Skript zur Vorlesung

Numerik für Differentialgleichungen

Sommersemester 2015

Literatur

Deuflhard, Bornemann: *Numerische Mathematik 2*, de Gruyter, 2008

Hairer, Nørsett, Wanner: *Solving Ordinary Differential Equations I*, Springer, 2009

Stoer, Bulirsch: *Numerische Mathematik 2*, Springer, 2005

1 Einleitung: Modellierung mit Differentialgleichungen

Unter (mathematischer) Modellierung versteht man das Übertragen von beobachteten (physikalischen) Gesetzmäßigkeiten in die Sprache der Mathematik.

Beispiel 1.1 (Der freie Fall).
G. Galilei beobachtete: ***Die Beschleunigung eines fallenden Körpers ist annähernd konstant.*** *Mathematisch formuliert bedeutet dies:*

$$y''(t) = -g \approx -9,8\,\frac{m}{s^2} \quad \text{für } t \in (0,T). \tag{1.1}$$

Führen wir zusätzlich die Geschwindigkeit $v := y'$ ein, so ist (1.1) *äquivalent zu:*

$$\begin{aligned} y'(t) &= v(t), \\ v'(t) &= -g, \end{aligned} \quad \text{für } t \in (0,T).$$

Nach dem Hauptsatz der Integral- und Differentialrechnung gilt dann:

$$\begin{aligned} v(t) &= \int_0^t (-g)\,dt + v(0) = v(0) - g\,t, \\ y(t) &= \int_0^t v(0) - g\,t\,dt + y(0) = y(0) + v(0)\,t - \frac{g}{2}\,t^2, \end{aligned} \quad \text{für } t \in (0,T).$$

Um eine eindeutige Lösung zu bekommen, müssen wir noch weitere Bedingungen vorschreiben, z.B. die Anfangsbedingungen

$$\begin{aligned} y(0) &= y_0, \\ v(0) &= v_0, \end{aligned}$$

für den Anfangsort y_0 und die Anfangsgeschwindigkeit v_0.

Zu gegebenen Anfangsdaten können wir nun die Vorhersage aus dem mathematischen Modell mit dem praktischen Experiment vergleichen. In der Realität gibt es eine endliche Maximalgeschwindigkeit v für einen durch Luft fallenden Körper. Das Modell in Beispiel 1.1 spiegelt diesen Effekt jedoch nicht wider: Wir haben die Luftreibung vernachlässigt!

Beispiel 1.2 (Freier Fall mit Stokes'scher Reibung).
Wir erweitern das Modell aus Beispiel 1.1 um einen Stokes'schen Reibungsterm:

$$y''(t) = -g - \sigma\, y'(t) \quad \text{für } t \in (0,T), \tag{1.2}$$

wobei $\sigma > 0$ der Reibungskoeffizient ist.
Um (1.2) zu lösen, schreiben wir wieder:

$$\begin{aligned} y'(t) &= v(t), \\ v'(t) &= -g - \sigma\, v(t), \end{aligned} \quad \text{für } t \in (0,T). \tag{1.3}$$

Um die Differentialgleichung

$$v'(t) + \sigma\, v(t) = -g$$

*zu lösen verwenden wir einen **integrierenden Faktor**. So bezeichnet man eine Funktion $\varphi : (0,T) \to \mathbb{R}$, $\varphi(t) \neq 0$ für $t \in (0,T)$ mit der Eigenschaft*

$$\big(v'(t) + \sigma\, v(t)\big)\, \varphi(t) = (v\varphi)'(t) \quad \text{für } t \in (0,T).$$

Dann erfüllt $w(t) := v(t)\, \varphi(t)$ die Differentialgleichung

$$w'(t) = -g\, \varphi(t) \quad \text{für } t \in (0,T). \tag{1.4}$$

In unserem Fall ist $\varphi(t) = e^{\sigma t}$ ein integrierender Faktor und wir erhalten: (1.4) gegeben durch

$$w(t) = w(0) - g \int_0^t e^{\sigma s}\, ds = w(0) + \frac{g}{\sigma} - \frac{g}{\sigma}\, e^{\sigma t}.$$

Daher bekommen wir folgende Lösung von (1.3):

$$\begin{aligned} v(t) &= \left(v(0) + \frac{g}{\sigma}\right) e^{-\sigma t} - \frac{g}{\sigma}, \\ y(t) &= \left(y(0) + \frac{v(0)}{\sigma} + \frac{g}{\sigma^2}\right) - \left(\frac{v(0)}{\sigma} + \frac{g}{\sigma^2}\right) e^{-\sigma t} - \frac{g}{\sigma}\, t, \end{aligned} \quad \text{für } t \in (0,T).$$

Für $t \to \infty$ ergibt sich $v(t) \to -\frac{g}{\sigma}$, d.h. dieses Modell kann eine maximale Fallgeschwindigkeit widerspiegeln.

Dieses Modell lässt sich weiter verfeinern. Beispielsweise ist der Reibungskoeffizient σ abhängig von der Geschwindigkeit: $\sigma(v) = \tilde{\sigma}\, |v|$. Einen solchen Modellierungszyklus kann man sich wie in Abbildung 1 veranschaulichen.

Intuitiv sollten umfassendere Modelle einfachere „erweitern" oder „enthalten". Mathematisch sprechen wir von **Modellkonvergenz**. Für die Beispiele 1.1 und 1.2 etwa gilt:

Lemma 1.3.
Zu gegebenen Anfangswerten y_0 und v_0 sei y die Lösung von (1.1) und y_σ die Lösung von (1.2) zu gegebenem $\sigma > 0$. Dann gilt für jedes $t \in (0,T)$:

$$\lim_{\sigma \to 0} y_\sigma(t) = y(t).$$

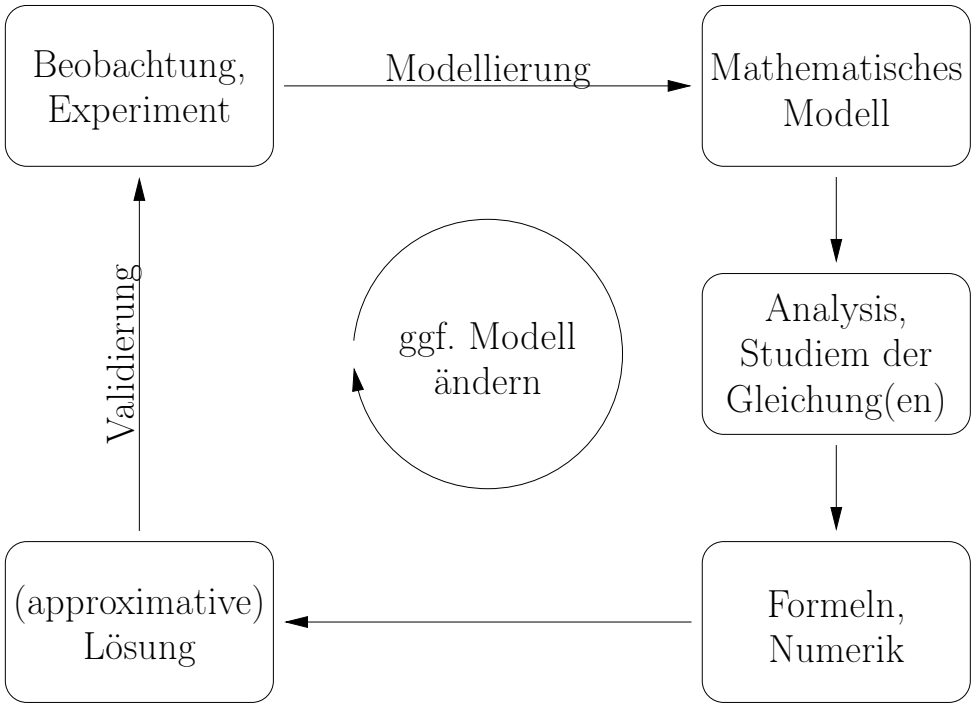

Abbildung 1: Skizze eines Modellierungszyklus

Beweis:
Aus den Beispielen 1.1 und 1.2 folgt:

$$y_\sigma(t) = -\frac{g}{\sigma}\,t - \left(\frac{v_0}{\sigma} + \frac{g}{\sigma^2}\right) e^{-\sigma t} + \left(y_0 + \frac{v_0}{\sigma} + \frac{g}{\sigma^2}\right),$$
$$y(t) = -\frac{g}{2}\,t^2 + v_0\,t + y_0.$$

Fassen wir die Terme in y_σ anders zusammen, so erhalten wir

$$y_\sigma(t) = -\frac{g}{\sigma^2}\left(\sigma\,t + e^{-\sigma t} - 1\right) + \frac{v_0}{\sigma}\left(-e^{-\sigma t} + 1\right) + y_0.$$

Nach dem Taylor'schen Satz gibt es $\xi, \vartheta \in (0, \sigma t)$, sodass

$$e^{-\sigma t} = 1 - \sigma\,t + \frac{\sigma^2\,t^2}{2}\,e^{-\xi}, \qquad\qquad e^{-\sigma t} = 1 - \sigma\,t\,e^{-\vartheta}.$$

Also haben wir für festes $t \in (0, T)$:

$$y_\sigma(t) = -\frac{g}{\sigma^2}\,\frac{\sigma^2\,t^2}{2}\,e^{-\xi} + \frac{v_0}{\sigma}\,\sigma\,t\,e^{-\vartheta} + y_0 = -\frac{g}{2}\,t^2\,\underbrace{e^{-\xi}}_{\to 1} + v_0\,t\,\underbrace{e^{-\vartheta}}_{\to 1} + y_0. \qquad \blacksquare$$

2 Differentialgleichungen erster Ordnung

Sei $G \subset \mathbb{R}^2$ ein Gebiet und sei $f \in C^0(G)$. Eine **Differentialgleichung erster Ordnung** (in expliziter Form) ist von der Gestalt:

$$u' = f(\cdot, u). \tag{2.1}$$

Eine auf einem Intervall I stetig differenzierbare Funktion u mit $\big(t, u(t)\big) \in G$ für alle $t \in I$ heißt **Lösung der Differentialgleichung**, falls gilt:

$$u'(t) = f\big(t, u(t)\big) \quad \text{für alle } t \in I.$$

Beispiel 2.1.
Beispiele für Differentialgleichungen erster Ordnung sind:

1. $u'(t) = \alpha(t)\, u(t) + \beta(t)$,
2. $u'(t) = \frac{1}{u(t)}$,
3. $u'(t) = a\, u(t)\, \big(b - u(t)\big)$ *(logistische Differentialgleichung).*

Ist $n \in \mathbb{N}$ und sind $f_1, \dots, f_n$ auf einem Gebiet $G \subset \mathbb{R}^{n+1}$ stetige Funktionen, so nennen wir

$$\begin{aligned} u_1' &= f_1(\cdot, u_1, \dots, u_n), \\ &\vdots \\ u_n' &= f_n(\cdot, u_1, \dots, u_n) \end{aligned} \tag{2.2}$$

ein **System von Differentialgleichungen erster Ordnung**. Eine Lösung von (2.2) ist ein n-Tupel $u = (u_1, \dots, u_n)$ stetig differenzierbarer Funktionen, die (2.2) punktweise genügen.

Beispiel 2.2 (Räuber-Beute-Modell).
Ein Beispiel für ein Populationsmodell ist das Räuber-Beute-Modell von Lotka und Volterra. Dieses modelliert die Entwicklung zweier Spezies: u_1 sei die Beute-Population (z.B. Kaninchen), u_2 sei die Räuber-Population (z.B. Füchse).

$$\begin{aligned} u_1'(t) &= a\, u_1(t) - b\, u_1(t)\, u_2(t), \\ u_2'(t) &= -c\, u_2(t) + d\, u_1(t)\, u_2(t), \end{aligned} \quad \textit{für } t \in (0, T).$$

Dabei ist

$a > 0$	*Geburtenrate der Beutetiere ohne Störung durch Räuber,*
$b > 0$	*Sterberate der Beutetiere pro Raubtier (Fressrate),*
$c > 0$	*Sterberate der Raubtiere in Abwesenheit von Beutetieren,*
$d > 0$	*Geburtenrate der Raubtiere pro Beutetier.*

Fassen wir $u = (u_1, \dots, u_n) \in \mathbb{R}^n$ als Vektor auf, und definieren $f : \mathbb{R} \times \mathbb{R}^n$ durch

$$f(t, u) = \begin{pmatrix} f_1(t, u_1, \dots, u_n) \\ \vdots \\ f_n(t, u_1, \dots, u_n) \end{pmatrix},$$

so können wir das System (2.2) wieder in der Form

$$u' = f(\cdot, u)$$

schreiben.

Definition 2.3 (Anfangswertproblem).
*Ein Punkt $(t_0, u_0) \in G \subset \mathbb{R}^{n+1}$ ist ein **Anfangswert**, wenn durch ihn der Wert $u(t_0)$ der Lösungsfunktion u an der Stelle t_0 vorgeschrieben wird.*
*Eine Differentialgleichung zusammen mit einem Anfangswert bildet ein **Anfangswertproblem**.*

2.1 Existenz und Eindeutigkeit der Lösung

Lemma 2.4 (Gronwall).
Seien $p, q \in C^0([0,T])$, wobei p nichtnegativ und q monoton wachsend ist. Genügt $u : [0,T] \to \mathbb{R}$ der Ungleichung

$$u(t) \le q(t) + \int_0^t p(s)\, u(s)\, ds \quad \textit{für alle } t \in [0,T],$$

so gilt:

$$u(t) \le q(t)\, \exp\left(\int_0^t p(s)\, ds \right) \quad \textit{für alle } t \in [0,T].$$

Beweis:
Wir definieren

$$P(t) := \int_0^t p(s)\, ds \quad \text{und} \quad w(t) := e^{-P(t)} \int_0^t p(s)\, u(s)\, ds.$$

Dann gilt:

$$w'(t) = \underbrace{e^{-P(t)}\, p(t)}_{\ge 0} \underbrace{\left[u(t) - \int_0^t p(s)\, u(s)\, ds \right]}_{\le q(t)} \le e^{-P(t)}\, p(t)\, q(t).$$

Durch Integration folgt:

$$\begin{aligned} w(t) - \underbrace{w(0)}_{=0} &= \int_0^t \underbrace{e^{-P(s)}\, p(s)}_{\ge 0}\, q(s)\, ds \le q(t) \int_0^t e^{-P(s)}\, p(s)\, ds \\ &= q(t) \left[-e^{-P(s)} \right]_0^t = q(t) \left(1 - e^{-P(t)} \right). \end{aligned}$$

Multiplikation mit $e^{P(t)}$ liefert dann:

$$u(t) - q(t) \le \int_0^t p(s)\, u(s)\, ds = e^{P(t)}\, w(t) \le q(t) \left(e^{P(t)} - 1 \right).$$

Dies ist gerade die Behauptung. ∎

Satz 2.5 (Picard-Lindelöf).
Es sei $t_0 \in I = [a,b]$, $a < b$ und $u_0 \in \mathbb{R}^n$. Für ein $R > 0$ sei $f \in C^0(I \times \overline{B_R(u_0)}, \mathbb{R}^n)$ eine Funktion mit folgenden Eigenschaften:

1. *Es gibt eine Konstante $M \ge 0$, sodass*

$$\left| f(t,u) \right| \le M \quad \textit{für alle } t \in I,\ u \in \overline{B_R(u_0)}. \tag{2.3}$$

2. *Es gibt eine Konstante $L \ge 0$, sodass*

$$\left| f(t,u_1) - f(t,u_2) \right| \le L\, |u_1 - u_2| \quad \textit{für alle } t \in I,\ u_1, u_2 \in \overline{B_R(u_0)}. \tag{2.4}$$

Ist dann $[t_1, t_2] \subset I$ ein Intervall mit $t_0 \in [t_1, t_2]$ und

$$M\,|t_2 - t_1| \le R, \tag{2.5}$$

so gibt es genau ein $u \in C^1([t_1, t_2], \overline{B_R(u_0)})$ mit

$$\begin{aligned} u'(t) &= f\big(t, u(t)\big) \quad \textit{für } t \in [t_1, t_2], \\ u(t_0) &= u_0 \end{aligned}$$

und es gilt:

$$\big|u(t)\big| \le |u_0| + \frac{M}{L}\,\big(e^{L|t-t_0|} - 1\big).$$

Beweis:
Seien $u, v \in C^1([t_1, t_2], \overline{B_R(u_0)})$ zwei Lösungen des Anfangswertproblems. Durch Integration gilt:

$$u(t) = u_0 + \int_{t_0}^{t} f\big(s, u(s)\big)\,ds \quad \text{bzw.} \quad v(t) = u_0 + \int_{t_0}^{t} f\big(s, v(s)\big)\,ds.$$

Also gilt für $\tau > 0$:

$$\begin{aligned} e(\tau) &:= \big|u(t_0+\tau) - v(t_0+\tau)\big| = \left|\int_{t_0}^{t_0+\tau} f\big(s, u(s)\big) - f\big(s, v(s)\big)\,ds\right| \\ &\le \int_{t_0}^{t_0+\tau} \big|f\big(s, u(s)\big) - f\big(s, v(s)\big)\big|\,ds \le L \int_{t_0}^{t_0+\tau} \big|u(s) - v(s)\big|\,ds = L \int_0^{\tau} e(s)\,ds. \end{aligned}$$

Aus dem Gronwall-Lemma 2.4 folgt dann

$$e(\tau) \le 0\, e^{L\,\tau} = 0$$

und damit $u(t) = v(t)$ für $t > t_0$. Analog verfährt man für $t < t_0$.
Um die Existenz zu zeigen, definieren wir induktiv eine Folge von Funktionen $u^k : [t_1, t_2] \to \mathbb{R}^n$ durch

$$\begin{aligned} u^0(t) &:= u_0, \\ u^{k+1}(t) &:= u_0 + \int_{t_0}^{t} f\big(s, u^k(s)\big)\,ds, \quad (t \in [t_1, t_2]). \end{aligned} \tag{2.6}$$

