to \mathbb{C}$, schreiben wir auch $f[x_i, \dots, x_{i+j}] = [y_i, \dots, y_{i+j}]$.

Die Berechnung geschieht im Dreiecksschema:

$$\begin{array}{llllll} x_0 & y_0{=}[y_0] & & & & \\ x_1 & y_1{=}[y_1] & [y_0,y_1] & & & \\ x_2 & y_2{=}[y_2] & [y_1,y_2] & [y_0,y_1,y_2] & \ddots & \\ \vdots & \vdots & \vdots & & \ddots & [y_0,\dots,y_n] \\ x_n & y_n{=}[y_n] & [y_{n-1},y_n] & \cdots & \cdots & \end{array}$$

Bei Hinzunahme neuer Stützstellen wird das Schema unten fortgesetzt.

Theorem 113 (Eigenschaften der Newton-Darstellung)**.** *Seien a_i die Koeffizienten der Newton-Darstellung.*

1. *$a_i = a_i(y_0, \dots, y_i)$ hängt nur von $y_0, \dots, y_i$ ab und ist der höchste Koeffizient des Interpolationspolynoms in Newton- und Monom-Darstellung zu den nur $i+1$ Stützstellen $x_0, \dots, x_i$.*
2. *$a_i = (w_0, \dots, w_i) \begin{pmatrix} y_0 \\ \vdots \\ y_i \end{pmatrix}$ für die* Knoten-Gewichte *$w_j = \prod_{k=0, k\neq j}^{i} \frac{1}{x_j - x_k}$.*
3. *$a_i = [y_0, \dots, y_i]$.*

Beweis. 1. $a_0, \dots, a_i$ lösen die ersten $i+1$ Zeilen von $V \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}$
$\Rightarrow$ dies sind die Gleichungen der Koeffizienten bei Interpolation mit nur $x_0, \dots, x_i$ & $y_0, \dots, y_i$.

2. Sei p_i Interpolationspolynom zu $x_0, \dots, x_i$ & $y_0, \dots, y_i$. Sein höchster Monom-Koeffizient ist wegen
$$p_i = \sum_{j=0}^{i} y_j l_j = \sum_{j=0}^{i} y_j \prod_{k=0, k\neq j}^{i} \frac{x - x_k}{x_j - x_k} = \sum_{j=0}^{i} y_j w_j \prod_{k=0, k\neq j}^{i} (x - x_k)$$
gleich $\sum_{j=0}^{i} y_j w_j$; der höchste Monom-Koeffizient ist aber auch der höchste Newton-Koeffizient a_i.

3. Sei $a_l(y_j, \dots, y_{j+l})$ der höchste Monom-/Newton-Koeffizient des Interpolationspolynoms $p_{j,l} \in \mathcal{P}_l$ durch (x_k, y_k), $k = j, \dots, j+l$. Zeige $a_l(y_j, \dots, y_{j+l}) = [y_j, \dots, y_{j+l}]$ per vollständiger Induktion in l.
$l = 0$: Trivial.
$l - 1 \rightsquigarrow l$: Neville-Schema $\Rightarrow p_{j,l}(x) = \frac{1}{x_{j+l} - x_j}[(x - x_j)p_{j+1,l-1}(x) - (x - x_{j+l})p_{j,l-1}(x)]$
$\Rightarrow$ höchster Monom-Koeffizient ist $\frac{1}{x_{i+l} - x_i}([y_{i+1}, \dots, y_{i+l}] - [y_i, \dots, y_{i+l-1}]) = [y_j, \dots, y_{j+l}]$. $\square$

Beispiel 114 ($(x_0, x_1, x_2) = (-1, 0, 2)$, $(y_0, y_1, y_2) = (1, 2, 3)$)**.**

- *Neville:* $\begin{smallmatrix} -1 & 1 & -x{\cdot}1{+}(x{+}1){\cdot}2{=}x{+}2 \\ 0 & 2 & \frac{2-x}{2}{\cdot}2{+}\frac{x}{2}{\cdot}3{=}\frac{x}{2}{+}2 \\ 2 & 3 & \end{smallmatrix}$ $\quad \frac{2-x}{3}(x+2) + \frac{x+1}{3}(\frac{x}{2}+2) = -\frac{x^2}{6} + \frac{5}{6}x + 2 = p(x)$
- *Lagrange:* $p(x) = 1 \cdot \frac{(x-0)(x-2)}{(-1-0)(-1-2)} + 2 \cdot \frac{(x+1)(x-2)}{(0+1)(0-2)} + 3 \cdot \frac{(x+1)(x-0)}{(2+1)(2+0)} = -\frac{x^2}{6} + \frac{5}{6}x + 2$
- *Newton:* $\begin{smallmatrix} -1 & 1 & \frac{2-1}{0-(-1)}{=}1 \\ 0 & 2 & \\ 2 & 3 & \frac{3-2}{2-0}{=}\frac{1}{2} \end{smallmatrix}$ $\quad \frac{1/2-1}{2-(-1)} = \frac{-1}{6} \quad \Rightarrow \quad p(x) = 1{+}1(x{-}(-1)){-}\frac{1}{6}(x{-}(-1))(x{-}0) = -\frac{x^2}{6}{+}\frac{5}{6}x{+}2$
- *Monom:* $V = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 2 & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = V^{-1} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -1/6 \\ 5/6 \\ 2 \end{pmatrix} \Rightarrow p(x) = -\frac{x^2}{6} + \frac{5}{6}x + 2$

5.2 Interpolationsfehler

Wenn eine Funktion f an Stützstellen $x_0, \dots, x_n$ ausgewertet und $f(x_0), \dots, f(x_n)$ durch ein Polynom p_n interpoliert werden, wie groß ist der Fehler $p_n - f$?

Definition 115 (B-Spline). *1. Die* abgeschnittene Potenz *ist definiert durch* $x_+^n := \begin{cases} x^n & \text{falls } x \geq 0, \\ 0 & \text{sonst.} \end{cases}$

2. Der B-Spline $(n-1)$*ten Grades zu Stützstellen* $x_0, \dots, x_n$ *ist* $B_{n-1}(t) := n \sum_{i=0}^{n} w_i (x_i - t)_+^{n-1}$ *(w_j die Knoten-Gewichte aus Theorem 113).*

Bemerkung 116 (B-Spline).

- *B-Spline steht für „basic spline“ (und kann auch zur Interpolation genutzt werden).*
- B_{n-1} *ist* $(n-2)$*mal differenzierbar.*
- *Sind* $x_0 < \ldots < x_n$*, kann man leicht zeigen* $B_{n-1} \geq 0$ *&* $B_{n-1} = 0$ *auf* $\mathbb{R} \setminus [x_0, x_n]$.

