

Numerik gewöhnlicher Differentialgleichungen

P. Spellucci

SS2007

Inhaltsverzeichnis

1	Das Anfangswertproblem gewöhnlicher DGLen	4
1.1	Einführung · Theoretischer Hintergrund ERG	4
1.2	Einige elementare Diskretisierungsverfahren	9
1.3	Allgemeine Theorie der Einschrittverfahren	14
1.4	Praktisch wichtige Verfahren	23
1.4.1	Taylorverfahren	23
1.4.2	Runge–Kutta–Verfahren	24
1.5	Implementierungsfragen · Schrittweitensteuerung, Fehlerschätzer	26
1.5.1	Lösung der impliziten Gleichungen	27
1.5.2	Schrittweitensteuerung und Schätzung des globalen Diskretisierungsfehlers	27
1.5.3	Kontinuierliche Näherungsformeln	43
1.6	Mehrschrittverfahren. Motivation und Herleitung einiger elementarer Verfahren	45
1.6.1	Verfahren, die auf numerischer Integration beruhen	45
1.6.2	Verfahren, die auf numerischer Differentiation beruhen	46
1.7	Lineare Differenzgleichungen mit konstanten Koeffizienten	47
1.8	Allgemeine Theorie der Mehrschrittverfahren	53
1.9	Prädiktor–Korrektor–Methoden	73
1.10	Adams–Bashforth–Moulton Prädiktor–Korrektor–Verfahren	76
1.11	Steife Differentialgleichungen	78
1.11.1	Einführung	78
1.11.2	Lineare Stabilität von Diskretisierungsverfahren für Anfangswertprobleme	83

1.11.3	Nichtlineare Stabilität*	90
1.11.4	Implizite R–K– und Rosenbrock–Verfahren*	93
1.12	Andere Problemstellungen bzw. Methoden ERG	101
2	Rand– und Eigenwertaufgaben gewöhnlicher Differentialgleichungen	104
2.1	Einführungsbeispiel. Einige theoretische Grundlagen. ERG	104
2.2	Das Schießverfahren und das Mehrfachschießverfahren	112
2.3	Elementare Differenzenverfahren	119
2.4	Kompakte Differenzenschemata für spezielle Randwertaufgaben zweiter Ordnung. ERG	136
2.5	Behandlung von Grenzschichten	138
2.6	Kollokationsmethoden	139
2.7	Variationsmethoden. Finite Elemente in einer Dimension	143
2.8	Sturm Liouville’sche Eigenwertprobleme (ERG)	161
2.9	Zusammenfassung	167
3	Software	168

Dieses Skriptum stellt den Inhalt der Vorlesung in einer sehr knappen, sicher nicht buchreifen Form dar. Es soll nicht das Studium der einschlägigen Lehrbücher ersetzen. Für Hinweise auf Fehler, unklare Formulierungen, wünschenswerte Ergänzungen etc. bin ich jederzeit dankbar. Man bedenke jedoch den Zeitrahmen der Veranstaltung, der lediglich 14 Doppelstunden umfasst, weshalb der eine oder andere Punkt wohl etwas zu kurz kommt oder auch einmal ganz wegfallen muss. Abschnitte, die im Kleindruck erscheinen, insbesondere eher technische Beweise, werden in der Vorlesung nicht vorgetragen. Sie sind aber für einen interessierten Leser zur Arbeitsvereinfachung hier aufgenommen worden. Diese Abschnitte sind durch eine Sequenz aus << und >> eingeklammert und ausserdem eingerückt, um die Orientierung zu erleichtern. Das Gleiche gilt für die mit ”ERG” gekennzeichneten Abschnitte.

Der vorläufige Vorlesungsplan ist der folgende:

V1 : 1.2;

V2 : 1.3;

V3 : 1.4, 1.5 bis Schätzung des globalen Fehlers,

V4 : Schrittweitensteuerung,

V5 : MSV Konstruktion und Differenzgleichungen,

V6 : MSV allgemeine Theorie bis zur 1. Dahlquist Schranke,

V7 : Absolute Stabilität, Prädiktor-Korrektor-Verfahren, ABM mit variabler Schrittweite und Ordnung

V8 : Steife DGLen, lineare und nichtlineare Stabilität

V9 : Implizite Runge-Kutta- und Rosenbrock-Verfahren

V10 : RWA Schieß-Verfahren

V11 : RWA Differenzenverfahren

V12 : Behandlung von Grenzschichten, lokale Kollokation

V13 : FEM 1D Theorie

V14 : FEM Praxis

Es wird empfohlen, vor der Vorlesung die entsprechenden Textteile durchzulesen, um in der Vorlesung gezielt Verständnisfragen klären zu können und nach der Vorlesung die ergänzenden Teile zu lesen. Als Ergänzung zur Vorlesung sollte eines der empfohlenen Fachbücher hinzugenommen werden.

Kapitel 1

Das Anfangswertproblem gewöhnlicher DGLen

1.1 Einführung · Theoretischer Hintergrund ERG

Wir betrachten hier die Anfangswertaufgabe für ein System (in der Regel nichtlinearer) Differentialgleichungen 1. Ordnung

$$(A) \quad \begin{cases} y' = f(t, y), & f : \mathcal{I} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \\ y(t_0) = y_0 & t_0 \in \mathcal{I} \end{cases}$$

wobei \mathcal{I} ein eigentliches oder uneigentliches reelles Intervall bedeutet. Auf Einschränkungen des Definitionsbereichs von f bezüglich der Variablen y wird aus Gründen der formalen Vereinfachung verzichtet. Es gelte stets

$$(V0) \quad \begin{cases} f \in C(\mathcal{I} \times \mathbb{R}^n), \\ \exists L \geq 0 : \quad \forall t \in \mathcal{I}, \quad \forall y_1, y_2 \in \mathbb{R}^n : \\ \quad \|f(t, y_1) - f(t, y_2)\| \leq L\|y_1 - y_2\| \end{cases}$$

Bekanntlich gilt:

Satz 1.1.1. *Unter der Voraussetzung (V0) besitzt (A) genau eine Lösung $y \in C^1(\mathcal{I})$. Falls darüberhinaus gilt:*

$$f \in C^p \quad (\mathcal{I} \times \mathbb{R}^n)$$

mit $p \geq 1$, dann ist $y \in C^{p+1}(\mathcal{I})$. □

Neben Systemen 1. Ordnung spielen in den Anwendungen auch vielfach Differentialgleichungen höherer Ordnung

$$y^{(m)} = f(t, y, y', \dots, y^{(m-1)})$$

eine Rolle. Bekanntlich kann man diese durch die Substitution

$$u_i := y^{(i-1)} \quad i = 1, \dots, m, \quad w := (u_1, \dots, u_m)^T$$

$$F(t, w) := \begin{pmatrix} u_2 \\ \vdots \\ u_m \\ f(t, u_1, \dots, u_m) \end{pmatrix}$$

in das nm -System 1. Ordnung

$$w' = F(t, w)$$

überführen. In der Regel ist diese Vorgehensweise auch für die numerische Lösung zu empfehlen. Im Falle der speziellen Gleichung

$$y^{(m)} = f(t, y),$$

insbesondere

$$y'' = f(t, y),$$

ist jedoch die Anwendung spezieller Methoden oft besser.

Im Zusammenhang mit Fehlerfragen und mit der (Mehr)-Zielmethode ist auch die Frage nach der Art der Abhängigkeit der Lösung von (A) vom Anfangswert y_0 zu behandeln. Bekanntlich gilt:

Satz 1.1.2. Über (V0) hinaus gelte: In $\mathcal{S} := \mathcal{I} \times \mathbb{R}^n$ existiere $f_y(t, y)$,¹ sei stetig und beschränkt:

$$\|f_y(t, y)\| \leq L \quad \forall (t, y) \in \mathcal{S}.$$

Dann hängt die Lösung $y(t; s)$ der AWA $y' = f(t, y)$, $y(t_0) = s$ ($t_0 \in \mathcal{I}$) stetig differenzierbar vom Anfangswert s ab. Es ist

$$Z(t; s) := y_s(t; s)$$

die Lösung der linearen Anfangswertaufgabe

$$Z'(t; s) = f_y(t, y(t; s))Z(t; s), \quad Z(t_0; s) = I$$

□

Anwendung der Taylor'schen Formel liefert unter den Vor. von Satz 1.1.2

$$y(t; s_1) - y(t; s_2) = \int_0^1 Z(t; s_2 + \tau(s_1 - s_2)) d\tau (s_1 - s_2)$$

also

$$\|y(t; s_1) - y(t; s_2)\| \leq \int_0^1 \|Z(t; s_2 + \tau(s_1 - s_2))\| d\tau \|s_1 - s_2\|$$

Die Norm von Z kann man mit Hilfe des folgenden Satzes abschätzen:

Satz 1.1.3. Sei $A : \mathcal{I} \rightarrow \mathbb{R}^{n \times n}$ stetig auf \mathcal{I} und $Y \in \mathbb{R}^{n \times n}$ Lösung der AWA

$$Y' = A(t)Y, \quad Y(t_0) = I \quad \text{mit} \quad t_0 \in \mathcal{I}.$$

Dann gilt

$$\|Y(t) - I\| \leq \exp\left(\int_{t_0}^t \alpha(\tau) d\tau\right) - 1$$

mit $\alpha(t) := \|A(t)\|$, falls $t_0 \leq t$.

□

<<

Beweis: Es gilt

$$Y(t) - \underbrace{Y(t_0)}_{=I} = \int_{t_0}^t Y'(\tau) d\tau = \int_{t_0}^t A(\tau)(Y(\tau) - I + I) d\tau$$

somit

$$\|Y(t) - I\| \leq \int_{t_0}^t \alpha(\tau)(\|Y(\tau) - I\| + 1) d\tau$$

oder mit $\varphi(t) := \|Y(t) - I\|$

$$\varphi(t) \leq \int_{t_0}^t \alpha(\tau)(\varphi(\tau) + 1) d\tau.$$

Wir setzen

$$\gamma(t) := \frac{\int_{t_0}^t \alpha(\tau)(\varphi(\tau) + 1) d\tau + 1}{\exp\left(\int_{t_0}^t \alpha(\tau) d\tau\right)}.$$

Dann wird

$$\begin{aligned} \gamma'(t) &= \alpha(t)(\varphi(t) + 1) \exp\left(-\int_{t_0}^t \alpha(\tau) d\tau\right) - \\ &\quad - \left(\int_{t_0}^t \alpha(\tau)(\varphi(\tau) + 1) d\tau + 1\right) \alpha(t) \exp\left(-\int_{t_0}^t \alpha(\tau) d\tau\right) \\ &= \alpha(t) \left(\varphi(t) - \int_{t_0}^t \alpha(\tau)(\varphi(\tau) + 1) d\tau\right) \exp\left(-\int_{t_0}^t \alpha(\tau) d\tau\right) \leq 0. \end{aligned}$$

Somit $\gamma(t) \leq \gamma(t_0) = 1$ für $t \geq t_0$ d.h.

$$\exp\left(\int_{t_0}^t \alpha(\tau) d\tau\right) - 1 \geq \int_{t_0}^t \alpha(\tau)(\varphi(\tau) + 1) d\tau \geq \varphi(t) \quad \text{q.e.d.}$$

□

>>

Bemerkung 1.1.1. In der Praxis treten oft Differentialgleichungen auf, bei denen die Matrix $f_y(t, y)$ nur Eigenwerte mit negativem Realteil besitzt (d.h. $\|Z\|$ klingt schnell ab), während zumindest einige Eigenwerte sehr großen Betrag haben (d.h. notwendig $\alpha(t) \gg 1$). In solchen Fällen ist die Aussage von Satz 1.1.3 praktisch wenig hilfreich, denn wegen $\alpha(t) \leq L$ folgt daraus letztlich nur

$$\|y(t; s_1) - y(t; s_2)\| \leq e^{L|t-t_0|} \|s_1 - s_2\|$$

und es ist $L \gg 1$, d.h. die rechte Seite der Abschätzung wird sehr groß. Es gibt für diesen Fall bessere Abschätzungen, siehe dazu die Spezialliteratur. \square

Bemerkung 1.1.2. In der Praxis ist die globale Lipschitzbeschränktheit von f bezüglich y fast nie gegeben. Dann existiert bei differenzierbarem bzw. in y lokal Lipschitzstetigen f eine eindeutige Lösung der DGL auch nur lokal. Beim numerischen Rechnen gehen wir jedoch davon aus, daß die Existenz der zu berechnenden Lösung bereits anderweitig gesichert ist.

Die obige Abschätzung liefert die Sensitivität der Lösung gegenüber Störungen im Anfangswert. Allgemeiner kann man nach Abschätzungen der Differenz $\|y(t) - z(t)\|$ fragen, wenn

$$y' = f(t, y), \quad y(t_0) = y_0 \tag{1.1}$$

und

$$z' = f(t, y) + r(t), \quad z(t_0) = y_0 + r_0 \tag{1.2}$$

Hierzu gilt

Satz 1.1.4. Sei

$$\|r_0\| \leq \varepsilon$$

und

$$\sup_{t \in \mathcal{I}} \|r(t)\| \leq \varepsilon$$

Dann gilt für die Lösungen von (1.1) und (1.2)

$$\|z(t) - y(t)\| \leq (1 + |t - t_0|) \exp(L(t - t_0)) \varepsilon$$

\square

Dieser Satz wird mit dem **Gronwall Lemma** bewiesen

Lemma 1.1.1. Sei $g : \mathcal{I} \rightarrow \mathbb{R}$ stetig und monoton nicht fallend sowie $\alpha : \mathcal{I} \rightarrow \mathbb{R}$ nichtnegativ und stetig sowie φ stetig auf \mathcal{I} . Wenn

$$\varphi(t) \leq g(t) + \int_{t_0}^t \alpha(\tau) \varphi(\tau) d\tau$$

dann ist für $t_0, t \in \mathcal{I}$

$$\varphi(t) \leq g(t) \exp\left(\int_{t_0}^t \alpha(\tau) d\tau\right).$$

\square

Zum Beweis siehe z.B. Quarteroni,A.,Valli,A.:Numerical Approximation for partial differential equations. Springer 1994. Lemma 1.4.1.

