Vorlesungsskript

N U M E R I K 1

erstellt von Frau Dr. Heike Faßbender basierend auf

- einer Vorlesung gehalten von Frau Prof. Dr. A. Bunse-Gerstner im SS94 an der Universität Bremen
- einer Vorlesung gehalten von Herrn Prof. Dr. V. Mehrmann im WS 94/95 an der Technischen Universität Chemnitz-Zwickau
- einer Vorlesung gehalten von Frau Dr. H. Faßbender im SS95 an der Universität Bremen

Bilder erstellt von Dipl.-Math. Gerd Banse

überarbeitet von Prof. Dr. Bunse-Gerstner im SS 96

Inhaltsverzeichnis

Kapitel 1

Einführung

Numerische Mathematik befaßt sich mit mathematischen Problemen, die (im weitesten Sinn) im Zusammenhang mit der (effizienten) Berechnung mathematischer Größen auftreten. Dabei sollen im allgemeinen die Berechnungen mit Hilfe eines Digitalrechners erfolgen. Wir betrachten einmal ein sehr einfaches Beispiel, um zu sehen, welche Probleme typischerweise auftreten.

Beispiel 1.0.1 *Freier Fall*

Ein Stein wird von der Spitze des Turmes von Pisa (Höhe $44.12m$) fallen gelassen. Nach wieviel Sekunden schlägt er auf?

Zur Lösung dieses Problems bilden wir ein mathematisches Modell. Zur Vereinfachung vernachlässigen wir die Reibung und nehmen an, daß der Stein eine Punktmasse darstellt. Dadurch machen wir natürlich einen gewissen (kleinen) Fehler (Modellierungsfehler). Sei nun $h(t)$ die Höhe des Steins zum Zeitpunkt t, dann ist die Geschwindigkeit des Steins zum Zeitpunkt t gegeben durch die erste Ableitung von h nach der Zeit:

$$v(t) = h'(t).$$

Die Beschleunigung $a(t)$ des Steins ist gerade definiert als die Ableitung der Geschwindigkeit nach der Zeit, also

$$a(t) = v'(t) = h''(t).$$

Die Erdbeschleunigung g hat näherungsweise den Wert $9.81 \frac{m}{S^2}$. Benutzen wir diesen Wert, so machen wir auch wieder einen (kleinen) Fehler (Datenfehler). Man weiß nun, daß gilt

$$a(t) = -g.$$

Die Geschwindigkeit des Steins läßt sich nun durch Integration bestimmen:

$$v(t) = \int a(t)dt = -gt + c_1$$

Weiter läßt sich nun $h(t)$ bestimmen als

$$\begin{aligned} h(t) &= \int v(t)dt \\ &= \int(-gt + c_1)dt \\ &= -\frac{g}{2}t^2 + c_1 t + c_2. \end{aligned}$$

Wir wissen weiter, daß $v(0) = 0\frac{m}{s}$ und $h(0) = 44.12m$ ist. Es muß daher gelten

$$v(0) = -g \cdot 0 + c_1 = 0\frac{m}{s}$$

$$\Longrightarrow c_1 = 0$$

und

$$h(0) = -\frac{g}{2} \cdot 0 + c_1 \cdot 0 + c_2 = 44.12m$$

$$\Longrightarrow c_2 = 44.12.$$

Wir können nun diese Werte einsetzen und erhalten als Lösung

$$h(t) = -\frac{g}{2}t^2 + 44.12.$$

Also ist $h(t) = 0$, wenn $t = \pm\sqrt{\frac{44.12}{g} \cdot 2}$.
Nun können wir einen Rechner benutzen, um diesen Zahlenwert zu ermitteln. Wir erhalten die numerische Lösung (3 Stellen):

$$t = 3.00s.$$

Wir haben hier einen weiteren (kleinen) Fehler gemacht, nämlich beim Wurzelziehen und beim Runden auf 3 Stellen. All diese Fehler können bei komplizierteren Problemen durchaus katastrophal groß werden. Was können wir tun, um sicherzustellen, daß das Ergebnis hinreichend korrekt ist?

Wir könnten nach Pisa fahren oder auf die entsprechende Höhe des Fallturms klettern und ein Experiment machen, oder wir könnten das Modell und unseren Lösungsalgorithmus mathematisch analysieren. Diesen zweiten Weg beschreitet die Numerische Mathematik, obwohl der erste sicher der unterhaltsamere ist.

Allgemein folgt der Lösungsweg vom Anwenderproblem bis zur Lösung, die der Rechner liefert, grob dem nachfolgenden Schema.

<table>
<tr><td colspan="2">Anwendungsproblem</td><td>Beispiel: Freier Fall</td><td>Fehler</td></tr>
<tr><td colspan="2">Modellbildung
Math. Problem</td><td>Diffgl. $h'' = 9.81$</td><td>Modellierungsfehler
Datenfehler</td></tr>
<tr><td>Math. Analyse
ana. Lsg.</td><td>Numer. Methode
numer. Lsg.
Impl. und Analyse</td><td>Lsg. von Diffgl.
$t = 3.00$
Programm Fehleranal.</td><td>Verfahrensfehler
Rundungsfehler</td></tr>
<tr><td>exakte Lsg.</td><td>Näherung</td><td></td><td>Gesamtfehler</td></tr>
</table>

Die Numerische Mathematik beschäftigt sich überwiegend mit dem Weg vom mathematischen Problem zum Ergebnis des Algorithmus. Ein großes Gewicht liegt dabei auf der Entwicklung und Analyse der numerischen Verfahren. Bei der Fehleranalyse und Effizienzbetrachtung spielt aber auch die Implementierung eine wichtige Rolle. Es sei hier bemerkt, daß als neuerer Aspekt hierbei die Rechnerarchitektur des benutzten Rechners bei parallel- vektoriellen Rechnern hinzukommt. Um in diesem Fall zu effizienten Verfahren zu kommen, berücksichtigt man insbesondere die Kommunikation zwischen den Prozessoren und die Vektorstruktur häufig schon bei der Entwicklung des numerischen Algorithmus.

Zur Entwicklung und Analyse numerischer Verfahren benötigt man also Kenntnisse in der Mathematik, Programmierkenntnisse sowie Grundkenntnisse über den Umgang mit Rechnern.

Nachdem wir im Mathematisches Praktikum schon etliche einfache numerische Verfahren kennengelernt haben, die als Bausteine für die numerische Behandlung von größeren Problemen verwendet werden, behandeln wir in dieser Vorlesung als ersten größeren Problembereich die numerische Lösung von gewöhnlichen Differentialgleichungen.

Kapitel 2

Numerische Lösung von Anfangswertaufgaben

2.1 Einleitung

2.1.1 Ein einfaches Beispiel: Transportmodelle für chemische Reaktionen

In einem Reaktor seien drei chemische Substanzen, C, E und D eingebracht, wobei sich C und E beim Aufeinandertreffen mit einer gewissen Rate zum Stoff D verbinden, symbolisch angedeutet durch:

$$C + E \longrightarrow D.$$

Seien c, e, d die Konzentrationen der Stoffe C, E, D in Mol/Volumen. Im allgemeinen hängen sie von den Raumvarianten x, y, z und der Zeit t ab, also $c = c(x, y, z, t), e = e(x, y, z, t)$ und $d = d(x, y, z, t)$.

Die folgenden Überlegungen gelten aber zu jeder Zeit und an jedem Ort, so daß x, y, z, t hier zunächst unterdrückt werden können. Sei g_D die Konzentration des Stoffes D, die pro Zeiteinheit entsteht. Wir nehmen das Gesetz

$$g_D = kce$$

an, wobei k ein Proportionalitätsfaktor ist und schreiben

$$\underline{\text{Reaktion I}}: \qquad C + E \xrightarrow{k} D.$$

Da wir in Mol/Volumen messen, werden entsprechend dem Aufbau von D die Stoffe C und E abgebaut, d.h.

$$g_E = -kce \quad \text{und} \quad g_C = -kce,$$

wobei g_E, g_C die Konzentration des Stoffes E bzw. C sei, die pro Zeiteinheit entsteht. Hängen c, d, e nun nur von der Zeit ab, so erhalten wir

$$\begin{aligned} \frac{d}{dt}c &= g_C = -kce, \\ \frac{d}{dt}e &= -kce, \\ \frac{d}{dt}d &= kce. \end{aligned} \tag{2.1.1}$$

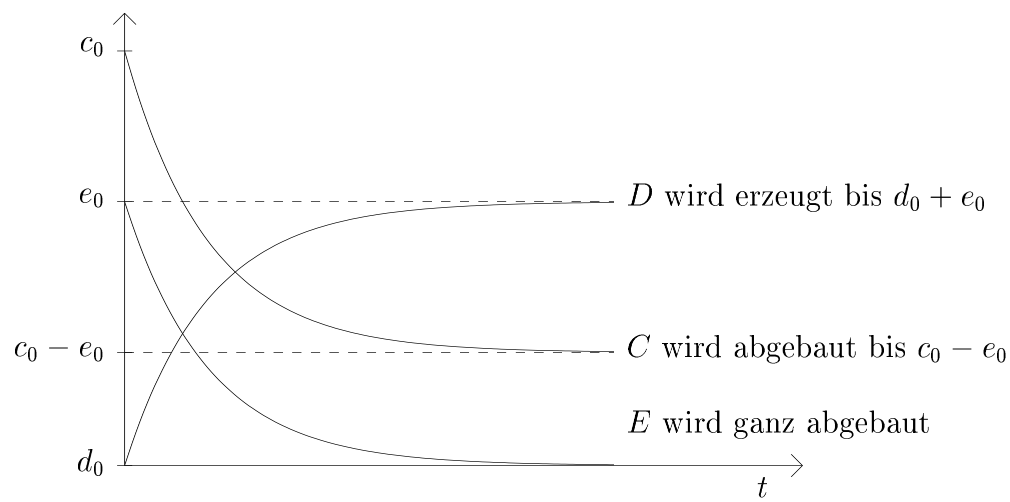
Gesucht sind hier nun die Konzentrationsverläufe in Abhängigkeit von der Zeit, also $c(t), e(t)$ und $d(t)$. (2.1.1) ist ein System von gewöhnlichen Differentialgleichungen für die gesuchten Größen. Die Konzentrationsverläufe werden eindeutig durch die Anfangskonzentrationen

$$\begin{aligned} c(0) &= c_0, \\ e(0) &= e_0, \\ d(0) &= d_0 \end{aligned} \tag{2.1.2}$$

bestimmt. Man bezeichnet (2.1.1) und (2.1.2) als Anfangswertaufgabe. Als Lösung erhält man im Fall $c_0 > e_0$, wie man leicht nachrechnet:

$$\begin{aligned} c(t) &= \frac{c_0(c_0 - e_0)}{c_0 - e_0 exp(-k(c_0 - e_0)t)} \\ e(t) &= c(t) - (c_0 - e_0) = (c_0 - e_0)\frac{e_0 exp(-k(c_0 - e_0)t)}{c_0 - e_0 exp(-k(c_0 - e_0)t)} \\ d(t) &= d_0 + \int_0^t kc(s)e(s)ds = d_0 - \int_0^t e'(s)ds = d_0 + e_0 - e(t) \end{aligned}$$

Graphisch dargestellt erhält man: $(d_0 = 0)$



Für $c_0 < e_0$ werden die Rollen von c und e vertauscht. Die obigen Formeln gelten weiterhin, aber man kürzt am besten durch $exp(-k(c_0 - e_0)t)$, so daß das Verhalten für $t \longrightarrow \infty$ ablesbar wird. Für $c_0 = e_0$ findet man $c(t) = \frac{c_0}{1-kc_0t}$ und $e(t) = \frac{e_0}{1-ke_0t}$.

Oft wirkt ein chemischer Stoff auch als Katalysator, d.h. er wird zwar zur Reaktion benötigt, geht aber unversehrt aus ihr hervor. Ist E etwa ein Katalysator für die Reaktion, so deutet man dies symbolisch an durch:

$$\underline{\text{Reaktion II}}: \qquad C + E \xrightarrow{k} D + E.$$

Hier lautet die Anfangswertaufgabe jetzt

$$\begin{array}{ll} c' = g_C = -kce & c(0) = c_0 \\ e' = g_E = 0 & e(0) = e_0 \\ d' = g_D = kce & d(0) = d_0 \end{array}$$

Man erhält hier als Lösung:

$$\begin{aligned} e(t) &= e_0 \\ c(t) &= c_0 e^{-ke_0t} \\ d(t) &= d_0 + c_0(1 - e^{-ke_0t}) \end{aligned}$$

Verfeinert man die Betrachtungen, indem man den durch Konzentrationsgefälle verursachten Substanztransport modelliert, so bekommt man partielle Differentialgleichungen für c, e, und d.

2.1.2 Allgemeine Problemstellung

Hier haben wir gesehen, daß die von nur einer Variablen abhängigen Funktionen durch ein Differentialgleichungssystem und die Festlegung eines Wertes (des Anfangswertes) eindeutig bestimmt waren. Das ist nicht für jede beliebige Kombination von Differentialgleichungssystemen und Anfangswerten der Fall. Wir betrachten daher zunächst kurz die Frage der Existenz und Eindeutigkeit von Lösungen der Differentialgleichung mit vorgebenen Anfangswert, wobei wir vorab noch die Problemstellung präzisieren.

Die allgemeine Problemstellung lautet:

Gegeben sei ein Intervall $\mathrm{I} = [t_0, t_0 + a] \subseteq \mathbb{R}, G \subseteq \mathbb{R}^n$, eine Funktion $f : I \times G \longrightarrow \mathbb{R}^n$ und einen Vektor $y_0 \in \mathbb{R}^n$.

Gesucht ist eine Funktion $y : [t_0, t_0 + a] \longrightarrow \mathbb{R}^n$ mit

$$y'(t) = f(t, y(t)) \quad \text{für alle} \quad t \in I \tag{2.1.3}$$

$$y(t_0) = y_0 \tag{2.1.4}$$

Die Aufgabe, eine solche Funktion y mit (2.1.3) und (2.1.4) zu bestimmen, heißt Anfangswertaufgabe zum System gewöhnlicher Differentialgleichungen (2.1.3).

In unserem Beispiel im Abschnitt 2.1.1 war $n = 3$ und

$$y(t) = \begin{pmatrix} c(t) \\ e(t) \\ d(t) \end{pmatrix} \quad f(t,y) = \begin{pmatrix} -ky_1y_2 \\ -ky_1y_2 \\ ky_1y_2 \end{pmatrix} \quad y_0 = \begin{pmatrix} c_0 \\ e_0 \\ d_0 \end{pmatrix}.$$

(2.1.3) nennt man eine gewöhnliche Differentialgleichung erster Ordnung, wobei „gewöhnlich" bedeutet, daß nur ein Parameter der Funktion als Variable betrachtet und nach diesem differenziert wird, (im Gegensatz zu „partiellen" Differentialgleichungen) und „erster Ordnung" bedeutet, daß nur die erste Ableitung auftritt. Man überzeugt sich aber leicht, daß Differentialgleichungen mit höherer Ordnung auf Differentialgleichungen erster Ordnung zurückgeführt werden können. Denn haben wir allgemeiner $g : \mathbb{R}^{n \cdot m+1} \longrightarrow \mathbb{R}^n$ gegeben, und ist $y : I \longrightarrow \mathbb{R}^n$ gesucht mit

$$\begin{aligned} y^{(m)}(t) &= g(t, y(t), y'(t), \ldots, y^{(m-1)}(t)) \\ y^{(i)}(t_0) &= y_{0i} \quad \text{für} \quad i = 0, \ldots, m-1 \end{aligned}$$

so bezeichnet man dies als Anfangswertaufgabe für eine Differentialgleichung m-ter Ordnung.

Setzt man nun:

$$z_1 := y, \quad z_i := y^{(i-1)} \quad \text{für} \quad i = 2, \ldots, m$$

dann gilt:

$$\begin{aligned} z_m'(t) &= y^{(m)}(t) = g(t, z_1(t), z_2(t), \ldots, z_m(t)) \\ z_i(t_0) &= y^{(i-1)}(t_0) = y_{0,i-1} \quad \text{für} \quad i = 1, \ldots, m \end{aligned}$$

Daher erhält man mit

$$z(t) = \begin{pmatrix} z_1(t) \\ z_2(t) \\ \vdots \\ z_m(t) \end{pmatrix}, z_0 = \begin{pmatrix} y_{0,0} \\ y_{0,1} \\ \vdots \\ y_{0,m-1} \end{pmatrix}:$$

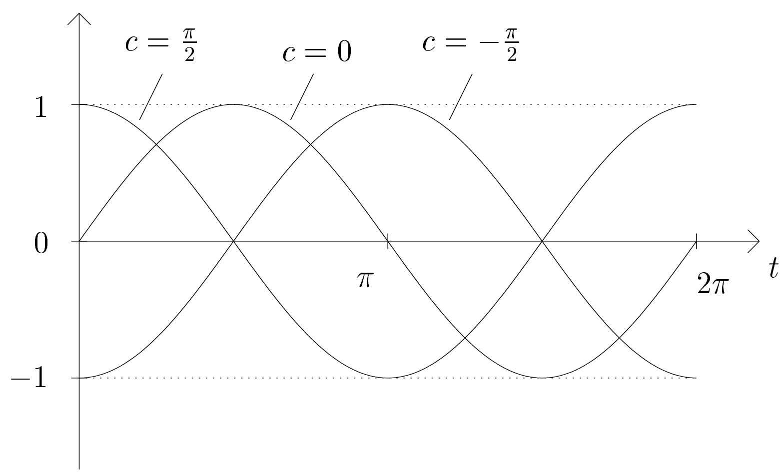
$$z'(t) = \begin{pmatrix} z_1'(t) \\ z_2'(t) \\ \vdots \\ z_m'(t) \end{pmatrix} = \begin{pmatrix} z_2(t) \\ z_3(t) \\ \vdots \\ z_m(t) \\ g(t, z_1(t), \ldots, z_m(t)) \end{pmatrix} =: f(t, z(t))$$

Es genügt also, im folgenden (2.1.3) und (2.1.4) zu betrachten. Im wesentlichen kann man sich dabei auch auf $n = 1$ beschränken.

2.1.3 Einiges aus der Theorie der gewöhnlichen Differentialgleichungen

In vielen Anwendungsproblemen läßt sich keine geschlossene Lösung $y(t)$ von (2.1.3), (2.1.4) angeben und man ist gezwungen, diese Aufgabe numerisch zu lösen. Dies wird das Thema dieser Veranstaltung sein. Nichtsdestotrotz ist es wichtig, die analytischen Resultate und Methoden zur Behandlung von gewöhnlichen Differentialgleichungen zu kennen. Daher werden wir zuerst eine kleine Wiederholung der theoretischen Resultate machen. Wichtige Fragen, die zunächst einmal zu klären sind, sind: Existiert eine Lösung von (2.1.3), (2.1.4)? Ist sie eindeutig? Hängt sie stetig von den gegebenen Daten ab?

Beispiel 2.1.1 *i)* $\quad y' = \pm\sqrt{1-y^2}$
Als Lösungen ergeben sich $y(t) = \sin(t+c)$ *für jedes* $c \in \mathbb{R}$*, so daß das Vorzeichen der Ableitung stimmt, sowie* $y \equiv 1$ *und* $y \equiv -1$ *sowie Kombinationen hieraus:*



Für $y(a) = 4$ *gibt es keine Lösung, a beliebig.*
Für $y(0) = 1$ *und den negativen Wurzelwert lauten die Lösungen:* $y(t) = \sin(t + \frac{\pi}{2})$ *und* $y \equiv 1$.
Für $y(0) = 0$ *und den positiven Wurzelwert lautet die Lösung:* $y(t) = \sin(t)$.

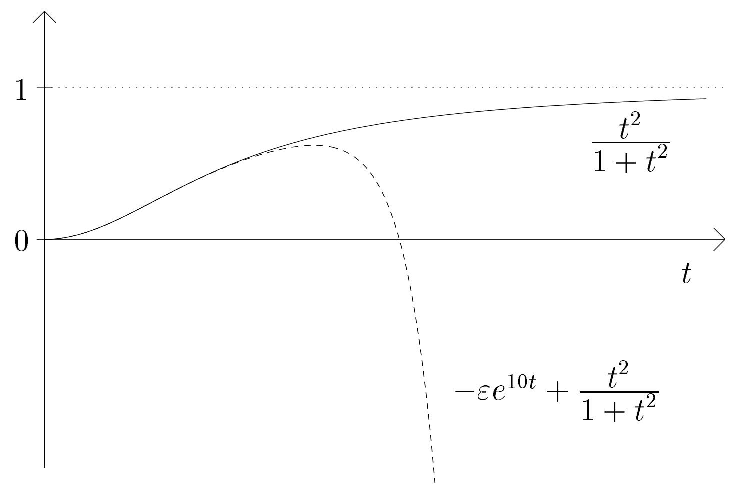
ii) $$y' = 10\left(y - \frac{t^2}{1+t^2}\right) + \frac{2t}{(1+t^2)^2} \qquad y(0) = y_0$$

Als Lösung ergibt sich: $\quad y(t) = y_0 e^{10t} + \dfrac{t^2}{1+t^2}.$

Für $y_0 = 0$ *lautet die Lösung:* $\dfrac{t^2}{1+t^2} = \hat{y}(t)$

Für $y_0 = -\varepsilon$ *lautet die Lösung:* $-\varepsilon e^{10t} + \dfrac{t^2}{1+t^2} = \tilde{y}(t)$

Graphisch dargestellt:



Hier existiert zu jedem ε eine Lösung, die eindeutig ist und die Lösung hängt stetig von den Anfangsdaten ε ab. Aber wir sehen hier eine sehr typische Erscheinung. Es ist nur lokal (d.h. in der Höhe des Anfangspunktes $t = 0$) richtig, daß bei kleiner Abweichung ε in den Daten auch die Lösung nur sehr wenig abweicht. Sobald wir uns von $t = 0$ weiter entfernen, wird die Abweichung sehr groß. Der „Fehler“ wächst exponentiell an. Kleine Störungen in den Anfangsdaten führen zu großen Störungen in der Lösung, d.h. die Lösung ist empfindlich gegenüber Störungen (Instabilität der Differentialgleichung, nicht zu verwechseln mit Instabilität eines numerischen Verfahrens).

Einer der Sätze, mit denen in der Theorie gewöhnlicher Differentialgleichungen die Existenz und Eindeutigkeit der Lösung gezeigt wird, sei hier kurz zitiert:

Satz 2.1.1 *Picard-Lindelöf*

$f : [t_0, t_0 + a] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *sei stetig auf* $R = \{(t, y) | t_0 \leq t \leq t_0 + a, \|y - y_0\| \leq b\}$, *wobei* $b > 0$ *und* $\|\cdot\|$ *eine beliebige Norm in* $\mathbb{R}^n$ *ist, und* f *genüge dort einer* <u>*Lipschitzbedingung*</u>*, d.h. es gibt eine Konstante L mit*

$$\|f(t, y_1) - f(t, y_2)\| \leq L\|y_1 - y_2\|.$$

Sei

$$M := \|f\|_{\infty,R} \quad \alpha := min\{a, \frac{b}{M}\}.$$

Dann hat die Anfangswertaufgabe

$$y' = f(t, y), \quad y(t_0) = y_0$$

eine eindeutige Lösung $y(t)$ *auf* $[t_0, t_0 + \alpha]$.

Beweis: siehe in den Büchern über gewöhnliche Differentialgleichungen z.B. Knobloch, Kappel; Walter; Heuser usw. auch Stoer, Bulirsch.

Der Beweis konstruiert eine Folge von Funktionen

$$\begin{aligned} y_0(t) &\equiv y_0 \\ y_{n+1}(t) &= y_0 + \int_{t_0}^{t} f(s, y_n(s))ds \end{aligned}$$

und zeigt, daß dies eine Cauchy-Folge ist und der Grenzwert $y(t)$ die Anfangswertaufgabe löst. Die wesentliche Ungleichung ist dabei

$$\|y_{\nu+1}(t) - y_\nu(t)\| \leq ML^\nu \frac{(t-t_0)^{\nu+1}}{(\nu+1)!},$$

die induktiv leicht zu zeigen ist. Man erhält damit dann

$$\|y_{n+\nu}(t) - y_\nu(t)\| \leq \frac{M}{L} \sum_{k=\nu}^{\infty} \frac{L^{k+1}(t-t_0)^{k+1}}{(k+1)!},$$

was hier bis auf den Faktor $\frac{M}{L}$ gerade der Rest der Reihenentwicklung von $exp(L(t-t_0))$ ist. Für gegebenes ε muß dieser Ausdruck daher bei hinreichend großem ν kleiner als ε werden. Damit haben wir hinreichende Bedingungen für Existenz und Eindeutigkeit. Wie aber bereits Beispiel 2.1.1 *ii*) zeigt, reicht dies in der Praxis eventuell nicht aus, weil kleine Störungen die Lösung stark verändern können. Um dieses Phänomen zu untersuchen, müssen wir die Abhängigkeit der Lösung von den Eingangsdaten genauer untersuchen. Dies ist die Frage nach der Kondition des Problems. Doch zunächst noch eine Bemerkung zum Satz von Picard-Lindelöf.

Bemerkung 2.1.1 - *Analoge Aussagen erhält man für* $t_0 - a \leq t \leq t_0$ *und* $t_0 - a \leq t \leq t_0 + a$.

- *Falls alle* $\frac{\partial f_i}{\partial y_j}$ *existieren und auf* R *stetig und beschränkt sind, so ist die Lipschitzbedingung erfüllt. Denn nach dem Mittelwertsatz gilt:*

$$\begin{aligned} \|f(t,y_1) - f(t,y_2)\| &\leq \|\int_0^1 D_y f(t, y_1 + \xi(y_1 - y_2))(y_1 - y_2) d\xi\| \\ &\leq \int_0^1 \|D_y f(t, y_1 + \xi(y_1 - y_2))\| \|y_1 - y_2\| d\xi \leq L\|y_1 - y_2\| \end{aligned}$$

wobei L *das Maximum von* $\|D_y f(t, \eta)\|$ *auf* R *ist.*

- *Sätze dieser Art zeigt man häufig mit $\{(t,y)|a \leq t \leq b, y \in \mathbb{R}^n\}$, einem 'Streifen', statt mit R wie in obigem Satz. In den Anwendungen ist oft f stetig differenzierbar auf solch einem Streifen, aber $\frac{\partial f_i}{\partial y_j}$ unbeschränkt. Es sind dann wieder Einschränkungen wie im Satz erforderlich.*
- *Die Einschränkung auf $[t_0, t_0+\alpha], \alpha \leq a$ ist nötig, da $\|y(t) - y_0\| = \| \int_{t_0}^t f(s, y(s))ds\| \leq M(t-t_0)$. Man bleibt in R nur für $t - t_0 \leq \frac{b}{M}$.*

Die folgende Gronwall'sche Ungleichung kann man benutzen, um im Satz von Picard-Lindelöf die Eindeutigkeit der Lösung zu beweisen. Man sieht mit ihrer Hilfe, wie die Lösungen mit den Anfangswerten abweichen können.

Satz 2.1.2 *Gronwall'sche Ungleichung*
Seien $u(t), v(t)$ nichtnegative stetige Funktionen auf $[t_0, t_0+a]$ und $C \geq 0$ eine Konstante. Sei

$$v(t) \leq C + \int_{t_0}^t u(s)v(s)ds \quad \text{für} \quad t_0 \leq t \leq t_0 + a \tag{2.1.5}$$

Dann gilt

$$v(t) \leq C\exp(\int_{t_0}^t u(s)ds) \quad \text{für} \quad t_0 \leq t \leq t_0 + a.$$

<u>Beweis:</u>

Zunächst sei $C > 0$. Setze $y(t) = C + \int_{t_0}^t u(s)v(s)ds$. Dann ist $v(t) \leq y(t)$ und $0 < C \leq y(t)$ für alle $t \in [t_0, t_0+a]$. Ferner gilt

$$y'(t) = u(t)v(t) \leq u(t)y(t) \quad \forall t \in [t_0, t_0+a].$$

Damit erhält man $\frac{y'(t)}{y(t)} \leq u(t)$, also auch

$$\int_{t_0}^t \frac{y'(s)}{y(s)} ds \leq \int_{t_0}^t u(s)ds.$$

Also gilt

$$ln(y(t)) - ln(y(t_0)) = ln\left(\frac{y(t)}{C}\right) \leq \int_{t_0}^t u(s)ds$$

und daher

$$y(t) \leq C\exp(\int_{t_0}^t u(s)ds).$$

Gilt (2.1.5) mit $C = 0$, so konstruiert man eine Folge $\{c_i\}_{i=1}^\infty, \lim_{i\to\infty} c_i = C, c_i > 0 \quad \forall i$. Anwendung des ersten Teils ergibt $v(t) \leq c_i \exp(\int_{t_0}^t u(s)ds)$ für alle i und damit folgt $v(t) \leq C(\exp(\int_{t_0}^t u(s)ds)$ auch für $C = 0$. $\surd$

Mit diesem Satz erhalten wir eine Aussage zur stetigen Abhängigkeit der Lösung von den Anfangsdaten. Seien die Voraussetzungen des Satzes von Picard-Lindelöf gegeben, wobei allerdings R durch $R_D = \{(t,y) | t \in [t_0, t_0 + a], \|y - z\| \leq b, z \in D\}$ ersetzt sei und D eine Umgebung des Anfangswertes y_0 ist. Dann existiert nach dem Satz von Picard-Lindelöf eine eindeutige Lösung $y(t,z)$ von

$$\begin{aligned} &y'(t) = f(t, y(t)) \quad \text{für jedes} \quad z \in D, \\ &y(t_0) = z. \end{aligned}$$

Satz 2.1.3 *Mit obigen Voraussetzungen gilt für alle* $z_1, z_2 \in D$*:*

$$\|y(t, z_1) - y(t, z_2)\| \leq e^{L(t-t_0)} \|z_1 - z_2\|$$

für alle $\quad t \in [t_0, t_0 + \alpha]$.

<u>Beweis:</u>

Es ist $y(t, z_i) = z_i + \int_{t_0}^t f(s, y(s, z_i)) ds \quad$ für $\quad i \in \{1, 2\}$.

Daher gilt:

$$\begin{aligned} \|y(t, z_1) - y(t, z_2)\| &\leq \|z_1 - z_2\| + \int_{t_0}^t \|f(s, y(s, z_1)) - f(s, y(s, z_2))\| ds \\ &\leq \|z_1 - z_2\| + L \int_{t_0}^t \|y(s, z_1) - y(s, z_2)\| ds. \end{aligned}$$

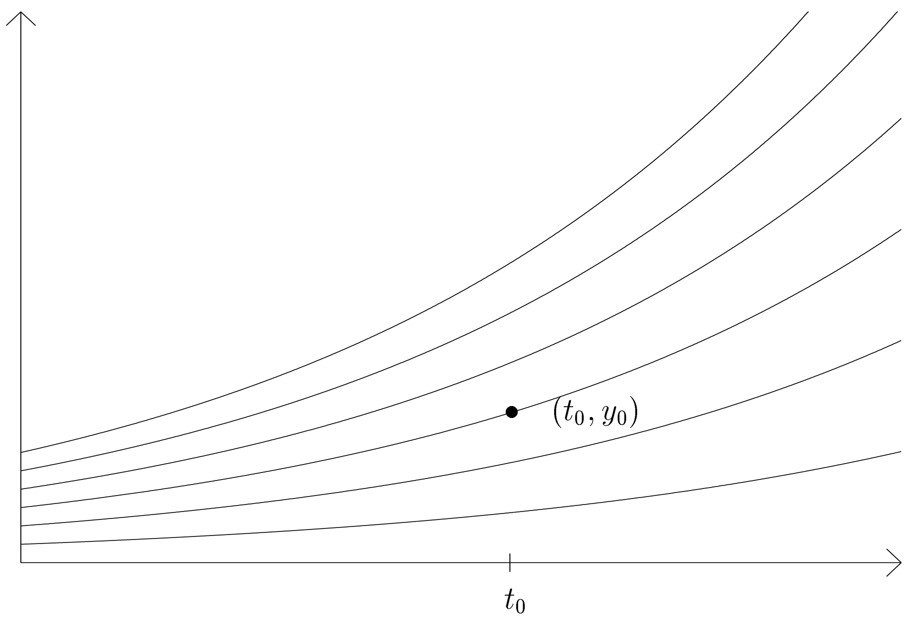
Die Behauptung folgt aus der Gronwall'schen Ungleichung mit $v(t) = \|y(t, z_1) - y(t, z_2)\|, C = \|z_1 - z_2\|$ und $u(t) \equiv L$. $\surd$

Man sieht, daß die Konstante e^{Lt} mit t exponentiell wächst, so daß für sehr große t die beiden Lösungen recht weit auseinanderliegen können (siehe Beispiel 2.1.1 *ii*). Das Problem der Anfangswertaufgabe ist also zumindest dann gut konditioniert, wenn e^{Lt} klein ist.

2.2 Einschrittverfahren

Zur Lösung von gewöhnlichen Differentialgleichungen gibt es viele verschiedene analytische Methoden: Variablensubstitution, Transformationsmethoden (wie Laplace-, Fouriertransformation), Potenzreihenansätze etc. Häufig gibt es jedoch keine geschlossene Lösung, so daß eine numerische Lösung erforderlich ist. Grundidee der numerischen Lösung ist die Diskretisierung des kontinuierlichen Problems. Im folgenden sei zunächst die Dimension $n = 1$. In jedem Punkt (t, y) kennen wir die Steigung der Lösung durch $y' = f(t, y)$. Die Lösung definiert also ein Richtungsfeld.

Beispiel 2.2.1 $y'(t) = 3y(t), y(0) = a \Rightarrow y(t) = ae^{3t}$



In jedem Punkt (t_0, y_0) gibt es eine Trajektorie $y(t)$, die die Differentialgleichung löst und durch (t_0, y_0) geht. Numerische Lösung heißt, daß wir diese Trajektorie approximieren müssen. Dies kann man auf viele verschiedene Arten tun. Die einfachsten Methoden sind die sogenannten expliziten Einzelschrittverfahren, die wir uns nun zuerst anschauen wollen.

2.2.1 Das Euler-Verfahren

Wir gehen dazu wie folgt vor: Sei $I = [t_0, t_0+a]$ und $I_h = \{t_0, t_1, \ldots, t_N = t_0+a\}$ eine Menge von Stützstellen t_i mit $t_i = t_{i-1} + h_i, i = 1, \ldots, N$ mit Schrittweiten h_i. Zunächst sei der Einfachheit halber $h_i = h$ für alle $i = 1, \ldots, N$, d.h. $h = \frac{a}{N}, t_i = t_0 + ih$. Gesucht sind nun Werte u_i, die die Werte $y(t_i)$ der exakten Lösung der Anfangswertaufgabe

$$y' = f(t, y(t)), \quad y(t_0) = y_0 \tag{2.2.6}$$

gut approximieren. (Vielfach wird die Notation $u_i = u_h(t_i)$ verwendet, wobei u_h eine Gitterfunktion auf I_h ist). Für kleine h ist

$$\frac{y(t+h)-y(t)}{h} = y'(t+\theta h)$$

mit einem $\theta \in]0,1[$. Ersetzt man die Ableitung in $y'(t) = f(t, y(t))$ durch diese Differenz, so erhält man

$$\frac{y(t+h)-y(t)}{h} \approx f(t, y(t))$$

und daher

$$y(t+h) \approx y(t) + hf(t, y(t)).$$

Da für $t = t_0$ der y-Wert bekannt ist, kann man einen Näherungswert u_1 für $y(t_0+h)$ ausrechnen:

$$u_1 = y_0 + hf(t_0, y_0).$$

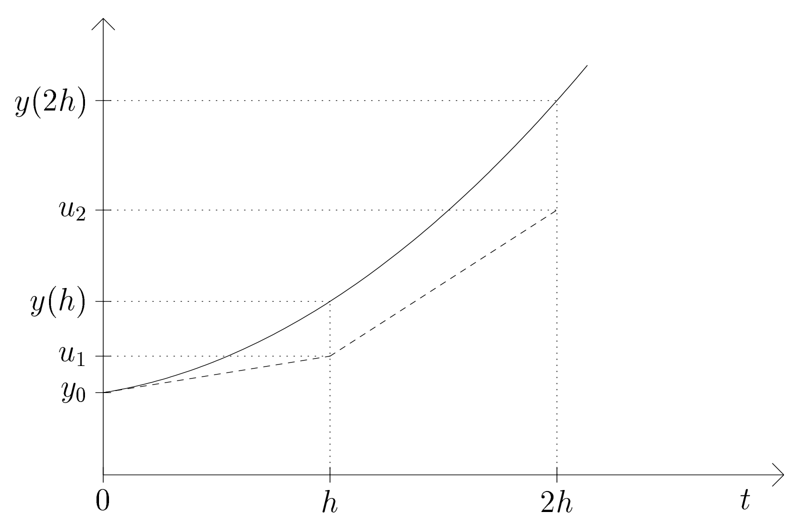
Mit diesem Näherungswert anstelle von $y(t_0+h)$ erhält man dann weiter einen Näherungswert u_2 für $y(t_0+2h)$ aus

$$u_2 = u_1 + hf(t_0+h, u_1)$$

Wir ersetzen also (2.2.6) durch

$$\begin{aligned} u_{i+1} &= u_i + hf(t_i, u_i) \quad i = 0, \ldots, N-1 && (2.2.7)\\ u_0 &= y_0 && (2.2.8) \end{aligned}$$

Geometrisch läßt sich das Verfahren folgendermaßen deuten:



Wir erhalten mit diesem Vorgehen das

Euler-Verfahren

$$\begin{aligned} \text{Setze} \quad h &:= \frac{a}{N}, \quad t_i := t_0 + ih \\ u_0 &:= y_0 \\ u_{i+1} &:= u_i + hf(t_i, u_i) \quad \text{für} \quad i = 0, \ldots, N-1 \end{aligned}$$

Das Euler-Verfahren ist der einfachste und älteste Vertreter der expliziten Einschrittverfahren. Allgemeiner kann man die Näherung durch eine allgemeinere Funktion $\phi(t_i, u_i, h, f)$, einer Inkrementfunktion, ersetzen. Man erhält:

Allgemeines explizites Einschrittverfahren:

$$\begin{aligned} \text{Setze} \quad h &:= \frac{a}{N}, \quad t_i := t_0 + ih \\ u_0 &:= y_0 \\ u_{i+1} &:= u_i + h\phi(t_i, u_i, h, f) \quad \text{für} \quad i = 0, \ldots, N-1 \end{aligned}$$

(Beim Euler-Verfahren: $\phi(t_i, u_i, h, f) = f(t_i, u_i)$)

2.2.2 Diskretisierungsfehler, Konsistenz

Wie gut approximieren die u_i die exakten Werte $y(t_i)$? D.h. man möchte Abschätzungen für den globalen Diskretisierungsfehler:

$$e_j = y(t_j) - u_j \quad j = 0, 1, 2, \ldots$$

Um ihn abschätzen zu können, muß man zunächst den lokalen Fehler betrachten. Sei dazu $y(t)$ die Lösung von $y'(t) = f(t, y(t)), y(\hat{t}) = z$. Das Einschrittverfahren mit $\hat{t}$ statt t_0 und $u_0 = z$ liefert als Näherung u_1 für $y(\hat{t} + h)$:

$$u_1 = u_0 + h\phi(\hat{t}, u_0, h, f).$$

Der lokale Fehler pro Schritteinheit ist dann:

$$\begin{aligned} \frac{y(\hat{t}+h) - u_1}{h} &= \frac{1}{h}[\underbrace{y(\hat{t})}_{=z} + y(\hat{t}+h) - y(\hat{t}) \underbrace{-u_0}_{=z} - h\phi(\hat{t}, u_0, h, f)] \\ &= \underbrace{\frac{y(\hat{t}+h) - y(\hat{t})}{h}}_{\text{Steigung der Sekante der Lösung}} - \underbrace{\phi(\hat{t}, u_0, h, f)}_{\text{Näherung der Steigung}} \\ &=: \tau(\hat{t}, z, h, f) \end{aligned}$$

Setze

$$\triangle(\hat{t}, z, h, f) := \begin{cases} \dfrac{y(\hat{t}+h) - y(\hat{t})}{h} & \text{für} \quad h \neq 0 \\ f(\hat{t}, z) & \text{für} \quad h = 0 \end{cases} .$$

Dies ist der Differenzenquotient der Lösung y zur Schrittweite h bei $\hat{t}$. Dann bezeichnet man

$$\tau(\hat{t}, z, h, f) = \triangle(\hat{t}, z, h, f) - \phi(\hat{t}, z, h, f)$$

als lokalen Diskretisierungsfehler.

Wegen

$$\underbrace{y(\hat{t}+h) = u_0 + h\phi(\hat{t}, u_0, h, f)}_{\text{Verfahrensgleichung für } y(\hat{t})} + \underbrace{h\tau(\hat{t}, z, h, f)}_{\text{Fehler } y(\hat{t}+h) - u_1}$$

gibt er an, wie gut die exakte Lösung die Gleichung des Einschrittverfahrens erfüllt. Damit ein Einzelschrittverfahren vernünftig ist, fordern wir

$$\lim_{h \to 0} \tau(\hat{t}, z, h, f) = 0. \tag{2.2.9}$$

Da aber $\lim_{h\to 0} \triangle(\hat{t}, z, h, f) = y'(\hat{t}) = f(\hat{t}, z)$ ist, so ist (2.2.9) equivalent zu

$$\lim_{h \to 0} \phi(\hat{t}, z, h, f) = f(\hat{t}, z).$$

Man nennt ϕ bzw. das Einschrittverfahren konsistent, wenn für alle $\hat{t} \in [t_0, t_0 + a], z \in \mathbb{R}$ und alle f mit stetigen beschränkten partiellen Ableitungen, (2.2.9) gilt.

Beispiel 2.2.2 *Das Euler-Verfahren ist konsistent. Genauer kann man sagen (falls f genügend oft differenzierbar ist):*

$$y(\hat{t}+h) = y(\hat{t}) + hy'(\hat{t}) + \frac{h^2}{2}y''(\hat{t}) + \ldots + \frac{h^n}{n!}y^{(n)}(\hat{t} + \theta h) \quad 0 < \theta < 1$$

Dabei ist nun $y(\hat{t}) = z, y'(\hat{t}) = f(\hat{t}, y(\hat{t}))$ und

$$\begin{aligned} y''(\hat{t}) &= \frac{d}{dt} f(t, y(t))|_{t=\hat{t}} \\ &= f_t(t, y(t))|_{t=\hat{t}} + f_y(t, y(t)) \cdot y'(t)|_{t=\hat{t}} \\ &= f_t(\hat{t}, z) + f_y(\hat{t}, z) f(\hat{t}, z) \\ y'''(\hat{t}) &= f_{tt}(\hat{t}, z) + 2f_{ty}(\hat{t}, z) f(\hat{t}, z) + f_{yy}(\hat{t}, z) f(\hat{t}, z)^2 + f_y(\hat{t}, z) y''(\hat{t}) \end{aligned}$$

Also erhält man:

$$\begin{aligned} \triangle(\hat{t}, z, h, f) &= y'(\hat{t}) + \frac{h}{2} y''(\hat{t}) + \ldots \\ &= f(\hat{t}, z) + \frac{h}{2}[f_t(\hat{t}, z) + f_y(\hat{t}, z) f(\hat{t}, z)] + \ldots \end{aligned}$$

und wegen $\phi(\hat{t}, z, h, f) = f(\hat{t}, z)$ ergibt sich hieraus für das Eulerverfahren

$$\begin{aligned} \tau(\hat{t}, z, h, f) &= \triangle(\hat{t}, z, h, f) - \phi(\hat{t}, z, h, f) \\ &= \frac{h}{2}[f_t(\hat{t}, z) + f_y(\hat{t}, z) f(\hat{t}, z)] + \ldots \\ &= O(h) \end{aligned}$$

Man sagt, ein Einschrittverfahren ist <u>konsistent von der Ordnung p,</u> wenn $\tau(\hat{t}, z, h, f) = O(h^p)$ für alle $\hat{t} \in [t_0, t_0 + a]$, alle $z \in \mathbb{R}$ und alle genügend glatten f gilt. Es folgt, daß das Euler-Verfahren konsistent von der Ordnung 1 ist. Verfahren höherer Ordnung könnte man zum Beispiel so gewinnen, daß man ϕ so wählt, daß die Taylorentwicklung von τ mit möglichst hoher $h-$Potenz beginnt. So ist etwa

$$\phi(\hat{t}, z, h, f) = f(\hat{t}, z) + \frac{h}{2}[f_t(\hat{t}, z) + f_y(\hat{t}, z) f(\hat{t}, z)]$$

ein Verfahren 2ter Ordnung. Ein Nachteil ist hierbei allerdings, daß Ableitungen von f benötigt werden, die im allgemeinen nicht verfügbar sind.

2.2.3 Modifiziertes Euler-Verfahren, Verfahren von Heun, klassisches Runge-Kutta-Verfahren

Einfache Verfahren höherer Ordnung erhält man mit dem Ansatz

$$\phi(t, z, h, f) = a_1 f(t, z) + a_2 f(t + p_1 h, z + p_2 h f(t, z))$$

Man bestimmt nun a_1, a_2, p_1, p_2 so, daß die Entwicklung von τ mit möglichst hoher $h-$Potenz beginnt. Die Taylorentwicklung von $\phi(t, z, h, f)$ ergibt

$$\phi(t, z, h, f) = (a_1 + a_2) f(t, z) + a_2 h [p_1 f_t(t, z) + p_2 f_z(t, z) f(t, z)] + O(h^2).$$

Für Verfahren 2-ter Ordnung ergibt daher sich:

$$a_1 + a_2 = 1 \quad a_2 p_1 = \frac{1}{2} \quad a_2 p_2 = \frac{1}{2}.$$

Lösungen sind:

1) $a_1 = \frac{1}{2}, a_2 = \frac{1}{2}, p_1 = p_2 = 1$
   Das ergibt das <u>Verfahren von Heun</u> (1900) mit

$$\phi(t, z, h, f) = \frac{1}{2}[f(t, z) + f(t + h, z + h f(t, z)],$$

   welches 2 Auswertungen von f pro Schritt hat.

2) $a_1 = 0, a_2 = 1, p_1 = p_2 = \frac{1}{2}$
Das ist das modifizierte Euler-Verfahren von Collatz (1960) mit

$$\phi(t,z,h,f) = f(t + \frac{h}{2}, z + \frac{h}{2} f(t,z))$$

das ebenfalls 2 Auswertungen von f pro Schritt hat.

Noch allgemeiner kann man ansetzen:

$$\phi(t,z,h,f) = \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4]$$

mit

$$\begin{aligned} k_1 &:= f(t,z) \\ k_2 &:= f(t + \frac{h}{2}, z + \frac{h}{2} k_1) \\ k_3 &:= f(t + \frac{h}{2}, z + \frac{h}{2} k_2) \\ k_4 &:= f(t + h, z + h k_3) \end{aligned}$$

So erhält man das klassische Runge-Kutta-Verfahren (1895). Es ist ein Einschrittverfahren der Ordnung 4. Pro Schritt sind 4 Auswertungen von f erforderlich.