Diese Funktionenfolge heißt **Picard-Iteration**.
Wir zeigen zunächst, dass $u^k \in C^1\big([t_1, t_2], \overline{B_R(u_0)}\big)$. Für u^0 ist dies trivial. Nehmen wir also an, die Aussage gilt für ein $k \in \mathbb{N}$. Dann gilt für $t \in [t_1, t_2]$:

$$\big|u^{k+1}(t) - u_0\big| \le \left|\int_{t_0}^{t} f\big(s, u^k(s)\big)\,ds\right| \le \left|\int_{t_0}^{t} \big|f\big(s, u^k(s)\big)\big|\,ds\right| \le M\,|t - t_0| \le M\,|t_2 - t_1| \le R.$$

Die Differenzierbarkeit folgt aus dem Hauptsatz der Integral- und Differentialrechnung. Also gilt: $u^{k+1} \in C^1\big([t_1, t_2], \overline{B_R(u_0)}\big)$.
Wiederum per Induktion zeigen wir:

$$\big|u^{k+1}(t) - u^k(t)\big| \le M\,L^k\,\frac{|t - t_0|^{k+1}}{(k+1)!}. \tag{$*$}$$

Für $k = 0$ gilt:

$$\left|u^1(t) - u^0(t)\right| = \left|u^1(t) - u_0\right| \leq M\,|t - t_0| = M\,L^0\,\frac{|t-t_0|^1}{0!}.$$

Gilt $(*)$ für ein $k \in \mathbb{N}$, so gilt:

$$\begin{aligned}\left|u^{k+2}(t) - u^{k+1}(t)\right| &\leq \left|\int_{t_0}^{t} \left|f\big(s, u^{k+1}(s)\big) - f\big(s, u^k(s)\big)\right| ds\right| \\ &\leq L\left|\int_{t_0}^{t} \left|u^{k+1}(s) - u^k(s)\right| ds\right| \leq \frac{ML^{k+1}}{(k+1)!}\left|\int_{t_0}^{t} |s-t_0|^{k+1}\, ds\right| \\ &= \frac{ML^{k+1}}{(k+1)!}\left|\int_{t_0}^{t} (s-t_0)^{k+1}\, ds\right| = \frac{ML^{k+1}}{(k+2)!}\,|t-t_0|^{k+2}.\end{aligned}$$

Damit haben wir für $k, m \in \mathbb{N}$:

$$\left|u^{k+m}(t) - u^k(t)\right| \leq \sum_{j=k}^{m-1}\left|u^{j+1}(t) - u^j(t)\right| \leq \frac{M}{L}\sum_{j=k}^{m-1}\frac{(L\,|t-t_0|)^{j+1}}{(j+1)!} \leq \frac{M}{L}\sum_{j=k+1}^{\infty}\frac{(L\,|t-t_0|)^{j}}{j!}.$$

Damit ist $(u^k)_{k\in\mathbb{N}}$ Cauchy-Folge in $X := C^0\big([t_1,t_2], \overline{B_R(u_0)}\big)$, denn

$$\|u^{k+m} - u^k\|_X = \max_{t\in[t_1,t_2]}\left|u^{k+m}(t) - u^k(t)\right| \leq \frac{M}{L}\sum_{j=k+1}^{\infty}\frac{(L\,|t_2-t_1|)^j}{j!} \overset{k\to\infty}{\longrightarrow} 0.$$

Aufgrund der Vollständigkeit von X, existiert ein $u \in X$ mit $\lim_{k\to\infty}\|u - u^k\|_X = 0$.
Aus der Iterationsvorschrift ergibt sich:

$$u(t) \overset{k\to\infty}{\longleftarrow} u^{k+1}(t) = u_0 + \int_{t_0}^{t} f\big(s, u^k(s)\big)\, ds \overset{k\to\infty}{\longrightarrow} u_0 + \int_{t_0}^{t} f\big(s, u(s)\big)\, ds.$$

Nach dem Hauptsatz der Integral- und Differentialrechnung ist u dann differenzierbar und es gilt:

$$u'(t) = f\big(t, u(t)\big).$$

Ferner gilt:

$$\left|u^m(t) - u^0(t)\right| \leq \frac{M}{L}\sum_{j=1}^{\infty}\frac{(L\,|t-t_0|)^j}{j!} = \frac{M}{L}\left(e^{L\,|t-t_0|} - 1\right).$$

Im Grenzwert $m \to \infty$ ergibt sich die a-priori Abschätzung (2.6). ∎

2.2 Analytische Lösungen

In einigen Fällen kann man solche Anfangswertprobleme für Differentialgleichungen exakt lösen. Der Einfachheit halber beschränken wir uns hier auf skalare Gleichungen.

Lineare Differentialgleichungen

$$\begin{aligned} u'(t) + p(t)\,u(t) &= q(t) \quad \text{für } t \in I, \\ u(t_0) &= u_0. \end{aligned} \tag{2.7}$$

Lemma 2.6.
Sei I ein offenes Intervall, seien $p, q \in C^0(I)$ und sei $(t_0, u_0) \in I \times \mathbb{R}$. Ist P eine Stammfunktion zu p in I, so ist die eindeutige Lösung von (2.7) gegeben durch

$$u(t) = u_0\, e^{-(P(t)-P(t_0))} + e^{-P(t)} \int_{t_0}^{t} e^{P(s)}\, q(s)\, ds. \tag{2.8}$$

Beweis:
Übungsaufgabe. ∎

Exakte Differentialgleichungen

Definition 2.7 (Exakte Differentialgleichungen).
Seien $I, J \subset \mathbb{R}$ Intervalle und sei $G = I \times J$. Eine Differentialgleichung der Form

$$N\big(t, u(t)\big)\, u'(t) + M\big(t, u(t)\big) = 0 \tag{2.9}$$

mit $M, N \in C^0(G)$ und $N(t,u) \neq 0$ für alle $(t,u) \in G$, heißt exakt in G, falls es eine Funktion $H \in C^1(G)$ gibt, sodass gilt:

$$\frac{\partial H}{\partial u}(t,u) = N(t,u) \quad \text{und} \quad \frac{\partial H}{\partial t}(t,u) = M(t,u) \quad \forall\ (t,u) \in G. \tag{2.10}$$

Hängt N nicht von t und M nicht von u ab, so heißt die Differentialgleichung separiert.

Lemma 2.8.
Seien $I, J \subset \mathbb{R}$ Intervalle und sei die Differentialgleichung (2.9) exakt auf $I \times J$. Dann ist $u : I \to J$ genau dann Lösung von (2.9), falls eine Konstante $c = c(u) \in \mathbb{R}$ existiert, sodass gilt:

$$H\big(t, u(t)\big) = c \quad \text{für alle } t \in I. \tag{2.11}$$

Beweis:
Erfüllt die Funktion u die Gleichung (2.11), so erhalten wir durch Differentiation nach t:

$$0 = \frac{\partial}{\partial t}\big[H\big(t, u(t)\big)\big] = \frac{\partial H}{\partial t}(t,u) + \frac{\partial H}{\partial u}(t,u)\, u'(t) = M(t,u) + N(t,u)\, u'(t),$$

d.h. u ist Lösung von (2.9).
Ist u eine Lösung von (2.9), so gilt für ein beliebiges $t_0 \in I$:

$$0 = \int_{t_0}^{t} N\big(s, u(s)\big)\, u'(s) + M\big(s, u(s)\big)\, ds = \int_{t_0}^{t} \frac{\partial}{\partial s}\big[H\big(s, u(s)\big)\big]\, ds = H\big(t, u(t)\big) - H\big(t_0, u(t_0)\big).$$

Setzen wir $c = H(t_0, u(t_0))$, so ergibt sich die Behauptung. ∎

Lemma 2.9.
Auf einem Rechteck G seien $M, N \in C^1(G)$. Dann ist (2.9) genau dann exakt auf G, falls gilt:

$$\frac{\partial N}{\partial t}(t,u) = \frac{\partial M}{\partial u}(t,u) \quad \text{für alle } (t,u) \in G.$$

Beweis:
Wir fixieren ein beliebiges $(t_0, u_0) \in G$ und definieren

$$H(t,u) := \int_{t_0}^{t} M(s,u_0)\,ds + \int_{u_0}^{u} N(t,v)\,dv.$$

Dann gilt:

$$\begin{aligned}
&\frac{\partial H}{\partial u}(t,u) = N(t,u),\\
&\frac{\partial H}{\partial t}(t,u) = M(t,u_0) + \int_{u_0}^{u} \frac{\partial N}{\partial t}(t,v)\,dv = M(t,u_0) + \int_{u_0}^{u} \frac{\partial M}{\partial v}(t,v)\,dv = M(t,u).
\end{aligned}$$

Also ist (2.9) exakt.
Die Umkehrung folgt direkt aus dem Satz von Schwarz:

$$\frac{\partial^2 H}{\partial t\,\partial u} = \frac{\partial^2 H}{\partial u\,\partial t}.$$

∎

Bernoulli'sche Differentialgleichungen

Definition 2.10 (Bernoulli'sche Differentialgleichung).
Sei $I = [a,b]$ ein Intervall und seien $p, q \in C^0(I)$. Eine Differentialgleichung der Form

$$u'(t) + p(t)\,u(t) = q(t)\,u(t)^{\alpha} \tag{2.12}$$

*heißt **Bernoulli'sche Differentialgleichung**.*

Bemerkung 2.11.
Der Ausdruck $u(t)^{\alpha}$ ist für $\alpha \in \mathbb{R}$ i.a. nur wohldefiniert, falls $u(t) \geq 0$ ist.
Die Funktion $u \equiv 0$ ist immer eine Lösung von (2.12). Ist u eine Lösung von (2.12) und gilt $u(t_0) = 0$ für ein t_0, so können wir für $\alpha \geq 1$ aus Satz 2.5 (Picard-Lindelöf) folgern, dass $u \equiv 0$. Aus dem Zwischenwertsatz folgt dann, dass $u(t) > 0$ für alle $t \in I$, falls es ein $t_0 \in I$ gibt, sodass $u(t_0) > 0$.

Satz 2.12.
Sei $\alpha \notin \{0,1\}$ und sei $I = [a,b]$. Eine positive Funktion $u \in C^1(I)$ ist Lösung von (2.12) genau dann, wenn die Funktion v mit $v(t) := u(t)^{1-\alpha}$ eine (positive) Lösung der linearen Differentialgleichung

$$v'(t) + (1-\alpha)\,p(t)\,v(t) = (1-\alpha)\,q(t) \tag{2.13}$$

ist.

Die Fälle $\alpha = 0$ und $\alpha = 1$ sind nur ausgenommen, da (2.12) in diesen Fällen bereits linear ist. Für $\alpha = 0$ fällt (2.13) mit (2.12) zusammen, für $\alpha = 1$ ergibt sich $v(t) = u(t)^0 = 1$ und man erhält u nicht aus v zurück.

Beweis:
Nach der Kettenregel ist v differenzierbar und es gilt:

$$v'(t) = (1-\alpha)\,u(t)^{-\alpha}\,u'(t).$$

Durch Multiplikation von (2.12) mit $(1-\alpha)\,u(t)^{-\alpha} \neq 0$ erhalten wir

$$\underbrace{(1-\alpha)\,u(t)^{-\alpha}\,u'(t)}_{=v'(t)} + (1-\alpha)\,p(t)\,\underbrace{u(t)^{1-\alpha}}_{=v(t)} = (1-\alpha)\,q(t).$$

Die Funktion v ist dann auch positiv.
Umgekehrt können wir (2.13) mit $\frac{1}{1-\alpha}\,u(t)^{\alpha}$, $u(t) = v(t)^{\frac{1}{1-\alpha}}$, multiplizieren, um (2.12) zu erhalten. ∎

Beispiel 2.13 (Logistische Differentialgleichung).
Wir betrachten das Anfangswertproblem

$$\begin{aligned} u'(t) &= a\,u(t)\,\big(b - u(t)\big), \\ u(0) &= u_0 \end{aligned} \tag{2.14}$$

mit $a, b > 0$, sowie $u_0 \in [0, b]$.
Dies ist eine Bernoulli'sche Differentialgleichung mit

$$p(t) = -a\,b, \qquad q(t) = -a, \qquad \alpha = 2.$$

Nach der Transformation $v(t) = \frac{1}{u(t)}$ löst v das Anfangswertproblem für die lineare Differentialgleichung

$$\begin{aligned} v'(t) + a\,b\,v(t) &= a, \\ v(0) &= \tfrac{1}{u_0}. \end{aligned}$$

Nach Lemma 2.6 lautet die Lösung:

$$v(t) = \frac{1}{u_0}\,e^{-abt} + e^{-abt}\int_0^t e^{abs}\,a\,ds = \frac{e^{-abt}}{u_0} + a\,e^{-abt}\frac{e^{abt}-1}{ab} = \frac{e^{-abt}}{u_0} + \frac{1}{b}\,(1 - e^{-abt}).$$

Durch Rücktransformation $u(t) = \frac{1}{v(t)}$ erhalten wir

$$u(t) = \frac{b\,u_0}{b\,e^{-abt} + u_0\,(1 - e^{-abt})}.$$

Lösungen zu verschiedenen Anfangswerten für $a = b = 1$ sind in Abbildung 2 dargestellt.

Ricatti'sche Differentialgleichungen

Definition 2.14 (Ricatti'sche Differentialgleichung).
Sei $I = [a, b]$ ein Intervall und seien $p, q, f \in C^0(I)$. Eine Differentialgleichung der Form

$$u'(t) + p(t)\,u(t) = q(t)\,u(t)^2 + f(t) \tag{2.15}$$

*heißt **Ricatti'sche Differentialgleichung**.*

Satz 2.15.
Sei $I = [a, b]$ ein Intervall und seien $p, q, f \in C^0(I)$. Ist $u : I \to \mathbb{R}$ eine Lösung von (2.15), so ist jede andere Lösung von (2.15) von der Form

$$v(t) = u(t) + \frac{1}{z(t)}.$$

Dabei ist z eine Lösung der linearen Differentialgleichung

$$z'(t) - \big[p(t) - 2\,u(t)\,q(t)\big]\,z(t) = -q(t).$$

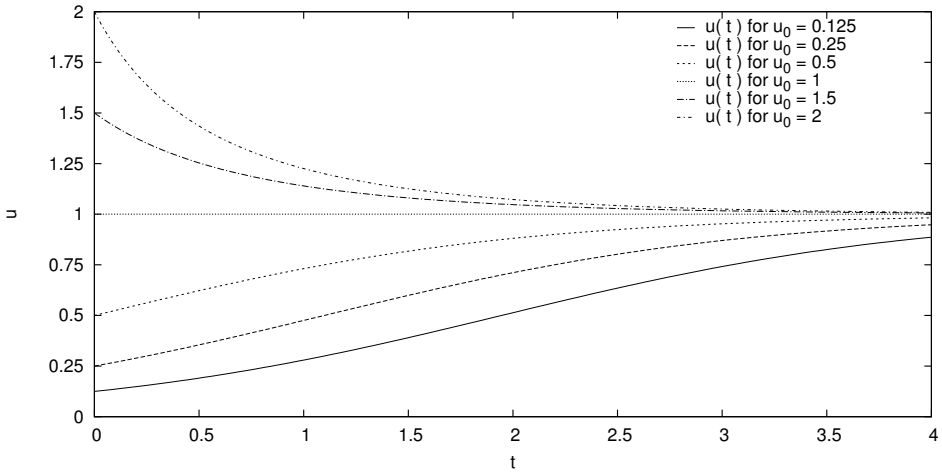

Abbildung 2: Lösungen von (2.14) für $a = b = 1$ und verschiedene Anfangswerte u_0.

Beweis:
Seien $u, v : I \to \mathbb{R}$ Lösungen von (2.15). Dann gilt für $w := v - u$:

$$w'(t)+p(t)\,w(t) = q(t)\,\big(v(t)^2-u(t)^2\big) = q(t)\,w(t)\,\big(u(t)+v(t)\big) = q(t)\,w(t)\,\big(2\,u(t)+w(t)\big).$$

Die Differenzfunktion w erfüllt also die Bernoulli'sche Differentialgleichung

$$w'(t) + \big[p(t) - 2\,u(t)\,q(t)\big]\,w(t) = q(t)\,w(t)^2.$$

Nach Satz 2.12 führt die Transformation $z(t) = \frac{1}{w(t)}$ auf die lineare Differentialgleichung

$$z'(t) - \big[p(t) - 2\,u(t)\,q(t)\big]\,z(t) = -q(t). \qquad \blacksquare$$

2.3 Das Euler'sche Polygonzugverfahren

In diesem Abschnitt wollen wir ein erstes numerisches Verfahren zur Approximation der Lösung des Anfangswertproblems

$$\begin{aligned} u' &= f(\cdot, u), \quad \text{in } [t_0, T], \\ u(0) &= u_0 \end{aligned} \tag{2.16}$$

für $f \in C^0([t_0, T])$ studieren.
Die Differentialgleichung $u' = f(\cdot, u)$ ist äquivalent zum System

$$\begin{aligned} t' &= 1, \\ u' &= f(t, u). \end{aligned} \quad \text{bzw.} \quad \begin{aligned} t'(s) &= 1, \\ u'(s) &= f(t(s), u(s)). \end{aligned}$$

Dieses ordnet jedem (t, u) eine Richtung (t', u') zu. Man bezeichnet daher die Abbildung

$$(t, u) \mapsto \big(1, f(t, u)\big)$$

als **Richtungsfeld** der Differentialgleichung $u' = f(\cdot, u)$.