Eine häufig auftretende Aufgabe ist die der Parameteridentifizierung bei gewöhnlichen Differentialgleichungen aus gemessenen Werten der Lösung. Das einfachste Beispiel dazu ist die Identifizierung der Parameter λ und y_∞ in

$$y' = \lambda y + y_\infty$$

aus Daten (t_i, y_i) , der sogenannte Exponentialfit. Zur Parameterabhängigkeit solcher Lösungen ist bekannt

Satz 1.1.5. Sei $f : \mathcal{I} \times \mathbb{R}^n \times \mathcal{U} \rightarrow \mathbb{R}^n$ k mal stetig differenzierbar nach y und u und

$$y' = f(t, y, u), \quad y(t_0) = y_0 \quad u \in \mathcal{U} \text{ fest.}$$

Dann ist $y = y(t, u)$ bezüglich t und u k -mal stetig differenzierbar und es ist

$$y_u(t)' = f_y(t, y, u)y_u + f_u(t, y, u), \quad y_u(t_0) = y_0.$$

□

Hier eine kleine Aufstellung vertiefender Lehrbücher zu Kapitel 1:

1. Butcher, J.C.:*Numerical Methods for Ordinary Differential Equations* 2nd ed. Wiley: 2003
2. Deuffhard, P. und Bornemann, F.A.: *Numerische Mathematik II. Gewöhnliche Differentialgleichungen.* 2nd revised ed. Walter de Gruyter: Berlin 2002
3. Grigorieff, R.D. : *Numerik gewöhnlicher Differentialgleichungen I,II.* Teubner: Stuttgart 1972 bzw. 1977
4. Haier, E., Nørsett , S.P.; Wanner,G.: *Solving ordinary differential equations I. Nonstiff problems.* 2nd ed. Springer : Berlin 1993
5. Hairer, E., Wanner, G.: *Solving ordinary differential equations II. Stiff and differential algebraic equations.* 2nd ed. Springer: Berlin 1996
6. Strehmel, K.; Weiner, R.; *Numerik gewöhnlicher Differentialgleichungen* Stuttgart: Teubner 1995

1.2 Einige elementare Diskretisierungsverfahren

Wir wollen unsere Betrachtungen an einigen sehr einfachen numerischen Verfahren, deren Eigenschaften sich leicht herleiten lassen, beginnen und alle auftretenden Begriffe daran veranschaulichen. (In der Praxis werden diese Verfahren wegen ihrer geringen Effizienz natürlich nicht angewendet.) Wir beginnen mit dem vom Beweis des Peano'schen Existenzsatzes her bekannten

Euler-Verfahren:

$$(A_h) \quad \begin{cases} \eta_0 := y_0 \\ \eta_{i+1} := \eta_i + hf(t_0 + ih, \eta_i) \quad i = 0, 1, \dots, N_h - 1, \end{cases}$$

wo wir noch zur Vereinfachung ein äquidistantes **Gitter**

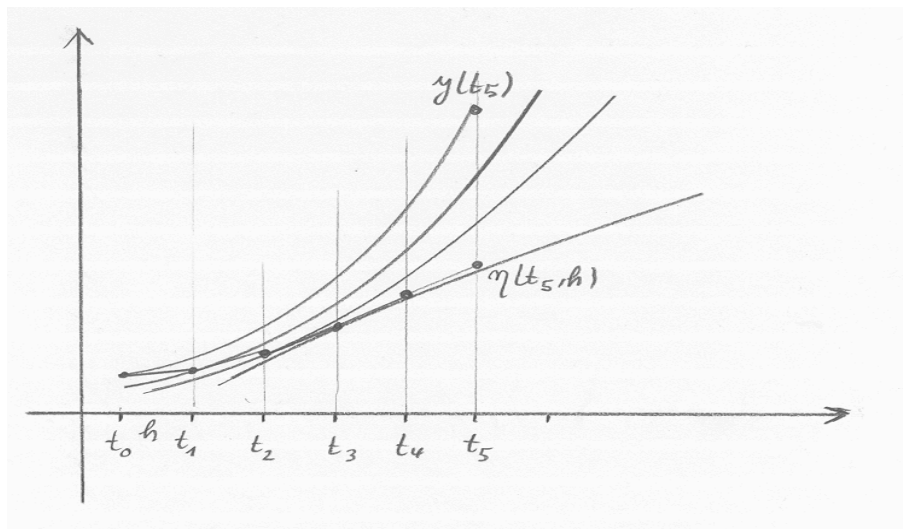
$$G_h = \{t_i : t_i = t_0 + ih, \quad 0 \leq i \leq N_h\}$$

angenommen haben. Durch das Verfahren (A_h) wird eine **Gitterfunktion** $\eta(\cdot; h) : G_h \rightarrow \mathbb{R}^n$ definiert, wo $\eta(t_i; h) := \eta_i$ gesetzt ist. Die stückweise linear Interpolierende dieser Gitterfunktion können wir dann als Approximation für die Lösung y von (A) auf $[t_0, t_0 + N_h h] \subset \mathcal{I}$ nehmen.

Dabei ist

$$t_0 + N_h h = t_E \in \mathcal{I} \quad \text{fest, d.h.} \quad h = \frac{t_E - t_0}{N_h}, \quad N_h \in \mathbb{N}.$$

Wir erwarten mit $N_h \rightarrow \infty$ (d.h. $h \rightarrow 0$) Konvergenz von $\eta(t_E; h)$ gegen $y(t_E)$ für bel. $t_E \in \mathcal{I}$. Wenn diese Konvergenz gleichmäßig in t_E ist, wobei $t_E \in \mathcal{I}_0$, \mathcal{I}_0 kompakt, dann ist dies gleichbedeutend damit, daß die Folge der linear Interpolierenden der Gitterfunktionen gleichmäßig gegen die Lösung y von (A) auf \mathcal{I}_0 konvergiert (wobei natürlich die Gitter \mathcal{I}_0 überdecken müssen.)



Um im Folgenden die Konvergenzbeweise möglichst einfach gestalten zu können, werden wir in der Regel stärkere Voraussetzungen formulieren, als eigentlich erforderlich

ist. Wir definieren F^k als Menge aller Funktionen, deren sämtliche partielle Ableitungen bis zur Ordnung k im betrachteten Steifen existieren, stetig und beschränkt sind:

$$\begin{aligned}
 F^k(\mathcal{I} \times \mathbb{R}^n) = & \{f \in C^k(\mathcal{I} \times \mathbb{R}^n) : \forall \mu, \nu, \\
 & 0 \leq \mu + \nu \leq k : \left\| \frac{\partial^{\nu+\mu}}{\partial^\nu t \partial^\mu y} f(t, y) \right\| \leq C_{\nu+\mu} \\
 & (\forall (t, y) \in \mathcal{I} \times \mathbb{R}^n)\} \tag{1.3}
 \end{aligned}$$

Satz 1.2.1. Sei $f \in F^1(\mathcal{I} \times \mathbb{R}^n)$, $\mathcal{I}_0 \subset \mathcal{I}$ kompakt und $\mathcal{I}_0 = [t_0, t_E]$.
Dann gilt für das Euler-Verfahren

$$(1) \quad \|\eta(t; h) - y(t)\| \leq (e^{(t_E-t_0)C_1} - 1)(C_0 + 1)h \quad t \in [t_0, t_E] \cap G_h$$

Falls sogar $f \in F^2(\mathcal{I} \times \mathbb{R}^n)$, dann gilt sogar eine asymptotische Entwicklung des globalen Diskretisierungsfehlers bezüglich h :

$$(2) \quad \eta(t; h) = y(t) + hz(t) + \mathcal{O}(h^2) \quad t \in [t_0, t_E] \cap G_h$$

wobei z die Lösung der linearen AWA

$$(3) \quad z' = f_y(t, y(t))z - \frac{1}{2}y''(t), \quad z(t_0) = 0$$

ist. (y Lösung von (A)) □

Beweis: Sei

$$\begin{aligned}
 \varepsilon_i & := \eta(t_i; h) - y(t_i), \quad i = 0, \dots, N, \\
 \delta_i & := y(t_i) + hf(t_i, y(t_i)) - y(t_{i+1}) \quad i = 0, \dots, N-1.
 \end{aligned}$$

ε_i bezeichnen wir als den globalen und δ_i als den lokalen Diskretisierungsfehler. Wegen

$$0 = \eta_i + hf(t_i, \eta_i) - \eta_{i+1}$$

gilt

$$-\delta_i = \varepsilon_i + h(f(t_i, \eta_i) - f(t_i, y(t_i))) - \varepsilon_{i+1}$$

also unter Anwendung der Lipschitzstetigkeit von f im zweiten Argument

$$\|\varepsilon_{i+1}\| \leq (1 + hC_1)\|\varepsilon_i\| + \|\delta_i\|$$

Taylorentwicklung liefert die Abschätzung für δ_i

$$\begin{aligned}\delta_i &= \int_0^h (f(t_i, y(t_i)) - f(t_i + \tau, y(t_i + \tau))) d\tau \\ &= \int_0^h (f(t_i, y(t_i)) - f(t_i, y(t_i + \tau)) + f(t_i, y(t_i + \tau)) - \\ &\quad f(t_i + \tau, y(t_i + \tau))) d\tau \\ \|\delta_i\| &\leq h^2 C_1 C_0 + h^2 C_1 = h^2 (C_0 + 1) C_1.\end{aligned}$$

Wegen $\varepsilon_0 = 0$ folgt

$$\begin{aligned}\|\varepsilon_N\| &\leq \frac{(1 + hC_1)^N - 1}{hC_1} h^2 (C_0 + 1) C_1 \\ &= \left(1 + \frac{hC_1 N}{N}\right)^N - 1 (C_0 + 1) h \\ &\leq (e^{hC_1 N} - 1) (C_0 + 1) h.\end{aligned}$$

Mit $hN = t - t_0$ folgt die erste Behauptung. Verfeinerung der Taylorentwicklung für δ_i liefert

$$\begin{aligned}\delta_i &= y(t_i) + hy'(t_i) - (y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(t_i) + \\ &\quad + \frac{h^3}{2} \int_0^1 (1 - \tau)^2 y^{(3)}(t_i + \tau h) d\tau) \\ &= -\frac{h^2}{2}y''(t_i) - \frac{h^3}{6}M_3\vartheta_i, \quad \|\vartheta_i\|_\infty \leq 1, \quad M_3 = \max_{t \in \mathcal{I}_0} \|y^{(3)}(t)\|_\infty.\end{aligned}$$

Also

$$\begin{aligned}\varepsilon_{i+1} &= \varepsilon_i + hf_y(t_i, y(t_i))\varepsilon_i - \frac{h^2}{2}y''(t_i) \\ &\quad + h \int_0^1 (1 - \tau) f_{yy}(t_i, y(t_i) + \tau\varepsilon_i) \varepsilon_i d\tau - \frac{h^3}{6}M_3\vartheta_i \\ &= (I + hf_y(t_i, y(t_i)))\varepsilon_i - \frac{h^2}{2}y''(t_i) + \mathcal{O}(h^3)\end{aligned}$$

Sei

$$\begin{aligned}e_i &:= \frac{\varepsilon_i}{h} \quad \text{und} \\ \tilde{e}_{i+1} &:= \tilde{e}_i + h(f_y(t_i, y(t_i)))\tilde{e}_i - \frac{1}{2}y''(t_i) \\ &\quad (\text{Euler-Verfahren für (3)})\end{aligned}$$

Nach Resultat (1) gilt $\|e_i\| \leq K_2$, K_2 geeignet, unabh. von i und h , und nach Def. von \tilde{e}_i

$$\|\tilde{e}_i - e_i\| \leq K_3 h, \quad K_3 \text{ geeignet, } i = 0, \dots, N.$$

Es ist aber wiederum nach (1)

$$\|\tilde{e}_i - z(t_i)\| \leq K_4 h, \quad K_4 \text{ geeignet}$$

Damit ist (2) bewiesen. □

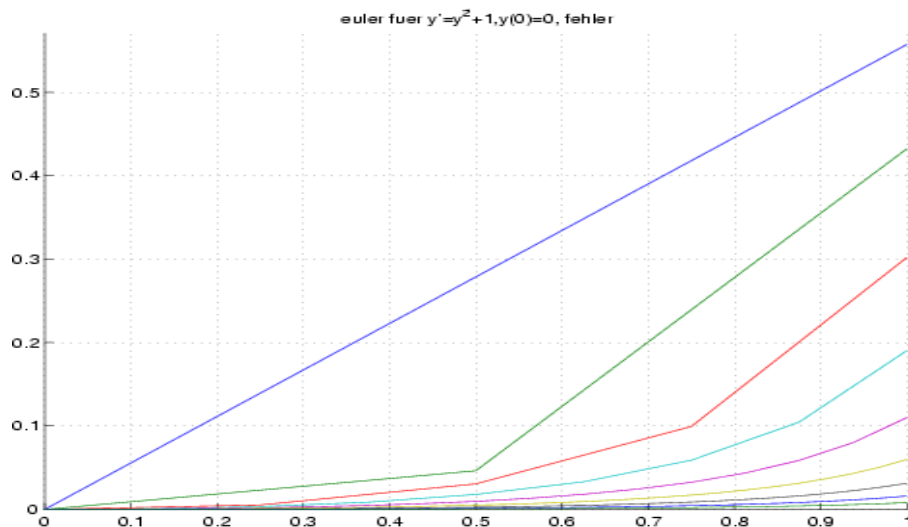
Gleichung (1) besagt, daß der **globale Diskretisierungsfehler**

$$\eta(t; h) - y(t)$$

mit h linear gegen null geht. Man spricht auch von einem **Verfahren erster Ordnung**. Der globale Diskretisierungsfehler hat eine um 1 kleinere h -Ordnung als der **lokale Diskretisierungsfehler**

$$\delta(t, z; h) := -y(t+h) + y(t) + hf(t, y(t)), \quad y(t) = z$$

der **entsteht, wenn man die exakte Lösung von (A) in die Verfahrensgleichung einsetzt**, wobei als Anfangswert an der Stelle t gerade y genommen wird. Aus (1) folgt weiter, daß der globale Diskretisierungsfehler mit wachsendem $t_E - t_0$ (höchstens) exponentiell anwächst. Die folgende Darstellung zeigt die Entwicklung des globalen Diskretisierungsfehlers des Eulerverfahrens bei fortgesetzter Schrittweitenhalbierung:



Die Entwicklung (2) heißt **asymptotische Entwicklung des globalen Diskretisierungsfehlers**. Sie liefert eine sehr viel feinere Aussage über das tatsächliche Fehlerverhalten als (1). Betrachten wir z.B. den Fall

$$n = 1, \quad f(t, y) = \lambda y, \quad (\text{d.h. } y(t) = y_0 e^{\lambda(t-t_0)})$$

dann wird

$$z' = \lambda z - \frac{1}{2} \lambda^2 y(t), \quad z(t_0) = 0$$

d.h.

$$z(t) = -\frac{1}{2} \lambda^2 (t - t_0) y(t)$$

$$\eta(t; h) = y(t) - \frac{h}{2} \lambda^2 (t - t_0) y(t) + \mathcal{O}(h^2)$$

Der **relative** Fehler in $\eta(t; h)$ bleibt also klein, solange $\mathcal{O}(h^2) \ll |y(t)|$ und $h(t-t_0)\lambda^2 \ll 1$, auch wenn $\lambda < 0$, d.h. man erhält also sehr kleine absolute Fehler für $\lambda < 0$. Formel (2) hat auch unmittelbare praktische Anwendungen. Es gilt für $f \in F^2$

$$\eta(t; h) = y(t) + hz(t) + \mathcal{O}(h^2)$$

Also definiert

$$\hat{\eta}_i := 2\eta(t_{2i}; \frac{h}{2}) - \eta(t_i; h)$$

ein **Verfahren zweiter Ordnung**. (Dies ist ein Spezialfall der sogenannten Richardsonextrapolation). Bezeichnet man $\hat{\eta}_{i+1}$ wieder mit η_{i+1} , dann ergibt sich das Verfahren

$$\eta_{i+1} = \eta_i + hf(t_i + \frac{h}{2}, \eta_i + \frac{h}{2}f(t_i, \eta_i))$$

(**modifiziertes Euler-Verfahren**). Dieses Verfahren kann man direkt aus der der DGL zugeordneten Volterra'schen Integralgleichung

$$y(t) = y(t_i) + \int_{t_i}^t f(\tau, y(\tau))d\tau \tag{1.4}$$

herleiten, wenn man $t = t_{i+1}$ setzt, das Integral nach der Rechteckregel annähert und den dabei auftretenden Wert $y(t_i + \frac{h}{2})$ durch seine Näherung aus dem Eulerverfahren ersetzt. Wendet man auf (1.4) die **Trapezregel** an, dann erhält man die Verfahrensvorschrift

$$\eta_{i+1} = \eta_i + \frac{1}{2}h(f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1}))$$

Dies ist nun eine **implizite Berechnungsvorschrift** für η_{i+1} (falls f nichtlinear von y abhängt) und man muß dann in jedem Schritt ein lineares bzw. nichtlineares Gleichungssystem lösen. Ersetzt man auf der rechten Seite η_{i+1} durch die Eulernäherung $\eta_i + hf(t_i, \eta_i)$, dann erhält man das

Verfahren von Heun:

$$\eta_{i+1} = \eta_i + \frac{1}{2}h(f(t_i, \eta_i) + f(t_{i+1}, \eta_i + hf(t_i, \eta_i)))$$

Auch die Trapezregel und das Verfahren von Heun besitzen globale Diskretisierungsfehler der Ordnung $\mathcal{O}(h^2)$.

Neben dem Fehlerverhalten der einzelnen Verfahren für $h \rightarrow 0$ interessiert auch das Verhalten der Gitterfunktion $\eta(t; h)$ bei festem $h > 0$ für $t \rightarrow \infty$, wenn z.B. die stationären Zustände eines durch die DGL beschriebenen physikalischen Systems gesucht sind. Oft besteht der instationäre Anteil der Lösung aus exponentiell abklingenden Anteilen und wir wollen deshalb das Verhalten der Gitterfunktionen für die AWA ($n = 1$)

$$\begin{aligned} y' &= \lambda y & \Re\lambda < 0 \\ y(0) &= y_0 \end{aligned}$$

untersuchen.

Das Eulerverfahren liefert die Gitterfunktion

$$\eta(t_i; h) = (1 + h\lambda)^i \eta_0$$

und $\eta(t_i; h) \rightarrow 0$ für $i \rightarrow \infty$ ist äquivalent mit der Bedingung $|1 + h\lambda| < 1$.
 Diese Bedingung heißt **Bedingung der absoluten Stabilität** (für das Euler–Verfahren).
 Für das modifizierte Eulerverfahren und das Heunverfahren erhält man

$$\eta(t_i; h) = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right)^i \eta_0,$$

Für die Trapezregel dagegen

$$\eta(t_i; h) = \left(\frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}}\right)^i \eta_0$$

Die Bedingung der absoluten Stabilität ist also in diesen Fällen

$$\left|1 + h\lambda + \frac{(h\lambda)^2}{2}\right| < 1$$

bzw.

$$\left|1 + \frac{h\lambda}{2}\right| < \left|1 - \frac{h\lambda}{2}\right|.$$

Für $\Re\lambda < 0$ ist die **letzte Bedingung stets erfüllt**, man nennt deshalb die Trapezregel **A–stabil**. Wenn $\Re\lambda < 0$ und $|\Re\lambda| \gg 1$ bedeutet die Bedingung der absoluten Stabilität für das Euler– bzw. das Heunverfahren eine sehr starke Einschränkung von h . Auf diese Problematik werden wir noch zurückkommen.

