

---

## 8 Gewöhnliche Differentialgleichungen

*Beispiel: Radioaktiver Zerfall*

Sei  $m(t)$  die Menge radioaktiven Materials zur Zeit  $t$ . Der Zerfall des Materials geschieht proportional zur vorhandenen Menge mit dem Proportionalitätsfaktor  $\lambda$ . Somit gilt:

$$\frac{dm(t)}{dt} = -\lambda m(t).$$

Alle Funktionen  $m(t) = ae^{-\lambda t}$ ,  $a \in \mathbb{R}$ , sind Lösungen dieser gewöhnlichen Differentialgleichung. Gibt man zur Zeit  $t_0$  die Menge  $m_0$  des Materials vor, d.h.  $m(t_0) = m_0$ , so ist  $m(t) = m_0 e^{-\lambda(t-t_0)}$  die eindeutige Lösung dieses sogenannten Anfangswertproblems.

*Problemstellung:*

**Definition:** Für eine gegebene Funktion  $f(t, x)$ , heisst

$$\dot{x} = f(t, x) \tag{8.1}$$

(skalare) *gewöhnliche Differentialgleichung (1. Ordnung)*.

Gesucht ist eine *Lösung*  $x(t)$  der Differentialgleichung (8.1), d.h. eine Funktion  $x(t)$ , für die  $\dot{x} = f(t, x(t))$  gilt. Differentialgleichungen haben viele Lösungen. Um eine bestimmte Lösung auszuzeichnen, müssen zusätzliche Bedingungen gestellt werden.

**Definition:** Verlangt man zusätzlich zu (8.1) noch

$$x(t_0) = x_0 \tag{8.2}$$

für vorgegebenes  $x_0, t_0$ , nennt man (8.1), (8.2) ein *Anfangswertproblem*.

Gesucht ist die *Lösung*  $x(t)$  des Anfangswertproblems (8.1), (8.2), d.h. eine Funktion  $x(t)$ , für die gilt:

i)  $\dot{x}(t) = f(t, x(t))$

ii)  $x(t_0) = x_0$ .

Unter vernünftigen Bedingungen an die Funktion  $f$  besitzt ein Anfangswertproblem eine *eindeutige* Lösung.

*Beispiel:* Gegeben sei das Anfangswertproblem

$$\dot{x} = \frac{x^2}{t}, \quad x(1) = 1.$$

---

Die Lösung dieses Anfangswertproblems ist

$$x(t) = \frac{1}{1 - \ln t}.$$

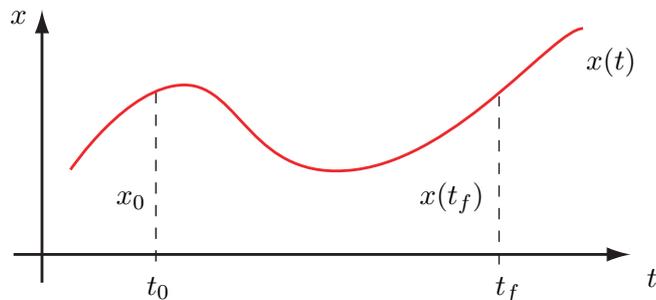
Da man nur in einfachen Fällen die Lösung eines Anfangswertproblems in geschlossener analytischer Form angeben kann, sind numerische Methoden gefragt.

*Numerische Fragestellung*

Gegeben seien das Anfangswertproblem

$$\dot{x} = f(t, x), \quad x(t_0) = x_0$$

sowie eine Endstelle  $t_f$ , und gesucht sei eine Approximation für  $x(t_f)$ .



*Grundidee:* Unterteile das Intervall  $[t_0, t_f]$  in Teilintervalle:  $t_0 < t_1 < \dots < t_n = t_f$ , und approximiere sukzessive  $x(t_1), x(t_2), \dots, x(t_n)$ . Damit wird das ursprünglich globale Problem in eine Folge lokaler Probleme zerlegt.

Dieser Zugang beruht auf der *Taylorformel* für die Funktion  $x$  an einer Stelle  $t + h$ :

$$x(t+h) = x(t) + \dot{x}(t)h + \ddot{x}(t)\frac{h^2}{2} + \dots + x^{(p)}(t)\frac{h^p}{p!} + \underbrace{x^{(p+1)}(\tau)\frac{h^{p+1}}{(p+1)!}}_{\text{Restglied}}, \quad \tau \in (t, t+h),$$

d.h. aus der Kenntnis der Funktion  $x$  an einer Stelle  $t$  (mit noch zusätzlichen Ableitungen) erhalten wir Information über die Funktion an der Stelle  $t + h$ .

Wir folgern aus der Taylorformel, dass das *Taylorpolynom vom Grad  $p$*  an der Stelle  $t$ ,

$$x(t) + \dot{x}(t)h + \dots + x^{(p)}(t)\frac{h^p}{p!},$$

$x(t+h)$  approximiert mit einem Fehler der Ordnung  $O(h^{p+1})$ .

## 8.1 Explizite Einschrittverfahren

### 8.1.1 Das Eulerverfahren

Wir approximieren  $x(t+h)$  durch  $x(t) + \dot{x}(t)h$ , d.h. durch das Taylorpolynom vom Grad 1, bzw. da  $\dot{x}(t) = f(t, x(t))$  durch den Ausdruck  $x(t) + hf(t, x(t))$ , und wir definieren die sogenannte *Verfahrensfunktion*  $F$  als

$$F(t, x, h) := x + hf(t, x).$$

Aus der Taylorformel wissen wir, dass die Funktion  $F(t, x(t), h)$  die Funktion  $x(t+h)$  mit einem Fehler der Ordnung  $O(h^2)$  approximiert.

*Bemerkung:* In der Literatur findet man oft die Verfahrensfunktion  $\Phi$ , die gegeben ist als  $x + h\Phi(t, x) := F(t, x, h)$ .

*Anwendung:*

Einteilung:  $t_0 < t_1 < t_2 \dots < t_{n-1} < t_n = t_f$

Lösung:  $x_0 = x(t_0), x_1 = x(t_1), x_2 = x(t_2), \dots, x_{n-1} = x(t_{n-1}), x_n = x(t_n)$

Schritt:  $h_0 = t_1 - t_0, h_1 = t_2 - t_1, h_2 = t_3 - t_2, \dots, h_{n-1} = t_n - t_{n-1}$

Näherung:  $\tilde{x}_0 = x_0, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{n-1}, \tilde{x}_n$

Das *Eulerverfahren* ist wie folgt definiert:

$$\begin{aligned} \tilde{x}_0 &= x_0 \\ \tilde{x}_{j+1} &= F(t_j, \tilde{x}_j, h_j), \quad j = 0, 1, \dots, n-1. \end{aligned}$$

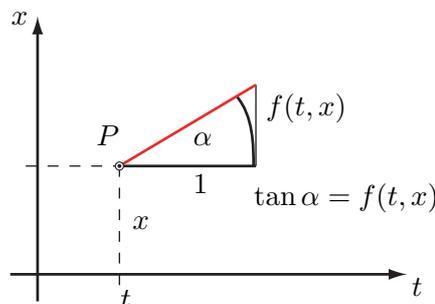
Da  $x(t_j + 1) \cong F(t_j, x(t_j), h_j) \cong F(t_j, \tilde{x}_j, h_j) = \tilde{x}_{j+1}$ , machen wir also zwei Vernachlässigungen (Fehler) pro Schritt des Eulerverfahrens.

*Bemerkung:* Oft werden (für die Theorie) die Schritte äquidistant gewählt:

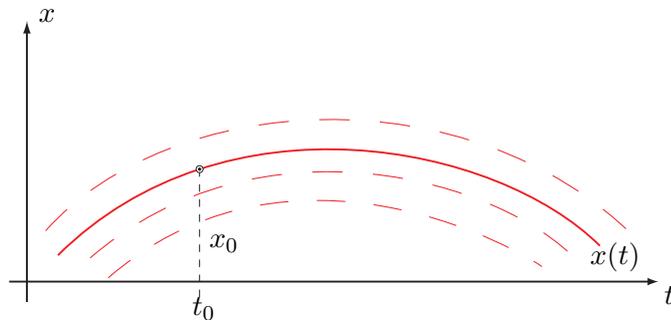
$$h := \frac{t_f - t_0}{n}, \quad t_j := t_0 + jh.$$

*Geometrische Interpretation des Eulerverfahrens*

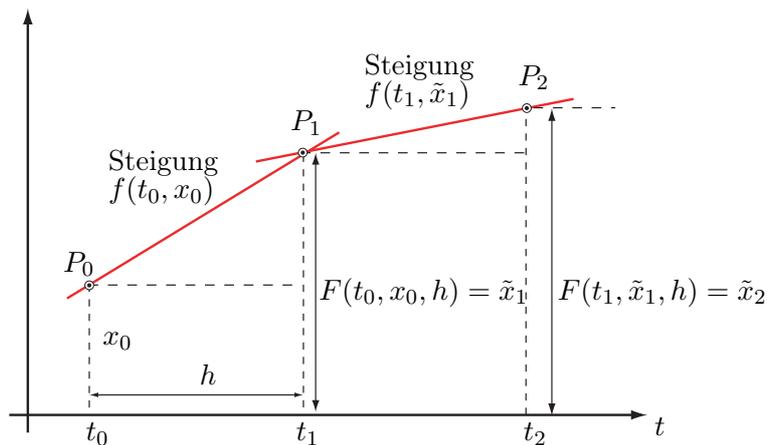
Für die Differentialgleichung (8.1) gilt, dass die Funktion  $f$  in jedem Punkt  $(x, t)$  ihres Definitionsbereiches eine Steigung (Richtung) definiert:



In jedem Punkt  $P$  stimmt die Steigung der Lösungskurve (die Richtung der Tangente der Lösungskurve) mit der Richtung dieses Richtungsfeldes überein:



Hingegen geht das Eulerverfahren von einem Punkt  $P$   $h$ -weit in Richtung des Richtungsfeldes in  $P$ . Durch das Eulerverfahren wird ein *Polygonzug* definiert:



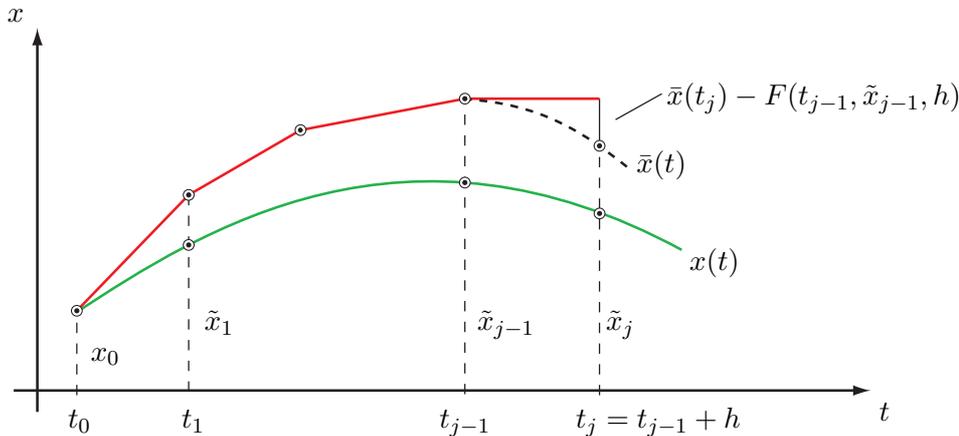
**Definition:** Der Fehler eines Einschrittverfahrens beim Schritt  $j$  heisst *lokaler (Diskretisations-)Fehler*.

*Bemerkung:* Eine übliche Bezeichnung für den lokalen Fehler im Schritt  $j + 1$  ist

$$\ell(t_j, \tilde{x}_j, h) := \bar{x}(t_j + h) - \underbrace{F(t_j, \tilde{x}_j, h)}_{=\tilde{x}_{j+1}},$$

wobei  $\bar{x}(t_j + h)$  die exakte Lösung an der Stelle  $t_j + h$  von (8.1) ist mit der Anfangsbedingung  $\bar{x}(t_j) = \tilde{x}_j$ .

**Definition:** Die Differenz  $x(t_j) - \tilde{x}_j$  heisst *globaler* (Diskretisations-)Fehler an der Stelle  $t_j$ .



**Definition:** Die *Fehlerordnung* eines Verfahrens mit Verfahrensfunktion  $F(t, x, h)$  ist gleich der Anzahl übereinstimmender Terme in den Taylorentwicklungen nach  $h$  von  $x(t+h)$  und von  $F(t, x(t), h)$ . Der *lokale Fehler* des Verfahrens ist dann von der Ordnung  $O(h^{p+1})$ .

*Bemerkungen:*

- Man kann zeigen: Ist der lokale Fehler eines Verfahrens von der Ordnung  $O(h^{p+1})$ , dann ist der *globale Fehler* von der Ordnung  $O(h^p)$ . (Dies rechtfertigt obige Definition der Fehlerordnung).
- Die *Fehlerkonstante* (der Koeffizient des  $h^{p+1}$ -Termes des lokalen Fehlers) hängt im Gegensatz zur Fehlerordnung von der Stelle  $j$  ab.

**Es gilt:** Der lokale Fehler des Eulerverfahrens ist von der Ordnung  $O(h^2)$ , falls die Schrittweite  $h$  ist. Der globale Fehler des Eulerverfahrens an der Stelle  $t_n = t_f$  ist  $O(h)$ , und somit ist die *Fehlerordnung* des Eulerverfahrens 1. Also konvergiert der Wert  $\tilde{x}_n$  für  $h = \frac{t_f - t_0}{n} \rightarrow 0$  (bei Abwesenheit von Rundungsfehlern) gegen den exakten Wert  $x(t_f)$  der Lösung. Man sagt, dass das Eulerverfahren *konsistent* ist.

*Beispiel:* Für das Anfangswertproblem

$$\dot{x} = \frac{x^2}{t}, \quad x(1) = 1$$

sei  $x(2)$  gesucht. Die exakte Lösung ist  $x(2) = (1 - \ln 2)^{-1} = 3.258891353 \dots$

Das Eulerverfahren für dieses Problem hat die Verfahrensfunktion

$$F(t, x, h) = x + h \frac{x^2}{t}.$$

Für eine äquidistante Einteilung erhalten wir damit:

$h$	$\tilde{x}_n$	Fehler (gerundet)
0.1	2.845	0.414
0.05	3.018	0.241
0.01	3.203	0.056

### 8.1.2 Taylor-Verfahren höherer Ordnung

*Ordnung 2:*

Wir approximieren  $x(t+h)$  durch  $x(t) + \dot{x}(t)h + \ddot{x}(t)\frac{h^2}{2}$ , d.h. durch das Taylorpolynom vom Grad 2. Da

$$\begin{aligned}\dot{x}(t) &= f(t, x(t)) \\ \ddot{x}(t) &= f_t(t, x(t)) + f_x(t, x(t))\dot{x}(t) = f_t(t, x(t)) + f_x(t, x(t))f(t, x(t)),\end{aligned}$$

definieren wir die Verfahrensfunktion

$$F(t, x, h) = x + hf(t, x) + \frac{h^2}{2} [f_t(t, x) + f_x(t, x)f(t, x)].$$

Das *Taylor-Verfahren der Ordnung 2* ist also wie folgt definiert:

$$\tilde{x}_{j+1} = F(t_j, \tilde{x}_j, h_j), \quad j = 0, 1, \dots, n-1.$$

**Es gilt:**

- $F(t, x(t), h)$  approximiert  $x(t+h)$  mit einem Fehler der Ordnung  $O(h^3)$ .
- Der lokale Fehler des Verfahrens ist also  $O(h^3)$  (äquidistante Schritte  $h$ ); der globale Fehler ist  $O(h^2)$ , d.h. die Fehlerordnung ist 2.
- Nachteil dieses Verfahrens: Es werden die partiellen Ableitungen  $f_t$  und  $f_x$  der Funktion  $f$  benötigt.

*Beispiel:* Für das Anfangswertproblem

$$\dot{x} = \frac{x^2}{t}, \quad x(1) = 1$$

sei wieder eine Approximation an  $x(2) = 1 - \ln 2)^{-1} = 3.258891353 \dots$  gesucht.

Das Taylor-Verfahren der Ordnung 2 hat die Verfahrensfunktion

$$F(t, x, h) = x + h \frac{x^2}{t} + h^2 (x - 0.5) \frac{x^2}{t^2}.$$

Das folgt aus

$$f = \frac{x^2}{t}, \quad f_t = \frac{-x^2}{t^2}, \quad f_x = \frac{2x}{t}.$$

Für eine äquidistante Einteilung erhalten wir

$h$	$\tilde{x}_n$	Fehler (gerundet)
0.1	3.21695	0.042
0.05	3.247044	0.012
0.01	3.25837	0.0005

*Ordnung  $p$ :*

Wir approximieren  $x(t+h)$  durch  $x(t) + \dot{x}(t)h + \dots + x^{(p)}(t) \frac{h^p}{p!}$ , d.h. durch das Taylorpolynom vom Grad  $p$ . Alle vorkommenden Ableitungen der Funktion  $x(t)$  lassen sich im Prinzip wieder durch  $f$  und partielle Ableitungen von  $f$  ausdrücken. Das ergibt die Verfahrensfunktion  $F(t, x, h)$ . Dieses Verfahren hat dann den lokalen Fehler  $O(h^{p+1})$  und den globalen Fehler  $O(h^p)$ , also Fehlerordnung  $p$ .

*Bemerkung:* Der Nachteil dieses Verfahrens ist natürlich, dass partielle Ableitungen von  $f$  bis zur Ordnung  $p-1$  benötigt werden.

### 8.1.3 Ableitungsfreie Verfahren höherer Ordnung

*Idee:* Simulation der partiellen Ableitungen durch 'Verschachtelung' von Funktionsauswertungen.

#### Das Verfahren von Heun

Wir definieren die Verfahrensfunktion:

$$F(t, x, h) = x + \frac{h}{2} [f(t, x) + f(t+h, x+hf(t, x))]$$

und damit das *Verfahren von Heun*

$$\tilde{x}_{j+1} = F(t_j, \tilde{x}_j, h_j), \quad j = 0, 1, \dots, n-1.$$

*Bemerkung:* Pro Schritt sind zwei Funktionsauswertungen nötig.