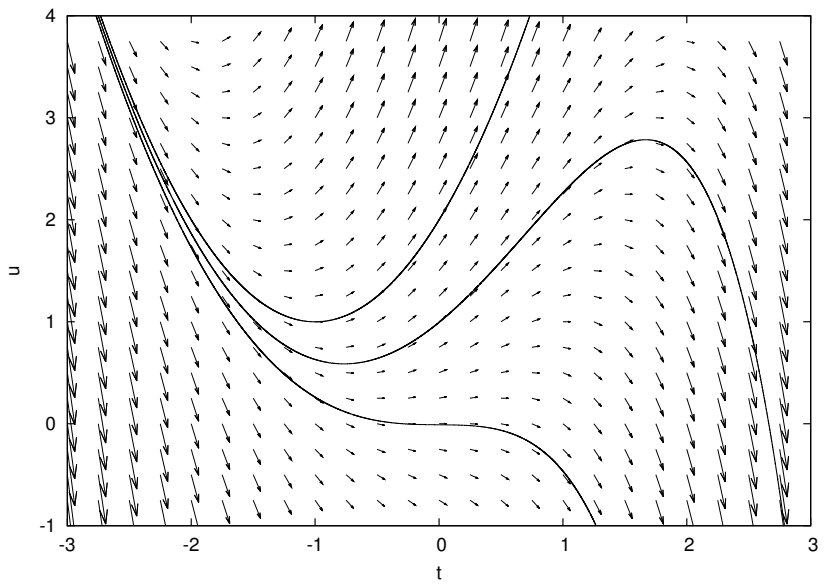

Abbildung 3: Richtungsfeld der Differentialgleichung $u'(t) = u(t) - t^2$

Abbildung 3 zeigt beispielsweise das Richtungsfeld für die Differentialgleichung

$$u'(t) = u(t) - t^2, \quad \text{d.h.} \quad f(t,u) = u - t^2,$$

sowie einige (approximative) Lösungskurven. Die Tangente an eine Lösungskurve im Punkt (t,u) ist also gerade

$$\tau_{t,u} = \Big\{ \big(1, f(t,u)\big)\, s + \big(t,u\big) \;\Big|\; s \in \mathbb{R} \Big\}.$$

Die Idee des **Euler'schen Polygonzugverfahrens** ist nun, die Lösung auf einem kleinen Intervall durch ihre Tangente zu approximieren. Dazu seien Zeitpunkte $t_0, \ldots, t_N$, sowie der Startwert $u_0 \in \mathbb{R}^n$ gegeben. Auf dem Intervall $[t_k, t_{k+1}]$ approximieren wir (t,u) durch die Tangente im Punkt (t_k, u_k), d.h.

$$\begin{aligned} t &= t_k + 1\,(t - t_k), \\ u(t) &\approx u_k + f(t_k, u_k)\,(t - t_k) \end{aligned} \qquad \text{für } t \in [t_k, t_{k+1}].$$

Definieren wir u_{k+1} als die Approximation von u in t_{k+1} und setzen $h_k := t_{k+1} - t_k$, so ergibt sich das folgende Verfahren:

$$\begin{aligned} t_{k+1} &:= t_k + h_k, \\ u_{k+1} &:= u_k + h_k\, f(t_k, u_k), \end{aligned} \qquad \text{für } k = 0, \ldots, N-1. \tag{2.17}$$

Diese Idee ist in Abbildung 4 skizziert.

Falls $h_k = h$ unabhängig von k ist, können wir den Fehler des Euler-Verfahrens leicht abschätzen.

Satz 2.16.
Sei $U \subset \mathbb{R}^n$ konvex und sei $u \in C^2([t_0,T], U)$ Lösung des Anfangswertproblems (2.16), wobei $f \in C^0([t_0,T] \times U)$ der Lipschitz-Bedingung

$$\big|f(t,u_1) - f(t,u_2)\big| \le L\,|u_1 - u_2| \quad \textit{für alle } t \in [t_0,T], u_1, u_2 \in U,$$

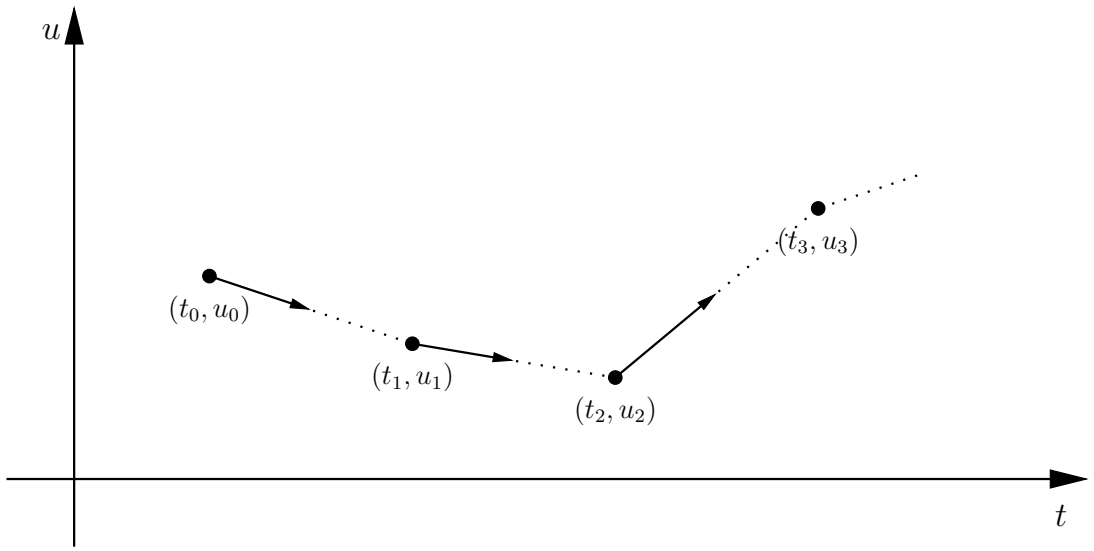

Abbildung 4: Skizze der Idee des expliziten Euler-Verfahrens

mit einem $L \geq 0$ genügt. Zu $N \in \mathbb{N}$ sei $h = \frac{T-t_0}{N}$, $t_k := t_0 + k\,h$ und u_k, $k = 0, \ldots, N$ sei die approximative Lösung des Euler-Verfahrens. Dann gilt:

$$\max_{k=0,\ldots,N} \big|u(t_k) - u_k\big| \leq h\,\frac{e^{L(T-t_0)} - 1}{L} \max_{[t_0,T]} |u''|.$$

Beweis:
Durch Taylorentwicklung erhalten wir

$$\begin{aligned} u(t_{k+1}) &= u(t_k) + h\,u'(t_k) + h^2\,u''(\xi_k) \\ &= u(t_k) + h\,f\big(t_k, u(t_k)\big) + h^2\,u''(\xi_k) \end{aligned} \tag{2.19}$$

mit einem $\xi_k \in (t_k, t_{k+1})$.
Wir definieren $e_k := |u(t_k) - u_k|$. Dann gilt:

$$\begin{aligned} e_{k+1} &= \big|u(t_k) + h\,f\big(t_k, u(t_k)\big) + h^2\,u''(\xi_k) - u_k - h\,f(t_k, u_k)\big| \\ &\leq \big|u(t_k) - u_k\big| + h\,\big|f\big(t_k, u(t_k)\big) - f(t_k, u_k)\big| + h^2|u''(\xi_k)| \\ &\leq \big|u(t_k) - u_k\big| + h\,L\,\big|u(t_k) - u_k\big| + h^2|u''(\xi_k)| \\ &\leq (1 + \underbrace{h\,L}_{=:A})\,e_k + \underbrace{h^2 \max_{[t_0,T]} |u''|}_{=:B}. \end{aligned}$$

Wir zeigen mittels vollständiger Induktion:

$$e_k \leq e^{kA}\,e_0 + \frac{e^{kA} - 1}{A}\,B \qquad\qquad \text{für } k \in \mathbb{N}.$$

Offenbar ist die Behauptung für $k = 0$ richtig. Nehmen wir an die Behauptung gilt für

ein $k \in \mathbb{N}$. Dann gilt:

$$\begin{aligned} e_{k+1} &\le (1+A)\left[e^{kA}\, e_0 + \frac{e^{kA}-1}{A}\, B\right] + B \\ &= (1+A)\left[e^{kA}\, e_0 + \frac{e^{kA}}{A}\, B\right] - \frac{1+A}{A}\, B + B \\ &\le e^{A}\left[e^{kA}\, e_0 + \frac{e^{kA}}{A}\, B\right] - \frac{1}{A}\, B. \end{aligned}$$

Dabei haben wir die Abschätzung $(1+A) \le e^A$ verwendet. Folglich haben wir gezeigt:

$$e_k \le e^{khL}\, e_0 + \frac{e^{khL}-1}{hL}\, h^2 \max_{[t_0,T]} |u''| \le \frac{e^{L(T-t_0)}-1}{L}\, h \max_{[t_0,T]} |u''|. \qquad \blacksquare$$

Das folgende Beispiel zeigt, dass eine Fehlerabschätzung allein nicht ausreicht, um ausreichende Qualität der numerischen Lösung sicherzustellen.

Beispiel 2.17.
Betrachten wir das Anfangswertproblem

$$\begin{aligned} u'(t) &= \lambda\, u(t), \quad \textit{für } t \ge 0, \\ u(0) &= 1, \end{aligned}$$

d.h. $f(t,u) = \lambda\, u$. *Die exakte Lösung ist gegeben durch* $u(t) = e^{\lambda t}$.
Das Eulerverfahren für dieses Anfangswertproblem lautet:

$$\begin{aligned} u_{k+1} &= u_k + h\, f(t_k, u_k) = u_k + \lambda\, h\, u_k = (1+\lambda\, h)\, u_k, \\ u_0 &= 1. \end{aligned}$$

Also ist $u_k = (1+\lambda\, h)^k$, *d.h.*

- *für* $1+\lambda\, h > 0$ *ist* u_k *positiv und approximiert die exakte Lösung zumindest qualitativ.*
- *für* $0 < h = -\frac{1}{\lambda}$ *ist* $u_k = 0$ *für* $k \ge 1$.
- *Für* $0 < h = -\frac{2}{\lambda}$ *ist* $u_k = (-1)^k$; *das Verfahren „oszilliert".*

Wir sagen auch, dass das Verfahren nur für $1 + \lambda\, h > 0$ *stabil ist.*
Dieses Verhalten widerspricht nicht Satz 2.16. Im Fall $\lambda < 0$ *ist* $t_0 = 0$, $L = -\lambda$ *und* $\max|u''| = \lambda^2$, *d.h.*

$$\max_{k=0,\dots,N} \big|u(t_k) - u_k\big| \le h\left(e^{-\lambda T} - 1\right)(-\lambda).$$

Wählen wir einen einzigen Zeitschritt, also $T = h$, *so gilt für den Fall* $h \ge -\frac{1}{\lambda}$ *bereits*

$$-\lambda\, h\left(e^{-\lambda T} - 1\right) \ge 1\big(e^1 - 1\big) \approx 1.718.$$

Nach dieser Abschätzung könnte der Fehler also bereits nach einem Zeitschritt deutlich größer sein als die exakte Lösung.

2.4 Einschrittverfahren

Das Euler-Verfahren ist ein Beispiel für ein Einschrittverfahren:

Definition 2.18 (Einschrittverfahren).
*Seien diskrete Zeitpunkte $t_0, \dots, t_N$ gegeben und sei $h_k = t_{k+1} - t_k$. Ein numerisches Verfahren für das Anfangswertproblem (2.16) heißt **Einschrittverfahren**, wenn es von der Form*

$$u_{k+1} = u_k + h_k \, \Phi(t_k, u_k; h_k)$$

*ist. Die Funktion Φ heißt **Verfahrensfunktion** des Einschrittverfahrens.*

Beispiel 2.19.
Einfache Beispiele für Einschrittverfahren sind:

1. *(explizites) Eulerverfahren $u_{k+1} = u_k + h_k \, f(t_k, u_k)$:*
$$\Phi(t, u; h) = f(t, u).$$
2. *implizites Eulerverfahren $u_{k+1} = u_k + h_k \, f(t_{k+1}, u_{k+1})$:*
$$\Phi(t, u; h) = f\big(t + h, u + h \, \Phi(t, u; h)\big).$$
Ebenso wie u_{k+1} ist Φ hier nur implizit definiert.
3. *Crank-Nicolson-Verfahren:*
$$\Phi(t, u; h) = \frac{1}{2} f(t, u) + \frac{1}{2} f\big(t + h, u + h \, \Phi(t, u; h)\big).$$
4. *Heun-Verfahren:*
$$\Phi(t, u; h) = \frac{1}{2} f(t, u) + \frac{1}{2} f\big(t + h, u + h \, f(t, u)\big).$$

In (2.19) haben wir gezeigt, dass für $u \in C^2([t_0, T])$ gilt:

$$u(t_{k+1}) = u(t_k) + h \, f\big(t_k, u(t_k)\big) + \mathcal{O}(h^2) = u(t_k) + h \, \Phi\big(t_k, u(t_k); h\big) + \mathcal{O}(h^2).$$

Dies motiviert folgende Definition:

Definition 2.20 (Konsistenz).
*Wir sagen ein Einschrittverfahren ist **konsistent von der Ordnung** p, falls für Lösungen $u \in C^{p+1}([t_0, T])$ gilt:*

$$u(t_{k+1}) = u(t_k) + h \, \Phi\big(t_k, u(t_k); h\big) + \mathcal{O}(h^{p+1}).$$

Satz 2.21.
Sei $U \subset \mathbb{R}^n$ konvex und sei $u \in C^{p+1}([t_0, T], U)$ Lösung des Anfangswertproblems (2.16). Zu $N \in \mathbb{N}$ sei $h = \frac{T - t_0}{N}$, $t_k := t_0 + k\,h$ und $u_k \in U$, $k = 1, \dots, N$ sei durch das Einschrittverfahren

$$u_{k+1} = u_k + h \, \Phi(t_k, u_k; h)$$

gegeben, wobei die Verfahrensfunktion $\Phi(\cdot, \cdot; h) \in C^0([t_0, T] \times U)$ der Lipschitz-Bedingung

$$\big|\Phi(t, u_1; h) - \Phi(t, u_2; h)\big| \le L \, |u_1 - u_2| \quad \text{für alle } t \in [t_0, T], u_1, u_2 \in U,$$

mit einem von h unabhängigen $L \ge 0$ genüge. Ist das Verfahren konsistent von der Ordnung p, so gilt:

$$\max_{k=0,\dots,N} \big|u(t_k) - u_k\big| \le C \, \frac{e^{L(T - t_0)} - 1}{L} \, h^p.$$

Beweis:
Wir definieren den Fehler $e_k := |u(t_k) - u_k|$. Da das Verfahren konsistent von der Ordnung p ist, gilt dann:

$$\begin{aligned} e_{k+1} &= \big|u(t_k) + h\,\Phi\big(t_k, u(t_k); h\big) + \mathcal{O}(h^{p+1}) - u_k - h\,\Phi(t_k, u_k; h)\big| \\ &\le \big|u(t_k) - u_k\big| + h\,\big|\Phi\big(t_k, u(t_k); h\big) - \Phi(t_k, u_k; h)\big| + \mathcal{O}(h^{p+1}) \\ &\le \big|u(t_k) - u_k\big| + h\,L\,\big|u(t_k) - u_k\big| + \mathcal{O}(h^{p+1}). \end{aligned}$$

Also existiert eine Konstante $C \ge 0$, sodass

$$e_{k+1} \le (1 + h\,L)\,e_k + C\,h^{p+1}.$$

Wie im Beweis von Satz 2.16 folgt dann

$$e_k \le e^{khL}\,e_0 + C\,\frac{e^{khL} - 1}{hL}\,h^{p+1} \le C\,\frac{e^{L(T-t_0)} - 1}{L}\,h^p. \qquad \blacksquare$$

Beispiel 2.22.
Falls $f \in C^0([t_0, T], U)$ der Lipschitz-Bedingung

$$\big|f(t, u_1) - f(t, u_2)\big| \le L\,|u_1 - u_2| \quad \textit{für alle } t \in [t_0, T], u_1, u_2 \in U$$

genügt, so erfüllen folgende Verfahren ebenfalls die Lipschitz-Bedingung:

1. *Heun-Verfahren:*

$$\begin{aligned} &\big|\Phi(t, u_1; h) - \Phi(t, u_2; h)\big| \\ &\le \frac{1}{2}\,\big|f(t, u_1) - f(t, u_2)\big| + \frac{1}{2}\,\big|f\big(t + h, u_1 + h\,f(t, u_1)\big) - f\big(t + h, u_2 + h\,f(t, u_2)\big)\big| \\ &\le \frac{L}{2}\,|u_1 - u_2| + \frac{L}{2}\,\big|u_1 + h\,f(t, u_1) - u_2 - h\,f(t, u_2)\big| \\ &\le L\,|u_1 - u_2| + \frac{L}{2}\,h\,\big|f(t, u_1) - f(t, u_2)\big| \\ &\le L\left(1 + \frac{L}{2}\,h\right)|u_1 - u_2|. \end{aligned}$$

2. *implizites Eulerverfahren:*

$$\begin{aligned} &\big|\Phi(t, u_1; h) - \Phi(t, u_2; h)\big| \\ &\le \big|f\big(t + h, u_1 + h\,\Phi(t, u_1; h)\big) - f\big(t + h, u_2 + h\,\Phi(t, u_2; h)\big)\big| \\ &\le L\,\big|u_1 + h\,\Phi(t, u_1; h) - u_2 - h\,\Phi(t, u_2; h)\big| \\ &\le L\,|u_1 - u_2| + h\,L\,\big|\Phi(t, u_1; h) - \Phi(t, u_2; h)\big|. \end{aligned}$$

Folglich gilt für $(1 - h\,L) > 0$:

$$\big|\Phi(t, u_1; h) - \Phi(t, u_2; h)\big| \le \frac{L}{1 - hL}\,|u_1 - u_2|.$$

2.5 Runge-Kutta-Verfahren

Definition 2.23 (Runge-Kutta-Verfahren).
*Ein s-**stufiges Runge-Kutta-Verfahren** ist ein Einschrittverfahren, dessen Verfahrensfunktion von der Form*

$$\Phi(t, u; h) = \sum_{i=1}^{s} b_i \, k_i$$

mit Stufen k_i von der Form

$$k_i = f\Big(t + c_i \, h, u + h \sum_{j=1}^{s} A_{ij} \, k_j\Big) \quad \text{für } i = 1, \ldots, s$$