Bei einem linearen **Differentialgleichungssystem** $y' = Ay$ mit A diagonalisierbar und $\Re\lambda_i < 0$ für alle Eigenwerte von A muß entsprechend die Stabilitätsbedingung für jeden Eigenwert erfüllt sein. Wenn die DGL selbst keine exponentiell abklingenden Lösungen besitzen, macht diese Betrachtungsweise natürlich keinen Sinn.

1.3 Allgemeine Theorie der Einschrittverfahren

Die bisher besprochenen elementaren Lösungsmethoden haben alle folgende allgemeine Form: **Einschrittverfahren (ESV)** mit fester Schrittweite

$$(A_h) \quad \begin{cases} \eta_0 := y_{0,h} , \\ \eta_{i+1} := \eta_i + h\Phi(\eta_i, \eta_{i+1}, t_i, h; f) \quad i = 0, \dots, N_h - 1 . \end{cases}$$

Ist Φ unabh. von η_{i+1} : dann hat man ein **explizites** Verfahren, sonst ein **implizites** Verfahren.

Der Wert η_{i+1} wird allein mit Hilfe des vorausberechneten Wertes η_i (und natürlich der DGL, d.h. f) berechnet. Deshalb heißen diese Verfahren **Einschrittverfahren, ESV**. Die **Schrittfunktion** Φ soll dabei dabei folgende Voraussetzungen erfüllen (bei obigen Beispielen ist dies auf Grund der Voraussetzungen an f gegeben):

$$(V1) \quad \begin{aligned} &\Phi \in C(\mathbb{R}^n \times \mathbb{R}^n \times \mathcal{I} \times [0, h_0]) \quad \text{mit } h_0 > 0 \\ &\exists L_1, L_2 : \quad \forall y_1, y_2 \quad z_1, z_2 \in \mathbb{R}^n, \quad \forall t \in \mathcal{I}, \quad \forall h \in [0, h_0] : \\ &\quad \|\Phi(y_1, z_1, t, h; f) - \Phi(y_2, z_2, t, h; f)\| \leq L_1 \|y_1 - y_2\| + L_2 \|z_1 - z_2\|. \end{aligned}$$

Normalerweise wird $y_{0,h} := y_0$ gewählt sein, doch können auch h -abhängige Fehler in y_0 zugelassen werden. Diese Betrachtungsweise ist z.B. sinnvoll, wenn man die später noch zu besprechenden Mehrschrittverfahren formal als ESV für Systeme schreibt. Aufgrund der Bedingung (V1) ist zumindest für $h_0 L_2 < 1$ η_{i+1} durch die Verfahrensvorschrift eindeutig definiert (Banach'scher Fixpunktsatz), sodaß man auch schreiben könnte:

$$\eta_{i+1} = \eta_i + h\tilde{\Phi}(\eta_i, t_i, h; f)$$

mit einer implizit definierten Funktion $\tilde{\Phi}$. Dies wird in der Literatur auch häufig so getan. Da jedoch alle Verfahrenseigenschaften aus Φ direkt abgeleitet werden können, ziehen wir diese Darstellung vor.

Definition 1.3.1. Das ESV (A_h) heißt **konvergent**, falls für jede Aufgabe (A) mit (V0) gilt:

$$\eta(t; h) - y(t) \rightarrow 0 \quad \text{für} \quad h \rightarrow 0 \quad \text{und} \quad y_{0,h} \rightarrow y_0$$

wo $t \in \mathcal{I}$ und $h = (t - t_0)/N$, $N \in \mathbb{N}$, $N \rightarrow \infty$ □

Definition 1.3.2. Das ESV (A_h) heißt **stabil**, falls für jede AWA (A) mit (V0) gilt: $\exists h_0 > 0$, $K \geq 0$:

$$0 < h \leq h_0, \quad t_0 \leq t \leq t_E \Rightarrow \|\tilde{\eta}(t; h) - \eta(t; h)\| \leq K(\|\tilde{\eta}(t_0; h) - \eta(t_0; h)\| + \rho)$$

worin $\tilde{\eta}$ die Lösung von

$$\tilde{\eta}_{i+1} = \tilde{\eta}_i + h\Phi(\tilde{\eta}_i, \tilde{\eta}_{i+1}, t_i, h; f) + hr(t_i, h), \quad \tilde{\eta}_0 := \tilde{\eta}(t_0; h)$$

mit $\|r(t_i, h)\| \leq \rho$ ist. □

Stabilität der das ESV beschreibenden Differenzgleichung bedeutet also u.a., daß die Gitterfunktion lipschitzstetig vom Anfangswert abhängt.

Satz 1.3.1. Es gelte (V1). Dann ist das ESV (A_h) stabil. □

<<

Beweis: Sei $h_0 L_2 < 1$ und $h \leq h_0$. Dann ist η_{i+1} bzw. $\tilde{\eta}_{i+1}$ nach dem Banach'schen Fixpunktsatz eindeutig bestimmt. Wir erhalten mit (V1)

$$\|\tilde{\eta}_{i+1} - \eta_{i+1}\| \leq \|\tilde{\eta}_i - \eta_i\| + h\rho + hL_1\|\tilde{\eta}_i - \eta_i\| + hL_2\|\tilde{\eta}_{i+1} - \eta_{i+1}\|,$$

also

$$\|\tilde{\eta}_{i+1} - \eta_{i+1}\| \leq \frac{1}{1 - hL_2}((1 + hL_1)\|\tilde{\eta}_i - \eta_i\| + h\rho)$$

Sei

$$K_1 := \frac{1 + hL_1}{1 - hL_2}, \quad K_2 := \frac{h\rho}{1 - hL_2}, \quad \varepsilon_{i+1} := K_1\varepsilon_i + K_2, \quad \varepsilon_0 := \|\tilde{\eta}_0 - \eta_0\|.$$

Dann gilt ersichtlich $\varepsilon_i \geq \|\tilde{\eta}_i - \eta_i\|$ ($\forall i$). Aber

$$\varepsilon_i = K_1^i \varepsilon_0 + \left(\sum_{j=0}^{i-1} K_1^j \right) K_2 = K_1^i \varepsilon_0 + \frac{K_1^i - 1}{K_1 - 1} K_2$$

Aber

$$K_1 = 1 + \frac{h(L_1 + L_2)}{1 - hL_2}$$

und daher

$$\begin{aligned} K_1^N &= \left(1 + \frac{(t - t_0)(L_1 + L_2)}{1 - hL_2} \cdot \frac{1}{N} \right)^N \\ &\leq e^{(t-t_0)(L_1+L_2)/(1-hL_2)} \\ &\leq e^{(t_E-t_0)(L_1+L_2)/(1-h_0L_2)} =: K_3 \end{aligned}$$

und dies ergibt

$$\begin{aligned} \|\tilde{\eta}_N - \eta_N\| &= \|\tilde{\eta}(t; h) - \eta(t; h)\| \\ &\leq K_3 \|\tilde{\eta}(t_0; h) - \eta(t_0; h)\| + (K_3 - 1) \cdot \frac{1}{L_1 + L_2} \rho \end{aligned}$$

d.h. die Behauptung mit $K := \max\{K_3, (K_3 - 1)/(L_1 + L_2)\}$ □

>>

Es bedeutet in (A_h)

$\eta_N = \eta(t; h)$ eine Näherung für $y(t)$

$\eta_{N-1} = \eta(t - h; h)$ eine Näherung für $y(t - h)$

und man wird erwarten, daß für $h \rightarrow 0$ nicht nur $\eta_N \rightarrow y(t)$ und

$\eta_{N-1} \rightarrow y(t)$, sondern auch

$$\frac{\eta_N - \eta_{N-1}}{h} \rightarrow y'(t) = \lim_{h \rightarrow 0} \frac{y(t) - y(t - h)}{h}.$$

Unter der Voraussetzung (V1) bedeutet dies, daß $\Phi(y, y, t, 0; f) = f(t, y)$ gelten muß:

Definition 1.3.3. *Das ESV (A_h) heißt konsistent, falls*

$$\Phi(y, y, t, 0; f) = f(t, y) \quad \forall t \in \mathcal{I}, \quad y \in \mathbb{R}^n,$$

und $\exists \lim_{h \rightarrow 0} y_{0,h} = y_0$. □

Satz 1.3.2. *Unter der Voraussetzung (V1) ist (A_h) genau dann konvergent, wenn es konsistent ist.* □

<<

Beweis: Sei

$$g(t, y) := \Phi(y, y, t, 0; f)$$

Wegen (V1) erfüllt g (V0), sodaß die AWA

$$z' = g(t, z), \quad z(t_0) = y_0$$

genau eine Lösung $z \in C^1(\mathcal{I})$ besitzt. Wir zeigen nun, daß

$$\max_{t \in \mathcal{I}_0} \|z(t) - \eta(t; h)\| \rightarrow 0$$

für $h \rightarrow 0$ und jedes kompakte $\mathcal{I}_0 := [t_0, t_E] \subset \mathcal{I}$. Es gilt

$$\begin{aligned} \eta_{i+1} &= \eta_i + h\Phi(\eta_i, \eta_{i+1}, t_i, h; f) \\ &= \eta_i + \int_{t_i}^{t_{i+1}} \Phi(\eta_i, \eta_{i+1}, t_i, h; f) d\tau \\ z_{i+1} &= z_i + \int_{t_i}^{t_{i+1}} \Phi(z(\tau), z(\tau), \tau, 0; f) d\tau \end{aligned}$$

somit mit $\varepsilon_i := \eta_i - z_i$

$$\begin{aligned} \varepsilon_{i+1} &= \varepsilon_i + \int_{t_i}^{t_{i+1}} (\Phi(\eta_i, \eta_{i+1}, t_i, h; f) - \Phi(z(\tau), z(\tau), \tau, 0; f)) d\tau \\ &= \varepsilon_i + \int_{t_i}^{t_{i+1}} (\Phi(\eta_i, \eta_{i+1}, t_i, h; f) - \Phi(z(\tau), \eta_{i+1}, t_i, h; f) \\ &\quad + \Phi(z(\tau), \eta_{i+1}, t_i, h; f) - \Phi(z(\tau), z(\tau), t_i, h; f) \\ &\quad + \Phi(z(\tau), z(\tau), t_i, h; f) - \Phi(z(\tau), z(\tau), \tau, h; f) \\ &\quad + \Phi(z(\tau), z(\tau), \tau, h; f) - \Phi(z(\tau), z(\tau), \tau, 0; f)) d\tau. \end{aligned}$$

Sei

$$M := \max_{t \in \mathcal{I}_0} \|g(t, z(t))\|.$$

Dann wird wegen

$$\begin{aligned} \|z_i - z(\tau)\|, \quad \|z_{i+1} - z(\tau)\| &\leq hM \quad \text{für } \tau \in [t_i, t_{i+1}] \\ \|\varepsilon_{i+1}\| &\leq \|\varepsilon_i\| + hL_1\|\varepsilon_i\| + hL_2\|\varepsilon_{i+1}\| + \\ &\quad + h^2M(L_1 + L_2) + h(\omega_3(\Phi; h) + \omega_4(\Phi; h)) \end{aligned}$$

wobei $\omega_j(\Phi; \cdot)$ den Stetigkeitsmodul von Φ bzgl. des j^{ten} Arguments bezeichnet auf $\mathbb{R}^n \times \mathbb{R}^n \times \mathcal{I}_0 \times [0, h_0]$. Also wird

$$\|\varepsilon_{i+1}\| \leq \frac{1 + hL_1}{1 - hL_2} \|\varepsilon_i\| + \frac{1}{1 - hL_2} h(hM(L_1 + L_2) + \omega_3(\Phi; h) + \omega_4(\Phi; h)).$$

Somit wegen $\varepsilon_0 = 0$, nach der gleichen Beweistechnik wie in Satz 1.3.1

$$\begin{aligned} \|\varepsilon_N\| &\leq \left(e^{(t_E - t_0)(L_1 + L_2)/(1 - h_0 L_2)} - 1 \right) \cdot \\ &\quad \cdot \frac{1}{L_1 + L_2} \cdot (hM(L_1 + L_2) + \omega_3(\Phi; h) + \omega_4(\Phi; h)) \end{aligned}$$

d.h. $\varepsilon_N \rightarrow 0$ mit $h \rightarrow 0$.

Ist das Verfahren also konsistent, dann wegen $g(t, y) = f(t, y)$ $z(t) \equiv y(t)$ und damit wegen $\|\eta(t; h) - y(t)\| \rightarrow 0$ auch konvergent.

Ist andererseits das Verfahren konvergent, dann muß wegen

$\eta(t; h) - y(t) \rightarrow 0$ und $\eta(t; h) - z(t) \rightarrow 0$ $z \equiv y$ sein auf \mathcal{I}_0 . Wir nehmen nun an, daß

$$f(\tilde{t}, \tilde{y}) \neq g(\tilde{t}, \tilde{y}) \quad \text{für ein } (\tilde{t}, \tilde{y}) \in \mathcal{I}_0 \times \mathbb{R}^n.$$

Dann gilt für die Lösung der AWA

$$y' = f(t, y), \quad y(\tilde{t}) = \tilde{y}, \quad z' = g(t, z), \quad z(\tilde{t}) = \tilde{y}, \quad y'(\tilde{t}) \neq z'(\tilde{t})$$

während nach obigem Beweis $y \equiv z$ (Widerspruch!) □

>>

Um die Konvergenzgüte der Verfahren beurteilen zu können, führen wir folgende Begriffe ein:

Definition 1.3.4. Sei $z \in C^1(\mathcal{I})$ die Lösung der AWA

$$z' = f(t, z), \quad z(t) = z_t .$$

Die Größe

$$\delta(t, z_t; h) = \eta_1 - z(t+h), \quad \eta_1 = z_t + h\Phi(z_t, \eta_1, t, h; f)$$

mit $t, t+h \in G_h \cap \mathcal{I}$ heißt der **lokale Diskretisierungsfehler** von (A_h) an der Stelle (t, z_t) und

$$\tau(t, z; h) = \begin{cases} \frac{1}{h}\delta(t-h, z; h) & \text{für } t \in G_h \setminus \{t_0\} \\ y_{0,h} - y_0 & \text{für } t = t_0 \end{cases}$$

der **lokale Abschneidefehler** des Verfahrens.