Theorem 117 (Peano-Darstellung dividierter Differenzen). *Ist* $f \in C^n([x_0, x_n])$*, gilt*

$$f[x_0, \dots, x_n] = \frac{1}{n!} \int_{x_0}^{x_n} f^{(n)}(t) B_{n-1}(t)\,\mathrm{d}t = \frac{f^{(n)}(\xi)}{n!} \int_{x_0}^{x_n} B_{n-1}(t)\,\mathrm{d}t = \frac{f^{(n)}(\xi)}{n!} \qquad \text{für ein } \xi \in [x_0, x_n].$$

Beweis.

- Taylor liefert $f(x) = \underbrace{\sum_{i=0}^{n-1} \frac{f^{(i)}(a)}{i!}(x-a)^i}_{=:p(x)} + \underbrace{\int_a^x \frac{(x-t)^{n-1}}{(n-1)!} f^{(n)}(t)\,\mathrm{d}t}_{=:r(x)}$.

  Bereits gezeigt: $f[x_0, \dots, x_n] = \sum_{i=0}^{n} w_i f(x_i)$, also $f[x_0, \dots, x_n] = \sum_{i=0}^{n} w_i p(x_i) + \sum_{i=0}^{n} w_i r(x_i)$.
  Da $p \in \mathcal{P}_{n-1}$, ist $\sum_{i=0}^{n} w_i p(x_i) = p[x_0, \dots, x_n] = (n\text{ter Koeffizient von } p) = 0$.
  $\Rightarrow f[x_0, \dots, x_n] = \sum_{i=0}^{n} w_i r(x_i) = \sum_{i=0}^{n} w_i \int_{x_0}^{x_i} \frac{(x_i - t)^{n-1}}{(n-1)!} f^{(n)}(t)\,\mathrm{d}t = \int_{x_0}^{x_n} \underbrace{\sum_{i=0}^{n} w_i \frac{(x_i - t)_+^{n-1}}{(n-1)!}}_{= B_{n-1}(t)/n!} f^{(n)}(t)\,\mathrm{d}t$

- Zweite Gleichung folgt aus erweitertem Mittelwertsatz.
- Für $q(x) = \prod_{i=0}^{n-1}(x - x_i)$ gilt $q[x_0, \dots, x_n] = \frac{1}{n!} \int_{x_0}^{x_n} B_{n-1}(t) q^{(n)}(t)\,\mathrm{d}t = \int_{x_0}^{x_n} B_{n-1}(t)\,\mathrm{d}t$.
  Außerdem ist die Newton-Darstellung der Interpolation von $q(x_0), \dots, q(x_n)$ gleich
  $q(x) = q[x_0] + q[x_0, x_1](x - x_0) + \ldots + q[x_0, \dots, x_n] \underbrace{(x - x_0) \cdots (x - x_{n-1})}_{= q(x)} \Rightarrow q[x_0, \dots, x_n] = 1$ □

Theorem 118 (Interpolationsfehler). *Für paarweise verschiedene Stützstellen* $x_0, \dots, x_n \in \mathbb{R}$ *sei* I *das kleinste Intervall mit* $x_0, \dots, x_n, x \in I$ *und das* Knotenpolynom *definiert durch* $\omega_{n+1}(x) = \prod_{k=0}^{n}(x - x_n)$. *Sei* p_n *das Interpolationspolynom von* $f(x_0), \dots, f(x_n)$*. Es gilt*

$$\begin{aligned} f(x) - p_n(x) &= f[x_0, \dots, x_n, x]\omega_{n+1}(x) \\ &= \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x) \qquad \text{für ein } \xi \in I, \qquad \text{falls } f \in C^{n+1}(I). \end{aligned}$$

Beweis. Sei $x_{n+1} := x$ und $p_i(t)$ das Interpolationspolynom zu $f(x_0), \dots, f(x_i)$, d.h. in Newton-Darstellung

$$p_i(t) = f[x_0] + f[x_0, x_1](t - x_0) + \ldots + f[x_0, \dots, x_i](t - x_0) \cdots (t - x_{i-1}).$$

Dann ist $f(x) - p_n(x) = p_{n+1}(x) - p_n(x) = f[x_0, \dots, x_n, x](x - x_0) \cdots (x - x_n)$. □

Beispiel 119.

- *Für* $f(x) = \cos(x)$ *ist* $|f^{(i)}(x)| \leq 1 \Rightarrow |f(x) - p_n(x)| \leq \frac{|\omega_{n+1}(x)|}{(n+1)!}$*. Zwischen den Stützstellen (dort wo* ω_{n+1} *klein ist) ist dies sehr klein! Außerhalb der Stützstellen wächst der Fehler wie die höchste Potenz von* ω_{n+1}*, also* x^{n+1}.

```
% Interpolation analytischer Funktion
n = 3;
x = linspace(-1,1,n+1); % Stuetzstellen
t = linspace(-2,2,100); % fuer Graph

p = polyfit(x,cos(x),n);
plot(t,cos(t),'k','Linewidth',3,t,polyval(p,t),'b','Linewidth',3);
```

- *Runge fand das Beispiel* $f(x) = \frac{1}{1+25x^2}$ *mit* $f^{(i)}$ *exponentiell wachsend (z.B.* $f^{(2n)}(0) = (-25)^n(2n)!$*)* $\Rightarrow$ *hier steigt der Fehler bei äquidistanten Stützstellen mit wachsendem* n *an; am Rand ergeben sich starke Oszillationen (*Runge-Phänomen*)!*

```
% Runges Beispiel
n = 21;
x = linspace(-1,1,n+1); % Stuetzstellen

p = polyfit(x,1./(1+25*x.^2),n);
plot(t,1./(1+25*t.^2),'k','Linewidth',3,t,polyval(p,t),'b','Linewidth',3);
axis([-1.5 1.5 -5 5]);
```

$\Rightarrow$ Nur für „nette“ Funktionen wird die Polynominterpolation für $n \to \infty$ konvergieren!
Ggfs. machen wir einen zusätzlichen Fehler bei der Messung/Auswertung der Stützwerte $f(x_0), \ldots, f(x_n)$.