Bemerkung 2.2.1 *i) Hängt f nicht von z ab, so ist die Lösung des Anfangswertproblems $y'(t) = f(t), \quad y(\hat{t}) = z$, gerade $y(t) = z + \int_{\hat{t}}^{t} f(s) ds$. Dann erhält man mit den obigen Verfahren Näherungen:*

- *Euler:* $\quad y(t) \approx z + hf(\hat{t})$
- *Heun:* $\quad y(t) \approx z + \frac{h}{2}[f(\hat{t}) + f(\hat{t} + h)]$
  Dies entspricht der Trapezsumme.
- *modifizierter Euler:* $\quad y(t) \approx z + hf(\hat{t} + \frac{h}{2})$
  Dies entspricht der Mittelpunktsregel.
- *Runge-Kutta:* $y(t) \approx z + \frac{1}{6}h[f(\hat{t}) + 4f(\hat{t} + \frac{h}{2}) + f(\hat{t} + h)]$
  Dies entspricht der Simpson-Regel.

ii) Für Systeme $(n > 1)$ überträgt man das Vorgehen komponentenweise, z. B. beim Euler-Verfahren:

$$U_{i+1} := \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{i+1} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_i + h \begin{bmatrix} f_1(t_i, U_i) \\ f_2(t_i, U_i) \\ \vdots \\ f_n(t_i, U_i) \end{bmatrix}$$

Beispiel 2.2.3 *Wir wollen uns die drei Verfahren Euler, Heun und modifizierter Euler im Vergleich anschauen. Dazu betrachten wir die Anfangswertaufgabe*

$$y' = -2xy^2, y(0) = 1$$

mit der exakten Lösung $y(x) = 1/(x^2+1)$. *Mit der Methode von Euler erhalten wir die in der Tabelle 1 zusammengestellten Näherungswerte* y_k *für verschiedene Schrittweiten* h *an gleichen Stellen* x_k *sowie die zugehörigen Fehler* $e_k := y(x_k) - y_k$. *Der Fehler nimmt etwa proportional zur Schrittweite* h *ab.*

		$h = 0.1$		$h = 0.01$		$h = 0.001$	
x_k	$y(x_k)$	y_k	e_k	y_k	e_k	y_k	e_k
0	1.00000	1.00000	–	1.00000	–	1.00000	–
0.1	0.99010	1.00000	−0.00990	0.99107	−0.00097	0.99020	−0.00010
0.2	0.96154	0.98000	−0.01846	0.96330	−0.00176	0.96171	−0.00018
0.3	0.91743	0.94158	−0.02415	0.91969	−0.00226	0.91766	−0.00022
0.4	0.86207	0.88839	−0.02632	0.86448	−0.00242	0.86231	−0.00024
0.5	0.80000	0.82525	−0.02525	0.80229	−0.00229	0.80023	−0.00023
0.6	0.73529	0.75715	−0.02185	0.73727	−0.00198	0.73549	−0.00020

Tabelle 1: Verfahren von Euler

Obige Anfangswertaufgabe ist mit dem Verfahren von Heun und dem modifiziertem Euler-Verfahren je mit den Schrittweiten $h = 0.1$ *und* $h = 0.005$ *behandelt worden. Die erhaltenen Näherungswerte* y_k *sind zusammen mit den zugehörigen globalen Fehlern* $g_k = y(x_k) - y_k$ *in der Tabelle 2 zusammengestellt. Die Resultate offenbaren die bedeutend kleineren Diskretisierungsfehler der beiden Methoden und ihre Fehlerordnung 2.*

	Modifiziertes Euler-Verfahren				*Methode von Heun*			
	$h = 0.1$		$h = 0.05$		$h = 0.1$		$h = 0.05$	
x_k	y_k	g_k	y_k	g_k	y_k	g_k	y_k	g_k
0	1.00000	–	1.00000	–	1.00000	–	1.00000	–
0.1	0.99000	0.00010	0.99007	0.00002	0.99000	0.00010	0.99009	0.00001
0.2	0.96118	0.00036	0.96145	0.00009	0.96137	0.00017	0.96152	0.00002
0.3	0.91674	0.00069	0.91727	0.00016	0.91725	0.00019	0.91742	0.00001
0.4	0.86110	0.00096	0.86184	0.00023	0.86195	0.00011	0.86208	−0.00001
0.5	0.79889	0.00111	0.79974	0.00026	0.80003	−0.00003	0.80004	−0.00004
0.6	0.73418	0.00111	0.73503	0.00026	0.73553	−0.00023	0.73538	−0.00009
0.7	0.67014	0.00100	0.67091	0.00023	0.67159	−0.00045	0.67128	−0.00014
0.8	0.60895	0.00080	0.60957	0.00018	0.61040	−0.00064	0.60993	−0.00018
0.9	0.55191	0.00058	0.55236	0.00013	0.55329	−0.00080	0.55270	−0.00021
1.0	0.49964	0.00036	0.49992	0.00008	0.50092	−0.00092	0.50024	−0.00024

Tabelle 2: modifiziertes Euler-Verfahren und Methode von Heun

2.2.4 Konvergenz und Gesamtfehler

Wir brauchen im folgenden mehrfach:

Lemma 2.2.1 *Falls* $\delta_j \geq 0, \eta_j \geq 0, z_j \geq 0$ *für alle* $j \in 0 \ldots, N$ *mit*

$$z_k \leq (1+\delta_{k-1})z_{k-1} + \eta_{k-1} \text{ für alle } k = 1, \ldots, N.$$

Dann gilt

$$z_k \leq (z_0 + \sum_{j=0}^{k-1} \eta_j) e^{\sum_{j=0}^{k-1} \delta_j} \text{ für alle } k = 1, \ldots, N$$

Beweis:

Mit vollständiger Induktion:

I.A: $z_1 \leq (1+\delta_0)z_0 + \eta_0 \leq (z_0 + \eta_0)e^{\delta_0} = (z_0+\eta_0)(1 + \delta_0 + \frac{\delta_0^2}{2} + \frac{\delta_0^3}{3!} + \ldots)$

I.V: Gelte die Behauptung für $k \geq 1$, dann

$$z_{k+1} \leq (1+\delta_k)z_k + \eta_k \leq (1+\delta_k)(z_0 + \sum_{j=0}^{k-1} \eta_j) e^{\sum_{j=0}^{k-1} \delta_j} + \eta_k$$

Es ist aber

$$(z_0 + \sum_{j=0}^{k} \eta_j) e^{\sum_{j=0}^{k} \delta_j} = (z_0 + \sum_{j=0}^{k-1} \eta_j) e^{\sum_{j=0}^{k-1} \delta_j} \underbrace{e^{\delta_k}}_{\geq 1+\delta_k} + \eta_k \underbrace{e^{\sum_{j=0}^{k} \delta_j}}_{\geq 1}$$

also folgt die Behauptung. $\surd$

Der folgende Satz gibt Aufschluß über das Konvergenzverhalten, d.h. das Verhalten des globalen Diskretisierungsfehlers bei kleiner werdender Schrittweite.

Satz 2.2.1 $\phi(t, y, h, f)$ *genüge auf dem Streifen* $[t_0, t_0 + a] \times \mathbb{R}$ *einer Lipschitzbedingung mit Lipschitzkonstante* L *bzgl.* y*. Dann gilt für den globalen Diskretisierungsfehler*

$$|e_N| \leq (|e_0| + (t_N - t_0)\tau_h) e^{L(t_N - t_0)}.$$

Dabei ist $\tau_h = max|\tau_j|, \tau_j$ *lokaler Diskretisierungsfehler bei* t_j*. (Das heißt der globale Diskretisierungsfehler verhält sich wie der lokale, solange* $|e_0|$ *klein genug ist und* $L(t_N - t_0)$ *nicht zu groß.)*

Beweis:

Sei zur Schreiberleichterung $y_j = y(t_j)$ für alle $j \in 0, \ldots, N$.
Für das Verfahren

$$u_{j+1} = u_j + h\phi(t_j, u_j, h, f)$$

erhalten wir als Lösung

$$y_{j+1} = y_j + h[\phi(t_j, y_j, h, f) + \underbrace{\tau(t_j, y_j, h, f)}_{\tau_j}]$$

Also

$$e_{j+1} = e_j + h[\phi(t_j, y_j, h, f) - \phi(t_j, u_j, h, f)] + h\tau_j.$$

Da ϕ einer Lipschitzbedingung genügt, folgt

$$|e_{j+1}| \leq (1 + Lh)|e_j| + h|\tau_j|. \tag{2.2.10}$$

Aus (2.2.10) erhält man mit Lemma 2.2.1

$$|e_N| \leq (|e_0| + h\sum_{j=0}^{N-1} |\tau_j|)e^{NLh} \leq (|e_0| + (t_N - t_0)\tau_h)e^{L(t_N - t_0)},$$

da $Nh = t_N - t_0$. $\checkmark$

Wir nennen ein Einschrittverfahren konvergent, falls

$$\lim_{N\to\infty} |e_N| = 0.$$

Falls also $\tau_h = Ch^P$, so gilt mit $e_0 = 0\ :\ |e_N| \leq (t_N - t_0)Ch^P e^{L(t_N - t_0)}$ und damit $lim_{h\to 0}|e_N| = 0$. Also hat man Konvergenz der Einschrittverfahren der Ordnung P, die in einem Streifen eine Lipschitzbedingung erfüllen.

Die Frage, die nun sich für die Praxis stellt ist: „Wie klein muß h sein, um eine bestimmte Genauigkeit zu erreichen? “. Da N für τ_h und L selbst beim Euler-Verfahren schwer zugänglich sind, erhält man mit diesem Satz hierauf keine Antwort. Schließlich müssen wir noch berücksichtigen, daß wir auf einem Rechner mit Gleitpunktarithmetik rechnen, während das obige Ergebnis ja von exakter Rechnung ausgeht. Der folgende Satz besagt, daß es nicht sinnvoll ist, bei Durchführung des Verfahrens auf einem Rechner, die Schrittweite h über eine gewisse Grenze hinaus zu verkleinern, da sonst die Rundungsfehlereinflüsse das Ergebnis stark verfälschen können.

Satz 2.2.2 *Rundungsfehlereinfluß beim expliziten Einzelschrittverfahren*

Seien u_i die (theoretischen) Verfahrenswerte, die gemäß

$$u_{i+1} = u_i + h\phi(t_i, u_i, h, f), \quad u_0 = y_0$$

gewonnen werden, und seien $\tilde{u}_i$ die vom Rechner hierfür gelieferten Werte, für die gilt

$$\tilde{u}_{i+1} = \tilde{u}_i + h\phi(t_i, \tilde{u}_i, h, f) + \epsilon_{i+1}, \quad \tilde{u}_0 = u_0 + \epsilon_0$$

ϕ genüge einer Lipschitzbedingung mit Konstante L bzgl. u und es sei $|\epsilon_i| \leq \epsilon$ für alle $i \geq 0$.

Dann gilt für $r_i = \tilde{u}_i - u_i$:

$$|r_N| \leq e^{L(t_N - t_0)}(|r_0| + \frac{t_N - t_0}{h}\epsilon)$$

Beweis:

Wir haben

$$r_{i+1} = r_i + h[\phi(t_i, \tilde{u}_i, h, f) - \phi(t_i, u_i, h, f)] + \epsilon_{i+1}$$

also

$$|r_{i+1}| \leq (1 + hL)|r_i| + \epsilon.$$

Nach Lemma 2.2.1 gilt daher

$$|r_N| \leq (|r_0| + N\epsilon)e^{NLh} = (|r_0| + \frac{t_N - t_0}{h}\epsilon)e^{L(t_N - t_0)},$$

da $Nh = t_N - t_0$. $\surd$

Satz 2.2.3 *Gesamtfehler $E_i := \tilde{u}_i - y(t_i)$ beim expliziten Einzelschrittverfahren*

Mit den Bezeichnungen und Voraussetzungen des vorherigen Satzes und mit $|\tau_i| \leq M$ folgt

$$|E_i| = |\tilde{u}_i - y(t_i)| \leq e^{L(t_N - t_0)}(|\epsilon_0| + (t_N - t_0)(M + \frac{\epsilon}{h}))$$

wobei $y(t_i)$ die exakten Werte seien.

Als Konsequenz aus diesem Satz erhalten wir: Falls h groß ist, so ist τ und damit M groß und $\frac{\epsilon}{h}$ groß. Falls h klein ist, so ist M klein, aber $\frac{\epsilon}{h}$ groß. Wir müssen h klein machen, damit τ klein ist, wir dürfen es aber nicht zu klein machen, weil sonst der Anteil der von den Rundungsfehlern herrührt dominiert. Wir brauchen daher Verfahren höherer Ordnung.

2.3 Explizite und implizite Runge-Kutta-Verfahren

Für die Entwicklung von Runge-Kutta-Formeln betrachtet man wie schon vorher häufig die Gleichung:

$$y(t_{j+1}) = y(t_j) + \int_{t_j}^{t_{j+1}} f(s, y(s))ds.$$

Das Einsetzen einer geeigneten Quadraturformel zur Näherung des Integral würde auf die Näherung

$$y(t_{j+1}) = y(t_j) + h_j \sum_{r=1}^{m} \gamma_r f(s_r, y(s_r))$$

führen, wobei die s_l m Stützstellen im Intervall $[t_j, t_{j+1}]$ und die γ_l Gewichte der gewählten Quadratur sind und $h_j = t_{j+1} - t_j$. Das Problem besteht hier noch darin, daß in den Ausdrücken $f(s_r, y(s_r))$ die Größe $y(s_r)$ unbekannt ist. Sie muß daher jeweils durch eine geeignete Approximation ersetzt werden.

Man macht dazu folgenden Ansatz, wobei wir wieder $y_j = y(t_j)$ setzen:

$$f(s_r, y(s_r)) \approx k_r(t_j, y_j)$$

mit

$$\begin{array}{rcl} s_1 = t_j & & s_r = t_j + \alpha_r h_j \quad \alpha_r = \sum_{l=1}^{r-1} \beta_{rl} \\ k_1(t_j, y_j) & = & f(t_j, y_j) \\ k_2(t_j, y_j) & = & f(t_j + \alpha_2 h_j, y_j + h_j \beta_{21} k_1(t_j, y_j)) \\ k_3(t_j, y_j) & = & f(t_j + \alpha_3 h_j, y_j + h_j(\beta_{31} k_1(t_j, y_j) + \beta_{32} k_2(t_j, y_j))) \\ & \vdots & \\ k_m(t_j, y_j) & = & f(t_j + \alpha_m h_j, y_j + h_j(\beta_{m1} k_1(t_j, y_j) + \ldots + \beta_{m,m-1} k_{m-1}(t_j, y_j))) \end{array} \tag{2.3.1}$$

Mit diesem Ansatz führt unsere Idee auf das folgende Verfahren:

$$\begin{array}{rcl} u_{j+1} & = & u_j + h_j(\gamma_1 k_1(t_j, u_j) + \ldots + \gamma_m k_m(t_j, u_j)) \\ u_0 & = & y_0 \end{array}$$

bzw. auf die zugehörige Inkrementfunktion:

$$\phi(t, y, h_j, f) = \gamma_1 k_1(t, y) + \ldots + \gamma_m k_m(t, y) \tag{2.3.2}$$

Ein Verfahren dieser Form heißt <u>explizites m-stufiges Runge-Kutta-Verfahren</u>. Es wird üblicherweise durch eine Tabelle angegeben:

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
\alpha_2 & \beta_{21} & & & & & \\
\vdots & \vdots & \ddots & & & & \\
\alpha_r & \beta_{r1} & \cdots & \beta_{r,r-1} & & & \\
\vdots & \vdots & & & \ddots & & \\
\alpha_m & \beta_{m1} & \cdots & \cdots & \cdots & \beta_{m,m-1} & \\
\hline
 & \gamma_1 & \cdots & \cdots & \cdots & \gamma_{m-1} & \gamma_m
\end{array}
\tag{2.3.3}
$$

Die α_r charakterisieren die Stützstellen der Quadratur, die γ_r sind ihre Gewichte.

Die Parameter γ_j, β_{rj} und α_j müssen gewissen Bedingungen genügen, damit man ein konsistentes, sinnvolles Verfahren erhält. Zunächst möchte man aus Gründen der numerischen Stabilität, daß $\alpha_r, \beta_{rj}, \gamma_j$ alle nichtnegativ sind. Außerdem soll $k_r(t_j, y_j)$ die Größe $f(s_r, y(s_r))$ approximieren. Also

$$
\begin{aligned}
k_r(t_j, y_j) &= f(t_j + \alpha_r h_j, y_j + h_j(\beta_{r1} k_1(t_j, y_j) + \ldots + \beta_{r,r-1} k_{r-1}(t_j, y_j)) \\
&\approx f(t_j + \alpha_r h_j, y(t_j + \alpha_r h_j)) \\
&= f(t_j + \alpha_r h_j, y_j + \alpha_r h_j y'(t_j) + O(h_j^2)).
\end{aligned}
$$

Durch Taylorentwicklung und Vergleich findet man, falls mindestens eine $O(h_j^2)-$Approximation gefordert wird, die Bedingung

$$\alpha_r = \beta_{r1} + \ldots + \beta_{r,r-1} \tag{2.3.4}$$

Schließlich erhält man aus der Konsistenzforderung $\phi(t, y, 0, f) = f(t, y)$ für stetige ϕ aus (2.3.2) die Bedingung

$$\gamma_1 + \gamma_2 + \ldots + \gamma_m = 1. \tag{2.3.5}$$

Denn für $h = 0$ gilt $k_l(t_j, y_j) = k_1(t_j, y_j) = f(t_j, y_j)$ für alle $l = 1, \ldots, m$. Wir erhalten damit sofort:

Lemma 2.3.1 *Ein Runge-Kutta-Verfahren gegeben durch (2.3.3) mit der Bedingung (2.3.5) ist konsistent.*

<u>Beweis:</u>

Es ist für ein $\hat{t} \in [t_0, t_0 + a]$, $z = y(\hat{t})$

$$
\begin{aligned}
|\tau(\hat{t}, z, h, f)| &= |\triangle(\hat{t}, z, h, f) - \phi(\hat{t}, z, h, f)| \\
&= \left|\frac{y(\hat{t} + h) - y(\hat{t})}{h} - \sum_{r=1}^{m} \gamma_r k_r(\hat{t}, z)\right| \\
&\leq \left|\frac{y(\hat{t} + h) - y(\hat{t})}{h} - f(\hat{t}, z)\right| + \left|f(\hat{t}, z) - \sum_{r=1}^{m} \gamma_r k_r(\hat{t}, z)\right|
\end{aligned}
$$

Der erste Summand konvergiert offensichtlich gegen 0 für h gegen 0. Der zweite Summand läßt sich nach oben abschätzen durch

$$\sum_{r=1}^{m} \gamma_r |(f(\hat{t}, z) - k_r(\hat{t}, z))|,$$

da $\sum_{r=1}^{m} \gamma_r = 1$. Aber

$$f(\hat{t}, z) - k_r(\hat{t}, z) = f(\hat{t}, z) - f(\hat{t} + \alpha_r h, z + h(\ldots))$$

und dies strebt für stetige f offenbar ebenfalls gegen 0 für h gegen 0. $\surd$

Wir haben die folgenden Spezialfälle:

i) Euler-Verfahren
$m = 1 \quad \gamma_1 = 1$
$u_{j+1} = u_j + h_j f(t_j, u_j)$
Ordnung 1, Quadratur basiert auf Rechteckregel

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

ii) modifiziertes Euler-Verfahren
$m = 2 \quad \gamma_1 = 0 \quad \gamma_2 = 1 \quad \beta_{21} = \alpha_2 = \frac{1}{2}$
$u_{j+1} = u_j + h_j f(t_j + \frac{1}{2} h_j, u_j + \frac{1}{2} h_j f(t_j, u_j))$
Ordnung 2, Quadratur basiert auf Mittelpunktsregel.

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

iii) Verfahren von Heun
$m = 2 \quad \gamma_1 = \gamma_2 = \frac{1}{2} \quad \beta_{21} = \alpha_2 = 1$
$u_{j+1} = u_j + \frac{h_j}{2}(f(t_j, u_j) + f(t_j + h_j, u_j + h_j f(t_j, u_j))$
Ordnung 2, Quadratur basiert auf Trapezregel

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

iv) Klassisches Runge-Kutta-Verfahren
$m = 4$
Ordnung 4, Quadratur basiert auf Simpsonregel

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Die Konsistenzordnung der Runge-Kutta-Verfahren ist i.a. schwierig zu bestimmen. Taylorentwicklung mit Koeffizientenvergleich führt auf nichtlineare Bedingungsgleichungen für die Parameter der Tabelle, deren Zahl mit wachsender Ordnung schnell wächst, d.h. die Anzahl der Funktionsauswertungen wächst auch sehr schnell mit wachsendem p (mehr Parameter nötig!)

Ordnung p	1	2	3	4	5	6	7	8
Zahl der Bed.	1	2	4	8	17	37	85	200

Die erreichbare Konsistenzordnung p einer expliziten Runge-Kutta-Formel ist

m	1	2	3	4	5	6	7	8	9	$m \geq 9$
$p(m)$	1	2	3	4	4	5	6	6	7	$\leq m - 2$

Für Diagramme und Bestimmung der Parameter siehe z.B.Hairer/Nørsett/ Wanner oder direkt bei Butcher, J.C.: On Runge-Kutta Processes of Higher Order, J. Austr. Math. Soc. 4 (1964) 179 - 194 und On The Admissable Order of Runge-Kutta Methods, Math. Comp. 19 (1965) 408 - 417.

Beispiel 2.3.1 *Für $m = 4$, $p = 4$ gibt es mehrere Lösungen*

klassisches Runge-Kutta-Verfahren
hat die wenigsten Parameter, da viele Nullen
Quadratur mittels Simpson-Regel

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & \\
1 & 0 & 0 & 1 & \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}$$

3/8-Regel
Quadratur basiert auf 3/8-Regel

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{3} & \frac{1}{3} & & & \\
\frac{2}{3} & -\frac{1}{3} & 1 & & \\
1 & 1 & -1 & 1 & \\
\hline
 & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8}
\end{array}$$

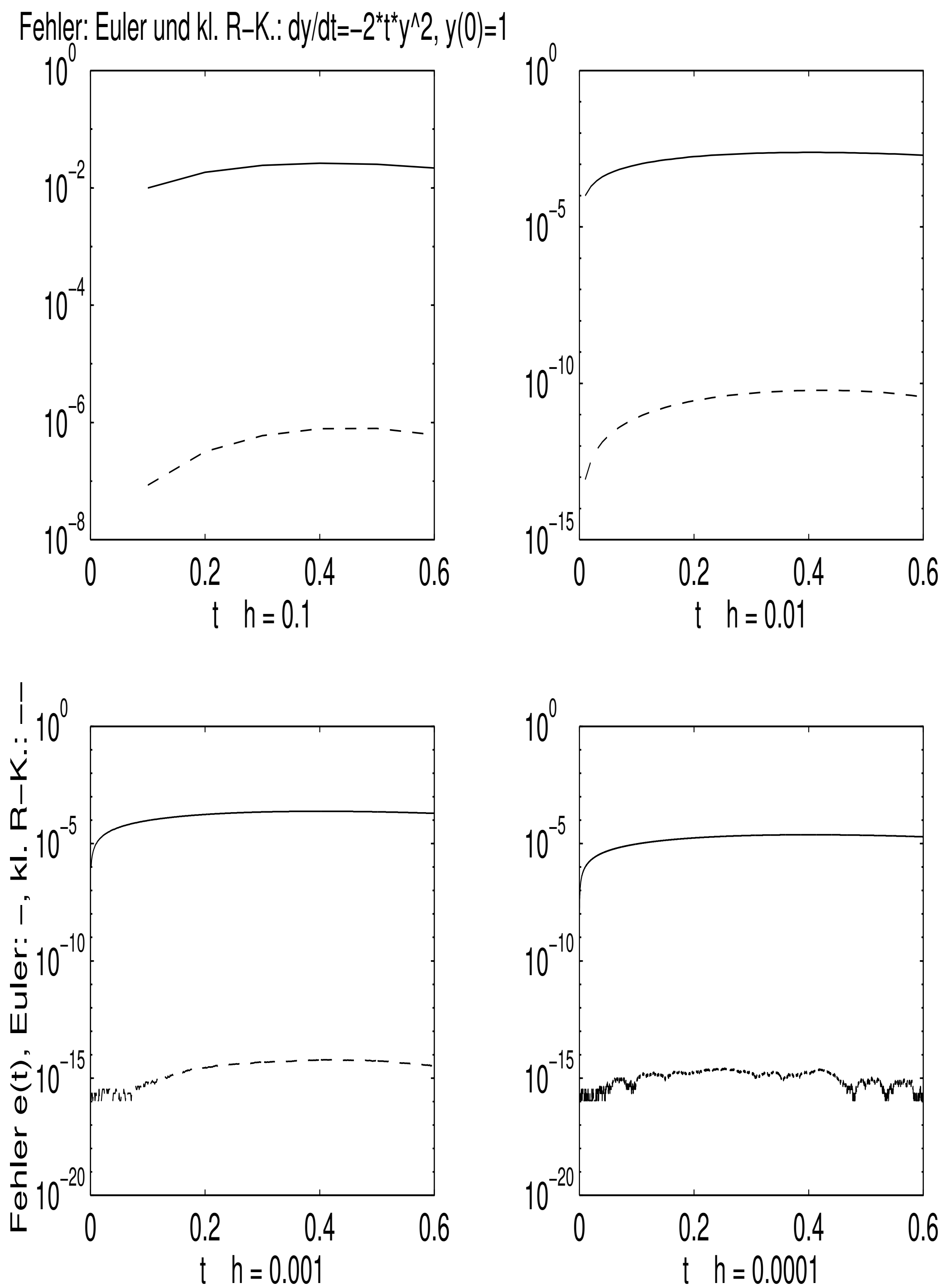
Optimale Formel von Kuntzmann

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{2}{5} & \frac{2}{5} & & & \\
\frac{3}{5} & -\frac{3}{20} & \frac{3}{4} & & \\
1 & \frac{19}{44} & -\frac{15}{44} & \frac{40}{44} & \\
\hline
 & \frac{55}{360} & \frac{125}{360} & \frac{125}{360} & \frac{55}{360}
\end{array}$$

Weitere Lösung: in den Übungen

Beispiel 2.3.2 *Vergleich zwischen dem Euler-Verfahren und dem klassischen Runge-Kutta-Verfahren an Hand der Anfangswertaufgabe*

$$y' = -2ty^2 \quad y(0) = 1$$



2.3.1 Implizite Runge-Kutta Formeln

Für eine Reihe von Problemklassen benötigt man Verfahren, die zwar viel höheren Rechenaufwand erfordern aber dafür gewisse Stabilitätseigenschaften (später) haben, dazu gehören die impliziten Runge-Kutta-Verfahren. Sie lassen sich als Erweiterung der expliziten Runge-Kutta-Verfahren ganz einfach durch die folgende Verallgemeinerung der Definition (2.3.1) entwickeln:

$$k_r(t_j, y_j) = f(t_j + \alpha_r h_j, y_j + h_j(\beta_{r1} k_1 + \ldots + \beta_{rm} k_m)) \quad \text{für} \quad r = 1, \ldots, m \tag{2.3.6}$$

und damit

$$u_{j+1} = u_j + h_j(\gamma_1 k_1(t_j, u_j) + \ldots + \gamma_m k_m(t_j, u_j))$$

Mit $\beta_{rl} = 0$ für $l \geq r$ ist k_r in (2.3.6) explizit aus $k_1, \ldots, k_{r-1}$ berechenbar, sonst ist (2.3.6) implizit in den Unbekannten $k_1, \ldots, k_m$. Man beachte: Nicht die obige Verfahrensgleichung ist eine implizite Gleichung für u_{j+1}, sondern die $k_1, \ldots, k_m$ müssen als Lösung des Gleichungssystems (2.3.6) z.B. mit dem Newton-Verfahren berechnet werden.

Implizite Runge-Kutta-Verfahren werden üblicherweise durch eine Tabelle angegeben:

$$\begin{array}{c|ccc}
\alpha_1 & \beta_{11} & \ldots & \beta_{1m} \\
\alpha_2 & \beta_{21} & \ldots & \beta_{2m} \\
\vdots & \vdots & & \vdots \\
\alpha_m & \beta_{m1} & \ldots & \beta_{mm} \\
\hline
 & \gamma_1 & \ldots & \gamma_m
\end{array}$$

Bei den impliziten Verfahren sind drei verschiedene Formeltypen von besonderem Interesse.

1. Gauß-Form : Alle Parameter $\alpha_j, \beta_{jl}, \gamma_j$ beliebig wählbar
2. Radau-Form : Entweder $\alpha_1 = \beta_{11} = \beta_{12} = \ldots = \beta_{1m} = 0$ oder $\alpha_m = 1$ und $\beta_{1m} = \beta_{2m} = \ldots = \beta_{mm} = 0$.
3. Lobatto-Form : $\alpha_1 = \beta_{11} = \beta_{12} = \ldots = \beta_{1m} = \beta_{2m} = \ldots = \beta_{mm} = 0, \alpha_m = 1$.

Die Namen der drei Typen von Verfahren leiten sich aus dem Umstand her, daß sie im Spezialfall einer von y unabhängigen Funktion f in die gleichnamigen Quadraturformeln übergehen. Die Radau-Formeln haben den Vorteil, daß entweder k_1 oder k_m explizit berechnet werden kann, bei den Lobatto-Formeln können sogar k_1 und k_m explizit berechnet werden, wodurch die Zahl der in jedem Schritt zu lösenden impliziten Gleichungen verringert wird. Dafür muß man eine geringere Konsistenzordnung bei gleicher Stufenzahl in Kauf nehmen. Sie beträgt in der angegebenen Reihenfolge 1. - 3. der Verfahren $p = 2m, p = 2m-1, p = 2m-2$.

Wir geben nun einige spezielle implizite Formeln der genannten Typen an.

Gauß-Form : $m = 1, p = 2$
$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Mit diesen Koeffizienten erhält man das folgende Verfahren :

$$\begin{aligned} k_1 &= f(t_j + \frac{h_j}{2}, u_j + \frac{h_j}{2} k_1) \\ u_{j+1} &= u_j + h_j k_1 \end{aligned}$$

Gauß-Form : $m = 2, p = 4$
$$\begin{array}{c|cc} \frac{(3-\sqrt{3})}{6} & \frac{1}{4} & \frac{(3-2\sqrt{3})}{12} \\ \frac{(3+\sqrt{3})}{6} & \frac{(3+2\sqrt{3})}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Radau-Form : $m = 1, p = 1$
$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

Bei dieser Formel handelt es sich also gerade um das Euler-Verfahren.

Radau-Form : $m = 2, p = 3$
$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{1}{3} & 0 \\ 1 & 1 & 0 \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Lobatto-Form : $m = 2, p = 2$
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Setze man diese Werte ein, so erhält man die (explizite) Formel

$$u_{j+1} = u_j + \frac{h_j}{2}(f(t_j, u_j) + f(t_j + h_j, u_j + f(t_j, u_j))).$$

Lobatto-Form : $m = 3, p = 4$
$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 1 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Zur Berechnung der k_r muß ein nichtlineares Gleichungssystem gelöst werden (Newton-Verfahren!). Dazu ist natürlich jeweils zunächst zu klären, ob das System überhaupt eindeutig lösbar ist.

Satz 2.3.1 *f genüge für alle $(t, y_1), (t, y_2) \in I \times \mathbb{R}^n, I = [t_0, t_0 + a]$*

$$\|f(t, y_1) - f(t, y_2)\| \leq L\|y_1 - y_2\|.$$

Dann existiert für alle Schrittweiten h mit

$$q = L \cdot h \cdot max_{j=1,\ldots,m}(\sum_{k=1}^{m} |\beta_{jk}|) < 1$$

und alle $(t, y) \in I \times \mathbb{R}^n$ eine eindeutig bestimmte Lösung $k_1(t, y), \ldots, k_m(t, y)$ von (2.3.6).

Beweis: mittels Banachschen Fixpunktsatz.

Bemerkung 2.3.1 *i) Wesentlicher Vorteil der impliziten gegenüber den expliziten Verfahren ist der größere Stabilitätsbereich (kommt später).*

ii) Die Lipschitz-Bedingung im letzten Satz läßt sich auf eine lokale Lipschitz-Bedingung abschwächen, die nur in einem Schlauch um die Lösung gilt.

iii) Um $q < 1$ zu garantieren, muß h genügend klein sein. Das führt oft dazu, daß der erhöhte Rechenaufwand die hohe Konsistenzordnung wieder aufhebt.

Da der Aufwand, Verfahren höherer Ordnung zu erstellen, recht hoch sein kann, wird man sich überlegen müssen, warum man überhaupt an Verfahren höherer Ordnung interessiert ist. Das Ziel bei der Entwicklung numerischer Verfahren zur Lösung von Differentialgleichung ist es, bei kurzer Rechenzeit eine hohe Genauigkeit zu erreichen. Wir hatten bereits im Satz 2.2.3 gesehen, daß der Rundungsfehlereinfluß der Verkleinerung der Schrittweite h Grenzen setzt. Eine Erhöhung der Genauigkeit kann man dann nur noch durch eine Erhöhung der Konsistenzordnung erreichen.

Darüberhinaus kann man sich überlegen, daß zum Erreichen einer bestimmten Genauigkeit der Rechenaufwand bei Verfahren höherer Ordnung oft geringer ist als bei Verfahren niedriger Ordnung. Dazu betrachten wir einmal grob die Gesamtrechenzeit Z als Funktion der Konsistenzordnung p des Verfahrens. Nach unseren bisherigen Überlegungen können wir annehmen, daß der Fehler im Endpunkt $t_0 + a$ sich als $\epsilon = Kh^p$ darstellt, wobei K eine Konstante ist und h die Schrittweite. Wir nehmen an, daß für einen Schritt der Rechenaufwand pZ_0 beträgt. Dies ist nicht unrealistisch, denn der Hauptaufwand besteht i.a. in den Funktionsauswertungen, von denen man bei einer Konsistenzordnung p häufig p pro Schritt hat. Gelegentlich aber muß man auch $p^j Z_0$ ansetzen. Bei vorgebener Genauigkeit ϵ erhält man aus $\epsilon = Kh^p$ als benötigte Schrittweite $h = \left(\frac{\epsilon}{K}\right)^{\frac{1}{p}}$. Dann ist aber

$$Z = \frac{a}{h} pZ_0 = apZ_0 \left(\frac{K}{\epsilon}\right)^{\frac{1}{p}}$$

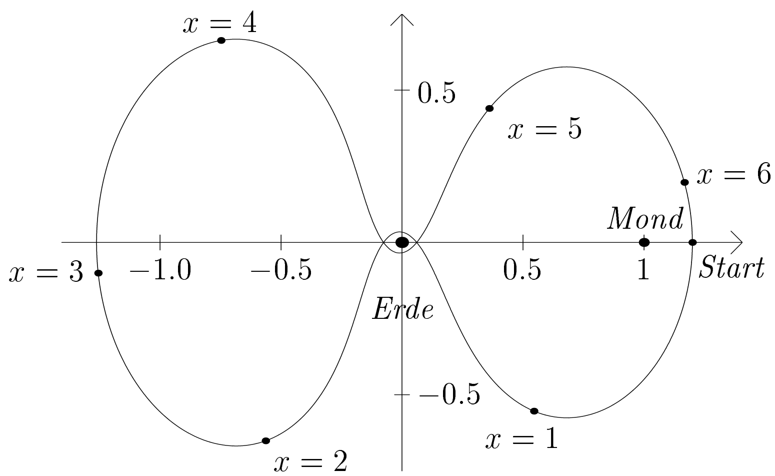
Durch Differenzieren dieser Funktion nach p sieht man, daß die Gesamtrechenzeit minimal wird für $p = [ln\left(\frac{K}{\varepsilon}\right)]$ oder $p = [ln\left(\frac{K}{\varepsilon}\right)] + 1$ ($[x]$ = ganzzahliger Anteil von x). Aus diesem Ergebnis ergibt sich, daß man bei größerere verlangter Genauigkeit, d.h. bei kleinem ϵ, mit entsprechend höherer Konsistenzordnung rechnen sollte, um minimale Gesamtrechenzeit zu erreichen. Eine zu groß gewählte Konsistenzordnung führt allerdings wieder zu einem Anwachsen von Z. Wir betrachten dazu das folgende Beispiel, das eine periodische Satellitenbahn im Erd-Mond-System beschreibt.

Beispiel 2.3.3 *(R.D. Grigorieff)*

Die Differentialgleichungen

$$
\begin{aligned}
u'' &= u + 2v' - (1-\mu)\frac{u+\mu}{[(u+\mu)^2+v^2]^{3/2}} - \mu\frac{u-(1-\mu)}{[(u-1+\mu)^2+v^2]^{3/2}} \\
v'' &= v - 2u' - (1-\mu)\frac{v}{[(u+\mu)^2+v^2]^{3/2}} - \mu\frac{v}{[(u-1+\mu)^2+v^2]^{3/2}} \\
u(0) &= 1.2, \\
v(0) &= u'(0) = 0, \\
v'(0) &= -1.049357509830350
\end{aligned}
$$

beschreiben eine periodische Satellitenbahn im Erde-Mond-System, wobei $\mu = \dfrac{1}{82.45}$ *die relative Mondmasse ist. In der Bewegungsebene hat die Bahnkurve das folgende Aussehen.*



Die Zeit für einen vollen Umlauf beträgt ungefähr $T = 6.19$. Das System ist mit dem Runge-Kutta-Fehlberg-Verfahren mit Konsistenzordnungen $p = 4, \ldots, 20$ integriert worden.

Runge-Kutta-Fehlberg-Verfahren bestehen aus einer Kombination der Methode der Taylorentwicklung (hier nicht näher behandelt) mit anschließenden Verfahren von Runge-Kutta. Besonders wichtig ist außerdem die Möglichkeit einer zeitsparenden Schrittweitensteuerung (siehe nächsten Abschnitt).

Mit r sei die Anzahl der Dezimalen, die in den mit Ordnung p bzw. $p+1$ berechneten Näherungen übereinstimmen sollen. Die folgende Tabelle gibt die Rechenzeiten in Abhängigkeit von p und r. Für jedes r sind die minimalen Rechenzeiten unterstrichen worden.

$p\backslash r$	4	5	6	7	8	9	10	11	12	13
4	72	98	131	184	287	414	639	1176	2905	3068
5	<u>71</u>	84	106	133	184	236	358	520	1247	1308
6	72	<u>76</u>	97	116	150	187	252	362	729	768
7	77	83	<u>96</u>	118	133	169	200	296	518	546
8	84	92	96	112	<u>127</u>	150	182	264	406	428
9	88	95	105	<u>111</u>	127	<u>147</u>	174	241	356	369
10	89	93	106	113	135	156	176	225	316	332
11	97	106	113	124	136	153	<u>171</u>	217	296	307
12	113	123	126	135	141	154	181	<u>212</u>	277	292
13	114	123	128	135	150	153	180	214	279	286
14	116	128	146	140	154	159	182	225	280	283
15	139	147	154	162	174	171	186	233	273	287
16	130	155	149	165	173	179	195	224	<u>268</u>	<u>281</u>
17	150	160	164	172	183	190	200	228	281	289
18	160	162	174	188	194	197	206	243	287	293
19	169	174	188	199	203	216	217	235	281	296
20	174	174	194	208	222	222	295	259	284	309

2.3.2 Schrittweitensteuerung bei Runge-Kutta-Verfahren

Wir haben gesehen, daß wir h nicht zu klein machen können (wegen der Rundungsfehler), aber auch die Ordnung p nicht zu groß, weil dann der Rechenaufwand stark ansteigt. In vielen Problemen ist außerdem ein kleines h oder großes p gar nicht notwendig, weil die Lösung schön glatt verläuft. Es wäre daher ideal, die Schrittweite an das Lösungsverhalten anzupassen. Dies sollte so geschehen, daß stets eine möglichst große Schrittweite gewählt wird, aber der Fehler garantiert unterhalb einer gewissen Schranke liegt. Da wir den globalen Fehler nicht kennen, verwenden wir den lokalen Fehler. Die Grundidee ist nun, den lokalen Fehler zu schätzen, falls diese Schätzung oberhalb einer Toleranzgrenze liegt, so wird die Schrittweite verkleinert, falls die Schätzung einen sehr kleinen Fehler ergibt, wird sie vergrößert. Aber wie erhalten wir eine Schätzung für den lokalen Fehler? Bei den Runge-Kutta-Verfahren ist das sehr einfach, wir verwenden ein Verfahren der Ordnung p zur Approximation der Lösung und eins der Ordnung $p+1$, welches durch möglichst wenig zusätzlichen Rechenaufwand berechnet werden kann, zur Schätzung des Fehlers.

Das beschriebene Vorgehen läßt sich wie folgt motivieren: Seien $\phi_p(t,z,h,f)$ und $\phi_{p+1}(t,z,h,f)$ die Inkrementfunktionen der Verfahren der Ordnung p bzw. $p+1$. Dann gilt:

$$\begin{aligned} \tau_p(t,z,h,f) &= \triangle(t,z,h,f) - \phi_p(t,z,h,f) \\ \tau_{p+1}(t,z,h,f) &= \triangle(t,z,h,f) - \phi_{p+1}(t,z,h,f) \end{aligned}$$

Also folgt

$$\tau_p(t,z,h,f) = \phi_{p+1}(t,z,h,f) - \phi_p(t,z,h,f) + \tau_{p+1}(t,z,h,f).$$

Man verwendet daher

$$\tau_p(t,z,h,f) \cong \phi_{p+1}(t,z,h,f) - \phi_p(t,z,h,f)$$

als Schätzung des Fehlers. Das heißt im Grunde, man tut so als, liefere das Verfahren der Ordnung $p+1$ gegenüber dem Verfahren der Ordnung p die exakten Werte.

Runge-Kutta-Verfahren, bei denen für die gleichen α_r, β_{rj} mit verschiedenen γ_r zwei unterschiedliche Ordnungen entstehen, nennt man <u>eingebettete Runge-Kutta-Verfahren</u>. Es gibt viele Beispiele, aber nicht alle sind brauchbar.

Nun geben wir zwei Beispiele von eingebetteten Runge-Kutta-Verfahren an. Das eingebettete Runge-Kutta-Fehlberg-4(5)-Verfahren hat die Form

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

Dieses Verfahren braucht keine zusätzlichen Funktionsauswertungen. Eine andere exzellente Methode ist das Verfahren von Dormand-Prince-5(4) (DOPRI5) mit den folgenden Koeffizienten

$$\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
y_1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
\hat{y}_1 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}$$

Für Details zu Schrittweitensteuerung bei Runge-Kutta-Verfahren und FORTRAN-Programme siehe E.Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I, Springer 1993.

Die allgemeine Vorgehensweise läßt sich nun wie folgt beschreiben:

Schrittweitensteuerung

Gegeben t_0, u_0, a
Anfangsschrittweite $h < a$
Toleranz τol
Intervall $[t_0, t_0 + a]$
2 Inkrementfunktionen ϕ_p, ϕ_{p+1}, zu Verfahren der Ordnung $p, p+1$

$i = 0$
$t = t_0 - h$
WHILE $t + h < t_0 + a$
    $t_i = t + h$
    berechne $\phi_1 = \phi_p(t_i, u_i, h, f)$
    berechne $\phi_2 = \phi_{p+1}(t_i, u_i, h, f)$
    berechne $\tau = |\phi_2 - \phi_1|$
    IF $\tau < \tau ol$
    THEN IF $\tau > \tau ol/100$
        THEN $u_{i+1} = u_i + h\phi_1$
            $t = t_i$
            $i = i + 1$
        ELSE $h = 2h$
        END IF
    ELSE $h = h/2$
    END IF
END WHILE
$h = t_0 + a - t$
$u_{i+1} = u_i + h\phi_p(t_i, u_i, h, f)$

Eine weitere klassische Methode zur Schrittweitensteuerung bietet die Richardson-Extrapolation. Hier ist die Idee einfach neben dem Verfahren mit Schrittweite h auch noch mit Schrittweite $h/2$ zu rechnen und dieses als Schätzwert zu verwenden. Auf diese Methode kommen wir im nächsten Kapitel zurück.

Beispiel 2.3.4 *Schrittweitensteuerung beim Verfahren von Heun:*

Das Verfahren von Heun ($p = 2$) *lautet:*

$$u_{j+1} = u_j + \frac{1}{2} h_j \{k_1 + k_2\},$$

$$\begin{aligned} k_1 &= f(x_j, u_j) \\ k_2 &= f(x_j + h_j, u_j + h_j k_1). \end{aligned}$$

Für die Schätzung brauchen wir ein Verfahren der Ordnung 3, daß k_1, k_2 *verwendet. Eine mögliche Lösung lautet*

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

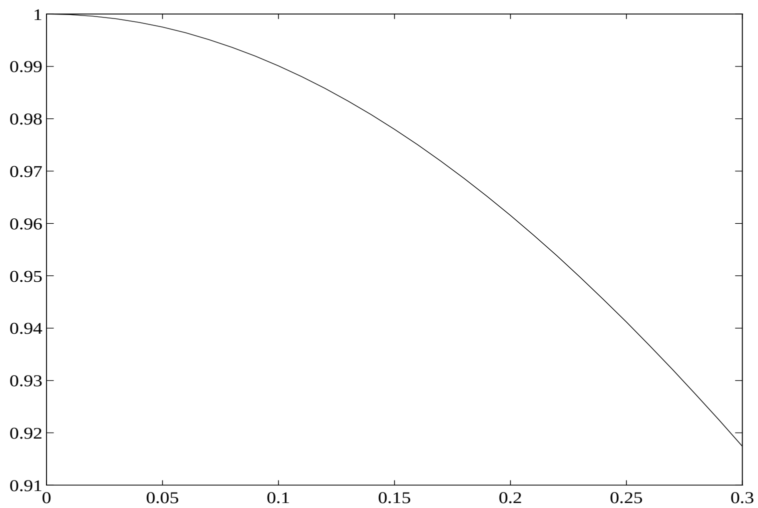
Also

$$\begin{aligned} \hat{u}_{j+1} &= \hat{u}_j + \frac{1}{6} h_j \{k_1 + k_2 + 4k_3\} \\ k_1 &= f(x_j, \hat{u}_j) \\ k_2 &= f(x_j + h_j, \hat{u}_j + h_j k_1) \\ k_3 &= f(x_j + \frac{1}{2} h_j, \hat{u}_j + \frac{1}{4} h_j (k_1 + k_2)) \end{aligned}$$

Diese beiden Verfahren wollen wir nun zur Lösung von

$$y' = -2ty^2, y(0) = 1$$

auf $[0, 0.3]$ *verwenden. Die explizite Lösung lautet* $y(t) = \frac{1}{x^2+1}$.



Der lokale Diskretisationsfehler soll in jedem Schritt betragsmäßig den Wert $\epsilon = 10^{-6}$ *nicht überschreiten. Fällt die Schätzung des lokalen Diskretisationsfehlers* $d_{k+1}^{(H)} = y(x_{k+1}) - y(x_k) -$

$h\phi(x_k, y(x_k), h, f) \approx h/3\{-k_1 - k_2 + 2k_3\}$ *betragsmäßig zu groß aus, soll die Schrittweite halbiert werden. Sinkt sie unter den Wert* 10^{-7}*, soll die Schrittweite verdoppelt werden. In Tabelle 3 ist der Rechengang zusammengestellt. Gestartet wird mit der Versuchsschrittweite* $h = 0.05$*, die im ersten Integrationsschritt sofort angepaßt werden muß. Mit der Schrittweite* $h = 0.025$ *sind sechs Schritte ausführbar. Da auf Grund der Werte von* k_1, k_2 *und* k_3 *an der Stelle* $x_k = 0.125$ *der Schätzwert für* $\mathrm{d}_{k+1}^{(H)}$ *betragsmäßig kleiner als* 10^{-7} *ausfällt, wird die Schrittweite versuchsweise verdoppelt. Doch wird der nächste Schätzwert zu groß, so daß die Schrittweite wieder auf den alten Wert verkleinert werden muß. Das scheinbare Versagen der Strategie hängt mit dem Vorzeichenwechsel des lokalen Diskretisationsfehlers zusammen. Nach diesem Schritt ist eine Integration mit* $h = 0.05$ *erfolgreich, und eine nochmalige Vergrößerung der Schrittweite wäre auf Grund des Schätzwertes* $\mathrm{d}_{k+1}^{(H)}$ *an sich möglich, doch erfolgt dies nur nach zwei aufeinanderfolgenden erfolgreichen Integrationsschritten. Bei* $x_k = 0.300$ *ist die Schrittweite zu halbieren. Der globale Fehler* $g_k = |u_k - y(x_k)|$ *wurde mit der exakten Lösung berechnet.*

x_k	u_k	$10^8 g_k$	h	k_1	k_2	k_3	$10^8 d_{k+1}^{(H)}$
0	1.00000	0	0.050	0	−0.1000	−0.0498	416(!)
			0.025	0	−0.0500	−0.0249	26
0.025	0.99937	39	0.025	−0.0499	−0.0996	−0.0747	26
0.050	0.99750	77	0.025	−0.0995	−0.1485	−0.1239	24
0.075	0.99440	114	0.025	−0.1483	−0.1962	−0.1722	21
0.100	0.99009	147	0.025	−0.1960	−0.2426	−0.2193	16
0.125	0.98461	173	0.025	−0.2423	−0.2872	−0.2648	9
0.150	0.97799	192	0.050	−0.2869	−0.3714	−0.3291	130(!)
			0.025	−0.2869	−0.3298	−0.3084	1
0.175	0.97028	199	0.050	−0.3295	−0.4093	−0.3694	9
0.225	0.95181	372	0.050	−0.4076	−0.4771	−0.4425	−301(!)
			0.025	−0.4076	−0.4433	−0.4255	−48
0.250	0.94117	327	0.025	−0.4429	−0.4757	−0.4593	−68
0.275	0.92968	260	0.025	−0.4753	−0.5054	−0.4904	−91
0.300	0.91742	172	0.025	−0.5050	−0.5321	−0.5186	−116(!)