Das Verfahren heißt **konsistent von der Ordnung p** , falls

$$\sup_{t \in G_h \cap \mathcal{I}_0} \|\tau(t, y(t-h); h)\| \leq K_1 h^p \quad \text{für } h \leq h_0$$

Es heißt **konvergent von der Ordnung p** , falls

$$\|y_{0,h} - y_0\| \leq K_2 h^p \quad \text{und} \quad \|\eta(t; h) - y(t)\| \leq K_3 h^p$$

($\forall h \leq h_0, \quad \forall t \in \mathcal{I}_0 \cap G_h$)

Dabei ist \mathcal{I}_0 ein kompaktes Teilintervall von \mathcal{I} und $t_0 \in \mathcal{I}_0$. □

Dazu gilt

Satz 1.3.3. Es gelte (V1). Dann ist die Konsistenzordnung gleich der Konvergenzordnung. □

Beweis: Nach Voraussetzung ist

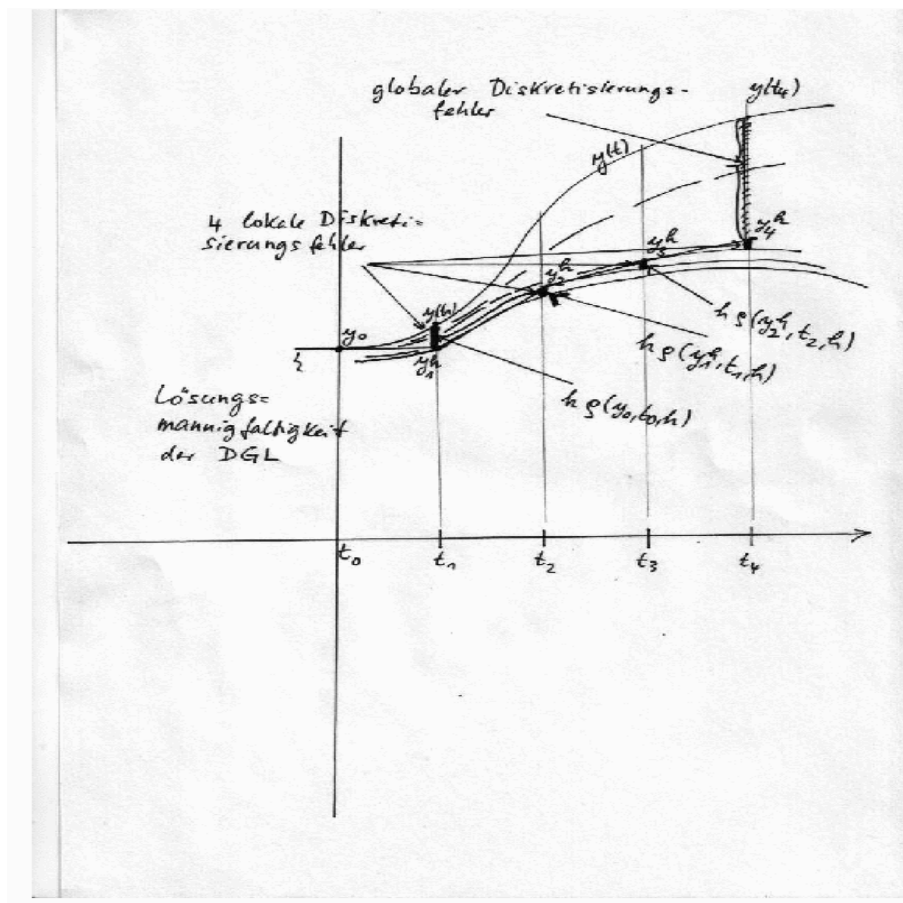
$$\begin{aligned} \eta_{i+1} &= \eta_i + h\Phi(\eta_i, \eta_{i+1}, t_i, h; f) \\ y_{i+1} &= y_i + h\Phi(y_i, y_{i+1}, t_i, h; f) - h\tau_{i+1}; \\ \delta_i &= \delta(t_i, y(t_i); h) \end{aligned}$$

also nach der gleichen Schlußweise wie in Satz 1.3.1

$$\begin{aligned} \|y_N - \eta_N\| &\leq K \underbrace{\left(\|y_0 - \eta_0\| + \sup_{t \in G_h \setminus \{t_0\}} \|\tau(t, y(t-h); h)\| \right)}_{\sup_{t \in G_h} \|\tau(t; h)\|} \\ &\leq K(K_2 + K_3)h^p \end{aligned}$$

□

Die folgende Skizze zeigt den Zusammenhang zwischen dem lokalen Diskretisierungsfehler (hier mit $h\varrho(\dots)$ bezeichnet) und dem globalen Diskretisierungsfehler: der globale Diskretisierungsfehler hängt ab vom lokalen Diskretisierungsfehler und den Stabilitätseigenschaften der Differentialgleichung:



Die Konsistenz- und damit auch die Konvergenzordnung von (A_h) hängt natürlich auch von den Regularitätseigenschaften der rechten Seite der DGL, d.h. f , ab. Man definiert weiter als Verfahrensordnung die Ordnung, die für alle beliebig glatten rechten Seiten f erreicht wird, formal

Definition 1.3.5. Die Größe

$$p^* := \max\{p \in \mathbb{N} : \|\tau(t, y(t-h); h)\| \leq Kh^p \\ (\forall h \leq h_0, \quad \forall t \in G_h \setminus \{t_0\}, \quad \forall f \in F^\infty(\mathcal{I} \times \mathbb{R}^n))\}$$

heißt die **Verfahrensordnung** von (A_h) . □

Die Verfahrensordnung hängt allein von der Konstruktion von Φ ab. Man ermittelt sie durch geeignete Taylorentwicklung, wobei man für f beliebig hohe Differenzierbarkeitseigenschaften annehmen kann.

Die Aussage von Satz 1.3.3 ist noch nicht sonderlich befriedigend, da die Konstante K exponentiell mit $t_E - t_0$ anwächst. Eine erheblich verfeinerte Aussage liefert

Satz 1.3.4. *Sei (V1) erfüllt und (A_h) konsistent von der Ordnung p . Ferner gelte*

$$\Phi(., ., ., .; f) \in C^2(\mathbb{R}^n \times \mathbb{R}^n \times \mathcal{I} \times [0, h_0]).$$

Der lokale Diskretisierungsfehler $\delta(t, y; h)$ besitze eine Darstellung der Form

$$\delta(t, y, h) = h^{p+1}\psi(t, y) + c(t; h; y; \Phi)h^{p+2}$$

wobei $\psi(t, y) \in C^1(\mathcal{I} \times \mathbb{R}^n)$ und $\|c(t; h; y; \Phi)\|$ beschränkt ist auf

$$\mathcal{I}_0 \times [0, h_0] \times \{z : \|y(t) - z\| \leq k_0 h_0^p \quad \forall t \in \mathcal{I}_0\} .$$

*Dann besitzt der globale Diskretisierungsfehler die Darstellung (**asymptotische Entwicklung**)*

$$\eta(t; h) - y(t) = h^p z(t) + \tilde{c}(t; h; y; \Phi)h^{p+1}$$

wobei $\|\tilde{c}(t; h; y; \Phi)\|$ beschränkt ist auf $\mathcal{I}_0 \times [0, h_0]$ und z die Lösung der (linear inhomogenen) AWA

$$z' = f_y(t, y(t))z + \psi(t, y(t)), \quad z(t_0) = \frac{y_{0,h} - y_0}{h^p}$$

ist. ($\mathcal{I}_0 = [t_0, t_E] \subset \mathcal{I}$)

<<

Beweis: Nach dem Hauptsatz über implizite Funktionen ist die Gleichung

$$F(y, z, t, h) = y - z - h\Phi(z, y, t, h; f) = 0$$

für $0 \leq h \leq h_0$ und h_0 hinreichend klein für beliebiges $(z, t, h) \in \mathbb{R}^n \times \mathcal{I}_0 \times [0, h_0]$ eindeutig auflösbar nach y , d.h. mit einer implizit definierten Funktion $\tilde{\Phi}$ ist

$$y = z + h\tilde{\Phi}(z, t, h; f)$$

und die Funktion $\tilde{\Phi}$ ist ebenfalls zweimal stetig differenzierbar bzgl. z, t, h . ($\tilde{\Phi}(z, t, h; f) = \Phi(z, y(z, t, h), t, h; f)$)

Dies ergibt die Darstellung

$$\eta_{i+1} = \eta_i + h\tilde{\Phi}(\eta_i, t_i; h; f)$$

mit einer zweimal stetig differenzierbaren Funktion $\tilde{\Phi}$. Sei nun $\varepsilon_i := \eta_i - y_i$. Dann ergibt

sich nach Definition von δ und ε (mit $y_i = y(t_i)$)

$$\begin{aligned}
\varepsilon_{i+1} &= \varepsilon_i + h(\tilde{\Phi}(\eta_i, t_i, h; f) - \tilde{\Phi}(y_i, t_i, h; f)) + \delta(t_i, y_i; h) + \mathcal{O}(h^{p+2}) \\
&= \varepsilon_i + h(\tilde{\Phi}(y_i + \varepsilon_i, t_i, h; f) - \tilde{\Phi}(y_i, t_i, h; f)) + \delta(t_i, y_i; h) + \mathcal{O}(h^{p+2}) \\
&= \varepsilon_i + h\tilde{\Phi}_y(y_i, t_i, h; f)\varepsilon_i + h \int_0^1 (1 - \tau)\tilde{\Phi}_{yy}(y_i + \tau\varepsilon_i, t_i, h; f)\varepsilon_i\varepsilon_i d\tau + \\
&\quad + \delta(t_i, y_i; h) + \mathcal{O}(h^{p+2}) \\
&= (I + h\tilde{\Phi}_y(y_i, t_i, 0; f))\varepsilon_i + h^{p+1}\psi(t_i, y_i) + \mathcal{O}(h^{p+2}) + \\
&\quad + h^2 \int_0^1 \tilde{\Phi}_{yh}(y_i, t_i, \tau h; f) d\tau \varepsilon_i + h \int_0^1 (1 - \tau)\tilde{\Phi}_{yy}(y_i + \tau\varepsilon_i, t_i, h; f)\varepsilon_i\varepsilon_i d\tau
\end{aligned}$$

Denn wegen

$$\begin{aligned}
\delta(t_i, y_i; h) &= y_i + h\Phi(y_i, y_{i+1}, t_i, h; f) - y_{i+1} \\
&= y_i + h\Phi(y_i, y_{i+1}, t_i, h; f) - y_{i+1} - (y_i + h\Phi(y_i, \hat{\eta}_{i+1}, t_i, h; f) - \hat{\eta}_{i+1}) \\
&= (I + \mathcal{O}(h))(\hat{\eta}_{i+1} - y_{i+1})
\end{aligned}$$

ist

$$\hat{\eta}_{i+1} - y_{i+1} = (I + \mathcal{O}(h))\delta(t_i, y_i; h) \quad \text{wobei} \quad \hat{\eta}_{i+1} := y_i + h\tilde{\Phi}(t_i, y_i, h; f).$$

Wegen $\varepsilon_i = \mathcal{O}(h^p)$ (Satz 1.3.3) und $p \in \mathbb{N}$ sind die letzten drei Terme alle $\mathcal{O}(h^{p+2})$. Also mit

$$\begin{aligned}
\tilde{\varepsilon}_i &:= \varepsilon_i / h^p \\
\tilde{\varepsilon}_{i+1} &= (I + h\tilde{\Phi}_y(y_i, t_i, 0; f))\tilde{\varepsilon}_i + h\psi(t_i, y_i) + h^2\sigma_i
\end{aligned}$$

mit $\|\sigma_i\| \leq \hat{\sigma} \quad \forall i \in \{0, \dots, N_h - 1\}, \quad 0 < h \leq h_0$.

Nun ist aber $\tilde{\Phi}_y(y, t, 0; f) = f_y(t, y)$ und $\|f_y(t, y)\| \leq L$.

Definieren wir also

$$\hat{\varepsilon}_{i+1} = (I + hf_y(t_i, y_i))\hat{\varepsilon}_i + h\psi(t_i, y_i), \quad \hat{\varepsilon}_0 = \tilde{\varepsilon}_0$$

dann gilt

1. $\|\hat{\varepsilon}_N - \tilde{\varepsilon}_N\| \leq K_1 \hat{\sigma} h$
2. $\|\hat{\varepsilon}_N - z(t)\| \leq K_2 h$

mit geeigneten Konstanten K_1, K_2 (vgl. Beweis von Satz 1.2.1 und 1.3.1)

Also wird

$$\varepsilon_N = \eta(t; h) - y(t) = h^p z(t) + \mathcal{O}(h^{p+1}) \quad q.e.d.$$

□

>>

Wie am Beispiel des Euler-Verfahrens bereits erläutert, kann die asymptotische Entwicklung des globalen Diskretisierungsfehlers sowohl zur Schätzung des lokalen (bzw. globalen) Fehlers als auch zur Konstruktion von Verfahren höherer Ordnung dienen. Ist etwa

$$\eta(t; h) = y(t) + h^p z(t) + c(t; h; y; \Phi)h^{p+1}$$

dann

$$\eta(t; \frac{h}{2}) = y(t) + 2^{-p} h^p z(t) + c(t; \frac{h}{2}; y; \Phi) 2^{-p-1} h^{p+1}$$

somit

$$\frac{2^p \eta(t; \frac{h}{2}) - \eta(t; h)}{2^p - 1} = y(t) + \mathcal{O}(h^{p+1}) \quad (1.5)$$

$$h^p z(t) = \frac{\eta(t; h) - \eta(t; \frac{h}{2})}{1 - 2^{-p}} + \mathcal{O}(h^{p+1}) \quad (1.6)$$

Neben dieser Schätzung (1.6) für den globalen Diskretisierungsfehler (rechnet man nur von t nach $t+h$, (und faßt dabei $\eta(t; h)$ als exakt auf), dann ist dies zugleich der lokale Diskretisierungsfehler) kann man die Formel (1.5) zur Konstruktion von **Verfahren der Konsistenz- und Konvergenzordnung $p+1$** benutzen, (**falls** $y_{0,h} - y_0 = \mathcal{O}(h^{p+1})$), und zwar auf zwei verschiedene Weisen:

A) globale Extrapolation, passive Extrapolation

(auf die Schrittweite $h=0$)

Berechne $\eta(t_i; h)$ $i = 1, \dots, N_h$.

Berechne $\eta(\tilde{t}_i; \frac{h}{2})$ $i = 1, \dots, 2N_h$ ($\tilde{t}_{2i} = t_i$ $i = 0, \dots, N_h$)

Setze $\hat{\eta}(t_i; h) := \frac{2^p \eta(\tilde{t}_{2i}; \frac{h}{2}) - \eta(t_i; h)}{2^p - 1}$ $i = 1, \dots, N_h$

(Extrapolation nachträglich)

B) lokale Extrapolation, aktive Extrapolation

$$\left. \begin{aligned} \hat{\eta}_0 &:= \eta_0 \\ \eta_{i+1} &:= \hat{\eta}_i + h\Phi(\hat{\eta}_i, \eta_{i+1}, t_i, h; f) \\ \tilde{\eta}_{i+\frac{1}{2}} &:= \hat{\eta}_i + \frac{h}{2}\Phi(\hat{\eta}_i, \tilde{\eta}_{i+\frac{1}{2}}, t_i, \frac{h}{2}; f) \\ \tilde{\eta}_{i+1} &:= \tilde{\eta}_{i+\frac{1}{2}} + \frac{h}{2}\Phi(\tilde{\eta}_{i+\frac{1}{2}}, \tilde{\eta}_{i+1}, t_i + \frac{h}{2}, \frac{h}{2}; f) \\ \hat{\eta}_{i+1} &:= (2^p \tilde{\eta}_{i+1} - \eta_{i+1}) / (2^p - 1) \end{aligned} \right\} i = 0, 1, \dots, N_h - 1$$

(Extrapolierte Werte werden als Startwerte für den nächsten Schritt verwendet)

Man würde zunächst vermuten, daß die aktive Extrapolation bessere Resultate liefert als die passive. Dies ist jedoch nicht immer der Fall. So definiert die Trapezregel mit passiver Extrapolation ein A -stabiles Verfahren der Ordnung 4, (h beliebig) (vgl. S. 11), während aktive Extrapolation ein nicht A -stabiles Verfahren ergibt.

Bemerkung 1.3.1. *Das Verfahren der Richardson-Extrapolation ist bekanntlich besonders effizient einzusetzen, wenn eine asymptotische Entwicklung hoher Ordnung für den Fehler existiert:*

$$\eta(t; h) = y(t) + \sum_{k=0}^m \psi_k(t) h^{p+k} + \mathcal{O}(h^{p+m+1})$$

(insbesondere, wenn zusätzlich ψ_k für $k+p$ ungerade verschwinden). Bei Grigorieff (Bd.I) wird gezeigt, daß solche Entwicklungen möglich sind, wenn Φ entsprechend oft differenzierbar ist und $\delta(t, y(t); h)$ eine Entwicklung gleicher Bauart besitzt. Eine Entwicklung nach geraden Potenzen von h tritt z.B. bei der Trapezregel auf. \square

Bemerkung 1.3.2. In der Praxis ist man aus Aufwandsgründen nicht daran interessiert, ein ESV tatsächlich mit äquidistanter Schrittweite zu rechnen. Vielmehr wird man versuchen, die Gitterbreite in Abhängigkeit vom Verhalten der Lösung zu wählen. Mangels anderer brauchbarer Kriterien bedient man sich hierbei der Technik, in t_i h_i so zu wählen, daß

$$\|\delta(t_i, \eta_i; h_i)\| \leq \varepsilon \quad \text{bzw.} \quad \|\delta(t_i, \eta_i; h_i)\| \leq \varepsilon \|\eta_i\| \quad (1.7)$$

wird. Dabei wird δ mit geeigneten Methoden geschätzt, z.B. mit der oben beschriebenen Richardsonextrapolation. Man erhält dann ein Verfahren

$$\begin{aligned} \eta_0 &:= y_{0,h} \\ \eta_{i+1} &:= \eta_i + h_i \Phi(\eta_i, \eta_{i+1}, t_i, h_i; f) \\ \text{mit } t_{i+1} &= t_i + h_i \end{aligned} \quad \sum_{i=0}^{N-1} h_i = t_E - t_0.$$

Die Sätze 1.3.1, 1.3.2 und 1.3.3 bleiben dann gültig, wenn man h als die maximal auftretende Schrittweite interpretiert.