Theorem 120 (Stabilität Polynominterpolation)**.** *Sei* $y_k = f(x_k) + \epsilon_k$*,* $k = 0, \ldots, n$*, mit* $|\epsilon_k| \leq \epsilon$*. Seien* $p, q \in \mathcal{P}_n$ *die Interpolationspolynome von* $f(x_0), \ldots, f(x_n)$ *bzw.* $y_0, \ldots, y_n$*. Es ist* $|p(x) - q(x)| \leq \epsilon L_n(x)$ *&* $\max_{x\in[a,b]} |p(x) - q(x)| \leq \epsilon \max_{x\in[a,b]} L_n(x)$ *für* $L_n(x) = \sum_{k=0}^{n} |l_k(x)|$*.*

Beweis. $p-q$ ist Interpolationspolynom zu $(x_0, \epsilon_0), \ldots, (x_n, \epsilon_n) \Rightarrow |p(x)-q(x)| = \left|\sum_{i=0}^{n} \epsilon_i l_i(x)\right| \leq \epsilon \sum_{k=0}^{n} |l_k(x)|$. □

Beispiel 121 (Interpolation mit Stützwert-Fehler)**.**
```
% Interpolation mit Stuetzwert-Fehler
n = 10;
x = linspace(-1,1,n+1); % Stuetzstellen
t = linspace(-1,1,100); % fuer Graph
y = cos(x) + .01 * randn(size(x));
p = polyfit(x,y,n);
plot(t,cos(t),'k','Linewidth',3,t,polyval(p,t),'b','Linewidth',3);

% Ln-Funktion
Ln = 0;
for i = 0:n
  y = zeros(1,n+1);
  y(i+1) = 1;
  l = polyfit(x,y,n);
  Ln = Ln + abs(polyval(l,t));
end
plot(t,Ln,'Linewidth',3);
```

5.3 Interpolationsversionen

Zusätzlich zu Funktionswerten kann man auch Ableitungen spezifizieren (sog. *Hermite-Interpolation*).

Theorem 122 (Hermite-Interpolation)**.** *Gegeben paarweise verschiedene Stützstellen* $x_0, \ldots, x_n$ *und Stützwerte* y_i^k*,* $i = 0, \ldots, n$*,* $k = 0, \ldots, n_i$*,* $N := \sum_{i=0}^{n} n_i + 1$*, gibt es ein eindeutiges Polynom* $p \in \mathcal{P}_{N-1}$ *mit* $p^{(k)}(x_i) = y_i^k \ \forall i, k$*.*

Beweis. Hausaufgabe □

Beispiel 123 (Hermite-Interpolation, $(x_0, x_1) = (0, 1)$, $y_0^0 = 0$, $y_0^1 = 1$, $y_1^0 = 0$)**.** *Sei* $p(x) = a_0 + a_1 x + a_2 x^2 \Rightarrow \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 0 & 1 & 2x_0 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \Rightarrow a_0 = 0, a_1 = 1, a_2 = -1$

Bemerkung 124 (Interpolationsfehler)**.** *Ähnlich zur normalen Interpolation findet man den Fehler* $f(x) - p_N(x) = \frac{f^{(N+1)}(\xi)}{(N+1)!}\omega_{N+1}(x)$ *mit* $\omega_{N+1}(x) = \prod_{i=0}^{n}(x - x_i)^{n_i+1}$.

Man kann Polynominter-/-extrapolation auch nutzen, um die Genauigkeit numerischer Simulationen zu erhöhen. Wollen wir z.B. eine gewöhnliche Differentialgleichung numerisch lösen, so müssen wir die Schrittweite h wählen. Wir können die Lösung $f(h)$ nur für $h > 0$ berechnen, die exakte Lösung wäre $\lim_{h\to 0} f(h)$. Angenommen, $f(h)$ hat die Taylorentwicklung $f(h) = f(0) + a_1 h + O(h^2)$ ($f(0)$ und a_1 unbekannt), dann kann man den $O(h)$-Term wie folgt eliminieren:

$$2f(h/2) - f(h) = f(0) + O(h^2)$$

$\Rightarrow$ man erhält eine bessere Fehlerordnung. Analog kann man höhere Ordnungen elimieren. Dieses Verfahren heißt *Richardson-Extrapolation.*

Algorithmus 1 Richardson-Extrapolation

Require: f habe in 0 eine Taylorentwicklung $f(x) = f(0) + \sum_{i=1}^{n} a_i x^i + o(x^n)$.
1: Wähle Stützstellen $x_i = z^i h$ für ein $z \in (0, 1)$, $h > 0$, $i = 0, \ldots, n$ (z.B. $x = h, \frac{h}{2}, \frac{h}{4}, \ldots$).
2: Berechne Polynominterpolation $p(x)$ zu $(x_i, f(x_i))$ & werte sie in $x = 0$ aus (mit Neville–Aitken).

Theorem 125 (Richardson-Extrapolation)**.** *Die Richardson-Extrapolation* $p(0)$ *erfüllt* $f(0)-p(0) = o(h^n)$.

Beweis.
- Peano-Darstellung des Fehlers: $f(0) - p(0) = f[x_0, \ldots, x_n, 0]\omega_{n+1}(0)$
- $\omega_{n+1}(0) = x_0 \cdots x_n = z^{n\frac{n+1}{2}} h^{n+1}$
- $f[x_0, \ldots, x_n, 0] = \frac{f[x_1,\ldots,x_n,0]-f[x_0,\ldots,x_n]}{0-x_0} = \frac{f^{(n)}(\xi_1)/n! - f^{(n)}(\xi_2)/n!}{-h}$ für $\xi_1, \xi_2 \in [0, h]$
  $= \frac{1}{n!h}(a_n + o(1) - a_n - o(1)) = \frac{o(1)}{n!h}$
- $\Rightarrow f(0) - p(0) = [z^{n\frac{n+1}{2}}/n!]o(h^n)$ □

Beispiel 126 ($g(z) = g(0) + a_1 z^2 + a_2 z^4 + O(z^6)$)**.** *Sezte* $x = z^2$, $f(x) = g(\sqrt{x}) = g(0) + a_1 x + a_2^2 + O(x^3)$.

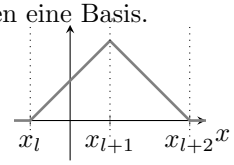
h	$f(h)$	$\frac{h/2}{-h/2}f(h) - \frac{h}{-h/2}f(h/2)$	$\frac{h/4}{-3h/4}(-f(h)+2f(h/2)) - \frac{h}{-3h/4}(-f(h/2)+2f(h/4)) = \frac{1}{3}f(h) - 2f(h/2) + \frac{8}{3}f(h/4) \approx g(0)$
$h/2$	$f(h/2)$	$\frac{h/4}{-h/4}f(h/2) - \frac{h/2}{-h/4}f(h/4)$	
$h/4$	$f(h/4)$		

Man kann auch mit anderen Funktionen interpolieren:

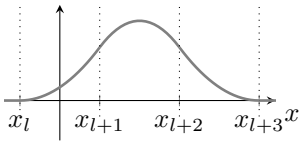
- *Rationale Funktionen* (etwa sog. *NURBS*). Üblicherweise sucht man zu Stützstellen $x_0, \ldots, x_n$ und -werten $y_0, \ldots, y_n$ zwei Polynome $p \in \mathcal{P}_m$, $q \in \mathcal{P}_{n-m}$ und löst dann das Gleichungssystem $y_i q(x_i) = p(x_i)$, $i = 0, \ldots, n$, $p(x_0) = 1$ für die $n + 2$ Koeffizienten von p und q.
- *Beliebige Ansatzfunktionen* $g_0, \ldots, g_n : \mathbb{C} \to \mathbb{C}$. Um eine Interpolierende $g(x) = \sum_{j=0}^{n} a_j g_j(x)$ zu erhalten, müssen die $n+1$ Gleichungen $a_0 g_0(x_j) + \ldots + a_n g_n(x_j) = y_j$, $j = 0, \ldots, n$, nach $a_0, \ldots, a_n$ gelöst werden.
- *Trigonometrische Funktionen* $g_j(x) = \exp(ijx)$. Dies ist die sog. *diskrete Fouriertransformation.*
- *Splines*, also stückweise Polynome vom Grad d, sodass die Gesamtfunktion in C^{d-1} liegt. Computerhardware kann auch Verzweigungen („if“) auswerten $\Rightarrow$ stückweise Definition ist effizient möglich. Hierdurch kann das Runge-Phänomen vermieden werden. Die B-Splines

$$B_{l,m}(x) = (m - l) \sum_{i=l}^{m} w_i (x_i - x)_+^{m-l-1} \qquad \text{mit } w_j = \prod_{\substack{k=l \\ k\neq j}}^{m} \frac{1}{x_j - x_k}$$

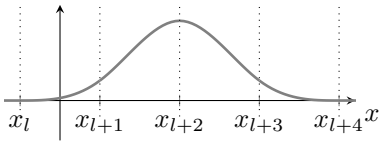
bilden eine Basis.



$m = l + 2$, linearer B-Spline



$m = l + 3$, quadratischer B-Spline



$m = l + 4$, kubischer B-Spline

5.4 Numerische Integration

Wie kann man näherungsweise das Volumen von Körpern anhand weniger Messwerte bestimmen? Allgemeiner, wie kann man das Integral einer Funktion anhand weniger Funktionswerte annähern? Dies bildet u.a. die Grundlage der numerischen Lösung von Differentialgleichungen.

Definition 127 (Quadratur)**.** *Sei $f : [a, b] \to \mathbb{R}$, $a \leq x_0 < \ldots < x_n \leq b$ Stützstellen. Eine Approximation*

$$I_n(f) = \sum_{i=0}^{n} a_i f(x_i)$$

an $I(f) = \int_a^b f(x)\,\mathrm{d}x$ heißt Quadraturformel *mit* Gewichten $a_0, \ldots, a_n$. *$|I_n(f) - I(f)|$ heißt* Quadraturfehler. *I_n heißt* exakt von Ordnung/Grad m, *falls $I_n(p) = I(p)\ \forall p \in \mathcal{P}_m$.*

Quadraturformeln erhält man z.B. mittels Polynominterpolation.

Definition 128 (Newton–Cotes-Formel)**.** *Sei $p_n^f(x) = \sum_{i=0}^n f(x_i) l_i(x)$ das Interpolationspolynom zu f in Lagrange-Darstellung. $I_n(f) := \int_a^b p_n^f(x)\,\mathrm{d}x = \sum_{i=0}^n a_i f(x_i)$ mit $a_i = \int_a^b l_i(x)\,\mathrm{d}x$ heißt* Newton–Cotes-Formel. *Ist $x_0 = a, x_n = b$ heißt sie* geschlossen, *im Fall $x_0 > a, x_n < b$* offen.

Bemerkung 129 (Newton–Cotes-Formel)**.**

- *Typischerweise verwendet man in der Newton–Cotes-Formel äquidistante Stützstellen.*
- *Offene Formeln eignen sich, wenn f einen Pol in a oder b hat.*
- *Die Newton–Cotes-Formel ist mindestens exakt vom Grad n, da $f \in \mathcal{P}_n \Rightarrow p_n^f = f$.*

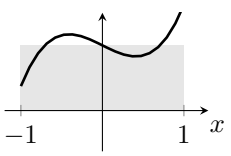
Bemerkung 130 (Quadraturgebiet)**.** *Wegen $\int_a^b f(s)\,\mathrm{d}s = \frac{b-a}{2} \int_{-1}^1 f(T(x))\,\mathrm{d}x$ mit $T(x) = a + (b-a)\frac{x+1}{2}$ reicht es/ist es üblich, die Stützstellen und Gewichte einer Quadraturformel nur für $[-1, 1]$ anzugeben.*

Beispiel 131 (Newton–Cotes-Formeln auf $[-1, 1]$)**.**

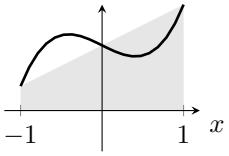
- $x_0 = 0$, $a_0 = \int_{-1}^1 l_0(x)\,\mathrm{d}x = \int_{-1}^1 1\,\mathrm{d}x = 2$
  $\Rightarrow I_0(f) = 2f(0)$ *heißt* Mittelpunktregel/Rechteckregel
- $x_0 = -1$, $x_1 = 1$, $a_0 = \int_{-1}^1 l_0(x)\,\mathrm{d}x = \int_{-1}^1 \frac{x-1}{-1-1}\,\mathrm{d}x = 1$, $a_1 = \int_{-1}^1 l_1(x)\,\mathrm{d}x = \int_{-1}^1 \frac{x+1}{1+1}\,\mathrm{d}x = 1$
  $\Rightarrow I_1(f) = f(-1) + f(1)$ *heißt* Trapezregel
- $-1 = x_0 < \ldots < x_n = 1$ *gleichverteilt:*

n	$(a_0\ \cdots\ a_n)$	*Name*	*Exaktheitsgrad*	$I_n(\cos\frac{\pi}{2}x)$
1	(1 1)	*Trapezregel*	1	0
2	(1 4 1)/3	*Simpson-/Keplersche Fassregel*	3	4/3
3	(1 3 3 1)/4	*Newtonsche $\frac{3}{8}$-Regel/Pulcherrima*	3	$1,2990\ldots$
5	(19 75 50 50 75 19)/144	*Milne-Regel*	5	$1,2727\ldots$
				$\approx \frac{4}{\pi} = 1,2732\ldots$

Für großes n können Gewichte negativ werden!



Mittelpunkt-/Rechteckregel



Trapezregel

Offenbar kann der Exaktheitsgrad von I_k auch größer als n sein – was weiß man darüber?

Theorem 132 (Maximaler Exaktheitsgrad)**.** *Sei I_n eine Quadraturformel zu Stützstellen $x_0, \dots, x_n$ mit Exaktheitsgrad m.*

1. *Ist $m \geq n$, so ist I_n die Newton–Cotes-Formel.*
2. *$m \leq 2n+1$.*

Beweis. 1. Das Lagrange-Polynom l_i hat Grad n, also $\int_a^b l_i \,\mathrm{d}x = I_n(l_i) = \sum_{j=0}^n a_j l_i(x_j) = a_i$.