Tabelle 3 : Methode von Heun mit Schrittweitensteuerung

2.4 Extrapolationsverfahren

Eine weitere Möglichkeit, Verfahren höherer Ordnung zu erhalten (sowie auch Methoden zur Schrittweitensteuerung) ist die Extrapolation. Ähnlich wie bei der Extrapolationsquadratur (siehe Mathematisches Praktikum, dort wurde das Prinzip erläutert und für Trapezregel benutzt) wird versucht, führende Terme bei der asymptotischen Entwicklung des Fehlers durch Kombination zu eliminieren. Besonders gut geht das bei Verfahren, die eine h^2-Entwicklung besitzen.

Wir betrachten wieder ein Einschrittverfahren in der Form

$$u_{n+1} = u_n + h\phi(t_n, u_n, h, f) \tag{2.4.1}$$

und schauen uns die asymptotische Entwicklung des globalen Fehlers an. Hat das Verfahren die Ordnung p, so können wir den lokalen Fehler für genügend glattes ϕ entwickeln

$$y(t+h) - y(t) - h\phi(t, y(t), h, f) = d_{p+1}(t)h^{p+1} + \ldots + d_{N+1}(t)h^{N+1} + O(h^{N+2}) \tag{2.4.2}$$

mit glatten Funktionen $d_{p+1}(t), \ldots, d_{N+1}(t)$. Nun setzen wir $u_n = u_h(t)$ für die numerische Lösung bei $t = t_0 + nh$. Wir wissen schon, daß sich der globale Fehler wie h^p verhält. Nehmen wir an, es gäbe eine Funktion $e_p(t)$ mit

$$y(t) - u_h(t) = e_p(t)h^p + O(h^{p+1}).$$

Wir hätten dann als Näherung an $y(t)$ lieber den Wert $\overline{u}_h(t) := u_h(t) + e_p(t)$.

Dazu stellt man sich einfach $\overline{u}_h(t)$ einmal als Lösung eines Einschrittverfahrens

$$\overline{u}_{k+1} = \overline{u}_k + h\overline{\phi}(t_k, \overline{u}_k, h, f),$$

vor, d.h. als

$$u_{k+1} + e_p(t_{k+1})h^p = u_k + e_p(t_k)h^p + h\overline{\phi}(t_k, \overline{u}_k, h, f).$$

Vergleich mit (2.4.1) liefert

$$\overline{\phi}(t_k, \overline{u}_k, h, f) = \phi(t_k, \overline{u}_k - e_p(t_k)h^p, h, f) + (e_p(t_{k+1}) - e_p(t_k))h^{p-1}$$

da $u_k = \overline{u}_k - e_p(t_k)h^p$, d.h.

$$\overline{\phi}(t, \overline{y}, h, f) = \phi(t, \overline{y} - e_p(t)h^p, h, f) + (e_p(t+h) - e_p(t))h^{p-1}.$$

Hiermit kann man nun e_p so bestimmen, daß $e_p(t_0) = 0$ und daß $\overline{\phi}$ ein Verfahren der Ordnung $p+1$ darstellt. Wir betrachten dazu

$$\begin{aligned}
y(t+h) - y(t) - h\overline{\phi}(t, y(t), h, f) &= y(t+h) - y(t) - h\phi(t, y(t), h, f) + h[\phi(t, y(t), h, f) - \overline{\phi}(t, y(t), h, f)] \\
&= d_{p+1}(t)h^{p+1} + h[\phi(t, y(t), h, f) - \phi(t, y(t) - e_p(t)h^p, h, f) \\
&\qquad\qquad -(e_p(t+h) - e_p(t))h^{p-1}] + O(h^{p+2}) \\
&= [d_{p+1}(t) + \frac{\partial\phi}{\partial y}(t, y(t), h, f)e_p(t) - e_p'(t)]h^{p+1} + O(h^{p+2})
\end{aligned}$$

Im letzten Schritt wurde eine Taylorentwicklung angewendet. Damit man ein Verfahren erhält, daß konsistent von der Ordnung $p+1$ ist, muss der erste Faktor des führenden Fehlerterms ein $O(h)$ sein. Dies errreicht man mit der Funktion e_p , für die gilt:

$$e_p'(t) = \frac{\partial f}{\partial y}(t, y(t))e_p(t) + d_{p+1}(t), \quad e_p(t_0) = 0 \tag{2.4.3}$$

denn $\lim_{h\to 0} \phi(t, y(t), h, f) = f(t, y(t))$.

Man kann nun diesen Prozess wiederholen, um die Ordnung nach oben zu treiben, sukzessive Anwendung ergibt:

Satz 2.4.1 *(Gragg 1964)*

Sei ein konsistentes Einzelschrittverfahren der Form (2.4.1) gegeben, für das eine Entwicklung der Form (2.4.2) existiert. Dann besitzt der globale Fehler eine asymptotische Entwicklung der Form

$$y(t) - u_h(t) = e_p(t)h^p + \ldots + e_N(t)h^N + E(t, h)h^{N+1} \tag{2.4.4}$$

wobei die Funktionen $e_j(t)$ Lösung inhomogener linearer Anfangswertaufgaben der Form (2.4.3) sind und $E(t, h)$ beschränkt auf $[t_0, t_0 + a] \times [0, H]$ mit $H > 0$.

Wir haben also wie bei der Quadratur eine asymptotische Entwicklung des Fehlers und können auf dieser Basis Extrapolationsmethoden entwickeln. Zuerst werden wir uns aber noch die Entwicklung des Fehlers für das zugehörige adjungierte Verfahren anschauen. Das zu einer Inkrementfunktion $\phi(t, y, h, f)$ eines Einschrittverfahrens gehörende adjungierte Verfahren ist definiert durch die Inkrementfunktion $\phi^\star(t, y, h, f)$ mittels Auflösung über die Formel

$$\begin{aligned} B &= A - h\phi(t + h, A, -h, f) \\ A &= B + h\phi^\star(t, B, h, f). \end{aligned}$$

Wir erhalten also das adjungierte Verfahren, indem wir in (2.4.1)

$$u_h(t + h) = u_h(t) + h\phi(t, u_h(t), h, f)$$

h durch $-h$ ersetzen,

$$u_{-h}(t - h) = u_{-h}(t) - h\phi(t, u_{-h}(t), -h, f)$$

t durch $t + h$ ersetzen

$$u_{-h}(t) = u_{-h}(t + h) - h\phi(t + h, u_{-h}(t + h), -h, f)$$

und dann mittels des Satzes über implizite Funktionen für kleine h nach $u_{-h}(t + h)$ auflösen. Man schreibt die Lösung in der Form

$$u_{-h}(t + h) = u_{-h}(t) + h\phi^\star(t, u_{-h}(t), h, f)$$

Beispiel 2.4.1 *Das zum expliziten Eulerverfahren adjungierte Verfahren ist das implizite Eulerverfahren* $u_{i+1} = u_i + hf(t_{i+1}, u_{i+1})$ *und umgekehrt.*

Satz 2.4.2 *Asymptotische Entwicklung des globalen Fehlers des adjungierten Verfahrens.*
Es gelten die Bezeichnungen und Voraussetzungen des letzten Satzes. Dann besitzt der globale Fehler des adjungierten Verfahrens die asymptotische Entwicklung wie das zugehörige Verfahren, aber mit h *durch* $-h$ *ersetzt.*

<u>Beweis:</u> analog zu Satz 2.4.1

Insbesondere erhalten wir, daß das adjungierte Verfahren dieselbe Ordnung wie das Verfahren hat.

Ein Einschrittverfahren mit Inkrementfunktion $\phi(t, u_h, h, f)$ heißt <u>symmetrisch</u> genau dann, wenn gilt

$$\phi^\star(t, u_h, h, f) = \phi(t, u_h, h, f)$$

Beispiel 2.4.2 *Die* <u>*implizite Mittelpunktsregel*</u>

$$\begin{aligned} u_{k+1} &= u_k + hk_1 \\ k_1 &= f(t_k + \frac{h}{2}, u_k + \frac{h}{2}k_1) \end{aligned}$$

und die <u>*implizite Trapezregel*</u>

$$u_{k+1} = u_k + \frac{h}{2}(f(t_k, u_k) + f(t_{k+1}, u_{k+1}))$$

sind symmetrische Verfahren.

Für symmetrische Verfahren erhalten wir nun sogar eine h^2 Entwicklung des Fehlers.

Satz 2.4.3 *Es gelten die Voraussetzungen und Bezeichnungen von Satz 2.4.1. Ist das Verfahren zusätzlich symmetrisch, so enthält die asymptotische Entwicklung (2.4.4) nur gerade Potenzen von* h*, d.h.*

$$y(t) - u_h(t) = \tilde{e}_{2q}(t)h^{2q} + \tilde{e}_{2q+2}(t)h^{2q+2} + \ldots \tag{2.4.5}$$

<u>Beweis:</u>

Aus der Symmetrie folgt für gleiche Startwerte, daß $u_{-h}(t) = u_h(t)$. Dann folgt (2.4.5) durch Addition von (2.4.4) und der asymptotischen Entwicklung des Fehlers des adjungierten Verfahrens (siehe Satz 2.4.2). $\surd$

Wir sehen, daß wir mit Hilfe von symmetrischen Verfahren sehr gute Fehlerentwicklungen bekommen, allerdings sind die einfachen Verfahren leider implizit. Was wir gerne hätten, ist ein explizites Verfahren mit Symmetrie. Das einfachste Verfahren dieser Art ist das Verfahren von Gragg (1963), welches wie folgt konstruiert wird:

$$\begin{array}{rcllr} u_1 & = & u_0 + hf(t_0, u_0) & & \text{expliziter Euler} \\ u_{i+1} & = & u_{i-1} + 2hf(t_i, u_i) & i = 1, 2, \ldots, 2n & \text{explizite Mittelpunktsregel} \\ S_h(t_{2n}) & = & \frac{1}{4}(u_{2n-1} + 2u_{2n} + u_{2n+1}) & & \text{Schlußschritt} \end{array}$$

Man kann zeigen, daß $S_h(t) - y(t)$ eine h^2-Entwicklung hat. Auf dieser Basis werden wir ein Extrapolationsverfahren entwickeln.

Satz 2.4.4 *Sei f hinreichend glatt ($f \in C^{2N+2}$) und seien u_j, S_h wie oben konstruiert, dann gilt*

$$u_h(t) - y(t) = \sum_{j=1}^{N} (u_j(t) + (-1)^{\frac{(t-t_0)}{h}} v_j(t)) h^{2j} + E(t,h) h^{2N+2}$$

mit glatten Funktionen u_j, v_j, wobei $u_j = 0$ oder $u_j = 1$ je nachdem, ob u_h eine Iterierte mit geradem oder ungeradem Index ist. Für $t = t_0 + 2nh$ gilt

$$y(t) - S_h(t) = \sum_{j=1}^{N} e_{2j}(t) h^{2j} + \overline{E}(t,h) h^{2N+2} \tag{2.4.6}$$

mit glatten Funktionen $e_{2j}(t)$, $e_{2j}(t_0) = 0$ und $E, \overline{E}$ beschränkt.

Beweis: siehe Hairer, Nørsett, Wanner.

Nachdem wir nun eine h^2−Entwicklung haben, können wir Extrapolation anwenden, um so Verfahren höherer Ordnung zu entwickeln. Sei H eine Basisschrittweite, sei $\{n_i\}_{i=1}^m$ eine streng monoton wachsende Folge von natürlichen Zahlen und sei $h_i = H/n_i, i = 1, \ldots, m$. Wir berechnen nun die numerischen Resultate unseres Anfangswertproblems durch n_i Schritte mit Schrittweite h_i, um

$$S_{h_i}(t_0 + H) =: T_{i1}$$

zu erhalten (dabei steht 'T' historisch für Trapezregel). Dann eliminieren wir soviele Terme wie möglich aus der asymptotischen Entwicklung (2.4.6) durch Berechnung des Interpolationspolynoms

$$p(h) := \sum_{j=0}^{m-1} \alpha_j h^{2j}, \tag{2.4.7}$$

so daß

$$p(h_i) = T_{i1} = S_{h_i}(t_0 + H) \quad i = 1, \ldots, m. \tag{2.4.8}$$

Als verbesserte Approximation an $y(t_0 + H)$ verwenden wir dann $\alpha_0 = p(0)$ (Extrapolation).

Satz 2.4.5 *Es gilt*

$$\alpha_0 - y(t_0 + H) = O(H^{2m+1})$$

Beweis:

Wegen (2.4.6) gilt

$$T_{i1} = y(t_0 + H) - \sum_{j=1}^{m-1} e_{2j}(t_0 + H)h_i^{2j} - \triangle_i \quad i = 1, \ldots, m$$

wobei $\triangle_i = O(H^{2m+1})$, denn

$$\begin{aligned} \triangle_i &= e_{2m}(t_0 + H)h_i^{2m} + O(h_i^{2m+2}) = O(H^{2m+1}) \\ e_{2m}(t_0) &= 0 \\ H &= n_i h_i. \end{aligned}$$

In Matrixschreibweise gilt

$$\begin{bmatrix} T_{11} \\ \vdots \\ T_{m1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & h_1^2 & \ldots & h_1^{2m-2} \\ \vdots & & & \\ 1 & h_{m-1}^2 & \ldots & h_{m-1}^{2m-2} \end{bmatrix}}_{V} \begin{bmatrix} y(t_0 + H) \\ -e_1(t_0 + H) \\ \vdots \\ -e_{2m-2}(t_0 + H) \end{bmatrix} - \underbrace{\begin{bmatrix} \triangle_1 \\ \vdots \\ \triangle_{m-1} \end{bmatrix}}_{\triangle}.$$

(2.4.7) und (2.4.8) ergeben

$$\begin{bmatrix} T_{11} \\ \vdots \\ T_{m,1} \end{bmatrix} = V \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}.$$

Daher folgt

$$\alpha_0 - y(t_0 + H) = [1, 0, \ldots, 0]V^{-1}\triangle.$$

Sei $\begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = V^{-T} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, so gilt mit der Cramerschen Regel

$$\begin{aligned} |w_k| &= |det(\begin{bmatrix} 1 & \ldots & 1 & 1 & 1 & \ldots & 1 \\ h_1^2 & \ldots & h_{k-1}^2 & 0 & h_{k+1}^2 & & h_{m-1}^2 \\ \vdots & & & \vdots & & & \vdots \\ h_1^{2m-2} & \ldots & h_{k-1}^{2m-2} & 0 & h_{k+1}^{2m-2} & & h_{m-1}^{2m-2} \end{bmatrix})/detV| \\ &= \prod_{i \neq k} h_i^2 \prod_{i>j,i,j\neq k} |h_i^2 - h_j^2| / \prod_{i>j} |h_i^2 - h_j^2| \\ &= c_1(H^2)^{m-2+\frac{1}{2}(m-2)(m-1)}/c_2(H^2)^{\frac{1}{2}(m-1)(m-2)} \\ &= O(1). \end{aligned}$$

Mit $\triangle = O(H^{2m})$ folgt dann die Behauptung. $\surd$

Zur Bestimmung von $\alpha_0 = p(0)$ verwenden wir das Nevill-Aitken-Schema

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{(\frac{h_{i-k}}{h_i})^2 - 1}$$

und $\alpha_0 = T_{mm}$. Als Folge $\{n_i\}$ verwendet man $n_i = 2i$ und erhält für die T_{ik} die Konsistenzordnung $2k$.

Bemerkung 2.4.1 *Man kann alle T_{ik} als explizite Runge-Kutta-Verfahren auffassen. Damit kann man für gerade p die Existenz eines expliziten Runge-Kutta-Verfahrens der Ordnung p mit $\frac{1}{4}\,p^2 + 1$ Stufen zeigen.*

Eigentlich ist das Gragg'sche Verfahren schon kein Einschrittverfahren mehr, wir haben nämlich nicht mehr die Form $u_{i+k} = u_i + h\phi$. Es ist ein sogenanntes Mehrschrittverfahren. Diese werden wir im folgenden Kapitel untersuchen.

2.4.1 Schrittweitensteuerung mittels Extrapolation

Leistungsfähige Verfahren zur Steuerung der Schrittweite lassen sich noch auf andere Weise als in 2.3.2 gewinnen. Anstatt 2 Näherungen zur selben Schrittweite, die aus (verschiedenen) Diskretisierungsverfahren der Ordnung p und $p+1$ stammen, zur Schrittweitensteuerung heranzuziehen, kann man auch 2 Näherungen mit unterschiedlicher Schrittweite aus denselben Diskretisierungsverfahren vergleichen. Im Prinzip ist dies folgendes Problem: Bestimme zu gegebenem t_0, y_0 eine möglichst große Schrittweite h, so daß der Diskretisierungsfehler $e(t_0+h;h)$ nach einem Schritt mit dieser Schrittweite noch unterhalb einer gewissen Schranke ϵ bleibt. Diese Schranke ϵ sollte man nicht kleiner als $K \cdot eps$ wählen, d.h. $\epsilon \geq K \cdot eps$, wobei eps die relative Maschinengenauigkeit und K eine Schranke für die Lösung $y(x)$ in dem in Frage kommenden Bereich ist

$$K \approx max\{|y(x)| : x \in [t_0, t_0 + h]\}.$$

Eine $\epsilon = K \cdot eps$ entsprechende Wahl von h garantiert dann, daß die gefundene Näherungslösung $u(t_0 + h; h)$ im Rahmen der Maschinengenauigkeit mit der exakten Lösung $y(t_0 + h)$ übereinstimmt. Eine solche Schrittweite $h = h(\epsilon)$ mit

$$|e(t_0 + h; h)| \approx \epsilon, \quad \epsilon \geq K \cdot eps,$$

kann oft näherungsweise gewonnen werden. Wie schon gesehen, läßt sich häufig zeigen, daß der Fehler eine asymptotische Entwicklung hat

$$e(t, h) = u_h(t) - y(t) = e_p(t)h^p + O(h^{p+1}) \tag{2.4.9}$$

mit $e_p(t_0) = 0$. Dabei ist $u_h(t)$ wieder eine Näherung an $y(t)$ berechnet mit Schrittweite h. Also gilt in 1. Näherung

$$e(t, h) \approx e_p(t)h^p.$$

Nun ist $e_p(t_0) = 0$, also in 1. Näherung

$$e_p(t) \approx (t - t_0) e_p'(t_0). \tag{2.4.10}$$

Es ist damit $e(t_0 + h, h) \approx \epsilon$, falls

$$\epsilon \approx |e_p(t_0 + h) h^p| \approx |h^{p+1} e_p'(t_0)|.$$

Daraus kann man h berechnen, wenn man $e_p'(t_0)$ kennt

$$h \approx \sqrt[p+1]{\frac{\epsilon}{|e_p'(t_0)|}}. \tag{2.4.11}$$

Einen Näherungswert für $e_p'(t_0)$ kann man wie folgt gewinnen. Benutzt man zunächst die Schrittweite H, um $u_H(t)$ und $u_{\frac{H}{2}}(t)$ für $t = t_0 + H$ zu berechnen, so gilt nach (2.4.9)

$$\begin{aligned} u_{\frac{H}{2}}(t) &= y(t) + e_p(t)(\frac{H}{2})^p + \ldots \\ u_H(t) &= y(t) + e_p(t) H^p + \ldots \end{aligned}$$

Also

$$\frac{u_H(t) - u_{\frac{H}{2}}(t)}{2^p - 1} \approx e_p(t)(\frac{H}{2})^p. \tag{2.4.12}$$

Insbesondere erhält man mit (2.4.10)

$$\frac{u_H(t_0 + H) - u_{\frac{H}{2}}(t_0 + H)}{2^p - 1} \approx \frac{H^p}{2^p} e_p'(t_0) H$$

und daher läßt sich $e_p'(t_0)$ mittels

$$e_p'(t_0) \approx \frac{2^p}{2^p - 1} \frac{1}{H^{p+1}} [u_H(t_0 + H) - u_{\frac{H}{2}}(t_0 + H)] \tag{2.4.13}$$

schätzen. Hiermit kann man nun wiederum h aus (2.4.11) schätzen.

$$\frac{H}{h} \approx \sqrt[p+1]{\frac{2^p}{2^p - 1} \frac{u_H(t_0 + H) - u_{\frac{H}{2}}(t_0 + H)}{\epsilon}}.$$

Praktisch wird man folgendermaßen vorgehen beim Schritt von t_0 nach $t_0 + H$: Man wählt eine Schrittweite H und berechnet $u_H(t_0 + H)$ und $u_{\frac{H}{2}}(t_0 + H)$. Hiermit ermittelt man die Schätzung für $e_p'(t_0)$ nach (2.4.13) und berechnet mit dieser Schätzung h aus (2.4.11). Falls $\frac{H}{2} \gg h$, so verwirft man das errechnete Ergebnis, denn der Fehler ist viel größer als das vorgeschriebene ϵ. Es ist daher zweckmäßig H zu reduzieren. Dazu ersetzt man H durch $2h$, berechnet mit dem neuen H wieder $u_H(t_0 + H), u_{\frac{H}{2}}(t_0 + H)$ und aus (2.4.11) das zugehörige h, bis schließlich $|\frac{H}{h}| \leq 2$ wird. Sobald dies der Fall ist, akzeptiert man $u_{\frac{H}{2}}(t_0 + H)$ als Näherungslösung von $y(t_0 + H)$ und geht zu dem nächsten Integrationsschritt über, indem man t_0, y_0 und H durch die neuen Startwerte $t_0 + H, u_{\frac{H}{2}}(t_0 + H)$ und $2h$ ersetzt.

Beispiel 2.4.3 *Betrachtet wurde das Anfangswertproblem*

$$y' = -200xy^2 \quad y(-3) = \frac{1}{901}$$

mit der exakten Lösung $y(x) = \frac{1}{1+100x^2}$. *Mit dem Runge-Kutta-Verfahren und der besprochenen „Schrittweitensteuerung“ wurde ein Näherungswert* u_h *für* $y(0) = 1$ *berechnet. Dabei wurde das Kriterium* $\frac{H}{h} \gg 2$ *durch die Abfrage* $\frac{H}{h} \geq 3$ *ersetzt. 12-stellige Rechnung liefert:*

$u_h - y(0)$	*Zahl der benötigten Runge-Kutta-Schritte*	*auftretende kleinste Schrittweite H*
$-0.13585 \cdot 10^{-6}$	1476	$0.1226 \ldots 10^{-2}$

Für die feste Schrittweite h liefert das Runge-Kutta-Verfahren:

h	$u_h(0) - y(0)$	*Zahl der Runge-Kutta-Schritte*
$\frac{3}{1476} = 0.2032.. \cdot 10^{-2}$	$-0.5594 \cdot 10^{-6}$	1476
$0.1226.. \cdot 10^{-2}$	$-0.4052 \cdot 10^{-6}$	$\frac{3}{0.1226 \ldots \cdot 10^{-2}} = 2446$

Eine feste Schrittweitenwahl liefert also bei gleichem bzw. größerem Rechenaufwand schlechtere Resultate! Im ersten Fall dürfte die Schrittweite $h = 0.2032 \ldots \cdot 10^{-2}$ *in dem „kritischen“ Bereich nahe 0 zu groß sein; der Diskretisierungsfehler wird zu groß. Die Schrittweite* $h = 0.1226 \ldots \cdot 10^{-2}$ *wird dagegen in dem „harmlosen“ Bereich von* -3 *bis nahe 0 zu klein sein; man macht unnötig viele Schritte und erhält dadurch einen großen Rundungsfehler.*

2.5 Einführung von Mehrschrittverfahren

Ein Nachteil der Einschrittverfahren ist die hohe Anzahl von Funktionsauswertungen, die man für eine hohe Konsistenzordnung benötigt. Die Ordnung eines bestimmten Verfahrens ist fest. Ein Vorteil der Einschrittverfahren besteht darin, daß die Schrittweitensteuerung recht einfach durchzuführen ist.

Im folgenden werden wir Mehrschrittverfahren betrachten, deren Vorteil darin besteht, daß bei geringen Rechenaufwand eine hohe Konsistenzordnung erreichbar ist. Die Ordnung ist für ein gewähltes Verfahren ebenfalls fest, aber Verfahren höherer Ordnung sind leicht konstruierbar. Die Schrittweite sollte nach Möglichkeit konstant sein, sonst ist der Rechenaufwand hoch.

Sei im folgenden $h = a/N$ fest, wobei a die Länge des betrachteten Intervalls sei, und sei $t_j = t_0 + jh, j = 0, \ldots, N$. Ein k-Schrittverfahren zur Lösung der Anfangswertaufgabe $y'(t) = f(t, y(t)), y(t_0) = y_0$ ist ein Verfahren zur Berechnung von Näherungswerten u_{j+k} an $y(t_{j+k})$ aus $u_j, \ldots, u_{j+k-1}$ durch Lösen der Gleichung

$$\frac{1}{h}\sum_{l=0}^{k} \alpha_l u_{j+l} = \phi(t_j, u_j, \ldots, u_{j+k}, h, f) \tag{2.5.1}$$

für $j = 0, \ldots, N-k$ mit geeigneten Startwerten $u_0, \ldots, u_{k-1}$. Dabei ist $\frac{1}{h}\sum_{l=0}^{k} \alpha_l u_{j+l}$ als eine Näherung an die Ableitung y' [bisher: $\frac{1}{h}(u_{j+1} - u_j)$] und $\phi(t_j, u_j, \ldots, u_{j+k}, h, f)$ als Näherung an $f(t, y(t))$ aufzufassen.

Ein lineares k-Schrittverfahren ist von der Form

$$\sum_{l=0}^{k} \alpha_l u_{j+l} = h\sum_{l=0}^{k} \beta_l f(t_{j+l}, u_{j+l}) \tag{2.5.2}$$

für $j = 0, \ldots, N-k$ mit geeigneten Startwerten $u_0, \ldots, u_{k-1}$. Das Verfahren heißt implizit, falls $\beta_k \neq 0$ ist, sonst explizit.

Bei impliziten linearen k-Schrittverfahren muß u_{j+k} aus (2.5.2) iterativ als Lösung dieser i.a. nichtlinearen Gleichung in u_{j+k} berechnet werden. (2.5.2) ist eindeutig lösbar, falls f einer Lipschitzbedingung genügt. Sei dazu o.B.d.A. $\alpha_k = 1$. Eine Fixpunktiteration zur Berechnung von u_{j+k} aus (2.5.2) berechnet Iterierte $u_{j+k}^{(r)}$ gemäß

$$\begin{aligned} u_{j+k}^{(r)} &= -\sum_{l=0}^{k-1} \alpha_l u_{j+l} + h\sum_{l=0}^{k-1} \beta_l f(t_{j+l}, u_{j+l}) + h\beta_k f(t_{j+k}, u_{j+k}^{(r-1)}) \\ &=: \psi(u_{j+k}^{(r-1)}) \end{aligned}$$

Aus der Lipschitzbedingung für f erhalten wir, daß $||\psi(y) - \psi(z)|| \leq |\beta_k| hL ||y - z||$ gilt für alle geeigneten y, z. Ist h hinreichend klein, so gilt auch $||\psi(y) - \psi(z)|| < ||y - z||$ und ψ ist eine Kontraktion, so daß der Banachsche Fixpunktsatz die Existenz einer eindeutigen Lösung u_{j+k} von (2.5.2) sichert.

Wir beschäftigen uns zunächst wieder mit der Frage, wann ein solches Mehrschrittverfahren konsistent ist, d.h. wann diese Diskretisierung der Differentialgleichung lokal vernünftig ist.

Betrachten wir wieder $y'(t) = f(t, y(t))$ mit $y(\hat{t}) = z$. Wir definieren den lokalen Diskretisierungsfehler in Analogie zu den Einschrittverfahren als den Fehler, den man erhält, wenn man die tatsächliche Lösung in die Verfahrensgleichung einsetzt.

Der lokale Diskretisierungsfehler des Mehrschrittverfahrens (2.5.1) ist

$$\tau(\hat{t}, z, h, f) = \frac{1}{h}\sum_{l=0}^{k} \alpha_l y_l - \phi(\hat{t}, y_0, \ldots, y_k, h, f),$$

wobei $y_r = y(\hat{t} + rh)$ ist für alle $r = 0, \ldots, k$.

Das Mehrschrittverfahren heißt konsistent, falls für jedes f, dessen partielle Ableitungen auf $[t_0, t_0 + a] \times \mathbb{R}^n$ existieren, stetig und beschränkt sind, gilt

$$\lim_{h\to 0} \tau(\hat{t}, z, h, f) = 0$$

für alle $\hat{t} \in [t_0, t_0{+}a]$ und alle $z \in \mathbb{R}^n$. Es heißt konsistent von der Ordnung p, falls $\tau(\hat{t}, z, h, f) = O(h^p)$.

Beispiel 2.5.1 *Entwicklung eines Mehrschrittverfahrens*

Betrachte die Taylor-Entwicklungen

$$\begin{aligned} y(t+h) &= y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y'''(t) + \ldots \\ y(t-h) &= y(t) - hy'(t) + \frac{h^2}{2}y''(t) - \frac{h^3}{6}y'''(t) + \ldots \end{aligned}$$

Durch Subtraktion erhält man:

$$\begin{aligned} y(t+h) - y(t-h) &= 2hy'(t) + O(h^3) \\ &= 2hf(t, y(t)) + O(h^3) \end{aligned}$$

Das hieraus gebildete Verfahren $u_{j+2} - u_j = 2hf_{j+1}$ ist ein lineares 2-Schrittverfahren und entspricht als Quadratur der Mittelpunktsregel. Seine Konsistenzordnung ist 2, denn

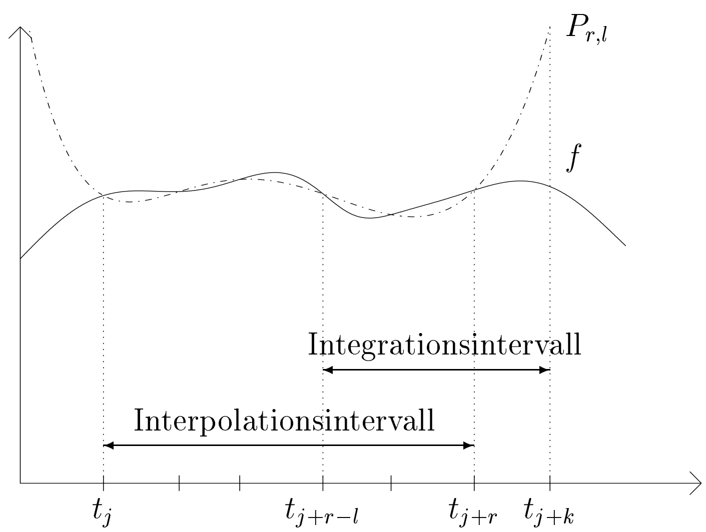
$$\tau(\hat{t}, z, h, f) = \frac{1}{h}[-y(\hat{t}) + y(\hat{t} + 2h)] - 2f(\hat{t} + h, y(\hat{t} + h)) = O(h^2).$$

Anders als bei Einschrittverfahren impliziert bei Mehrschrittverfahren die Konsistenz nicht immer Konvergenz. Hierauf werden wir später noch genau eingehen. Zunächst betrachten wir die Entwicklung der wichtigsten Mehrschrittverfahren, die man über numerische Integration herleiten kann wie die Runge-Kutta-Verfahren.

Es gilt

$$y(t_{j+k}) - y(t_{j+r-l}) = \int_{t_{j+r-l}}^{t_{j+k}} f(s, y(s))ds. \tag{2.5.3}$$

$y(t_{j+k})$ ist der Wert, der genähert werden soll und es ist eine naheliegende Idee, den Integranden f zu ersetzen durch das Interpolationspolynom $P_{r,l}(s) = P_f(s|t_j, \ldots, t_{j+r})$ vom Grad r, das bei $t_j, \ldots, t_{j+r}$ interpoliert. Dabei sind übliche Wahlen $r = k, r = k-1, l = 1, 2$ oder $l = 0, 1$. Das Integral „lebt“ auf $[t_{j+r-l}, t_{j+k}]$, während das Polynom, dessen Integral nun hierauf berechnet wird, auf $[t_j, t_{j+r}]$ interpoliert. Also graphisch betrachtet



Im Mathematischen Praktikum wurde die Polynominterpolation besprochen. Dabei wurden zwei Darstellungsformen für das Interpolationspolynom vorgestellt, die Lagrange-Form und die Newton-Form.

In der Lagrange-Form sieht das Polynom $P_{r,l}(s)$ folgendermaßen aus:

$$P_{r,l}(s) = \sum_{i=0}^{r} f(t_{j+i}, y_{j+i}) \prod_{\substack{p=0 \\ p \neq i}}^{r} \frac{s - t_{j+p}}{(i-p)h}.$$

Also erhalten wir, falls nun $P_{r,l}(s)$ in (2.5.3) für f ersetzt wird:

$$y_{j+k} - y_{j+r-l} \approx h \sum_{i=0}^{r} f(t_{j+i}, y_{j+i}) \beta_i^{(r,l)},$$

wobei

$$h\beta_i^{(r,l)} = \int_{t_{j+r-l}}^{t_{j+k}} \prod_{\substack{p=0 \\ p \neq i}}^{r} \frac{s - t_{j+p}}{(i-p)h} ds = h \int_{-l}^{k-r} \prod_{\substack{p=0 \\ p \neq i}}^{r} \left(\frac{r - p + \tilde{s}}{i - p} \right) d\tilde{s},$$

wie man sofort durch Substitution $s = t_j + (r + \tilde{s})h$ ersieht.

Das Verfahren lautet dann also

$$u_{j+k} = u_{j+r-l} + h \sum_{i=0}^{r} \beta_i^{(r,l)} f_{j+i}, \tag{2.5.4}$$

wobei $f_q = f(t_q, u_q)$ ist.

In der Newton-Darstellung sieht das Polynom $P_{r,l}(s)$ so aus:

$$P_{r,l}(s) = \sum_{i=0}^{r} \underbrace{(s - t_{j+r-i}) \dots (s - t_{j+r-1})}_{:= 1 \text{ für } \ i=0} [t_{j+r-i}, \dots, t_{j+r}] f$$

wobei $[t_{j+r-i}, \dots, t_{j+r}]f$ die i-te dividierte Differenz von f zu den Knoten $t_{j+r-i}, \dots, t_{j+r}$ ist.

In der Notation des Mathematischen Praktikums bzw. der einschlägigen Texte über Polynominterpolation in Büchern sieht man das f in den Knoten $x_0, \dots, x_n$ interpolierende Polynom P_n vom Grade n als

$$P_n(x) = f(x_0) + (x - x_0)[x_0, x_1]f + (x - x_0)(x - x_1)[x_0, x_1, x_2]f + \dots$$

Man kann zeigen, daß sich die dividierten Differenzen wegen des stets gleichen Abstands zwischen den Knoten j bequem mit Hilfe von rückwärtsgenommen Differenzen darstellen lassen. Es ist nämlich mit

$$\nabla f(t_q) = f(t_q) - f(t_{q-1}) \quad \text{und} \quad \nabla^\nu f(t_q) = \nabla(\nabla^{\nu-1} f(t_q))$$

gerade $[t_{j+r-i}, \dots, t_{j+r}]f = \frac{1}{i! h^i} \nabla^i f(t_{j+r})$.

Man erhält damit analog zu (2.5.4) das Verfahren beschrieben durch

$$u_{j+k} = u_{j+r-l} + h \sum_{i=0}^{r} \gamma_i^{(r,l)} \nabla^i f_{j+r}$$

wobei

$$\gamma_i^{(r,l)} \quad = \quad (-1)^i \int_{-l}^{k-r} \binom{-s}{i} ds,$$

was man wieder wie oben mit Substitution herleiten kann. Hierbei ist

$$\binom{-s}{i} = \frac{-s(-s-1)(-s-2)\dots(-s-(i-1))}{i!}.$$

Vorteile der Newton-Form:

- Die numerische Schätzung des Diskretisierungsfehlers kann man wegen der hier einfachen Restglieddarstellung der Polynominterpolation recht leicht erhalten.

- Die Variation der Ordnung ist leicht möglich durch „Anhängen“ von höheren Differenzen.
- Numerisch ist diese Darstellung vorteilhafter, da die Differenzen ∇^i immer kleiner werden und so der Rundungsfehlereinfluß gemindert wird, während bei der Lagrange-Darstellung die verschiedenen Vorzeichen der Koeffizienten problematisch werden können.

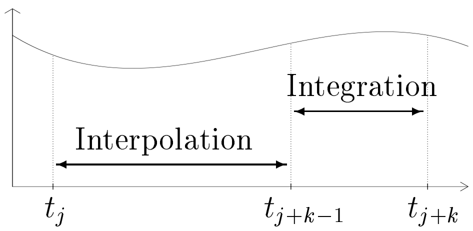
Vorteile der Lagrange-Darstellung:

- weniger Operationen pro Schritt
- Änderung der Schrittweite einfacher

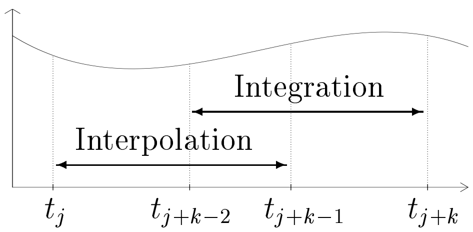
Übliche Wahlen für r und l:

r	l	Name	Art
k - 1	0	Adams-Bashforth	Extrapolation, explizites Verfahren
k - 1	1	Nyström	Extrapolation, explizites Verfahren
k	1	Adams-Moulton	Interpolation, implizites Verfahren
k	2	Milne-Simpson	Interpolation, implizites Verfahren

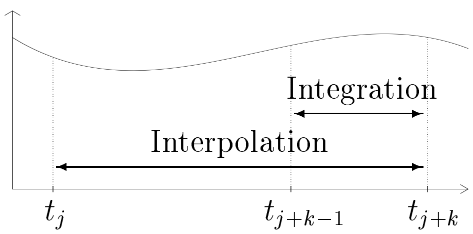
Adams-Bashforth:



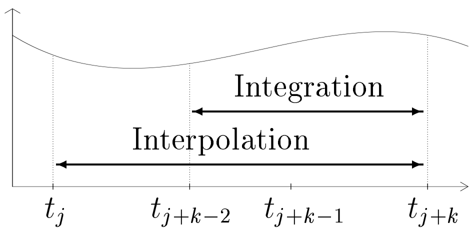
Nyström:



Adams-Moulton:



Milne-Simpson:



Beispiel 2.5.2
Adams-Bashforth: $u_{j+k} = u_{j+k-1} + h\sum_{i=0}^{k-1} \beta_i^{(k-1,0)} f_{j+i}$

$r = k-1 = 0:$ *Hiermit erhält man:* $u_{j+1} = u_j + hf_j$ *(Euler-Verfahren)*

$r = k-1 = 1:$ *Hiermit erhält man:* $u_{j+2} = u_{j+1} + \frac{h}{2}(3f_{j+1} - f_j)$

Nyström: $u_{j+k} = u_{j+k-2} + h\sum_{i=0}^{k-1} \beta_i^{(k-1,1)} f_{j+i}$

$r = k-1 = 0:$ *Hiermit erhält man:* $u_{j+1} = u_{j-1} + 2hf_j$ *(Mittelpunktsregel)*

Adams-Moulton: $u_{j+k} = u_{j+k-1} + h(\sum_{i=0}^{k-1} \beta_i^{(k,1)} f_{j+i} + \beta_k^{(k,1)} f_{j+k})$

u_{j+k} *taucht auf beiden Seiten auf und muß daher durch die Lösung eines nichtlinearen Gleichungssystems iterativ bestimmt werden.*

$r = k = 0$ *Hiermit erhält man:* $u_j = u_{j-1} + hf_j$ *(implizierter Euler)*

$r = k = 1:$ *Hiermit erhält man:* $u_{j+1} = u_j + \frac{h}{2}(f_{j+1} + f_j)$

$r = k = 2:$ *Hiermit erhält man:* $u_{j+2} = u_{j+1} + \frac{h}{12}(5f_{j+2} + 8f_{j+1} - f_j)$

Milne-Simpson: $u_{j+k} = u_{j+k-2} + h\sum_{i=0}^{k} \beta_i^{(k,2)} f_{j+i}$

Bemerkung 2.5.1 *Es ist möglich, Mehrschrittverfahren auch über numerische Differentiation herzuleiten. Die Idee ist dabei, y in $t_j, \dots, t_{j+k}$ durch das Interpolationspolynom P_k vom Grade k zu ersetzen und $P_k'(t_{j+k-r})$ als Näherung für $y'(t_{j+k-r})$ einzusetzen. Mit Hilfe des Newton-*

Interpolationsschemas erhält man

$$P_k(t) = \sum_{i=0}^{k} (-1)^i \begin{pmatrix} -s \\ i \end{pmatrix} \nabla^i y(t_{j+k}) \qquad t = t_{j+k} + sh$$

wobei ∇^i *wieder die dividierte Differenzen seien. Damit erhalten wir*

$$P_k'(t_{j+k-r}) = \frac{1}{h} \sum_{i=0}^{k} \rho_{ri} \nabla^i y(t_{j+k})$$

mit

$$\rho_{ri} = (-1)^i \frac{d}{ds} \begin{pmatrix} -s \\ i \end{pmatrix} |_{s=-r}$$

Als Verfahren ergibt sich

$$\sum_{i=0}^{k} \rho_{ri} \nabla^i u_{j+k} = h f_{j+k-r}$$

Mit $r = 0, \rho_{r0} + \ldots + \rho_{rk} \neq 0$ *ergeben sich implizite Verfahren, mit* $r \neq 0, \rho_{r0} + \ldots + \rho_{rk} \neq 0$ *ergeben sich explizite Verfahren. Außer für* $r = 0$ *und* $k \leq 6; r = 1$ *und* $k \leq 2$ *führt dies aber auf numerisch unstabile Verfahren. Näheres siehe Grigorieff, Band 2.*

2.6 Konsistenz von Mehrschrittverfahren

Für die aus der Polynominterpolation hergeleiteten Verfahren kann man mit Hilfe der Restglieddarstellung der Polynominterpolation leicht die Konsistenzordnung bestimmen.

Satz 2.6.1 *Die Mehrschrittverfahren vom Typ* (r,l) *sind* *konsistent* *für* $f \in C^{r+1}(U)$, *wobei* U *eine Umgebung der Lösung der Anfangswertaufgabe ist, von der* *Ordnung* $p = r+1$, *d.h. insbesondere*

- *das Adams-Bashforth-Verfahren und das Nyström-Verfahren sind konsistent von der Ordnung* $p = k$,
- *das Adams-Moulton-Verfahren und das Milne-Simpson-Verfahren sind konsistent von der Ordnung* $p = k+1$.

Bemerkung 2.6.1 *Diese Ordnungen sind exakt für die drei erstgenannten Verfahren für alle* k *und für das Milne-Simpson-Verfahren für* $k \geq 3$.

Beweis von Satz 2.6.1:

Zur Schreibvereinfachung sehen wir $\tau_j := \tau(t_j, y, h, f)$ und betrachten anstelle eines beliebigen $\hat{t}, z$ wie vorher hier $\hat{t} = t_j$ und $z = y(t_j) = y_j$ und $n = 1$. Dann gilt:

$$\begin{aligned} \tau_j &= \frac{1}{h}(y_{j+k} - y_{j+r-l}) - \sum_{i=0}^{r} \gamma_i^{(r,l)} \nabla^i f_{j+r} \\ &= \frac{1}{h}\int_{t_{j+r-l}}^{t_{j+k}} f(s, y(s))ds - \frac{1}{h}\int_{t_{j+r-l}}^{t_{j+k}} P_{r,l}(s)ds \\ &= \frac{1}{h}\int_{t_{j+r-l}}^{t_{j+k}} (s-t_j)\dots(s-t_{j+r})\frac{D^{r+1}f(\cdot, y(\cdot))(\zeta(s))}{(r+1)!}ds, \end{aligned}$$

wobei $\zeta(s)$ eine Zwischenstelle in dem von $t_j, \dots, t_{j+r}$ und s aufgespannter Intervall ist. Dies ergibt sich gerade aus der Darstellung des Fehlers $f(s, y(s)) - P_{r,l}(s)$ bei der Polynominterpolation. Durch die Substitution $s = t_{j+r} + hx$ erhält man, daß sich obiges Integral schreiben läßt als

$$\frac{1}{(r+1)!}\int_{-l}^{k-r} h^{r+1} x(x+1)\dots(x+r) D^{r+1} f(\cdot, y(\cdot))(\zeta(t_{j+r} + hx))dx.$$

Also finden wir, daß

$$\tau_j = h^{r+1} \underbrace{(-1)^{r+1}\int_{-l}^{k-r} \binom{-x}{r+1}}_{\gamma_{r+1}^{(r,l)}} D^{r+1} f(\cdot, y(\cdot))(\zeta(t_{j+r} + hx))dx.$$

Eine gute Näherung an die Ableitung im Integral ist die $r+1$-te Differenz. $\surd$

Betrachten wir Mehrschrittverfahren wieder allgemein, so stellt man fest, daß bei den Konsistenzüberlegungen zwei Polynome eine wesentliche Rolle übernehmen. Sei also allgemein wieder vorliegend die Verfahrensgleichung

$$\begin{aligned} \frac{1}{h}\sum_{i=0}^{k}\alpha_i u_{j+i} &= \phi(t_j, u_j, \ldots, u_{j+k}, h, f) \qquad (2.6.1)\\ \text{speziell} &= \sum_{i=0}^{k}\beta_i f_{j+i}. \end{aligned}$$

Betrachten wir die Taylorentwicklung bei Einsetzen der Lösung in die Verfahrensgleichung, so erhalten wir aus

$$\begin{aligned} \frac{1}{h}\sum_{i=0}^{k}\alpha_i y(\hat{t}+ih) &= \phi(\hat{t}, y(\hat{t}), \ldots, y(\hat{t}+kh), h, f) + \tau(\hat{t}, y(\hat{t}), h, f)\\ \text{speziell} &= \sum_{i=0}^{k}\beta_i y'(\hat{t}+ih) + \tau(\hat{t}, y(\hat{t}), h, f) \end{aligned}$$

gerade

$$\begin{aligned} \frac{1}{h}\sum_{i=0}^{k}\alpha_i\{y(\hat{t}) + ihy'(\hat{t}) &+ \frac{i^2h^2}{2}y''(\hat{t}) + \ldots\} =\\ &= \phi(\hat{t}, y(\hat{t}), \ldots, y(\hat{t}+kh), h, f) + \tau(\hat{t}, y(\hat{t}), h, f) \qquad (2.6.2)\\ \text{speziell} &= \sum_{i=0}^{k}\beta_i\{y'(\hat{t}) + ihy''(\hat{t}) + \frac{i^2h^2}{2}y'''(\hat{t}) + \ldots\} + \tau(\ldots). \end{aligned}$$

Möchte man hier, daß $\tau(\hat{t}, y(\hat{t}), h, f) = O(h)$ gilt, so muß offenbar mindestens gelten:

$$\sum_{i=0}^{k}\alpha_i = 0.$$

Im linearen Fall ist dann mit der zusätzlichen Bedingung

$$\sum_{i=0}^{k}\alpha_i i = \sum_{i=0}^{k}\beta_i$$

die Konsistenz gesichert. $\tau(\hat{t}, y(\hat{t}), h, f) = O(h^2)$ erhält man, wenn weiter gilt:

$$\sum_{i=0}^{k}\alpha_i i^2 = \sum_{i=0}^{k}2\beta_i i$$

usw.

Definiert man

$$\rho(z) := \sum_{i=0}^{k}\alpha_i z^i, \qquad \sigma(z) := \sum_{i=0}^{k}\beta_i z^i$$

so gilt:

$$\rho(1) = \sum_{i=0}^{k} \alpha_i, \qquad \rho'(1) = \sum_{i=0}^{k} i\alpha_i, \qquad \sigma(1) = \sum_{i=0}^{k} \beta_i.$$

$\rho(z)$ heißt erstes charakteristisches Polynom des k-Schrittverfahrens (2.6.1). $\sigma(z)$ heißt zweites charakteristisches Polynom des linearen k-Schrittverfahrens (2.6.1).