Die Aufrechterhaltung von Satz 1.3.4 erfordert die Konstruktion

$$h_i = h(1 - h\gamma(t_i, h)), \quad h = \sup h_i$$

mit einer in $\mathcal{I}_0 \times [0, h_0]$ beschränkten Funktion γ , d.h. für $h \rightarrow 0$ konstante Schrittweite. Diese Konstruktion entspricht **nicht** der in der Praxis üblichen Vorgehensweise, wo man in der Regel systematische Schrittweitenverdoppelung bzw. -halbierung anwendet, um die Forderung (1.7) zu erfüllen. Man kann dann die auf Satz 1.3.4 beruhenden Techniken nur anwenden, wenn man die Schrittweite **stückweise konstant hält**. \square

1.4 Praktisch wichtige Verfahren

1.4.1 Taylorverfahren

Mit

$$y' = f(t, y), \quad f \in C^{p+1}(\mathcal{I} \times \mathbb{R}^n)$$

ist

$$y^{(\nu+1)}(t) = \frac{d^\nu}{dt^\nu} f(t, y(t)), \quad \nu = 0, \dots, p+1.$$

Für

$$\Phi(y, z, t, h; f) \stackrel{\text{def}}{=} \sum_{k=0}^p \frac{h^k}{(k+1)!} \frac{d^k}{dt^k} f(t, y)$$

wird nach der Taylorformel

$$\delta(t, y(t); h) = \mathcal{O}(h^{p+2})$$

d.h. es entsteht ein **explizites** Verfahren der Ordnung $p + 1$. $p = 0$ ergibt das Euler-Verfahren. Bei Funktionen f , deren totale Ableitungen höherer Ordnung einfach berechenbar sind (z.B. durch automatische symbolische Differentiation oder durch Potenzreihenentwicklungen), kann man diese Verfahren erfolgreich einsetzen. Meistens wird dieser Ansatz jedoch an den formalen Schwierigkeiten bei der Aufstellung der höheren totalen Ableitungen scheitern. Das gilt erst recht für Verfahren des Typs

$$\Phi(y, z, t, h; f) \stackrel{\text{def}}{=} \sum_{k=0}^p \frac{h^k}{(k+1)!} \left(\alpha_k \frac{d^k}{dt^k} f(t, y) + \beta_k \frac{d^k}{dt^k} f(t+h, z) \right)$$

bei denen durch geeignete Wahl der α 's und β 's die Ordnung $2p + 2$ erreichbar ist. Z.B. hat das Verfahren

$$\begin{aligned} \eta_{i+1} &= \eta_i + h \cdot \frac{1}{2} (f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1})) + \\ &\quad + \frac{h^2}{12} ((f_t + f_y f)(t_i, \eta_i) - (f_t + f_y f)(t_{i+1}, \eta_{i+1})) \end{aligned}$$

die Ordnung 4. Diese Verfahren werden als Obreschkov-Verfahren bezeichnet.

1.4.2 Runge-Kutta-Verfahren

Bereits in Abschnitt 6.1 hatten wir ein Verfahren der Ordnung 2 konstruiert durch eine Linearkombination von f -Werten der Form

$$\begin{aligned} k_1 &\stackrel{\text{def}}{=} f(t, y) \\ k_2 &\stackrel{\text{def}}{=} f(t + \alpha_2 h, y + h\beta_{21} f(t, y)), \end{aligned}$$

nämlich die Methode von Heun. Die Verallgemeinerung dieser Konstruktion führt zu den expliziten Runge-Kutta-Verfahren:

$$\begin{aligned} k_j(t, y) &\stackrel{\text{def}}{=} f\left(t + \alpha_j h, y + h \sum_{i=1}^{j-1} \beta_{ji} k_i(t, y)\right) \quad j = 1, \dots, m \\ \Phi(y, z, t, h; f) &\stackrel{\text{def}}{=} \sum_{j=1}^m \gamma_j k_j(t, y) \end{aligned}$$

explizites m -stufiges Runge-Kutta-Verfahren.

Die Parameter $\alpha_j, \gamma_j, \beta_{ji}$ kann man unter verschiedenen Gesichtspunkten wählen, wobei der naheliegendste sicher die Erzielung einer möglichst hohen Verfahrensordnung p^* ist. Hinreichend und notwendig für $p^* \geq 1$ ist die Bedingung

$$\sum_{j=1}^m \gamma_j = 1.$$

Sinnvollerweise fordert man auch

$$\sum_{i=1}^{j-1} \beta_{ji} = \alpha_j \quad j = 1, \dots, m.$$

(d.h. auch die sogenannte Stufenordnung ist mindestens 1). Aus der Forderung

$$-y(t+h) + y(t) + h \sum_{i=1}^m \gamma_i k_i(t, y(t)) = \mathcal{O}(h^{p^*+1}) \quad \text{für } f \in F^\infty(\mathcal{I} \times \mathbb{R}^n)$$

erhält man dann ein nichtlineares Gleichungssystem für die Koeffizienten. Für die Aufstellung dieser Gleichungen gibt es eine spezielle Technik, die sogenannten “Butcher-Bäume“. Die **erreichbare** Ordnung p^* hängt natürlich von m ab. Folgende Resultate sind bekannt: (Butcher (1985))

$$\begin{array}{cccccccccccc} m & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & > 9 \\ \hline p^* & 1 & 2 & 3 & 4 & 4 & 5 & 6 & 6 & 7 & \leq m-3 \end{array}$$

Die Anzahl der Gleichungen steigt mit p^* sehr rasch an. Für $p^* = 8$ handelt es sich bereits um 200 nichtlineare Gleichungen. Für $m = p^* = 4$ lauten diese Gleichungen (Konsistenzrelationen)

$$\begin{aligned} 1 &= \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 \\ \frac{1}{2} &= \alpha_2 \gamma_2 + \alpha_3 \gamma_3 + \alpha_4 \gamma_4 \\ \frac{1}{3} &= \alpha_2^2 \gamma_2 + \alpha_3^2 \gamma_3 + \alpha_4^2 \gamma_4 \\ \frac{1}{4} &= \alpha_2^3 \gamma_2 + \alpha_3^3 \gamma_3 + \alpha_4^3 \gamma_4 \\ \frac{1}{8} &= \alpha_3 \alpha_4 \beta_{43} \gamma_4 + \alpha_2 \alpha_4 \beta_{42} \gamma_4 + \alpha_2 \alpha_3 \beta_{32} \gamma_3 \\ \frac{1}{12} &= \alpha_3^2 \beta_{43} \gamma_4 + \alpha_2^2 \beta_{42} \gamma_4 + \alpha_2^2 \beta_{32} \gamma_3 \\ \frac{1}{24} &= \alpha_2 \beta_{32} \beta_{43} \gamma_4 \end{aligned}$$

Die Parameter sind durch diese Gleichungen nicht eindeutig bestimmt und man kann die freien Parameter unter verschiedenen Gesichtspunkten wählen. Rechnerisch besonders einfach ist das “klassische” Runge–Kutta–Verfahren der Ordnung 4:

$$\begin{array}{l|l} \alpha_1 = 0 & \\ \vdots & \beta_{21} \\ \vdots & \vdots \\ \alpha_m & \beta_{m1} \quad \cdots \quad \beta_{m,m-1} \\ \hline & \gamma_1 \qquad \qquad \qquad \gamma_{m-1} \quad \gamma_m \end{array} \qquad \begin{array}{l|ll} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 1 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Von besonderem praktischen Interesse sind sogenannte “eingebettete” Runge–Kutta–Verfahren, bei denen durch verschiedene Wahl der γ_i bei gleichen k_j (d.h. α_j, β_{jl}) Verfahren verschiedener Ordnung entstehen:

z.B. bei $m = 6$

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	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
$p^* = 4$	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
$p^* = 5$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Diese Koeffizientensatz ist als Runge–Kutta–Fehlberg-4-5 bekannt. Es gibt viele Formel-paare dieses Types.

Runge–Kutta–Verfahren sind rechnerisch einfach zu handhaben und bei nicht zu hohen Genauigkeitsansprüchen auch hinreichend effizient, so daß sie in der Praxis am häufigsten verwendet werden. Es sind viele Verfahren dieses Typs mit noch besseren Eigenschaften entwickelt worden, so von Dormand und Prince, Shampine und anderen. Auf die praktische Bedeutung der ineinander eingebetteten Verfahren gehen wir im folgenden Abschnitt ein.

Während bei den Taylormethoden der Hauptteil des **lokalen** Diskretisierungsfehlers stets die Form

$$\delta(t, y(t); h) = Ch^{p+1}y^{(p+1)}(t) + \mathcal{O}(h^{p+2})$$

hat und deshalb mittels dividierter Differenzen von f geschätzt werden kann, ist dies bei den Runge–Kutta–Methoden nicht der Fall. Man kennt zwar a priori Abschätzungen, die aber keine praktische Bedeutung haben, da sie globale Abschätzungen vieler verschiedener partieller Ableitungen von f (nicht nur der Ableitungen von y) benötigen.

1.5 Implementierungsfragen · Schrittweitensteuerung, Fehlerschätzer

Bei der Anwendung von ESV (und natürlich auch von den noch zu besprechenden MSV) in der Praxis hat man im wesentlichen zwei Probleme zu lösen:

- a) Falls das Verfahren implizit ist, wie soll dann die implizite Gleichung gelöst werden?
- b) Wie soll die Schrittweite h_j (lokal) bemessen werden, so daß eine vorgegebene Endgenauigkeit in $\eta(t_E; h)$ ($h \stackrel{def}{=} \max_j h_j$) erreicht wird?

- c) Wie vereinbart man die Forderung einer schrittweitengesteuerten Integration mit der nach der Ausgabe auf einem vordefinierten Gitter?

1.5.1 Lösung der impliziten Gleichungen

Für die implizite Gleichung

$$\eta_{i+1} = \eta_i + h_i \Phi(\eta_i, \eta_{i+1}, t_i, h_i; f)$$

bietet sich zunächst direkte Iteration an:

$$\eta_{i+1}^{[k+1]} \stackrel{\text{def}}{=} \eta_i + h_i \Phi(\eta_i, \eta_{i+1}^{[k]}, t_i, h_i; f)$$

mit $\eta_{i+1}^{[0]} \stackrel{\text{def}}{=} \eta_i$ oder $\eta_{i+1}^{[0]} \stackrel{\text{def}}{=} \eta_i + h_i f(t_i, \eta_i)$. Ist L_2 die Lipschitzkonstante von Φ bzgl. des zweiten Arguments, dann lautet die Konvergenzbedingung

$$hL_2 < 1, \quad h_i \leq h.$$

Nun liegt aber L_2 stets in der Größenordnung von $\|f_y(t, y)\|$. Hat $f_y(t, y)$ Eigenwerte mit betragsmäßig sehr großem negativen Realteil, dann ist notwendig $\|f_y(t, y)\|$ und damit $L_2 \gg 1$, d.h. h muß dann sehr klein gewählt werden. Dies ist in der Praxis unerwünscht. Man zieht dann die Anwendung des vereinfachten Newton-Verfahrens vor:

$$(I - h_i \partial_2 \Phi(\eta_i, \eta_{i+1}^{[0]}, t_i, h_i; f))(\eta_{i+1}^{[k+1]} - \eta_{i+1}^{[k]}) = -\eta_{i+1}^{[k]} + \eta_i + h_i \Phi(\eta_i, \eta_{i+1}^{[k]}, t_i, h_i; f)$$

↑

Funktionalmatrix von Φ
bzgl. der 2. Vektorvariablen

Die Fortführung der Iteration bis auf volle Genauigkeit (sofern dies numerisch überhaupt erreichbar ist) ist nicht sinnvoll, da sie den Aufwand des Verfahrens enorm erhöhen würde. Da durch die Diskretisierung ohnehin pro Schritt ein hoher Diskretisierungsfehler entsteht, kann man die Iteration abbrechen, sobald

$$\|\eta_{i+1}^{[k]} - \eta_i - h_i \Phi(\eta_i, \eta_{i+1}^{[k]}, t_i, h_i; f)\| \leq h_i \varepsilon$$

wobei ε die bei der Schrittweitensteuerung benutzte Schranke für den lokalen Diskretisierungsfehler bedeutet. Der Iterationsfehler ist dadurch immer eine h -Ordnung kleiner.

1.5.2 Schrittweitensteuerung und Schätzung des globalen Diskretisierungsfehlers

In den bisherigen Überlegungen sind wir davon ausgegangen, daß die Näherungslösung η_i für $y(t_i)$ auf einem äquidistanten Gitter erzeugt wird und daß die Schrittweite h *genügend klein* ist. Mit der Formel aus Satz 1.3.3 hat man zwar eine Fehlerschranke,

aber diese ist aus mehreren Gründen für die Praxis wertlos. Zunächst ist festzuhalten, daß die Voraussetzung V0 bei den in der Praxis auftretenden Systemen in der Regel nicht erfüllt ist. Eine zu V0 analoge Bedingung gilt nur in einem Streifen mit $|y_i - y_i(x)| \leq \varepsilon$, $i = 1, 2, \dots, n$, $x \in [a, b]$, wobei aber ε und L nicht praktisch auswertbar sind. Aber selbst wenn die Konstanten bekannt wären, würde doch fast immer die Fehlerschranke unrealistisch groß und eine nach ihr bemessene Schrittweite h viel zu klein. Im Idealfall wird man ja h so bemessen wollen, daß

$$|\eta(t_i; h) - y(t_i)| \leq \varepsilon |y(t_i)| + \varrho, \quad i = 0, \dots, N, \quad (1.8)$$

bzw.

$$\|\eta(t_i; h) - y(t_i)\| \leq \varepsilon \|y(t_i)\| + \varrho, \quad i = 0, \dots, N, \quad (1.9)$$

gilt, mit vorgegebenem (kleinem) ε als relativer Genauigkeit und ϱ als absolutem Fehler. Für den Praktiker stellt sich mit (1.9), also dem Fall eines Differentialgleichungssystems, sogleich ein neues Problem, nämlich das der Konstruktion einer geeigneten Norm. Wenn das Differentialgleichungssystem etwa aus einer Differentialgleichung höherer Ordnung entstanden ist, hat man in den Komponenten von y in ihrer physikalischen Bedeutung etwa Auslenkung, Geschwindigkeit, Beschleunigung vorliegen, also Größen, die um viele Zehnerpotenzen variieren können. Es ist unrealistisch, anzunehmen, man könne alle diese Größen simultan mit der gleichen Genauigkeit integrieren. Andererseits wäre es auch unsinnig, in einem solchen Fall in (1.9) etwa die euklidische Norm zu verwenden, weil das bedeuten würde, daß nur die betragsmäßig größten Komponenten von y *genau* integriert werden. Bei der zu verwendenden Norm wird es sich also stets um eine gewichtete Norm, z.B.

$$\|y\| = \left(\sum_{i=1}^n w_i y_i^2 \right)^{1/2}$$

mit vom Benutzer definierten, von Fall zu Fall verschiedenen Gewichten $w_i > 0$ handeln. Wir gehen im Folgenden jedoch der Einfachheit halber von der euklidischen Norm aus.

Die Genauigkeitsforderung (1.8) bzw. (1.9) ist in der Regel nicht streng einhaltbar. Im Gegensatz zur numerischen Quadratur, wo der Gesamtquadraturfehler die Summe des Einzelquadraturfehler auf den Teilintervallen ist, (vgl. Einführung in die Numerische Mathematik 1, adaptive Quadratur), können bei der Integration einer Differentialgleichung bzw. eines Differentialgleichungs-Systems zurückliegende Fehler, dies sind die pro Integrationsschritt auftretenden Abschneidefehler (und natürlich Rundungsfehler), in den folgenden Schritten sowohl gedämpft als auch verstärkt werden. Bei der Fehlerschranke in Satz 1.3.3 liegt die Annahme einer fortwährenden Fehlerverstärkung, die durch den Faktor $\exp(L(x - x_0))$ beschrieben wird, vor. Da das Fehlerdämpfungs/verstärkungsverhalten einer Differentialgleichung auch noch von der Integrationsstelle abhängt, kann es eine zuverlässige a-priori-Strategie zur Einhaltung von (1.9) bzw. (1.8) nicht geben, weil man ja der Differentialgleichung sozusagen nicht *ansieht*, ob in den späteren Schritten Fehlerdämpfung oder -verstärkung eintreten wird.