2. Das quadrierte Knotenpolynom $(\omega_{n+1}(x))^2 = \prod_{i=0}^n (x-x_i)^2$ hat Grad $2n+2$ und erfüllt $I(\omega_{n+1}^2) - I_n(\omega_{n+1}^2) = \int_a^b \omega_{n+1}(x)^2 \,\mathrm{d}x - \sum_{i=0}^n a_i \omega_{n+1}(x_i)^2 = \int_a^b \omega_{n+1}(x)^2 \,\mathrm{d}x > 0$. □

Können wir eine Newton–Cotes-Formel I_n (d.h. Stützstellen) von Exaktheitsgrad $2n+1$ finden? Sei $p \in \mathcal{P}_{2n+1}$ und $\omega_{n+1}(x) = \prod_{i=0}^n (x - x_i)$ das Knotenpolynom. Mittels Euklidischer Division gilt

$$p = q_n \omega_{n+1} + r_n \qquad \text{mit } q_n, r_n \in \mathcal{P}_n.$$

Der Quadraturfehler $R_n(f) = I_n(f) - I(f)$ erfüllt

$$R_n(p) = R_n(q_n\omega_{n+1}) + R_n(r_n) = R_n(q_n\omega_{n+1}) = \sum_{i=0}^n a_i q_n(x_i)\omega_{n+1}(x_i) - \int_a^b q_n\omega_{n+1} \,\mathrm{d}x = -\int_a^b q_n\omega_{n+1} \,\mathrm{d}x.$$

Theorem 133 (L^2-Skalarprodukt)**.** *Die Abbildung $p, q \mapsto \langle p, q\rangle = \int_a^b p(x)q(x) \,\mathrm{d}x$ definiert ein Skalarprodukt auf dem Vektorraum $C^0([a,b])$ der stetigen Funktionen auf $[a,b]$.*

Wenn das Knotenpolynom ω_{n+1} orthogonal zu allen $q_n \in \mathcal{P}_n$ bzgl. des Skalarprodukts ist, ist $R_n(p) = 0$ $\forall p \in \mathcal{P}_{2n+1}$ und somit I_n exakt von Ordnung $2n+1$. Ein solches Polynom ω_{n+1} erhält man durch Gram–Schmidt-Orthogonalisierung der $m_i(x) = x^i$, $i = 0, 1, \dots$, d.h. $\tilde{\omega}_i = m_i - \sum_{j=0}^{i-1} \langle m_i, \omega_j\rangle \omega_j$, $\omega_i = \frac{\tilde{\omega}_i}{\langle \tilde{\omega}_i, \tilde{\omega}_i\rangle^{1/2}}$.

Theorem 134 (Legendre-Polynome)**.** *Sei $[a,b] = [-1,1]$. Das Gram–Schmidt-Verfahren auf $m_0, m_1, \dots$ liefert die* Legendre-Polynome $\omega_i(x) = \frac{i!}{(2i)!} \frac{\mathrm{d}^i}{\mathrm{d}x^i}[(x^2-1)^i]$.

Beweis.
- $\omega_i \in \mathcal{P}_i = \operatorname{span}\{m_0, \dots, m_i\}$
- Definiere $\Omega_i^0 = \omega_i$, $\Omega_i^j(x) = \int_{-1}^x \Omega_i^{j-1}(t) \,\mathrm{d}t = \frac{i!}{(2i)!} \frac{\mathrm{d}^{i-j}}{\mathrm{d}x^{i-j}}[(x^2-1)^i]$ für $j = 1, \dots, i$,
  $\Rightarrow \Omega_i^j$ hat in 1 und -1 eine j-fache Nullstelle.
- $p \in \mathcal{P}_{i-1} \qquad \Rightarrow \qquad \langle \omega_i, p\rangle = \int_{-1}^1 \omega_i(x)p(x) \,\mathrm{d}x = [\Omega_i^1(x)p(x)]_{-1}^1 - \int_{-1}^1 \Omega_i^1(x)p^{(1)}(x) \,\mathrm{d}x$
  $= -\int_{-1}^1 \Omega_i^1(x)p^{(1)}(x) \,\mathrm{d}x = \dots = (-1)^i \int_{-1}^1 \Omega_i^i(x)p^{(i)}(x) \,\mathrm{d}x = 0$ □

Wählen wir also Stützstellen $x_0, \dots, x_n$, sodass das Knotenpolynom ω_{n+1} gleich dem $(n+1)$ten Legendre-Polynom ist, dann erhalten wir eine Quadraturformel mit Exaktheitsgrad $2n+1$. Allerdings ist das Knotenpolynom nur zulässig, wenn die $x_0, \dots, x_n$ paarweise verschieden sind. Dies ist der Fall.

Theorem 135 (Nullstellen orthogonaler Polynome)**.** *Das Legendre-Polynom ω_i hat genau i verschiedene Nullstellen in $(-1,1)$.*

Beweis. Angenommen, ω_i habe weniger Nullstellen
$\Rightarrow \omega_i$ ändert Vorzeichen nur an $m < i$ Stellen $x_0, \dots, x_{m-1}$
$\Rightarrow q\omega_i$ hat konstantes Vorzeichen für $q(x) = \prod_{j=0}^{m-1}(x - x_j) \in \mathcal{P}_m$
$\Rightarrow \langle q, \omega_i\rangle = \int_{-1}^1 q\omega_i \,\mathrm{d}x > 0$ (außer $\omega_i = 0$) ↯ □

Definition 136 (Gauß-Quadratur). *Die Quadraturformel von Exaktheitsgrad $2n+1$, deren Stützstellen die Nullstellen des $(n+1)$ten Legendre-Polynoms sind, heißt* Gauß-Quadratur.

Die Gauß-Quadratur hat außerdem positive Gewichte, was wichtig für numerische Stabilität ist.

Theorem 137 (Gauß-Gewichte). *Die Gewichte a_i der Gauß-Quadratur auf $[-1,1]$ erfüllen*

$$a_i = \int_{-1}^{1} (l_i(x))^2 \,\mathrm{d}x > 0.$$

Beweis. $l_i^2 \in \mathcal{P}_{2n} \Rightarrow$ Gauß-Quadratur ist exakt mit $\int_{-1}^{1}(l_i(x))^2\,\mathrm{d}x = \sum_{j=0}^{n} a_j l_i(x_j)^2 = a_i$. □

Beispiel 138 (Gauß-Quadratur-Formeln).

n	0	1	2
$(x_0,\dots,x_n)$	(0)	$(-\sqrt{3}/3\ \sqrt{3}/3)$	$(-\sqrt{3/5}\ 0\ \sqrt{3/5})$

Schließlich möchten wir wissen, wie groß der Quadraturfehler ist.