*und Koeffizienten $A \in \mathbb{R}^{s \times s}$, $b, c \in \mathbb{R}^s$. Ist A eine strikte untere Dreiecksmatrix, so heißt das Verfahren **explizit**; andernfalls heißt das Verfahren **implizit**. Im häufigen Spezialfall einer (nicht-strikten) unteren Dreiecksmatrix A heißt das Verfahren **diagonal-implizit**.*

Runge-Kutta-Verfahren werden häufig durch ihr Butcher-Tableau dargestellt:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

Beispiel 2.24.
Die bisher erwähnten Verfahren sind Runge-Kutta-Verfahren:

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

Euler (explizit) *Euler (implizit)* *Crank-Nicolson* *Heun*

Weitere (explizite) Runge-Kutta-Verfahren:

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|cc} 0 & & \\ \frac{1}{3} & \frac{1}{3} & \\ \frac{2}{3} & 0 & \frac{2}{3} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ 1 & 0 & 0 & 1 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \quad \frac{1}{6} \end{array}$$

Runge (2. Ordnung) *Runge (3. Ordnung)* *Heun (3. Ordnung)* *klassisches Runge-Kutta (4. Ordnung)*

Lemma 2.25.
Seien $A \in \mathbb{R}^{s \times s}$ und $b, c \in \mathbb{R}^s$ die Koeffizienten eines s-stufigen Runge-Kutta Verfahrens, das Konsistent von der Ordnung p ist. Dann gilt:

$$\sum_{i=1}^{s} b_i = 1.$$

Beweis:
Wendet man das Verfahren auf das Anfangswertproblem

$$\begin{aligned} u' &= 1 =: f(t,u), \\ u(0) &= 0 \end{aligned}$$

an, so folgt $k_i = 1$, $i = 1, \ldots, s$ und es gilt:

$$u(t+h) = u(t) + h \sum_{i=1}^{s} b_i \, k_i + \mathcal{O}(h^{p+1}).$$

Setzen wir die exakte Lösung $u(t) = t$ ein, so erhalten wir

$$t + h = t + h \sum_{i=1}^{s} b_i + \mathcal{O}(h^{p+1}) \quad \text{bzw.} \quad 1 - \sum_{i=1}^{s} b_i = \mathcal{O}(h^p).$$

Im Grenzwert $h \to 0$ folgt die Behauptung. ∎

Wendet man ein Runge-Kutta-Verfahren auf die beiden äquivalenten Systeme

$$u' = f(\cdot, u) \quad \text{und} \quad \begin{pmatrix} t' \\ u' \end{pmatrix} = \begin{pmatrix} 1 \\ f(t,u) \end{pmatrix}$$

an, so erhält man für die Zwischenschritte:

$$k_i = f\Big(t + c_i \, h, u + h \sum_{j=1}^{s} A_{ij} \, k_j\Big) \quad \text{bzw.} \quad \begin{pmatrix} l_i \\ k_i \end{pmatrix} = \begin{pmatrix} 1 \\ f\Big(t + h \sum_{j=1}^{s} A_{ij} \, l_j, u + h \sum_{j=1}^{s} A_{ij} \, k_j\Big) \end{pmatrix}.$$

Beide Vorschriften ergeben genau dann dasselbe Verfahren, wenn gilt:

$$c_i = \sum_{j=1}^{s} A_{ij}.$$

Auch wenn diese Bedingung i.A. nicht zwingend für die Konsistenz ist, so wird sie von den meisten Runge-Kutta-Verfahren erfüllt.

Lemma 2.26.
Sei $U \subset \mathbb{R}^n$ konvex und sei $f \in C^0([t_0, T], U)$ mit

$$\big|f(t,u_1) - f(t,u_2)\big| \leq L \, |u_1 - u_2| \quad \textit{für alle } t \in [t_0, T], u_1, u_2 \in U.$$

Dann erfüllt jedes Runge-Kutta-Verfahren für hinreichend kleines $h > 0$ die Lipschitz-Bedingung

$$\big|\Phi(t,u_1;h) - \Phi(t,u_2;h)\big| \leq L_\Phi \, |u_1 - u_2| \quad \textit{für alle } t \in [t_0, T], u_1, u_2 \in U.$$

Beweis:
Seien $t \in [t_0, T]$, $u_1, u_2 \in U$ und bezeichnen k_i^1, k_i^2 die Runge-Kutta-Stufen bzgl. u_1 bzw. u_2, d.h.

$$k_i^l = f\Big(t + c_i \, h, u_l + h \sum_{j=1}^{s} A_{ij} \, k_j^l\Big) \quad \text{für } i = 1, \ldots, s.$$

Dann gilt:

$$|k_i^1 - k_i^2| \le L \left| u_1 + h \sum_{j=1}^{s} A_{ij}\, k_j^1 - u_2 - h \sum_{j=1}^{s} A_{ij}\, k_j^2 \right| \le L\,|u_1 - u_2| + L\,h \sum_{j=1}^{s} |A_{ij}|\,|k_j^1 - k_j^2|.$$

Summation über alle i liefert:

$$\sum_{i=1}^{s} \Big(1 - h\,L \sum_{j=1}^{s} |A_{ji}|\Big)\,|k_i^1 - k_i^2| \le s\,L\,|u_1 - u_2|.$$

Unter der Bedingung

$$h\,L \sum_{j=1}^{s} |A_{ji}| \le \frac{1}{2} \quad \text{für } i = 1, \dots, s$$

gilt also die Abschätzung $|k_i^1 - k_i^2| \le 2\,s\,L\,|u_1 - u_2|$ und damit

$$\big|\Phi(t, u_1; h) - \Phi(t, u_2; h)\big| \le \sum_{i=1}^{s} |b_i|\,|k_i^1 - k_i^2| \le 2\,s\,L \Big(\sum_{i=1}^{s} |b_i|\Big)\,|u_1 - u_2|. \qquad \blacksquare$$

Zusammen mit Satz 2.21 folgt also, dass jedes konsistente Runge-Kutta-Verfahren konvergiert, falls f in u Lipschitz-stetig ist.

Bemerkung 2.27.
Wir können die Stufen eines Runge-Kutta-Verfahrens auch anders schreiben, indem wir folgende Beobachtung ausnutzen:

$$k_i = f\Big(t + c_i\,h, \underbrace{u_k + h \sum_{j=1}^{s} A_{ij}\, k_j}_{=:u^{(i-1)}}\Big)$$

Damit lautet das Runge-Kutta-Verfahren dann:

$$u_{k+1} = u_k + h_k\,\Phi(t_k, u_k; h) = u_k + h_k \sum_{i=1}^{s} b_i\, f\big(t_k + c_i\,h_k, u_k^{(i-1)}\big),$$

$$u_k^{(i-1)} = u_k + h_k \sum_{j=1}^{s} A_{ij}\, f\big(t_k + c_j\,h_k, u_k^{(j-1)}\big).$$

Die Runge-Kutta Zwischenschritte $u_k^{(i-1)}$ haben die gleiche Form wie der vollständige Runge-Kutta Schritt. Wir können daher $u_k^{(i-1)}$ als Approximation des Wertes $u(t_k + c_i\,h_k)$ interpretieren.

Lemma 2.28.
Seien $A \in \mathbb{R}^{s\times s}$ und $b, c \in \mathbb{R}^s$ die Koeffizienten eines s-stufigen Runge-Kutta-Verfahrens, das Konsistent von der Ordnung p ist. Dann ist die Quadraturformel

$$\int_0^1 q(s)\,ds \approx \sum_{i=1}^{s} b_i\, q(c_i)$$

exakt auf $\mathbb{P}_{p-1}$, d.h. für $q \in \mathbb{P}_{p-1}$ gilt hier Gleichheit.

Beweis:
Sei $q \in \mathbb{P}_{p-1}$. Wenden wir das Verfahren auf das Anfangswertproblem

$$\begin{aligned} u' &= q, \\ u(0) &= 0 \end{aligned}$$

an, so erhält man

$$\int_0^h q(s)\,ds = u(h) = h \sum_{i=1}^{s} b_i\, q(c_i\, h) + \mathcal{O}(h^{p+1}).$$

Folglich gilt:

$$P(h) := \int_0^h q(s)\,ds - h \sum_{i=1}^{s} b_i\, q(c_i\, h) = \mathcal{O}(h^{p+1}).$$

Nun ist $P \in \mathbb{P}_p$, d.h. es gibt $\alpha_i \in \mathbb{R}$, sodass

$$P(h) = \sum_{i=0}^{p} \alpha_i\, h^i.$$

Ist $P \neq 0$, so wähle i minimal mit $\alpha_i \neq 0$ und erfülle h die Bedingung $(p+1)\,|\alpha_j|\,h^{j-i} \leq |\alpha_i|$ für alle $j > i$. Dann gilt:

$$C\,h^{p+1} \geq |P(h)| = \Big|\sum_{i=0}^{p} \alpha_i\, h^i\Big| \geq |\alpha_i|\,h^i - \sum_{j=i+1}^{p} |\alpha_j|\,h^j \geq \Big[|\alpha_i| - \sum_{j=i+1}^{p} |\alpha_i|\Big]\,h^i \geq \frac{1}{p+1}|\alpha_i|\,h^i.$$

Dies ist aber ein Widerspruch, sodass $P \equiv 0$ sein muss. Aus $P(1) = 0$ und der Beliebigkeit von $q \in \mathbb{P}_{p-1}$ folgt die Behauptung. ∎

Umgekehrt verwendet man eine Quadraturformel

$$\int_0^1 q(s)\,ds \approx \sum_{i=1}^{s} b_i\, q(c_i),$$

um ein Runge-Kutta-Verfahren herzuleiten:

$$u(t_k + h_k) = u(t_k) + \int_{t_k}^{t_k+h_k} f\big(t, u(t)\big)\,dt \approx u(t_k) + h_k \sum_{i=1}^{s} b_i\, f\big(t + c_i\, h_k, u(t + c_i\, h_k)\big).$$

Wie in Bemerkung 2.27 verwenden wir nun weitere Quadraturformeln, um die notwendigen Werte $u(t + c_i\, h_k)$ zu erhalten:

$$u(t + c_i\, h_k) \approx u(t_k) + c_i\, h_k \sum_{j=1}^{s} \frac{A_{ij}}{c_i}\, f\big(t + c_j\, h_k, u(t + c_j\, h_k)\big).$$

Dabei müssen die gleichen Quadraturpunkte $(c_i)_{i=1,\ldots,s}$ verwendet werden.

Satz 2.29.
Für $p \geq 5$ existiert kein p-stufiges explizites Runge-Kutta-Verfahren der Ordnung p.

Beweis:
Siehe z.B. Hairer, Nørsett, Wanner: *Solving Ordinary Differential Equations I.* ∎

Bemerkung 2.30.
Bisher haben wir die Wahl einer geeigneten Zeitschrittweite h_k außer Acht gelassen. Als Indikator für eine geeignete Zeitschrittweite kann man für einen Zeitschritt aus dem Startwert u_k zwei Approximationen u_{k+1} und $\tilde{u}_{k+1}$ mit Hilfe zweier Runge-Kutta-Verfahren der Ordnungen p und $\tilde{p}$ berechnen. Der Zeitschritt wird dann so gewählt, dass der Fehler $|u_{k+1} - \tilde{u}_{k+1}|$ hinreichend klein ist.

Um den Rechenaufwand nicht unnötig in die Höhe zu treiben, versucht man aus den Zwischenergebnissen des genaueren Verfahrens zusätzlich eine Approximation niedrigerer Ordnung abzuleiten. Man sagt auch ein Runge-Kutta-Verfahren niedrigerer Ordnung wird in ein Runge-Kutta-Verfahren höherer Ordnung eingebettet.

In das Verfahren 3. Ordnung von Heun lässt sich beispielsweise leicht ein Verfahren 2. Ordnung einbetten:

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

2.6 Mehrschrittverfahren

Runge-Kutta-Verfahren benötigen pro Zeitschritt mehrere Auswertungen der Funktion f. Eine andere Idee besteht darin, neben u_k auch $u_{k-1}, \ldots, u_{k-m+1}$ zu verwenden, um u_{k+1} mit höherer Ordnung zu approximieren.

Um ein solches Verfahren herzuleiten, betrachten wir wieder die integrale Form des Problems:

$$u(t_{k+1}) = u(t_k) + \int_{t_k}^{t_{k+1}} f\big(t, u(t)\big)\, dt.$$

Aus den bereits bekannten Werten rekonstruieren wir nun ein Polynom $f^{(k)} \in \mathbb{P}_{m-1}$ mit

$$f^{(k)}(t_{k-i}) = f(t_{k-i}, u_{k-i}) \quad \text{für } i = 0, \ldots, m-1.$$

Dieses Vorgehen liefert das **Adams-Bashforth Verfahren**

$$u_{k+1} := u_k + \int_{t_k}^{t_{k+1}} f^{(k)}(t)\, dt,$$

wobei das Integral nun exakt berechnet werden kann. Mit Hilfe das Lagrange'schen Interpolationspolynoms erhalten wir

$$f^{(k)}(t) = \sum_{i=0}^{m-1} f(t_{k-i}, u_{k-i})\, L_i(t - t_k) \quad \text{mit} \quad L_i(t) = \prod_{j \neq i} \frac{t - t_{-j}}{t_{-i} - t_{-j}} = \prod_{j \neq i} \frac{t + jh}{(j-i)h}$$

Dies liefert:

$$u_{k+1} := u_k + \sum_{i=0}^{m-1} f(t_{k-i}, u_{k-i}) \int_{t_k}^{t_{k+1}} L_i(t - t_k)\, dt = u_k + \sum_{i=0}^{m-1} f(t_{k-i}, u_{k-i}) \int_0^h L_i(t)\, dt$$
$$= u_k + h \sum_{i=0}^{m-1} f(t_{k-i}, u_{k-i}) \underbrace{\int_0^1 \prod_{j \neq i} \frac{t+j}{j-i}\, dt}_{=:\beta_i}$$

Die Werte β_i für kleine m lauten wie folgt:

m	β_0	β_1	β_2	β_3
1	1			
2	$\frac{3}{2}$	$-\frac{1}{2}$		
3	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$

Solche Verfahren lassen sich wie folgt verallgemeinern:

Definition 2.31 (Mehrschrittverfahren).
Ein Verfahren der Form

$$\sum_{i=0}^{m} \alpha_i\, u_{k+i} = h\, \Phi(t_k, u_k, \ldots, u_{k+m}; h)$$

*mit Koeffizienten $\alpha_i \in \mathbb{R}$ und $\alpha_m = 1$ heißt m-**Schrittverfahren**. Das Verfahren heißt explizit, falls Φ nicht von u_{k+m} abhängt; andernfalls heißt es implizit. Das Mehrschrittverfahren heißt linear, falls es $\beta_i \in \mathbb{R}$ gibt, sodass*

$$\Phi(t_k, u_k, \ldots, u_{k+m}; h) = \sum_{i=0}^{m} \beta_i\, f(t_{k+i}, u_{k+i}).$$

Achtung: Um ein Mehrschrittverfahren zu starten benötigt man neben u_0 die Werte $u_1, \ldots, u_{m-1}$. Diese können z.B. durch ein Einschrittverfahren (der passenden Ordnung) bestimmt werden.

In dieser Darstellung lautet das Adams-Bashforth-Verfahren:

$$u_{k+m} - u_{k+m-1} = \int_{t_{k+m-1}}^{t_{k+m}} \sum_{i=0}^{m-1} f(t_{k+i}, u_{k+i}) \prod_{j \neq i} \frac{t - t_{k+j}}{t_{k+i} - t_{k+j}}\, dt$$

Nehmen wir nun den Punkt (t_{k+m}, u_{k+m}) als Interpolationspunkt hinzu, so erhalten wir das **Adams-Moulton-Verfahren**:

$$u_{k+m} - u_{k+m-1} = \int_{t_{k+m-1}}^{t_{k+m}} \sum_{i=0}^{m} f(t_{k+i}, u_{k+i}) \prod_{j \neq i} \frac{t - t_{k+j}}{t_{k+i} - t_{k+j}}\, dt$$

Im entarteten Fall $m = 0$ vereinfacht sich dieses Verfahren zum impliziten Euler-Verfahren:

$$u_k - u_{k-1} = \int_{t_{k-1}}^{t_k} f(t_k, u_k)\, dt = h\, f(t_k, u_k).$$

Auch der Fall $m = 1$ führt auf ein bereits bekanntes Verfahren:

$$\begin{aligned}
u_{k+1} - u_k &= \int_{t_k}^{t_{k+1}} \sum_{i=0}^{1} f(t_{k+i}, u_{k+i}) \prod_{j \neq i} \frac{t - t_{k+j}}{t_{k+i} - t_{k+j}} \, dt \\
&= \int_{t_k}^{t_{k+1}} f(t_k, u_k) \, \frac{t - t_{k+1}}{t_k - t_{k+1}} + f(t_{k+1}, u_{k+1}) \, \frac{t - t_k}{t_{k+1} - t_k} \, dt \\
&= \frac{1}{h} \left(f(t_{k+1}, u_{k+1}) \int_{t_k}^{t_{k+1}} (t - t_k) \, dt - f(t_k, u_k) \int_{t_k}^{t_{k+1}} (t - t_{k+1}) \, dt \right) \\
&= \frac{1}{h} \left(f(t_{k+1}, u_{k+1}) \Big[\frac{1}{2} (t - t_k)^2 \Big]_{t_k}^{t_{k+1}} - f(t_k, u_k) \Big[\frac{1}{2} (t - t_{k+1})^2 \Big]_{t_k}^{t_{k+1}} \right) \\
&= \frac{h}{2} f(t_{k+1}, u_{k+1}) + \frac{h}{2} f(t_k, u_k).
\end{aligned}$$

Wir erhalten also das Crank-Nicholson-Verfahren.