Im Gegensatz zum globalen Diskretisierungsfehler $\eta_i - y(t_i)$ kann man jedoch den lokalen Abschneidefehler $\tau(x, y; h)$ recht gut kontrollieren und wird deshalb versuchen, zunächst

durch Kontrolle von $\tau(x, y; h)$ eine Schrittweitensteuerung so zu generieren, daß (1.8) bzw. (1.9) wenigstens approximativ gilt. In einem zweiten Rechengang muß man dann a-posteriori sich in geeigneter Weise eine Kontrolle des globalen Diskretisierungsfehlers verschaffen. Um zu zeigen, daß in vielen praktisch wichtigen Fällen eine Kontrolle des lokalen Diskretisierungsfehlers tatsächlich erfolversprechend ist, benötigen wir einen neuen Begriff, den der *logarithmischen Norm* einer Matrix:

Definition 1.5.1. $\|\cdot\|$ sei eine Vektornorm auf \mathbb{C}^n bzw. die zugeordnete Matrixnorm. Dann heißt

$$\mu(A) \stackrel{\text{def}}{=} \lim_{h \searrow 0} \frac{\|I + hA\| - 1}{h}$$

die logarithmische Norm von A . □

Für einige Vektornormen kann man die zugeordnete logarithmische Matrixnorm explizit angeben, z.B. für

$$\begin{aligned} \|\cdot\|_2 : \quad \mu(A) &= \frac{1}{2} \lambda_{\max}(A^H + A), \\ \|\cdot\|_\infty : \quad \mu(A) &= \max_k \{ \Re a_{kk} + \sum_{i \neq k} |a_{ki}| \}. \end{aligned}$$

Ferner gilt für diese Normen

$$\frac{\|I + hA\| - 1}{h} = \mu(A) + \mathcal{O}(h).$$

Wir kehren nun zurück zur rekursiven Abschätzung der globalen Diskretisierungsfehler $\varepsilon_i = y(t_i) - \eta_i$:

Die Berechnung der η_i erfolgt auf einem nichtäquidistanten Gitter:
 $t_0 = a < t_1 < \dots < t_N = b$,

$$t_{i+1} - t_i = h_i, \quad \sum_{i=0}^{N-1} h_i = b - a,$$

gemäß

$$\eta_{i+1} = \eta_i + h_i \tilde{\Phi}(\eta_i, t_i, h_i; f).$$

Nach Def. 1.3.4 ist

$$y(t_{i+1}) = y(t_i) + h_i \tilde{\Phi}(y(t_i), t_i, h_i; f) - h_i \tau(t_i, y(t_i); h_i).$$

Wir setzen wie zuvor voraus, daß

$$\tilde{\Phi} \in C^2(U \times (0, \bar{h}])$$

mit einer Umgebung U der wahren Lösung der Differentialgleichung

$$U = \{(\eta, t) : t \in [a, b], \|\eta - y(t)\| \leq r\}$$

für ein geeignetes $r > 0$ und $\bar{h} > 0$ gilt.

Für die Funktion τ gelte

$$\tau(t, \eta; h) = h^p \psi(t, \eta) + \mathcal{O}(h^{p+1}) \quad (1.10)$$

mit $\psi \in C^1(U)$. Bei den praktisch interessierenden Verfahren ist dies stets der Fall. Dann wird

$$\begin{aligned} \varepsilon_{i+1} &= \varepsilon_i + h_i (\tilde{\Phi}(y(t_i), t_i, h_i; f) - \tilde{\Phi}(\eta_i, t_i, h_i; f)) + \\ &\quad - h_i^{p+1} \psi(t_i, y(t_i)) + \mathcal{O}(h_i^{p+2}) \end{aligned}$$

und weiter unter Ausnutzung der zweimaligen stetigen Differenzierbarkeit von $\tilde{\Phi}$ und $\tilde{\Phi}(\eta, t, 0) \equiv f(t, \eta)$, d.h. $\partial_1 \tilde{\Phi}(\eta, t, 0) \equiv \partial_2 f(t, \eta)$,² und dem Vorwissen von $\|\varepsilon_i\| = \mathcal{O}(h_{max}^p)$

$$\varepsilon_{i+1} = \varepsilon_i + h_i \partial_2 f(t_i, y(t_i)) \varepsilon_i - h_i^{p+1} \psi(t_i, y(t_i)) + h_i \mathcal{O}(h_{max}^{p+1}).$$

Also erhalten wir

$$\begin{aligned} \|\varepsilon_{i+1}\| &\leq \|\varepsilon_i\| + h_i \frac{\|I + h_i \partial_2 f(t_i, y(t_i))\| - 1}{h_i} \|\varepsilon_i\| + h_i^{p+1} \|\psi(t_i, y(t_i))\| + h_i \mathcal{O}(h_{max}^{p+1}) \\ &\leq (1 + h_i \mu(\partial_2 f(t_i, y(t_i)))) \|\varepsilon_i\| + h_i^{p+1} \|\psi(t_i, y(t_i))\| + \mathcal{O}(h_i^{p+2}), \end{aligned}$$

und unter Auflösung der Rekursion und Beachtung von $\sum_{j=0}^i h_j \mathcal{O}(h_{max}^{p+1}) = \mathcal{O}(h^{p+1})$ mit $h = \max h_j$ und $\varepsilon_0 = 0$

$$\begin{aligned} \|\varepsilon_{i+1}\| &\leq \sum_{j=0}^i \exp\left(\sum_{k=j}^i h_k \mu(\partial_2 f(t_k, y(t_k)))\right) h_j^{p+1} \|\psi(t_j, y_j^h)\| + \mathcal{O}(h^{p+1}) \\ &\leq \sum_{j=0}^i \exp\left(\int_{t_j}^{t_{i+1}} \mu(\partial_2 f(\xi, y(\xi))) d\xi\right) h_j^{p+1} \|\psi(t_j, y_j^h)\| + \mathcal{O}(h^{p+1}). \end{aligned} \quad (1.11)$$

Dies ist die erwünschte verfeinerte Fehlerabschätzung. Für $\mu(\dots) \leq 0$ ist $\|\varepsilon_{i+1}\|$ also im wesentlichen abschätzbar durch $\sum_{j=0}^i h_j \tau(t_j, y(t_j); h_j)$, und auch wenn μ nicht wesentlich größer als 0 ist, beeinflusst eine Steuerung von ϱ die Größe ε in überschaubarer Weise. Die Summe ist hierin von der Größenordnung $\mathcal{O}(h^p)$. Falls

$$\mu(\partial_2 f(t, y(t))) \leq q < 0 \quad (1.12)$$

gilt, dann kann man weiter vereinfachen zu

$$\|\varepsilon_{i+1}\| \leq \sum_{j=0}^i \exp(-|q| |t_{i+1} - t_{j+1}|) h_j^{p+1} \|\psi(t_j, y_j^h)\| + \mathcal{O}(h^{p+1}).$$

Man erkennt, daß dann zurückliegende lokale Abschneidefehler immer stärker weggedämpft werden. In vielen Fällen ist wenigstens $\mu(\partial_2 f(\cdot))$ nicht wesentlich größer als

² $\partial_2 f$ bezeichnet die partielle Ableitung der Vektorfunktion f nach der zweiten Variablen, also hier nach dem Vektor y . $\partial_2 f$ ist also eine $n \times n$ -Matrix. Gelegentlich schreibt man hierfür auch f_y .

null, während oft $L \gg 1$ ist, so daß der Vorteil von (1.11) gegenüber Satz 1.3.3 unmittelbar einsichtig ist. Um (1.9) wenigstens approximativ zu erreichen, ist folgende Strategie üblich: Wähle an der Stelle (t_i, η_i) die Schrittweite h_i so, daß

$$h_i^{p+1} \|\psi(t_i, \eta_i)\| \lesssim (\varepsilon \|\eta_i\| + \varrho) \quad (1.13)$$

bzw. wenn man davon ausgehen muß, daß (1.12) nicht gilt, etwas vorsichtiger

$$h_i^{p+1} \|\psi(t_i, \eta_i)\| \lesssim (\varepsilon \|\eta_i\| + \varrho) h_i / (b - a). \quad (1.14)$$

Um h_i hieraus zu berechnen, muß natürlich die Größe ψ , d.h. i.w. der lokale Abschneidefehler, zumindest zuverlässig schätzbar sein.

Dies ist aber tatsächlich der Fall. Hierzu stehen praktisch drei Methoden zur Verfügung.

Ein Spezialfall ist

$$\delta(t, y(t); h) = Ch^{p+1} y^{(p+1)}(t) + \mathcal{O}(h^{p+2}), \quad (1.15)$$

wobei C eine bekannte Konstante ist. p ist natürlich auch bekannt.

Bei manchen Verfahrenstypen (z.B. den Taylormethoden und den noch zu besprechenden MSV) hat der lokale Diskretisierungsfehler diese Form, d.h.

$$\delta(t_i, \eta_i; h_i) = Ch_i^{p+1} \left(\frac{d^p}{dt^p} f \right) (t_i, \eta_i) + \mathcal{O}(h_i^{p+2})$$

Nun ist aber mit $j \stackrel{def}{=} i, i-1, \dots, i-p$

$$\begin{aligned} z_j &\stackrel{def}{=} f(t_j, \eta_j) & \tilde{z}_j &\stackrel{def}{=} f(t_j, y_j) & y' &= f(t, y), & y(t_i) &= \eta_i \\ z_j - \tilde{z}_j &= \mathcal{O}(h^{p+1}) & h &\stackrel{def}{=} \max_{j=i-p, \dots, i} h_i \\ [\tilde{z}_{i-p}, \dots, \tilde{z}_i] &= \frac{1}{p!} \left(\frac{d^p}{dt^p} f \right) (t_i, \eta_i) + \mathcal{O}(h) \\ [z_{i-p}, \dots, z_i] &= [\tilde{z}_{i-p}, \dots, \tilde{z}_i] + \mathcal{O}(h) \end{aligned}$$

so daß man erhält

$$\delta(t_i, \eta_i; h_i) = p! Ch_i^{p+1} [z_{i-p}, \dots, z_i] + \mathcal{O}(h^{p+2}),$$

d.h. man kann den Hauptteil des lokalen Diskretisierungsfehlers aus dividierten Differenzen zurückliegender f -Werte schätzen.

Der vielleicht eleganteste Zugang besteht im Vergleich der Resultate zweier Einschrittverfahren verschiedener Ordnung. Seien also durch $\tilde{\Phi}_1(\eta, t, h; f)$ und $\tilde{\Phi}_2(\eta, t, h; f)$ zwei Einschritt-Verfahren definiert; die zu $\tilde{\Phi}_1$ gehörende Ordnung sei p_1 und die zu $\tilde{\Phi}_2$ sei $p_2 > p_1$.

Ausgehend von t_i, y_i^h wird nun mit einer Vorschlagsschrittweite \tilde{h}_i gerechnet:

$$\begin{aligned} \eta_{i+1}^{[1]} &\stackrel{def}{=} \eta_i + \tilde{h}_i \tilde{\Phi}_1(\eta_i, t_i, \tilde{h}_i; f), \\ \eta_{i+1}^{[2]} &\stackrel{def}{=} \eta_i + \tilde{h}_i \tilde{\Phi}_2(\eta_i, t_i, \tilde{h}_i; f). \end{aligned}$$

Ferner sei $y_{[i]}(t)$ die Lösung der Anfangswertaufgabe

$$y'_{[i]} = f(t, y_{[i]}), \quad y_{[i]}(t_i) = \eta_i.$$

$y_{[i]}$ löst also die Differentialgleichung und hat an der Stelle t_i den verfälschten Wert η_i als exakten Anfangswert. Dann ist nach Definition

$$\begin{aligned} y_{[i]}(t_{i+1}) - \eta_i - \tilde{h}_i \tilde{\Phi}_1(\eta_i, t_i, \tilde{h}_i; f) &= -\tilde{h}_i \tau_1(t_i, \eta_i; \tilde{h}_i) = \mathcal{O}(\tilde{h}_i^{p_1+1}), \\ y_{[i]}(t_{i+1}) - \eta_i - \tilde{h}_i \tilde{\Phi}_2(\eta_i, t_i, \tilde{h}_i; f) &= -\tilde{h}_i \tau_2(t_i, \eta_i; \tilde{h}_i) = \mathcal{O}(\tilde{h}_i^{p_2+1}), \end{aligned}$$

und daher wegen $p_2 \geq p_1 + 1$

$$\begin{aligned} \eta_{i+1}^{[2]} - \eta_{i+1}^{[1]} &= -\tilde{h}_i \tau_1(t_i, \eta_i; \tilde{h}_i) + \mathcal{O}(\tilde{h}_i^{p_1+2}) \\ &= -\tilde{h}_i^{p_1+1} \psi_1(t_i, \eta_i) + \mathcal{O}(\tilde{h}_i^{p_1+2}) \end{aligned}$$

unter Ausnutzung von (1.10). Somit folgt

$$\|\psi_1(t_i, \eta_i)\| \lesssim \|\eta_{i+1}^{[1]} - \eta_{i+1}^{[2]}\| / \tilde{h}_i^{p_1+1}.$$

Diese Schätzung kann nun auf der linken Seite von (1.13) oder (1.14) eingesetzt werden, um die aktuelle Schrittweite h_i zu ermitteln. Für die Forderung (1.13) testet man, ob

$$\|\eta_{i+1}^{[1]} - \eta_{i+1}^{[2]}\| \leq \varepsilon \|\eta_i\| + \varrho. \quad (1.16)$$

Ist dies der Fall, dann war \tilde{h}_i klein genug und der Schritt wird akzeptiert, d.h.

$h_i = \tilde{h}_i$, $\eta_{i+1} \stackrel{def}{=} \eta_{i+1}^{[2]}$ z.B. (Es ist üblich, mit dem – vermeintlich – genaueren Wert weiterzurechnen, obwohl das Gitter für die Integration mit $\tilde{\Phi}_1$ gesteuert wird.) Die Vorschlagsschrittweite für den nächsten Schritt errechnet man dann aus

$$\tilde{h}_{i+1} \stackrel{def}{=} \min\{\alpha h_i, \left(\frac{\varepsilon \|\eta_i\| + \varrho}{\|\eta_{i+1}^{[1]} - \eta_{i+1}^{[2]}\|}\right)^{1/(p_1+1)} h_i, h_{\max}\}, \quad 1 < \alpha,$$

(mit $\alpha = 2$ z.B., um zu verhindern, daß die Schrittweite zu schnell vergrößert wird.)

Gilt (1.16) nicht, muß die Vorschlagsschrittweite \tilde{h}_i für den laufenden Schritt verkleinert werden und die Rechnung ist zu wiederholen, etwa mit

$$\tilde{h}_i \stackrel{def}{=} \max\{\beta \tilde{h}_i, \left(\frac{\varepsilon \|\eta_i\| + \varrho}{\|\eta_{i+1}^{[1]} - \eta_{i+1}^{[2]}\|}\right)^{1/(p_1+1)} \tilde{h}_i\}, \quad 0 < \beta < 1,$$

(mit $\beta = 1/2$ z.B., um zu verhindern, daß \tilde{h}_i zu schnell zu klein wird). Unterschreitet dabei \tilde{h}_i einen Wert h_{\min} (sinnvoll ist z.B. $h_{\min} = \sqrt{\text{epsmach}}$, epsmach=relative Maschinengenauigkeit), dann wird man die Integration mit h_{\min} fortsetzen und einen entsprechenden Vermerk setzen oder auch die Rechnung ganz abbrechen. Für die Forderung (1.14) läuft alles analog.

Auf diese Art wird also wie bei der adaptiven Quadratur ein variables Gitter erzeugt. In diesem Zusammenhang ist es von großer Bedeutung, daß es möglich ist, sogenannte

eingebettete Runge–Kutta–Verfahren zu konstruieren, bei denen unter Benutzung der gleichen k_j –Werte durch Variation der Gewichte γ_i Verfahren verschiedener Ordnung entstehen.

Die ersten Verfahren dieser Art sind bei England bzw. Fehlberg zu finden, vgl. die ausführliche Diskussion in Hairer, E.; Norsett, S.P. und Wanner, G.: Solving Ordinary Differential Equations I. Nonstiff Problems, Springer, (1987).