Theorem 139 (Quadraturfehler). *Die Newton–Cotes-Formel mit Stützstellen $x_0,\dots,x_n \in [a,b]$ erfüllt*

$$|I_n(f) - I(f)| \le \left| \int_a^b f[x_0,\dots,x_n,x]\omega_{n+1}(x)\,\mathrm{d}x \right| \le \|f[x_0,\dots,x_n,x]\|_\infty \int_a^b |\omega_{n+1}(x)|\,\mathrm{d}x$$
$$\le \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \int_a^b |\omega_{n+1}(x)|\,\mathrm{d}x \qquad \textit{falls } f \in C^{n+1}([a,b]).$$

Beweis. Folgt aus $I_n(f) - I(f) = \int_a^b p_n^f - f\,\mathrm{d}x$ für das Interpolationspolynom p_n^f und der Darstellung des Interpolationsfehlers. □

Für Formeln mit Exaktheitsgrad $> n$ erwartet man noch kleinere Fehler. Um dies zu sehen, benutzen wir eine Darstellung ähnlich zu der Peano-Darstellung dividierter Differenzen.

Definition 140 (Monospline). *Seien $a_0,\dots,a_n$ die Gewichte zu Stützstellen $x_0,\dots,x_n \in [a,b]$ einer Quadratur mit Exaktheitsgrad $\ge k$. $M_{k,n}(x) = (b-x)_+^{k+1} - (a-x)_+^{k+1} - (k+1)\sum_{i=0}^{n} a_i (x_i - x)_+^k$ heißt k-ter Monospline.*

Theorem 141 (Monospline).
1. *$M_{k,n} = 0$ auf $\mathbb{R}\setminus[a,b]$*
2. *Für $x_0 > a$ hat $M_{k,n}$ in a eine $(k+1)$-fache Nullstelle, für $x_0 = a$ eine k-fache.*
3. *Analoges gilt für b.*

Beweis. 1. Für $x > b$ ist $M_{k,n}(x) = 0$, für $x < a$ ist

$$M_{k,n}(x) = (b-x)^{k+1} - (a-x)^{k+1} - (k+1)\sum_{i=0}^{n} a_i (x_i - x)^k = (b-x)^{k+1} - (a-x)^{k+1} - (k+1)\int_a^b (t-x)^k \mathrm{d}t = 0$$

2. Auf (a, x_0) ist $M_{k,n}(x) = (a-x)^{k+1}$ (ähnlich für $x_0 = a$).
3. Auf (x_n, b) ist $M_{k,n}(x) = (b-x)^{k+1}$ (ähnlich für $x_n = b$). □

Theorem 142 (Peanodarstellung des Quadraturfehlers). *Sei I_n eine Quadraturformel mit Stützstellen $x_0,\dots,x_n \in [a,b]$ und Exaktheitsgrad $\ge k$, so ist der Quadraturfehler für $f \in C^{k+1}([a,b])$ gegeben durch $I_n(f) - I(f) = -\int_a^b f^{(k+1)}(t)\frac{M_{k,n}(t)}{(k+1)!}\,\mathrm{d}t$.*

Beweis. Partielle Integration liefert für $x \in [a,b]$

$$\int_a^b f^{(k+1)}(t)(x-t)_+^m \,\mathrm{d}t = [f^{(k)}(t)(x-t)^m]_a^x + m\int_a^x f^{(k)}(t)(x-t)^{m-1}\,\mathrm{d}t = -f^{(k)}(a)(x-a)^m + m\int_a^b f^{(k)}(t)(x-t)_+^{m-1}\,\mathrm{d}t,$$

somit
$$\begin{aligned}\int_a^b f^{(k+1)}(t)\frac{M_{k,n}(t)}{(k+1)!}\,\mathrm{d}t &= -f^{(k)}(a)\frac{M_{k,n}(a)}{(k+1)!} + \int_a^b f^{(k)}\frac{M_{k-1,n}(t)}{k!}\,\mathrm{d}t\\ &= \int_a^b f^{(k)}(t)\frac{M_{k-1,n}(t)}{k!}\,\mathrm{d}t = \ldots = \int_a^b f'(t)\frac{M_{0,n}(t)}{1!}\,\mathrm{d}t\\ &= \int_a^b f'(t)(b-t) - \sum_{i=0}^n a_i f'(t)(x_i-t)_+^0\,\mathrm{d}t\\ &= [f(t)(b-t)]_a^b + \int_a^b f(t)\,\mathrm{d}t - \sum_{i=0}^n a_i(f(x_i)-f(a))\\ &= f(a)\underbrace{\left[a-b+\sum_{i=0}^n a_i\right]}_{=I_1(1)-I(1)=0} + I(f) - I_n(f)\end{aligned}$$ □

Beispiel 143 (Quadraturfehler)**.**

- *Mittelpunktregel auf* $[-1,1]$*:* $n=0$*,* $k=1$*,* $M_{k,n}(t) = (1-t)_+^2 - (-1-t)_+^2 - 2\cdot 2\cdot(0-t)_+^1$
  $\Rightarrow$ *Ist* $f\in C^2([-1,1])$*,* $\exists\,\xi\in[-1,1]$*:* $I(f)-I_0(f) = \int_{-1}^1 f''(t)\frac{M_{1,0}(t)}{2}\,\mathrm{d}t = \frac{f''(\xi)}{2}\int_{-1}^1 M_{1,0}(t)\,\mathrm{d}t = \frac{f''(\xi)}{3}$*.*
- *Trapezregel auf* $[-1,1]$*:* $k=n=1$*,* $M_{1,1}(t) = (1-t)_+^2 - (-1-t)_+^2 - 2[(-1-t)_+ + (1-t)_+] = \min\{0, t^2-1\}$
  $\Rightarrow$ *Für* $f\in C^2([-1,1])$ *gilt* $I(f)-I_1(f) = \int_{-1}^1 f''(t)\frac{t^2-1}{2}\,\mathrm{d}t$*.*
- *Man kann zeigen, dass* $M_{2n+1,n}\geq 0$ *(Gauß-Quadratur) und* $\int_{-1}^1 M_{2n+1,n}(t)\,\mathrm{d}t = \frac{2^{2n+3}}{2n+3}\frac{[(n+1)!]^4}{[(2n+2)!]^2}$
  $\Rightarrow I(f)-I_n(f) = \int_{-1}^1 f^{(2n+2)}(t)\frac{M_{2n+2,n}(t)}{(2n+2)!}\,\mathrm{d}t = \frac{2^{2n+3}}{2n+3}\frac{[(n+1)!]^4}{[(2n+2)!]^3}f^{(2n+2)}(\xi)$ *für ein* $\xi\in[-1,1]$*.*

Wie schon bei der Interpolation ist es oft genauer (z.B. wenn die zu integrierende Funktion nicht oft differenzierbar ist oder die Ableitungen stark anwachsen, sodass Runges Phänomen auftritt), statt Polynomen stückweise Polynome für die Quadratur zu nutzen. Hierzu teilt man $[a,b]$ in N kleine Teilintervalle auf und approximiert auf jedem das Integral mit Quadratur. Sei z.B. $I_n(f) = \sum_{i=0}^n a_i f(x_i)$ eine Quadraturformel auf $[-1,1]$, dann ist $I_n^{cd}(f) = \frac{d-c}{2}\sum_{i=0}^n a_i f(c+\frac{x_i+1}{2}(d-c))$ eine Quadratur auf $[c,d]$. Setzen wir nun

$$z_i = a+hi \qquad \text{für } h=\frac{b-a}{N},\ i=0,\ldots,N,\ x_i^j = z_i + \frac{h}{2}(x_j+1),\ j=0,\ldots,n,$$

so ist $I(f) = \int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{2}\sum_{i=0}^{N-1}\sum_{j=0}^n a_j f(x_i^j) =: I_n^N(f)$.