Einen ähnlichen Ansatz verfolgen die **Backward-Differentiation-Formulas** (kurz BDF-Verfahren): Wir fordern, dass das Interpolationspolynom $q_k \in \mathbb{P}_m$ mit $q_k(t_k + i) = u_{k+i}$, $i = 0, \ldots, m$, die Differentialgleichung im Punkt t_{k+m} erfüllt, d.h.

$$q_k'(t_{k+m}) = f\big(t_{k+m}, q(t_{k+m})\big) = f(t_{k+m}, u_{k+m}).$$

Definition 2.32 (Konsistenz).
*Wir sagen ein m-Schrittverfahren ist **konsistent von der Ordnung** p, falls für Lösungen $u \in C^{p+1}([t_0, T])$ gilt:*

$$\sum_{i=0}^{m} \alpha_i \, u(t_{k+i}) = h \, \Phi\big(t_k, u(t_k), \ldots, u(t_{k+m}); h\big) + \mathcal{O}(h^{p+1}).$$

Satz 2.33.
Ein lineares m-Schittverfahren

$$\sum_{i=0}^{m} \alpha_i \, u_{k+i} = h \sum_{i=0}^{m} \beta_i \, f(t_{k+i}, u_{k+i})$$

ist genau dann konsistent von der Ordnung p, wenn

$$\sum_{i=0}^{m} \alpha_i = 0 \quad \textit{und} \quad \sum_{i=0}^{m} \alpha_i \, i^q = q \sum_{i=0}^{m} \beta_i \, i^{q-1} \quad \textit{für alle } q = 1, \ldots, p. \tag{2.20}$$

Beweis:
Sei $u \in C^{p+1}([t_k, t_{k+m}])$ Lösung des Anfangswertproblems. Dann gilt:

$$\begin{aligned}
&\sum_{i=0}^{m} \alpha_i \, u(t_{k+i}) - h \sum_{i=0}^{m} \beta_i \, f\big(t_{k+i}, u(t_{k+i})\big) \\
&= \sum_{i=0}^{m} \alpha_i \, u(t_{k+i}) - h \sum_{i=0}^{m} \beta_i \, u'(t_{k+i}) \\
&= \sum_{i=0}^{m} \alpha_i \sum_{q=0}^{p} \frac{(ih)^q}{q!} \, u^{(q)}(t_k) - h \sum_{i=0}^{m} \beta_i \sum_{q=0}^{p-1} \frac{(ih)^q}{q!} \, u^{(q+1)}(t_k) + \mathcal{O}(h^{p+1}) \\
&= u(t_k) \Big(\sum_{i=0}^{m} \alpha_i \Big) + \sum_{q=1}^{p} \frac{h^q}{q!} \, u^{(q)}(t_k) \Big(\sum_{i=0}^{m} i^q \, \alpha_i - q \sum_{i=0}^{m} i^{q-1} \, \beta_i \Big) + \mathcal{O}(h^{p+1}).
\end{aligned}$$

Dieser Ausdruck ist genau dann von der Ordnung $\mathcal{O}(h^{p+1})$, wenn die geklammerten Terme verschwinden, d.h., wenn (2.20) gilt. ∎

Beispiel 2.34 (Ordnung des Adams-Bashforth-Verfahrens).
Betrachten wir für $q = 0, \dots, m$ die Differentialgleichung

$$u'(t) = q\, t^{q-1} =: f\big(t, u(t)\big).$$

Die exakte Lösung lautet $u(t) = t^q$. Sei $u_i = u(t_i)$, $i = 0, \dots, m-1$ gegeben. Das Adams-Bashforth-Verfahren lautet dann:

$$u_m - u_{m-1} = \int_{t_{m-1}}^{t_m} \tilde{f}(t)\, dt,$$

wobei $\tilde{f} \in \mathbb{P}_{m-1}$ die Punkte $(t_i, f(t_i, u_i))$, $i = 0, \dots, m-1$, interpoliert. Nun gilt: $\tilde{f}(t) = q\, t^{q-1}$ und damit $u_m = u(t_m)$, d.h. das Verfahren liefert die exakte Lösung:

$$\sum_{i=0}^{m} \alpha_i\, (i\, h)^q = \sum_{i=0}^{m} \alpha_i\, t_i^q = \sum_{i=0}^{m} \alpha_i\, u_i = h \sum_{i=0}^{m} \beta_i\, f(t_i, u_i) = h \sum_{i=0}^{m} \beta_i\, q\, t_i^{q-1} = h \sum_{i=0}^{m} \beta_i\, q\, (i\, h)^{q-1}.$$

Nach Division durch h^q erhalten wir (2.20). Das Adams-Bashforth-Verfahren ist also mindestens von der Ordnung m.

Analog sieht man ein, dass das Adams-Moulton-Verfahren mindestens von der Ordnung $m+1$ ist.

Beispiel 2.35 (Adams-Bashforth-Moulton-Verfahren).
Das Adams-Bashforth-Verfahren und das Adams-Moulton-Verfahren lassen sich zu einem expliziten Verfahren der Ordnung $m+1$ kombinieren. Dazu führen wir folgende Schritte aus:

1. *Prädiktorschritt (Adams-Bashforth):*

$$\tilde{u}_{k+m} = u_{k+m-1} + h \sum_{i=0}^{m-1} \beta_i^{AB}\, f(t_{k+i}, u_{k+i}).$$

2. *Korrektorschritt (Adams-Moulton):*

$$u_{k+m} = u_{k+m-1} + h \sum_{i=0}^{m-1} \beta_i^{AM}\, f(t_{k+i}, u_{k+i}) + h\, \beta_m^{AM}\, f(t_{k+m}, \tilde{u}_{k+m}).$$

Das Adams-Bashforth-Moulton Verfahren lässt sich als explizites Mehrschrittverfahren schreiben:

$$u_{k+m} - u_{k+m-1} = h\, \Phi^{AM}\big(t_k, u_k, \dots, u_{k+m-1}, u_{k+m-1} + h\, \Phi^{AB}(t_k, u_k, \dots, u_{k+m-1}; h); h\big).$$

Dieses Verfahren ist nicht linear.

Satz 2.36.
Gegeben seien explizites und ein implizites Mehrschrittverfahren

$$\sum_{i=0}^{m} \alpha_i^{ex} u_{k+i} = h\,\Phi^{ex}(t_k, u_k, \ldots, u_{k+m-1}; h),$$
$$\sum_{i=0}^{m} \alpha_i^{im} u_{k+i} = h\,\Phi^{im}(t_k, u_k, \ldots, u_{k+m}; h).$$

Die Verfahren seien konsistent von der Ordnung p^{ex} bzw. p^{im}. Ist Φ^{im} Lipschitz-stetig bezüglich u_{k+m}, so ist das Prädiktor-Korrektor Verfahren

$$\tilde{u}_{k+m} + \sum_{i=1}^{m} \alpha_i^{ex} u_{k+i} = h\,\Phi^{ex}(t_k, u_k, \ldots, u_{k+m-1}; h)$$
$$\sum_{i=0}^{m} \alpha_i^{im} u_{k+i} = h\,\Phi^{im}(t_k, u_k, \ldots, u_{k+m-1}, \tilde{u}_{k+m}; h)$$

konsistent von der Ordnung $p^{pc} := \min\{p^{ex}+1, p^{im}\}$.

Beweis:
Sei $u \in C^{p^{pc}+1}([t_k, t_{k+m}])$ und sei $u_{k+i} = u(t_{k+i})$, $i = 0, \ldots, m-1$. Dann gilt:

$$\tilde{u}_{k+m} - u(t_{k+m}) = \mathcal{O}(h^{p^{ex}+1}).$$

Ferner folgt aus der Konsistenz des impliziten Verfahrens:

$$\begin{aligned}
&|u_{k+m} - u(t_{k+m})| \\
&= h\left|\Phi(t_k, u_k, \ldots, u_{k+m-1}, \tilde{u}_{k+m}; h) - \Phi\big(t_k, u_k, \ldots, u_{k+m-1}, u(t_{k+m}); h\big)\right| + \mathcal{O}(h^{p^{im}+1}) \\
&\le h\,L\left|\tilde{u}_{k+m} - u(t_{k+m})\right| + \mathcal{O}(h^{p^{im}+1}) = \mathcal{O}(h^{p^{ex}+2}) + \mathcal{O}(h^{p^{im}+1}).
\end{aligned}$$

Dies ist gerade die Behauptung. ∎

Korollar 2.37.
Das m-Schritt Adams-Bashforth-Moulton-Verfahren ist von der Ordnung $m+1$.

Bemerkung 2.38.
Das 1-Schritt Adams-Bashforth-Moulton-Verfahren verwendet das explizite Euler-Verfahren als Prädiktor und das Crank-Nicholson-Verfahren als Korrektor. Das Verfahren lautet also:

$$\begin{aligned}
\tilde{u}^{k+1} &= u_k + h\,f(t_k, u_k), \\
u^{k+1} &= u_k + \frac{h}{2} f(t_k, u_k) + \frac{h}{2} f(t_{k+1}, \tilde{u}_{k+1}) \\
&= u_k + \frac{h}{2} f(t_k, u_k) + \frac{h}{2} f\big(t_{k+1}, u_k + h\,f(t_k, u_k)\big).
\end{aligned}$$

Dies ist gerade das Heun-Verfahren.

Bemerkung 2.39.
Seien zwei Mehrschrittverfahren wie in Satz 2.36 gegeben. Man kann auch mehrere Korrektor-Schritte ausführen:

$$u_{k+m}^{(0)} + \sum_{i=1}^{m} \alpha_i^{ex} u_{k+i} = h\,\Phi^{ex}(t_k, u_k, \ldots, u_{k+m-1}; h),$$
$$u_{k+m}^{(s)} + \sum_{i=1}^{m} \alpha_i^{im} u_{k+i} = h\,\Phi^{im}(t_k, u_k, \ldots, u_{k+m-1}, u_{k+m}^{(s-1)}; h), \quad s = 1, \ldots, \sigma.$$

Der Korrektor-Schritt entspricht dabei einer Fixpunktiteration. Durch rekursive Anwendung von Satz 2.36 kann man zeigen, dass das resultierende Prädiktor-Multikorrektor-Verfahren konsistent von der Ordnung $\min\{p^{ex} + \sigma, p^{im}\}$ *ist.*

Definition 2.40 (Konvergenz).
Das Mehrschrittverfahren

$$\sum_{i=0}^{m} \alpha_i\, u_{k+i} = h\,\Phi(t_k, u_k, \ldots, u_{k+m}; h)$$

heißt konvergent, wenn für alle Anfangswertprobleme

$$u' = f(\cdot, u) \quad \text{in } [t_0, T], \quad u(t_0) \ = 0,$$

die den Bedingungen des Satzes von Picard-Lindelöf genügen, gilt:

$$\max_k |u_k - u(t_k)| \xrightarrow{h \to 0} 0,$$

falls dies bereits für $k = 0, \ldots, m-1$ *erfüllt ist.*
Das Verfahren heißt konvergent von der Ordnung p*, wenn für Anfangswertprobleme mit Lösungen* $u \in C^{p+1}([t_0, T])$ *gilt:*

$$\max_k |u_k - u(t_k)| = \mathcal{O}(h^p)$$

falls dies bereits für $k = 0, \ldots, m-1$ *erfüllt ist.*

Beispiel 2.41.
Betrachten wir das explizite Zweischrittverfahren

$$u_{k+2} + 4\,u_{k+1} - 5\,u_k = 4\,h\,f(t_{k+1}, u_{k+1}) + 2\,h\,f(t_k, u_k). \qquad (*)$$

Für die Koeffizienten gilt:

$$\alpha_2 = 1, \quad \alpha_1 = 4, \quad \alpha_0 = -5, \quad \beta_2 = 0, \quad \beta_1 = 4, \quad \beta_0 = 2.$$

Nach Satz 2.33 ist dieses Verfahren konsistent von der Ordnung 3, denn

$$\begin{aligned}
1 + 4 - 5 &= 1 \cdot \alpha_2 + 1 \cdot \alpha_1 + 1 \cdot \alpha_0 = 0,\\
2 + 4 &= 2 \cdot \alpha_2 + 1 \cdot \alpha_1 + 0 \cdot \alpha_0 = 1 \cdot \big(1 \cdot \beta_2 + 1 \cdot \beta_1 + 1 \cdot \beta_0\big) = 4 + 2,\\
4 + 4 &= 4 \cdot \alpha_2 + 1 \cdot \alpha_1 + 0 \cdot \alpha_0 = 2 \cdot \big(2 \cdot \beta_2 + 1 \cdot \beta_1 + 0 \cdot \beta_0\big) = 2 \cdot 4,\\
8 + 4 &= 8 \cdot \alpha_2 + 1 \cdot \alpha_1 + 0 \cdot \alpha_0 = 3 \cdot \big(4 \cdot \beta_2 + 1 \cdot \beta_1 + 0 \cdot \beta_0\big) = 3 \cdot 4.
\end{aligned}$$

Betrachten wir nun das triviale Anfangswertproblem

$$u' = 0,$$
$$u(0) = u_0,$$

mit exakter Lösung $u(t) = u_0$, so lautet das Verfahren:

$$u_{k+2} + 4\,u_{k+1} - 5\,u_k = 0.$$

Machen wir nun den Ansatz $u_k = \lambda^k\,u_0$, so ergibt sich: $\lambda^2 + 4\,\lambda - 5 = 0$. Dieses Polynom besitzt die Nullstellen

$$\lambda = -2 \pm \sqrt{4+5} \quad \textit{d.h.} \quad \lambda_1 = 1,\ \ \lambda_2 = -5.$$

Sind nun u_0 und u_1 gegeben, so können wir diese wie folgt darstellen:

$$u_0 = \lambda_1^0\,z_1 + \lambda_2^0\,z_2,$$
$$u_1 = \lambda_1^1\,z_1 + \lambda_2^1\,z_2.$$

Dieses Gleichungssystem für z_1, z_2 ist immer lösbar. Aufgrund der Linearität von $(*)$ lautet die Lösung dann

$$u_k = \lambda_1^k\,z_1 + \lambda_2^k\,z_2.$$

Wählen wir nun $u_0 = 1$ und $u_1 = 1 + \delta$, so erhalten wir $z_1 = 1 + \frac{\delta}{6}$ und $z_2 = -\frac{\delta}{6}$, d.h.

$$u_k = 1 + \frac{\delta}{6} - (-5)^k\,\frac{\delta}{6} \overset{k\to\infty}{\longrightarrow} \begin{cases} -\infty, & k \text{ gerade}, \\ \infty, & k \text{ ungerade}. \end{cases}$$

Eine beliebig kleine Störung δ des exakten Wertes $u_1 = 1$ wird also verstärkt; die Lösung oszilliert. Für den Fehler gilt:

$$|u_k - u(k\,h)| = \frac{\delta}{6}\,|1 - (-5)^k|$$

Nun wächst $5^k = 5^{\frac{t}{h}}$ schneller als jedes Polynom in $\frac{1}{h}$. Ist $\delta = \mathcal{O}(h^q)$ eine polynomiale Störung, so konvergiert das Verfahren im allgemeinen nicht.

Durch dieses Beispiel motiviert, betrachten wir die Anwendung eines allgemeinen, linearen Mehrschrittverfahrens

$$\sum_{i=0}^{m} \alpha_i\,u_{k+i} = h\,\sum_{i=0}^{m} \beta_i\,f(t_{k+i}, u_{k+i})$$

auf die Differentialgleichung $u' = 0$. Dann gilt:

$$\sum_{i=0}^{m} \alpha_i u_{k+i} = 0. \tag{2.21}$$

Gegeben $u_0, \ldots, u_{m-1}$ liefert uns diese Differenzengleichung eine Folge $(u_k)_{k\in\mathbb{N}}$, die die Lösung $u \equiv u_0$ approximieren soll.

Definition 2.42 (Nullstabilität).
*Ein Mehrschrittverfahren heißt **nullstabil**, falls jede Lösung von* (2.21) *beschränkt ist.*

Lemma 2.43.
Seien $\lambda_1, \ldots, \lambda_l$ die Nullstellen des Polynoms

$$q(\lambda) = \sum_{i=0}^{m} \alpha_i \, \lambda^i$$

mit Vielfachheiten $\sigma_1, \ldots, \sigma_l$. Dann lautet die allgemeine Lösung von (2.21)

$$u_k = \sum_{i=1}^{l} p_i(k) \, \lambda_i^k \tag{2.22}$$

mit Polynomen $p_i \in \mathbb{P}_{\sigma_i - 1}$, $i = 1, \ldots, l$.

Beweis:
Wir zeigen zunächst, dass für eine σ-Fache Nullstelle λ von q durch

$$u_k := p(k) \, \lambda^k$$

für jedes $p \in \mathbb{P}_{\sigma-1}$ eine Lösung von (2.21) gegeben ist. Mit Hilfe des Newton'schen Interpolationspolynoms können wir p in der Form

$$p(k+t) = \gamma_0 + \sum_{j=1}^{\sigma-1} \gamma_j \prod_{s=0}^{j-1} (t - s)$$

darstellen. Da λ eine σ-fache Nullstelle von q ist, verschwinden $q^{(0)}(\lambda), \ldots, q^{(\sigma-1)}(\lambda)$ und es gilt:

$$\begin{aligned}
\sum_{i=0}^{m} \alpha_i \, u_{k+i} &= \sum_{i=0}^{m} \alpha_i \, p(k+i) \, \lambda^{k+i} = \lambda^k \sum_{i=0}^{m} \alpha_i \left[\gamma_0 + \sum_{j=1}^{\sigma-1} \gamma_j \prod_{s=0}^{j-1} (i - s) \right] \lambda^i \\
&= \lambda^k \, \gamma_0 \sum_{i=0}^{m} \alpha_i \, \lambda^i + \sum_{j=1}^{\sigma-1} \gamma_j \, \lambda^{k+j} \sum_{i=0}^{m} \left[\prod_{s=0}^{j-1} (i - s) \right] \lambda^{i-j} \, \alpha_i \\
&= \lambda^k \, \gamma_0 \, q(\lambda) + \sum_{j=1}^{\sigma-1} \gamma_j \, \lambda^{k+j} \, q^{(j)}(\lambda) = 0.
\end{aligned}$$

Aufgrund der Linearität von (2.21) folgt dann, dass durch (2.22) eine Lösung der Differenzengleichung gegeben ist.