Ein einfaches Beispiel ist

$$\begin{array}{c|ccc}
 0 & 0 & 0 & 0 \\
 1 & 1 & 0 & 0 \\
 \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\
 \hline
 \gamma_i & \frac{1}{2} & \frac{1}{2} & 0 \\
 \hat{\gamma}_i & \frac{1}{6} & \frac{1}{6} & \frac{4}{6}
 \end{array}$$

d.h.

$$\begin{aligned}
 k_1 &= f(t, y), \\
 k_2 &= f(t + h, y + hk_1), \\
 k_3 &= f(t + \frac{h}{2}, y + \frac{h}{4}(k_1 + k_2)).
 \end{aligned}$$

Hier ist

$$\tilde{\Phi}_1(y, t, h; f) = \frac{1}{2}(k_1 + k_2) \quad \text{von der Ordnung 2}$$

und

$$\tilde{\Phi}_2(y, t, h; f) = \frac{1}{6}(k_1 + k_2 + 4k_3) \quad \text{von der Ordnung 3.}$$

Sehr gute Resultate erzielt man mit der Formel von Dormand und Prince (Dormand, J.R.; Prince, P.J.: A family of embedded Runge-Kutta-formulae, Comp. Appl. Math.,

Vol. 6, (1980), S. 19-26), die durch das folgende Koeffizientenschema beschrieben wird.

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	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
γ_i	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
$\hat{\gamma}_i$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Es ist also

$$k_j = f(t + \alpha_j h, y + h \sum_{s=1}^{j-1} \beta_{js} k_s), \quad j = 1, \dots, 7.$$

Hier hat die Formel

$$\tilde{\Phi}_1(y, t, h; f) = \sum_{i=1}^7 \gamma_i k_i \quad \text{die Ordnung 5}$$

und

$$\tilde{\Phi}_2(y, t, h; f) = \sum_{i=1}^7 \hat{\gamma}_i k_i \quad \text{die Ordnung 4.}$$

Die Fehlerkonstanten der Methode der Ordnung 5 sind hierbei minimiert, so daß die obige Fehlerschätzungsmethode sehr zuverlässig ist. Ferner beachte man, daß hier k_7 zugleich k_1 für den nächsten Schritt ist, so daß effektiv nur 6 Funktionsauswertungen pro Schritt zu leisten sind.

Die dritte Methode zur Schätzung des lokalen Abschneidefehlers beruht auf der Ausnutzung asymptotischer Entwicklungen. Dazu gilt der folgende

Satz 1.5.1. Die Lösung y der Anfangswertaufgabe

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

existiere auf $[a, b]$ und U sei eine Umgebung der Lösungskurve:

$$U = \{(t, \eta) : t \in [a, b], \quad \|\eta - y(t)\| \leq r\}.$$

Es sei $\tilde{\Phi} \in C^3(U \times (0, \bar{h}])$ mit $\bar{h} > 0$, $\tilde{\Phi}(t, \eta; 0) \equiv f(t, \eta)$. Das durch $\tilde{\Phi}$ beschriebene Verfahren besitze die Ordnung p . Der lokale Abschneidefehler besitze die Entwicklung

$$\tau(t, \eta; h) = h^p \psi_1(t, \eta) + h^{p+1} \psi_2(t, \eta) + \mathcal{O}(h^{p+2})$$

mit $\psi_1 \in C^2(U)$, $\psi_2 \in C^1(U)$. $y^h(t)$ bezeichne schließlich die mit dem Einschritt-Verfahren bei konstanter Schrittweite h berechnete Näherung an der Stelle $t \geq a$. Dann gilt

$$y^h(t) - y(t) = h^p z_1(t) + h^{p+1} z_2(t) + \mathcal{O}(h^{p+2}), \quad (1.17)$$

wobei z_1 und z_2 selbst differenzierbare Funktionen sind.

Zum Beweis vgl. Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen I, Teubner, (1972). \square

Diesen Satz wenden wir nun zur Schätzung von $\psi_1(t_i, \eta_i)$ auf die Aufgabe $y'_{[i]} = f(t, y_{[i]})$, $y_{[i]}(t_i) = \eta_i$ an. In diesem Fall ist dann $z_1(t_i) = z_2(t_i) = 0$, d.h. $z_1(t_{i+1}) = \mathcal{O}(h_i)$, $z_2(t_{i+1}) = \mathcal{O}(h_i)$. Von der Stelle (t_i, η_i) aus berechnen wir nun einmal η_{i+1} , d.h.

$$\eta_{i+1} = \eta_i + h_i \tilde{\Phi}(\eta_i, t_i, h_i; f),$$

sowie die zweite Näherung mit der Schrittweite $h_i/2$

$$\begin{aligned} \tilde{\eta}_{i+\frac{1}{2}} &= \eta_i + \frac{h_i}{2} \tilde{\Phi}(\eta_i, t_i, \frac{h_i}{2}; f), \\ \tilde{\eta}_{i+1} &= \tilde{\eta}_{i+\frac{1}{2}} + \frac{h_i}{2} \tilde{\Phi}(\tilde{\eta}_{i+\frac{1}{2}}, t_i, \frac{h_i}{2}; f). \end{aligned}$$

Damit ergibt (1.17) wegen $z_2(t_{i+1}) = \mathcal{O}(h_i)$

$$\eta_{i+1} - \tilde{\eta}_{i+1} = h_i^p z_1(t_{i+1}) - \left(\frac{h_i}{2}\right)^p z_1(t_{i+1}) + \mathcal{O}(h_i^{p+2}),$$

also

$$z_1(t_{i+1}) = \frac{2^p(\eta_{i+1} - \tilde{\eta}_{i+1})}{(2^p - 1)h_i^p} + \mathcal{O}(h_i^2)$$

bzw.

$$\eta_{i+1} - y_{[i]}(t_{i+1}) = \frac{2^p(\eta_{i+1} - \tilde{\eta}_{i+1})}{(2^p - 1)} + \mathcal{O}(h_i^{p+2}).$$

Andererseits ist

$$\begin{aligned} h_i \tau(t_i, \eta_i; h_i) &= y_{[i]}(t_{i+1}) - \eta_i - h_i \tilde{\Phi}(\eta_i, t_i, h_i; f) \\ &= y_{[i]}(t_{i+1}) - \eta_{i+1} \\ &= h_i^{p+1} \psi_1(t_i, \eta_i) + \mathcal{O}(h_i^{p+2}). \end{aligned}$$

Somit wird

$$\frac{2^p}{2^p - 1}(\tilde{\eta}_{i+1} - \eta_{i+1}) = h_i^{p+1}\psi_1(t_i, \eta_i) + \mathcal{O}(h_i^{p+2}), \quad (1.18)$$

die linke Seite von (1.18) kann also als Schätzung von $h_i^{p+1}\psi_1(t_i, \eta_i)$ in (1.13) bzw. (1.14) benutzt werden. Dies ist also wieder eine Anwendung der Richardson-Extrapolation.

Es ist selbstverständlich, daß alle hier vorgestellten Methoden zur Kontrolle des lokalen Abschneidefehlers nur wirksam sind, wenn die Terme höherer Ordnung $\mathcal{O}(h_i^{p+2})$ gegenüber $h_i^{p+1}\psi_1(t_i, \eta_i)$ tatsächlich vernachlässigbar sind, wobei ein Fehler von vielleicht 10% noch tolerierbar ist. Dies erfordert, daß grundsätzlich die angewandten Integrationsschritte niemals *zu groß* werden, also etwa unter einer Schranke $(b - a)/10$ oder $(b - a)/100$ bleiben, wobei $b - a$ die Größenordnung 1 besitzt. Wie diese Sicherheitswerte zu wählen sind, kann man in einem konkreten Fall nur aufgrund numerischer Erfahrung entscheiden.

Wir haben nun geklärt, wie in der Praxis eine nichtäquidistante Gittereinteilung erzeugt wird mit dem Ziel, die Genauigkeitsforderung (1.9) (bzw. (1.8)) einzuhalten. Eine Garantie für die Einhaltung dieser Genauigkeitsforderung bietet die entwickelte Strategie nur dann, wenn $\mu(\partial_2 f(\cdot)) < 0$.

Um den globalen Diskretisierungsfehler zu kontrollieren, gibt es verschiedene Ansätze. Wir beschränken uns auf eine einfache, aber recht zuverlässige Vorgehensweise. Hierbei wird simultan zur Berechnung der approximativen Lösung $(t_i, \eta(t_i; h))$ noch eine zweite, in der Regel genauere Lösung $(t_i, \eta(t_i; h/2))$ erzeugt, indem jedesmal, wenn ein Integrationsschritt akzeptiert werden kann, zwei Schritte des gleichen Verfahrens mit der Schrittweite $h_i/2$ ausgeführt werden, also

$$\begin{aligned} \eta(t_i + \frac{h}{2}, \frac{h}{2}) &= \eta(t_i; h/2) + \frac{h_i}{2} \tilde{\Phi}(\eta(t_i; \frac{h}{2}), t_i, \frac{h_i}{2}; f) \\ \eta(t_{i+1}; \frac{h}{2}) &= \eta(t_i + \frac{h}{2}; \frac{h}{2}) + \frac{h_i}{2} \tilde{\Phi}(\eta(t_i + \frac{h}{2}, \frac{h}{2}), t_i + \frac{h_i}{2}, \frac{h_i}{2}; f). \end{aligned}$$

Dabei wird auch der lokale Abschneidefehler von $\eta(t_{i+1}; \frac{h_i}{2})$ geschätzt. Der Schritt wird endgültig nur dann akzeptiert, wenn diese beiden Teilschritte erfolgreich waren und die Schätzung des lokalen Abschneidefehlers des Verfahrens mit der Schrittweite $h_i/2$ kleiner ausfällt als die Schätzung des lokalen Abschneidefehlers für das Verfahren mit der Schrittweite h_i , etwa bei Verwendung von zwei Verfahren verschiedener Ordnung, wenn

$$\|\eta^{[1]}(t_{i+1}; \frac{h_i}{2}) - \eta^{[2]}(t_{i+1}; \frac{h_i}{2})\| \leq \frac{1}{2} \|\eta^{[1]}(t_{i+1}; h_i) - \eta^{[2]}(t_{i+1}; h_i)\|.$$

Man beachte, daß, von der Akzeptierungsregel für h_i abgesehen, die Näherungen $\eta(t_i; h)$ und $\eta(t_i; \frac{h}{2})$ völlig entkoppelt berechnet werden. $\eta(t_i; \frac{h}{2}) - \eta(t_i; h)$ dient dann als Schätzung des globalen Diskretisierungsfehlers $y(t_i) - \eta(t_i; h)$. Die Werte $\eta(t_i; \frac{h}{2})$ werden nie benutzt, um $\eta(t_i; h)$ zu verbessern, d.h. es wird keine lokale Extrapolation verwendet.

Strenge Fehlerschranken erhält man dadurch natürlich *nicht*, aber die Methode hat sich doch als sehr zuverlässig erwiesen.

Eine andere, oft bewährte Methode beruht auf der sogenannten *Defektkorrektur*. Aus den berechneten diskreten Werten (t_i, η_i) , $i = 0, \dots, N_h$, konstruiert man durch stück-

weise Polynominterpolation oder Spline-Interpolation eine zumindest stückweise beliebig oft differenzierbare kontinuierliche Approximation

$$\tilde{y}^h(t), \quad t \in [t_0, t_{N_h}], \quad \text{mit } \tilde{y}^h(t_i) = \eta_i, \quad i = 0, \dots, N_h,$$

für $y(t)$. Wir interessieren uns für den Fehler

$$\varepsilon(t; h) = y(t) - \tilde{y}^h(t).$$

Offensichtlich ist in jedem Intervall, in dem \tilde{y}^h differenzierbar ist, auch ε differenzierbar. ε löst die Anfangswertaufgabe

$$\begin{aligned} \varepsilon' &= f(t, \tilde{y}^h(t) + \varepsilon(t)) - (\tilde{y}^h)'(t), \\ \varepsilon(t_0) &= y(t_0) - y_0^h. \end{aligned} \tag{1.19}$$

Durch numerische Integration dieses Anfangswertproblems kann man also auch eine Schätzung für ε erhalten. Die sinnvolle Anwendung dieser Idee wird allerdings dadurch erschwert, daß der Interpolationsfehler für $\tilde{y}^h(t)$ mindestens wie h^{p+q} gegen null gehen muß, wenn p die Konvergenzordnung in η_i und q die Ordnung des Verfahrens zur Integration von (1.19) ist, und daß bei der Integration der Anfangswertaufgabe (1.19) alle Punkte, an denen $\tilde{y}^h(t)$ nicht von hoher Ordnung differenzierbar ist, auch Gitterpunkte des Gitters für ε_i^h werden müssen.

Beispiel 1.5.1. *Mit der oben beschriebenen Strategie der simultanen Integration des Systems mit den Schrittweiten h_i und $h_i/2$ wurde das von Hairer, Norsett und Wanner beschriebene Dreikörperproblem unter Benutzung der oben angegebenen Formeln von Dormand und Prince integriert. Bei diesem Problem handelt es sich um das Anfangswertproblem*

$$\begin{aligned} y_1'' &= y_1 + 2y_2' - \mu_1(y_1 + \mu_2)/D_1 - \mu_2(y_1 - \mu_1)/D_2, \\ y_2'' &= y_2 - 2y_1' - \mu_1 y_2/D_1 - \mu_2 y_2/D_2, \\ D_1 &= ((y_1 + \mu_2)^2 + y_2^2)^{3/2}, \\ D_2 &= ((y_1 - \mu_1)^2 + y_2^2)^{3/2}, \\ \mu_1 &= 1 - \mu_2, \quad \mu_2 \stackrel{\text{def}}{=} 0.012277471, \\ y_1(0) &= 0.994, \quad y_1'(0) \stackrel{\text{def}}{=} 0, \quad y_2(0) \stackrel{\text{def}}{=} 0, \quad y_2'(0) \stackrel{\text{def}}{=} -2.00158510637908252, \end{aligned}$$

mit einer periodischen Lösung der Periode $T = 17.0652165601579625$. Das System wurde umgeschrieben in ein System 1. Ordnung mit

$$Y_1 = y_1, \quad Y_2 = y_1', \quad Y_3 = y_2, \quad Y_4 = y_2'.$$

Zur Schrittweitensteuerung diente die Forderung

$$\left(\frac{1}{n} \sum_{j=1}^n \left(\frac{\eta_{i+1,j}^{[1]} - \eta_{i+1,j}^{[2]}}{\max\{|\eta_{i,j}^{[2]}|, |\eta_{i+1,j}^{[2]}|, \varrho\}} \right)^2 \right)^{1/2} \leq \varepsilon$$

mit $\varepsilon = TOL$ und $\rho = SMALLY$ in den unten angegebenen Resultatlisten. ($\eta_{i,j}$ bezeichnet die j -te Komponente des Vektors η_i) Man erkennt aus den unten angegebenen Resultaten, daß die erzielte Schätzung des globalen Fehlers sehr realistisch ist. Der globale Fehler liegt hier aber weit über der Toleranzgrenze für den lokalen Abschneidefehler. Die Genauigkeit der Lösung hängt dabei sehr kritisch vom Parameter ρ ab, weil kleine Fehler in der Approximation von y'_1 in der Nähe von 0 einen extremen Einfluß auf die Genauigkeit der Lösungstrajektorie haben. $\rho = 1$ z.B. wäre völlig ungeeignet. GLOBERR bezeichnet die euklidische Norm des geschätzten globalen Diskretisierungsfehlers, NFE die akkumulierte Anzahl der benutzten Funktionsauswertungen und Y die vier Lösungskomponenten. Für $\varepsilon = TOL = 10^{-6}$ werden 238 Integrationsschritte ausgeführt, darunter 31 wiederholte. Insbesondere an y_2 und y_3 erkennt man das relativ hohe Fehlerniveau, denn zur Endzeit sollte ja eigentlich wieder $y_2 = y_3 = 0$ sein. Für $\varepsilon = TOL = 10^{-8}$ erhöht sich der Aufwand nur etwa auf das Doppelte, die Lösung ist jetzt viel besser. Insgesamt werden 497 Integrationsschritte ausgeführt, davon 27 wiederholte.