**Es gilt:** Das Verfahren von Heun hat die Fehlerordnung 2.

*Begründung:*

$$\begin{aligned} f(t + \delta, x + \Delta) &= f(t, x) + f(t, x)\delta + f_x(t, x)\Delta \\ &\quad + O(\delta^2) + O(\delta\Delta) + O(\Delta^2) \\ f(t + h, x + hf(t, x)) &= f(t, x) + f_t(t, x)h + f_x(t, x)hf(t, x) + O(h^2) \\ \frac{h}{2}f(t + h, x + hf(t, x)) &= \frac{h}{2}f(t, x) + \frac{h^2}{2}[f_t(t, x) + f_x(t, x)f(t, x)] + O(h^3) \end{aligned}$$

Daraus folgt

$$F(t, x(t), h) = x(t) + \underbrace{f(t, x(t))}_{\dot{x}(t)}h + \underbrace{[f_t(t, x(t)) + f_x(t, x(t))f(t, x(t))]}_{\ddot{x}(t)}\frac{h^2}{2} + O(h^3).$$

Also hat das Verfahren (mindestens) den lokalen Fehler der Ordnung  $O(h^3)$ . (Ganz genau müsste man noch zeigen, dass sich die  $O(h^3)$ -Terme der beiden Taylor-Entwicklungen unterscheiden).  $\square$

*Beispiel:* Das Verfahren von Heun zur Approximation von  $x(2)$  des Anfangswertproblems

$$\dot{x} = \frac{x^2}{t}, \quad x(1) = 1$$

ergibt:

$h$	$\tilde{x}_n$	Fehler (gerundet)
0.1	3.22279	0.0361
0.05	3.24898	0.0099
0.01	3.25847	0.0004

Das *Verfahren von Heun* kann auch wie folgt geschrieben werden:

$$\begin{aligned} k_1 &= f(t, x) \\ k_2 &= f(t + h, x + hk_1) \\ \bar{x} &= F(t, x, h) = x + \frac{h}{2}[k_1 + k_2], \end{aligned}$$

mit den zwei *Stufen*  $k_1, k_2$  (zwei Funktionsauswertungen). Es ist ein *explizites Runge-Kutta-Verfahren* mit zwei *Stufen* und der Fehlerordnung 2.

Das *Eulerverfahren* ist ein explizites Runge-Kutta-Verfahren mit einer Stufe und der Fehlerordnung 1:

$$\begin{aligned}k_1 &= f(t, x) \\ \bar{x} &= F(t, x, h) = x + h k_1.\end{aligned}$$

### Das klassische Runge-Kutta-Verfahren

ist wie folgt definiert:

$$\begin{aligned}k_1 &= f(t, x) \\ k_2 &= f\left(t + \frac{h}{2}, x + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t + \frac{h}{2}, x + \frac{h}{2}k_2\right) \\ k_4 &= f(t + h, x + h k_3) \\ \bar{x} &= F(t, x, h) = x + \frac{h}{6}[k_1 + 2k_2 + 2k_3 + k_4].\end{aligned}$$

Es ist explizit und besitzt 4 Stufen.

**Es gilt:** Das klassische Runge-Kutta-Verfahren hat die Fehlerordnung 4. (Die Entwicklung nach  $h$  von  $F(t, x(t), h)$  stimmt bis zu den Gliedern der Ordnung  $p = 4$  in  $h$  mit der Entwicklung von  $x(t + h)$  überein).

*Beispiel:* Das klassische Runge-Kutta-Verfahren zur Approximation der Lösung  $x(2)$  des Anfangswertproblems

$$\dot{x} = \frac{x^2}{t}, \quad x(1) = 1$$

ergibt:

$h$	$\tilde{x}_n$	Fehler (gerundet)
0.1	3.25882141	$7 \cdot 10^{-5}$
0.05	3.25888661	$5 \cdot 10^{-6}$
0.01	3.25889134	$1 \cdot 10^{-8}$

### Das Butcher-Tableau

Ein explizites Runge-Kutta-Verfahren mit  $s$  Stufen kann durch sein Butcher-Tableau wie folgt beschrieben werden:

$$\begin{array}{c|cccc}
 0 & & & & \\
 c_2 & a_{21} & & & \\
 c_3 & a_{31} & a_{32} & & \\
 \vdots & \vdots & \vdots & \ddots & \\
 c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\
 \hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
 \end{array}
 \qquad
 c_i = \sum_{j=1}^{i-1} a_{ij}$$

$$k_i = f(t + c_i h, x + h \sum_{j=1}^{i-1} a_{ij} k_j), \quad i = 1, \dots, s \quad (\text{Stufen})$$

$$\bar{x} = x + h \sum_{i=1}^s b_i k_i \quad (\text{Approximation bei } t + h)$$

*Beispiele:*

Eulerverfahren:

$$\begin{array}{c|c}
 0 & \\
 \hline
 1 & 1
 \end{array}
 \qquad
 \begin{array}{l}
 s = 1 \\
 p = 1
 \end{array}$$

Verfahren von Heun:

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

Klassisches Runge-Kutta-Verfahren:

$$\begin{array}{c|ccc}
 0 & & & \\
 \frac{1}{2} & \frac{1}{2} & & \\
 \frac{1}{2} & 0 & \frac{1}{2} & \\
 1 & 0 & 0 & 1 \\
 \hline
 & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6}
 \end{array}
 \qquad
 \begin{array}{l}
 s = 4 \\
 p = 4
 \end{array}$$

*Bemerkung: (Butcher Barrier)* Für  $p \geq 5$  existiert kein explizites Runge-Kutta-Verfahren der Fehlerordnung  $p$  mit  $s = p$  Stufen.

### 8.1.4 Variable Schrittweiten

*Wunsch:* Wenn die Lösung stark variiert, möchte man eine hohe Auflösung (kleine Schritte), wenn sie wenig variiert, eine geringere Auflösung (grosse Schritte).

*Ziel:* Die Schrittweiten  $h_j$  sollen so gewählt werden, dass der lokale Fehler gleich einer vorgegeben Toleranz  $TOL$  ist:

$$|\ell(t_j, \tilde{x}_j, h_j)| = TOL.$$

Da  $\ell$  nicht bekannt ist, begnügt man sich mit einer Schätzung  $\bar{\ell}$  des lokalen Fehlers.

*Idee für die Schätzung  $\bar{\ell}$ :*

- Berechne  $\tilde{x}_{j+1} = F(t_j, \tilde{x}_j, h)$  mit einem ersten Verfahren  $F$ .
- Berechne  $\hat{\tilde{x}}_{j+1} = \hat{F}(t_j, \tilde{x}_j, h)$  mit einem Vergleichsverfahren  $\hat{F}$ .
- Setze  $\bar{\ell}(t_j, \tilde{x}_j, h) := \hat{F}(t_j, \tilde{x}_j, h) - F(t_j, \tilde{x}_j, h)$ .

**Es gilt:** Falls  $F$  von der Fehlerordnung  $p$  ist und  $\hat{F}$  von der Fehlerordnung  $\hat{p}$  mit  $\hat{p} > p$ , dann ist

$$\bar{\ell}(t_j, \tilde{x}_j, h) = O(h^{p+1})$$

und

$$\ell(t_j, \tilde{x}_j, h) = \bar{\ell}(t_j, \tilde{x}_j, h) + O(h^{\hat{p}+1}).$$

*Begründung:* Sei  $z(t)$  die Lösung von (8.1) mit der Anfangsbedingung  $z(t_j) = \tilde{x}_j$ , und sei  $z_{j+1} := z(t_j + h)$ . Dann gilt

$$\begin{aligned} \ell(t_j, \tilde{x}_j, h) &= z_{j+1} - \tilde{x}_{j+1} = z_{j+1} - \hat{\tilde{x}}_{j+1} + \hat{\tilde{x}}_{j+1} - \tilde{x}_{j+1} \\ &= \hat{\ell}(t_j, \tilde{x}_j, h) + \bar{\ell}(t_j, \tilde{x}_j, h) = \bar{\ell}(t_j, \tilde{x}_j, h) + O(h^{\hat{p}+1}), \end{aligned}$$

da  $\hat{\ell}(t_j, \tilde{x}_j, h) = O(h^{\hat{p}+1})$ . Da  $\ell(t_j, \tilde{x}_j, h) = O(h^{p+1})$  ist, folgt daraus mit  $\hat{p} > p$  auch, dass  $\bar{\ell}(t_j, \tilde{x}_j, h) = O(h^{p+1})$ . □

#### Variante A (*Richardson-Extrapolation*)

Wir illustrieren das Vorgehen ohne Einschränkung der Allgemeinheit für  $p = 1$  und das Eulerverfahren für  $F$ . Für  $\hat{F}$  wählen wir zwei Euler-Schritte mit Schrittweite  $\frac{h}{2}$ :

$$\begin{aligned} \hat{\tilde{x}}_{j+1} &= \tilde{x}_j + \frac{h}{2}f(t_j, \tilde{x}_j) + \frac{h}{2}f(t_j + \frac{h}{2}, \tilde{x}_j + \frac{h}{2}f(t_j, \tilde{x}_j)) \\ &\stackrel{\text{Taylor}}{=} \tilde{x}_j + \frac{h}{2}f(t_j, \tilde{x}_j) + \frac{h}{2}[f(t_j, \tilde{x}_j) + \frac{h}{2}f_t(t_j, \tilde{x}_j) + \frac{h}{2}f_x(t_j, \tilde{x}_j)f(t_j, \tilde{x}_j) + \dots] \\ &= \tilde{x}_j + hf(t_j, \tilde{x}_j) + \frac{h^2}{4} \underbrace{[f_t(\dots) + f_x(\dots)f(\dots)]}_{\ddot{z}(t_j)} + O(h^3), \end{aligned}$$

wobei  $z(t)$  die Lösung von (8.1) sei mit der Anfangsbedingung  $z(t_j) = \tilde{x}_j$ . Da  $z(t_j + h)$  die Taylorentwicklung

$$z(t_j + h) = \tilde{x}_j + hf(t_j, \tilde{x}_j) + \frac{h^2}{2}\ddot{z}(t_j) + O(h^3)$$

besitzt, erhalten wir für den lokalen Fehler des Verfahrens  $F$

$$z(t_j + h) - \tilde{x}_{j+1} = \frac{h^2}{2}\ddot{z}(t_j) + O(h^3). \tag{8.3}$$