Mit obigen Überlegungen erhält man folgende Ergebnisse:

Lemma 2.6.1 *Sei $y \in C^1(I = [t_0, t_0 + a]), \rho(1) = 0$. Es gelte*

$$\lim_{h \to 0} \max_{t \in I} |\phi(t, y(t), \ldots, y(t + kh), h, f) - \rho'(1) f(t, y(t))| = 0.$$

Dann ist das Verfahren (2.6.1) konsistent.

Beweis: Klar nach (2.6.2) $\surd$

Die obige Bedingung heißt, daß ϕ eine „vernünftige” Inkrementfunktion ist, z.B. bei ESV: $u_{j+1} = u_j + h\phi, \rho(z) = z - 1, \rho'(1) = 1$, und wir erhalten wieder $\lim_{h \to 0} \phi(t, y(t), h, f) = f(t, y(t))$

Lemma 2.6.2 *Das lineare k-Schrittverfahren (2.6.1) ist konsistent genau dann, wenn $\rho(1) = 0$ und $\rho'(1) = \sigma(1)$.*

Beweis: klar nach (2.6.2) und unseren Überlegungen. $\surd$

Die Konsistenz ist hier also äquivalent zu einer algebraischen Bedingung. Im linearen Fall gilt dies sogar für höhere Konsistenzordnungen, wie der folgende Satz zeigt.

Satz 2.6.2 *Folgende Aussagen sind für das lineare k-Schrittverfahren (2.6.1) paarweise äquivalent.*

i. $\sum_{i=0}^{k} (i^l \alpha_i - l i^{l-1} \beta_i) = 0$ *für alle* $l = 0, \ldots, p$.

ii. Das Mehrschrittverfahren ist für alle $f \in C^p$ konsistent von der Ordnung p.

iii. Die Konsistenzordnung für $f(t, y) = y, y(0) = 1$ ist p.

iv. $\frac{\rho(z)}{ln(z)} - \sigma(z)$ *hat eine $p-$fache Nullstelle in 1.*

Beweis:

$''i \Longleftrightarrow ii''$
Nach (2.6.2) ist

$$\begin{aligned}\tau(\hat{t}, y(\hat{t}), h, f) &= \frac{y(\hat{t})}{h}\sum_{i=0}^{k}\alpha_i + y'(\hat{t})\sum_{i=0}^{k}(i\alpha_i - \beta_i) + \frac{y''(\hat{t})}{2}h\sum_{i=0}^{k}(i^2\alpha_i - 2i\beta_i) + \\ &\quad \ldots + \frac{y^{(p)}(\hat{t})}{p!}h^{p-1}\sum_{i=0}^{k}(i^p\alpha_i - pi^{p-1}\beta_i) + O(h^p).\end{aligned}$$

Hieraus folgt sofort: $i \Longleftrightarrow ii$.

$''iv \Longleftrightarrow iii''$
Die Lösung für $y'(t) = y(t), y(0) = 1$ ist $y(t) = e^t$.

Der lokale Diskretisierungsfehler ist für diesen speziellen Fall:

$$\begin{aligned}\tau(t, e^t, h, f) &= \frac{1}{h}\sum_{i=0}^{k}\alpha_i e^{t+ih} - \sum_{i=0}^{k}\beta_i e^{t+ih} \\ &= e^t[\frac{1}{h}\rho(e^h) - \sigma(e^h)] \\ &= e^t[\frac{1}{ln(z)}\rho(z) - \sigma(z)] \\ &= e^t\psi(z)\end{aligned}$$

für $z = e^h$. Dies ist von der Ordnung p genau dann, wenn $\psi(z)$ in 1 eine p-fache Nullstelle hat, d.h.

$$\psi(z) = h^p\psi^{(p)}(1) + \ldots.$$

$''iii \Longrightarrow i''$
Betrachte die Entwicklung in $''i \Longleftrightarrow ii''$ speziell für $y = e^t$. Gilt iii, dann muß für alle h gelten

$$e^t[\frac{1}{h}\sum_{i=0}^{k}\alpha_i + \sum_{i=0}^{k}(i\alpha_i - \beta_i) + \ldots + \frac{h^{p-1}}{p!}\sum_{i=0}^{k}(i^p\alpha_i - pi^{p-1}\beta_i)] = 0.$$

Also gilt i.

$''i \Longrightarrow iii''$
wissen $i \Longrightarrow ii$, $ii \Longrightarrow iii$ ist trivial. $\surd$

Die Größe

$$C_p = \frac{1}{(p+1)!}\sum_{i=0}^{k}(i^{p+1}\alpha_i - (p+1)i^p\beta_i)$$

heißt Fehlerkonstante des linearen k-Schrittverfahrens (2.6.1) der Ordnung p. Ist p minimal mit $C_p \neq 0$, so ist p die genaue Konsistenzordnung. $\hat{C}_p = C_p/\sigma(1)$ heißt normierte Fehlerkonstante.

Korollar 2.6.1 *i) Für $0 \le r \le k$ existieren eindeutig bestimmte $\rho(z) = \sum_{i=0}^{k} \alpha_i z^i$ mit $\alpha_k = 1$ und $\sigma(z) = \sum_{i=0}^{r} \beta_i z^i$, so daß das zugehörige k-Schrittverfahren die Ordnung $k+r$ hat (d.h. $p = 2k$ ist erreichbar).*

ii) Zu vorgegebenem $\rho(z) = \sum_{i=0}^{k} \alpha_i z^i$ mit $\rho(1) = 0$ existiert ein eindeutig bestimmtes $\sigma(z) = \sum_{i=0}^{r} \beta_i z^i, r \le k$, so daß das zugehörige Verfahren die Ordnung $p = r+1$ hat.

Beweis:

Betrachte für α_i, β_i das folgende Gleichungssystem:

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 2 & \dots & k & -1 & -1 & -1 & \dots & -1 \\ 0 & 1 & 2^2 & \dots & k^2 & 0 & -2\cdot 1 & -2\cdot 2 & \dots & -2\cdot r \\ 0 & 1 & 2^3 & \dots & k^3 & 0 & -3\cdot 1 & -3\cdot 2^2 & \dots & -3\cdot r^2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2^p & \dots & k^p & 0 & -p\cdot 1 & -p\cdot 2^{p-1} & \dots & -p\cdot r^{p-1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} = 0. \tag{2.6.3}$$

Nach Satz 2.6.2 ist zu zeigen, daß dieses System für $p = k + r$ eindeutig lösbar ist, wenn $\alpha_k = 1$ festgelegt wird. In diesem Fall kann man das System schreiben als

$$A \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_{k-1} \\ \beta_0 \\ \vdots \\ \beta_r \end{bmatrix} = - \begin{bmatrix} 1 \\ k \\ k^2 \\ \vdots \\ k^p \end{bmatrix},$$

wobei A aus obiger Matrix durch Streichen der $(k+1)$-ten Spalte entsteht. A ist dann eine Matrix mit $p+1$ Zeilen und $k+r+1$ Spalten und damit für $p = k+r$ quadratisch.

Betrachte das homogene Gleichungssystem

$$A^T \begin{bmatrix} \gamma_0 \\ \vdots \\ \gamma_p \end{bmatrix} = 0. \tag{2.6.4}$$

Mit $q(z) = \sum_{i=0}^{p} \gamma_i z^i$ kann man diese $p+1$ Gleichungen schreiben als

$$\begin{aligned} q(l) &= 0 \quad \text{für} \quad l = 0, \dots, k-1 \\ -q'(l) &= 0 \quad \text{für} \quad l = 0, \dots, r \end{aligned}$$

d.h. das Polynom q vom Grad p hat $k+r+1 = p+1$ Nullstellen. Das ist aber nur für $q \equiv 0$ möglich. Also hat (2.6.4) nur die triviale Lösung. A^T bzw. A ist damit nichtsingulär und das Gleichungssystem (2.6.3) ist daher mit $\alpha_k = 1$ eindeutig lösbar. Also gilt i).

Zu ii) Sei ρ mit $\rho(1) = 0$ gegeben. Dann sind in (2.6.3) die Werte $\alpha_0, \ldots, \alpha_k$ bekannt, und da $\sum_{i=0}^{k} \alpha_i = 0$ gilt, kann man aus dem System die erste Gleichung streichen und erhält durch Umschreiben das folgende Gleichungssystem für die Koeffizienten $\beta_0, \ldots, \beta_r$:

$$-\begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 0 & 2\cdot 1 & 2\cdot 2 & \ldots & 2\cdot r \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & p\cdot 1 & p\cdot 2^{p-1} & \ldots & p\cdot r^{p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \vdots \\ \beta_r \end{bmatrix} = b$$

mit passendem Vektor b. Analog zu obigen Überlegungen zeigt man, daß dieses System eindeutig lösbar ist. $\surd$

Bemerkung 2.6.2 *Wir werden (später) sehen, daß $\rho(z)$ für Stabilitätsaussagen geeignet ist. Teil ii) von obigem Korollar sagt, daß es zu jedem guten $\rho(z)$ ein $\sigma(z)$ gibt, so daß die Ordnung $p = r+1$ erreicht wird.*

2.7 Prediktor-Korrektor-Verfahren

Wie kombinieren nun implizite und explizite Verfahren zu sogenannten Prediktor-Korrektor-Verfahren. Dabei dient das explizite Verfahren dazu, einen Näherungswert $u_{j+k}^{(0)}$ für $y(t_{j+k})$ zu verschaffen, der dann als Startnäherung für das iterative Verfahren benutzt wird, das zur Lösung der Gleichung des impliziten Verfahrens in diesem Schritt eingesetzt werden muß. Dabei gibt es verschiedene Möglichkeiten von Kombinationen. Wir wollen dies hier präzisieren und betrachten, welche Konsistenzordnung Prediktor-Korrektor-Verfahren haben.

Gegeben seien ein explizites Verfahren (Prediktor):

$$\sum_{i=0}^{k} \alpha_i^\star u_{j+i} = h \sum_{i=0}^{k-1} \beta_i^\star f_{j+i}, \quad \alpha_k^\star = 1$$

und ein implizites Verfahren (Korrektor):

$$\sum_{i=0}^{k} \alpha_i u_{j+i} = h \sum_{i=0}^{k} \beta_i f_{j+i}, \quad \alpha_k = 1.$$

Dann erhalten wir hiermit das folgende Verfahren:

<u>Prediktor-Korrektor-Verfahren</u>

1. Berechnung der Näherung u_{j+k}
   Sind $u_j, \ldots, u_{j+k-1}$ akzeptierte Näherungen, so berechne
   $u_{j+k} = -\sum_{i=0}^{k-1} \alpha_i^\star u_{j+i} + h \sum_{i=0}^{k-1} \beta_i^\star f_{j+i}$ (Anwendung des Prediktors)
2. Setze $u_{j+k}^{(0)} := u_{j+k}$
3. Iteriere gemäß
   $u_{j+k}^{(m+1)} = -\sum_{i=0}^{k-1} \alpha_i u_{j+i} + h \sum_{i=0}^{k-1} \beta_i f_{j+i} + h\beta_k f(t_{j+k}, u_{j+k}^{(m)})$
   für $m = 0, 1, \ldots$ (Anwendung des Korrektors)

Im allgemeinen iteriert man nicht bis zur Konvergenz, sondern nur eine feste Anzahl m_0 von Schritten, meist ein oder zwei. Dabei gibt es zwei Varianten: Nach Anwendung des Korrektors wird $f(t_{j+k}, u_{j+k}^{(m_0)})$ noch ausgewertet. Dieser f-Wert wird dann bei der Anwendung des Prediktors für den nächsten Gitterpunkt verwendet. Dieses Vorgehen wird mit $\underline{P(EC)^{m_0}E}$ bezeichnet. (Dabei steht P für Prediktor, E für Auswerten von f (Evaluation) und C für Korrektor). Die andere Möglichkeit ist, nach Anwendung des Korrektors mit $f(t_{j+k}, u_{j+k}^{(m_0-1)})$ weiterzurechnen, um eine Auswertung von f zu sparen. Dieses Vorgehen wird mit $\underline{P(EC)^{m_0}}$ bezeichnet.

$P(EC)^{m_0}E$:

$u_j^{(m_0)}, \dots, u_{j+k-1}^{(m_0)}$ als Näherung akzeptiert.

$u_{j+k}^{(0)} = -\sum_{i=0}^{k-1} \alpha_i^\star u_{j+i}^{(m_0)} + h \sum_{i=0}^{k-1} \beta_i^\star f_{j+i}^{(m_0)}$

`FOR` $s = 0, 1, \dots, m_0 - 1$ `DO`

$f_{j+k}^{(s)} = f(t_{j+k}, u_{j+k}^{(s)})$

$u_{j+k}^{(s+1)} = -\sum_{i=0}^{k-1} \alpha_i u_{j+i}^{(m_0)} + h \sum_{i=0}^{k-1} \beta_i f_{j+i}^{(m_0)} + h\beta_k f_{j+k}^{(s)}$

`END FOR`

$f_{j+k}^{(m_0)} = f(t_{j+k}, u_{j+k}^{(m_0)})$

$P(EC)^{m_0}$:

$u_j^{(m_0)}, \dots, u_{j+k-1}^{(m_0)}$ als Näherungen akzeptiert.

$u_{j+k}^{(0)} = -\sum_{i=0}^{k-1} \alpha_i^\star u_{j+i}^{(m_0)} + h \sum_{i=0}^{k-1} \beta_i^\star f_{j+i}^{(m_0-1)}$

`FOR` $s = 0, 1, \dots, m_0 - 1$ `DO`

$f_{j+k}^{(s)} = f(t_{j+k}, u_{j+k}^{(s)})$

$u_{j+k}^{(s+1)} = -\sum_{i=0}^{k-1} \alpha_i u_{j+i}^{(m_0)} + h \sum_{i=0}^{k-1} \beta_i f_{j+i}^{(m_0-1)} + h\beta_k f_{j+k}^{(s)}$

`END FOR`

Beispiel 2.7.1 *Betrachte die Anfangswertaufgabe $y' = \lambda y, y(0) = 1$. Als Korrektor wollen wir die Trapezregel*

$$u_{j+1} = u_j + \frac{h}{2}(f_j + f_{j+1})$$

verwenden. Hier ist also $\alpha_0 = -1, \alpha_1 = 1, \beta_0 = \beta_1 = \frac{1}{2}, k = 1$. Nach Satz 2.6.2 ist die Konsistenzordnung $p = 2$, denn

$$\begin{aligned} l = 0 &: \alpha_0 + \alpha_1 = 0 \\ l = 1 &: \alpha_1 - \beta_0 - \beta_1 = 0 \\ l = 2 &: \alpha_1 - 2\beta_1 = 0 \\ l = 3 &: \alpha_1 - 3\beta_1 \neq 0 \end{aligned}$$

Als Prediktor verwenden wir das Eulerverfahren

$$u_{j+1} = u_j + hf_j,$$

dessen Konsistenzordnung $p = 1$ ist. Wir behandeln (zunächst) den $P(EC)^{m_0}E$-Modus. Für obige Anfangswertaufgabe ist $f_j = \lambda u_j$. Mit $q = h\lambda$ erhält man zunächst aus dem Prediktor-Schritt

$$u_{j+1}^{(0)} = (1+q)u_j^{(m_0)}$$

Einsetzen in die Korrektor-Gleichung ergibt

$$\begin{aligned}
u_{j+1}^{(1)} &= u_j^{(m_0)} + \frac{h}{2}f_j^{(m_0)} + \frac{h}{2}f_{j+1}^{(0)} \\
&= u_j^{(m_0)} + \frac{1}{2}qu_j^{(m_0)} + \frac{1}{2}qu_{j+1}^{(0)} \\
&= u_j^{(m_0)} + \frac{1}{2}qu_j^{(m_0)} + \frac{1}{2}q(1+q)u_j^{(m_0)} \\
&= (1+q+\frac{1}{2}q^2)u_j^{(m_0)} \\
u_{j+1}^{(2)} &= (1+q+\frac{q^2}{2}+\frac{q^3}{4})u_j^{(m_0)}
\end{aligned}$$

Die Lösung der Anfangswertaufgabe ist $y(t) = e^{\lambda t}$, so daß für die entsprechenden lokalen Diskretisierungsfehler gilt

$$\begin{aligned}
\tau^{(0)}(t_j, e^{\lambda t_j}, h, \lambda y) &= \frac{1}{h}(e^{\lambda t_{j+1}} - (1+h\lambda)e^{\lambda t_j}) \\
&= \frac{h}{2}\lambda^2 e^{\lambda t_j} + O(h^2)
\end{aligned}$$

da

$$e^{\lambda t_{j+1}} = e^{\lambda t_j} + h\lambda e^{\lambda t_j} + \frac{1}{2}h^2\lambda^2 e^{\lambda t_j} + \frac{h^3\lambda^3}{6}e^{\lambda t_j} + O(h^4).$$

Weiter gilt

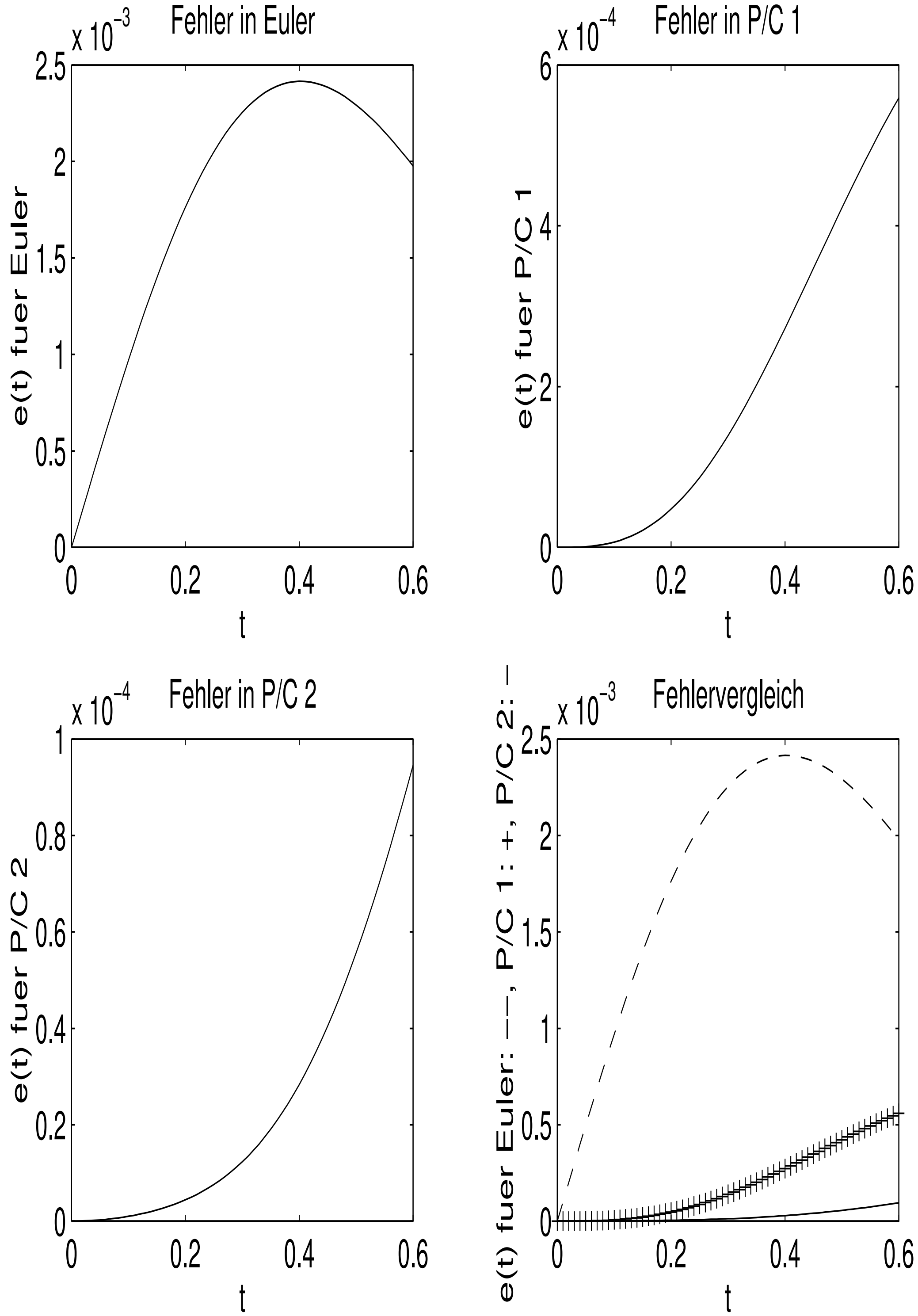
$$\begin{aligned}
\tau^{(1)}(t_j, e^{\lambda t_j}, h, \lambda y) &= \frac{1}{h}(e^{\lambda t_{j+1}} - (1+h\lambda+\frac{h^2\lambda^2}{2})e^{\lambda t_j}) \\
&= \frac{h^2\lambda^3}{6}e^{\lambda t_j} + O(h^3) \\
\tau^{(2)}(t_j, e^{\lambda t_j}, h, \lambda y) &= \frac{1}{h}(e^{\lambda t_{j+1}} - (1+h\lambda+\frac{h^2\lambda^2}{2}+\frac{h^3\lambda^3}{4})e^{\lambda t_j}) \\
&= -\frac{h^2\lambda^3}{12}e^{\lambda t_j} + O(h^3)
\end{aligned}$$

Mit der ersten Korrektur erhöht sich die Konsistenzordnung auf die des Korrektorverfahrens $p = 2$. Eine weitere Iteration verringert jetzt nur die Fehlerkonstante, die ab da unverändert bleibt. Für den $P(EC)^{m_0}$-Modus gestaltet sich die Berechnung der Näherungen komplizierter. Näheres siehe Grigorieff, Band 2, Abschnitt 4.2.

Beispiel 2.7.2 *Die folgenden Grafiken zeigen den Fehler, der bei der Lösung von*

$$y' = -2ty^2, y(0) = 1$$

im Intervall $[0, .6]$ *beim Euler-Verfahren und zwei Prediktor-Korrektor-Verfahren entsteht. Als Prediktor-Verfahren wurde, wie im letzten Beispiel betrachet, das Euler-Verfahren, als Korrektor-Verfahren die Trapezregel verwendet. P/C 1 verwendet einen Korrektor-Schritt, P/C 2 zwei. Die exakte Lösung ist gegeben durch* $y = (1 + t^2)^{-1}$.



Satz 2.7.1 *Konsistenzordnung von Prediktor-Korrektor-Verfahren*

Ist f hinreichend oft differenzierbar, genügt es insbesondere einer Lipschitzbedingung bzgl. y, dann ist die Ordnung des $P(EC)^{m_0}E$ und $P(EC)^{m_0}$ Verfahrens $h^{min(p,m_0+p^\star)}$, wobei $p^\star$ und p die Konsistenzordnung des Prediktor- bzw. Korrektorverfahrens ist, d.h.

$$\begin{aligned} \tau^\star(t+kh,y,h,f) &= c_p^\star h^{p^\star} y^{(p^\star+1)}(t) + O(h^{p^\star+1}) \quad \textit{Prediktor} \\ \tau(t+kh,y,h,f) &= c_p h^p y^{(p+1)}(t) + O(h^{(p+1)}) \quad \textit{Korrektor} \end{aligned}$$

Bemerkung 2.7.1 *Falls $m_0 + p^\star > p$, so ist die Fehlerkonstante die des Korrektors. Es kann also ein Prediktor mit kleinerer Ordnung gewählt und mittels genügender Iteration die Ordnung des Korrektors mit dessen Fehlerkonstante erreicht werden.*

Beispiel 2.7.3 *Prediktor-Korrektor-Verfahren*
Das sogenannte Adams-Bashforth-Moulton-Verfahren, kurz als A-B-M-Methode bezeichnet, ist eine Kombination des impliziten Adams-Moulton-Verfahren mit dem expliziten Adams-Bashforth-Verfahren. Eine solche Kombination von zwei 3-Schrittverfahren zu einer Prediktor-Korrektor-Methode lautet

$$\boxed{\begin{aligned} u_{k+1}^{(P)} &= u_k + \frac{h}{12}\{23f_k - 16f_{k-1} + 5f_{k-2}\} && \textit{Adams-Bashforth} \quad O(h^3) \\ u_{k+1} &= u_k + \frac{h}{24}\{9f(x_{k+1},u_{k+1}^{(P)}) + 19f_k - 5f_{k-1} + f_{k-2}\} && \textit{Adams-Moulton} \quad O(h^4) \end{aligned}} \tag{2.7.1}$$

Als weiteres Beispiel formulieren wir das Prediktor-Korrektor-Verfahren, welches aus der expliziten 4-Schrittmethode von Adams-Bashforth als Prediktor und aus der impliziten 3-Schrittmethode von Adams-Moulton besteht.

$$\boxed{\begin{aligned} u_{k+1}^{(P)} &= u_k + \frac{h}{24}\{55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}\} && \textit{Adams-Bashforth} \quad O(h^4) \\ u_{k+1} &= u_k + \frac{h}{24}\{9f(x_{k+1},u_{k+1}^{(P)}) + 19f_k - 5f_{k-1} + f_{k-2}\} && \textit{Adams-Moulton} \quad O(h^4) \end{aligned}} \tag{2.7.2}$$

Wir behandeln die Anfangswertaufgabe $y' = xe^{x-y}, y(1) = 0$ mit den beiden A-B-M-Methoden 2.7.1 und 2.7.2 und wählen die Schrittweite $h = 0.1$. Um die beiden Verfahren unter gleichen Bedingungen vergleichen zu können, werden in beiden Fällen die drei Startwerte u_1, u_2, u_3 mit der klassischen Runge-Kutta-Methode berechnet. Die untenstehende Tabelle enthält die Näherungswerte, die sowohl mit dem Runge-Kutta-Verfahren vierter Ordnung als auch mit den beiden Prädiktor-Korrektor-Methoden erhalten worden sind, wobei für diese Verfahren auch die Prädiktorwerte $u_{k+1}^{(P)}$ angegeben sind. Die Mehrschrittmethoden arbeiten etwas ungenauer im Vergleich zum Runge-Kutta-Verfahren.

		Runge-Kutta	*A-B-M* 2.7.1		*A-B-M* 2.7.2	
t_k	$u(t_k)$	u_k	$u_k^{(P)}$	u_k	$u_k^{(P)}$	u_k
1.0	0	0				
1.1	0.2626847	0.26268				
1.2	0.5092384	0.5092357				
1.3	0.7423130	0.7423100				
1.4	0.9639679	0.9639648	0.9641567	0.9639538	0.9638945	0.9639751
1.5	1.1758340	1.1758309	1.1759651	1.1758147	1.1757894	1.1758460
1.6	1.3792243	1.3792213	1.3793138	1.3792027	1.3792028	1.3792382
1.7	1.5752114	1.5752087	1.5752742	1.5751894	1.5751995	1.5752258
1.8	1.7646825	1.7646799	1.7647268	1.7646610	1.7646775	1.7646965
1.9	1.9483792	1.9483769	1.9484106	1.9483588	1.9483785	1.9483924
2.0	2.1269280	2.1269258	2.1269503	2.1269089	2.1269298	2.1269401
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2.5	2.9587436	2.9587422	2.9587472	2.9587315	2.9587478	2.9587508
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
3.0	3.7177359	3.7177350	3.7177359	3.7177287	3.7177388	3.7177399

Tabelle : Runge-Kutta-Methode und A-B-M-Verfahren

2.8 Konvergenz von Mehrschrittverfahren

2.8.1 Ein instabiles Mehrschrittverfahren

Wir haben gesehen, daß es sehr einfach ist, lineare Mehrschrittverfahren von hoher Konsistenzordnung zu konstruieren. Anders als im Fall der Einschrittverfahren ist es aber hier so, daß die Konsistenz allein noch keine Konvergenz impliziert. Dazu wollen wir in diesem Abschnitt ausführlich ein Beispiel betrachten, um zu sehen, woran die Konvergenz scheitern kann.

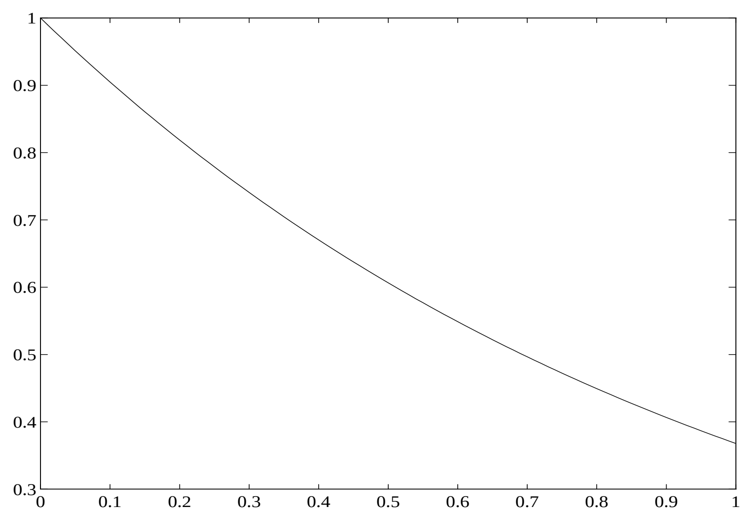
Zur Lösung der Anfangswertaufgabe $y' = -y, y(0) = 1$ betrachten wir das lineare 2-Schrittverfahren

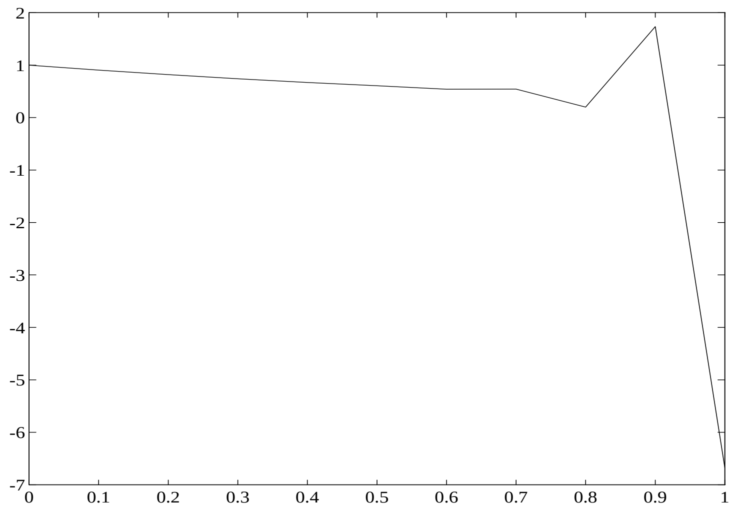
$$u_{j+2} + 4u_{j+1} - 5u_j = h(4f_{j+1} + 2f_j)$$

mit exakten Anfangswerten $u_0 = 1, u_1 = e^{-h}$. Das Verfahren ist nach Satz 2.6.2 von der Ordnung 3. Denn es ist $\alpha_2 = 1, \alpha_1 = 4, \alpha_0 = -5, \beta_0 = 2, \beta_1 = 4$ und daher

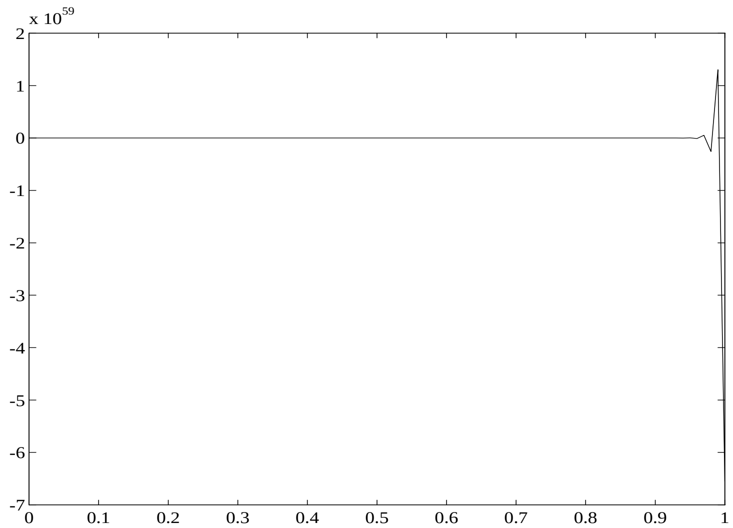
$$\begin{aligned}
l = 0 &: \sum \alpha_i = 0 \\
l = 1 &: \sum i\alpha_i - \sum \beta_i = 4 + 2 - (4 + 2) = 0 \\
l = 2 &: \sum i^2\alpha_i - \sum 2i\beta_i = 4 + 4 - 2 \cdot 4 = 0 \\
l = 3 &: \sum i^3\alpha_i - \sum 3i^2\beta_i = 4 + 8 - 3 \cdot 4 = 0 \\
l = 4 &: \sum i^4\alpha_i - \sum 4i^3\beta_i = 4 + 16 - 4 \cdot 4 \neq 0.
\end{aligned}$$

Wir rechnen dies einmal mit der Schrittweite $h = 0.1$. Die erste Abbildung zeigt noch einmal den Verlauf der Lösungskurve, die sehr harmlos aussieht. Die zweite Abbildung zeigt die Näherungskurve.





Ab 0.8 ist die Näherung überhaupt nicht mehr akzeptabel. Vielleicht ist einfach die Schrittweite zu groß. Rechnen wir noch einmal mit $h = 0.01$, so erhalten wir folgendes Bild:



Beachte die Skalierung!

Wenn wir die Lösungen betrachten, so sehen wir, daß der Fehler in der Nähe von 1 schlechter wird, wenn h kleiner wird. Dies widerspricht ganz unserer Erfahrung bei Einzelschrittverfahren. Rundungsfehler können dieses Verhalten nicht erklären, denn $h = 0.01$ ist noch nicht sehr klein. Um zu sehen, was passiert, schauen uns das Verfahren genauer an.

Da $f(t,u) = -u$, kann man das Verfahren schreiben als

$$u_{j+2} + (4+4h)u_{j+1} - (5-2h)u_j = 0. \tag{2.8.1}$$

Um eine Lösung für diese Differenzengleichung zu erhalten machen wir den Ansatz $u_l = z^l$. Damit erhält man aus der Differenzengleichung:

$$z^l(z^2 + (4+4h)z - (5-2h)) = 0.$$

Diese Gleichung ist erfüllt für

$$\begin{aligned} z_1 &= -2 - 2h + 3\sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2} \\ z_2 &= -2 - 2h - 3\sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2} \end{aligned} \tag{2.8.2}$$

und offenbar auch für alle Linearkombinationen, d.h.

$$u_l = c_1 z_1^l + c_2 z_2^l$$

löst (2.8.1). Da außerdem noch gelten soll $u_0 = 1, u_1 = e^{-h}$, erhalten wir Gleichungen für c_1, c_2

$$\begin{aligned} u_0 &= 1 = c_1 + c_2 \\ u_1 &= e^{-h} = c_1 z_1 + c_2 z_2. \end{aligned}$$

Damit ergibt sich

$$c_1 = \frac{z_2 - e^{-h}}{z_2 - z_1}, \quad c_2 = \frac{e^{-h} - z_1}{z_2 - z_1}. \tag{2.8.3}$$

Also erhalten wir für unsere Näherungen $u_l = c_1 z_1^l + c_2 z_2^l$ mit z_1, z_2 aus (2.8.2), c_1, c_2 aus (2.8.3).

Betrachten wir eine feste Stelle $\hat{t} \in I, \hat{t} = lh$, so bedeutet Konvergenz gerade, daß mit wachsendem $l = \frac{\hat{t}}{h}$ der Wert u_l gegen $y(\hat{t})$ konvergiert. Wir untersuchen daher nun, wie sich u_l mit wachsendem l verhält. Dazu entwickeln wir z_1, z_2 und c_1, c_2 nach h um $h = 0$.

Entwicklung von z_1, z_2 :

Benutze: $(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}x - \frac{1}{4}\frac{x^2}{2} + \frac{3}{8}\frac{x^3}{6} + \dots$.

Entwickeln wir hiermit für z_1 und z_2 die Wurzel, so erhalten wir:

$$\begin{aligned} z_1 &= 1 - h + \frac{h^2}{2} - \frac{h^3}{6} + \frac{h^4}{72} + O(h^5) \\ z_2 &= -5 - 3h + O(h^2). \end{aligned} \tag{2.8.4}$$

Entwicklung von c_1, c_2 :

Benutze: $\frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots$

$$\begin{aligned}
c_1 &= \frac{z_2 - e^{-h}}{z_2 - z_1} \\
&= \frac{-5 - 3h + O(h^2) - e^{-h}}{-6 - 2h + O(h^2)} \\
&= -\frac{1}{6}[-5 - 3h + O(h^2) - e^{-h}][1 - \frac{1}{3}h + O(h^2)] \\
&= 1 + \frac{1}{3}h - \frac{1}{3}h + O(h^2) \\
&= 1 + O(h^2). \qquad (2.8.5)
\end{aligned}$$

Denn $-e^{-h} = -(1 - h + O(h^2))$, so daß $-\frac{1}{6}[-5 - 3h + O(h^2) - e^{-h}]$ gerade $1 + \frac{1}{3}h + O(h^2)$ wird.

$$\begin{aligned}
c_2 &= \frac{e^{-h} - z_1}{z_2 - z_1} \\
&= \frac{e^{-h} - 1 + h - \frac{h^2}{2} + \frac{h^3}{6} - \frac{h^4}{72} + O(h^5)}{-6(1 + \frac{1}{3}h + O(h^2))}.
\end{aligned}$$

Wegen $e^{-h} = 1 - h + \frac{h^2}{2} - \frac{h^3}{6} + \frac{h^4}{24} - \dots$ kann man weiter entwickeln

$$\begin{aligned}
c_2 &= \frac{\frac{h^4}{36} + O(h^5)}{-6(1 + \frac{1}{3}h + O(h^2))} \\
&= [-\frac{1}{216}h^4 + O(h^5)][1 - \frac{1}{3}h + O(h^2)] \\
&= -\frac{1}{216}h^4 + O(h^5). \qquad (2.8.6)
\end{aligned}$$

Wir erhalten also für festes $\hat{t}$ und $h = \frac{\hat{t}}{l}$ mit (2.8.4), (2.8.5) und (2.8.6)

$$\begin{aligned}
u_l &= c_1 z_1^l + c_2 z_2^l \\
&= \{1 + O((\frac{\hat{t}}{l})^2)\}\{1 - \frac{\hat{t}}{l} + O((\frac{\hat{t}}{l})^2)\}^l \qquad (2.8.7) \\
&\quad + \{-\frac{1}{216}(\frac{\hat{t}}{l})^4 + O((\frac{\hat{t}}{l})^5)\}\{-5 - 3\frac{\hat{t}}{l} + O((\frac{\hat{t}}{l})^2)\}^l.
\end{aligned}$$

Nach de l'Hospital (Ableitung nach l) gilt

$$\lim_{l\to\infty} \frac{ln(1 - \frac{\hat{t}}{l})}{\frac{1}{l}} = \lim_{l\to\infty} \frac{\frac{1}{1-\frac{\hat{t}}{l}} \cdot \frac{\hat{t}}{l^2}}{-\frac{1}{l^2}} = \lim_{l\to\infty} \frac{-\hat{t}}{1 - \frac{\hat{t}}{l}} = -\hat{t}$$

und daher $(1-\frac{\hat{t}}{l})^l = exp(l \cdot ln(1-\frac{\hat{t}}{l})) \longrightarrow e^{-\hat{t}}$ für $l \longrightarrow \infty$. Daher konvergiert der erste Term in (2.8.7) gegen $e^{-\hat{t}}$ für $l \longrightarrow \infty$. Der zweite ist näherungsweise

$$-\frac{\hat{t}^4}{216}\frac{(-5)^l(1+\frac{3}{5}\frac{\hat{t}}{l})^l}{l^4} \approx -\frac{\hat{t}^4}{216}\frac{(-5)^l}{l^4}e^{\frac{3\hat{t}}{5}}.$$

Dieser Ausdruck geht aber betragsmäßig für $l \longrightarrow \infty$ gegen $\pm\infty$. Er oszilliert! Daher wird bei Verkleinerung der Schrittweite der Fehler exponentiell größer, wie wir auch in unserer Rechnung gesehen haben. Der Grund dafür liegt darin, daß $\rho(z) = z^2 + 4z - 5$ außer der Wurzel 1 noch die Wurzel -5 hat. Beachte, daß $z_1 \longrightarrow 1, z_2 \longrightarrow -5$ für $h \longrightarrow 0$. Das Wachstum von u_l wird aber im wesentlichen von z_1^l, z_2^l gesteuert.

Bei Einzelschrittverfahren hatten wir $\rho(z) = z - 1$ und daher nur eine Wurzel, die gerade 1 war. Beim Mehrschrittverfahren erhält man durch die Approximation der Ableitung durch höhere Differenzen $\frac{1}{h}\sum_{i=0}^{k}\alpha_i y(t+ih)$ weitere Nullstellen und es schleichen sich weitere Lösungszweige der Form $c_j z_j^l$ ein, die zu anderen, "parasitären" Wurzeln (spurious roots) außer der Hauptwurzel $z_1 = 1$ (principal roots) gehören.

2.8.2 Einiges aus der Theorie der Differenzengleichungen

Um das im letzten Abschnitt gesehene Verhalten von Mehrschrittverfahren allgemeiner zu studieren, machen wir einen kleinen Exkurs in Differenzengleichungen. Eine Gleichung der Form

$$\alpha_k u_{j+k} + \alpha_{k-1}u_{j+k-1} + \ldots + \alpha_0 u_j = 0, \quad \alpha_k \neq 0 \tag{2.8.8}$$

für alle $j \in \mathbb{N}$, mit $\alpha_\nu, u_\nu \in \mathbb{C}$ oder $\mathbb{R}$, nennt man lineare homogene Differenzengleichung k-ter Ordnung mit konstanten Koeffizienten. Eine Folge $(u_l)_{l\in\mathbb{N}}$ heißt Lösung, wenn die u_ν der Gleichung (2.8.8) genügen. Die Lösungen bilden einen Vektorraum.

Man sieht sofort:

- Es gibt zu jedem Satz von Startwerten $u_0, \ldots, u_{k-1}$ genau eine Lösung.
- Es gibt k linear unabhängig Vektoren $u^{(1)}, u^{(2)}, \ldots, u^{(k)}, u^{(i)} = (u_0^{(i)}, \ldots, u_k^{(i)})^T \in \mathbb{C}^{k+1}$, die senkrecht auf $(\alpha_0, \ldots, \alpha_k)^T$ stehen.
- Zu jedem $u^{(i)}$ gibt es eine Lösungsfolge, die gerade mit $u^{(i)}$ beginnt. Der Lösungsraum hat die Dimension k.

Wir wollen in diesem Abschnitt die allgemeine Lösung von (2.8.8) darstellen. Zur Motivation betrachten wir einmal für den Fall $\alpha_k = 1$, den wir ohne Beschränkung der Allgemeinheit annehmen können, folgende aus (2.8.8) resultierende Beziehung

$$U_{j+1} := \begin{bmatrix} u_{j+1} \\ \vdots \\ u_{j+k} \end{bmatrix} = \begin{bmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \dots & -\alpha_{k-2} & -\alpha_{k-1} \end{bmatrix} \begin{bmatrix} u_j \\ \vdots \\ u_{j+k-1} \end{bmatrix} =: AU_j.$$

Damit erhält man

$$U_n = AU_{n-1} = AAU_{n-2} = \dots = A^n U_0 = X^{-1} J^n X U_0,$$

wobei $XAX^{-1} = J$ die Jordansche Normalform von A ist. Falls $J = diag(z_1, \dots, z_k), X^{-1} = [y_1, \dots, y_k]$ und $XU_0 = (b_1, \dots, b_k)^T$, so gilt

$$U_n = [y_1 z_1^n, y_2 z_2^n, \dots, y_k z_k^n] X U_0 = (b_1 y_1) z_1^n + \dots + (b_k y_k) z_k^n.$$

U_n läßt sich also mit Hilfe der Eigenvektoren und Eigenwerte von A angeben. Falls J nicht diagonal ist, so hat man in J^n Blöcke der Form

$$\begin{bmatrix} z_p & 1 & & \\ & z_p & \ddots & \\ & & \ddots & 1 \\ & & & z_p \end{bmatrix}^n = \begin{bmatrix} z_p^n & \binom{n}{1} z_p^{n-1} & \binom{n}{2} z_p^{n-2} & \cdots & \binom{n}{\sigma-1} z_p^{n-\sigma+1} \\ & z_p^n & & & \vdots \\ & & z_p^n & & \vdots \\ & & & \ddots & z_p^n \end{bmatrix},$$

wenn der Block eine $\sigma \times \sigma$-Matrix ist. Zur Erinnerung sei erwähnt, daß $\binom{n}{r} = \frac{n!}{(n-r)!r!} = \frac{n(n-1)\dots(n-r-1)}{r!}$ ist.

Betrachte nun das zur Differenzengleichung (2.8.8) gehörige Polynom

$$\rho(z) = \alpha_k z^k + \dots + \alpha_1 z + \alpha_0.$$

Seien $z_1, \dots, z_l$ die paarweise verschiedenen Nullstellen mit den Vielfachheiten $\sigma_1, \dots, \sigma_l$ mit $\sigma_1 + \dots + \sigma_l = k$, d.h.

$$\rho(z) = \alpha_k (z - z_1)^{\sigma_1} \dots (z - z_l)^{\sigma_l}.$$

Definiere die Folge

$$u^{(p,q)} = (j(j-1)\dots(j-q+1) z_p^{j-q})_{j \in \mathbb{N}}$$

mit der Vereinbarung $j(j-1)\dots(j-q+1) = 1$ für $q = 0$.

Es ist also

$$u^{(p,0)} = (1, z_p, z_p^2, z_p^3, \dots)$$

und für $q \neq 0$

$$u^{(p,q)} = (0, 0, \dots, 0, q!, (q+1)q \dots 2 z_p, (q+2)(q+1) \dots 3 z_p^2, \dots) \tag{2.8.9}$$

mit q führenden Nullen.

Die allgemeine Lösung von (2.8.8) läßt sich hiermit beschreiben.

Lemma 2.8.1 *Jede Lösung von (2.8.8) hat die Form*

$$u = \sum_{p=1}^{l} \sum_{q=0}^{\sigma_p - 1} c_{pq} u^{(p,q)}$$

mit Koeffizienten $c_{pq} \in \mathbb{R}$.

Beweis:

Im ersten Schritt zeigen wir, daß $u^{(p,q)}$ eine Lösung von (2.8.8) ist. Da z_p eine σ_p-fache Nullstelle von $\rho(z)$ ist, gilt für $0 \leq q \leq \sigma_p - 1$

$$\begin{aligned}
0 &= (z^j \rho(z))|_{z=z_p}^{(q)} \\
&= \sum_{\nu=0}^{k} \alpha_\nu z^{j+\nu}|_{z=z_p}^{(q)} \\
&= \sum_{\nu=0}^{k} \alpha_\nu (j+\nu)(j+\nu-1)\dots(j+\nu-q+1) z_p^{j+\nu-q} \\
&= \sum_{\nu=0}^{k} \alpha_\nu u_{j+\nu}^{(p,q)}.
\end{aligned}$$

Im zweiten Schritt zeigen wir, daß die k Folgen $u^{(p,q)}$ für $p = 1, \dots, l$ und $0 \leq q \leq \sigma_p - 1$ linear unabhängig sind. Für die ersten k Elemente von $u^{(p,q)}$ aus (2.8.9) finden wir

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ q! \\ (q+1)q\cdots 2z_p \\ (q+2)\cdots 3z_p^2 \\ \vdots \\ k(k-1)\cdots(k-q+1)z_p^{k-q} \end{bmatrix} = \begin{bmatrix} 1 \\ z \\ z^2 \\ \vdots \\ z^{q-2} \\ z^{q-1} \\ z^q \\ z^{q+1} \\ \vdots \\ z^{k-1} \end{bmatrix}_{|z=z_p}^{(q)}.$$

Betrachten wir das Hermite-Interpolationspolynom mit

$$\sum_{i=0}^{k-1} \beta_i z^i|_{z=z_p}^{(q)} = f_{pq}$$

für $p = 1, \dots, l, q = 0, \dots, \sigma_p - 1$, so sieht man, daß die obigen Vektoren gerade die Zeilen in der Matrix der Bestimmungsgleichung für die $\beta_0, \dots, \beta_{k-1}$ sind. Aus der eindeutigen Lösbarkeit der Hermite-Interpolationsaufgabe folgt damit die lineare Unabhängigkeit der obigen Vektoren für $q = 0, \dots, \sigma_p - 1, p = 1, \dots, l$. $\surd$

Bemerkung 2.8.1 *Insbesondere hat man für $q = 1$:*

$$u^{(p,1)} = \begin{bmatrix} 0 \\ 1 \\ 2z_p \\ 3z_p^2 \\ 4z_p^3 \\ \vdots \end{bmatrix}.$$

2.8.3 Konvergenz und Stabilitätsbedingungen für Mehrschrittverfahren

Wir wollen uns nun der Frage der Konvergenz bei Mehrschrittverfahren zuwenden. Dabei werden wir den Einfluß von Rundungsfehlern gleich mit berücksichtigen.