Da $\sigma_1 + \ldots + \sigma_l = m$ ist, haben $p_1, \ldots, p_l$ zusammen genau m Freiheitsgrade. Die Darstellbarkeit jeder Lösung folgt dann aus der linearen Unabhängigkeit der u_k (Übungsaufgabe). ∎

Satz 2.44 (Dahlquist'sche Wurzelbedingung).
Ein Mehrschrittverfahren ist genau dann nullstabil, wenn das charakteristische Polynom

$$q(\lambda) = \sum_{i=0}^{m} \alpha_i \, \lambda^i$$

die folgende Wurzelbedingung erfüllt:

1. *Alle Nullstellen (Wurzeln) $\lambda \in \mathbb{C}$ von q liegen im Einheitskreis, d.h. $|\lambda| \leq 1$.*

2. *Alle Nullstellen $\lambda \in \mathbb{C}$ von q, die auf dem Rand des Einheitskreises liegen ($|\lambda| = 1$) sind einfach.*

Beweis:
Für $|\lambda| < 1$ gilt:

$$u_k = p(k)\,\lambda^k \overset{k\to\infty}{\longrightarrow} 0,$$

da λ^k ein Exponential ist. Im Fall einer einfachen Nullstelle mit $|\lambda| = 1$ gilt:

$$|u_k| = |\gamma_0\,\lambda^k| = |\gamma_0|.$$

In beiden Fällen bleibt die Folge beschränkt.
Existiert umgekehrt eine σ-fache Nullstelle λ mit $|\lambda| = 1$, so wählen wir:

$$u_k = k\,\lambda^k.$$

Dann gilt $|u_k| = k$ und die Folge ist unbeschränkt. Analog können wir für eine Nullstelle λ mit $|\lambda| > 1$ die Lösung

$$u_k = \lambda^k$$

wählen, um $|u_k| \to \infty$ für $k \to \infty$ zu erhalten. ∎

Beispiel 2.45.
Für Adams-Verfahren mit m Schritten lautet das charakteristische Polynom

$$q(\lambda) = \lambda^m - \lambda^{m-1} = (\lambda - 1)\,\lambda^{m-1}.$$

Wir haben also $\lambda = 0$ als $m-1$-fache Nullstelle und $\lambda = 1$ als einfache Nullstelle. Die Verfahren genügen also der Wurzelbedingung und sind daher nullstabil.

Bemerkung 2.46.
Das m-schrittige BDF-Verfahren ist stabil für $m \leq 6$. Für $m \geq 7$ ist es instabil.

Lemma 2.47.
Das Mehrschrittverfahren

$$\sum_{i=0}^{m} \alpha_i\,u_{k+i} = h\,\Phi(t_k, u_k, \ldots, u_{k+m}; h)$$

sei konvergent und erfülle die Bedingung

$$\Phi(t, u_0, \ldots, u_m; h) \equiv 0, \quad \textit{falls } f \equiv 0.$$

Dann ist das Verfahren nullstabil.

Beweis:
Nehmen wir an, das Mehrschrittverfahren ist nicht nullstabil. Dann existiert eine Folge $(u_k)_{k\in\mathbb{N}}$, sodass

$$\sum_{i=0}^{m} \alpha_i\,u_{k+i} = 0 \quad \forall k \in \mathbb{N} \qquad \text{und} \qquad u_k \overset{k\to\infty}{\longrightarrow} \infty.$$

Wir approximieren nun die Lösung von $u' = 0$ auf $[0, T]$. Für $n \in \mathbb{N}$ setzen wir $h_n := \frac{T}{n}$ und definieren

$$u_k^{(n)} := \frac{1}{|u_n|}\, u_k.$$

Dann gilt:

$$\max_{k=0,\dots,m-1} |u_k^{(n)}| = \frac{1}{|u_n|} \max_{k=0,\dots,m-1} |u_k| \overset{n\to\infty}{\longrightarrow} 0.$$

Andererseits ist $u_k^{(n)}$, $k = 0, \dots, n$, die approximative Lösung von $u' = 0$ mit Schrittweite h_n. Für den Fehler gilt:

$$\max_{k=0,\dots,n} |u_k^{(n)} - u(k\, h_n)| \geq |u_n^{(n)} - u(T)| = |1 - 0| = 1,$$

d.h. das Verfahren ist nicht konvergent. ∎

Ein Mehrschrittverfahren lässt sich auch als Einschrittverfahren umschreiben. Dazu schreiben wir das Mehrschrittverfahren von der Form

$$u_{k+m} = -\sum_{i=0}^{m-1} \alpha_i\, u_{k+i} + h\, \Phi(t_k, u_k, \dots, u_{k+m}; h).$$

Nun setzen wir $\bar{\alpha} = (\alpha_0, \dots, \alpha_{m-1})^T$ und definieren für $t \in [t_0, T]$ und $U \in \mathbb{R}^m$ implizit

$$\Psi(t, U; h) := \Phi(t, U, -\bar{\alpha} \cdot U + h\, \Psi(t, U; h); h) \tag{2.23}$$

Mit $U_k := (u_k, \dots, u_{k+m-1})^T$ können wir das Mehrschrittverfahren dann wie folgt schreiben:

$$U_{k+1} = \underbrace{\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{m-1} \end{pmatrix}}_{=:A} U_k + h \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \Psi(t_k, U_k; h) \end{pmatrix} \tag{2.24}$$

Dabei haben wir der Einfachheit halber angenommen, dass $u_k \in \mathbb{R}$ skalar ist. Durch Entwicklung nach der letzten Zeile ergibt sich für das charakteristische Polynom χ_A:

$$\chi_A(\lambda) = \lambda^m + \sum_{i=0}^{m-1} \alpha_i\, \lambda^i = \sum_{i=0}^{m} \alpha_i\, \lambda^i = q(\lambda).$$

Wir benötigen nun folgendes Lemma:

Lemma 2.48.
Zu jeder Matrix $A \in \mathbb{C}^{n\times n}$ und jedem $\epsilon > 0$ existiert eine induzierte Matrixnorm $\|\cdot\|$, sodass gilt:

$$\|A\| \leq |\lambda|_{max}(A) + \epsilon.$$

Falls für alle Eigenwerte λ von A mit $|\lambda| = |\lambda|_{max}(A)$ die geometrische Vielfachheit mit der algebraischen übereinstimmt, so existiert eine induzierte Matrixnorm mit:

$$\|A\| = |\lambda|_{max}(A).$$

Beweis:
Siehe z.B. Stoer, Bulirsch: *Numerische Mathematik 2*, Springer, 2005. ∎

Korollar 2.49.
Falls das Mehrschrittverfahren

$$\sum_{i=0}^{m} \alpha_i \, u_{k+i} = h \, \Phi(t_k, u_k, \ldots, u_{k+m}; h)$$

nullstabil ist, so existiert eine Norm $\|\cdot\|$ auf $\mathbb{C}^m$ (und damit auch auf $\mathbb{R}^m$), sodass für die Matrix A in (2.24) gilt:

$$\|A\,U\| \leq \|U\| \quad \textit{für alle } U \in \mathbb{C}^m.$$

Beweis:
Wir müssen zeigen, dass eine induzierte Matrixnorm existiert, sodass $\|A\| \leq 1$ ist.
Da das Verfahren nullstabil ist, erfüllt das charakteristische Polynom von A die Dahlquist'sche Wurzelbedingung:

- Entweder: Alle Eigenwerte erfüllen $|\lambda| < 1$. Dann existiert eine induzierte Matrixnorm mit
$$\|A\| \leq |\lambda|_{max}(A) + \epsilon < 1.$$
- Oder: Alle Eigenwerte erfüllen $|\lambda| \leq 1$ und Eigenwerte mit $|\lambda| = 1$ haben die algebraische Vielfachheit 1. Wegen $|\lambda|_{max}(A) = 1$ haben alle maximalen Eigenwerte eine Übereinstimmende algebraische und geometrische Vielfachheit und es existiert eine induzierte Matrixnorm mit
$$\|A\| = 1.$$ ∎

Satz 2.50.
Das Mehrschrittverfahren

$$\sum_{i=0}^{m} \alpha_i \, u_{k+i} = h \, \Phi(t_k, u_k, \ldots, u_{k+m}; h)$$

sei konsistent von der Ordnung p und die Verfahrensfunktion Φ sei Lipschitz-stetig in $u_k, \ldots, u_{k+m}$. Dann ist das Verfahren konvergent von der Ordnung p, wenn es nullstabil ist.

Beweis:
Für hinreichend kleines h ist die Funktion Ψ aus (2.23) ebenfalls Lipschitz-stetig:

$$\begin{aligned}
&|\Psi(t, U_1; h) - \Psi(t, U_2; h)| \\
&= \left|\Phi\big(t, U_1, -\bar{\alpha} \cdot U_1 + h\,\Psi(t, U_1; h); h\big) - \Phi\big(t, U_2, -\bar{\alpha} \cdot U_2 + h\,\Psi(t, U_2; h); h\big)\right| \\
&\leq C\,|U_1 - U_2| + C\left|-\bar{\alpha} \cdot (U_1 - U_2) + h\,\big(\Psi(t, U_1; h) - \Psi(t, U_2; h)\big)\right| \\
&\leq C\,|U_1 - U_2| + C\,|\bar{\alpha}|\,|U_1 - U_2| + C\,h\,|\Psi(t, U_1; h) - \Psi(t, U_2; h)|.
\end{aligned}$$

Folglich gilt:

$$|\Psi(t, U_1; h) - \Psi(t, U_2; h)| \leq \frac{C}{1 - C\,h}\,|U_1 - U_2|.$$

Anders geschrieben lautet das Verfahren

$$u_{k+m} = -\bar{\alpha} \cdot U_k + h\,\Psi(t_k, U_k; h).$$

Wir zeigen, dass dieses Verfahren für hinreichend kleines h ebenfalls konsistent von der Ordnung p ist. Für eine exakte Lösung $u \in C^{p+1}(t_0, T)$ gilt:

$$\begin{aligned}
&\big|u(t_{k+m}) + \bar{\alpha} \cdot U(t_k) - h\,\Psi\big(t_k; U(t_k); h\big)\big| \\
&= \big|h\,\Phi\big(t_k, U(t_k), u(t_{k+m}); h\big) - h\,\Phi\big(t_k; U(t_k), -\bar{\alpha} \cdot U(t_k) + h\,\Psi(t_k, U(t_k); h); h\big)\big| + C\,h^{p+1} \\
&\leq C\,h\,\big|u(t_{k+m}) + \bar{\alpha} \cdot U(t_k) - h\,\Psi\big(t_k, U(t_k)\big)\big| + C\,h^{p+1}.
\end{aligned}$$

wobei $U(t_k) := \big(u(t_k), ..., u(t_{k+m-1})\big)^T$ ist.
Nun läuft der Beweis analog zu dem eines Einschrittverfahrens. Sei $E_k := \|U_k - U(t_k)\|$, wobei $\|\cdot\|$ die Norm aus Korollar 2.49 ist. Dann gilt:

$$\begin{aligned}
\|E_{k+1}\| &\leq \|A\,(U_k - U(t_k))\| + h\,\|(0, \ldots, 0, \Psi(t, U_k; h)) - \Psi(t, U(t_k); h)^T\| \\
&\leq E_k + h\,L\,E_k + C\,h^{p+1}.
\end{aligned}$$

Ferner ist $E_0 = \mathcal{O}(h^p)$. Wie im Beweis von Satz 2.16 folgt dann

$$E_k \leq e^{khL}\,E_0 + C\,\frac{e^{khL} - 1}{hL}\,h^{p+1} \leq C\,e^{L(T-t_0)}\,h^p + C\,\frac{e^{L(T-t_0)} - 1}{L}\,h^p.$$

Das Verfahren konvergiert also von der Ordnung p. ∎

Bemerkung 2.51.
Vereinfacht lassen sich Lemma 2.47 und Satz 2.50 in folgender Aussage zusammenfassen:

$$\textbf{\textit{Konsistenz}} + \textbf{\textit{Stabilität}} = \textbf{\textit{Konvergenz}}.$$

Dieser Zusammenhang wird uns häufiger begegnen.

2.7 Stabilität autonomer Differentialgleichungen

Definition 2.52.
Eine Differentialgleichung der Form

$$u' = f(u)$$

heißt ***autonom****. In diesem Fall heißt ein Punkt $p \in \mathbb{R}^n$ mit $f(p) = 0$* ***Gleichgewichtspunkt*** *von $u' = f(u)$. Die Lösung $u \equiv p$ nennt man dann* ***Gleichgewichtslösung*** *oder* ***stationäre Lösung****.*

Bemerkung 2.53.
Lösungen autonomer Differentialgleichungen sind ***translationsinvariant****, d.h. ist u eine Lösung, so auch $u(\cdot + \alpha)$, $\alpha \in \mathbb{R}$.*
Ist $I \subset \mathbb{R}$ und $u : I \to \mathbb{R}^n$ eine Lösung von $u' = f(u)$, so bezeichnet man die Menge

$$u(I) = \big\{u(t) \;\big|\; t \in I\big\} \subset \mathbb{R}^n$$

als zu u gehörige(n) ***Orbit****,* ***Trajektorie*** *oder* ***Phasenkurve****. Ist f Lipschitz-stetig, so sind Orbits entweder disjunkt oder identisch (Picard-Lindelöf).*
Die Gesamtheit aller Phasenkurven heißt das ***Phasenportrait*** *von $u' = f(u)$.*

Starten wir ein numerisches Verfahren in einem Gleichgewichtspunkt, so möchten wir, dass das numerische Verfahren dieses Gleichgewicht erhält. Durch kleine Rechenungenauigkeiten wird dieses jedoch gestört. Wir hoffen jedoch, dass diese kleinen Störungen sich nicht sehr auf das Ergebnis auswirken. Dies ist leider nicht immer der Fall, wie wir am Beispiel des Eulerverfahrens gesehen haben.

2.7.1 Lineare Differentialgleichungen

Eine lineare autonome Differentialgleichung ist von der Form

$$u' = A\,u$$

mit einer Matrix $A \in \mathbb{R}^{n\times n}$. In diesem Fall ist die Menge aller Gleichgewichtspunkte gerade der Kern von A; insbesondere ist 0 immer ein Gleichgewichtspunkt.

Lemma 2.54.
Die Lösung des Anfangswertproblems

$$\begin{aligned} u' &= A\,u \quad \text{in } \mathbb{R}, \\ u(t_0) &= u_0 \end{aligned}$$

mit $A \in \mathbb{R}^{n\times n}$ lautet

$$u(t) = e^{A(t-t_0)}\,u_0,$$

wobei e^A das Exponential einer Matrix $A \in \mathbb{K}^{n\times n}$, $\mathbb{K} = \mathbb{R}, \mathbb{C}$, bezeichnet:

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!}\,A^k.$$

Dabei ist $A^0 = \mathbb{I}$.

Beweis:
Übungsaufgabe. ∎

Bemerkung 2.55.
Ist $A, B, T \in \mathbb{K}^{n\times n}$, $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Dann gilt:

1. *Ist A blockdiagonal, d.h.*

$$A = \operatorname{diag}(A_1, \ldots, A_l) = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_l \end{pmatrix}$$

   mit $A_k \in \mathbb{K}^{n_k\times n_k}$, so gilt:

$$e^A = \operatorname{diag}(e^{A_1}, \ldots, e^{A_l}).$$

2. *Falls $A\,B = B\,A$, so gilt $e^{A+B} = e^A\,e^B$.*
3. *e^A ist invertierbar und es gilt $(e^A)^{-1} = e^{-A}$.*
4. *Ist $A = T\,B\,T^{-1}$, so gilt $e^A = T\,e^B\,T^{-1}$.*

Lemma 2.56.
Ist $A \in \mathbb{C}^{n\times n}$ ein Jordanblock mit Eigenwert λ, so gilt:

$$e^{At} = e^{\lambda t} \begin{pmatrix} 1 & t & \cdots & \cdots & \frac{t^{n-2}}{(n-2)!} & \frac{t^{n-1}}{(n-1)!} \\ & 1 & t & \cdots & \ddots & \frac{t^{n-2}}{(n-2)!} \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & t & \vdots \\ & & & & 1 & t \\ & & & & & 1 \end{pmatrix}.$$

Beweis:
Sei A also ein Jordanblock mit Eigenwert λ, d.h. $A = \lambda\mathbb{I} + N$ mit

$$N = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

N ist nilpotent mit $N^n = 0$. Allgemeiner ist N^k von der Form:

$$N^k = \begin{pmatrix} 0 & \cdots & 0 & 1 & & \\ & \ddots & & \ddots & \ddots & \\ & & \ddots & & \ddots & 1 \\ & & & \ddots & & 0 \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{pmatrix}.$$

Wegen $\lambda\,\mathbb{I}\,N = N\,\lambda\,\mathbb{I}$ gilt:

$$e^A = e^{\lambda\mathbb{I}+N} = e^{\lambda}\,\mathbb{I}\,e^N \quad \text{bzw.} \quad e^{At} = e^{\lambda t}\,\mathbb{I}\,e^{Nt}.$$

Wir müssen also nur noch e^{Nt} berechnen. Es gilt:

$$e^{Nt} = \sum_{k=0}^{n-1} \frac{t^k}{k!} N^k = \begin{pmatrix} 1 & t & \cdots & \cdots & \frac{t^{n-2}}{(n-2)!} & \frac{t^{n-1}}{(n-1)!} \\ & 1 & t & \cdots & \ddots & \frac{t^{n-2}}{(n-2)!} \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & t & \vdots \\ & & & & 1 & t \\ & & & & & 1 \end{pmatrix}.$$

Damit ist die Behauptung gezeigt. ∎

Definition 2.57 (Lyapunov, ca. 1890).
Wir nennen den Gleichgewichtspunkt $0 \in \mathbb{R}^n$ von $u' = A\,u$

1. ***stabil**, wenn alle Lösungen von $u' = A\,u$ beschränkt bleiben (d.h. $\|u(t)\| \le C$).*

2. ***asymptotisch stabil**, wenn für jede Lösung von $u' = A\,u$ gilt: $u(t) \to 0$.*