Da  $\dot{z}(t) = f(t, z)$  ist, ergibt sich

$$\ddot{z}(t) = f_t(t, z) + f_z(t, z)\dot{z}(t) = f_t(t, z) + f_z(t, z)f(t, z).$$

Also ist der lokale Fehler von  $\hat{F}$

$$z(t_j + h) - \hat{x}_{j+1} = \frac{h^2}{4}\ddot{z}(t_j) + O(h^3). \tag{8.4}$$

Subtrahiert man vom Zweifachen der Gleichung (8.4) die Gleichung (8.3), erhält man

$$\underbrace{z(t_j + h) - \hat{x}_{j+1}}_{=\hat{\ell}(t_j, \tilde{x}_j, h)} = \underbrace{\hat{x}_{j+1} - \tilde{x}_{j+1}}_{=:\bar{\ell}(t_j, \tilde{x}_j, h)} + O(h^3).$$

Daraus folgt, dass  $\hat{F}(t_j, \tilde{x}_j, h) - F(t_j, \tilde{x}_j, h)$  eine Schätzung des lokalen Fehlers  $\hat{\ell}(t_j, \tilde{x}_j, h)$  des Verfahrens  $\hat{F}$  liefert zur Ordnung  $O(h^3)$ .

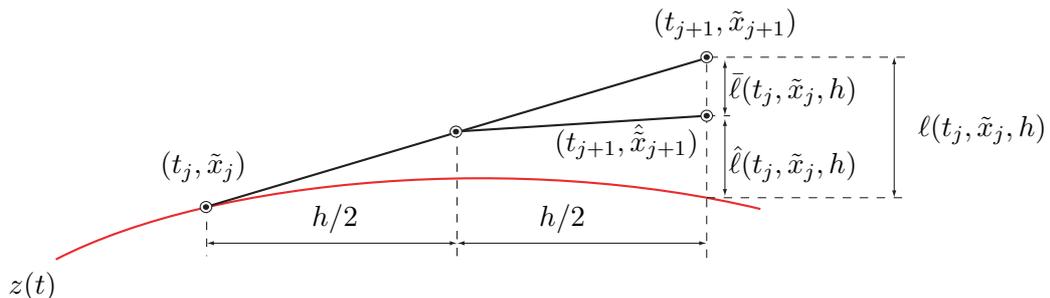
*Bemerkung:*

- $2\bar{\ell}(t_j, \tilde{x}_j, h)$  ergibt eine Schätzung zur Ordnung  $O(h^3)$  des lokalen Fehlers  $\ell(t_j, \tilde{x}_j, h)$  des Verfahrens  $F$ .
- Für allgemeine Ordnung  $p$  gilt:

$$z(t_j + h) - \hat{x}_{j+1} = \underbrace{\frac{\hat{x} - \tilde{x}_{j+1}}{2^p - 1}}_{=\bar{\ell}(t_j, \tilde{x}_j, h)} + O(h^{p+1}).$$

Dieses Vorgehen heisst *Richardson-Extrapolation*, da dabei der unbekannte Fehlerterm  $\frac{h^2}{2}\ddot{z}(t_j)$  in (8.3) und (8.4) eliminiert wird (vergleiche die Romberg-Extrapolation bei der numerischen Integration).

Die Richardson-Extrapolation für das Euler-Verfahren lässt sich wie folgt skizzieren:



**Variante B (Eingebettete Runge-Kutta-Formeln)**

Wir betrachten zwei Runge-Kutta-Verfahren (für  $\tilde{x}_{j+1}$  und für  $\hat{x}_{j+1}$ ) mit den gleichen Funktionsauswertungen und den Fehlerordnungen  $p$  und  $\hat{p}$ , wobei  $\hat{p} = p + 1$  (oder  $\hat{p} = p - 1$ ). (Die bekannteste Methode ist DOPRI5 von Dormand/Prince mit  $p = 5$ ,  $\hat{p} = 4$  ( $s = 7$ )).

0					
$c_2$	$a_{21}$				
$c_3$	$a_{31}$	$a_{32}$			
$\vdots$	$\vdots$				
$c_s$	$a_{s1}$	$a_{s2}$	$\cdots$	$a_{s,s-1}$	
	$b_1$	$b_2$	$\cdots$	$b_{s-1}$	$b_s$
	$\hat{b}_1$	$\hat{b}_2$	$\cdots$	$\hat{b}_{s-1}$	$\hat{b}_s$

$$\tilde{x}_{j+1} = \tilde{x}_j + h \sum_{i=1}^s b_i k_i \quad (\text{Fehlerordnung } p)$$

$$\hat{x}_{j+1} = \tilde{x}_j + h \sum_{i=1}^s \hat{b}_i k_i \quad (\text{Fehlerordnung } \hat{p} = p + 1)$$

Damit gilt:

$$\bar{\ell}(t_j, \tilde{x}_j, h) = \hat{x}_{j+1} - z(t_j+1) + z(t_j+1) - \tilde{x}_{j+1} = \underbrace{-\hat{\ell}(t_j, \tilde{x}_j, h)}_{O(h^{p+2})} + \underbrace{\ell(t_j, \tilde{x}_j, h)}_{O(h^{p+1})} = O(h^{p+1})$$

und

$$\ell(t_j, \tilde{x}_j, h) = \bar{\ell}(t_j, \tilde{x}_j, h) + O(h^{\hat{p}+1}),$$

und  $\bar{\ell}(t_j, \tilde{x}_j, h) = \hat{F}(t_j, \tilde{x}_j, h) - F(t_j, \tilde{x}_j, h)$  liefert also eine Schätzung des lokalen Fehlers  $\ell(t_j, \tilde{x}_j, h)$  des Verfahrens  $F$  von der Ordnung  $O(h^{\hat{p}+1}) = O(h^{p+2})$ .

**Praktische Schrittweitensteuerung**

Wähle ein Verfahren  $F$  der Fehlerordnung  $p$  und ein Verfahren  $\hat{F}$  der Fehlerordnung  $\hat{p}$  mit  $\hat{p} \geq p$  und bilde

$$\bar{\ell}(t_j, \tilde{x}_j, h) = \hat{F}(t_j, \hat{x}_j, h) - F(t_j, \tilde{x}_j, h).$$

(Bei der Richardson-Extrapolation bilde  $\bar{\ell} = \frac{1}{2^{p-1}}[\hat{F} - F]$ ).

Es gilt:

$$\begin{aligned}
 |\bar{\ell}| &\cong C \cdot h^{p+1} \Rightarrow C \cong \frac{|\bar{\ell}|}{h^{p+1}} \\
 C \cdot h_{opt}^{p+1} &\cong TOL \Rightarrow h_{opt} \cong \left(\frac{TOL}{C}\right)^{\frac{1}{p+1}} \\
 &\Rightarrow h_{opt} \cong h \left(\frac{TOL}{|\bar{\ell}|}\right)^{\frac{1}{p+1}},
 \end{aligned}$$

dabei ist  $h$  die aktuell durchgeführte Schrittlänge und  $h_{opt}$  das optimale  $h$  für diesen Schritt.

**Algorithmus:** (Schrittweitensteuerung)

1. Berechne  $\bar{\ell}(t_j, \tilde{x}_j, h_j) = \hat{\tilde{x}}_{j+1} - \tilde{x}_{j+1}$  [für Richardson:  $\cdot \frac{1}{2^{p-1}}$ ]
2. Falls  $|\bar{\ell}| > TOL$ :
  - Verwerfe die Approximation  $\tilde{x}_{j+1}$
  - Wähle eine neue (kleinere) Schrittweite

$$h^* = h_j \left(\frac{TOL}{|\bar{\ell}|}\right)^{\frac{1}{p+1}} \cdot Fak,$$

wobei  $Fak$  ein Sicherheitsfaktor ist, z.B.  $Fak = 0.8$ .