Beispiel 144 (Trapezregel)**.** *Für die Trapezregel ist* $x_i^0 = z_i = a+hi$*,* $x_i^1 = z_{i+1} = a+h(i+1)$*,* $a_0 = a_1 = 1$
$\Rightarrow I_1^N(f) = \frac{h}{2}[f(x_0^0)+f(x_0^1)+f(x_1^0)+f(x_1^1)+\ldots+f(x_{N-1}^0)+f(x_{N-1}^1)] = \frac{(b-a)}{N}[\frac{f(a)+f(b)}{2}+\sum_{i=1}^{N-1} f(z_i)]$

Theorem 145 (Fehler summierter Formeln)**.** *Sei* I_n *eine Quadraturformel auf* $[-1,1]$ *mit Exaktheitsgrad* k*,* $f\in C^{k+1}([a,b])$*. Für den Fehler der summierten Quadraturformel* I_n^N *gilt*

$$|I_n^N(f) - I(f)| \leq \max_{x\in[a,b]}|f^{(k+1)}(x)|\frac{(h/2)^{k+1}}{2(k+1)!}\int_{-1}^1 |M_{k,n}(t)|\,\mathrm{d}t$$

für $h=\frac{b-a}{N}$ *und* $M_{k,n}$ *den Monospline zu* I_n*.*

Beweis. Sei $I_n(f) = \sum_{i=0}^n a_i f(x_i)$ und sei $z_i = a+h_i$ und $M_{k,n}^i$ der Monospline zur Quadraturformel auf $[z_i, z_{i+1}]$, also $M_{k,n}^i(t) = (z_{i+1}-t)_+^{k+1} - (z_i-t)_+^{k+1} - (k+1)\sum_{i=0}^n a_i\frac{h}{2}(z_i+\frac{h}{2}(x_i+1)-t)_+^k$. Es ist

$$\begin{aligned}|I(f)-I_n^N(f)| &= \left|\sum_{i=0}^{N-1} I(f|_{[z_i,z_{i+1}]}) - I_n(f|_{[z_i,z_{i+1}]})\right| = \left|\sum_{i=0}^{N-1}\int_{z_i}^{z_{i+1}}\frac{f^{(k+1)}(t)}{(k+1)!}M_{k,n}^i(t)\,\mathrm{d}t\right|\\ &\leq \max_{x\in[a,b]}\frac{|f^{(k+1)}(x)|}{(k+1)!}\sum_{i=0}^{N-1}\int_{z_i}^{z_{i+1}}|M_{k,n}^i(t)|\,\mathrm{d}t.\end{aligned}$$

Mit der Variablentransformation $t = z_i + \frac{h}{2}(x+1)$ folgt $M_{k,n}^i(t) = (\frac{h}{2})^{k+1}M_{k,n}(x)$
$\Longrightarrow \sum_{i=0}^N\int_{z_i}^{z_{i+1}}|M_{k,n}^i(t)|\,\mathrm{d}t = \sum_{i=0}^N\int_{-1}^1(\frac{h}{2})^{k+1}|M_{k,n}(x)|\frac{h}{2}\,\mathrm{d}x = (\frac{h}{2})^{k+1}\frac{1}{2}\int_{-1}^1|M_{k,n}(x)|\,\mathrm{d}x.$ □

Der Fehler ist offenbar eine Potenz von h (bzw. hat vielleicht sogar eine höhere Taylorentwicklung in h), somit kann der Fehler mit Richardson–Extrapolation verkleinert werden. Dies ist die Idee in Folgendem.

Algorithmus 2 Romberg-Quadratur

Require: Sei $f \in C^{2m+2}([a,b])$ und I_1^N für $N = \frac{b-a}{h}$ die summierte Trapezregel, d.h.

$$I_1^N(f) = h\left[\frac{f(a)+f(b)}{2} + \sum_{i=1}^{N-1} f(a+ih)\right] =: g(z) \qquad \text{für } z = h^2.$$

1: Berechne $g(z), g(\frac{z}{4}), \ldots, g(\frac{z}{4^m})$ (also die Quadratur zu $h, \frac{h}{2}, \ldots, \frac{h}{2^m}$).
2: Approximiere $g(0) = I(f)$ mittels Richardson-Extrapolation $\tilde{g}(0)$.

Bemerkung 146 (Romberg-Quadratur)**.** *Insgesamt muss f nur an $2^m \frac{b-a}{h}$ Stellen ausgewertet werden (den Stützstellen zu $h/2^m$), da die Stützstellen zu $h, \frac{h}{2}, \ldots, \frac{h}{2^{m-1}}$ eine Teilmenge hiervon bilden.*

Theorem 147 (Romberg-Quadratur)**.** *Für die Romberg-Quadratur gilt $|\tilde{g}(0) - I(f)| = O(h^{2m+2})$.*

Beweis. Folgt aus Fehlerabschätzung für Richardson-Extrapolation, wenn wir zeigen können, dass $g(z)$ eine Taylorentwicklung $g(0) + \sum_{i=1}^m a_i z^i + O(z^{m+1})$ hat, also

$$I_1^{\frac{b-a}{h}}(f) = I(f) + \sum_{i=1}^m a_i h^{2i} + O(h^{2m+2}).$$

Dies trifft zu, da

$$\begin{aligned} I(f) - I_1^N(f) &= \sum_{i=0}^{N-1} \int_{z_i}^{z_{i+1}} f''(x) \frac{M_{1,1}^i(x)}{2} \,\mathrm{d}x = \sum_{i=0}^{N-1} \int_{z_i}^{z_{i+1}} f''(x) \frac{(x-z_i)(x-z_{i+1})}{2} \,\mathrm{d}x \\ &= \sum_{i=0}^{N-1} \int_{z_i}^{z_{i+1}} f''(x) \frac{(x-z_i)(x-z_{i+1}) + h^2/6}{2} \,\mathrm{d}x - \frac{h^2}{12} \int_a^b f''(x) \,\mathrm{d}x \\ &= \sum_{i=0}^{N-1} \int_{-h/2}^{h/2} f''\left(x + \frac{z_i + z_{i+1}}{2}\right) \frac{x^2 - h^2/12}{2} \,\mathrm{d}x - \frac{h^2}{12}[f'(b) - f'(a)]. \end{aligned}$$