3. ***neutral stabil**, falls er stabil ist, aber nicht asymptotisch stabil.*

4. ***instabil**, falls es eine unbeschränkte Lösung gibt.*

Satz 2.58 (Eigenwerttest auf Stabilität).
Sei $A \in \mathbb{R}^{n\times n}$ mit Eigenwerten $\lambda_1, \dots, \lambda_l \in \mathbb{C}$, algebraischen Vielfachheiten $a_1, \dots, a_l$ und geometrischen Vielfachheiten $g_1, \dots, g_l$. Dann gilt für $u' = A\,u$:

1. *0 ist asymptotisch stabil $\Leftrightarrow \operatorname{Re}\lambda_j < 0$ für alle $j = 1, \dots, l$.*

2. *0 ist neutral stabil $\Leftrightarrow \operatorname{Re}\lambda_j \le 0$, $g_j = a_j$ im Fall $\operatorname{Re}\lambda_j = 0$ und für mindestes ein j gilt $\operatorname{Re}\lambda_j = 0$.*

3. *0 ist instabil $\Leftrightarrow$ für mindestens ein j gilt $\operatorname{Re}\lambda_j > 0$ oder $\operatorname{Re}\lambda_j = 0$ und $g_j < a_j$.*

Beweis:
Bringen wir A in die Jordan-Normalform $A = T\,J\,T^{-1}$, wobei T die Matrix aus Eigenvektoren (Hauptvektoren) und $J = \operatorname{diag}(J_1, \dots, J_l)$ mit Jordanblöcken $J_1, \dots, J_l$ ist. Wegen

$$e^{At} = T\,\operatorname{diag}(e^{J_1 t}, \dots, e^{J_l t})\,T^{-1}$$

können wir o.E. annehmen, dass A bereits ein Jordanblock mit Eigenwert λ ist.
Ist $\operatorname{Re}\lambda < 0$, so gilt

$$e^{At} = e^{\lambda t}\,e^{Nt} = e^{\lambda t}\,q(t)$$

mit $q \in \mathbb{P}_n$. Für $t \to \infty$ folgt also $e^{At} \to 0$ (und damit $e^{At}\,u_0 \to 0$ für alle $u_0 \in \mathbb{R}^n$). Das Gleichgewicht ist also asymptotisch stabil.
Ist hingegen $\operatorname{Re}\lambda = 0$, so folgt $|e^{\lambda t}| = 1$ für alle $t \in \mathbb{R}$. Folglich ist $e^{At}\,e_n = e^{\lambda t}\,e_n$ lediglich beschränkt. Ferner gilt:

$$e^{At}\,(e_{n-1} + e_n) = e^{\lambda t}\,\big((1+t)\,e_{n-1} + e_n\big).$$

Diese Lösung ist unbeschränkt für $t \to \infty$, d.h. das Gleichgewicht ist neutralstabil genau dann, wenn $n = 1$ ist. Dies ist genau dann der Fall, wenn die algebraische Vielfachheit des Eigenwertes gleich der geometrischen ist.
Ist $\operatorname{Re}\lambda > 0$, so ist offenbar bereits $e^{\lambda t}$ unbeschränkt für $t \to \infty$, sodass das Gleichgewicht instabil ist.
Da die Fälle sich gegenseitig ausschließen muss auch die Umkehrung gelten. Ist etwa $\operatorname{Re}\lambda \ge 0$, so kann das system nur neutral stabil oder instabil sein. Dies schließt aber asymptotische Stabilität aus. ∎

2.7.2 Nichtlineare Differentialgleichungen

Definition 2.59.
Das System $u' = f(u)$ heißt

1. ***stabil** an einem Gleichgewichtspunkt p, falls zu jedem $\epsilon > 0$ ein $\delta > 0$ existiert, sodass für jede Lösung u des Systems gilt:*

$$\big\|u(0) - p\big\| < \delta \quad \Rightarrow \quad \forall t \ge 0 : \big\|u(t) - p\big\| < \epsilon.$$

Anschaulich bedeutet dies, dass u für alle Zeiten „in der Nähe" von p bleibt.

2. ***asymptotisch stabil*** *an einem Gleichgewichtspunkt p, falls es stabil bei p ist und ein $\delta > 0$ existiert, sodass*

$$\big\|u(0) - p\big\| < \delta \quad \Rightarrow \quad \big\|u(t) - p\big\| \overset{t\to\infty}{\longrightarrow} 0.$$

3. ***neutral stabil*** *an einem Gleichgewichtspunkt p, falls es stabil bei p ist, aber nicht asymptotisch stabil.*

4. ***instabil*** *an einem Gleichgewichtspunkt p, falls es nicht stabil ist.*

Satz 2.60.
Sei f zweifach stetig differenzierbar auf $\mathbb{R}^n$ und sei p ein Gleichgewichtspunkt von $u' = f(u)$, d.h. $f(p) = 0$. $DF(u)$ bezeichne die Jacobi-Matrix von f an der Stelle u. Dann ist das System bei p

1. *asymptotisch stabil, falls alle Eigenwerte von $DF(p)$ negative Realteile haben.*

2. *instabil, falls mindestens ein Eigenwert von $DF(p)$ positiven Realteil hat.*

Die analogen Aussagen für verschwindende Eigenwerte in Satz 2.58 übertragen sich im allgemeinen nicht.

2.8 Steife Differentialgleichungen

2.8.1 Exkurs: Reaktionskinetik

Betrachten wir exemplarisch den Zerfall von Ozon (O_3) zu Sauerstoff:

$$2O_3 \rightleftharpoons 3O_2$$

Kennen wir die Konzentration $u_2 := [O_2]$ und $u_3 := [O_3]$ zum Zeitpunkt t_0, so würden wir die chemische Reaktion gerne als Differentialgleichung schreiben:

$$u_2' = f_2(u_2, u_3),$$
$$u_3' = f_3(u_2, u_3).$$

Da O_3-Moleküle in Abwesenheit von O_2 aber nicht miteinander reagieren, können wir der Reaktionsgleichung nicht ansehen, von welcher Form die Funktionen f_2 und f_3 sind. Wir müssten sie experimentell bestimmen.

Ein weiterer Modellierungsschritt, der Begriff der **Elementarreaktion**, erlaubt uns aber, das Problem weiter aufzuspalten. Eine Elementarreaktion ist eine chemische Reaktion, die immer stattfindet, wenn notwendigen Moleküle aufeinandertreffen. In unserem Fall lauten die Elementarreaktionen:

$$O_2 + O_3 \rightleftharpoons O + 2O_2$$
$$O_3 + O \longrightarrow 2O_2$$

Man bezeichnet die Zerlegung in Elementarreaktionen auch als **Reaktionsmechanismus**.

Die Wahrscheinlichkeit, dass passende Moleküle (auf passende Art) aufeinandertreffen ist nun proportional zum Produkt ihrer Konzentrationen ($[\frac{mol}{l}]$). Bezeichnen wir $u_1 := [O]$, so ist die Geschwindigkeit der Elementarreaktion $O_2 + O_3 \longrightarrow O + 2O_2$ gerade

$$v_1 = k_1\, u_2\, u_3.$$

Dabei wird k_1 als **Reaktionsrate** bezeichnet. Da hierbei genau jeweils ein O_2 und ein O_3 Molekül verbraucht werden und ein O und zwei O_2 Moleküle entstehen, lässt sich die Einzelreaktion wie folgt bilanzieren:

$$\begin{aligned} u_1' &= 1\, v_1, \\ u_2' &= (2-1)\, v_1, \\ u_3' &= (-1)\, v_1 \end{aligned}$$

Dabei hängen die Vorfaktoren gerade von den **stöchiometrischen** Koeffizienten ν_i' der Edukte und ν_i'' der Produkte ab:

i	Reaktionsgleichung	k_i	ν_i'	ν_i''
1	$O_2 + O_3 \longrightarrow O + 2O_2$	$2.01 \cdot 10^{11}\ \exp(\frac{-24000}{RT})\ \frac{l}{mol\, s}$	$(0,1,1)$	$(1,2,0)$
2	$O + 2O_2 \longrightarrow O_2 + O_3$	$2.96 \cdot 10^{7}\ \exp(\frac{890}{RT})\ \frac{l^2}{mol^2\, s}$	$(1,2,0)$	$(0,1,1)$
3	$O_3 + O \longrightarrow 2O_2$	$3.37 \cdot 10^{10}\ \exp(\frac{-5700}{RT})\ \frac{l}{mol\, s}$	$(1,0,1)$	$(0,2,0)$

Quelle: Benson, Axworthy, *Reconsideration of the Rate Constants from the Thermal Decomposition of Ozone*, Journal of Chemical Physics 42, 1965
Bemerkung: k_1 abgeleitet aus $\frac{k_1}{k_2} = 6.8 \cdot 10^4\ \exp(\frac{-24890}{RT})$

Dabei ist $R = 8.314\,462\ \frac{J}{mol\, K}$ die allgemeine Gaskonstante. Bei $25°C = 298.15K$ gilt also

$$k_1 \approx 1.25 \cdot 10^7\ \tfrac{mol}{l\, s}, \qquad k_2 \approx 4.24 \cdot 10^7\ \tfrac{mol}{l\, s}, \qquad k_3 \approx 3.38 \cdot 10^9\ \tfrac{mol}{l\, s}.$$

Allgemein gilt für eine Elementarreaktion:

$$v_i = k_i \prod_{k=1}^{3} u_k^{\nu_{ik}'}$$

Die Differentialgleichung für den gesamten Reaktionsmechanismus lautet dann:

$$u_j' = \sum_{i=1}^{3} \left(\nu_{ij}'' - \nu_{ij}'\right) v_i.$$

Die gesamte Differentialgleichung lautet also

$$\begin{aligned} u_1' &= k_1\, u_2\, u_3 - k_2\, u_1\, u_2^2 - k_3\, u_1\, u_3, \\ u_2' &= k_1\, u_2\, u_3 - k_2\, u_1\, u_2^2 + 2\, k_3\, u_1\, u_3, \\ u_3' &= -k_1\, u_2\, u_3 + k_2\, u_1\, u_2^2 - k_3\, u_1\, u_3. \end{aligned}$$

Die Lösung für $u_0 = \frac{1}{23}\,(0, 0.99, 0.01)$ ist in Abbildung 5 dargestellt. Man sieht, dass die Reaktion auf unterschiedlichen Skalen abläuft: Das Gleichgewicht zwischen O_3 und O stellt sich nach kurzer Zeit ein und danach passiert fast nichts mehr. Dies ist typisch für chemische Reaktionen.

2.8.2 Steife Probleme

Um dieses Problem besser zu verstehen betrachten wir ein einfaches Beispiel:

$$\begin{aligned} u_1' &= \frac{\lambda_1 + \lambda_2}{2}\, u_1 + \frac{\lambda_1 - \lambda_2}{2}\, u_2, \\ u_2' &= \frac{\lambda_1 - \lambda_2}{2}\, u_1 + \frac{\lambda_1 + \lambda_2}{2}\, u_2 \end{aligned}$$

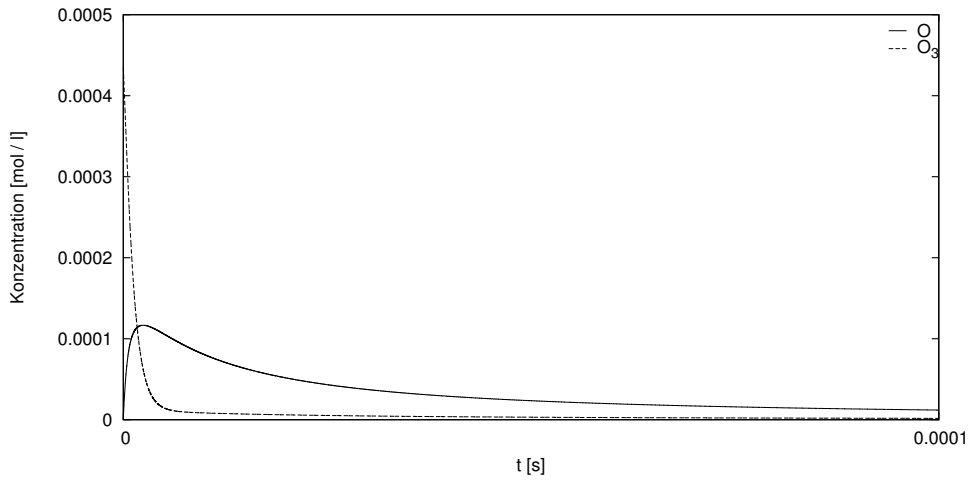

Abbildung 5: Lösung des Ozonzerfalls für $u_0 = \frac{1}{23}(0, 0.99, 0.01)$

mit $\lambda_1, \lambda_2 < 0$. Die allgemeine Lösung dieser Differentialgleichung lautet

$$u_1(t) = C_1\, e^{\lambda_1 t} + C_2\, e^{\lambda_2 t},$$
$$u_2(t) = C_1\, e^{\lambda_1 t} - C_2\, e^{\lambda_2 t}.$$

Ist $\lambda_1 = -1$ und $\lambda_2 = -1000$, so haben wir ein ähnliches Verhalten wie bei dem Beispiel aus der Reaktionskinetik: Der Term $C_2\, e^{-1000\, t}$ sorgt anfänglich für eine starke Änderung der Lösung, danach passiert fast nichts mehr.
Die numerische Lösung aus dem Euler-Verfahren lautet dann

$$u_{1,k} = C_1\,(1 + \lambda_1\, h)^k + C_2\,(1 + \lambda_2\, h)^k,$$
$$u_{2,k} = C_1\,(1 + \lambda_1\, h)^k - C_2\,(1 + \lambda_2\, h)^k$$

Damit keine Oszillationen auftreten muss die Zeitschrittweite hinreichend klein sein:

$$h \leq -\frac{1}{\min\{\lambda_1, \lambda_2\}}.$$

Im Fall $\lambda_1 = -1$ und $\lambda_2 = -1000$, so ist $h \leq \frac{1}{1000}$. Dies gilt unabhängig von C_2. Obwohl also $C_2\, e^{-1000\, t}$ keinen nennenswerten Beitrag zur Lösung liefert, dominiert doch seine Approximation die Zeitschrittweite. Dieses Verhalten bezeichnen wir als **steif**.

2.8.3 Absolute Stabilität

Ziel: Das numerische Verfahren soll die Stabilität von Gleichgewichten erhalten.
Um dieses Problem zu studieren, betrachten wir wieder eine lineare Differentialgleichung der Form

$$u' = A\, u$$

und nehmen an, dass alle Eigenwerte von A negativen Realteil haben. In diesem Fall ist 0 asymptotisch stabiler Gleichgewichtspunkt und wir erwarten, dass für die numersiche Lösung gilt:

$$u_k \overset{k\to\infty}{\longrightarrow} 0.$$

Definition 2.61 (Stabilitätsfunktion).
Lässt sich das Einschrittverfahren

$$u_{k+1} = u_k + h_k \, \Phi(t_k, u_k; h_k)$$

angewandt auf die Differentialgleichung $u' = A\,u$ in der Form

$$u_{k+1} = R(h\,A)\,u_k$$

mit einer rationalen Funktion R schreiben, so heißt R Stabilitätsfunktion des Verfahrens.

Beispiel 2.62.

1. *explizites Euler-Verfahren:*

$$u_{k+1} = u_k + h\,A\,u_k = (\mathbb{I} + h\,A)\,u_k.$$

Die Stabilitätsfunktion lautet also $R(z) = 1 + z$.

2. *implizites Euler-Verfahren:*

$$u_{k+1} = u_k + h\,A\,u_{k+1} \quad \Leftrightarrow \quad (\mathbb{I} - h\,A)\,u_{k+1} = u_k \quad \Leftrightarrow \quad u_{k+1} = (\mathbb{I} - h\,A)^{-1}\,u_k$$

Die Stabilitätsfunktion lautet dann $R(z) = \frac{1}{1-z}$.

Satz 2.63.
Ist ein explizites Runge-Kutta Verfahren konsistent von der Ordnung p, so gilt für die Stabilitätsfunktion

$$R(z) = \sum_{k=0}^{p} \frac{z^k}{k!} + \mathcal{O}(z^{p+1}).$$

Beweis:
Siehe z.B. Hairer, Wanner: *Solving Ordinary Differential Equations II.* ∎

Mit Hilfe der Stabilitätsfunktion können wir das numerische Verfahren wie folgt darstellen:

$$u_k = R(h\,A)^k\,u_0.$$

Ist $R(h\,A)^m < 1$ für ein $m \geq 1$, so haben wir die Konvergenz $u_k \to 0$ für $k \to \infty$. Da R rational ist, und alle Eigenwerte $\lambda_1, \ldots, \lambda_l$ von A negativen Realteil haben ist dies erfüllt falls $R(h\,\lambda_j) < 1$, $j = 1, \ldots, l$ ist.
Dies motiviert folgende Definition:

Definition 2.64 (absolute Stabilität).
Ein Einschrittverfahren

$$u_{k+1} = u_k + h_k \, \Phi(t_k, u_k; h_k)$$

mit Stabilitätsfunktion R heißt absolut stabil, falls

$$|R(z)| < 1 \quad \textit{für alle } z \in \mathbb{C} \textit{ mit } Re(z) < 0.$$

Zur Lösung steifer Differentialgleichungen sollte man also ein absolut stabiles Verfahren verwenden.

Beispiel 2.65.

1. *explizites Euler-Verfahren:* $R(z) = 1 + z$.

$$|R(z)| < 1 \quad \Leftrightarrow \quad z \in B_1(-1).$$

Das Verfahren ist nur bedingt stabil.

2. *implizites Euler-Verfahren:* $R(z) = \frac{1}{1-z}$.

$$|R(z)| = \frac{1}{|1-z|} < 1 \quad \Leftrightarrow \quad |1-z| > 1.$$

Dies ist für alle z mit negativem Realteil erfüllt; das Verfahren ist absolut stabil.

Bemerkung 2.66.
Es gibt kein absolut stabiles explizites Verfahren.