- Setze  $h_j := h^*$ ; goto 1.
3. Falls  $|\bar{\ell}| \leq TOL$ :
    - Akzeptiere  $\tilde{x}_{j+1}$  als Approximation für  $x(t_{j+1})$
    - Bestimme den Vorschlag für  $h_{j+1}$  als

$$h_{j+1} = h_j \left(\frac{TOL}{|\bar{\ell}|}\right)^{\frac{1}{p+1}} \cdot Fak.$$

## 8.2 Implizite Einschrittverfahren

### Die Trapezmethode

Wir betrachten wie bisher das *Anfangswertproblem*

$$\dot{x} = f(t, x), \quad x(t_0) = x_0$$

und die Einteilung  $t_0 < t_1 < \dots < t_n = t_f$ . Die Integration beider Seiten der Differentialgleichung von  $t_0$  bis  $t_1$  ergibt die zum Anfangswertproblem äquivalente (auf  $[t_0, t_1]$ ) Integralgleichung:

$$x(t_1) - x(t_0) = \int_{t_0}^{t_1} f(t, x(t)) dt.$$

*Idee:* Approximiere das Integral durch eine Quadraturformel, z.B. durch den Trapezwert:

$$\int_{t_0}^{t_1} f(t, x(t)) dt \cong \frac{(t_1 - t_0)}{2} [f(t_0, x(t_0)) + f(t_1, x(t_1))].$$

Dies führt zur *Trapezmethode*:

$$\tilde{x}_{j+1} = \tilde{x}_j + \frac{h_j}{2} [f(t_j, \tilde{x}_j) + f(t_{j+1}, \tilde{x}_{j+1})] =: F(t_j, \tilde{x}_j; t_{j+1}, \tilde{x}_{j+1}; h_j), \quad j = 0, 1, \dots, n-1 \quad (8.5)$$

*Bemerkung:* Für eine allgemeine nichtlineare Funktion  $f(t, x)$  stellt (8.5) eine implizite Gleichung für den unbekanntem Näherungswert  $\tilde{x}_{j+1}$  dar. Diese nichtlineare Gleichung muss in jedem Schritt gelöst werden, z.B. mit dem Newtonverfahren oder in diesem Fall einfacher mit Fixpunktiteration:

$$\begin{aligned} \tilde{x}_{j+1}^{(0)} &= \tilde{x}_j + hf(t_j, \tilde{x}_j) && \text{(Start mit dem Eulerverfahren)} \\ \tilde{x}_{j+1}^{(k+1)} &= \tilde{x}_j + \frac{h_j}{2} [f(t_j, \tilde{x}_j) + f(t_{j+1}, \tilde{x}_{j+1}^{(k)})], \quad k = 0, 1, \dots \end{aligned}$$

*Abbruchkriterium:*

$$|\tilde{x}_{j+1}^{(k+1)} - \tilde{x}_{j+1}^{(k)}| \leq |\tilde{x}_{j+1}^{(k+1)}| \cdot TOL + TOL$$

Da der Bachnach'sche Fixpunktsatz für genügend kleine  $h_j$  erfüllt ist, konvergiert diese Iteration für  $h_j$  klein genug.

**Es gilt:** Die Fehlerordnung der Trapezmethode ist 2 (lokaler Fehler von der Ordnung  $O(h^3)$ ).

*Bemerkung:* Implizite Verfahren sind im Allgemeinen den expliziten Verfahren überlegen, dafür aber aufwendiger. Sie besitzen Stabilitätseigenschaften, die für eine grosse Klasse von Differentialgleichungen wichtig sind. Implizite Runge-Kutta-Verfahren spielen in diesem Zusammenhang heute eine wichtige Rolle.

### Das Butcher-Tableau für allgemeine Runge-Kutta-Verfahren

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}; \quad \text{Runge-Kutta Matrix } A \in \mathbb{R}^{s \times s}; \quad b, c \in \mathbb{R}^s$$

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

$$k_i = f\left(t + c_i h, x + h \sum_{j=1}^s a_{ij} k_j\right), \quad i = 1, \dots, s$$

$$\bar{x} = x + h \sum_{i=1}^s b_i k_i$$

**Definition:** Ein Runge-Kutta-Verfahren heisst *explizit*, falls  $a_{ij} = 0$  für  $i \leq j$ , und *implizit* sonst.

*Beispiele:*

Implizites Eulerverfahren:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \quad \begin{array}{l} \bar{x} = x + h f(t + h, \bar{x}) \\ s = 1 \\ p = 1 \end{array} \quad \begin{array}{l} k_1 = f(t + h, x + h k_1) \\ \bar{x} = x + h k_1 \end{array}$$

Trapezmethode:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \begin{array}{l} \bar{x} = x + \frac{h}{2} [f(t, x) + f(t + h, \bar{x})] \\ s = 2 \\ p = 2 \end{array} \quad \begin{array}{l} k_1 = f(t, x) \\ k_2 = f(t + h, x + \frac{h}{2} k_1 + \frac{h}{2} k_2) \\ \bar{x} = x + \frac{h}{2} k_1 + \frac{h}{2} k_2 \end{array}$$

Theta-Verfahren :

$$\bar{x} = x + h f(t + \vartheta h, x + \vartheta h k_1)$$

$\vartheta$	$\vartheta$	$s = 1$	$k_1 = f(t + \vartheta h, x + \vartheta h k_1)$
	1	$p = 1(2)$	$\bar{x} = x + h k_1$

$\vartheta = 0$  : explizites Eulerverfahren ( $p = 1$ )

$\vartheta = \frac{1}{2}$  : implizite Mittelpunktsregel ( $p = 2$ )

$\vartheta = 1$  : implizites Eulerverfahren ( $p = 1$ )

Implizite Mittelpunktsregel:

$$\bar{x} = x + h f\left(t + \frac{h}{2}, \frac{x + \bar{x}}{2}\right)$$

$\frac{1}{2}$	$\frac{1}{2}$	$s = 1$	$k_1 = f\left(t + \frac{h}{2}, x + \frac{h}{2} k_1\right)$
	1	$p = 2$	$\bar{x} = x + h k_1$

### 8.3 Systeme gewöhnlicher Differentialgleichungen

Wir betrachten das Anfangswertproblem

$$\dot{\underline{x}} = \underline{f}(t, \underline{x}), \quad \underline{x}(t_0) = \underline{x}^0$$

mit der Lösung  $\underline{x}(t)$ . Für ein  $n$ -dimensionales System ist

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \underline{f}(t, \underline{x}) = \begin{pmatrix} f_1(t, x_1, \dots, x_n) \\ \vdots \\ f_n(t, x_1, \dots, x_n) \end{pmatrix}, \text{ etc.}$$

Für die numerische Integration  $n$ -dimensionaler Differentialgleichungssysteme gelten alle bisherigen Formeln analog, z.B:

Euler-verfahren:  $\underline{F}(t, \underline{x}, h) = \underline{x} + h \underline{f}(t, \underline{x})$

Verfahren von Heun:  $\underline{F}(t, \underline{x}, h) = \underline{x} + \frac{h}{2} [\underline{f}(t, \underline{x}) + \underline{f}(t + h, \underline{x} + h \underline{f}(t, \underline{x}))]$

Klass. Runge-Kutta-Verfahren:  $\underline{k}_1 = \underline{f}(t, \underline{x})$

$$\underline{k}_2 = \underline{f}\left(t + \frac{h}{2}, \underline{x} + \frac{h}{2} \underline{k}_1\right)$$

$$\underline{k}_3 = \underline{f}\left(t + \frac{h}{2}, \underline{x} + \frac{h}{2} \underline{k}_2\right)$$

$$\underline{k}_4 = \underline{f}(t + h, \underline{x} + h \underline{k}_3)$$

$$\underline{F}(t, \underline{x}, h) = \underline{x} + \frac{h}{6} [\underline{k}_1 + 2\underline{k}_2 + 2\underline{k}_3 + \underline{k}_4]$$

Das Butcher-Tableau:

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

charakterisiert ein Runge-Kutta-Verfahren also auch für den Fall eines Differentialgleichungssystems.

*Beispiel:* Wir betrachten die Pendelgleichung von weiter unten und wenden darauf das implizite Eulerverfahren

$$\tilde{x}^{j+1} = \tilde{x}^j + h_j \underline{f}(t_{j+1}, \tilde{x}^{j+1}), \quad j = 0, 1, 2, \dots$$

an mit konstantem Schritt  $h$ :

$$\begin{aligned} \tilde{x}_1^{j+1} &= \tilde{x}_1^j + h \tilde{x}_2^{j+1} \\ \tilde{x}_2^{j+1} &= \tilde{x}_2^j - h \sin \tilde{x}_1^{j+1}. \end{aligned}$$

*Bemerkung:* Auch die Schrittweitensteuerung funktioniert für Systeme wie im skalaren Fall. Der Betrag  $|\cdot|$  muss dabei durch eine Norm  $\|\cdot\|$  ersetzt werden.

### Differentialgleichungen höherer Ordnung

Beispiel:  $\ddot{x} = g(t, x, \dot{x}, \ddot{x})$

Definiere:  $x_1 = x, x_2 = \dot{x}, x_3 = \ddot{x}$

Dann gilt:  $\ddot{x} = g(t, x, \dot{x}, \ddot{x})$  ist äquivalent zum 3-dimensionalen System

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= g(t, x_1, x_2, x_3). \end{aligned}$$

*Folgerung:* Differentialgleichungen der Ordnung  $k$  kann man als  $k$ -dimensionale Systeme 1. Ordnung schreiben. Es genügt also, wenn man Systeme von Differentialgleichungen erster Ordnung numerisch integrieren kann.