Statt u_j berechnen wir

$$\tilde{u}_j = y(t_j) + \epsilon_j \qquad j = 0, 1, \ldots, k-1. \tag{2.8.10}$$

Für die weiteren $\tilde{u}_j$ erhalten wir dann aus dem Mehrschrittverfahren

$$\sum_{r=0}^{k} \alpha_r \, \tilde{u}_{j+r} = h\, \phi(t_j, \tilde{u}_j, \ldots, \tilde{u}_{j+k}, h, f) + h\epsilon_{j+k}. \tag{2.8.11}$$

Wir betrachten nun wieder eine feste Stelle $\hat{t} \in I = [t_0, t_0 + a]$ und setzen $h_j = \frac{\hat{t}-t_0}{j}, j = 0, 1, 2, \ldots$. Für $j \to \infty$ erhalten wir $\tilde{u}_j$ als berechnete Näherung für $y(\hat{t})$. Für diese feste Stelle $\hat{t}$ setzen wir

$$\epsilon(\hat{t}, h_j) = \epsilon_j, \quad \tilde{u}_{h_j}(\hat{t}) = \tilde{u}_j(\hat{t}, \epsilon)$$

Das durch $\rho(z) = \sum\limits_{r=0}^{k} \alpha_r z^r$ und die Funktion ϕ gegebene Mehrschrittverfahren heißt <u>konvergent</u>, falls für alle $t \in [t_0, t_0 + a]$, für alle $f \in C^1(U)$ (U Umgebung der Lösung) und für alle $\epsilon(t, h)$, für die es ein $\psi(h) \geq 0$ gibt mit $\psi(h) \to 0$ für $h \to 0$, so daß

$$|\epsilon(t, h)| \leq \psi(h)$$

gilt, daß

$$\lim_{j\to\infty} \tilde{u}_{h_j}(t, \epsilon) = y(t).$$

Der folgende Satz gibt eine notwendige Bedingung für die Konvergenz von Mehrschrittverfahren an.

Satz 2.8.1 *Wurzelbedingung*
Falls $\phi(t, \eta_j, \ldots, \eta_{j+k}, h, 0) = 0$ für alle $(t, \eta_j, \ldots, \eta_{j+k})$ (gilt z.B. für lineare Mehrschrittverfahren) und falls das zu $\rho(z) = \sum\limits_{r=0}^{k} \alpha_r z^r$ und ϕ gehörende Mehrschrittverfahren konvergent ist, so gilt die Stabilitätsbedingung (auch Wurzelbedingung):

Ist z eine Nullstelle von ρ, so ist $|z| < 1$ oder $|z| = 1$ und in diesem Fall ist die Nullstelle einfach.

<u>Beweis:</u>

Da das Mehrschrittverfahren konvergent sein soll, ist es insbesondere konvergent für $y' = 0, y(t_0) = 0$ mit der exakten Lösung $y \equiv 0$. Sei z eine Nullstelle von $\rho(z)$. Seien $\epsilon_\nu = h_j z^\nu$ für $\nu = 0, 1, \ldots, k-1$ spezielle Anfangsfehler und $\epsilon_\nu = 0$ für alle $\nu \geq k$. Die $\tilde{u}_j$ sind hier wegen $\phi(t, u_j, \ldots, u_{j+k}, h, 0) \equiv 0$ durch die Rekursionsformel

$$\begin{aligned} \tilde{u}_i &= \epsilon_i \qquad\qquad i = 0, 1, \ldots, k-1 \\ \sum_{\nu=0}^{k} \alpha_\nu \, \tilde{u}_{j+\nu} &= h_j \epsilon_{j+k} = 0 \end{aligned}$$

bestimmt. Nach Lemma 2.8.1 löst $(1, z, z^2, z^3, \ldots)$ obige Differenzengleichung. Wegen $\tilde{u}_i = \epsilon_i = h_j z^i$ für $i = 0, \ldots, k-1$ folgt daher $\tilde{u}_\nu = h_j z^\nu$. Die Konvergenz des Mehrschrittverfahrens ergibt nun

$$0 = \lim_{j\to\infty} \tilde{u}_j = \lim_{j\to\infty} h_j z^j = \lim_{j\to\infty} \frac{\hat{t} - t_0}{j} \, z^j .$$

Dies geht nur, wenn $|z|^j$ beschränkt ist, d.h. falls $|z| \leq 1$ gilt. Angenommen nun gilt, daß z mit $|z| = 1$ eine mehrfache Nullstelle ist. Wähle dann $\epsilon_0 = 0, \epsilon_\nu = \nu h_j z^{\nu-1}$ für $\nu = 1, \ldots, k-1$ so folgt, daß

$$\rho'(z) = k\alpha_k z^{k-1} + (k-1)\alpha_{k-1}\alpha_{k-1} \, z^{k-2} + \ldots + \alpha_1 = 0$$

von $\tilde{u}_\nu = \nu h_j z^{\nu-1}$ erfüllt wird. Die Konvergenz des Mehrschrittverfahrens ergibt nun

$$0 = \lim_{j\to\infty} \tilde{u}_j = \lim_{j\to\infty} j \cdot h_j \, z^{j-1} = \lim_{j\to\infty} (t - t_0) z^{j-1} .$$

Dies geht nur für $|z| < 1$, also haben wir einen Widerspruch und z mit $|z| = 1$ kann nur eine einfache Nullstelle sein. $\surd$

Bemerkung 2.8.2 *Ergebnis für Differenzengleichungen*
Insbesondere haben wir gezeigt: Falls für alle Startwerte $\delta = (\delta_0, \ldots, \delta_{k-1})^T$ die Lösung $(u_j(\delta))_{j\in\mathbb{N}}$ der Differenzengleichung $\alpha_k u_{j+k} + \ldots + \alpha_0 u_j = 0$ der Bedingung

$$\lim_{j\to\infty} \frac{u_j(\delta)}{j} = 0 \tag{2.8.12}$$

genügt, dann erfüllt $\rho(z)$ die Stabilitätsbedingung. Umgekehrt: Gilt die Stabilitätsbedingung, so sieht man sofort, daß gilt $\sup_j |u_j(\delta)| < \infty$, insbesondere also auch (2.8.12).

Wir zeigen jetzt, daß für konsistente Verfahren die Stabilitätsbedingung auch hinreichend für die Konvergenz ist. Dazu benötigen wir das folgende Lemma:

Lemma 2.8.2 *Sei* $\rho(M) = \max\limits_{\lambda \text{ Eigenwert von } M} |\lambda|$ *der Spektralradius von* M *für* $M \in \mathbb{C}^{n\times n}$.

1. *Zu jedem* $M \in \mathbb{C}^{n\times n}$ *und jedem* $\epsilon > 0$ *gibt es eine Vektornorm* $||\cdot||$*, so daß für die zugehörige Matrixnorm* $||M|| = \sup\limits_{||x||=1} ||Mx||$ *gilt*

$$||M|| \leq \rho(M) + \epsilon$$

2. *Sei* λ *Eigenwert von* M *mit* $|\lambda| = \rho(M)$*, und sei die algebraische Vielfachheit von* λ *gleich der geometrischen (d.h.* λ *hat nur einfache Elementarteiler), dann existiert eine Vektornorm* $||\cdot||$*, so daß für die zugehörige Matrixnorm* $||M|| = \sup\limits_{||x||=1} ||Mx||$ *gilt*

$$||M|| = \rho(M).$$

Beweis:

Sei $TMT^{-1} = J$ die Jordansche Normalform von M mit

$$J = diag(c_k(\lambda_i)), \qquad c_k(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{k\times k}.$$

Mit $D_\epsilon = diag\,(1, \epsilon, \epsilon^2, \ldots, \epsilon^{k-1})$, $\epsilon > 0$ folgt

$$D_\epsilon^{-1} J D_\epsilon = diag\,(c_k^{(\epsilon)}(\lambda_i))$$

mit

$$c_k^{(\epsilon)}(\lambda_i) = \begin{bmatrix} \lambda_i & \epsilon & & \\ & \lambda_i & \ddots & \\ & & \ddots & \epsilon \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{k\times k}.$$

Dann gilt für $||\cdot||_\infty$

$$\begin{aligned} ||D_\epsilon^{-1} TMT^{-1} D_\epsilon||_\infty &= ||D_\epsilon^{-1} J D_\epsilon||_\infty \\ &\leq \rho(M) + \epsilon. \end{aligned}$$

Allgemein gilt für nichtsinguläres B und einer Vektornorm $||\cdot||$, daß $||x||_B := ||Bx||$ eine Vektornorm und die zugehörige Operatornorm $||A||_B := ||BAB^{-1}||$ ist. Die in 1. gesuchte Vektornorm ist daher

$$||x|| := ||D_\epsilon^{-1} T x||_\infty.$$

Zur Konstruktion der in 2. gesuchten Vektornorm ordne die Eigenwerte von M absteigend an: $|\lambda_1| = |\lambda_2| = \ldots = |\lambda_s| > |\lambda_{s+1}| \geq \ldots \geq |\lambda_n|$ und wähle $\epsilon = \rho(M) - |\lambda_{s+1}|$. Dann ist

$$J = \begin{bmatrix} \lambda_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_s & & & & & \\ & & & \lambda_{s+1} & 1 & & & \\ & & & & \lambda_{s+1} & \ddots & & \\ & & & & & \ddots & 1 & \\ & & & & & & \lambda_{s+1} & \\ & & & & & & & \ddots \end{bmatrix}$$

und mit D_ϵ und T wie eben ist dann

$$\begin{aligned} ||D_\epsilon^{-1} T M T^{-1} D_\epsilon||_\infty &= ||D_\epsilon^{-1} J D_\epsilon||_\infty \\ &= \rho(M). \end{aligned}$$

Die gesuchte Vektornorm ist daher dieselbe wie eben. $\surd$

Wir brauchen hier wieder eine Lipschitzbedingung für ϕ (wie bei den Einschrittverfahren). Diese hat die Form (wieder nur der Einfachheit halber für $n = 1$) : Für alle $f \in C^1(I \times \mathbb{R})$ gelte: Es gibt ein $h_0 > 0$ und ein $L < \infty$ mit

$$|\phi(t, \eta_j, \ldots, \eta_{j+k}, h, f) - \phi(t, \tilde{\eta}_j, \ldots, \tilde{\eta}_{j+k}, h, f)| \leq L \sum_{\nu=0}^{k} |\eta_{j+\nu} - \tilde{\eta}_{j+\nu}| \tag{2.8.13}$$

für alle $h < h_0$ und alle $\eta_{j+\nu}, \tilde{\eta}_{j+\nu} \in \mathbb{R}$.

Satz 2.8.2 *Das durch $\rho(z)$ und ϕ gegebene Mehrschrittverfahren sei konsistent und erfülle (2.8.13). Dann gilt: Das Verfahren ist genau dann konvergent, wenn es die Stabilitätsbedingung erfüllt.*

<u>Beweis:</u>

$''\Longrightarrow''$: nach Satz 2.8.1.
$''\Longleftarrow''$: Sei nun die Stabilitätsbedingung erfüllt. Seien $\tilde{u}_j, \epsilon_j$ wie in (2.8.10) und (2.8.11). Dann gilt für den Fehler $e_j = \tilde{u}_j - y(t_j)$

$$e_j = \epsilon_j \quad \text{für} \quad j = 0, \ldots, k-1$$

und

$$\begin{aligned} \sum_{\nu=0}^{k} \alpha_\nu e_{j+\nu} &= h[\phi(t_j, \tilde{u}_j, \ldots, \tilde{u}_{j+k}, h, f) - \phi(t_j, y(t_j), \ldots, y(t_{j+k}), h, f)] \\ &\qquad + h(\epsilon_{j+k} - \tau_{j+k}) \\ &= : c_{j+k} \end{aligned} \tag{2.8.14}$$

(mit $\tau_{j+k} = \tau(t_{j+k}, y, h, f)$).
Wegen der Lipschitzbedingung (2.8.13) gilt

$$|c_{j+k}| \leq hL \sum_{\nu=0}^{k} |e_{j+\nu}| + h(|\epsilon_{j+k}| + |\tau_{j+k}|). \tag{2.8.15}$$

Definiere

$$E_j = \begin{bmatrix} e_j \\ e_{j+1} \\ \vdots \\ e_{j+k-1} \end{bmatrix} \quad C_{j+k} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ c_{j+k} \end{bmatrix} \quad M = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \dots & -\alpha_{k-2} & -\alpha_{k-1} \end{bmatrix} \in \mathbb{C}^{k\times k}.$$

Dabei sei o.B.d.A. $\alpha_k = 1$. (2.8.14) ist dann äquivalent zu

$$\begin{aligned} E_{j+1} &= ME_j + C_{j+k} \\ \begin{bmatrix} e_{j+1} \\ e_{j+2} \\ \vdots \\ e_{j+k-1} \\ e_{j+k} \end{bmatrix} &= \begin{bmatrix} e_{j+1} \\ e_{j+2} \\ \vdots \\ e_{j+k-1} \\ c_{j+k} - \sum_{\nu=0}^{k-1} \alpha_\nu e_{j+\nu} \end{bmatrix}. \end{aligned} \tag{2.8.16}$$

Nun ist $(-1)^k\rho(z)$ das charakteristische Polynom von M. Aus der Konsistenz des Mehrschrittverfahrens folgt $\rho(1) = 0$. Daher ist 1 Eigenwert von M. Wegen der Stabilitätsbedingung sind alle Eigenwert vom Betrag ≤ 1 und die vom Betrag 1 sind einfach. Insbesondere gilt also $\rho(M) = 1$. Nach Lemma 2.8.2 gibt es Vektornorm $||\cdot||_\sim$ mit $||M||_\sim = \sup\limits_{||x||=1} ||Mx||_\sim = \rho(M) = 1$, also

$$||Mx||_\sim \leq ||x||_\sim \quad \text{für alle} \quad x \in \mathbb{C}^k. \tag{2.8.17}$$

Da in $\mathbb{C}^k$ alle Normen äquivalent sind, gibt es eine Konstante R mit

$$\frac{1}{R}||x||_\sim \leq \underbrace{\sum_{\nu=1}^{k} |x_\nu|}_{=:||x||_1} \leq R||x||_\sim$$

für alle $x = [x_1, \dots, x_k]^T \in \mathbb{C}^k$. Damit folgt

$$||E_j||_1 = \sum_{\nu=0}^{k-1} |e_{j+\nu}| \leq R||E_j||_\sim$$

und

$$||E_{j+1}||_1 = \sum_{\nu=1}^{k} |e_{j+\nu}| \leq R||E_{j+1}||_\sim.$$

Aus (2.8.15) folgt

$$|c_{j+k}| \le hLR(||E_j||_\sim + ||E_{j+1}||_\sim) + h(|\epsilon_{j+k}| + |\tau_{j+k}|), \tag{2.8.18}$$

da $\sum\limits_{\nu=0}^{k} |e_{j+\nu}| \le \sum\limits_{\nu=1}^{k-1} |e_{j+\nu}| + \sum\limits_{\nu=1}^{k} |e_{j+\nu}|$.

Anwendung von (2.8.16) und (2.8.17) ergibt

$$\begin{aligned} ||E_{j+1}||_\sim &\le ||E_j||_\sim + ||C_{j+k}||_\sim \\ &\le ||E_j||_\sim + R||C_{j+k}||_1 \\ &\le ||E_j||_\sim + R|c_{j+k}| \\ &\le ||E_j||_\sim + hLR^2(||E_j||_\sim + ||E_{j+1}||_\sim) + hR(|\epsilon_{j+k}| + |\tau_{j+k}|) \end{aligned}$$

wegen (2.8.18). Also

$$(1 - hLR^2)||E_{j+1}||_\sim \le (1 + hLR^2)||E_j||_\sim + hR(|\epsilon_{j+k}| + |\tau_{j+k}|).$$

Dies ist äquivalent zu

$$||E_{j+1}||_\sim \le \frac{1 + hLR^2}{1 - hLR^2}||E_j||_\sim + \frac{hR}{1 - hLR^2}(|\epsilon_{j+k}| + |\tau_{j+k}|). \tag{2.8.19}$$

Für $h \le \frac{1}{2LR^2}$ ist nun $(1 - hLR^2) \ge \frac{1}{2}$ und

$$\frac{1 + hLR^2}{1 - hLR^2} \le 1 + 4hLR^2.$$

(2.8.19) ergibt somit für $h \le \frac{1}{2LR^2}$

$$||E_{j+1}||_\sim \le (1 + 4hLR^2)||E_j||_\sim + 2hR(|\epsilon_{j+k}| + |\tau_{j+k}|).$$

Anwendung von Lemma 2.2.1 ergibt

$$\begin{aligned} ||E_{j+1}||_\sim &= (||E_0||_\sim + \sum_{\nu=0}^{j} 2hR(|\epsilon_{j+k}| + |\tau_{j+k}|))e^{\sum\limits_{\nu=0}^{j} 4hLR^2} \\ &\le (||E_0||_\sim + 2hR(j+1)(\epsilon + \tau))e^{(j+1)4hLR^2} \end{aligned}$$

mit $\epsilon = \max\limits_j |\epsilon_{j+k}|, \tau = \max\limits_j |\tau_{j+k}|$. D.h. man hat für $t \ne t_0, h = \frac{t-t_0}{j+k}, h < \frac{1}{2LR^2}$

$$||E_{j+1}||_\sim \le (||E_0||_\sim + 2R(t - t_0)\frac{j+1}{j+k}(\epsilon + \tau))e^{(t-t_0)4LR^2\frac{j+1}{j+k}}.$$

Da das Mehrschrittverfahren konsistent ist, folgt $\tau \to 0$ für $j \to \infty$. Falls nun die ϵ_l gegen 0 gehen für $h \to 0$ (d.h. $\epsilon \to 0$), so geht nach der letzten Ungleichung auch $||E_{j+1}||_\sim \to 0$ und damit $e_{j+k} \to 0$. $\surd$

Korollar 2.8.1 *Ist das Mehrschrittverfahren ein konsistentes Verfahren $p-$ter Ordnung und sind die Anfangsfehler ϵ_i ebenfalls von der Ordnung p, dann gilt auch für den globalen Diskretisierungsfehler*

$$|u(t,\epsilon,h_n)-y(t)| = O(h_n^p)$$

für alle $h_n = \frac{t-t_0}{n}$, n genügend groß.

Bemerkung 2.8.3 *Genügt $\rho(z)$ der Stabilitätsbedingung, so heißt das Verfahren auch null-stabil. Diese Stabilität hängt nur von ρ, nicht von ϕ ab.*

Alle k-Schrittverfahren von Adams-Bashforth und Adams-Moulton erfüllen die Bedingung der Nullstabilität, denn ihre ersten charakteristischen Polynome $\rho(z) = z^m - z^{m-1} = z^{m-1}(z-1)$ haben die einfache Nullstelle $z_1 = 1$ und die $(m-1)$-fache Nullstelle $z_2 = z_3 = \ldots = z_m = 0$. Dasselbe trifft zu für die Nystrom- und Milne-Simpson-Methoden, da ihre ersten charakteristischen Polynome $\rho(z) = z^m - z^{m-2} = z^{m-2}(z^2-1)$ die einfachen Nullstellen $z_1 = 1$ und $z_2 = -1$ auf dem Rand des Einheitskreises und die $(m-2)$-fache Nullstelle $z_3 = \ldots = z_m = 0$ im Nullpunkt besitzen.

Beispiel 2.8.1 *Wir wollen noch ein explizites $2-$Schrittverfahren*

$$a_0u_{k-1} + a_1u_k + a_2u_{k+1} = h\{b_0f_{k-1} + b_1f_k\}$$

ohne jede Einschränkung an die Koeffizienten herleiten, welches maximale Ordnung haben soll. Mit den vier freien Parametern kann die Ordnung $p = 3$ erreicht werden, falls sie Lösung des folgenden Gleichungssystems sind (siehe Satz 2.6.2)

$$\begin{aligned}
a_0 + a_1 + a_2 &= 0 \\
a_1 + 2a_2 - b_0 - b_1 &= 0 \\
a_1 + 4a_2 - 2b_1 &= 0 \\
a_1 + 8a_2 - 3b_1 &= 0
\end{aligned}$$

Seine Lösung $a_0 = -5, a_1 = 4, a_2 = 1, b_0 = 2, b_1 = 4$ ergibt die $2-$Schrittmethode

$$u_{k+1} = 5u_{k-1} - 4u_k + h\{4f_k + 2f_{k-1}\}.$$

Obwohl die Ordnung des Verfahrens $p = 3$ ist, wird sich zeigen, daß die Methode nicht brauchbar ist, weil die Mehrschrittmethode mit dem ersten charakteristischen Polynom $\rho(z) = z^2 + 4z - 5$ und den Nullstellen $z_1 = 1, z_2 = -5$ nicht null-stabil ist. Zur Konkretisierung der Situation sind in untenstehender Tabelle für $y' = -y, y(0) = 1$ mit der exakten Lösung $y(x) = e^{-x}$ und einige Werte von h die Lösungen z_1 und z_2 der Gleichung $z^2 + 4(1+h)z - (5-2h) = 0$ und die Koeffizienten c_1 und c_2 der Lösung $u_k = c_1z_1^k + c_2z_2^k$ der Differenzengleichung angegeben, falls die exakten Anfangsstartwerte $u_0 = 1, u_1 = e^{-h}$ verwendet werden. Die Zahlwerte zeigen, daß der zweite Anteil in y_k für $k \to \infty$ oszillierend zunimmt, während der erste Anteil die gesuchte Lösung qualitativ richtig wiedergeben würde, denn z_1 ist eine gute Approximation für e^{-h}.

h	e^{-h}	z_1	z_2	c_1	c_2
0.4	0.670320046	0.669870315	-6.269870	1.000064805	$-6.481 \cdot 10^{-5}$
0.2	0.818730753	0.818695388	-5.618695	1.000005494	$-5.494 \cdot 10^{-6}$
0.1	0.904837418	0.904834939	-5.304835	1.000000399	$-3.992 \cdot 10^{-7}$
0.05	0.951229425	0.951229261	-5.151229	1.000000027	$-2.688 \cdot 10^{-8}$
0.02	0.980198673	0.980198669	-5.060199	1.000000001	$-7.195 \cdot 10^{-10}$
0.01	0.990049834	0.990049834	-5.030050	1.000000000	$-4.634 \cdot 10^{-11}$
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$		
0	1	1	-5		

Tabelle: Zur Instabilität der Methode

Wir hatten gesehen, daß für lineare Mehrschrittverfahren die Ordnung $p = 2k$ erreichbar ist (siehe Korollar 2.6.1). Es gilt jedoch

Satz 2.8.3 *Dahlquist*
Ein lineares k-Schrittverfahren erfülle die Stabilitätsbedingung und besitze für $f \in C^p(U)$ die Ordnung p. Dann ist $p \leq k + 2$.

Beweis: siehe Grigorieff. Band 2.

2.9 Stabilitätsbegriffe, Stabilitätsbereiche

Bisher haben wir Sätze über Konvergenz hergeleitet, die die Verhältnisse bei gegen Null gehenden Schrittweiten wiedergeben. Jede Rechnung wird jedoch mit einem Schrittweitenvektor $h \neq 0$ durchgeführt, von dem man i.a. nicht weiß, ob er genügend klein ist, daß die asymptotische Theorie das Verhalten der Lösungen (wenigstens) qualitativ richtig beschreibt. Statt an Aussagen für genügend kleines h ist man daher primär an einer Beschreibung des Verhaltens der Näherungslösung für die gerade gewählte Schrittweite interessiert.

Es gibt sehr viele Stabilitätskonzepte, siehe z.B. bei Grigorieff. Wir diskutieren hier zwei Konzepte. Im ersten wird ein Verfahren als stabil bezeichnet, wenn die absoluten Fehler $u_j - y(t_j)$ mit wachsendem j kleiner werden. Dies ist die sogenannte absolute Stabilität. Im zweiten Fall wird ein Verfahren als stabil bezeichnet, wenn die Fehler nicht stärker wachsen als die Lösung. Dies ist die relative Stabilität.

Vorab wollen wir nochmal auf die folgende spezielle Frage zur Stabilität eingehen. Wie hängt die Lösung $y(t)$ der Anfangswertaufgabe von Störungen im Anfangswert ab? Die Antwort darauf gibt der Satz 2.1.3

$$||y(t, z_1) - y(t, z_2)|| \leq e^{L(t-t_0)}||z_1 - z_2||.$$

Als Beispiel hatten wir schon die Differentialgleichung

$$y' = 10(y - \frac{t^2}{1+t^2}) + \frac{2t}{(1+t^2)^2}$$

betrachtet. Für $y(0) = 0$ ergibt sich $y(t) = \frac{t^2}{1+t^2}$, für $y(0) = \epsilon$ ergibt sich $\tilde{y}(t) = \frac{t^2}{1+t^2} - \epsilon e^{10t}$ als Lösung. Mit wachsendem t entfernt sich $\tilde{y}$ immer mehr von $y, ||y - \tilde{y}|| = |\epsilon||e^{10t}|$. Dies ist ein Problem der Anfangswertaufgabe und unabhängig vom Verfahren. Dies wird manchmal als inhärente Instabilität bezeichnet. Um dies in den Griff zu bekommen, muß man mit Verfahren hoher Konsistenzordnung und mit hoher Rechengenauigkeit arbeiten, um dies unabwendbare Anwachsen des Fehlers so gering wie möglich zu halten.

Die 0-Stabilität (Wurzelbedingung) eines Verfahrens sichert uns zusammen mit der Konsistenz die Konvergenz eines Mehrschrittverfahrens für $h \to 0$. Wegen der Rundungsfehlereinflüsse kann man aber h nicht beliebig verkleinern. Wir betrachten daher das Verhalten konsistenter, 0-stabiler Verfahren für eine feste Schrittweite $h > 0$.

Sei nun

$$\sum_{r=0}^{k} \alpha_r u_{j+r} = h \sum_{r=0}^{k} \beta_r f_{j+r}$$

ein 0-stabiles, konsistentes Verfahren. Da wir Rundungsfehler beim Rechnen machen, erhalten wir statt der u_{j+r} Werte $\tilde{u}_{j+r}$ mit

$$\sum_{r=0}^{k} \alpha_r \tilde{u}_{j+r} = h \sum_{r=0}^{k} \beta_r f_{j+r} + h\epsilon_{j+k}. \tag{2.9.1}$$

Es sei wieder $e_j = y(t_j) - \tilde{u}_j$ für $j = 0, 1, 2, \ldots$. Beachte, daß für $j = 0, \ldots, k-1$ dies gerade die Startfehler sind, die aus dem Einschrittverfahren kommen. Da gilt

$$\sum_{r=0}^{k} \alpha_r y(t_{j+r}) = h \sum_{r=0}^{k} \beta_r f(t_{j+r}, y(t_{j+r})) + h\tau_{j+k}$$

(mit $\tau_{j+k} = \tau(t_{j+k}, y, h, f)$), folgt mit (2.9.1)

$$\begin{aligned}\sum_{r=0}^{k} \alpha_r e_{j+r} &= h \sum_{r=0}^{k} \beta_r (f(t_{j+r}, y(t_{j+r})) - f(t_{j+r}, \tilde{u}_{j+r})) + h(\tau_{j+k} - \epsilon_{j+k}) \\ &= h \sum_{r=0}^{k} \beta_r \frac{\partial f}{\partial y}(t_{j+r}, \eta_{j+r}) e_{j+r} + h(\tau_{j+k} - \epsilon_{j+k})\end{aligned}$$

wobei η_{j+r} eine geeignete Zwischenstelle sei. Als vereinfachende Annahme wollen wir vereinbaren, daß gelte

$$\begin{aligned}h(\tau_{j+k} - \epsilon_{j+r}) &= \Phi = konstant \\ \frac{\partial f}{\partial y} &= \lambda = konstant.\end{aligned}$$

Damit erhält man die Differenzengleichung

$$\sum_{r=0}^{k} (\alpha_r - h\beta_r\lambda) e_{j+r} = \Phi. \tag{2.9.2}$$

Eine spezielle Lösung von (2.9.2) ist

$$\hat{e}_j = -\frac{1}{h\lambda \sum\limits_{r=0}^{k} \beta_r} \Phi \quad \text{für alle } j, \text{ denn } \sum_{r=0}^{k} \alpha_r = 0.$$

Dann ist die allgemeine Lösung von (2.9.2)

$$e_j = \eta_j - \underbrace{\frac{1}{h\lambda \sum\limits_{r=0}^{k} \beta_r}}_{konstant} \Phi \tag{2.9.3}$$

wobei $(\eta_j)_{j \in \mathbb{N}}$ die Lösungen der homogenen Differenzengleichung

$$\sum_{r=0}^{k} (\alpha_r - h\lambda\beta_r) \eta_{j+r} = 0 \tag{2.9.4}$$

sind. Die η_j bestimmen das Wachstumsverhalten der Fehler e_j. Daher untersuchen wir jetzt die Lösungen von (2.9.4).

Bemerkung 2.9.1 *i) Die Annahme $\frac{\partial f}{\partial y} = \lambda$ kann man als lokale Linearisierung auffassen. Denn $y' = f(t,y) = f(t,y_0) + \frac{\partial f}{\partial y}(t,y_0)(y-y_0) + \ldots$. Wenn man vereinfachend $\frac{\partial f}{\partial y}(t,y_0)$ als konstant annimmt (lokal eine gute Approximation), so betrachtet man die angenäherte Anfangswertaufgabe $y' = f(t,y_0) + \frac{\partial f}{\partial y}(t,y_0)(y-y_0)$. Man kann diese Einschränkung aber auch verstehen als Untersuchung der Testaufgabe $y' = \lambda y$.*

ii) Falls man ein System von Differentialgleichungen hat, so ist die entsprechende Linearisierung offenbar $y' = Ay + b$, wobei A die $n \times n$ Jacobi-/Funktionalmatrix $\frac{\partial f}{\partial y}(t,y_0)$ ist. Falls A diagonalisierbar ist, d.h. falls eine Matrix T mit $T^{-1}AT = diag(\lambda_1, \ldots, \lambda_n)$ existiert, so gilt für die transformierte Differentialgleichung ($x = T^{-1}y$, $c = T^{-1}b$)

$$x' = T^{-1}\,y' = diag(\lambda_1, \ldots, \lambda_n)x + c$$

und man hat n unabhängige Differentialgleichungen vom obigen Typ. Untersuchungen für den eindimensionalen Fall geben also auch Auskunft über das Verhalten von Systemen (allerdings sind die λ_j dabei i.a. komplex).

Nun betrachten wir den Fall $n = 1$ und untersuchen η_j. Man nennt

$$\prod(z) = \sum_{r=0}^{k} (\alpha_r - h\lambda\beta_r) z^r = \rho(z) - h\lambda\sigma(z)$$

das Stabilitätspolynom des Mehrschrittverfahrens.

Wir setzen im folgenden 0-Stabilität und Konsistenz voraus. Wir hatten bereits vorher gesehen, daß ohne 0-Stabilität ein Mehrschrittverfahren divergieren kann. Aber auch mit 0-Stabilität kann es Probleme geben.

Beispiel 2.9.1 *Betrachte für $y' = \lambda y, y(0) = 1$ das Verfahren $u_{j+2} = u_j + 2h\lambda u_{j+1}$ (Mittelpunktsregel). Das ist ein 0-stabiles Verfahren der Ordnung 2, denn $\rho(z) = z^2 - 1 = (z+1)(z-1)$ und mit $\alpha_0 = -1, \alpha_1 = 0, \alpha_2 = 1, \beta_0 = 0, \beta_1 = 2, \beta_2 = 0$*

$$\begin{array}{ll} l = 0 & \sum \alpha_i = \rho(1) = 0 \\ l = 1 & \alpha_1 + 2\alpha_2 - \beta_0 - \beta_1 - \beta_2 = 0 \\ l = 2 & 4\alpha_2 - 2\beta_1 - 4\beta_2 = 0. \end{array}$$

Die u_j sind Lösungen der Differenzengleichung

$$u_{j+2} - 2h\lambda u_{j+1} - u_j = 0.$$

Mit den exakten Anfangswerten $u_0 = 1, u_1 = e^{h\lambda}$ erhalten wir mit Lemma 2.8.1

$$u_j = c_1 z_1^j + c_2 z_2^j$$

wobei $z_{1/2} = h\lambda \pm \sqrt{1 + h^2\lambda^2}$ die Nullstellen von $\prod(z) = z^2 - 2h\lambda z - 1$ sind und c_1, c_2 aus $u_0 = 1, u_1 = e^{\lambda h}$ bestimmt werden:

$$c_1 = \frac{z_2 - e^{\lambda h}}{z_2 - z_1} \quad c_2 = \frac{e^{\lambda h} - z_1}{z_2 - z_1}.$$

Falls man wieder z_1, z_2, c_1 und c_2 nach h entwickelt, so erhält man für $u_h(t)$, der Näherung an der Stelle t mit $h = \frac{t}{j}$

$$u_h(t) = \underbrace{e^{\lambda t}(1 - \frac{\lambda^2}{2}h^2 + \frac{\lambda^3}{12}h^3)}_{\text{Entwicklung von } c_1 z_1^j} + \underbrace{(-1)^{t/h}\, e^{-\lambda t}\frac{\lambda^3}{12}h^3}_{\text{Entwicklung von } c_2 z_2^j} + O(h^4).$$

Falls nun $\lambda < 0$, so wächst der zweite Term für wachsende t oszillierend an und überdeckt den abfallenden Teil $y(t) = e^{\lambda t}$. Beachte: Dies ist kein Widerspruch zur Konvergenz. Dort betrachtet man t fest und $h \to 0$, hier betrachten wir h fest und $t \to \infty$. Der Grund für das hier auftretende Problem ist die parasitäre Wurzel -1 von ρ. Für die das Fehlerwachstum bestimmenden Nullstellen von $\prod$ gilt

$$\begin{aligned} z_1 &= 1 + h\lambda + O(h^2\lambda^2) \\ z_2 &= -1 + h\lambda + O(h^2\lambda^2) \end{aligned}$$

und für $\lambda < 0$ und h klein ist dann

$$|z_1| < 1 \quad \textit{aber} \quad |z_2| > 1.$$

Nach (2.9.3) , (2.9.4) und Lemma 2.8.1 ist

$$e_j = c_1 z_1^j + c_2 z_2^j - \frac{1}{h\lambda \sum_{r=0}^k \beta_r}\Phi.$$

Dabei gilt $|c_2 z_2^j| \to \infty$ für $j \to \infty$.

Dieses Beispiel motiviert die folgende Definition. Das lineare $k-$Schrittverfahren heißt <u>stark stabil</u>, falls $\rho(1) = 0$ gilt und alle $(k-1)$ parasitären Wurzeln vom Betrag < 1 sind. Dies soll beinhalten, daß 1 einfache Nullstelle ist. Es heißt <u>schwach stabil</u>, falls die Stabilitätsbedingung gilt und mehrere Nullstellen den Betrag 1 haben.

Beispiel 2.9.2 *Die Mittelpunktsregel und die Milne-Simpson-Verfahren ($\rho(z) = z^k - z^{k-2}$) sind also schwach stabil.*

Beachte: Diese Begriffe sind problemunabhängig, es kann jedoch sein, daß sie erst für sehr kleine h greifen. Es steht ja die Idee dahinter, daß die Wurzeln $z_j(h\lambda)$ von $\prod(z, h\lambda)$ für $h \to 0$ gegen die Wurzeln z_j von $\rho(z)$ streben. Insbesondere heißt dann die Wurzel $z_1(h\lambda)$ von $\prod(z, h\lambda)$ für die $\lim\limits_{h\to 0} z_1(h\lambda) = 1$ gilt, <u>Hauptwurzel von $\prod$</u>.

Die Konvergenz läßt sich für stark stabile Verfahren genauer beschreiben. Falls das Verfahren die Ordnung p hat, so gilt für die Hauptwurzel $z_1(h\lambda)$ von $\prod$

$$z_1(h\lambda) = e^{h\lambda} + O(h^{p+1}) \tag{2.9.5}$$

Dies sieht man so: Für festes t ist für $f(x,y) = \lambda y, y(0) = 1$

$$\begin{aligned} h\tau(t+kh,y,h,f) &= \sum_{r=0}^{k} \alpha_r e^{\lambda(t+rh)} - h\sum_{r=0}^{k} \beta_r \lambda e^{\lambda(t+rh)} \\ &= O(h^{p+1}). \end{aligned}$$

Also folgt

$$e^{\lambda t} \underbrace{\sum_{r=0}^{k} (\alpha_r e^{r\lambda h} - h\lambda\beta_r e^{r\lambda h})}_{\prod(e^{\lambda h}, h\lambda)} = O(h^{p+1})$$

und daher $\prod(e^{\lambda h}, h\lambda) = O(h^{p+1})$. Seien $z_1(\lambda h), \ldots, z_k(\lambda h)$ die Nullstellen von $\prod(z, h\lambda)$. Dann gilt

$$\prod(e^{\lambda h}, h\lambda) = (\alpha_k - h\lambda\beta_k)(e^{h\lambda} - z_1(h\lambda)) \ldots (e^{h\lambda} - z_k(h\lambda)). \tag{2.9.6}$$

Wegen $e^{\lambda h} \overset{h\to 0}{\longrightarrow} 1$ und starker Stabilität gilt

$$e^{\lambda h} - z_j(h\lambda) \overset{h\to 0}{\not\longrightarrow} 0 \quad \text{für} \quad j > 1.$$

D.h.

$$|(\underbrace{\alpha_k}_{\neq 0} - h\lambda\beta_k)(e^{h\lambda} - z_2(h\lambda)) \ldots (e^{h\lambda} - z_k(h\lambda))| > M$$

für kleine h. Daher folgt aus (2.9.6)

$$e^{h\lambda} - z_1(h\lambda) = O(h^{p+1})$$

und damit ergibt sich (2.9.5). Für genügend kleines h ist wegen der starken Stabilität

$$|z_1(h\lambda)| > |z_l(h\lambda)|$$

für alle $l = 2, \ldots, k$. Das Wachstum der Fehler richtet sich also nach dem Wachstum von $[z_1(h\lambda)]^j$. Nach (2.9.5) ist nun für $\lambda < 0$ und kleine $h > 0$ $\quad |z_1(h\lambda)| < 1$. In diesem Fall haben wir für den Fehler e_j sogar eine Fehlerdämpfung, da

$$|z_i(h\lambda)|^j \overset{j\to\infty}{\longrightarrow} 0 \quad \text{für alle} \quad i \in \{1, \ldots, k\}$$

Beachte: Dies sind die absoluten Fehler und nicht die relativen Fehler. Die Dämpfung der absoluten Fehler folgt damit nur dem Abklingen der Lösung $e^{\lambda t}$.

Das Verfahren heißt <u>absolut stabil</u> für $h\lambda$, falls $|z_j(\lambda h)| < 1$ für alle $j = 1, \ldots, k$, andernfalls heißt es <u>absolut instabil</u>. Der Bereich $D_\lambda = \{h\lambda |\ |z_i(h\lambda)| < 1$ für alle $i = 1, \ldots, k\}$ heißt <u>Bereich absoluter Stabilität</u>.

Bemerkung 2.9.2 *Wegen (2.9.5) ist jedes konsistente, 0-stabile Verfahren für alle hinreichend kleinen positiven λh absolut instabil. In diesem Fall wächst der Fehler, aber auch die Lösung. Der Begriff der absoluten Stabilität ist ein Konzept für $h\lambda$ negativ, d.h. für abklingende Lösungen $e^{\lambda t}$ mit $\lambda < 0$. Solche Anteile sind in vielen Differentialgleichungen zumindest lokal (in der Linearisierung $y' = Ay + f$) enthalten.*

Es stellt sich nun die Frage, ob es immer einen Bereich absoluter Stabilität gibt. Dazu schauen wir uns ein Beispiel an.

Beispiel 2.9.3 *Wir betrachten wieder (wie im letzten Beispiel) die Mittelpunktsregel*

$$u_{j+2} = u_j + 2h\lambda u_{j+1}$$

mit $\prod(z) = z^2 - 2h\lambda z - 1$. *Die Nullstellen von* $\prod$ *sind gegeben durch*

$$\begin{aligned} z_1 &= h\lambda + \sqrt{1+h^2\lambda^2} \\ z_2 &= h\lambda - \sqrt{1+h^2\lambda^2}. \end{aligned}$$

Für $h\lambda \leq 0$ *folgt* $|z_2| > 1$, *für* $h\lambda > 0$ *folgt* $|z_1| > 1$. *Also kann nie* $|z_1| < 1$ *und* $|z_2| < 1$ *gelten. Es gibt daher keinen Bereich absoluter Stabilität. Dasselbe folgt für das Milne-Simpson-Verfahren*

$$\begin{aligned} u_{j+2} &= u_j + \frac{h}{3}(f_{j+2} + 4f_{j+1} + f_j) \\ &= u_j + h\lambda(\frac{1}{3}u_{j+2} + \frac{4}{3}u_{j+1} + \frac{1}{3}u_j) \end{aligned}$$

mit $\prod(z) = (1 - \frac{h\lambda}{3})z^2 - \frac{4}{3}h\lambda z - (1 + \frac{1}{3}h\lambda)$ *und den Nullstellen*

$$z_{1/2} = \frac{2h\lambda \pm \sqrt{3(3+h^2\lambda^2)}}{3 - h\lambda}.$$

Bemerkung 2.9.3 *i) Man kann zeigen, daß "optimale" Verfahren, das sind solche mit* $p = k + 2$, *keine Bereiche absoluter Stabilität besitzen.*

ii) Alle stark stabilen Verfahren haben Bereiche absoluter Stabilität, da alle parasitären Nullstellen von $\rho(z)$ *betraglich* < 1 *sind, und* $lim_{h\to 0} \prod(z, h\lambda) = \rho(z)$ *für alle z gilt.*

iii) Falls alle parasitären Nullstellen in 0 liegen, erwartet man einen großen Bereich absoluter Stabilität, z.B. bei Adams-Moulton-Verfahren $u_{j+k} - u_{j+k-1} = h\sum \ldots$, *denn* $\rho(z) = z^k - z^{k-1}(z-1)$. *Bei impliziten Verfahren sind diese Bereiche zum Teil erheblich (bis zu 10 mal) größer als bei expliziten Verfahren (siehe Grigorieff.)*

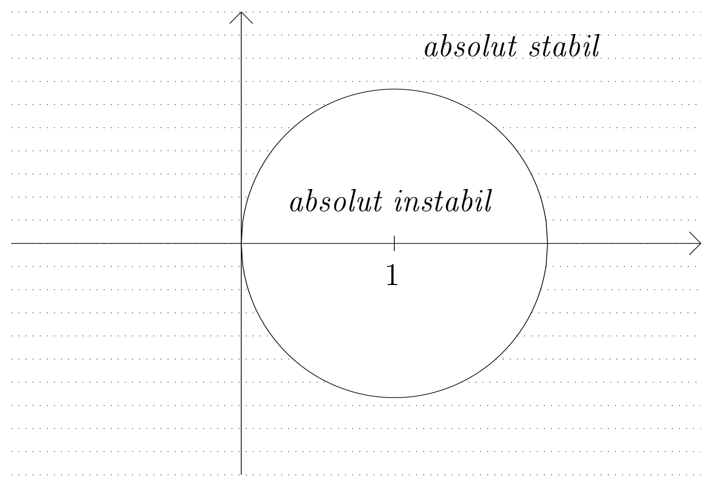
Beispiel 2.9.4 *Bereiche absoluter Stabilität beim impliziten und expliziten Euler-Verfahren*

1. *Implizites Euler-Verfahren*

$$\begin{aligned} u_{j+1} &= u_j + hf(t_{j+1}, u_{j+1}) \\ &= u_j + h\lambda u_{j+1} \quad (f\ddot{u}r\, f(t,y) = \lambda y) \\ \prod(z) &= (1 - h\lambda)z - 1 \\ z(h\lambda) &= \frac{1}{1 - h\lambda} \end{aligned}$$

Der Bereich absoluter Stabilität ist daher gegeben durch

$$D_\lambda = \{h\lambda | \quad |1 - h\lambda| > 1\}.$$

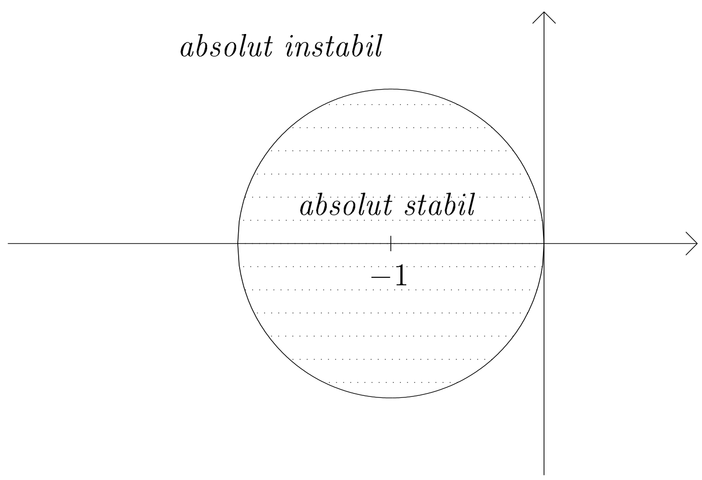


2. *Explizites Euler-Verfahren*

$$\begin{aligned} u_{j+1} &= u_j + h\lambda u_j \\ \prod(z) &= z - (1 + h\lambda) \\ z(h\lambda) &= 1 + h\lambda \end{aligned}$$

Der Bereich absoluter Stabilität ist daher gegeben durch

$$D_\lambda = \{h\lambda| \quad |1 + h\lambda| < 1\}.$$



Daß der Bereich der absoluten Stabilität für den impliziten Fall positive Werte $h\lambda$ enthält, ist kein Widerspruch zu Bemerkung 2.9.2. Dort wurde ausgesagt: Jedes konsistente 0-stabile Verfahren ist für alle hinreichend kleinen $h\lambda$ absolut instabil.

Bisher war unsere Blickrichtung auf $e^{\lambda t}$ mit λ negativ ausgerichtet. λ stand dabei für die Linearisierung $\frac{\partial f}{\partial y}$. Im allgemeinen ist $\frac{\partial f}{\partial y}$ nicht konstant und man wählt λ als geeignete Schranke für $\frac{\partial f}{\partial y}$ oder als typischen Wert, der über einem Teil des Integrationsbereichs stimmt. Falls λ verändert werden muß, so muß sich das maximal erlaubte h ändern, so daß $h\lambda$ noch im Bereich absoluter Stabiltät liegt.

Betrachten wir nun $\lambda > 0$. In diesem Fall sagt (2.9.3) und Bemerkung 2.9.2, daß die Fehler e_j wachsen. Da dies die Lösung der linearisierten Gleichung $y' = \lambda y + g(t)$ auch tut, ist das Anwachsen des Fehlers akzeptabel, ja sogar unvermeidbar, da Rundungsfehler stets in der Größenordnung macheps * Größe der Werte liegen. Der absolute Fehler sollte jedoch nicht wesentlich schneller als die Lösung wachsen.

Daher die folgende Definition. Das lineare Mehrschrittverfahren heißt für $h\lambda$ <u>relativ stabil</u>, falls $|z_j(h\lambda)| < |z_1(h\lambda)|$ für alle $j = 2, \ldots, k$ (z_1 Hauptwurzel), sonst heißt es <u>relativ instabil</u>. Für relativ stabile Verfahren wachsen die parasitären Zweige langsamer als der Hauptzweig, die Hauptwurzel darf größer als 1 sein. Den <u>Bereich relativer Stabilität</u> definiert man als

$$R_\lambda = \{h\lambda | \quad |z_j(h\lambda)| < |z_1(h\lambda)| \quad \text{für alle} \quad j > 1\}.$$

Der Bereich kann aus disjunkten Mengen bestehen.