2.9 Randwertprobleme

Bisher haben wir Anfangswertprobleme betrachtet:

$$\begin{aligned} u' &= f(\cdot, u) \quad \text{für } t \in I = [a, b], \\ u(t_0) &= u_0. \end{aligned}$$

Aber: Wenn $u = (v, w) \in \mathbb{R}^2$ ist, warum sollten wir dann nicht den Wert von v in a und den Wert von w in b vorschreiben können?
Allgemein können wir offenbar Randbedingungen der Form

$$g\big(u(a), u(b)\big) = 0$$

fordern mit einer Funktion $g : \mathbb{R}^{2n} \to \mathbb{R}^n$. Ist $g(u_a, u_b) = A\,u_a + B\,u_b - c$ mit $A, B \in \mathbb{R}^{n\times n}$ und $c \in \mathbb{R}^n$, so heißen die Randbedingungen (affin-)**linear**. Anfangswerte sind dann der Spezialfall

$$g(u_a, u_b) = u_a - u_0$$

Während Anfangswertprobleme zumindest lokal eindeutig lösbar sind, ist dies für Randwertprobleme nicht immer der Fall:

Beispiel 2.67.
Betrachten wir die Differentialgleichung

$$\begin{aligned} u_1' &= u_2, \\ u_2' &= -u_1. \end{aligned}$$

Die allgemeine Lösung lautet

$$\begin{aligned} u_1(t) &= c_1 \sin t + c_2 \cos t, \\ u_2(t) &= c_1 \cos t - c_2 \sin t. \end{aligned}$$

Schreiben wir nun die Randbedingung

$$u_1(0) = 0, \quad u_2(\tfrac{\pi}{2}) = c$$

vor, so haben wir im Falle

- $c = 0$ *beliebig viele Lösungen der Form* $u_1(t) = c_1 \sin t$, $u_2(t) = c_1 \cos t$,
- $c \neq 1$ *gar keine Lösung.*

Wählen wir hingegen die Randbedingung

$$u_1(0) = 0, \quad u_2(\pi) = -c,$$

so gibt es genau die Lösung: $u_1(t) = c \sin t$, $u_2(t) = c \cos t$.

Beispiel 2.68 (Randwertprobleme mit freiem Rand).
Finde ein u *mit*

$$u'(t) = f(t, u(t)) \quad \text{für } t \in [a, b]$$

und passenden Randbedingungen, wobei der Randpunkt b *selbst eine Unbekannte des Problems ist.*
Um dieses Problem zu lösen, transformieren wir das Problem zunächst auf das feste Intervall $I = [0, 1]$, *indem wir* $t = a\,(1 - s) + b\,s$, $s \in I$, *und* $w(s) := u(t)$ *setzen. Dann gilt:*

$$w'(s) = (b - a)\,u'(t) = (b - a)\,f\big(t, u(t)\big) = (b - a)\,f\big(a + (b - a)\,s, w(s)\big).$$

Setzen wir nun $z := (b - a)$, *so erfüllt* (w, z) *die Differentialgleichung*

$$\begin{aligned} w'(s) &= z\,f\big(a + z\,s, w(s)\big), \\ z'(s) &= 0. \end{aligned}$$

In diesem Fall benötigen wir $n + 1$ *Randbedingungen, d.h. für ein* $g : \mathbb{R}^{2n} \to \mathbb{R}^{n+1}$ *muss gelten:*

$$g\big(u(a), u(b)\big) = 0.$$

Das einfache Schießverfahren

Ziel: Löse das Randwertproblem

$$u' = f(\cdot, u), \quad g(u(a), u(b)) = 0.$$

Idee: Wir können bereits Anfangswertprobleme

$$u' = f(\cdot, u), \quad u(a) = u_a$$

lösen. Dies definiert uns eine Abbildung $F : \mathbb{R}^n \to \mathbb{R}^n : u_a \mapsto u(b)$. Die Randbedingung ist dann äquivalent zu

$$G(u_a) := g\big(u_a, F(u_a)\big) = 0.$$

Wir suchen also eine Nullstelle der Funktion G. Numerisch können wir hierzu jedes Verfahren zur Nullstellenbestimmung verwenden, etwa das Newton-Verfahren. Allerdings gilt:

$$\mathbf{D}G(z) = \mathbf{D}_{u_a} g\big(z, F(z)\big) + \mathbf{D}_{u_b} g\big(z, F(z)\big)\,\mathbf{D}F.$$

Um eine Richtungsableitung $\partial_\xi F(u_a)$, $\xi \in \mathbb{R}^n$, zu bestimmen verwenden wir die Lösung u zu Anfangsdaten u_a und die Lösung w_h zu Anfangsdaten $u_a + h\,\xi$. Dann gilt:

$$w_h' - u' = f(\cdot, w_h) - f(\cdot, u_h) = \mathbf{D}f(\cdot, u)\,(w_h - u) + \mathbf{o}(|w_h - u|).$$

Aufgrund der a-priori Abschätzung gilt $\mathbf{o}(|w_h - u|) = \mathbf{o}(|h|)$. Setzen wir nun $\Psi_h := \frac{w_h - u}{|h|}$, so erfüllt Ψ_h die Differentialgleichung

$$\Psi_h' = \mathbf{D}f(\cdot, u)\,\Psi_h + \mathbf{o}(1).$$

Für die Richtungsableitung gilt also $\partial_\xi F(u_a) = \Psi_0(b)$, wobei Ψ Lösung des linearen Anfangswertproblems

$$\begin{aligned} \Psi' &= \mathbf{D}f(\cdot, u)\,\Psi, \\ \Psi(a) &= \xi \end{aligned}$$

ist.

3 Differentialgleichungen höherer Ordnung

Sei $n \in \mathbb{N}$, sei $G \subset \mathbb{R}^{n+1}$ ein Gebiet und sei $f \in C^0(G)$. Eine (skalare) **Differentialgleichung n-ter Ordnung** (in expliziter Form) ist von der Gestalt

$$u^{(n)} = f(\cdot, u, u', \ldots, u^{(n-1)}). \tag{3.1}$$

Eine Funktion $u \in C^n(I)$ mit einem Intervall $I \subset \mathbb{R}$ und $\big(t, u(t), u'(t), \ldots, u^{(n-1)}(t)\big) \in G$ für alle $t \in I$ heißt **Lösung der Differentialgleichung**, falls gilt:

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

Durch die Festlegung $u_1 = u$, $u_2 = u'$, …, $u_n = u^{(n-1)}$ kann eine Differentialgleichung n-ter Ordnung in ein System

$$\begin{aligned} u_1' &= u_2, \\ &\vdots \\ u_{n-1}' &= u_n, \\ u_n' &= f(\cdot, u_1, \ldots, u_n) \end{aligned}$$

von Differentialgleichungen erster Ordnung überführt werden. Damit können wir alle Verfahren für Differentialgleichungen erster Ordnung anwenden.

Bei Differentialgleichungen höherer Ordnung hat man es häufig mit Randwertproblemen zu tun. Wir haben schon gesehen, dass sich Verfahren für Anfangswertprobleme hier nur beschränkt eignen. Daher soll hier eine weitere Diskretisierungstechnik, die Finiten Differenzen, vorgestellt werden.

Wir betrachten exemplarisch das lineare Randwertproblem

$$\begin{aligned} -u''(x) + q(x)\,u(x) &= f(x) \quad \text{in } (a, b), \\ u(a) &= u_a, \\ u(b) &= u_b, \end{aligned} \tag{3.2}$$

mit Funktionen $q, f \in C^0([a, b])$, $q \geq 0$.

3.1 Das Maximumprinzip

Um die Notation zu vereinfachen führen wir den folgenden Differentialoperator ein:

$$L\,u(x) := -u''(x) + q(x)\,u(x), \tag{3.3}$$

Satz 3.1 (schwaches Maximumprinzip).
Sei L wie in (3.3) und sei $u \in C^2((a,b)) \cap C^0([a,b])$ mit

$$L\,u = f \leq 0 \quad \text{in } [a,b].$$

1. *Ist $q = 0$, so nimmt u sein Maximum auf dem Rand an, d.h.*

$$\max_{x\in[a,b]} u(x) = \max\big\{u(a), u(b)\big\}.$$

2. *Ist $q \geq 0$, so gilt:*

$$\max_{x\in[a,b]} u(x) = \max\big\{u^+(a), u^+(b)\big\},$$

wobei $\alpha^+ = \max\{\alpha, 0\}$ der Positivteil von $\alpha \in \mathbb{R}$ ist.

Beweis:
Betrachten wir zunächst den Fall $q = 0$. Nehmen wir an, dass u sein Maximum in $\bar{x} \in [a,b]$ annimmt. Dann gilt $f(\bar{x}) = 0$, denn:

$$0 \geq f(\bar{x}) = -u''(\bar{x}) \geq 0.$$

Zu $\delta > 0$ definieren wir eine Funktion u_δ durch

$$u_\delta(x) := u(x) + \frac{\delta}{2}\,(x - \bar{x})^2.$$

Dann gilt:

$$L\,(u_\delta)(x) = -u_\delta''(x) = -u''(x) - \delta = f(x) - \delta =: f_\delta(x) < 0.$$

Wäre nun $u(\bar{x}) > \max\big\{u(a), u(b)\big\}$, so gäbe es ein $\delta > 0$ sodass u_δ sein Maximum in $x_\delta \in (a,b)$ annimmt. Wir haben aber gezeigt, dass dann $f_\delta(x_\delta) = 0$ wäre. Dies ist ein Widerspruch.
Im Fall $q \geq 0$ betrachten wir ein Intervall $(\tilde{a}, \tilde{b}) \subset (a,b)$ mit $u > 0$ in $(\tilde{a}, \tilde{b})$. Dann gilt in $(\tilde{a}, \tilde{b})$:

$$-u'' = f - q\,u =: \tilde{f} \leq 0.$$

Also ist

$$\max_{x\in[\tilde{a},\tilde{b}]} u(x) \leq \max\big\{u(\tilde{a}), u(\tilde{b})\big\} \leq \max\big\{u(a), u(b), 0\big\}. \qquad \blacksquare$$

Korollar 3.2 (Eindeutigkeit für das Randwertproblem).
Sei L wie in (3.3) und seien $u, v \in C^2((a,b)) \cap C^0([a,b])$ mit

$$L\,u = L\,v \quad \text{in } (a,b).$$

Dann gilt:

$$\max_{x\in[a,b]} |u - v| \leq \max_{x\in\{a,b\}} \big|u(x) - v(x)\big|.$$

Insbesondere ist die Lösung des Randwertproblems (3.2) eindeutig.

Beweis:
Die Differenz $w := u - v$ erfüllt $L\,w = 0$. Aus dem Maximumprinzip folgt:

$$\max_{x\in[a,b]}\big(u(x)-v(x)\big) = \max_{x\in[a,b]} w(x) \le \max_{x\in\{a,b\}} w^+(x) \le \max_{x\in\{a,b\}}\big|w(x)\big| = \max_{x\in\{a,b\}}\big|u(x)-v(x)\big|$$

Die Aussage folgt durch Vertauschen der Rollen von u und v unter Ausnutzung der Definition $|\alpha| = \max\{\alpha, -\alpha\}$, $\alpha \in \mathbb{R}$. ∎

Bemerkung 3.3.
Beispiel 2.67 lässt sich umschreiben zu

$$-u''(x) + q(x)u(x) = 0$$

mit $q(x) = -1$. Wir haben gezeigt, dass jenach Wahl der Randwerte kein, eine oder unendlich viele Lösungen existieren. Dies ist kein Widerspruch zu Korollar 3.2, da q negativ ist.

3.2 Finite Differenzen-Verfahren

Wir diskretisieren dieses Randwertproblem, indem wir das Intervall (a, b) gleichmäßig in $N \in \mathbb{N}$ Teilintervalle (x_{i-1}, x_i) mit $x_i = a + h\,i$ und $h = \frac{b-a}{N}$ einteilen. Wir bezeichnen mit u_i, $i = 0, \dots, N$, die approximativen Werte von $u(x_i)$. Nun diskretisieren wir den Operator L wie folgt:

$$\begin{aligned} L_h\,u_i := -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + q(x_i)\,u_i = f(x_i) \quad \text{für } i = 1, \dots, N-1, \\ u_0 = u_a, \\ u_N = u_b. \end{aligned} \tag{3.4}$$

Lemma 3.4.
Die Approximation (3.4) *ist konsistent von zweiter Ordnung, d.h. für $u \in C^4([a, b])$ gilt mit $u_i = u(x_i)$:*

$$L\,u(x) - L_h u_i = \mathcal{O}(h^2) \quad \textit{für } i = 1, \dots, N-1.$$

Beweis:
Übungsaufgabe. ∎

Satz 3.5 (Diskretes Maximumprinzip).
Sei u_i, $i = 0, \dots, N$ eine Lösung des diskreten Problems

$$L_h\,u_i \le 0 \quad \textit{für } i = 1, \dots, N-1.$$

Dann gilt:

$$\max_{i=0,\dots,N} u_i \le \max\{u_0^+, u_N^+\}.$$

Beweis:
Nehmen wir an, es gibt ein $0 < k < N$ mit $u_k = \max_{i=0,\dots,N} u_i \ge 0$. Andernfalls ist die Aussage bewiesen. Dann gilt:

$$0 \ge L_h\,u_k = -\frac{u_{k+1} - 2\,u_k + u_{k-1}}{h^2} + \underbrace{q(x_k)\,u_k}_{\ge 0} \ge \frac{1}{h^2}\left(2\,u_i - \big(\underbrace{u_{i+1} + u_{i-1}}_{\le 2u_i}\big)\right) \ge 0.$$

Folglich gilt überall Gleichheit und damit:

$$u_k = u_{k+1} = u_{k-1}.$$

Induktiv folgt, dass $u_0 = u_1 = \ldots = u_{N-1} = u_N$ ist. Wegen $u_k \geq 0$ folgt die Aussage. ∎

Korollar 3.6 (Existenz und Eindeutigkeit der diskreten Lösung).
Es existiert genau eine Lösung u_i, $i = 0, \ldots, N$ des diskreten Randwertproblems (3.4).

Beweis:
Seien u_i, v_i, $i = 0, \ldots, N$ zwei Lösungen von (3.4). Dann gilt:

$$\begin{aligned} L_h\,(u-v)_i &= f(x_i) - f(x_i) = 0 \quad \text{für } i = 1, \ldots, N-1, \\ (u-v)_0 &= u_a - u_a = 0, \\ (u-v)_N &= u_b - u_b = 0. \end{aligned}$$

Aus dem diskreten Maximumprinzip folgt dann

$$\max_{i=0,\ldots,N} (u-v)_i \leq \max\big\{(u-v)_0^+, (u-v)_N^+\big\} = 0.$$

Also ist $u_i \leq v_i$ für $i = 0, \ldots, N$. Gleichheit folgt, wenn man die Rollen von u_i und v_i vertauscht.
Wir haben also ein quadratisches Gleichungssystem (für jeden Gitterpunkt gibt es eine Gleichung), das eindeutig lösbar ist, also vollen Rang hat. Aus der Dimensionsformel folgt dann die Existenz einer Lösung. ∎

Lemma 3.7 (Stabilität).
Die Diskretisierung (3.4) *ist* ***stabil****, d.h.*

$$\max_{i=0,\ldots,N} |u_i| \leq C\big(\max_{i=1,\ldots,N-1} |L_h\,u_i| + \max\{u_0, u_N\}\big) \quad \textit{für alle } (u_i)_{i=0,\ldots,N}$$

mit einer von h unabhängigen Konstanten $C \geq 0$.

Beweis:
Sei $r = \frac{1}{2}\,(b-a)$ und $m := \frac{1}{2}\,(a+b)$. Dann gilt für $w(x) := r^2 - \frac{1}{2}\,(x-m)^2$:

$$L\,w(x) = -w''(x) + q(x)\,w(x) = 1 + \underbrace{q(x)\,w(x)}_{\geq 0} \geq 1.$$

Setzen wir $w_i := w(x_i)$, so folgt aus der Konsistenz von L_h für hinreichend kleine h:

$$|L_h\,w_i - L\,w(x)| \leq 1 \quad \text{für } i = 1, \ldots, N-1.$$

Seien $(u_i)_{i=0,\ldots,N}$ fest gewählt und setze

$$v_i := \max\big\{|u_0|, |u_N|\big\} + \frac{w_i}{2} \max_{j=1,\ldots,N-1} |L_h\,u_j|.$$

Dann ist $L_h\,(u-v)_i \leq 0$, $i = 1, \ldots, N-1$ und $(u-v)_0, (u-v)_N \leq 0$. Aus dem diskreten Maximumprinzip folgt $u_i - v_i \leq 0$, $i = 0, \ldots, N$. Analoges gilt für $-u_i - v_i$ und wir erhalten für $i = 0, \ldots, N$:

$$|u_i| \leq v_i \leq \max\big\{|u_0|, |u_N|\big\} + \frac{r^2}{2} \max_{j=1,\ldots,N-1} |L_h\,u_j|.$$ ∎

Satz 3.8.
Die Diskretisierung (3.4) *ist von zweiter Ordnung konvergent, d.h. ist* $u \in C^4([a,b])$ *die exakte Lösung von* (3.2) *und* $(u_i^h)_{i=0,\dots,N}$ *die Lösung von* (3.4)*, so gilt:*

$$\max_{i=0,\dots,N} \left| u_i^h - u(x_i) \right| = \mathcal{O}(h^2).$$

Beweis:
Aus der Stabilität folgt (mit $u_i = u(x_i)$):

$$\max_{i=0,\dots,N} |u_i^h - u_i| \leq C\Big(\max_{i=1,\dots,N-1} |L_h\, u_i^h - L_h\, u_i| + \underbrace{\max\big\{ |u_0^h - u_0|, |u_N^h - u_N| \big\}}_{=0} \Big).$$

Mit Hilfe der Konsistenz können wir weiter abschätzen:

$$|L_h\, u_i^h - L_h\, u_i| \leq |L_h\, u_i^h - L\, u(x_i)| + |L\, u(x_i) - L_h\, u_i| = |f(x_i) - f(x_i)| + \mathcal{O}(h^2).$$

Zusammensetzen beider Abschätzungen liefert die Behauptung. ∎