*Beispiel:* Die Differentialgleichung des mathematischen Pendels

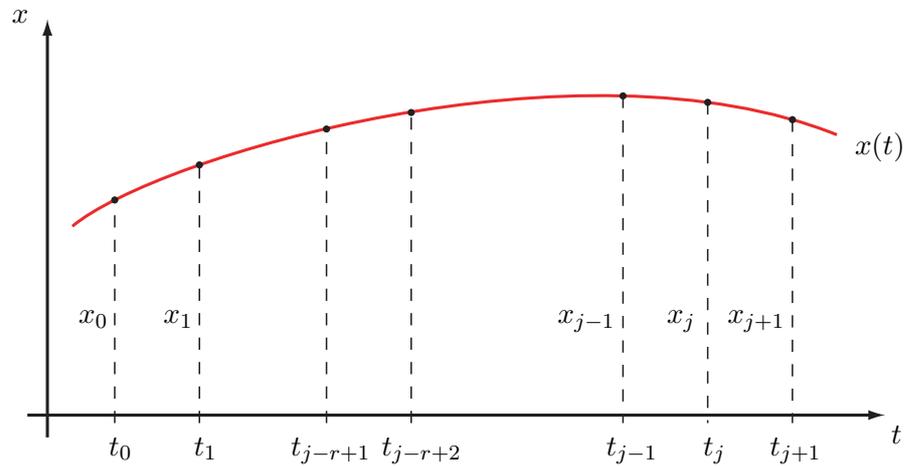
$$\ddot{x} + \sin x = 0$$

ist äquivalent zum 2-dimensionalen System

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\sin x_1. \end{aligned}$$

## 8.4 Mehrschrittverfahren

*Idee:* Versuche höhere Fehlerordnung eines Verfahrens anstatt durch Funktionsschachtelung durch Einbeziehen von Information an mehreren Stellen zu erreichen.

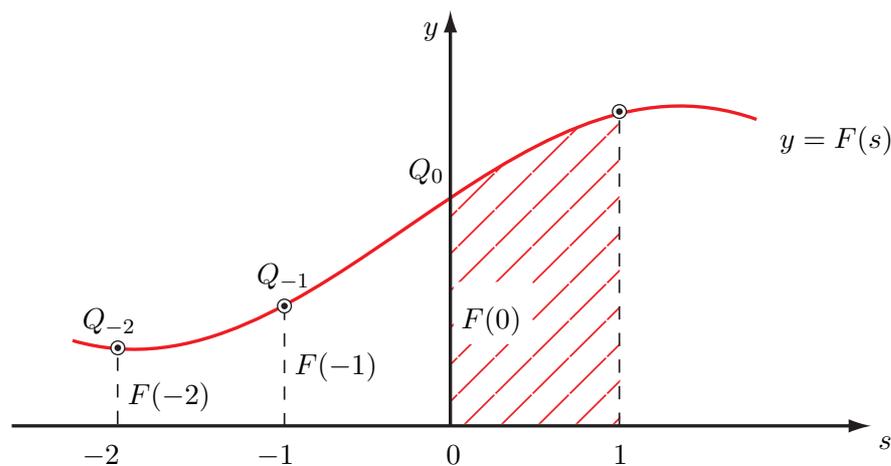


Wir benutzen die Approximationen  $\tilde{x}_{j-r+1}, \tilde{x}_{j-r+2}, \dots, \tilde{x}_j$  zur Bestimmung der Approximation  $\tilde{x}_{j+1}$ . Das ergibt ein explizites  $r$ -Schritt-Verfahren.

### Herleitung einer Adams-Bashforth-Formel

Integrieren der Differentialgleichung  $\dot{x} = f(t, x)$  von  $t_j$  bis  $t_{j+1} = t_j + h$  ergibt mit der Substitution  $t = t_j + sh, dt = h ds$ :

$$\underbrace{x(t_{j+1})}_{x_{j+1}} - \underbrace{x(t_j)}_{x_j} = \int_{t_j}^{t_{j+1}} f(t, x(t)) dt = h \int_0^1 \underbrace{f(t_j + sh, x(t_j + sh))}_{=: F(s)} ds$$



$$\begin{aligned} F(0) &= f(t_j, x_j) &=: f_j \\ F(-1) &= f(t_{j-1}, x_{j-1}) &=: f_{j-1} \\ F(-2) &= f(t_{j-2}, x_{j-2}) &=: f_{j-2} \end{aligned}$$

*Idee:* Lege Interpolationspolynom  $P_2(s)$  durch  $Q_{-2}, Q_{-1}, Q_0$  und approximiere  $F(s)$  für  $s \in [0, 1]$  durch  $P_2(s)$ .

Wir führen Lagrange-Interpolation durch:

$$\begin{aligned} P_2(s) &= f_j \frac{(s+1)(s+2)}{2} - f_{j-1} s(s+2) + f_{j-2} \frac{s(s+1)}{2} \\ \int_0^1 F(s) ds &\cong \int_0^1 P_2(s) ds = f_j \frac{23}{12} - f_{j-1} \frac{16}{12} + f_{j-2} \frac{5}{12}. \end{aligned}$$

Wir definieren damit die *3-Schritt-Adams-Bashfort-Formel*

$$\tilde{x}_{j+1} = \tilde{x}_j + \frac{h}{12} [23f(t_j, \tilde{x}_j) - 16f(t_{j-1}, \tilde{x}_{j-1}) + 5f(t_{j-2}, \tilde{x}_{j-2})], \quad j = 2, 3, \dots$$

*Bemerkungen:*

- Dies ist ein explizites lineares 3-Schritt-Verfahren mit der Fehlerordnung 3 (lokaler Fehler von der Ordnung  $O(h^4)$ ).
- Neben dem Startwert  $x_0$  zur Zeit  $t_0$  braucht dieses Verfahren noch weitere Startwerte zur Zeit  $t_0 + h$  und  $t_0 + 2h$ . Diese Approximationen  $\tilde{x}_1$  und  $\tilde{x}_2$  können durch ein Einschrittverfahren der gleichen Fehlerordnung oder durch eines der Fehlerordnung 1 mit kleinen Schritten gewonnen werden.
- Explizite Mehrschrittverfahren sind billig, da sie nur eine Funktionsauswertung pro Schritt brauchen. Aber sie benötigen ein Startverfahren und die Schrittweitensteuerung ist schwieriger als bei Runge-Kutta-Verfahren.

### Eine Adams-Moulton-Formel

Nimmt man für die Interpolation den Punkt  $Q_1 = (1, F(1))$  hinzu, bekommt man ein implizites Mehrschrittverfahren. Die Interpolation durch  $Q_{-2}, Q_{-1}, Q_0, Q_1$  ergibt eine *3-Schritt-Adams-Moulton-Formel*:

$$\tilde{x}_{j+1} = \tilde{x}_j + \frac{h}{24} [9f(t_{j+1}, \tilde{x}_{j+1}) + 19f(t_j, \tilde{x}_j) - 5f(t_{j-1}, \tilde{x}_{j-1}) + f(t_{j-2}, \tilde{x}_{j-2})]$$

für  $j = 2, 3, \dots$

*Bemerkungen:*

- Dieses Verfahren hat die Fehlerordnung 4.
- Die implizite Gleichung kann gut durch Fixpunktiteration mit Startwert  $\tilde{x}_{j+1}^0 = \tilde{x}_j$  gelöst werden.
- Eine andere Möglichkeit ist,  $\tilde{x}_{j+1}$  auf der rechten Seite mit der 3-Schritt-Adams-Bashforth-Formel zu bestimmen. Das ergibt eine *Prädiktor-Korrektor-Methode* der Fehlerordnung 4, die aber nicht mehr implizit ist.

## 8.5 Stabilität von Einschrittverfahren

Wir betrachten das folgende *Modellproblem*:

$$\dot{x} = \lambda x, \quad x(0) = x_0$$

für  $\lambda \in \mathbb{C}$ . Die Lösung dieses Anfangswertproblems ist

$$x(t) = e^{\lambda t} x_0.$$

*Bemerkung:* Für  $\operatorname{Re}(\lambda) < 0$  gilt:  $x(t) \rightarrow 0$  für  $t \rightarrow +\infty$ .

Wir führen die Stabilitätsdiskussion an zwei konkreten Einschrittverfahren durch, einem expliziten und einem impliziten Verfahren.

### Verfahren von Heun

$$\tilde{x}_{j+1} = \tilde{x}_j + \frac{h}{2} [f(t_j, \tilde{x}_j) + f(t_{j+1}, \tilde{x}_j + hf(t_j, \tilde{x}_j))], \quad j = 0, 1, 2, \dots,$$

mit  $\tilde{x}_0 = x_0$ , angewandt auf das Modellproblem mit  $f(t, x) = \lambda x$  ergibt:

$$\begin{aligned} \tilde{x}_{j+1} &= \tilde{x}_j + \frac{h}{2} [\lambda \tilde{x}_j + \lambda(\tilde{x}_j + h\lambda \tilde{x}_j)] \\ &= \tilde{x}_j + h\lambda \tilde{x}_j + \frac{(h\lambda)^2}{2} \tilde{x}_j = \left[ 1 + h\lambda + \frac{(h\lambda)^2}{2} \right] \tilde{x}_j \end{aligned}$$

**Definition:** *Stabilitätsfunktion* des Verfahrens von Heun:

$$R(\mu) = 1 + \mu + \frac{\mu^2}{2} \in \mathbb{C}, \quad \mu := \lambda h$$

**Es gilt:**

$$\tilde{x}_{j+1} = R(h\lambda)\tilde{x}_j \quad j = 0, 1, 2, \dots, \text{ bzw. } \tilde{x}_n = R(h\lambda)^n x_0.$$