Beispiel 2.9.5 *Wir betrachten wieder die Mittelpunktsregel*

$$u_{j+2} = u_j + 2h\lambda u_{j+1}$$

mit $\prod(z) = z^2 - 2h\lambda z - 1$ *und den Nullstellen* $z_1 = h\lambda + \sqrt{1 + h^2\lambda^2}$ *und* $z_2 = h\lambda - \sqrt{1 + h^2\lambda^2}$. *Für* $h\lambda > 0$ *ist* $|z_2| < |z_1|$. *Der Bereich relativer Stabilität ist daher die gesamte rechte Halbachse.*

Beispiel 2.9.6 *Betrachte das konsistente und* 0*-stabile Verfahren*

$$u_{j+2} - u_j = \frac{1}{2}h(f_{j+1} + 3f_j)$$

wieder für $f(t, y) = \lambda y$. *Hier ist* $\prod(z, h\lambda) = z^2 - \frac{1}{2}h\lambda z - (1 + \frac{3}{2}h\lambda)$ *mit den Wurzeln*

$$\begin{aligned} z_1 &= \frac{1}{4}h\lambda + \sqrt{\frac{1}{16}h^2\lambda^2 + 1 + \frac{3}{2}h\lambda} \\ z_2 &= \frac{1}{4}h\lambda - \sqrt{\frac{1}{16}h^2\lambda^2 + 1 + \frac{3}{2}h\lambda}. \end{aligned}$$

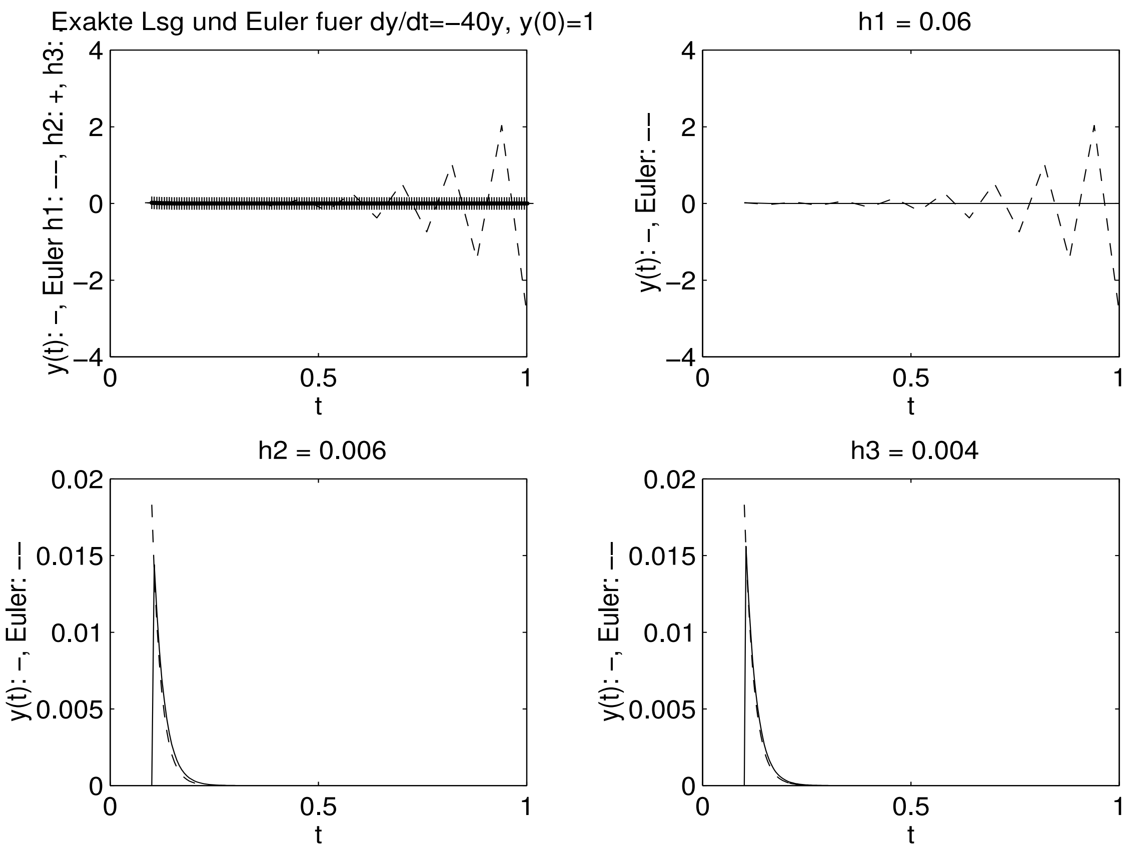
Entwickeln wir wie üblich nach h ($\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{4}\frac{x^2}{2} + \frac{3}{8}\frac{x^3}{6} + \ldots$) *ergibt sich*

$$\begin{aligned} z_1 &= 1 + h\lambda + O(h^2) \\ z_2 &= -1 - \frac{1}{2}h\lambda + O(h^2). \end{aligned}$$

Intervall absoluter Stabilität ist daher $]\alpha, 0[$ *mit geeignetem* $\alpha < 0$, *Interval relativer Stabilität ist* $]0, \beta[$ *mit geeignetem* $\beta > 0$. *Nach dem absoluten Stabilitätskriterium würden wir dieses Verfahren also nur für* $\frac{\partial f}{\partial y} < 0$ *verwenden. Dann nimmt der absolute Fehler ab, aber nicht so schnell wie die Lösung, denn der Fehler nimmt ab wie* $|z_2^j| = |1 + \frac{1}{2}h\lambda - O(h^2)|^j$, *die Lösung* $e^{\lambda t_j} = e^{\lambda(t_0+jh)} = e^{\lambda t_0}(e^{\lambda h})^j = e^{\lambda t_0}(1 + h\lambda + O(h^2))^j$ *nimmt ab wie* $|z_1|^j = |1 + h\lambda + O(h^2)|^j$. *Nach dem relativen Stabilitätsbegriff würden wir dieses Verfahren nur für* $\frac{\partial f}{\partial y} > 0$ *verwenden und einen Fehler akzeptieren, der wächst, aber nicht schneller als die Lösung. Hier scheint also das Konzept der relativen Stabilität günstiger zu sein.*

Beide Konzepte haben Vor- und Nachteile. Die Aussagen bei absoluter Stabilität sind tendenziell zu optimistisch, die der relativen Stabilität oft schwer herzuleiten und gelegentlich pessimistisch. Die Menge der "guten" $h\lambda$ ist in Wahrheit größer als der Stabilitätsbereich angibt. Zusammengefassend kann man sagen: Für kleine $h\lambda > 0$ sind konsistente, 0-stabile Verfahren absolut instabil ($z_1(h\lambda) = e^{\lambda h} + O(h^{p+1})$). Ein relatives Konzept ist daher angebracht. Für $h\lambda < 0$ sind relativ stabile Mehrschrittverfahren für kleine $h\lambda$ auch absolut stabil (wegen $z_1(h\lambda) = e^{\lambda h} + O(h^{p+1})$). Das relative Konzept ist also i.a. vorzuziehen. Das absolute Konzept ist jedoch für steife Differentialgleichungen wichtig (siehe nächsten Abschnitt).

Beispiel 2.9.7 *Wir wollen die Anfangswertaufgabe* $y' = -40y, y(0) = 1$ *im Intervall* $[0.1, 1]$ *mit dem Eulerverfahren für Schrittweiten* $0.06, 0.004, 0.006$ *lösen. Man sieht, daß bei* 0.06 *die numerische Lösung stark oszilliert, während die exakte Lösung fast* 0 *ist. Es ist* $D_{-40} = \{-40h| \quad |1 - 40h| < 1\}$, *also ist* $-40h \in D_{-40}$ *für* $h < 0.05$!



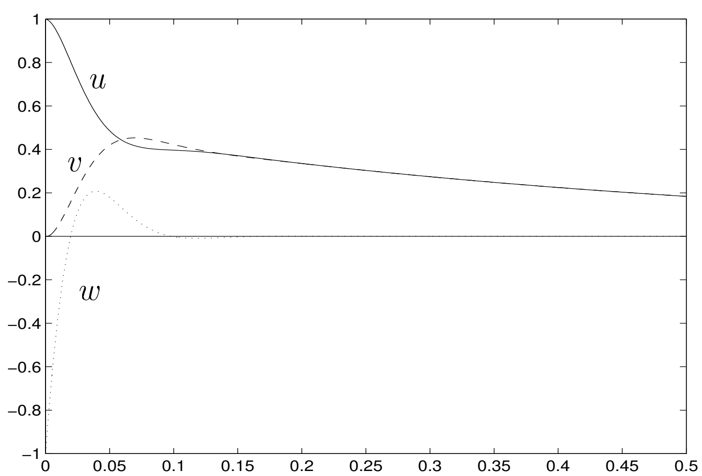
2.10 Steife Differentialgleichungen

Probleme der Kontrolltheorie und des Ablaufs chemischer Prozesse zeigen häufig ein Phänomen, wie es im folgenden Beispiel auftritt.

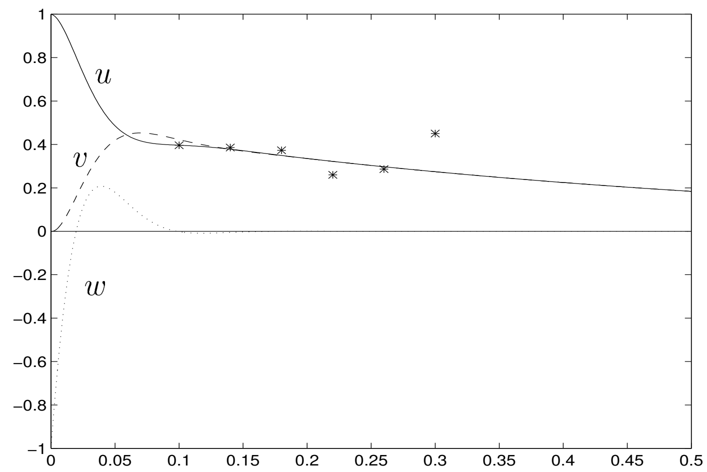
Beispiel 2.10.1 *(siehe Lambert, Computational Methods in Ordinary Differential Equations, 1973, Seite 230).*

Sei $y' = Ay, y(0) = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$ *mit* $A = \begin{bmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{bmatrix}$. *Die Eigenwerte von* A *lauten* $-2, -40+40i, -40-40i$. *Die exakte Lösung des Differentialgleichungssystems ist gegeben durch* $y = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$ *mit*

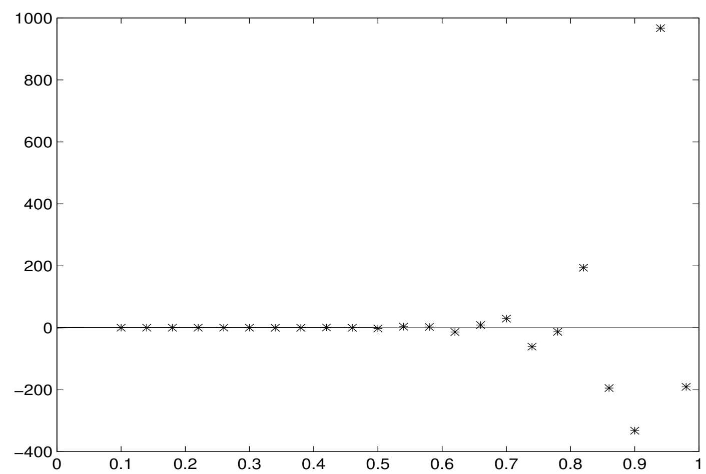
$$\begin{aligned} u(t) &= \frac{1}{2}e^{-2t} + \frac{1}{2}e^{-40t}(\cos 40t + \sin 40t) \\ v(t) &= \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-40t}(\cos 40t + \sin 40t) \\ w(t) &= -e^{-40t}(\cos 40t - \sin 40t) \end{aligned}$$



Ab 0.1 *gibt es keine Schwankungen mehr im Lösungsverlauf. Man möchte daher eine nicht zu kleine Schrittweite bei der numerischen Lösung benutzen. Betrachten wir das Eulerverfahren mit der Schrittweite* $h = 0.04$, *so erhalten wir für* $u(t)$ *auf* $[0.1, 0.3]$ *die Näherungen * in der folgenden Grafik*



Hier ist der letzte Näherungswert nicht mehr akzeptabel. Viel schlimmer sieht die Situation aus, wenn wir das größere Intervall $[0.1, 1]$ *betrachten. (Beachte die geänderte Skalierung!)*



Diese Näherung ist offensichtlich völlig sinnlos.

Beispiel 2.10.2 *Wenn man dagegen das Problem mit*

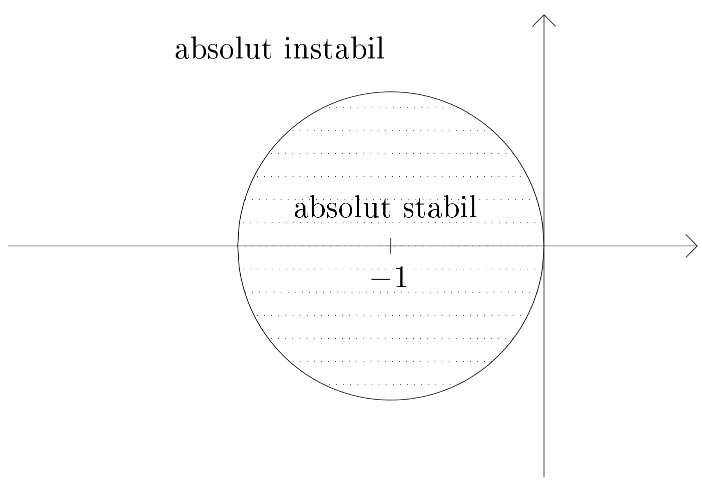
$$A = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; y(0.1) = \begin{bmatrix} \frac{1}{2}e^{-0.2} \\ \frac{1}{2}e^{-0.2} \\ 0 \end{bmatrix}$$

betrachtet, so ist die Lösung gegeben durch

$$y = \begin{bmatrix} \frac{1}{2}e^{-2t} \\ \frac{1}{2}e^{-2t} \\ 0 \end{bmatrix}.$$

Dies ist für $t > 0.1$ nicht von der Lösung aus dem ersten Beispiel zu unterscheiden. Wendet man hierauf das Euler-Verfahren mit Schrittweite $h = 0.04$ an, so erhält man eine befriedigende Näherungslösung.

Der Grund des so unterschiedlichen Verhaltens liegt in den verschiedenen Bereichen der absoluten Stabilität. Der Bereich absoluter Stabilität für das explizite Euler-Verfahren ist, wie wir schon gesehen haben



Damit für das erste Beispiel $h\lambda$ für alle Eigenwerte von A in diesem Bereich liegen, muß $h < 0.025$ sein. Verantwortlich für diese Einschränkung sind die Eigenwerte $-40 \pm 40i$, obwohl ihre Anteile an der Lösung ab 0.1 fast verschwinden. Wählt man h größer, so wächst der absolute Fehler immer stärker. Für das zweite Beispiel sind die Eigenwerte $-2, -2, 0$ und $h < 1$ sichert absolute Stabilität.

Wir betrachten nun allgemein das $n \times n$ System

$$y' = Ay + \phi(t)$$

(z.B. aus der Linearisierung eines Differentialgleichungssystems). A sei diagonalisierbar und $Re(\lambda_i) < 0$ für alle Eigenwerte λ_i von A. Dann hat die Lösung die Form

$$y(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t} v^{(i)} + \psi(t),$$

wobei $v^{(i)}$ der Eigenvektor von A zum Eigenwert λ_i ist. Wegen $Re(\lambda_i) < 0$ geht der erste Teil für $t \to \infty$ gegen 0. Für $t \to \infty$ strebt $y(t)$ also gegen $\psi(t)$. $\psi(t)$ heißt stationäre Lösung.

Bei Aufgaben dieser Art ist es das Ziel, diese stationäre Lösung zu bestimmen. Wir müssen dazu die numerische Lösung solange verfolgen, bis der am langsamsten abklingende Term $c_i e^{\lambda_i t} v^{(i)}$ vernachlässigbar klein ist. Falls $|Re(\lambda_i)|$ sehr klein ist, ist dieser Integrationsweg sehr lang. Falls außerdem ein $|Re(\lambda_j)|$ sehr groß ist, so muß für diesen langen Weg eine sehr kleine Schrittweite h verwendet werden, damit $h\lambda_j$ im Bereich absoluter Stabilität liegt. Falls also

$$|Re(\lambda_1)| \geq \ldots \geq |Re(\lambda_n)| \quad \text{und} \quad |Re(\lambda_1)| \gg |Re(\lambda_n)|,$$

so muß sehr lange mit sehr kleiner Schrittweite gerechnet werden. Das hier beschriebene Problem bezeichnet man als das Problem der Steifheit (stiffness, auch Steifigkeit bei manchen Autoren).

Ein System $y' = Ay + \phi(t)$ heißt steif, falls

i) $Re(\lambda_i) < 0$ für alle Eigenwerte λ_i von A

ii) $\max\limits_{i=1,\ldots,n} |Re(\lambda_i)| \gg \min\limits_{i=1,\ldots,n} |Re(\lambda_i)|$.

Der Quotient $\frac{\max |Re(\lambda_i)|}{\min |Re(\lambda_i)|}$ heißt Steifigkeitsquotient.

Allgemeiner redet man für

$$y' = f(t, y) = f(t, y_0) + \frac{\partial f}{\partial y}(t, y_0)(y - y_0) + \ldots$$

von der Steifheit (auf dem Intervall I), falls die Eigenwerte von $\frac{\partial f}{\partial y}$ $i)$ und $ii)$ auf I erfüllen.

Im ersten Beispiel dieses Abschnitts ist der Steifigkeitsquotient 20, in der Praxis ist 10^6 nicht unüblich. Das Problem ist dann, λh in den Bereich der absoluten Stabilität zu bekommen.

Ein Verfahren heißt A-stabil, falls der Bereich absoluter Stabilität die gesamte linke Halbebene umfaßt. Bei einem A-stabilen Verfahren entfallen also die Schwierigkeiten mit den steifen Systemen. Aber

Satz 2.10.1 *Dahlquist*

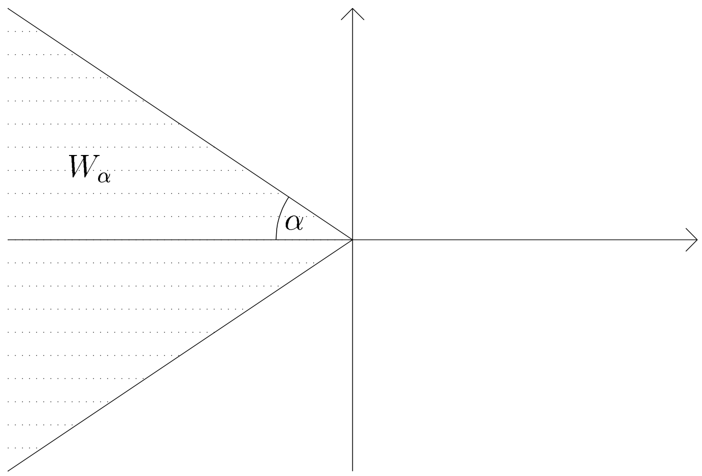
i) Ein explizites lineares Mehrschrittverfahren ist niemals A-stabil.

ii) Die Ordnung eines A-stabilen impliziten linearen Mehrschrittverfahrens ist ≤ 2.

iii) Das A-stabile Verfahren 2. Ordnung mit der kleinsten Fehlerkonstante ist die Trapezregel $u_{j+1} - u_j = \frac{h}{2}(f_{j+1} - f_j)$.

Beweis: siehe z.B. Hairer, Wanner.

Dieser Satz motiviert die folgenden Abschwächungen. Ein Verfahren heißt $A(\alpha)$-stabil, $\alpha \in]0, \pi/2[$, falls der Bereich absoluter Stabilität $W_\alpha = \{h\lambda | -\alpha < \pi - arg(h\lambda) < \alpha\}$ umfaßt. Es heißt $A(0)$-stabil, falls es $A(\alpha)$-stabil ist für hinreichend kleines $\alpha \in]0, \frac{\pi}{2}[$.



Beachte: Für λ mit $Re(\lambda) < 0$ liegt $h\lambda$ entweder für alle in h in W_α oder für alle h außerhalb. Zum Beispiel ist $h\lambda$ für $\lambda = iz, z \in \mathbb{R}$, stets außerhalb von W_α, für $\lambda = z, z \in \mathbb{R}$ ist $h\lambda$ stets innerhalb von W_α. Falls wir wissen, daß alle Eigenwerte eines steifen Systems in W_α liegen, so kann ein $A(\alpha)$-stabiles Verfahren ohne Stabilitätseinschränkungen (an die Schrittweite) verwendet werden. Falls alle Eigenwerte reell sind, z.B. bei symmetrischen Jacobimatrizen $\frac{\partial f}{\partial y}$, so kann jedes $A(0)$-stabiles Verfahren benutzt werden. Man kann zeigen:

Satz 2.10.2 *Widlund*

i) Ein explizites lineares Mehrschrittverfahren ist niemals $A(0)$-stabil.

ii) Die Trapezregel ist das einzige $A(0)$-stabile k-Schrittverfahren mit Ordnung $\geq k+1$.

iii) Für alle $\alpha \in [0, \frac{\pi}{2}[$ gibt es ein $A(\alpha)$- stabiles lineares k-Schrittverfahren der Ordnung p mit $k = p = 3$ und $k = p = 4$.

Bemerkung 2.10.1 *Explizite Runge-Kutta-Verfahren haben i.a. einen sehr kleinen Bereich absoluter Stabilität. Implizite Runge-Kutta-Methoden sind A-stabil.*

Beispiel 2.10.3 *Implizites Runge-Kutta-Verfahren mit $m = 1$ für $y' = \lambda y$.*

Wir betrachten das Verfahren

$$u_{j+1} = u_j + k_1$$

mit

$$k_1 = hf(t_j + \frac{1}{2}h, u_j + \frac{1}{2}k_1).$$

In diesem Fall also

$$k_1 = \lambda h(u_j + \frac{1}{2}k_1),$$

so daß gilt

$$(1 - \frac{1}{2}h\lambda)k_1 = h\lambda u_j.$$

Damit ergibt sich

$$u_{j+1} = u_j + \frac{h\lambda}{1 - \frac{1}{2}h\lambda}u_j = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}u_j$$

und

$$\prod(z, h\lambda) = z - \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}.$$

Falls $\lambda < 0$, so ist die Nullstelle stets betraglich < 1. Analog für $m = 2$:

$$u_{j+2} = \frac{1 + \frac{1}{2}h\lambda + \frac{1}{12}h^2\lambda^2}{1 - \frac{1}{2}h\lambda + \frac{1}{12}h^2\lambda^2}\, u_j.$$

Allgemein gilt: Bei expliziten Runge-Kutta-Verfahren sind die Koeffizienten in $\prod(z, h\lambda)$ Polynome in $h\lambda$ (Taylorpolynom von $e^{h\lambda}$). Bei impliziten Runge-Kutta-Verfahren sind sie rationale Funktionen in $h\lambda$ (Padé-Approximation von $e^{h\lambda}$). Diese sind vom Betrag < 1 für $Re(\lambda) < 0$, also sind die Verfahren A-stabil.

Beispiel 2.10.4 *Explizites Runge-Kutta-Verfahren mit $m = 2$ für $y' = \lambda y$.*

Wir betrachten das modifizierte Euler-Verfahren

$$\begin{aligned} u_{j+1} - u_j &= hf(t_j + \frac{1}{2}h, u_j + \frac{1}{2}hf(t_j, u_j)) \\ &= h\lambda(1 + \frac{1}{2}h\lambda)u_j \\ &= (h\lambda + \frac{h^2\lambda^2}{2})u_j. \end{aligned}$$

Damit ergibt sich

$$\prod(z, h\lambda) = z - (1 + h\lambda + \frac{h^2\lambda^2}{2}).$$

Das Intervall absoluter Stabilität ist daher $]-2, 0[$.

Mehr über steife Differentialgleichungen findet man in Hairer, Wanner.

Kapitel 3

Numerische Lösung von Randwertaufgaben

3.1 Verschiedene Randwertprobleme

Allgemeiner als Anfangswertaufgaben sind sogenannte Randwertprobleme. Es gibt sehr vielfältige Probleme, die auf Randwertaufgaben zurückgeführt werden können. Wir stellen daher zunächst einige Typen von Randwertaufgaben vor. Viele der anschließend diskutierten Lösungsmethoden lassen sich leicht auf partielle Differentialgleichungen übertragen (mehr dazu in Numerik II). Die theoretischen Grundlagen für Randwertaufgaben sind sehr viel komplizierter als für Anfangswertaufgaben, daher verzichten wir hier im wesentlichen auf Theorie. Wir wollen hier Randwertaufgaben der folgenden Form behandeln: Finde y auf $[a, b] \subseteq \mathbb{R}$ mit

$$y' = f(t, y), \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad f(t, y) = \begin{bmatrix} f_1(t, y_1, \ldots, y_n) \\ \vdots \\ f_n(t, y_1, \ldots, y_n) \end{bmatrix} \tag{3.1.1}$$

mit den Randbedingungen

$$r(y(a), y(b)) = 0, \quad r(u, v) = \begin{bmatrix} r_1(u_1, \ldots, u_n, v_1, \ldots, v_n) \\ \vdots \\ r_n(u_1, \ldots, u_n, v_1, \ldots, v_n) \end{bmatrix}. \tag{3.1.2}$$

In vielen Fällen haben wir <u>lineare Randbedingungen</u>

$$Ay(a) + By(b) = c \quad \text{mit} \quad A, B \in \mathbb{R}^{n \times n}, c \in \mathbb{R}^n,$$

oder sogar <u>separierte Randbedingungen</u> ($A = \binom{A_1}{0}, B = \binom{0}{B_2}, c = \binom{c_1}{c_2}$)

$$A_1 y(a) = c_1, \quad B_2 y(b) = c_2 \quad \text{mit} \quad A_1 \in \mathbb{R}^{m \times n}, B_2 \in \mathbb{R}^{(n-m) \times n}.$$

(So lassen sich Anfangswertprobleme als spezielle Randwertprobleme auffassen ($A_1 = I, B_2 = 0, c = x_0$). In der Praxis treten jedoch auch Mehrpunktbedingungen wie

$$y(\frac{a+b}{2}) = \frac{1}{2}(y(a) + y(b))$$

oder Integralbedingungen wie

$$\int_a^b y(x)dx = 1$$

auf. Diese beiden Probleme behandeln wir hier nicht.

In dem Spezialfall, daß f affin linear und die Randbedingungen linear in $y(a), y(b)$ sind, nennt man die Randwertaufgabe linear. Ein weiterer Spezialfall sind periodische Randbedingungen

$$y(a) = y(b), \quad y'(a) = y'(b)$$

oder

$$y(a) = -y(b), \quad y'(a) = -y'(b).$$

In vielen Fällen erhält man dann auch Lösungen, die periodisch in $[a, b]$ sind.

Wie üblich können auch Differentialgleichungen höherer Ordnung durch Einführung neuer Variablen auf Probleme der Form (3.1.1), (3.1.2) zurückgeführt werden. Um die Schwierigkeiten mit der Theorie zu analysieren, betrachten wir das folgende Beispiel:

Beispiel 3.1.1 *Die Differentialgleichung*

$$w'' + w = 0$$

für die reelle Funktion $w : \mathbb{R} \to \mathbb{R}$ *läßt sich mit den Abkürzungen* $y_1(t) := w(t), y_2(t) := w'(t)$ *in der Form*

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} y_2 \\ -y_1 \end{bmatrix}$$

schreiben. Sie besitzt die allgemeine Lösung

$$w(t) = c_1 \cdot \sin t + c_2 \cdot \cos t \quad c_1, c_2 \quad \textit{beliebig.}$$

Die spezielle Lösung $w(t) := \sin t$ *ist die einzige Lösung, die den Randbedingungen*

$$w(0) = 0, \quad w(\pi/2) = 1$$

genügt. Sämtliche Funktionen $w(t) := c_1 \cdot \sin t, c_1$ *beliebig, erfüllen die Randbedingungen*

$$w(0) = 0, \quad w(\pi) = 0,$$

während es keine Lösung $w(t)$ *gibt, die den Randbedingungen*

$$w(0) = 0, \quad w(\pi) = 1$$

genügt.

Das vorstehende Beispiel zeigt, daß es keinen ähnlich allgemeinen Satz für die Existenz und Eindeutigkeit von Lösungen von Randwertproblemen geben wird wie bei den Anfangswertproblemen.

Andere Formen von Differentialgleichungsaufgaben lassen sich auf Randwertaufgaben zurückführen. Betrachte z.B. das Eigenwertproblem

$$y' = f(t, y, \lambda) \quad \text{mit} \quad r(y(a), y(b), \lambda) = 0. \tag{3.1.3}$$

Aufgabe ist es, die Eigenwerte λ zu bestimmen, für die die Lösung für dieses Problem eindeutig ist. Man macht den Ansatz

$$y_{n+1}(t) = \lambda, \quad y'_{n+1}(t) = 0.$$

Dann ist (3.1.3) äquivalent zu

$$\overline{y}' = \overline{f}(t, \overline{y}), \quad \overline{r}(\overline{y}(a), \overline{y}(b)) = 0$$

mit

$$\overline{y} = \begin{bmatrix} y \\ y_{n+1} \end{bmatrix}, \quad \overline{f}(t, \overline{y}) = \begin{bmatrix} f(t, y, y_{n+1}) \\ 0 \end{bmatrix}$$

und

$$\overline{r}(u_1, \ldots, u_n, u_{n+1}, v_1, \ldots, v_n, v_{n+1}) = r(u_1, \ldots, u_n, v_1, \ldots, v_n, v_{n+1}).$$

Ein anderes wichtiges Anwendungsproblem sind Randwertprobleme mit freiem Rand. Bei diesen Problemen ist lediglich eine Randabszisse a gegeben und es ist b so zu bestimmen, daß das System von n gewöhnlichen Differentialgleichungen

$$y' = f(t, y)$$

eine Lösung y besitzt, die $n+1$ Randbedingungen erfüllt

$$r(y(a), y(b)) = 0, \quad r(u, v) = \begin{bmatrix} r_1(u, v) \\ \vdots \\ r_{n+1}(u, v) \end{bmatrix}.$$

Beispiele hierfür sind Schmelzvorgänge oder Diffusionsvorgänge in porösen Medien. Man macht zunächst den Ansatz

$$z_{n+1} = b - a$$

und führt eine neue Variable x ein durch

$$t - a = x \cdot z_{n+1}, \quad 0 \leq x \leq 1.$$

Dann gilt

$$z'_{n+1} = \frac{dz_{n+1}}{dx} = 0.$$

Setzt man $z(x) = y(a + xz_{n+1})$, so erhalten wir

$$z'(x) = \frac{d}{dx} z = \frac{d}{dt} y(a + xz_{n+1}) z_{n+1} = f(a + xz_{n+1}, z(t)) z_{n+1}$$

Die freie Randwertaufgabe ist daher äquivalent zu

$$\begin{bmatrix} z_1 \\ \vdots \\ z_n \\ z_{n+1} \end{bmatrix}' = \begin{bmatrix} z_{n+1} f_1(a + x z_{n+1}, z_1, \ldots, z_n) \\ \vdots \\ z_{n+1} f_n(a + x z_{n+1}, z_1, \ldots, z_n) \\ 0 \end{bmatrix}$$

mit den Randbedingungen

$$r_i(z_1(0), \ldots, z_n(0), z_1(1), \ldots, z_n(1)) = 0 \quad \text{für} \quad i = 1, \ldots, n+1.$$

Es gibt verschiedene theoretische Methoden zur Lösung von Randwertaufgaben, z.B. mittels Green'scher Funktion oder über die adjungierte Gleichung. Diese Methoden lassen sich oft mit numerischer Methoden kombinieren.

3.2 Das einfache Schießverfahren

Die einfachste Methode zur Lösung von Randwertaufgaben ist die Zurückführung auf Anfangswertaufgaben. Der erste und einleuchtendste Ansatz ist dabei das einfache Schießverfahren. Wir betrachten zuerst den eindimensionalen Fall. Aufgabe ist die Lösung von

$$y'' = f(t, y, y'), \quad y(a) = \alpha, \quad y(b) = \beta \tag{3.2.1}$$

zu bestimmen. Betrachte dazu das Anfangswertproblem

$$y'' = f(t, y, y'), \quad y(a) = \alpha, \quad y'(a) = s \tag{3.2.2}$$

und bestimme ein $\hat{s}$ so, daß für die Lösung $y(t, s)$ gilt

$$y(b, \hat{s}) = \beta.$$

Mit anderen Worten: wir wollen eine Nullstelle $\hat{s}$ von

$$F(s) = y(b, s) - \beta$$

finden. Für jedes s läßt sich $F(s)$ durch Lösung der Anfangswertaufgabe (3.2.2) bestimmen. Zur Bestimmung der Nullstelle $\hat{s}$ von $F(s)$ kann man im Prinzip alle im Mathematischen Praktikum besprochenen Methoden zur Lösung nichtlinearer Gleichungen benutzen. Im allgemeinen verwendet man aber das Newton-Verfahren, da $y(t, s)$ und damit $F(s)$ i.a. stetig differenzierbare Funktionen von s sind. Dazu hat man ausgehend von einem Startwert $s^{(0)}$ iterativ Werte $s^{(i)}$ nach der Vorschrift

$$s^{(i+1)} = s^{(i)} - \frac{F(s^{(i)})}{F'(s^{(i)})} \tag{3.2.3}$$

zu berechnen. $y(b, s^{(i)})$ und damit $F(s^{(i)})$ kann man durch Lösung des Anfangswertproblems (3.2.2) bestimmen. Den Wert der Ableitung von F

$$F'(s) = \frac{\partial}{\partial s} y(b, s),$$

für $s = s^{(i)}$ erhält man z.B. durch Behandlung eines weiteren Anfangwertproblems: Man kann zeigen, daß für die Funktion $v(x) :\equiv v(x, s) = \frac{\partial}{\partial s} y(x, s)$

$$v'' = f_y(x, y, y')v + f_{y'}(x, y, y')v', \quad v(a) = 0, \quad v'(a) = 1$$

gilt. Wegen der partiellen Ableitungen $f_y, f_{y'}$ ist dieses Anfangswertproblem i.a. wesentlich komplizierter als (3.2.2). Aus diesem Grund ersetzt man in der Newton-Formel (3.2.3) die Ableitung $F'(s^{(i)})$ durch einen Differenzenquotienten $\triangle F(s^{(i)})$,

$$\triangle F(s^{(i)}) := \frac{F(s^{(i)} + \triangle s^{(i)}) - F(s^{(i)})}{\triangle s^{(i)}},$$

wobei $\triangle s^{(i)}$ „genügend" klein zu wählen ist. $F(s^{(i)} + \triangle s^{(i)})$ berechnet man wie $F(s^{(i)})$ durch Lösung eines Anfangswertproblems. Dabei ergeben sich folgende Schwierigkeiten: Wählt man $\triangle s^{(i)}$ zu groß, ist $\triangle F(s^{(i)})$ nur eine schlechte Approximation für $F'(s^{(i)})$ und die Iteration:

$$s^{(i+1)} = s^{(i)} - \frac{F(s^{(i)})}{\triangle F(s^{(i)})}$$

konvergiert wesentlich schlechter gegen $\hat{s}$ als (3.2.3) (falls sie überhaupt konvergiert). Wählt man $\triangle s^{(i)}$ zu klein, wird $F(s^{(i)}+\triangle s^{(i)}) \approx F(s^{(i)})$ und bei der Subtraktion $F(s^{(i)}+\triangle s^{(i)})-F(s^{(i)})$ tritt Auslöschung auf, so daß selbst kleine Fehler bei der Berechnung von $F(s^{(i)}), F(s^{(i)} + \triangle s^{(i)})$ das Resultat $\triangle F(s^{(i)})$ stark verfälschen. Die Lösung der Anfangswertprobleme, d.h. die Berechnung von F, hat man daher möglichst genau vorzunehmen. Der relative Fehler von $F(s^{(i)}), F(s^{(i)} + \triangle s^{(i)})$ darf nur die Größenordnung der Maschinengenauigkeit *eps* haben. Diese Genauigkeit kann man mit Extrapolationsverfahren erreichen.

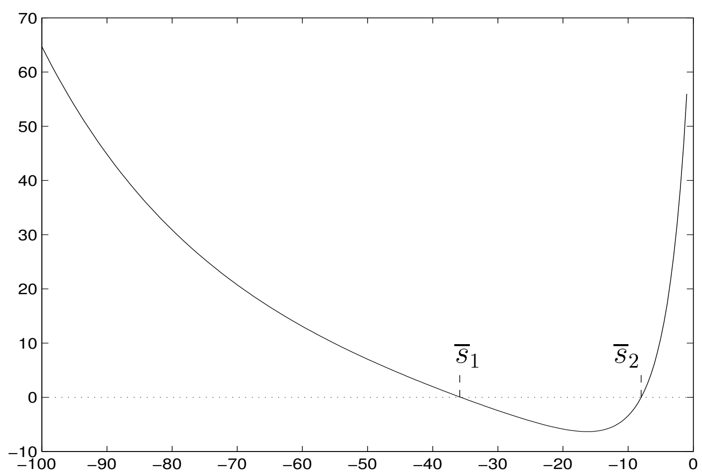
Beispiel 3.2.1 *Zu lösen sei das Randwertproblem*

$$y'' = \frac{3}{2}y^2, \quad y(0) = 4, \quad y(1) = 1.$$

Nach unseren Überlegungen berechnet man zunächst die Lösung y_s von dem Anfangswertproblem

$$y'' = \frac{3}{2}y^2, \quad y(0) = 4, \quad y'(0) = s.$$

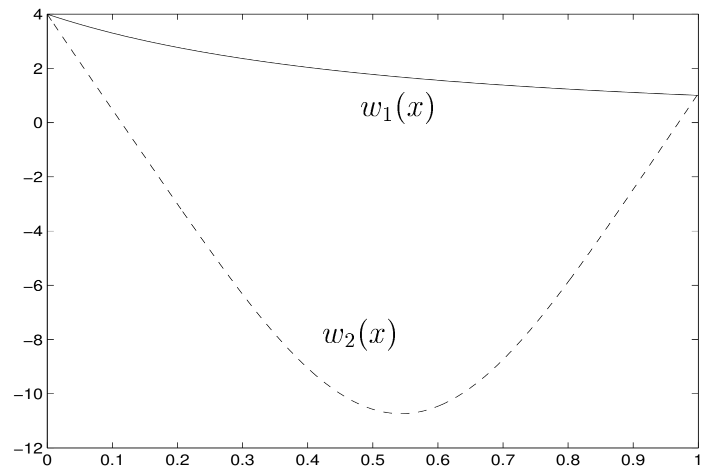
Dann ist $F(s) = y_s(1) - 1$, die folgende Grafik zeigt den Graphen von F.



Man sieht, daß $F(s)$ zwei Nullstellen hat. Newton-Iteration mit angenäherter Ableitung liefert

$$\begin{aligned} \overline{s}_1 &= -35.8585487278 \\ \overline{s}_2 &= -8.0000000000. \end{aligned}$$

Die Lösungen $y_{s_1} = \omega_2, y_{s_2} = \omega_1$ sehen so aus:



Ähnlich wie in dem obigen Beispiel verfährt man bei der Lösung eines allgemeinen Randwertproblems für n gesuchte Funktionen $y_i(x), i = 1, \ldots, n,$

$$y' = f(x,y), \quad r(y(a), y(b)) = 0, \quad y = (y_1, \ldots, y_n)^T$$

wo $f(x,y)$ und $r(u,v)$ Vektoren von n Funktionen sind. Man versucht wieder einen Startvektor $s \in \mathbb{R}^n$ für das Anfangswertproblem

$$y' = f(x,y), \quad y(a) = s \tag{3.2.4}$$

so zu bestimmen, daß die Lösung $y(x) = y(x,s)$ den Randbedingungen

$$r(y(a,s), y(b,s)) \equiv r(s, y(b,s)) = 0$$

genügt. Dazu hat man eine Lösung $s = (\sigma_1, \sigma_2, \ldots, \sigma_n)^T$ der Gleichung

$$F(s) = 0, \quad F(s) := r(s, y(b,s))$$

zu finden. Dies macht man z.B mit Hilfe des allgemeinen Newton-Verfahrens

$$s^{(i+1)} = s^{(i)} - DF(s^{(i)})^{-1} \cdot F(s^{(i)}).$$

Dabei ist in jedem Iterationsschritt $F(s^{(i)})$, die Funktionalmatrix

$$DF(s^{(i)}) = \left[\frac{\partial F_j}{\partial \sigma_k}\right]_{s=s^{(i)}}$$

und die Lösung $d^{(i)} := s^{(i)} - s^{(i+1)}$ des linearen Gleichungssystems $DF(s^{(i)})d^{(i)} = F(s^{(i)})$ zu berechnen. Zur Berechnung von $F(s^{(i)}) = r(s^{(i)}, y(b, s^{(i)}))$ hat man $y(b, s^{(i)})$ zu bestimmen, d.h.

das Anfangswertproblem (3.2.4) für $s = s^{(i)}$ zu lösen. Zur Berechnung von $DF(s^{(i)})$ beachte man

$$DF(s) = D_u r(s, y(b)) + D_v r(s, y(b, s)) \cdot Z(b, s)$$

mit den Matrizen

$$\begin{aligned} D_u r(u, v) &= \left[\frac{\partial r_i(u, v)}{\partial u_j}\right], \\ D_v r(u, v) &= \left[\frac{\partial r_i(u, v)}{\partial v_j}\right], \\ Z(b, s) &= D_s y(b, s) = \left[\frac{\partial y_i(b, s)}{\partial \sigma_j}\right]. \end{aligned}$$

Bei nichtlinearen Funktionen f und r wird man $DF(s)$ jedoch nicht mittels dieser komplizierten Formeln berechnen, sondern stattdessen durch Differenzenquotienten approximieren. Dazu ersetzt man $DF(s)$ näherungsweise durch die Matrix

$$\triangle F(s) = (\triangle_1 F(s), \ldots, \triangle_n F(s))$$

mit

$$\triangle_j F(s) = \frac{F(\sigma_1, \ldots, \sigma_j + \triangle\sigma_j, \ldots, \sigma_n) - F(\sigma_1, \ldots, \sigma_j, \ldots, \sigma_n)}{\triangle\sigma_j}.$$

Wegen $F(s) = r(s, y(b, s))$ hat man zur Berechnung von $\triangle_j F(s)$ $y(b, s) = y(b, \sigma_1, \ldots, \sigma_n)$ und $y(b, \sigma_1, \ldots, \sigma_j + \triangle\sigma_j, \ldots, \sigma_n)$ durch Lösung der entsprechenden Anfangswertprobleme zu bestimmen.

Bei der Ersetzung von $DF(s^{(i)})$ durch $\triangle F(s^{(i)})$ geht i.a. die (lokale) quadratische Konvergenz des Newton-Verfahrens verloren. Das Ersatzverfahren ist in der Regel nur noch (lokal) linear konvergent und zwar umso besser, je besser $\triangle F(s^{(i)})$ und $DF(s^{(i)})$ übereinstimmen. In dem Sonderfall *linearer Randwertprobleme* ist nun $DF(s) = \triangle F(s)$ für alle s (bei beliebiger Wahl der $\triangle\sigma_j$) und damit stimmen das Newton-Verfahren und das Ersatzverfahren überein. Es sei

$$y' = T(t) \cdot y + g(t)$$

$$Ay(a) + By(b) = c$$

mit einer $n \times n-$Matrix $T(t)$, einer Funktion $g : \mathbb{R} \to \mathbb{R}^n$ und konstanten $n \times n-$Matrizen A und B. Wir setzen im folgenden voraus, daß $T(t)$ und $g(t)$ auf $[a, b]$ stetige Funktionen sind. Mit $y(t, s)$ bezeichnen wir wieder die Lösung des Anfangswertproblems

$$y' = T(t) \cdot y + g(t), \quad y(a, s) = s. \tag{3.2.5}$$

Für $y(t, s)$ läßt sich eine explizite Formel angeben

$$y(t, s) = Y(t) \cdot s + y(t, 0), \tag{3.2.6}$$

wobei die $n \times n$-Matrix $Y(t)$ die Lösung des Anfangswertproblems

$$Y' = T(t) \cdot Y, \quad Y(a) = I$$

ist. Bezeichnet man die rechte Seite von (3.2.6) mit $u(t,s)$, so gilt nämlich

$$\begin{aligned} u(a,s) &= Y(a)\cdot s + y(a,0) = I\cdot s + 0 = s \\ D_t u(t,s) &= u'(t,s) = Y'(t)\cdot s + y'(t,0) \\ &= T(t)\cdot Y(t)\cdot s + T(t)\cdot y(t,0) + g(t) \\ &= T(t)\cdot u(t,s) + g(t) \end{aligned}$$

d.h. $u(t,s)$ ist eine Lösung von (3.2.5). Da unter den oben getroffenen Voraussetzungen über $T(t)$ und $g(t)$ das Anfangswertproblem eindeutig lösbar ist, folgt $u(t,s) = y(t,s)$. Mit (3.2.6) ergibt sich für die Funktion $F(s)$

$$F(s) = As + By(b,s) - c = [A + BY(b)]\cdot s + By(b,0) - c.$$

$F(s)$ ist also auch eine lineare Funktion von s. Daher gilt

$$DF(s) = \triangle F(s) = A + BY(b) = \triangle F(0).$$

Die Lösung $\overline{s}$ von $F(s) = 0$ ist (die Existenz von $\triangle F(0)^{-1}$ vorausgesetzt)

$$\begin{aligned} \overline{s} &= -[A + BY(b)]^{-1}\cdot[By(b,0) - c] \\ &= 0 - \triangle F(0)^{-1}F(0) \end{aligned}$$

oder etwas allgemeiner

$$\overline{s} = s^{(0)} - \triangle F(s^{(0)})^{-1}F(s^{(0)})$$

für beliebiges $s^{(0)} \in \mathbb{R}^n$. Mit anderen Worten, die Lösung $\overline{s}$ von $F(s) = 0$ und damit die Lösung des linearen Randwertproblems wird ausgehend von beliebigen Startwerten $s^{(0)}$ durch das einfache Schießverfahren in *einem* Iterationsschritt geliefert.

3.3 Mehrzielmethode

Einige Schwierigkeiten beim einfachen Schießverfahren sind (siehe z.B. Stoer, Bulirsch):

a) Die Lösung von $y' = f(t,y), y(a) = s$ kann sehr empfindlich von s abhängen. Man hat $|y(t,s) - y(t,\tilde{s})| \le e^{L(t-a)}|s - \tilde{s}|$ mit Lipschitzkonstante L für f. Für $b \gg a$ kann eine Ungenauigkeit im Anfangswert erhebliche Abweichungen in $y(b,s)$ bewirken.

b) Häufig hat f zwar partielle Ableitungen nach y, aber $||D_y f(t,y)||$ ist unbeschränkt auf $[a,b] \times \mathbb{R}^n$. In diesem Fall braucht die Lösung y des Anfangwertproblems $y' = f(t,y), y(t_0) = s$ nur noch in einer gewissen Umgebung von t_0 definiert zu sein, deren Größe von s abhängt. Dann existieren eventuell $y(a,t_0,s)$ und $y(b,t_0,s)$ nur für die Werte s aus einer kleinen Menge M. Dabei ist in der Regel M nicht bekannt. Hat man $s^{(0)} \notin M$, so bricht das einfache Schießverfahren zusammen.

Die Idee der Mehrzielmethode ist nun, das Interval $[a,b]$ in Teilintervalle zu zerlegen und das einfache Schießverfahren auf alle Teilintervalle gleichzeitig anzuwenden. Sei dazu

$$a = t_1 < t_2 < \ldots < t_m = b$$

eine Zerlegung des Intervalls $[a,b]$. Seien $s_1, \ldots, s_m$ gegebene Anfangswerte. Löse nun auf jedem Teilintervall $[t_j, t_{j+1}],\ j = 1, \ldots, m-1$

$$y' = f(t,y), \quad y(t_j) = s_j$$

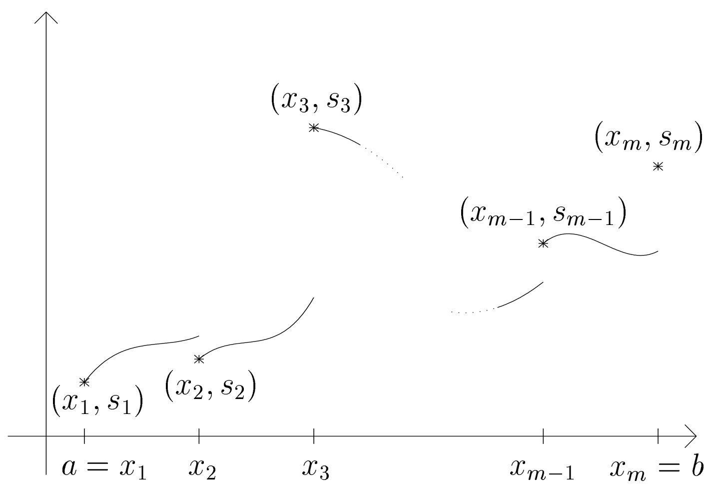
(bezeichne die Lösung mit $y(t,t_j,s_j)$) und bestimme die s_j dann so, daß die aus den $y(t,t_j,s_j), j = 1, \ldots, m-1$ zusammengesetzte Funktion

$$y(t) = \begin{cases} y(t,t_i,s_i) & \text{für } t \in [t_i, t_{i+1}] \\ s_m & \text{für } t = b \end{cases}$$

stetig ist und der Differentialgleichung

$$y' = f(t,y), \quad r(y(a), y(b)) = 0$$

genügt.