Sei nun $p_2(x) = \frac{x^2 - h^2/12}{2}$, $p_{i+1}'(x) = p_i(x)$ mit $\int_{-h/2}^{h/2} p_{i+1}(x) \,\mathrm{d}x = 0$.
Offenbar ist p_{2i} gerade ($p_{2i}(x) = p_{2i}(-x)$) und p_{2i+1} ungerade ($p_{2i+1}(x) = -p_{2i+1}(-x)$) $\forall i$, außerdem $p_k(\frac{h}{2}) - p_k(-\frac{h}{2}) = \int_{-h/2}^{h/2} p_k'(x) \,\mathrm{d}x = 0 \; \forall k$.
Mittels $(2m)$-facher partieller Integration folgt

$$\begin{aligned} I(f) - I_1^N(f) =& \sum_{i=0}^{N-1} [f''(x + \tfrac{z_i+z_{i+1}}{2}) p_3(x) - f'''(x + \tfrac{z_i+z_{i+1}}{2}) p_4(x) + \ldots - f^{(2m+1)}(x + \tfrac{z_i+z_{i+1}}{2}) p_{2m+2}(x)]_{x=-h/2}^{h/2} \\ &+ \int_{-h/2}^{h/2} f^{(2m+2)}(x + \tfrac{z_i+z_{i+1}}{2}) p_{2m+2}(x) \,\mathrm{d}x - \tfrac{h^2}{12}[f'(b) - f'(a)] \\ =& \sum_{j=0}^{m} [f^{(2j+1)}(a) p_{2j+2}(-\tfrac{h}{2}) - f^{(2j+1)}(b) p_{2j+2}(\tfrac{h}{2})] + \sum_{i=0}^{N-1} \int_{-h/2}^{h/2} f^{(2m+2)}(x + \tfrac{z_i+z_{i+1}}{2}) p_{2m+2}(x) \,\mathrm{d}x \\ =& \, c_1 h^2 + c_2 h^4 + \ldots + c_m h^{2m} + O(h^{2m+2}). \end{aligned}$$

□

Bemerkung 148 (Periodische Romberg-Quadratur)**.** *Ist f periodisch, d.h. $f^{(k)}(a) = f^{(k)}(b)$, $k = 0, \ldots, 2m+2$, dann zeigt obige Rechnung wegen $p_{2j}(-h) = p_{2j}(h)$ sogar $I(f) - I_1^N(f) = O(h^{2m+2})$. Ist f analytisch, konvergiert $I_1^N(f)$ also sogar schneller gegen $I(f)$ als jede Potenz von h!*

Wie bei numerischen Verfahren für gewöhnliche Differentialgleichungen nennt man *Konsistenz*, dass der Diskretisierungsfehler (durch die Wahl ausreichend vieler Stützstellen) beliebig klein gemacht werden kann (z.B. ist Polynominterpolation mit äquidistanten Stützstellen konsistent für analytische Funktionen, nicht jedoch für beliebige wegen des Runge-Phänomens), und *Stabilität*, dass die akkumulierten Rundungsfehler nicht größer werden als der unvermeidbare Fehler, den man bei exakter Rechnung macht, wenn die Eingabedaten in gleichem Maße gestört werden.

Theorem 149 (Stabilität)**.** *Konsistente Quadraturformeln mit positiven Gewichten sind stabil.*

Beweis. Sei $\tilde{f}$ eine Näherung an f mit $|f - \tilde{f}| \leq \epsilon$, die Quadraturformel I_n habe Stützstellen $x_0, \ldots, x_n$ und Gewichte $a_0, \ldots, a_n > 0$. Wir müssen zeigen $|I_0(\tilde{f}) - I_n(f)| \lesssim \epsilon$: $|I_n(\tilde{f}) - I_n(f)| = |\sum_{i=0}^{n} a_i(\tilde{f}(x_i) - f(x_i))| \leq \epsilon \sum_{i=0}^{n} |a_i||f(x_i)| = \epsilon I_n(|f|) \overset{\text{Konsistenz}}{\approx} \epsilon I(|f|)$. □

Beispiel 150 (Unterschiedliche Quadraturformeln)**.**

```
function exampleQuadrature
  % Integranden
  f = cell(3);
  f{1} = @(x) 1./(1+25*x.^2);
  f{2} = @(x) exp(x);
  f{3} = @(x) sin(x*pi+1)-cos(2*x*pi);
  % Integral auf [-1,1]
  val = [2*atan(5)/5 exp(1)-exp(-1) 0];

  % geschlossene Newton-Cotes-Quadratur
  n = 50;
  error = zeros(3,n);
  for i = 1:n
    % Stuetzstellen
    x = linspace(-1,1,i+1);
    % Gewichte
    weights = computeQuadratureWeights( x, -1, 1 );
    % Quadratur
    for j = 1:3
      error(j,i) = abs(weights*f{j}(x')-val(j));
    end
  end
  subplot(1,3,1);
  semilogy(1:n,error,'Linewidth',3);

  % Gauss-Quadratur
  error = zeros(3,n);
  for j = 2:n
    % Legendre-Polynom-Nullstellen
    syms x;
    roots = vpasolve( legendreP(j,x) == 0 );
    % Gewichte
    weights = computeQuadratureWeights( roots, -1, 1 );
    % Quadratur
    for i = 1:3
      error(i,j) = abs(weights*f{i}(roots)-val(i));
    end
  end
  subplot(1,3,2);
  semilogy(1:n,error,'Linewidth',3);

  % summierte Trapezregel
```

```
  error = zeros(3,n);
  for j = 1:n
    % Stuetzstellen
    x = linspace(-1,1,j+1);
    % Quadratur
    for i = 1:3
      q = (sum(f{i}(x(2:end-1)))+f{i}(x(1))/2+f{i}(x(end))/2)*2/j;
      error(i,j) = abs(q-val(i));
    end
  end
  subplot(1,3,3);
  semilogy(1:n,error,'Linewidth',3);
end

% Quadratur-Gewicht-Berechnung
function weights = computeQuadratureWeights( knots, a, b )
  n = numel(knots)-1;
  weights = zeros(1,n+1);
  syms t lagrangeP;
  for j = 1:n+1
    % Lagrange Polynom
    lagrangeP = 1;
    for l = 1:n+1
      if ( l ~= j )
        lagrangeP = lagrangeP * (t-knots(l)) / (knots(j)-knots(l));
      end
    end
    % Integral des Lagrange Polynoms
    weights(j) = int(lagrangeP,'t',a,b);
  end
end
```