Das Verfahren von Heun approximiert die Lösung des Modellproblems zur Zeit  $t = nh$ , nämlich

$$e^{\lambda t} x_0 = e^{\lambda nh} x_0 = \left(e^{\lambda h}\right)^n x_0,$$

durch

$$(R(\lambda h))^n x_0, \text{ wobei } R(\mu) = 1 + \mu + \frac{\mu^2}{2}.$$

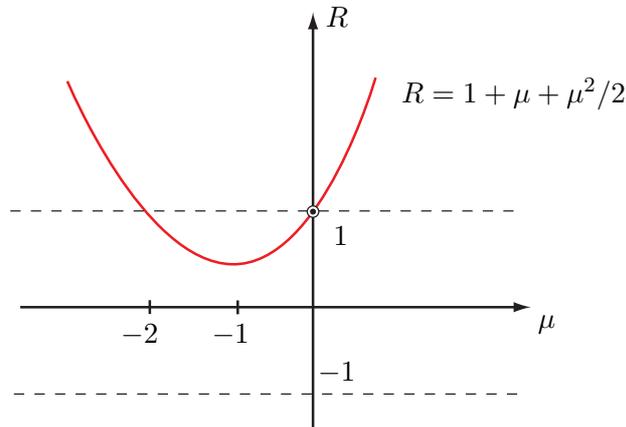
*Bemerkungen:*

- Während die Lösung des Modellproblems in jedem  $h$ -Schritt mit dem Faktor  $e^{\lambda h}$  multipliziert wird, wird die Approximation in jedem Schritt mit dem Faktor  $R(\lambda h)$  multipliziert.
- Es gilt:

$$e^\mu = 1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6} + \dots = R(\mu) + O(\mu^3).$$

Wir betrachten zuerst den Fall  $\lambda \in \mathbb{R}$ . Für  $\lambda < 0$  ist  $|e^{\lambda h}| < 1$ , falls  $h > 0$ , und deshalb gilt  $(e^{\lambda h})^n x_0 \rightarrow 0$  für  $n \rightarrow +\infty$ .

**Frage:** Für welche  $h$  gilt  $|R(\lambda h)| < 1$ , falls  $\lambda < 0$  ?



**Definition:** *Reelles Stabilitätsintervall* eines Einschrittverfahrens:

$$B = \{\mu \in \mathbb{R} \mid |R(\mu)| < 1\} \subset \mathbb{R}$$

*Folgerung:* Das reelle Stabilitätsintervall des Verfahrens von Heun ist  $B_{\text{Heun}} = (-2, 0)$ .

Wir betrachten im Folgenden den allgemeinen Fall  $\lambda \in \mathbb{C}$ .

**Definition:** *Stabilitätsgebiet* eines Einschrittverfahrens:

$$A = \{\mu \in \mathbb{C} \mid |R(\mu)| < 1\} \subset \mathbb{C}$$

Für das Verfahren von Heun gilt  $R(\mu) = 1 + \mu + \frac{\mu^2}{2}$ . Wir bestimmen

$$\partial A = \{\mu \in \mathbb{C} \mid |R(\mu)| = 1\} = \text{Rand von } A.$$

Da das reelle Stabilitätsintervall  $B_{\text{Heun}} = (-2, 0)$  ist, machen wir den Ansatz  $\mu = -1 + re^{i\varphi}$ . Dies ergibt:

$$\begin{aligned} \mu^2 &= 1 - 2re^{i\varphi} + r^2e^{i2\varphi} \\ R(\mu) &= re^{i\varphi} + \frac{1}{2} - re^{i\varphi} + \frac{r^2}{2}e^{i2\varphi} = \frac{1}{2}(1 + r^2e^{i2\varphi}). \end{aligned}$$

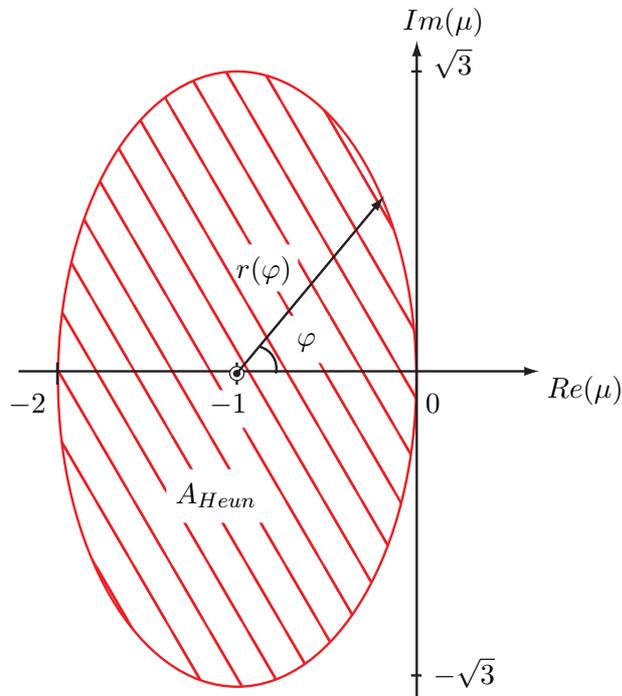
Aus

$$1 = R \cdot \bar{R} = |R(\mu)|^2 = \frac{1}{4}(1 + 2r^2 \cos 2\varphi + r^4)$$

erhält man die folgende Gleichung für  $r^2$ :

$$r^2 = -\cos 2\varphi + \sqrt{\cos^2 2\varphi + 3},$$

d.h.  $r^2$  und damit  $r$  ist eine  $\pi$ -periodische Funktion in  $\varphi$  mit folgenden Eigenschaften: Für  $\varphi = 0, \pi$  ist  $r = 1$ , für  $\varphi = \frac{\pi}{2}, \frac{3\pi}{2}$  ist  $r = \sqrt[4]{3}$ . Also hat das Stabilitätsgebiet des Verfahrens von Heun,  $A_{\text{Heun}}$ , die unten skizzierte Form:



**Trapezmethode**

$$\tilde{x}_{j+1} = \tilde{x}_j + \frac{h}{2} [f(t_j, \tilde{x}_j) + f(t_{j+1}, \tilde{x}_{j+1})], \quad j = 0, 1, 2, \dots,$$

mit  $\tilde{x}_0 = x_0$ , angewandt auf das Modellproblem mit  $f(t, x) = \lambda x$  ergibt:

$$\tilde{x}_{j+1} = \tilde{x}_j + \frac{h}{2} [\lambda \tilde{x}_j + \lambda \tilde{x}_{j+1}] = \tilde{x}_j + \frac{h\lambda}{2} \tilde{x}_j + \frac{h\lambda}{2} \tilde{x}_{j+1}$$

bzw.

$$\left(1 - \frac{h\lambda}{2}\right) \tilde{x}_{j+1} = \left(1 + \frac{h\lambda}{2}\right) \tilde{x}_j$$

und also

$$\tilde{x}_{j+1} = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \tilde{x}_j.$$

**Definition:** *Stabilitätsfunktion* der Trapezmethode:

$$R(\mu) := \frac{1 + \frac{\mu}{2}}{1 - \frac{\mu}{2}} \in \mathbb{C}, \quad \mu := \lambda h$$

**Es gilt:**

$$\tilde{x}_{j+1} = R(h\lambda) \tilde{x}_j \quad j = 0, 1, 2, \dots, \quad \text{bzw.} \quad \tilde{x}_n = R(h\lambda)^n x_0.$$

Die Trapezmethode approximiert die Lösung des Modellproblems zur Zeit  $t = nh$ , nämlich

$$e^{\lambda t} x_0 = e^{\lambda nh} x_0 = (e^{\lambda h})^n x_0,$$

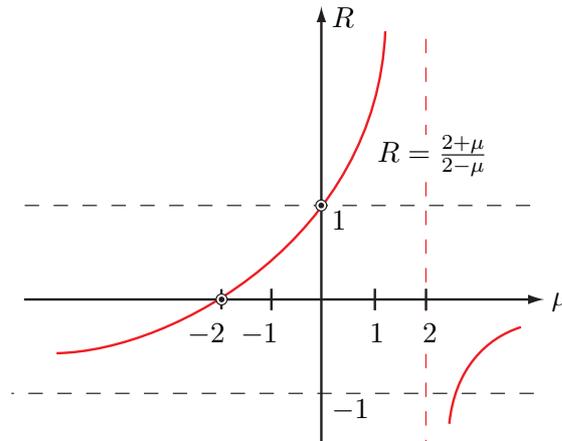
durch

$$(R(\lambda h))^n x_0, \quad \text{wobei} \quad R(\mu) = \frac{2 + \mu}{2 - \mu}.$$

*Bemerkung:* Es gilt:

$$R(\mu) = \left(1 + \frac{\mu}{2}\right) \left(1 + \frac{\mu}{2} + \frac{\mu^2}{4} + \dots\right) = 1 + \mu + \frac{\mu^2}{2} + \dots = e^\mu + O(\mu^3)$$

Für das reelle Stabilitätsintervall der Trapezmethode erhält man (siehe Skizze)  $B_{\text{Trapez}} = (-\infty, 0)$ .



Um das Stabilitätsgebiet  $A$  der Trapezmethode zu bestimmen ( $\lambda \in \mathbb{C}$ ), berechnen wir wieder den Rand  $\partial A = \{\mu \in \mathbb{C} \mid |R(\mu)| = 1\}$  für  $R(\mu) = \frac{2+\mu}{2-\mu}$ .