Diese Forderung führt zu den folgenden $n \times m$ Bedingungen

$$\begin{aligned} y(t_{i+1}, t_i, s_i) &= s_{i+1} \qquad i = 1, \ldots, m-1 \\ r(s_1, s_m) &= 0. \end{aligned}$$

Dies ist ein nichtlineares Gleichungssystem für $s = [s_1, \ldots, s_m]^T$ der Form

$$F(s) := \begin{bmatrix} F_1(s_1, s_2) \\ F_2(s_2, s_3) \\ \vdots \\ F_{m-1}(s_{m-1}, s_m) \\ F_m(s_1, s_m) \end{bmatrix} := \begin{bmatrix} y(t_2, t_1, s_1) - s_2 \\ y(t_3, t_2, s_2) - s_3 \\ \vdots \\ y(t_m, t_{m-1}, s_{m-1}) - s_m \\ r(s_1, s_m) \end{bmatrix} = 0. \tag{3.3.1}$$

Es kann wieder mit Hilfe des Newton-Verfahrens

$$s^{(i+1)} = s^{(i)} - DF(s^{(i)})^{-1} F(s^{(i)}) \tag{3.3.2}$$

iterativ gelöst werden. In jedem Schritt des Verfahrens müssen $F(s)$ und $DF(s)$ für $s = s^{(i)}$ berechnet werden. Zur Berechnung von $F(s)$ hat man für $k = 1, 2, \ldots, m-1$ durch Lösung der Anfangswertprobleme

$$y' = f(t, y), \quad y(t_k) = s_k$$

die Werte $y(t_{k+1}, t_k, s_k)$, $k = 1, 2, \ldots, m-1$, zu bestimmen, und $F(s)$ gemäß (3.3.1) zu berechnen. Die Funktionalmatrix

$$DF(s) = [D_{s_k} F_i(s)]_{i,k=1,\ldots,m}$$

hat wegen der besonderen Struktur der F_i (3.3.1) die Gestalt

$$DF(s) = \begin{bmatrix} G_1 & -I & 0 & & & 0 \\ 0 & G_2 & -I & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & G_{m-1} & -I \\ A & 0 & & & 0 & B \end{bmatrix},$$

wobei die $n \times n$-Matrizen $A, B, G_k,\ k = 1, \ldots, m-1$, selbst wieder Funktionalmatrizen sind:

$$\begin{aligned} G_k &:= D_{s_k} F_k(s) = D_{s_k} y(t_{k+1}, t_k, s_k), \qquad k = 1, 2, \ldots, m-1, \\ B &:= D_{s_m} F_m(s) = D_{s_m} r(s_1, s_m), \\ A &:= D_{s_1} F_m(s) = D_{s_1} r(s_1, s_m). \end{aligned}$$

Die G_k werden wieder durch Differentialquotienten ersetzt. Für jedes $G_k, k = 1, \ldots, m-1$ ist 1 Anfangswertaufgabe zu lösen. Mit den $m-1$ Anfangswertaufgaben zur Bestimmung von $F(s^{(i)})$ sind also insgesamt weitere $(m-1)$ Anfangswertprobleme zu lösen.

Die Berechnung von $s^{(i+1)}$ aus $s^{(i)}$ kann man dann wie folgt durchführen. Setze

$$\begin{bmatrix} \triangle s_1 \\ \vdots \\ \triangle s_m \end{bmatrix} := s^{(i+1)} - s^{(i)} \quad \text{und} \quad F_j = F_j(s_j^{(i)}, s_{j+1}^{(i)}).$$

Das Gleichungssystem (3.3.2) des Newton-Verfahrens lautet dann

$$DF(s^{(i)}) \begin{bmatrix} \triangle s_1 \\ \vdots \\ \triangle s_m \end{bmatrix} = - \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix}.$$

Multipliziert man beide Seiten nacheinander mit

$$\begin{bmatrix} I & & & \\ G_2 & I & & \\ & & \ddots & \\ & & & I \end{bmatrix}, \begin{bmatrix} I & & & & \\ & I & & & \\ & G_3 & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix}, \ldots, \begin{bmatrix} I & & & & & \\ & I & & & & \\ & & \ddots & & & \\ & & & I & & \\ & & & G_{m-1} & I & \\ & & & & 0 & I \end{bmatrix}, \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & B & I \end{bmatrix}$$

(Block-Gauß-Elimination) so ergibt sich

$$\begin{bmatrix} G_1 & -I & 0 & & & \\ G_2 G_1 & 0 & -I & & & \\ G_3 G_2 G_1 & 0 & 0 & \ddots & & \\ \vdots & \vdots & \vdots & & \ddots & \\ G_{m-1} \ldots G_1 & 0 & 0 & & & -I \\ A + B G_{m-1} \ldots G_1 & 0 & 0 & & & 0 \end{bmatrix} \begin{bmatrix} \triangle s_1 \\ \triangle s_2 \\ \triangle s_3 \\ \vdots \\ \triangle s_{m-1} \\ \triangle s_m \end{bmatrix} =$$

$$
- \begin{bmatrix} F_1 \\ F_2 + G_2 F_1 \\ F_3 + G_3 F_2 + G_3 G_2 F_1 \\ \vdots \\ F_{m-1} + G_{m-1} F_{m-2} + \ldots + G_{m-1} \ldots G_2 F_1 \\ F_m + B F_{m-1} + B G_{m-1} F_{m-2} + \ldots + B G_{m-1} \ldots G_2 F_1 \end{bmatrix}
$$

Löse zuerst die letzte Gleichung mit dem Gaussverfahren

$$
(A + B G_{m-1} \ldots G_1) \triangle s_1 = -F_m - B \sum_{j=1}^{m-1} G_{m-1} \ldots G_{j+1} F_j
$$

und löse dann mit dem nun bekannten $\triangle s_m$ sukzessive nach $\triangle s_m, \triangle s_{m-1}, \ldots, \triangle s_2$ auf.

3.4 Differenzenverfahren für Randwertaufgaben

Die allen Differenzenverfahren zugrunde liegende Idee ist, in einer Differentialgleichung die Differentialquotienten durch passende Differenzenquotienten zu ersetzen und die so erhaltenen diskretisierten Gleichungen zu lösen.

Die Idee zur Lösung des Randwertproblems

$$y' = f(t, y) \quad r(y(a), y(b)) = 0$$

ist die folgende (in Numerik II ganz analog zur Lösung von partiellen Differentialgleichungen). Wir bilden eine Knotenverteilung ("Gitter")

$$a = t_0 < t_1 < \ldots < t_{N+1} = b,$$

betrachten die Differentialgleichung nur an den Stellen t_i und ersetzen die Ableitungen durch Differenzen

$$y'(t_i) = \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} + O(t_{i+1} - t_i) = f(t_i, y(t_i)).$$

So erhält man Näherungen $u_0, \ldots, u_{N+1}$ an $y(t_0), \ldots, y(t_{N+1})$ mit

$$\begin{aligned} u_{i+1} &= u_i + (t_{i+1} - t_i)\ f(t_i, u_i) \quad i = 0, \ldots, N \\ r(u_0, u_{N+1}) &= 0 \end{aligned}$$

Dies ergibt ein nichtlineares Gleichungssystem für $u_0, \ldots, u_{N+1}$. Bei Differentialgleichungen höherer Ordnung kann man entweder zum System übergehen oder direkt eine Ersetzung der höheren Ableitung durch geeignete Differenzen vornehmen. Wir wollen das Vorgehen an folgendem Spezialfall genauer betrachten.

$$-y'' + q(t)y = g(t) \quad y(a) = \alpha \quad y(b) = \beta \tag{3.4.1}$$

mit $q, g \in C[a, b]$ (d.h. q und g sind stetige Funktionen auf $[a, b]$) und $q(t) \geq 0$ für $t \in [a, b]$. Unter diesen Voraussetzungen läßt sich zeigen, daß eine eindeutige bestimmte Lösung existiert. Um obiges Problem zu diskretisieren, teilen wir $[a, b]$ in $N + 1$ äquidistante Teile

$$a = t_0 < t_1 < \ldots < t_{N+1} = b, \quad t_j = a + jh, \quad h = \frac{b - a}{N + 1}$$

und ersetzen mit der Abkürzung $y_j := y(t_j)$ den Differentialquotienten $y_i'' = y''(t_i)$ für $i = 1, 2, \ldots, N$ durch

$$\triangle^2 y_i := \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$$

Wir wollen nun den lokalen Diskretisierungsfehler $\tau_i(y) := y''(t_i) - \triangle^2 y_i$ abschätzen. Dazu setzen wir voraus, daß y auf $[a, b]$ 4 mal stetig differenzierbar ist, $y \in C^4[a, b]$. Man findet dann durch Taylorentwicklung von $y(t_i \pm h)$ um t_i

$$y_{i\pm1} = y_i \pm h y_i' + \frac{h^2}{2!} y_i'' \pm \frac{h^3}{3!} y_i''' + \frac{h^4}{4!} y^{(4)}(t_i \pm \theta_i^{\pm} h), \quad 0 < \theta_i^{\pm} < 1,$$

und damit

$$\triangle^2 y_i = y_i'' + \frac{h^2}{24}[y^{(4)}(t_i + \theta_i^+ h) + y^{(4)}(t_i - \theta_i^- h)].$$

Da $y^{(4)}(t)$ noch stetig ist, folgt

$$\tau_i(y) = y''(t_i) - \triangle^2 y_i = -\frac{h^2}{12} y^{(4)}(t_i + \theta_i h) \quad \text{für ein} \quad |\theta_i| < 1. \tag{3.4.2}$$

Damit erfüllen die Werte $y_i = y(t_i)$ die Gleichungen

$$\begin{aligned} y_0 &= \alpha \\ \frac{2y_i - y_{i-1} - y_{i+1}}{h^2} + q(t_i) y_i &= g(t_i) + \tau_i(y), \quad i = 1, 2, \ldots, N, \\ y_{N+1} &= \beta. \end{aligned} \tag{3.4.3}$$

Mit den Abkürzungen $q_i := q(t_i), g_i := g(t_i)$, den Vektoren

$$\overline{y} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \tau(y) := \begin{bmatrix} \tau_1(y) \\ \tau_2(y) \\ \vdots \\ \tau_N(y) \end{bmatrix}, \quad k := \begin{bmatrix} g_1 + \frac{\alpha}{h^2} \\ g_2 \\ \vdots \\ g_{N-1} \\ g_N + \frac{\beta}{h^2} \end{bmatrix}$$

und der symmetrischen $N \times N-$ Tridiagonalmatrix

$$A := \frac{1}{h^2} \cdot \begin{bmatrix} 2 + q_1 h^2 & -1 & & & & \\ -1 & 2 + q_2 h^2 & -1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 + q_N h^2 \end{bmatrix} \tag{3.4.4}$$

ist (3.4.3) äquivalent mit

$$A\overline{y} = k + \tau(y). \tag{3.4.5}$$

Das Differenzenverfahren besteht nun darin, in Gleichung (3.4.5) den Fehlerterm $\tau(y)$ fortzulassen und die Lösung $u = (u_1, \ldots, u_N)^T$ des erhaltenen linearen Gleichungssystems

$$Au = k \tag{3.4.6}$$

als Approximation für $\overline{y}$ zu nehmen.

Wir wollen zunächst einige Eigenschaften der Matrix A zeigen. Dazu schreiben wir $A \leq B$ für zwei $n \times n-$Matrizen, falls $a_{ij} \leq b_{ij}$ für $i, j = 1, 2, \ldots, n$. Es gilt nun:

Satz 3.4.1 *Ist* $q_i \geq 0$ *für* $i = 1, \ldots, n$, *so ist* A *positiv definit und es gilt* $0 \leq A^{-1} \leq \tilde{A}^{-1}$ *mit der positiv definiten* $n \times n-$*Matrix*

$$\tilde{A} := \frac{1}{h^2} \cdot \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

Zur Erinnerung ein kurzer Auszug aus dem Skript zum Mathematischen Praktikum

Eine symmetrische Matrix A *ist genau dann positiv definit, falls*

$$x^T A x > 0 \quad \textit{für alle} \quad x \in \mathbb{R}^n, x \neq 0$$

gilt. Wir nennen solche Matrizen auch kurz spd−Matrizen.

Satz 6 *Für jede spd−Matrix* $A \in \mathbb{R}^{n \times n}$ *gilt:*

i) Alle Eigenwerte von A *sind positiv.*

ii) A *ist invertierbar.*

iii) Alle führenden Hauptabschnittsdeterminanten von A *sind positiv.*

Beweis:

Wir wollen zunächst zeigen, daß A positiv definit ist. Dazu betrachten wir die $N \times N$-Matrix $A_N := h^2 \cdot \tilde{A}$. Ist λ ein Eigenwert von A_N, so gilt nach dem Satz von Gerschgorin

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \quad \text{für ein} \quad i \in \{1, \ldots, N\}$$

mit $A_N = (a_{ij})_{i,j=1}^N$. Also folgt für die Eigenwerte λ_i von A_N die Abschätzung $|\lambda_i - 2| \leq 2$, d.h. $0 \leq \lambda_i \leq 4$. Wäre $\lambda_i = 0$ Eigenwert von A_N, so hätte das $det(A_N) = 0$ zur Folge; es gilt aber $det(A_N) = N + 1$, wie man sich leicht mittels der Rekursionsformel det $(A_{N+1}) = 2det(A_N) - det(A_{N-1})$ überzeugt (man erhält diese sofort durch Entwicklung von $det(A_{N+1})$ nach der 1. Zeile). Andererseits ist auch $\lambda_i = 4$ kein Eigenwert von A_N, denn dann wäre $A_N - 4I$ singulär. Nun gilt aber:

$$A_N - 4I = \begin{bmatrix} -2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & -2 \end{bmatrix} = -D A_N D^{-1},$$

$$D := diag(1, -1, 1, \ldots, \pm 1)$$

also $|det(A_N - 4I)| = |det(A_N)|$, und damit ist mit A_N auch $A_N - 4I$ nichtsingulär. Es folgt die Abschätzung $0 < \lambda_i < 4$ für die Eigenwerte von A_N. Dies zeigt insbesondere, daß $\tilde{A}$ positiv definit ist. Wegen

$$z^H A z = z^H \tilde{A} z + \sum_{i=1}^{N} q_i |z_i|^2, \quad z^H \tilde{A} z > 0 \quad \text{für} \quad z \neq 0 \quad \text{und} \quad q_i \geq 0$$

folgt sofort $z^H A z > 0$ für $z \neq 0$, also die positive Definitheit von A. Da A und $\tilde{A}$ positiv definit sind, sind sie auch invertierbar. Es bleibt noch die Ungleichung $0 \leq A^{-1} \leq \tilde{A}^{-1}$ zu beweisen.

Dazu betrachten wir die Matrizen $D, \tilde{D}, J, \tilde{J}$ mit

$$\begin{aligned} h^2 A &= D(I - J), \\ D &= diag(2 + q_1 h^2, \ldots, 2 + q_N h^2), \\ h^2 \tilde{A} &= \tilde{D}(I - \tilde{J}), \\ \tilde{D} &= 2I. \end{aligned}$$

Offensichtlich gelten wegen $q_i \geq 0$ die Ungleichungen

$$0 \leq \tilde{D} \leq D,$$

$$0 \leq J = \begin{bmatrix} 0 & \frac{1}{2+q_1h^2} & 0 & \ldots & 0 \\ \frac{1}{2+q_2h^2} & 0 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{1}{2+q_{N-1}h^2} \\ 0 & \ldots & 0 & \frac{1}{2+q_Nh^2} & 0 \end{bmatrix} \leq$$

$$\leq \tilde{J} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \ldots & 0 \\ \frac{1}{2} & 0 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{1}{2} \\ 0 & \ldots & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Wegen $\tilde{J} = \frac{1}{2}(-A_N + 2I)$ und der Abschätzung $0 < \lambda_i < 4$ für die Eigenwerte von A_N gilt $-1 < \eta_i < 1$ für die Eigenwerte η_i von $\tilde{J}$, d.h. man hat für den Spektralradius von $\tilde{J} : \rho(\tilde{J}) < 1$. Daraus folgt leicht die Konvergenz der Reihe

$$(0 \leq)\ I + \tilde{J} + \tilde{J}^2 + \tilde{J}^3 + \ldots = (I - \tilde{J})^{-1}.$$

Wegen $0 \leq J \leq \tilde{J}$ konvergiert dann auch

$$(0 \leq)\ I + J + J^2 + J^3 + \ldots = (I - J)^{-1} \leq (I - \tilde{J})^{-1}$$

und es gilt $0 \le D^{-1} \le \tilde{D}^{-1}$ und daher

$$0 \le (h^2 A)^{-1} = (I - J)^{-1} D^{-1} \le (I - \tilde{J})^{-1} \tilde{D}^{-1} = (h^2 \tilde{A})^{-1},$$

was zu zeigen war. $\surd$

Das Gleichungssystem (3.4.6) hat also eine eindeutige Lösung. Da A eine positiv definite symmetrische Tridiagonalmatrix ist, kann das Gleichungssystem z.B. mittels des Cholesky-Verfahrens (siehe Mathematisches Praktikum) gelöst werden. Die Anzahl der Operation ist (wegen der Tridiagonalgestalt von A) proportional zu N.

Wir wollen nun eine Abschätzung für die Fehler

$$y_i - u_i$$

der aus (3.4.6) gewonnenen Näherungswerte u_i für die exakten Lösungen $y_i = y(t_i)$, $i = 1, \ldots, N$, herleiten.

Satz 3.4.2 *Das Randwertproblem (3.4.1) habe eine Lösung $y(t) \in C^4[a,b]$ und es gelte $|y^{(4)}(t)| \le M$ für $t \in [a,b]$. Ferner sei $q(t) \ge 0$ für $t \in [a,b]$ und $u = (u_1, \ldots, u_N)^T$ die Lösung von (3.4.6). Dann gilt*

$$|y(t_i) - u_i| \le \frac{Mh^2}{24}(t_i - a)(b - t_i)$$

für $i = 1, 2, \ldots, N$.

<u>Beweis:</u>

Aus (3.4.5) und (3.4.6) hat man für $\overline{y} - u$

$$A(\overline{y} - u) = \tau(y).$$

Satz 3.4.1 und die Darstellung (3.4.2) für $\tau(y)$ ergeben mit der Schreibweise $|y| := (|y_1|, \ldots, |y_N|)^T$

$$\begin{aligned} |\overline{y} - u| &= |A^{-1}\tau(y)| \\ &\le \tilde{A}^{-1}|\tau(y)| \qquad (3.4.7) \\ &\le \frac{Mh^2}{12}\tilde{A}^{-1}e \end{aligned}$$

mit $e = (1, 1, \ldots, 1)^T$. Der Vektor $\tilde{A}^{-1}e$ kann mittels folgender Beobachtung sofort angegeben werden: Das spezielle Randwertproblem

$$-y'' = 1, \quad y(a) = y(b) = 0$$

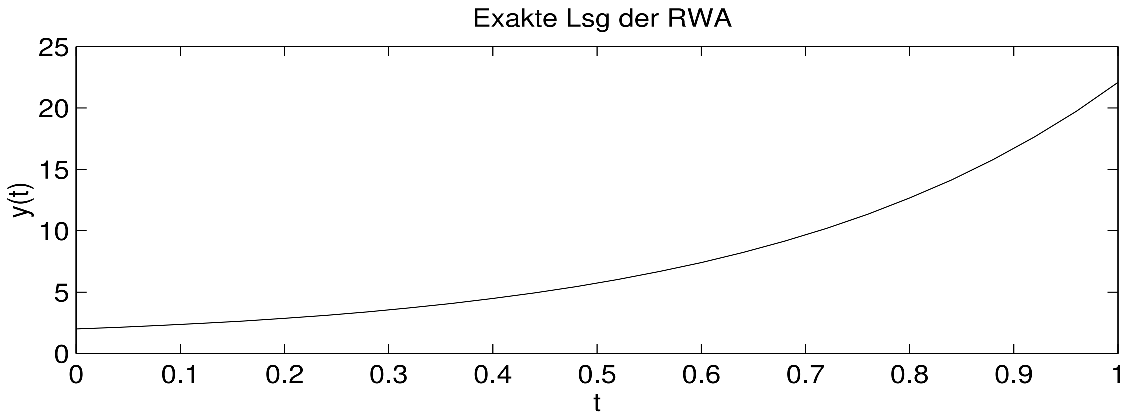
des Types (3.4.1) besitzt die exakte Lösung $y(t) = \frac{1}{2}(t - a)(b - t)$. Für dieses Randwertproblem gilt aber wegen (3.4.2) $\tau(y) = 0$, und es stimmen die diskrete Lösung u von (3.4.6) und die exakte Lösung $\overline{y}$ von (3.4.5) überein. Überdies ist für dieses spezielle Randwertproblem die Matrix A gerade die Matrix $\tilde{A}$ und weiter $k = e$. Also gilt $\tilde{A}^{-1}e = u, u_i = \frac{1}{2}(t_i - a)(b - t_i)$. Zusammen mit (3.4.7) ergibt sich die Behauptung des Satzes. $\surd$

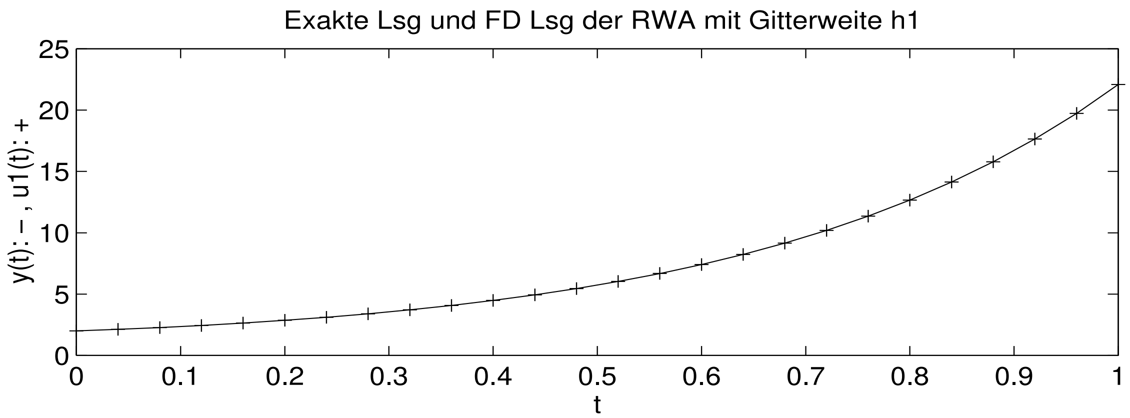
Unter den Voraussetzungen des obigen Satzes gehen die Diskretisierungsfehler also mit h^2 gegen 0, das Differenzenverfahren ist ein Verfahren 2. Ordnung.

Beispiel 3.4.1 *Wir wollen die Randwertaufgabe*

$$y'' = 9y - 9t^2 - 7, \quad y(0) = 2, \quad y(1) = e^3 + 2$$

mit dem Differenzenverfahren lösen. Die exakte Lösung lautet $y(t) = e^{3t} + t^2 + 1$. *Als Schrittweite wurde* $h_1 = 0.04$ *gewählt.*





3.4.1 Allgemeine Konvergenztheorie

Sei $(R, ||\cdot||)$ ein normierter Vektorraum von Funktionen auf einem Gebiet $\Omega \subseteq \mathbb{R}$. $M : R \to C(\Omega)$ sei ein linearer Operator (z.B. $My = -y'' + q(t)y$, dann ist R der Raum zweimal stetig differenzierbarer Funktionen y mit $y(a) = y(b) = 0$ (andere Randbedingungen lassen sich durch Umtransformation auf diese transformieren)). Das zu lösende Problem lautet hier: Finde zu gegebenem $r \in C(\Omega)$ ein $y \in R$ mit

$$My = r. \tag{3.4.8}$$

Zur Diskretisierung von (3.4.8) ersetzt man Ω durch ein Gitter I_n auf Ω und wählt eine passende Approximation M_n von M auf I_n:

$$M_n : R|_{I_n} \to L^\infty(I_n).$$

Der lokale Diskretisierungsfehler ist hier

$$\tau_n = M_n y_n - r_n$$

mit $y_n = y|_{I_n}$ und $r_n = r|_{I_n}$. Das durch

$$M_n u_n = r_n$$

gegebene Verfahren heißt <u>konsistent von der Ordnung p,</u> falls

$$||\tau_n||_\infty(I_n) = \sup_{t \in I_n} |\tau_n(t)| \leq K h^p.$$

Man löst also $M_n u_n = r_n$ und erhält die Näherung u_n an y_n. Für den Fehler $e_n = y_n - u_n$ erhalten wir mit

$$\tau_n = M_n(y_n - u_n) = M_n e_n$$

gerade

$$|e_n| = |M_n^{-1}\tau_n| \leq lub_\infty(M_n^{-1})||\tau_n||_\infty(I_n).$$

Konvergenz gilt also genau dann, wenn

$$lub_\infty(M_n^{-1})||\tau_n||_\infty(I_n) \xrightarrow{h \to 0} 0,$$

z.B. wenn M_n^{-1} gleichmäßig beschränkt in h ist. Ist M_n etwa eine Matrix (wie in dem vorhin betrachtetem Spezialfall), so ist die Bestimmung von $lub_\infty(M_n^{-1})$ einfacher, falls $M_n^{-1} \geq 0$. Dann hat man nämlich komponenterweise

$$|M_n^{-1}x| \leq M_n^{-1}|x| \leq ||x||_\infty M_n^{-1}e$$

und daher $lub_\infty(M_n^{-1}) = \max_i |M_n^{-1}\, e|_i$. In dem vorhin betrachtetem Spezialfall konnten wir dies abschätzen.

3.5 Variationsmethoden (Finite Elemente)

Eine der populärsten Methoden zur Lösung von Randwertaufgaben (insbesondere für partielle Differentialgleichungen) ist die Finite-Elemente-Methode. Wir wollen sie hier nur exemplarisch an einem speziellen Problem darstellen, dem Sturm-Liouville-Problem. Damit wird der Grundstock gelegt für die Methoden zur Lösung elliptischer partieller Differentialgleichungen (Numerik II). Die Grundidee unterscheidet sich wesentlich von den bisher betrachteten Methoden. Anstatt die Lösung an diskreten Punkten zu bestimmen, werden Funktionen bestimmt, die die Lösung der Randwertaufgabe approximieren. Zu diesem Zweck führt man im allgemeinen die Randwertaufgabe in eine Variationsaufgabe über und bestimmt die Lösung dieser Variationsaufgabe in einem geeigneten Funktionenraum. Dies ist immer dann möglich, wenn die Lösung der Randwertaufgabe gewisse Minimalitätseigenschaften hat.

Das Problem, das wir hier exemplarisch betrachten wollen, ist das verallgemeinerte Sturm-Liouville-Problem

$$Ly := -[p(t)y'(t)]' + q(t)y(t) = r(t), \quad y(a) = \alpha, \quad y(b) = \beta \tag{3.5.1}$$

für $t \in [a,b]$. Dabei gelte für die Koeffizientenfunktionen

$$\begin{array}{lll} p \in C^1[a,b] & p(t) \geq p_0 > 0 \quad \text{in} \quad [a,b] \\ q, r \in C[a,b] & q(t) \geq 0 \quad \text{in} \quad [a,b] \end{array} \tag{3.5.2}$$

In diesem Fall kann man zeigen, daß es eine eindeutige Lösung der Randwertaufgabe gibt.

Ist $y(t)$ die Lösung von (3.5.1), so ist $u(t) = y(t) - l(t)$ mit

$$l(t) = \alpha\,\frac{b-t}{b-a} + \beta\,\frac{t-a}{b-a}, \quad l(a) = \alpha, \quad l(b) = \beta$$

die Lösung eines Randwertproblems der Form

$$\begin{aligned} -(pu')' + qu &= -[p(y-l)]' + qy - ql \\ &= -(py')' + (pl')' + qy - ql \\ &= r + (pl')' - ql \\ &=: f \end{aligned}$$

mit $u(a) = u(b) = 0$. Wir können also im folgenden oBdA

$$Ly := -(py')' + qy = f, \quad y(a) = y(b) = 0 \tag{3.5.3}$$

betrachten. Der Differentialoperator L bildet die Menge

$$D_L := \{v \in C^2[a,b] | v(a) = v(b) = 0\}$$

aller zweimal auf $[a,b]$ stetig differenzierbaren reellen Funktionen, die die Randbedingung $v(a) = v(b) = 0$ erfüllen, in die Menge $C[a,b]$ aller auf $[a,b]$ stetigen Funktionen ab. Das Sturm-Liouville-Problem ist also damit äquivalent, eine Lösung von

$$Ly = f, \quad y \in D_L$$

zu finden. (Beachte: Die Randbedingungen sind jetzt in die Definition von D_L "eingebaut", ohne die Homogenisierung wäre D_L kein linearer Raum). Offensichtlich ist D_L ein reeller Vektorraum und L ein linearer Operator auf D_L: Mit $u, v \in D_L$ gehört auch $\alpha u + \beta v$ zu D_L und es ist $L(\alpha u + \beta v) = \alpha L(u) + \beta L(v)$ für alle reellen Zahlen α, β. Auf der Menge $L_2[a,b]$ aller auf $[a,b]$ quadratisch integrierbaren Funktionen führen wir nun durch die Definition

$$(u,v) := \int_a^b u(x) \cdot v(x) dx, \quad ||u||_2 := (u,u)^{1/2}$$

eine Bilinearform und eine Norm ein. Der Differentialoperator L hat einige für das Verständnis der Variationsmethoden wichtige Eigenschaften. Es gilt

Satz 3.5.1 *L ist ein symmetrischer Operator auf D_L, d.h. es gilt*

$$(u, L(v)) = (L(u), v) \quad \textit{für alle} \quad u, v, \in D_L.$$

Beweis:

Durch partielle Integration findet man

$$\begin{aligned}
(u, L(v)) &= \int_a^b u(x)[-(p(x)v'(x))' + q(x)v(x)]dx \\
&= -u(x)p(x)v'(x)|_a^b + \int_a^b [p(x)u'(x)v'(x) + q(x)u(x)v(x)]dx \\
&= \int_a^b [p(x)u'(x)v'(x) + q(x)u(x)v(x)]dx
\end{aligned}$$

weil $u(a) = u(b) = 0$ für $u \in D_L$ gilt. Aus Symmetriegründen folgt ebenso

$$(L(u), v) = \int_a^b [p(x)u'(x)v'(x) + q(x)u(x)v(x)]dx$$

und damit die Behauptung. $\surd$

$\int_a^b [pu'v' + quv]dx$ ist eine symmetrische Bilinearform auf

$$D = \{u \in C[a,b] | u' \in L_2[a,b], u(a) = u(b) = 0\}.$$

Insbesondere gehören aller stückweise stetig differenzierbaren Funktionen, die die Randbedingungen erfüllen, zu D. D ist wieder ein reeller Vektorraum mit $D \supseteq D_L$. Durch

$$\begin{aligned}
[u,v] &:= \int_a^b [p(x)u'(x)v'(x) + q(x)u(x)v(x)]dx \\
&= (pu', v') + (qu, v)
\end{aligned}$$

wird auf D ein Skalarprodukt definiert, welches für $u, v \in D_L$ mit $(u, L(v))$ übereinstimmt. Wie eben zeigt man für $y \in D_L, u \in D$ durch partielle Integration

$$(u, L(y)) = [u, y].$$

Bezüglich des auf D_L eingeführten Skalarproduktes ist L ein positiv definiter Operator in folgendem Sinn:

Satz 3.5.2 *Unter den Voraussetzungen (3.5.2) ist*

$$(L(u), u) = [u, u] > 0 \quad \textit{für alle} \quad u \neq 0, u \in D_L.$$

(Koerzivität der Bilinearform). Es gilt sogar die Abschätzung

$$\gamma ||u||_\infty^2 \leq [u, u] \leq \Gamma ||u'||_\infty^2 \quad \textit{für alle} \quad u \in D \tag{3.5.4}$$

mit der Norm $||u||_\infty := \sup\limits_{a \leq x \leq b} |u(x)|$ *und den Konstanten*

$$\gamma := \frac{p_0}{b-a}, \Gamma := ||p||_\infty (b-a) + ||q||_\infty (b-a)^3.$$

Beweis:

Wegen $\gamma > 0$ genügt es, (3.5.4) zu zeigen. Für $u \in D$ gilt wegen $u(a) = 0$

$$u(x) = \int_a^x u'(\xi) d\xi \quad \text{für} \quad x \in [a, b].$$

Die Schwarzsche Ungleichung liefert die Abschätzung

$$\begin{aligned} u(x)^2 &\leq \int_a^x 1^2 d\xi \cdot \int_a^x u'(\xi)^2 d\xi \\ &= (x-a) \int_a^x u'(\xi)^2 d\xi \\ &\leq (b-a) \cdot \int_a^b u'(\xi)^2 d\xi \end{aligned}$$

und daher

$$||u||_\infty^2 \leq (b-a) \int_a^b u'(x)^2 dx \leq (b-a)^2 ||u'||_\infty^2.$$

Nun ist nach Voraussetzung $p(x) \geq p_0 > 0, q(x) \geq 0$ für $x \in [a, b]$; also folgt

$$\begin{aligned} [u, u] &= \int_a^b [p(x) u'(x)^2 + q(x) u(x)^2] dx \\ &\geq p_0 \int_a^b u'(x)^2 dx \\ &\geq \frac{p_0}{b-a} ||u||_\infty^2. \end{aligned}$$

Schließlich ist auch

$$\begin{aligned} [u, u] &= \int_a^b [p(x) u'(x)^2 + q(x) u(x)^2] dx \\ &\leq ||p||_\infty (b-a) ||u'||_\infty^2 + ||q||_\infty (b-a) ||u||_\infty^2 \\ &\leq \Gamma ||u'||_\infty^2 \end{aligned}$$

was zu zeigen war. $\surd$

Bemerkung 3.5.1 *Dieser Satz impliziert die Eindeutigkeit der Lösung, denn für* $L(y_1) = L(y_2) = f, y_1, y_2 \in D_2$ *ist* $L(y_1 - y_2) = 0$ *und daher*

$$\begin{aligned} 0 &= (y_1 - y_2, L(y_1 - y_2)) \\ &\geq \gamma ||y_1 - y_2||_\infty^2 \\ &\geq 0, \end{aligned}$$

was direkt $y_1 = y_2$ *nach sich zieht.*

Für $u \in D$ definieren wir das quadratische Funktional

$$\begin{aligned} F \quad : \quad D &\rightarrow \mathbb{R} \\ u &\longmapsto F(u) = [u, u] - 2(f, u). \end{aligned}$$

Grundlegend für die Variationsmethoden ist die Beobachtung, daß die Funktion F ihren kleinsten Wert genau für die Lösung y von (3.5.3) annimmt.

Satz 3.5.3 *Es sei* y *die Lösung von (3.5.3). Dann gilt*

$$F(u) > F(y)$$

für alle $u \in D, u \neq y$.

<u>Beweis:</u>

Es ist $L(y) = f$ und daher

$$\begin{aligned} F(u) &= [u, u] - 2(u, f) \\ &= [u, u] - 2(u, L(y)) \\ &= [u, u] - 2[u, y] + [y, y] - [y, y] \\ &= [u - y, u - y] - [y, y] \\ &\geq -[y, y] \\ &= F(y) \end{aligned}$$

$\surd$

Bemerkung 3.5.2 *Ist* y *Lösung der Randwertaufgabe (3.5.3), so löst* y *die Minimierungsaufgabe* $\min\limits_{u \in D} F(u)$.

Ist umgekehrt y Lösung der Minimierungsaufgabe, so gilt für alle $v \in D, \epsilon \in \mathbb{R} (v \neq 0, \epsilon \neq 0)$:

$$\begin{aligned} F(y) &< F(y + \epsilon v) \\ &= [y + \epsilon v, y + \epsilon v] - 2(f, y + \epsilon v) \\ &= [y, y] + 2\epsilon[y, v] + \epsilon^2[v, v] - 2(f, y) - 2\epsilon(f, v) \\ &= F(y) + \epsilon^2[v, v] + 2\epsilon([y, v] - (f, v)). \end{aligned}$$

Also gilt

$$\epsilon^2[v,v] > 2\epsilon((f,v)-[y,v])$$

bzw.

$$[v,v] > \frac{2}{\epsilon}((f,v)-[y,v]). \tag{3.5.5}$$

Ist nun $v \in D$ fest und $(f,v)-[y,v] \neq 0$, so kann man

$$\epsilon_v = sign((f,v)-[y,v])\eta$$

mit

$$0 < \eta < \frac{2}{[v,v]}|(f,v)-[y,v]|$$

wählen. Dann gilt

$$\begin{aligned} \frac{2}{\epsilon_v}((f,v)-[y,v]) &= \frac{2}{|\epsilon_v|}|(f,v)-[y,v]| \\ &= \frac{2}{\eta}|(f,v)-[y,v]| \\ &> [v,v] \end{aligned}$$

im Widerspruch zu (3.5.5). Also gilt

$$(Ly-f,v) = [y,v]-(f,v) = 0$$

für alle $v \in D$.

Man kann zeigen, daß D dicht in $L_2[a,b]$ ist, d.h. man hat auch $(Ly-f,v)=0$ für alle $v \in L_2[a,b]$,d.h. aber $Ly-f=0$ im L_2-Sinne (d.h. bis auf eine Menge vom Maß 0). Man sagt, $y \in D$, das die Minimierungsaufgabe löst, ist eine <u>schwache Lösung</u> von (3.5.3). Ist y sogar aus D_L, so folgt, daß y (3.5.3) punktweise löst.

3.5.1 Das Ritz'sche Verfahren zur Lösung der Minimierungsaufgabe

Die Idee des Ritz'schen-Verfahren ist es, einen Teilraum $S_n \subset D$ mit $dim\ S_n = n$ zu wählen und F über S_n zu minimieren. Dazu sei $\rho_1, \ldots, \rho_n$ eine Basis für S_n. Dann müssen wir

$$u^* = \sum_{i=1}^{n} \alpha_i^* \rho_i$$

bestimmen mit

$$F(u^*) = \min_{u \in Sn} F(u).$$

Dann ist $[u^* - y, u^* - y] = F(u^*) - F(y)$ minimal, d.h. u^* ist die beste Approximation aus S_n an y bzgl. $[\cdot,\cdot]$. Für $u = \sum_{i=1}^{n} c_i \rho_i \in S_n$ gilt

$$\begin{aligned}
\phi(c_1,\ldots,c_n) &:= F(u) \\
&= F(c_1\rho_1 + \ldots + c_n\rho_n) \\
&= [\sum_{i=1}^{n} c_i\,\rho_i, \quad \sum_{i=1}^{n} c_i\rho_i] - 2(f, \sum_{i=1}^{n} c_i\rho_i) \\
&= \sum_{i,k=1}^{n} [\rho_i,\rho_k] c_i c_k - 2\sum_{i=1}^{n} (f,\rho_i) c_i
\end{aligned}$$

Dies müssen wir nun bzgl. $c = (c_1,\ldots,c_n)^T$ minimieren. In Matrixform lautet obige Gleichung mit

$$A = ([\rho_i,\rho_k])_{i,k=1}^{n} \quad \text{und} \quad b = \begin{bmatrix} (f,\rho_1) \\ \vdots \\ (f,\rho_n) \end{bmatrix}$$

$$\phi(c) = c^T A c - 2b^T c.$$

Die Matrix A ist positiv definit, denn A ist symmetrisch und es gilt für alle Vektoren $d \neq 0$ auch $u := d_1\rho_1 + \ldots + d_n\rho_n \neq 0$ und daher

$$d^T A d = \sum_{i,k} [\rho_i,\rho_k] d_i d_k = [u,u] > 0.$$

Das lineare Gleichungssystem

$$Ax = z$$

besitzt somit eine eindeutige Lösung $x = \overline{x}$, die man z.B. mit Hilfe des Cholesky-Verfahrens berechnen kann.

Notwendige Bedingung für ein Minimum von ϕ ist

$$0 = \nabla\phi(c) := grad(\phi(c)) = 2(Ac - b)$$

d.h.

$$Ac = b.$$

Da A positiv definit ist, liegt ein eindeutiges Minimum vor.

Bemerkung 3.5.3 *Fehlerorthogonalisierung*

$Ac = b$ bedeutet $\sum_{i=1}^{n} [\rho_i,\rho_j] c_i - (f,\rho_j) = 0$ für alle $j = 1,\ldots,n$. Für $\rho_i \in D_L$ folgt dann

$$0 = \sum_{i=1}^{n} (L\rho_i,\rho_j) c_i - (f,\rho_j) = (L\{\sum_{i=1}^{n} \rho_i c_i\} - f, \rho_i)$$

für alle $j = 1,\ldots,n$. Die c_i sind so gewählt, daß der Fehler $L\{\sum_{i=1}^{n} c_i\rho_i\} - f$ orthogonal zu ρ_j ist für alle j $(L\{\sum_{i=1}^{n} c_i\rho_i\} - f \perp \rho_j)$.

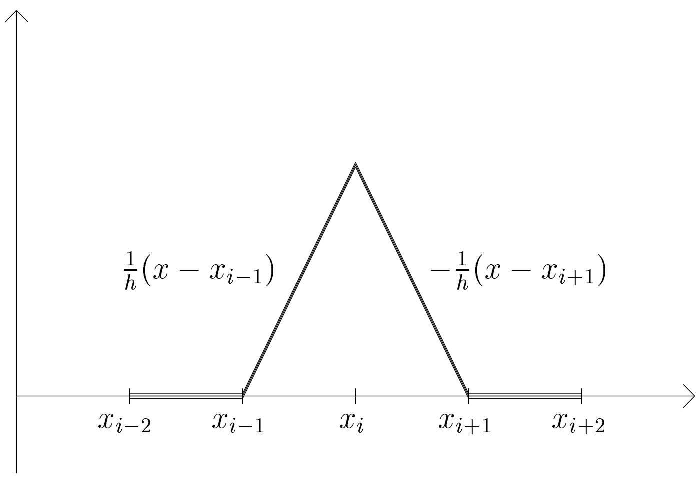
Beispiel 3.5.1 *Wähle ein Gitter auf* $[a, b]$

$$x_i = a + ih \quad mit \quad h = \frac{b-a}{N+1}.$$

S_N sei die Menge aller stetigen Streckenzüge u mit $u(a) = u(b) = 0$ mit Knickstellen $x_1, \ldots, x_N$, d.h.

$$S_N = \{u \in C[a, b] | u(a) = u(b) = 0 \quad und \quad u|_{[x_i, x_{i+1}]} \quad linear\ für \quad i = 0, \ldots, N\} \subset D$$

Wegen der 0-Randbedingung folgt $dim S_N = N$. Eine Basis für S_N ist die Menge der Shape-Funktionen $\rho_i \in S_N$ mit $\rho_i(x_j) = \delta_{ij}$



(Die ρ_i haben einen kleinen Träger, daher werden wir in der Matrix A eine kleine Bandbreite erhalten).

Für $u \in S_N$ gilt nun $u(x) = \sum_{i=1}^{N} u(x_i)\rho_i(x)$. Betrachte nun (3.5.3) mit $p = 1$, und $[a, b] = [0, 1]$

$$-u'' + q(x)u = f, \quad u(0) = u(1) = 0.$$

Es folgt dann

$$[u, v] = \int_0^1 u'v' + quv \quad dx$$

und

$$[\rho_i, \rho_j] = \int_0^1 (\rho_i'\rho_j' + q\rho_i\rho_j)dx = \begin{cases} 0 & für \quad |i - j| > 1 \\ \int\limits_{x_{i-1}}^{x_i} (-\frac{1}{h^2} + q\frac{(x_i - x)(x - x_{i-1})}{h^2})dx & für \quad j = i - 1 \\ \int\limits_{x_{i-1}}^{x_i} (\frac{1}{h^2} + q\frac{(x - x_{i-1})^2}{h^2}dx + \int\limits_{x_i}^{x_{i+1}} (\frac{1}{h^2} + q\frac{(x_{i+1} - x)^2}{h^2})dx & für \quad j = i \end{cases}$$

A ist also tridiagonal.

Mit der Lösung y von $L(y) = f$ folgt wegen $F(u_s) = \min_{u \in S_N} F(u)$ aus dem Beweis von Satz 3.5.3 sofort

$$[u_s - y, u_s - y] = \min_{u \in S_N} [u - y, u - y]. \tag{3.5.6}$$

Wir wollen diese Beziehung zu einer Abschätzung für den Fehler $||u_s - y||_\infty$ benutzen. Es gilt:

Satz 3.5.4 *Seien γ, Γ wie im Satz 3.5.2, y Lösung von (3.5.3). Ferner sei $S \subset D$ ein endlich dimensionaler Teilraum von D und es sei $F(u_s) = \min_{u \in S} F(u)$. Dann gilt*

$$||u_s - y||_\infty \leq \sqrt{\frac{\Gamma}{\gamma}} \; ||u' - y'||_\infty$$

für alle $u \in S$.

Beweis:

Für beliebiges $u \in S \subseteq D$ ergibt sich sofort wegen (3.5.4) und (3.5.6)

$$\begin{aligned} \gamma ||u_s - y||_\infty^2 &\leq [u_s - y, u_s - y] \\ &\leq [u - y, u - y] \\ &\leq \Gamma ||u' - y'||_\infty^2 \end{aligned}$$

$\surd$

In der Praxis verwendet man $u \in S$, für welches $u' - y'$ klein ist, häufig ist u Projektion von y auf S oder Interpolierende an y. Dies sind Methoden aus der Funktionalanalysis.

Beispiel 3.5.2 *Wir wählen als Teilmenge S von D die Menge*

$$Sp_\triangle := \{s \in C^2[a,b] | s|_{[x_i, x_{i+1}]} \in \Pi_3 \quad \textit{für} \quad i = 0, \ldots, n-1, s(a) = s(b) = 0\}$$

aller kubischen Splinefunktionen (siehe Mathematisches Praktikum), die zu einer festen Unterteilung

$$\triangle : a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

des Intervalls $[a, b]$ gehören und an den Endpunkten a, b verschwinden. Offensichtlich ist $Sp_\triangle \subseteq D_L \subseteq D$. Wir bezeichnen mit $||\triangle||$ das maximale Teilintervall der Zerlegung $\triangle$,

$$||\triangle|| := \max_{1 \leq i \leq n} (x_i - x_{i-1}),$$

und setzen weiter

$$K := \max_{1 \leq i \leq n} \frac{||\triangle||}{x_i - x_{i-1}}$$

(für äquidistante Einteilungen ist $K = 1$). Sei die Splinefunktion $u := S_\triangle$ gegeben durch

$$\begin{aligned} u(x_i) &= y(x_i) \quad i = 0, 1, \ldots, n, \\ u'(\xi) &= y'(\xi) \quad \textit{für} \quad \xi = a, b, \end{aligned}$$

wobei y wieder die exakte Lösung ist. Offenbar gilt $S_\triangle \in Sp_\triangle$ und aus der Fehlerabschätzung der Splineinterpolation folgt

$$||S'_\triangle - y'||_\infty \leq \frac{7}{4} K ||y^{(4)}||_\infty || \triangle ||^3$$

sofern $y \in C^4[a, b]$. Zusammen mit Satz 3.5.4 folgt daher

$$||u_s - y||_\infty \leq \sqrt{\frac{\Gamma}{\gamma}} \, \frac{7}{4} K ||y^{(4)}||_\infty || \triangle ||^3.$$

Die Fehlerschranke geht also mit der 3. Potenz der Feinheit $\triangle$ gegen 0, das Verfahren ist somit in dieser Hinsicht den Differenzenverfahren des letzten Abschnitts überlegen.