Es gilt:

$$|R(\mu)| = 1 \Leftrightarrow |R(\mu)|^2 = R \cdot \bar{R} = 1.$$

Wir setzen  $\mu =: u + iv$  und erhalten damit

$$R \cdot \bar{R} = \frac{(2+u+iv)(2+u-iv)}{(2-u-iv)(2-u+iv)} = \frac{(2+u)^2 + v^2}{(2-u)^2 + v^2} \stackrel{!}{=} 1$$

$$\Leftrightarrow (2+u)^2 = (2-u)^2$$

$$\Leftrightarrow 4 + 4u + u^2 = 4 - 4u + u^2$$

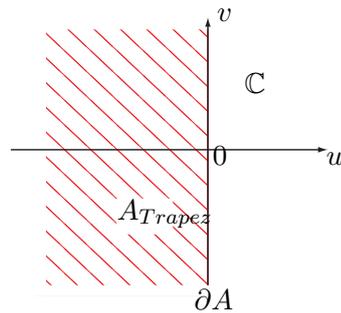
$$\Leftrightarrow 8u = 0$$

$$\Leftrightarrow u = 0$$

Damit haben wir gezeigt, dass das Stabilitätsgebiet der Trapezmethode

$$A_{\text{Trapez}} = \{\mu \in \mathbb{C} \mid \operatorname{Re} \mu < 0\}$$

ist (d.h. die linke Halbebene von  $\mathbb{C}$ ). Man nennt deshalb die Trapezmethode *A-stabil* (absolut stabil).



**Es gilt:** Für ein Runge-Kutta-Verfahren der Fehlerordnung  $p$  ist

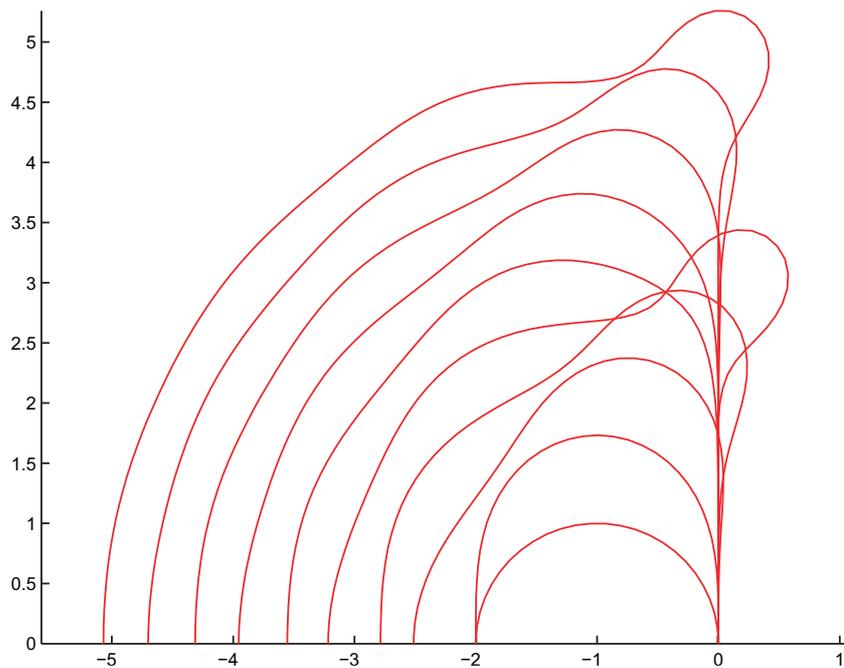
$$R(\mu) = 1 + \mu + \frac{\mu^2}{2!} + \dots + \frac{\mu^p}{p!} + O(\mu^{p+1}),$$

bzw.

$$R(\mu) - e^\mu = O(\mu^{p+1}).$$

*Bemerkung:* Die Umkehrung muss nicht gelten, d.h. wenn ein Einschrittverfahren die obige Beziehung erfüllt, folgt daraus nicht, dass es die Fehlerordnung  $p$  besitzt.

In der folgenden Figur sind die Stabilitätsgebiete des Taylorverfahrens (vergl. Abschn. 8.1.2) der Fehlerordnung  $p = 1, \dots, 9$  dargestellt. (Sie sind symmetrisch bezüglich der  $x$ -Achse.)



### Anwendung des Stabilitätskonzeptes auf steife Differentialgleichungen

*Beispiel:* Wir betrachten das Anfangswertproblem einer Differentialgleichung 2. Ordnung, die wir als zweidimensionales System schreiben:

$$\ddot{x} + 101 \dot{x} + 100 x = 0$$

$$x(0) = 0$$

$$\dot{x}(0) = 0.99$$

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -100 x_1 - 101 x_2$$

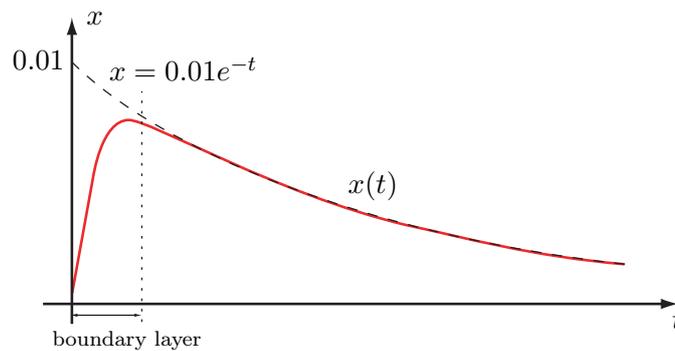
$$x_1(0) = 0, x_2(0) = 0.99$$

Wir lösen das Eigenwertproblem der Koeffizientenmatrix

$$A := \begin{pmatrix} 0 & 1 \\ -100 & -101 \end{pmatrix}$$

und erhalten  $\lambda_1 = -1$ ,  $\lambda_2 = -100$ . Das ergibt folgende Lösung für das obige Anfangswertproblem:

$$x(t) = 0.01e^{-t} - 0.01e^{-100t}.$$



*Numerische Lösung des Anfangswertproblems:*

1. In der boundary-layer-Region muss mit sehr kleinen Schritten integriert werden, um die rasche Änderung der Lösung zu erfassen.
2. Für die Wahl der Schrittweite ausserhalb der boundary-layer-Region machen wir die folgende Betrachtung:

$$t = 0.2 : \quad e^{-100t} = e^{-20} \cong 2 \cdot 10^{-9}$$

$$x(0.2 + nh) = 0.01 \cdot e^{-0.2}(e^{-h})^n - 0.01 \cdot e^{-20}(e^{-100h})^n,$$

d.h. nach der Zeit  $t = 0.2$  ist der 'schnelle Mode'  $e^{-100t}$  der Lösung  $x(t)$  schon kleiner als die Maschinengenauigkeit *eps (IEEE)*,  $x(t)$  'besteht' nur mehr aus dem langsamen Mode  $e^{-t}$ . Für die numerische Approximation ausserhalb der boundary-layer-Region gilt:

$$\tilde{x}(0.2 + nh) = 0.01e^{-0.2}(R(-h))^n - 0.01 \cdot e^{-20}(R(-100h))^n.$$

Für die Schrittweite  $h = 0.05$  erhalten wir:

$$e^{-h} = 0.9512294, \quad R_H(-h) = 0.95125, \quad R_T(-h) = 0.951220$$

$$e^{-100h} = 6.7379447 \cdot 10^{-3}, \quad R_H(-100h) = 8.5, \quad R_T(-100h) = -0.4285714.$$

Beim Verfahren von Heun schaukelt sich der Term  $0.01e^{-20}(R(-100h))^n$  auf, weil  $-100h = -5$  ausserhalb des Stabilitätsgebiets liegt.

**Definition:** Ein lineares (inhomogenes) Differentialgleichungssystem

$$\dot{\underline{x}} = A\underline{x} + \underline{b}, \quad A \text{ } n \times n \text{ - Matrix; } \underline{x}, \underline{b} \in \mathbb{R}^n \quad (8.6)$$

heisst *stif*, falls ein Eigenwert von  $A$  mit negativem Realteil existiert, so dass der Betrag dieses Realteils sehr viel grösser ist als der Betrag des grössten Realteils der anderen Eigenwerte von  $A$ . (Z.B.  $n = 3$ ,  $\lambda_1 = -1000$ ,  $\lambda_2 = -10$ ,  $\lambda_3 = 3$ ;  $|\lambda_1| = 1000 \gg |\lambda_3| = 3$ ).

*Bemerkungen:*

- Nichtlineare Differentialgleichungssysteme  $\dot{\underline{x}} = \underline{f}(t, \underline{x})$  lassen sich lokal (d.h. in der Nähe von  $(t_j, \tilde{x}^j)$ ) immer in der Form (8.6) schreiben, wobei  $A$  die Jacobi-Matrix ist,  $A = A(t_j) = \frac{\partial \underline{f}}{\partial \underline{x}}(t_j, \tilde{x}^j)$ . Somit wird das qualitative Verhalten der Lösung des nichtlinearen Systems lokal durch die Eigenwerte der Jacobi-Matrix beschrieben.
- Wenn man nicht weiss, dass ein Problem nicht-steif ist, sollte man zur numerischen Approximation der Lösung ein implizites Verfahren wählen mit einem 'guten Stabilitätsgebiet', z.B. ein A-stabiles Verfahren.