Die Finite-Element-Methode ist natürlich nicht auf Probleme wie das Sturm-Liouville-Problem beschränkt. Allerdings ist es oft nicht leicht Konvergenz oder Fehlerabschätzungen zu zeigen. Es ist klar, daß man andere als die hier betrachteten, einfachen Ansatzfunktionen wählen kann. Insbesondere bei partiellen Differentialgleichungen ist dies von großer Bedeutung.

Die Hauptarbeit bei Verfahren dieses Typs besteht darin, die Koeffizienten des linearen Gleichungssystems mittels Integration zu berechnen, nachdem man sich für eine geeignete Basis $\rho_1, \ldots, \rho_m$ von S entschieden hat. Für Randwertprobleme bei gewöhnlichen Differentialgleichungen sind diese Verfahren in der Regel zu aufwendig und zu ungenau und mit der Mehrzielmethode nicht konkurrenzfähig (s. den nächsten Abschnitt für vergleichende Resultate). Ihr Wert zeigt sich erst bei Randwertproblemen für partielle Differentialgleichungen.

Ein anderer Ansatz für Randwertprobleme, der i.a. besser analysiert werden kann, sind die Kollokationsmethoden, welche wir nach einem Vergleich der bisher besprochenen Verfahren, erläutern wollen.

3.6 Vergleich der Methoden zur Lösung von Randwertproblemen für gewöhnliche Differentialgleichungen

Die in den vorausgegangenen Abschnitten beschriebenen Methoden, nämlich

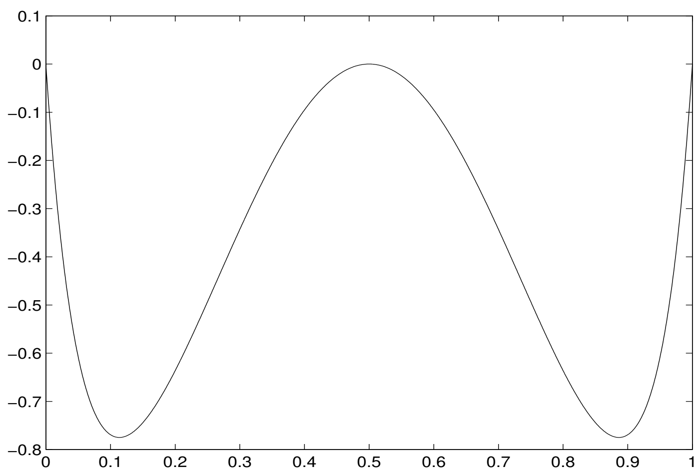
1. Einfachschießverfahren,
2. Mehrzielmethode,
3. Differenzenverfahren,
4. Variationsverfahren,

sollen anhand des folgenden Beispiels verglichen werden. Gegeben sei das lineare Randwertproblem mit linearen separierten Randbedingungen

$$\begin{aligned} -y'' + 400y &= -400cos^2\pi x - 2\pi^2 cos(2\pi x) \\ y(0) &= y(1) = 0 \end{aligned} \tag{3.6.1}$$

mit der exakten Lösung

$$y(x) = \frac{e^{-20}}{1+e^{-20}}\, e^{20x} + \frac{1}{1+e^{-20}} e^{-20x} - cos^2(\pi x). \tag{3.6.2}$$



Obwohl dieses Problem verglichen mit den meisten Problemen aus der Praxis als sehr einfach bezeichnet werden muß, liegen bei ihm folgende Besonderheiten vor, die Schwierigkeiten bei seiner Lösung erwarten lassen:

1. Die zu (3.6.1) gehörige homogene Differentialgleichung $-y'' + 400y = 0$ besitzt exponentiell stark wachsende bzw. fallende Lösungen der Form $y(x) = c \cdot e^{\pm 20x}$. Dies führt zu Schwierigkeiten für das Einfachschießverfahren.

2. Die Ableitungen $y^{(i)}(x), i = 1, 2, \ldots,$ der exakten Lösung (3.6.2) sind für $x \approx 0$ und $x \approx 1$ sehr groß. Die Fehlerabschätzungen für das Differenzenverfahren bzw. für die Variationsmethode lassen deshalb große Fehler erwarten.

Bei 12-stelliger Rechnung wurden folgende Ergebnisse gefunden (bei der Bewertung der Fehler, insbesondere der relativen Fehler, beachte man den Verlauf der Lösung: $\max\limits_{0 \le x \le 1} |y(x)| = 0.77\ldots, y(0) = y(1) = 0, y(0.5) = +0.907998593379 \cdot 10^{-4})$:

1. *Einfachschießverfahren*: Nach einer Iteration wurden die Anfangswerte

$$y(0) = 0, \quad y'(0) = -20\frac{1 - e^{-20}}{1 + e^{-20}} = -19.99999\,99176\ldots$$

   der exakten Lösung mit einem relativen Fehler $\le 3.2 \cdot 10^{-11}$ gefunden. Löst man dann das (3.6.1) entsprechende Anfangswertproblem mit Hilfe eines Extrapolationsverfahren, wobei man als Anfangswerte die auf Maschinengenauigkeit gerundeten exakten Anfangswerte $y(0), y'(0)$ nimmt, so erhält man statt der exakten Lösung $y(x)$ eine Näherungslösung $\tilde{y}(x)$ mit dem absoluten Fehler $\triangle y(x) = \tilde{y}(x) - y(x)$ bzw. dem relativen Fehler $\epsilon_y(x) = \dfrac{\tilde{y}(x) - y(x)}{y(x)}$:

x	$\lvert \triangle y(x) \rvert$	$\lvert \epsilon_y(x) \rvert$
0.1	$1.9 \cdot 10^{-11}$	$2.5 \cdot 10^{-11}$
0.2	$1.5 \cdot 10^{-10}$	$2.4 \cdot 10^{-10}$
0.3	$1.1 \cdot 10^{-9}$	$3.2 \cdot 10^{-9}$
0.4	$8.1 \cdot 10^{-9}$	$8.6 \cdot 10^{-8}$
0.5	$6.0 \cdot 10^{-8}$	$6.6 \cdot 10^{-4}$
0.6	$4.4 \cdot 10^{-7}$	$4.7 \cdot 10^{-5}$
0.7	$3.3 \cdot 10^{-6}$	$9.6 \cdot 10^{-6}$
0.8	$2.4 \cdot 10^{-5}$	$3.8 \cdot 10^{-5}$
0.9	$1.8 \cdot 10^{-4}$	$2.3 \cdot 10^{-4}$
1.0	$1.3 \cdot 10^{-3}$	∞

$$\max_{x \in [0,1]} |\triangle y(x)| \approx 1.3 \cdot 10^{-3}.$$

   Hier wirkt sich deutlich der Einfluß der exponentiell anwachsenden Lösung $y(x) = c \cdot e^{20x}$ des homogenen Problems aus.

2. *Mehrzielmethode*: Um den Einfluß der exponentiell anwachsenden Lösung $y(x) = c \cdot e^{20x}$ des homogenen Problems zu mindern, wurde ein großes $m, m = 21$ gewählt

$$0 = x_1 < x_2 < \ldots < x_{21} = 1, \quad x_k = \frac{k-1}{20}.$$

3 Iterationen der Mehrzielmethode ergaben die absoluten bzw. relativen Fehler

x	$\|\triangle y(x)\|$	$\|\epsilon_y(x)\|$
0.1	$9.2 \cdot 10^{-13}$	$1.2 \cdot 10^{-12}$
0.2	$2.7 \cdot 10^{-12}$	$4.3 \cdot 10^{-12}$
0.3	$4.4 \cdot 10^{-13}$	$1.3 \cdot 10^{-12}$
0.4	$3.4 \cdot 10^{-13}$	$3.6 \cdot 10^{-12}$
0.5	$3.5 \cdot 10^{-13}$	$3.9 \cdot 10^{-9}$
0.6	$1.3 \cdot 10^{-12}$	$1.4 \cdot 10^{-11}$
0.7	$1.8 \cdot 10^{-12}$	$5.3 \cdot 10^{-12}$
0.8	$8.9 \cdot 10^{-13}$	$1.4 \cdot 10^{-12}$
0.9	$9.2 \cdot 10^{-13}$	$1.2 \cdot 10^{-12}$
1.0	$5.0 \cdot 10^{-12}$	∞

$$, \quad \max_{x \in [0,1]} |\triangle y(x)| \approx 5 \cdot 10^{-12}.$$

3. *Differenzenverfahren:* Das Verfahren liefert für die Schrittweiten $h = \frac{1}{n+1}$ folgende maximale absolute Fehler $\triangle y = \max_{0 \le i \le n+1} |\triangle y(x_i)|, x_i = i \cdot h$ (für h wurden Potenzen von 2 gewählt, damit die Matrix A (3.4.4) in der Maschine exakt berechnet werden kann):

h	$\triangle y$
2^{-4}	$2.0 \cdot 10^{-2}$
2^{-6}	$1.4 \cdot 10^{-3}$
2^{-8}	$9.0 \cdot 10^{-5}$
2^{-10}	$5.6 \cdot 10^{-6}$

.

Diese Fehler stehen im Einklang mit der Abschätzung von Satz 3.4.2. Eine Halbierung von h reduziert den Fehler auf $\frac{1}{4}$ des alten Werts. Um Fehler zu erreichen, die mit denen der Mehrzielmethode vergleichbar sind, $\triangle y \approx 5 \cdot 10^{-12}$, müßte man $h \approx 10^{-6}$ wählen!

4. *Variationsmethoden:* In der Methode wurden als Räume S Räume $Sp_\triangle$ von kubischen Splinefunktionen zu äquidistanten Einteilungen

$$\triangle : 0 = x_0 < x_1 < \ldots < x_n = 1, \quad x_i = ih, \quad h = \frac{1}{n},$$

gewählt. Folgende maximale absolute Fehler $\triangle y = ||u_s - y||_\infty$ wurden gefunden:

h	$\triangle y$
1/10	$6.0 \cdot 10^{-3}$
1/20	$5.4 \cdot 10^{-4}$
1/30	$1.3 \cdot 10^{-4}$
1/40	$4.7 \cdot 10^{-5}$
1/50	$2.2 \cdot 10^{-5}$
1/100	$1.8 \cdot 10^{-6}$

Um Fehler der Größenordnung $\triangle y \approx 5 \cdot 10^{-12}$ wie bei der Mehrzielmethode zu erreichen, müßte $h \approx 10^{-4}$ gewählt werden. Zur Berechnung allein der Bandmatrix A wären für diese Schrittweite ca. $4 \cdot 10^4$ Integration nötig!

Diese Ergebnisse zeigen eindeutig die Überlegenheit der Mehrzielmethode selbst bei einfachen linearen separierten Randwertproblemen. Differenzenverfahren und Variationsmethoden kommen selbst hier nur dann in Frage, wenn die Lösung nicht sehr genau berechnet werden muß. Um die gleiche Genauigkeit zu erreichen, hat man bei Differenzenverfahren Gleichungssysteme größerer Dimension zu lösen als bei Variationsmethoden. Dieser Vorteil der Variationsmethoden fällt jedoch erst bei kleineren Schrittweiten h ins Gewicht und wird in seiner Bedeutung erheblich dadurch eingeschränkt, daß die Berechnung der Koeffizienten der Gleichungssysteme viel aufwendiger ist als bei den einfachen Differenzenverfahren. Zur Behandlung nichtlinearer Randwertprobleme für gewöhnliche Differentialgleichungen kommt eigentlich nur die Mehrzielmethode in Frage.

3.7 Kollokationsverfahren

Beim Ritzverfahren haben wir eine Approximation $u_s \in S$ an y gesucht mittels

$$(Lu_s - f, \rho_i) = 0$$

für eine Basis $\rho_1, \ldots, \rho_n$ von $S \subset D$. Dies macht auch Sinn, wenn die Voraussetzungen des Ritzverfahrens nicht erfüllt sind. Man kann dann das sogenannte Galerkin-Verfahren anwenden. Die Idee dabei ist wie folgt: Bestimme $u_s \in S$ mit

$$(Lu_s - f, \rho_i) = 0.$$

Dabei ist

$$\begin{aligned} (\cdot, \rho_i) : C[a,b] &\to \mathbb{R} \\ v &\mapsto (v, \rho_i) \end{aligned}$$

ein lineares Funktional. Noch allgemeiner könnte man verlangen: Seien $l_1, ..., l_n$ lineare Funktionale. Finde u mit

$$l_i(Lu - f) = 0 \quad \text{für} \quad i = 1, \ldots, n.$$

Häufig sind die l_i Punktfunktionale, d.h. $l_i(u) = u(t_i), i = 1, \ldots, n$. In diesem Fall bezeichnet man diese Methode als Kollokation. Wollen wir also die Lösung y durch Kollokation approximieren, so müssen wir die folgende Aufgabe lösen: Finde $u_n(t) = \alpha_1\rho_1(t) + \ldots + \alpha_n\rho_n(t) \in S$, so daß u_n das System

$$Lu_n(t_i) - f(t_i) = \sum_{j=1}^{n} \alpha_j L\rho_j(t_i) - f(t_i) = 0$$

für $i = 1, \ldots, n$ und $t_1, \ldots, t_n$ paarweise verschieden löst. In Matrixschreibweise erhält man mit $A = (L\rho_j(t_i))_{i,j=1}^n$

$$A \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_n) \end{bmatrix}.$$

Beispiel 3.7.1 *Wir wollen*

$$\begin{aligned} Lu(t) &= u''(t) - u(t) = 1 \\ u(0) &= u(1) = 0 \end{aligned}$$

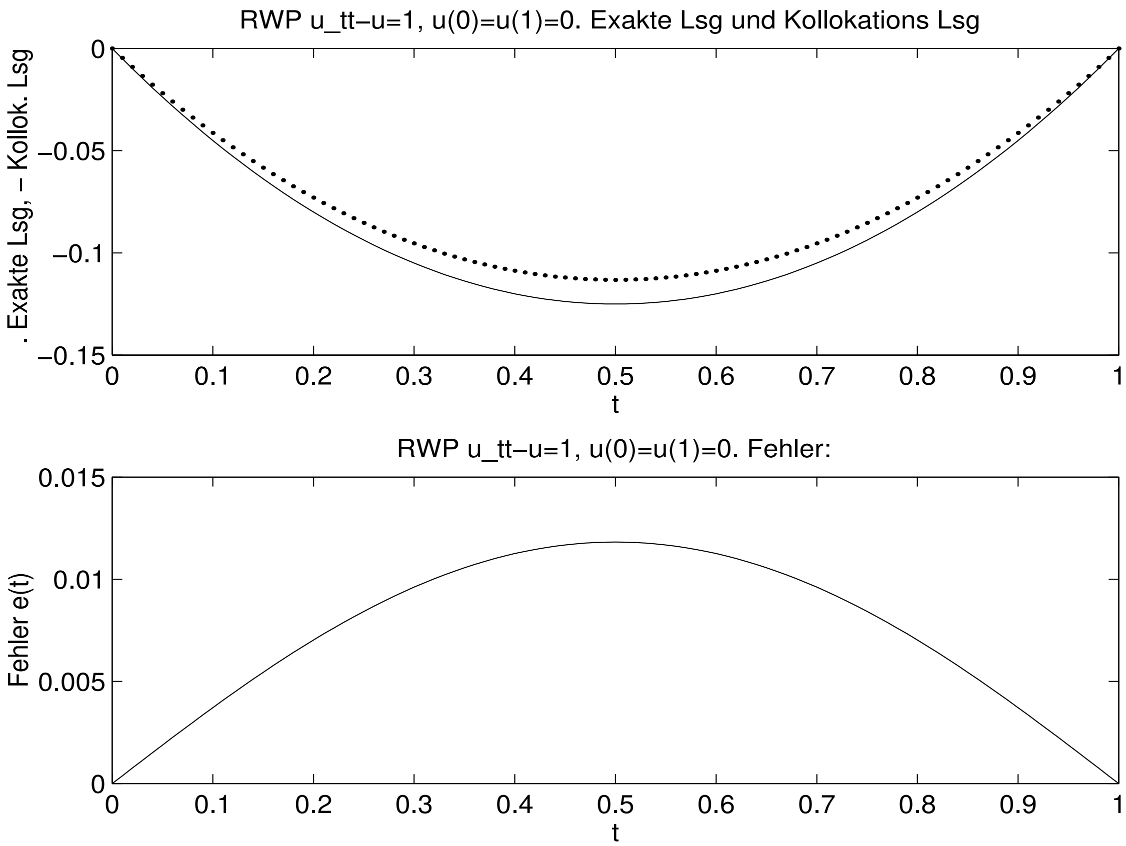
lösen. Sei $\rho_1(t) = t(t-1), \rho_2(t) = t^2(t-1)$. *Dann ist*

$$\begin{array}{lll} L\rho_1(t) = 2 - t(t-1) & L\rho_1(0) = 2 & L\rho_1(1) = 2 \\ L\rho_2(t) = 6t - 2 - t^2(t-1) & L\rho_2(0) = -2 & L\rho_2(1) = 4 \end{array}$$

Ansatz für unsere Approximation u^* ist $u^*(t) = \alpha_1\rho_1(t) + \alpha_2\rho_2(t)$. Das zu lösende Gleichungssystem lautet

$$\begin{bmatrix} 2 & -2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Es ergibt sich $\alpha_1 = \frac{1}{2}$ und $\alpha_2 = 0$ und damit $u^*(t) = \frac{1}{2}t(t-1)$. Die exakte Lösung lautet $y(t) = \frac{1}{e+1}(e^t + e^{1-t}) - 1$. Diese simple Approximation ist bereits auf eine Stelle genau.



Bemerkung 3.7.1 *Da für das zu lösende System $A\alpha = f$ die Matrix $A = (L\rho_j(t_i))$ ist, ist es auch hier günstig, Basisfunktionen mit kleinem Träger zu wählen (dann hat A eine kleine Bandbreite). Gegenüber dem Ritz- oder Galerkin- Verfahren ist hier die Matrix A viel leichter aufzustellen, da nur Auswertungen von $L\rho_j$ an den Stellen t_i erforderlich sind.*

Wir schauen uns nun die Existenz einer Kollokationsapproximation und die Fehlerschätzung an einem Beispiel an.

Beispiel 3.7.2 *Wir betrachten nun*

$$\begin{aligned} Lu &= u'' - \tau(t)u(t) = f(t) \\ u(a) &= u(b) = 0 \end{aligned}$$

mit $\tau(t) \geq 0$. *Sei*

$$\triangle : a = t_0 < t_1 < \ldots < t_n = b \quad mit \quad t_{i+1} - t_i = h = \frac{b-a}{n}$$

wieder eine Unterteilung von $[a,b]$. $Sp_\triangle$ *sei der Raum der kubischen Splines zu* $\triangle$ *mit Randwerten* $s(a) = s(b) = 0$

$$Sp_\triangle := \{s \in C^2[a,b] | s|_{[t_i,t_{i+1}]} \in \Pi_3 \quad für \quad i = 0, \ldots, n-1, s(a) = s(b) = 0\}$$

Als Basis verwendet man hier häufig sogenannte B-Splines

$$\begin{aligned}
B_i(t) &= \begin{cases} (t-t_{i-2})^3 & t_{i-2} \leq t \leq t_{i-1} \\ h^3 + 3h^2(t-t_{i-1}) + 3h(t-t_{i-1})^2 - 3(t-t_{i-1})^3 & t_{i-1} \leq t \leq t_i \\ h^3 + 3h^2(t_{i+1}-t) + 3h(t_{i+1}-t)^2 - 3(t_{i+1}-t)^3 & t_i \leq t \leq t_{i+1} \\ (t_{i+2}-t)^3 & t_{i+1} \leq t \leq t_{i+2} \\ 0 & sonst \end{cases} \\
B_i(t_{i-2}) &= B_i(t_{i+2}) = 0 \\
B_i(t_{i-1}) &= B_i(t_{i+1}) = h^3 \\
B_i(t_i) &= 4h^3
\end{aligned}$$

Hier müssen noch $B_{-1}, B_0, B_1, B_{n-1}, B_n, B_{n+1}$, *modifiziert werden, um die Randbedingungen zu erfüllen. Alternativ können wir* $B_{-1}, \ldots, B_{n+1}$ *als Basis verwenden und zusätzlich die Bedingungen* $u(a) = u(b) = 0$ *für das Gleichungssystem nehmen. Also ist* $u^* = \sum_{i=-1}^{n+1} \alpha_i B_i$ *gesucht mit*

$$\begin{aligned}
u^*(a) &= \sum_{i=-1}^{n+1} \alpha_i B_i(a) = 0 \\
Lu^*(t_j) &= \sum_{i=-1}^{n+1} \alpha_i LB_i(t_j) = f(t_j) \quad f\ddot{u}r \quad j = 0, \ldots, n \\
u^*(b) &= \sum_{i=-1}^{n+1} \alpha_i B_i(b) = 0
\end{aligned}$$

Aus der Definition der B_i *folgt*

$$\begin{aligned}
LB_i(t_j) &= B_i''(t_j) - \tau(t_j) B_i(t_j) \\
&= \begin{cases} 0 & j \leq i-2 \\ 6h - \tau(t_{i-1})h^3 & j = i-1 \\ -12h - 4\tau(t_i)h^3 & j = i \\ 6h - \tau(t_{i+1})h^3 & j = i+1 \\ 0 & j \geq i+2 \end{cases}
\end{aligned}$$

Damit ergibt sich

$$0 = \sum_{i=-1}^{n+1} \alpha_i B_i(a) = h^3(\alpha_{-1} + 4\alpha_0 + \alpha_1)$$

$$\begin{aligned}
f(t_j) &= \sum_{i=-1}^{n+1} \alpha_i LB_i(t_j) = (6h-\tau(t_j)h^3)\alpha_{j-1} - (12h+4\tau(t_j)h^3)\alpha_j + (6h-\tau(t_j)h^3)\alpha_{j+1} \\
0 &= \sum_{i=-1}^{n+1} \alpha_i B_i(b) = h^3(\alpha_{n-1}+4\alpha_n+\alpha_{n+1})
\end{aligned}$$

Das Gleichungssystem hat also die Form

$$h^3 \begin{bmatrix} 1 & 4 & 1 & & & \\ v_0 & w_0 & v_0 & & & \\ & v_1 & w_1 & v_1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & v_n & w_n & v_n \\ & & & 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \\ \alpha_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ f(t_0) \\ f(t_1) \\ \vdots \\ f(t_n) \\ 0 \end{bmatrix}$$

mit

$$\begin{aligned}
w_j &= -(\frac{12}{h^2} + 4\tau(t_j)) \\
v_j &= \frac{6}{h^2} - \tau(t_j)
\end{aligned}$$

Ein Schritt Gauss-Elimination von oben und unten liefert das System

$$h^3 \begin{bmatrix} 1 & 4 & 1 & & & & \\ 0 & w_0-4v_0 & 0 & & & & \\ & v_1 & w_1 & v_1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & v_{n-1} & w_{n-1} & v_{n-1} & \\ & & & & 0 & w_n-4v_n & 0 \\ & & & & 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \\ \alpha_n \\ \alpha_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ f(t_0) \\ f(t_1) \\ \vdots \\ f(t_{n-1}) \\ f(t_n) \\ 0 \end{bmatrix}$$

Das Subsystem

$$h^3 \begin{bmatrix} w_0-4v_0 & & & & \\ v_1 & w_1 & v_1 & & \\ & \ddots & \ddots & \ddots & \\ & & v_{n-1} & w_{n-1} & v_{n-1} \\ & & & & w_n-4v_n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \\ \alpha_n \end{bmatrix} = \begin{bmatrix} f(t_0) \\ f(t_1) \\ \vdots \\ f(t_{n-1}) \\ f(t_n) \end{bmatrix}$$

ist bis auf Multiplikation mit -1 *diagonaldominant und damit nach Gershgorin nichtsingulär, denn*

$$\begin{aligned}
h^3(w_0-4v_0) &= -12h - 4h^3\tau(t_0) - 4(6h-\tau(t_0)h^3) \\
&= -36h \\
h^3(w_n-4v_n) &= -12h - 4h^3\tau(t_n) - 4(6h-\tau(t_n)h^3) \\
&= -36h
\end{aligned}$$

und da $\tau(t) \geq 0$, *folgt* $h^3 w_j < 2h^3 v_j$. *Mit dem Subsystem ist das gesamte System nichtsingulär. Es gibt also eine eindeutige Lösung* u^* *des Kollokationsproblems.*

Satz 3.7.1 *Die Kollokationsapproximation u^* aus dem Raum der kubischen Splines zu den Knoten $a = t_0 < t_1 < \ldots < t_n = b$, an die Lösung des linearen Randwertproblems*

$$-y''(t) + \tau(t)y(t) = f(t) \quad y(a) = y(b) = 0$$

mit $\tau(t) \geq 0$ existiert. Ist $f \in C^2[a, b]$, so gilt

$$||y - u^*||_\infty \leq Mh^2$$

für hinreichend kleines h.

Bemerkung 3.7.2 *Beim Ritz-Verfahren kann man als Schranke sogar h^4 zeigen. Das ist theoretisch also besser. Praktisch sind die beiden Methoden aber etwa gleich gut. Die Kollokationsmatrix ist jedoch leichter aufzustellen als das Gleichungssystem beim Ritzverfahren.*

3.8 Zusammenfassung Randwertaufgaben

Wir haben folgende Methoden betrachtet

Schießverfahren Mehrzielmethode	im wesentlichen basierend auf der Lösung von Anfangswertaufgaben
Differenzenverfahren Variationsmethoden Kollokationsmethode	spezifisch für Randwertaufgaben auf partielle Differentialgleichungen übertragbar

Alle Ansätze führen auf lineare oder nichtlineare Gleichungssysteme, die zu lösen sind. Aussagen zur Konvergenz erhält man mittels funktionalanalytischer Methoden.

Anhang A

A.0.1 Wissenswertes über Taylor-Entwicklungen

Erinnere (aus Analysis I)

Taylorsche Formel: Sei $I \subset \mathbb{R}$ ein Intervall und $f : I \to \mathbb{R}$ eine $(n+1)$-mal stetige differenzierbare Funktion. Dann gilt für $a \in I$ und $x \in I$

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1}(x),$$

wobei

$$R_{n+1}(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)dt.$$

Lagrangesche Form des Restglieds: Sei $f : I \to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion und $a, x \in I$. Dann existiert ein ξ zwischen a und x, so daß

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}.$$

Korollar: Sei $f : I \to \mathbb{R}$ eine n-mal stetig differenzierbare Funktion und $a \in I$. Dann gilt für alle $x \in I$

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \eta(x)(x-a)^n,$$

wobei η eine Funktion mit $\lim\limits_{x\to a} \eta(x) = 0$ ist.

Wir verwenden häufig die folgende Darstellung für hinreichend oft stetig differenzierbare Funktionen f :

$$f(x+h) = f(x) + \frac{f'(x)}{1!}h + \frac{f''(x)}{2!}h^2 + \ldots + \frac{f^{(n)}(x)}{n!}h^n + O(h^{n+1}),$$

dabei bezeichnet $O(h^k)$ die Menge der Funktionen g für die ein $c \in \mathbb{R}^+$ existiert, so daß $|g(h)| \leq c \cdot h^k$ gilt (d.h. $O(h^k)$ ist die Menge der Funktionen, die nicht schneller als h^k wächst).

In den nun folgenden Beispielen wählen wir $a = 0$. Die Funktionen

$$f_1(t) := e^t, \qquad f_2(t) := sin\, t \qquad \text{und} \quad f_3(t) := cos\, t$$

sind auf ganz $\mathbb{R}$ beliebig oft differenzierbar, und zwar ist für jedes ganze $k \geq 0$

$$\begin{aligned}
f_1^{(k)}(t) &= e^t, \quad \text{also} \quad f_1^{(k)}(0) = 1;\\
f_2^{(2k)}(t) &= (-1)^k \sin t, \quad \text{also} \quad f_2^{(2k)}(0) = 0;\\
f_2^{(2k+1)}(t) &= (-1)^k \cos t, \quad \text{also} \quad f_2^{(2k+1)}(0) = (-1)^k;\\
f_3^{(2k)}(t) &= (-1)^k \cos t, \quad \text{also} \quad f_3^{(2k)}(0) = (-1)^k;\\
f_3^{(2k+1)}(t) &= (-1)^{k+1} \sin t \quad \text{also} \quad f_3^{(2k+1)}(0) = 0.
\end{aligned}$$

Aufgrund des Taylorschen Satzes erhält man also für jedes reelle x und jedes natürliche n die folgenden Darstellungen, wobei ϑ eine geeignete Zahl zwischen 0 und 1 ist, die von x, n und der betrachteten Funktion f_m abhängt:

$$\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} e^{\vartheta x},\\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^{n-1} \frac{x^{2n-1}}{(2n-1)!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \cos(\vartheta x),\\
\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + (-1)^{n-1} \frac{x^{2n-2}}{(2n-2)!} + (-1)^n \frac{x^{2n}}{(2n)!} \cos(\vartheta x).
\end{aligned}$$

Die Funktionen

$$f_4(t) := ln(1+t), \qquad f_5(t) := (1+t)^\alpha$$

sind auf $(-1, +\infty)$ beliebig oft differenzierbar, und zwar ist für jedes natürliche k

$$\begin{aligned}
f_4^{(k)}(t) &= (-1)^{k-1} \frac{(k-1)!}{(1+t)^k}, \quad \text{also} \quad \frac{f_4^{(k)}(0)}{k!} = \frac{(-1)^{k-1}}{k}\\
f_5^{(k)}(t) &= \alpha(\alpha-1)\ldots(\alpha-k+1)(1+t)^{a-k}, \quad \text{also} \quad \frac{f_5^{(k)}(0)}{k!} = \begin{pmatrix} \alpha \\ k \end{pmatrix}.
\end{aligned}$$

Für jedes reelle $x > -1$ und jedes natürliche n bestehen also die folgenden Darstellungen, wobei ϑ wieder eine geeignete Zahl zwischen 0 und 1 ist, die von x, n und der jeweils betrachteten Funktion abhängt:

$$ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n-1} \frac{x^n}{n} + (-1)^n \frac{x^{n+1}}{n+1} \frac{1}{(1+\vartheta x)^{n+1}},$$

$$(1+x)^\alpha = 1 + \begin{pmatrix} \alpha \\ 1 \end{pmatrix} x + \begin{pmatrix} \alpha \\ 2 \end{pmatrix} x^2 + \ldots + \begin{pmatrix} \alpha \\ n \end{pmatrix} x^n + \begin{pmatrix} \alpha \\ n+1 \end{pmatrix} x^{n+1} (1+\vartheta x)^{\alpha-n-1}.$$

Wir verwenden die Taylor-Entwicklung häufig, um den Wert von $f(x+h)$ durch die bekannten Werte von f und seinen Ableitungen an der Stelle x zu approximieren, z.B.

$$f(x+h) \approx f(x) + f'(x)h + \frac{1}{2} f''(x) h^2.$$

Bei diesem Beispiel machen wir einen Fehler in der Größenordnung von h^3, d.h. der Approximationsfehler beträgt hier $O(h^3)$.

In unserer Anwendung betrachten wir häufig $y' = f(t, y(t))$. Zur Berechnung der für die Taylor-Entwicklung von y benötigten Ableitungen von f muß man die Kettenregel für totale Differentiale verwenden. Daher wiederholen wir hier zunächst das Wichtigste zu totalen Differentialen und zur Kettenregel.

Erinnere (aus Analysis I):

Definition: Sei $U \subset \mathbb{R}^n$ eine offene Menge und $f : U \to \mathbb{R}$ eine reelle Funktion. f heißt im Punkt $x \in U$ <u>partiell differenzierbar</u> bzgl. der $i-$ten Koordinatenrichtung, falls der Limes

$$D_i f(x) := \lim_{h \to 0} \frac{f(x + he_i - f(x)}{h}$$

existiert. Dabei ist $e_i \in \mathbb{R}^n$ der $i-$te Einheitsvektor,

$$e_i = \underbrace{(0, \ldots, 0}_{i-1\,\text{Nullen}}, 1, 0, \ldots, 0)$$

und für den Limes $h \to 0$ hat man sich auf solche $h \in \mathbb{R}$ zu beschränken, für die $h \neq 0$ und $x + he_i \in U$. $D_i f(x)$ heißt die <u>i-te partielle Ableitung</u> von f in x. Statt $D_i f$ schreibt man auch

$$D_i f(x) = \frac{\partial f}{\partial x_i}(x) = \frac{\partial f(x)}{\partial x_i} = f_{x_i(x)}.$$

Die partiellen Ableitungen einer Funktion $f : U \to \mathbb{R}$ kann man als gewöhnliche Ableitungen von Funktionen einer Veränderlichen interpretieren: Sei $x = (x_1, \ldots, x_n) \in U$ ein fester Punkt. Für $i = 1, \ldots, n$ betrachten wir die Funktionen

$$\xi \to f_i(\xi) := f(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n).$$

Aus der Definition der partiellen Ableitung folgt

$$D_i f(x) = \lim_{h \to 0} \frac{f_i(x_i + h) - f_i(x_i)}{h} = f_i'(x_i).$$

Die partielle Ableitung bzgl. der i-ten Koordinatenrichtung ist also nichts anderes als die gewöhnliche Ableitung bzgl. der i-ten Variablen bei Festhaltung der übrigen $n - 1$ Veränderlichen. Deshalb gelten für die partiellen Ableitungen analoge Rechenregeln wie für die gewöhnlichen Ableitungen.

Betrachten wir als Beispiel die Funktion $r : \mathbb{R}^n \to \mathbb{R}$

$$r(x) := ||x|| = \sqrt{x_1^2 + \ldots + x_n^2}$$

für $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Es gilt

$$\frac{\partial}{\partial x_i}(x) = \frac{x_i}{r(x)} \quad \text{für} \quad x \neq 0.$$

Dies folgt daraus, daß die Funktion einer Variablen

$$\xi \to \sqrt{x_1^2 + \ldots + x_{i-1}^2 + \xi^2 + x_{i+1}^2 + \ldots + x_n^2}$$

differenzierbar ist. Mit der Kettenregel für Funktionen einer Veränderlichen erhält man ($x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ sind dabei als Konstante zu betrachten)

$$\begin{aligned} \frac{\partial r}{\partial x_i} &= \frac{\partial}{\partial x_i}(x_1^2 + \ldots + x_i^2 + \ldots + x_n^2)^{\frac{1}{2}} \\ &= \frac{1}{2}(x_1^2 + \ldots + x_i^2 + \ldots + x_n^2)^{-\frac{1}{2}} \cdot 2x_i \\ &= \frac{x_i}{r}. \end{aligned}$$

Sei nun $f : \mathbb{R}^+ \to \mathbb{R}$ eine beliebige differenzierbare Funktion. Dann ist die zusammengesetzte Funktion $x \to f(r(x))$, die meist kurz mit $f(r)$ bezeichnet wird, auf $\mathbb{R}^n \backslash \{0\}$ definiert und dort partiell differenzierbar. Aus der Kettenregel für Funktionen einer Veränderlichen folgt

$$\frac{\partial}{\partial x_i} f(r) = f'(r)\frac{\partial r}{\partial x_i} = f'(r)\frac{x_i}{r}.$$

In unseren Anwendungen betrachten wir häufig Funktionen $f(t, y(t))$, z.B.:

$$f(t, y(t)) = t^2 + 3ty(t).$$

Wollen wir hier die partiellen Ableitungen berechnen, so "vergessen" wir, daß y von t abhängt und betrachten

$$f(t, y) = t^2 + 3ty.$$

Dann ergibt sich

$$\begin{aligned} \frac{\partial f}{\partial t}(t, y) &= 2t + 3y \\ \frac{\partial f}{\partial y}(t, y) &= 3t. \end{aligned}$$

Also

$$\begin{aligned} \frac{\partial f}{\partial t}(t, y(t)) &= 2t + 3y(t) \\ \frac{\partial f}{\partial y}(t, y(t)) &= 3t. \end{aligned}$$

Definition Sei $U \subset \mathbb{R}^n$ eine offene Menge und $f : U \to \mathbb{R}^m$ eine Abbildung. f heißt im Punkt $x \in U$ (total) differenzierbar, falls es eine lineare Abbildung $A : \mathbb{R}^n \to \mathbb{R}^m$ gibt, so daß in einer Umgebung von x gilt

$$f(x + \xi) = f(x) + A\xi + \rho(\xi),$$

wobei ρ eine in einer Umgebung von Null definierte Funktion mit Werten in $\mathbb{R}^m$ ist mit

$$\lim_{\xi \to 0} \frac{\rho(\xi)}{||\xi||} = 0.$$

(Für $n = m = 1$ liefert dies die übliche Differenzierbarkeit von Funktionen einer Veränderlichen.)

Satz: Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^m$ eine Abbildung, die im Punkt $x \in U$ differenzierbar sei und zwar gelte

$$f(x+\xi) = f(x) + A\xi + \rho(\xi)$$

mit der Matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ und ρ wie eben. Dann gilt: Alle Komponenten $f_i : U \to \mathbb{R}, 1 \leq i \leq m$, von f sind in x partiell differenzierbar mit

$$\frac{\partial f_i}{\partial x_j}(x) = a_{ij}.$$

Man nennt A das Differential oder die Jacobi-Matrix oder auch die Funktional-Matrix von f im Punkte x, Schreibweisen:

$$(Df)(x) = J_f(x) = (\frac{\partial f_i}{\partial x_j}(x))_{i=1,\ldots,m \quad j=1,\ldots,n}$$

Wir erhalten also für $f : \mathbb{R}^n \to \mathbb{R}^m$

$$x = (x_1, \ldots, x_n) \to \begin{pmatrix} f_1(x_1,\ldots,x_n) \\ \vdots \\ f_m(x_1,\ldots,x_n) \end{pmatrix},$$

daß gilt

$$f(x+\xi) \approx f(x) + (Df)(x)\xi$$

mit

$$(Df)(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \ldots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \ldots & \frac{\partial f m(x)}{\partial x_n} \end{bmatrix}.$$

Als Beispiel betrachten wir für $n = 2, m = 3$

$$f(x) = f(x_1, x_2) = \begin{bmatrix} f_1(x_1,x_2) \\ f_2(x_1,x_2) \\ f_3(x_1,x_2) \end{bmatrix} = \begin{bmatrix} x_1 - x_2 \\ x_1 \\ x_1 x_2^2 \end{bmatrix}.$$

Dann folgt

$$(Df)(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) \\ \frac{\partial f_3}{\partial x_1}(x) & \frac{\partial f_3}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ x_2^2 & 2x_1x_2 \end{bmatrix}$$

und für ein kleines $\xi = (\xi_1, \xi_2)$

$$f(x+\xi) = f(x_1+\xi_1, x_2+\xi_2) = \begin{bmatrix} x_1 - x_2 + \xi_1 - \xi_2 \\ x_1 + \xi_1 \\ (x_1+\xi_1)(x_2+\xi_2)^2 \end{bmatrix},$$

wobei

$$\begin{aligned}(x_1+\xi_1)(x_2+\xi_2)^2 &= (x_1+\xi_1)(x_2^2+2x_2\xi_2+\xi_2^2)\\ &= x_1x_2^2+x_2^2\xi_1+2x_1x_2\xi_2+\underbrace{2x_2\xi_1\xi_2+x_1\xi_2^2+\xi_1\xi_2^2}_{\text{klein}}.\end{aligned}$$

Außerdem

$$\begin{aligned}f(x)+(Df)(x)\xi &= \begin{bmatrix} x_1-x_2\\ x_1\\ x_1x_2^2\end{bmatrix}+\begin{bmatrix}1 & -1\\ 1 & 0\\ x_2^2 & 2x_1x_2\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}\\ &= \begin{bmatrix} x_1-x_2+\xi_1-\xi_2\\ x_1+\xi_1\\ x_1x_2^2+x_2^2\xi_1+2x_1x_2\xi_2\end{bmatrix}.\end{aligned}$$

Dieses Beispiel zeigt also (wie erwartet) $f(x+\xi)\approx f(x)+(Df)(x)\xi$.

Nun zur

Kettenregel: Seien $U\subset\mathbb{R}^n$ und $V\subset\mathbb{R}^m$ offene Mengen und $g:U\to\mathbb{R}^m$ und $h:V\to\mathbb{R}^k$ Abbildungen mit $g(U)\subset V$. Die Abbildung g sei im Punkt $x\in U$ differenzierbar und die Abbildung h in Punkt $y:=g(x)$ differenzierbar. Dann ist die zusammengesetzte Abbildung

$$h\circ g:U\to\mathbb{R}^k$$

im Punkt x differenzierbar und für ihr Differential gilt

$$D(h\circ g)(x)=Dh(g(x))\cdot Dg(x).$$

Wir betrachten häufig $f(t,y(t))$. Für diesen Spezialfall sei

$$h(z_1,z_2)=f(z_1,z_2)$$

$$g(t)=\begin{bmatrix}g_1(t)\\ g_2(t)\end{bmatrix}=\begin{bmatrix}t\\ y(t)\end{bmatrix}.$$

Dann gilt für $h\circ g$:

$$\begin{aligned}(h\circ g)(t) &= h(g(t))\\ &= h(g_1(t),g_2(t))\\ &= f(g_1(t),g_2(t))\\ &= f(t,y(t)).\end{aligned}$$

Außerdem erhalten wir mit $z=(z_1,z_2)$

$$\begin{aligned}(Dh)(z) &= \begin{bmatrix}\frac{\partial f}{\partial z_1}(z), & \frac{\partial f}{\partial z_2}(z)\end{bmatrix}\\ (Dg)(t) &= \begin{bmatrix}\frac{\partial g_1}{\partial t}(t)\\ \frac{\partial g_2}{\partial t}(t)\end{bmatrix}=\begin{bmatrix}1\\ y'(t)\end{bmatrix}\end{aligned}$$

und daher

$$\begin{aligned}
(Df)(t) = D(h \circ g)(t) &= \left[\ \tfrac{\partial f}{\partial z_1}(g(t)), \tfrac{\partial f}{\partial z_2}(g(t))\ \right] \begin{bmatrix} 1 \\ y'(t) \end{bmatrix} \\
&= \frac{\partial f}{\partial z_1}(g(t)) + \frac{\partial f}{\partial z_2}(g(t))y'(t) \\
&= f_t(t, y(t)) + f_y(t, y(t)) \cdot y'(t).
\end{aligned}$$

Ganz analog läßt sich $(Df_t)(t), (Df_y)(t)$ oder z.B. $(Df_{tty})(t)$ berechnen. Allgemein gilt für differenzierbare Funktionen $h : V \to \mathbb{R}$ mit $V \subset \mathbb{R}^m$ offen und $g : U \to \mathbb{R}^m$ mit $U \subset \mathbb{R}^n$ offen und $g(U) \subset V$: $f := h \circ g : U \to \mathbb{R}$ ist partiell differenzierbar und für $i = 1, \ldots, n$

$$\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) = \sum_{j=1}^{m} \frac{\partial h}{\partial y_j}(g_1(x), \ldots, g_m(x)) \frac{\partial g_j}{\partial x_i}(x_1, \ldots, x_n).$$

Betrachte nun unsere Differentialgleichung $y' = f(t, y(t))$ und entwickle die Lösung y:

$$y(\hat{t} + h) = y(\hat{t}) + hy'(\hat{t}) + \frac{h^2}{2!}y''(\hat{t}) + \ldots + \frac{h^n}{n!}y^{(n)}(t + \theta h) \quad \text{mit} \quad 0 < \theta < 1.$$

Dabei sei nun $y(\hat{t}) = z, y'(\hat{t}) = f(\hat{t}, y(\hat{t}))$ und es ergibt sich für die Ableitungen von y

$$\begin{aligned}
y(\hat{t}) &= z \\
y'(\hat{t}) &= f(\hat{t}, z) \\
y''(\hat{t}) &= f_t(\hat{t}, z) + f_y(\hat{t}, z)f(\hat{t}, z) \\
y'''(t) &= f_{tt}(t, y(t)) + 2f_{ty}(t, y(t))y'(t) + f_{yy}(t, y(t))y'(t)^2 + f_y(t, y(t))y''(t) \\
&= f_{tt}(\hat{t}, z) + 2f_{ty}(\hat{t}, z)f(\hat{t}, z) + f_{yy}(\hat{t}, z)f(\hat{t}, z)^2 + f_t(\hat{t}, z)f_y(\hat{t}, z) + f_y(\hat{t}, z)^2 f(\hat{t}, z) \\
y''''(t) &= \frac{d}{dt}\left[f_{tt}(t, y(t)) + 2f_{ty}(t, y(t))y'(t) + f_{yy}(t, y(t))y'(t)^2 + f_y(t, y(t))y''(t)\right] \\
&= f_{ttt}(t, y(t)) + f_{tty}(t, y(t))y'(t) + 2[f_{tty}(t, y(t)) + f_{tyy}(t, y(t))y'(t)]y'(t) + 2f_{ty}(t, y(t))y''(t) \\
&\quad +[f_{tyy}(t, y(t)) + f_{yyy}(t, y(t))y'(t)]y'(t)^2 + f_{yy}(t, y(t))2y'(t)y''(t) \\
&\quad +[f_{ty}(t, y(t)) + f_{yy}(t, y(t))y'(t)]y''(t) + f_y(t, y(t))y'''(t) \\
&= f_{ttt}(t, y(t)) + 3f_{tty}(t, y(t))y'(t) + 3f_{tyy}(t, y(t))y'(t)^2 + f_{yyy}(t, y(t))y'(t)^3 + 3f_{ty}(t, y(t))y''(t) \\
&\quad +3f_{yy}(t, y(t))y'(t)y''(t) + f_y(t, y(t))y'''(t).
\end{aligned}$$

Hier nochmal in Kurzfassung das Wichtigste über das totale Differential :

Sei

$$\begin{aligned}
f : D_f \subseteq \mathbb{R}^n &\longrightarrow \mathbb{R}^m \\
(x_1, \ldots, x_n) &\mapsto \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix}
\end{aligned}$$

differenzierbar. Dann ist

$$Df(x) := \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \dots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}$$

das totale Differential von f. Es gilt in Analogie zum eindimensionalen Fall die Kettenregel: Seien $g : D_g \subset \mathbb{R}^n \longrightarrow \ R^m,\ h : D_h \subseteq g[D_g] \subseteq \mathbb{R}^m \longrightarrow \mathbb{R}^k$ differenzierbare Funktionen, dann ist

$$D(h \circ g)(x) = Dh(y)g(x)\ Dg(x),$$

d.h. das totale Differential der Komposition der beiden Funktionen ist gleich dem Matrizenprodukt der totalen Differentiale. Für unser Beispiel $\frac{d}{dt}f(t, y(t))$ ist $h \equiv f,\ g(t) = \binom{t}{y(t)}$. Somit gilt:

$$\begin{aligned} \boxed{\frac{d}{dt}f(t, y(t))} &= Df(t, z)(t, y(t))\ Dg(t) = \big(f_t(t, y(t)),\ f_y(t, y(t))\big) \begin{pmatrix} 1 \\ y'(t) \end{pmatrix} \\ &= \boxed{f_t(t, y(t)) + f_y(t, y(t))y'(t)} \end{aligned}$$

Entsprechend sind die Ableitungen für partielle Ableitungen $f_t, f_y, \dots$ von f zu bilden.

Literaturverzeichnis

[1] J.C. Butcher. On Runge-Kutta Processes of Higher Order, *J. Austr. Math. Soc.* 4 (1964) 179 - 194

[2] J.C. Butcher. On The Admissable Order of Runge-Kutta Methods, *Math. Comp.* 19 (1965) 408 - 417.

[3] R.D. Grigorieff *Numerik gewöhnlicher Differentialgleichungen 1 und 2*, Teubner Verlag

[4] E.Hairer, S.P. Nørsett und G. Wanner. *Solving Ordinary Differential Equations I*, Springer Verlag

[5] E.Hairer und G. Wanner. *Solving Ordinary Differential Equations II*, Springer Verlag

[6] H. Heuser. *Lehrbuch der Analysis, Teil 1*, Teubner Verlag

[7] H.W. Knobloch und F. Kappel. *Gewöhnliche Differentialgleichungen*, Teubner Verlag

[8] J. Stoer und R. Bulirsch. *Einführung in die Numerische Mathematik II*, Springer Verlag

[9] W. Walter. *Gewöhnliche Differentialgleichungen*, Springer Verlag