DREIKOERPERPROBLEM, PERIODISCHE LOESUNG

INTERVALL 0.0000000000000000E+00 17.06521656015796

ANFANGSWERT

0.9940D+00 0.0000D+00 0.0000D+00 -0.2002D+01

TOL= 1.0000000000000000E-06

HMAX= 1.0000000000000000

HSTART= 1.0000000000000000E-05

SMALLY= 1.0000000000000000E-07

t=	GLOBERR=	NFE=	Y=			
.3413D+00	.4788D-05	524	.7801D+00	-.5482D+00	-.1632D-01	.2496D+00
.6826D+00	.1013D-04	598	.5797D+00	-.6689D+00	.1316D+00	.6037D+00
.1024D+01	.2015D-04	672	.2880D+00	-.1073D+01	.3639D+00	.6533D+00
.1365D+01	.2161D-04	788	-.1060D+00	-.1100D+01	.5092D+00	.2024D+00
.1707D+01	.1597D-04	880	-.4152D+00	-.7097D+00	.5547D+00	.1326D+00
.2048D+01	.1482D-04	954	-.5991D+00	-.3808D+00	.6212D+00	.2636D+00
.2389D+01	.1454D-04	1010	-.6815D+00	-.1090D+00	.7313D+00	.3705D+00
.2730D+01	.1384D-04	1066	-.6776D+00	.1256D+00	.8649D+00	.3981D+00
.3072D+01	.1265D-04	1104	-.6011D+00	.3135D+00	.9935D+00	.3424D+00
.3413D+01	.1111D-04	1142	-.4711D+00	.4360D+00	.1091D+01	.2197D+00
.3754D+01	.9449D-05	1180	-.3129D+00	.4760D+00	.1139D+01	.5700D-01
.4096D+01	.7994D-05	1218	-.1565D+00	.4251D+00	.1129D+01	-.1137D+00
.4437D+01	.7391D-05	1256	-.3299D-01	.2842D+00	.1064D+01	-.2624D+00
.4778D+01	.8709D-05	1318	.2832D-01	.6268D-01	.9546D+00	-.3707D+00
.5120D+01	.1295D-04	1380	.2278D-02	-.2245D+00	.8146D+00	-.4480D+00
.5461D+01	.2145D-04	1436	-.1284D+00	-.5384D+00	.6445D+00	-.5668D+00
.5802D+01	.3429D-04	1492	-.3478D+00	-.6862D+00	.4075D+00	-.8483D+00
.6143D+01	.3743D-04	1584	-.5442D+00	-.4244D+00	.8187D-01	-.9763D+00
.6485D+01	.2859D-04	1694	-.6550D+00	-.2765D+00	-.2083D+00	-.6915D+00
.6826D+01	.2447D-04	1750	-.7557D+00	-.3268D+00	-.3865D+00	-.3586D+00
.7167D+01	.2301D-04	1806	-.8787D+00	-.3867D+00	-.4579D+00	-.6644D-01
.7509D+01	.2209D-04	1862	-.1012D+01	-.3841D+00	-.4365D+00	.1839D+00
.7850D+01	.2130D-04	1900	-.1132D+01	-.3074D+00	-.3383D+00	.3816D+00
.8191D+01	.2075D-04	1938	-.1215D+01	-.1703D+00	-.1840D+00	.5096D+00
.8533D+01	.2071D-04	1976	-.1245D+01	-.2343D-05	-.2072D-04	.5540D+00
.8874D+01	.2155D-04	2014	-.1215D+01	.1703D+00	.1840D+00	.5096D+00
.9215D+01	.2371D-04	2052	-.1132D+01	.3074D+00	.3383D+00	.3816D+00
.9557D+01	.2775D-04	2090	-.1012D+01	.3841D+00	.4365D+00	.1839D+00
.9898D+01	.3448D-04	2158	-.8787D+00	.3867D+00	.4578D+00	-.6646D-01
.1024D+02	.4568D-04	2214	-.7557D+00	.3269D+00	.3864D+00	-.3587D+00
.1058D+02	.6607D-04	2270	-.6550D+00	.2765D+00	.2082D+00	-.6915D+00
.1092D+02	.1015D-03	2356	-.5441D+00	.4245D+00	-.8196D-01	-.9763D+00
.1126D+02	.1119D-03	2448	-.3477D+00	.6863D+00	-.4076D+00	-.8482D+00
.1160D+02	.8221D-04	2522	-.1284D+00	.5384D+00	-.6445D+00	-.5667D+00
.1195D+02	.6490D-04	2578	.2345D-02	.2245D+00	-.8146D+00	-.4480D+00
.1229D+02	.5725D-04	2634	.2838D-01	-.6269D-01	-.9546D+00	-.3707D+00
.1263D+02	.5307D-04	2690	-.3293D-01	-.2842D+00	-.1064D+01	-.2624D+00
.1297D+02	.5062D-04	2746	-.1564D+00	-.4251D+00	-.1129D+01	-.1137D+00
.1331D+02	.4963D-04	2802	-.3128D+00	-.4760D+00	-.1139D+01	.5699D-01

.1365D+02	.5034D-04	2840	-.4710D+00	-.4360D+00	-.1091D+01	.2197D+00
.1399D+02	.5329D-04	2878	-.6010D+00	-.3135D+00	-.9935D+00	.3424D+00
.1433D+02	.5932D-04	2916	-.6775D+00	-.1255D+00	-.8649D+00	.3981D+00
.1468D+02	.6988D-04	2984	-.6815D+00	.1090D+00	-.7313D+00	.3704D+00
.1502D+02	.8837D-04	3040	-.5990D+00	.3809D+00	-.6212D+00	.2636D+00
.1536D+02	.1254D-03	3096	-.4151D+00	.7098D+00	-.5547D+00	.1326D+00
.1570D+02	.2082D-03	3170	-.1058D+00	.1100D+01	-.5092D+00	.2026D+00
.1604D+02	.2563D-03	3286	.2881D+00	.1072D+01	-.3638D+00	.6535D+00
.1638D+02	.1605D-03	3378	.5797D+00	.6688D+00	-.1315D+00	.6037D+00
.1672D+02	.1175D-03	3452	.7802D+00	.5482D+00	.1641D-01	.2496D+00
.1707D+02	.1596D-01	4012	.9940D+00	.1537D-01	.9493D-04	-.1997D+01

ERGEBNIS AUS DORPRI54

T= 17.06521656015796

LOESUNG

Y(1)= 0.9940310D+00 Y(2)= 0.1536854D-01

Y(3)= 0.9493243D-04 Y(4)=-0.1996608D+01

SCHAETZUNG DES GLOBALEN FEHLERS 1.5963071908337041E-02

DREIKOERPERPROBLEM, PERIODISCHE LOESUNG

INTERVALL 0.0000000000000000E+00 17.06521656015796

ANFANGSWERT

0.9940D+00 0.0000D+00 0.0000D+00 -0.2002D+01

TOL= 1.0000000000000000E-08

HMAX= 1.0000000000000000

HSTART= 1.0000000000000000E-06

SMALLY= 1.0000000000000000E-07

t=	GLOBERR=	NFE=	Y=			
.3413D+00	.3221D-07	1190	.7801D+00	-.5482D+00	-.1632D-01	.2496D+00
.6826D+00	.6840D-07	1372	.5797D+00	-.6689D+00	.1316D+00	.6037D+00
.1024D+01	.1431D-06	1572	.2880D+00	-.1073D+01	.3639D+00	.6533D+00
.1365D+01	.1495D-06	1850	-.1059D+00	-.1100D+01	.5092D+00	.2024D+00
.1707D+01	.1088D-06	2050	-.4152D+00	-.7097D+00	.5547D+00	.1326D+00
.2048D+01	.1033D-06	2196	-.5991D+00	-.3808D+00	.6212D+00	.2636D+00
.2389D+01	.1027D-06	2306	-.6815D+00	-.1090D+00	.7313D+00	.3704D+00
.2730D+01	.9817D-07	2440	-.6776D+00	.1256D+00	.8649D+00	.3981D+00
.3072D+01	.9012D-07	2532	-.6011D+00	.3135D+00	.9935D+00	.3424D+00
.3413D+01	.8016D-07	2606	-.4710D+00	.4360D+00	.1091D+01	.2197D+00
.3754D+01	.7009D-07	2680	-.3129D+00	.4760D+00	.1139D+01	.5699D-01
.4096D+01	.6108D-07	2778	-.1565D+00	.4251D+00	.1129D+01	-.1137D+00
.4437D+01	.5337D-07	2870	-.3299D-01	.2842D+00	.1064D+01	-.2624D+00
.4778D+01	.4601D-07	3010	.2832D-01	.6268D-01	.9546D+00	-.3707D+00
.5120D+01	.3841D-07	3162	.2286D-02	-.2245D+00	.8146D+00	-.4480D+00
.5461D+01	.4467D-07	3290	-.1284D+00	-.5384D+00	.6445D+00	-.5668D+00
.5802D+01	.8599D-07	3436	-.3478D+00	-.6862D+00	.4076D+00	-.8482D+00
.6143D+01	.1083D-06	3636	-.5442D+00	-.4244D+00	.8190D-01	-.9763D+00
.6485D+01	.8771D-07	3842	-.6550D+00	-.2765D+00	-.2083D+00	-.6915D+00
.6826D+01	.8490D-07	3970	-.7557D+00	-.3269D+00	-.3865D+00	-.3586D+00

.7167D+01	.8872D-07	4080	-.8787D+00	-.3867D+00	-.4578D+00	-.6645D-01
.7509D+01	.8895D-07	4214	-.1012D+01	-.3841D+00	-.4365D+00	.1839D+00
.7850D+01	.8521D-07	4306	-.1132D+01	-.3074D+00	-.3383D+00	.3816D+00
.8191D+01	.7915D-07	4380	-.1215D+01	-.1703D+00	-.1840D+00	.5096D+00
.8533D+01	.7280D-07	4478	-.1245D+01	-.1421D-07	-.2928D-07	.5540D+00
.8874D+01	.6753D-07	4588	-.1215D+01	.1703D+00	.1840D+00	.5096D+00
.9215D+01	.6403D-07	4680	-.1132D+01	.3074D+00	.3383D+00	.3816D+00
.9557D+01	.6200D-07	4772	-.1012D+01	.3841D+00	.4365D+00	.1839D+00
.9898D+01	.5942D-07	4894	-.8787D+00	.3867D+00	.4578D+00	-.6645D-01
.1024D+02	.5332D-07	5004	-.7557D+00	.3269D+00	.3865D+00	-.3586D+00
.1058D+02	.4805D-07	5114	-.6550D+00	.2765D+00	.2083D+00	-.6915D+00
.1092D+02	.7436D-07	5314	-.5442D+00	.4244D+00	-.8190D-01	-.9763D+00
.1126D+02	.9743D-07	5514	-.3478D+00	.6862D+00	-.4076D+00	-.8482D+00
.1160D+02	.7855D-07	5678	-.1284D+00	.5384D+00	-.6445D+00	-.5668D+00
.1195D+02	.7285D-07	5806	.2286D-02	.2245D+00	-.8146D+00	-.4480D+00
.1229D+02	.7949D-07	5958	.2832D-01	-.6268D-01	-.9546D+00	-.3707D+00
.1263D+02	.8447D-07	6086	-.3299D-01	-.2842D+00	-.1064D+01	-.2624D+00
.1297D+02	.8697D-07	6178	-.1565D+00	-.4251D+00	-.1129D+01	-.1137D+00
.1331D+02	.8945D-07	6270	-.3129D+00	-.4760D+00	-.1139D+01	.5699D-01
.1365D+02	.9437D-07	6344	-.4710D+00	-.4360D+00	-.1091D+01	.2197D+00
.1399D+02	.1028D-06	6418	-.6011D+00	-.3135D+00	-.9935D+00	.3424D+00
.1433D+02	.1146D-06	6492	-.6776D+00	-.1256D+00	-.8649D+00	.3981D+00
.1468D+02	.1280D-06	6614	-.6815D+00	.1090D+00	-.7313D+00	.3704D+00
.1502D+02	.1407D-06	6724	-.5991D+00	.3808D+00	-.6212D+00	.2636D+00
.1536D+02	.1568D-06	6852	-.4152D+00	.7097D+00	-.5547D+00	.1326D+00
.1570D+02	.2211D-06	7034	-.1059D+00	.1100D+01	-.5092D+00	.2024D+00
.1604D+02	.2526D-06	7318	.2880D+00	.1073D+01	-.3639D+00	.6533D+00
.1638D+02	.1526D-06	7518	.5797D+00	.6689D+00	-.1316D+00	.6037D+00
.1672D+02	.8206D-07	7694	.7801D+00	.5482D+00	.1632D-01	.2496D+00
.1707D+02	.9599D-05	8722	.9940D+00	.8256D-05	.5247D-07	-.2002D+01

ERGEBNIS AUS DORPRI54

T= 17.06521656015796

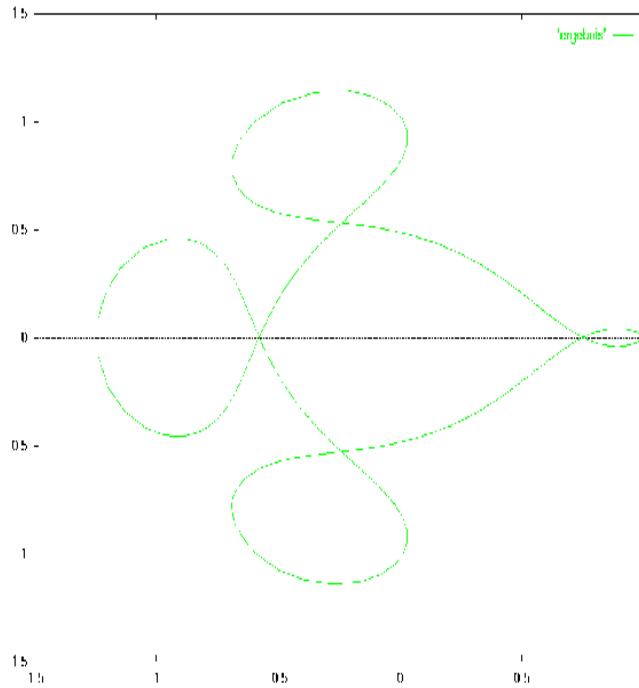
LOESUNG

Y(1)= 0.9940000D+00 Y(2)= 0.8255578D-05

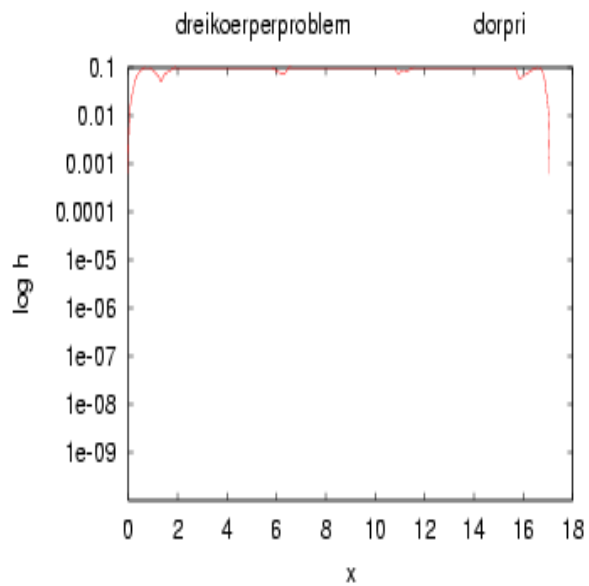
Y(3)= 0.5247239D-07 Y(4)=-0.2001586D+01

SCHAETZUNG DES GLOBALEN FEHLERS 9.5986607319577849E-06

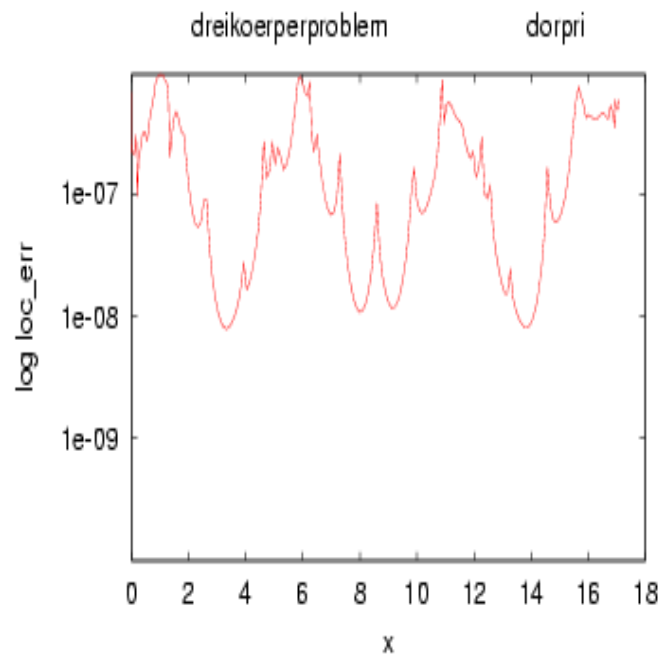
Die folgenden Abbildungen zeigen das Phasendiagramm $(y_1, y_2)(t)$ der Lösung sowie den Verlauf der Schrittweite, des lokalen und des globalen Fehlers über der Zeitachse.



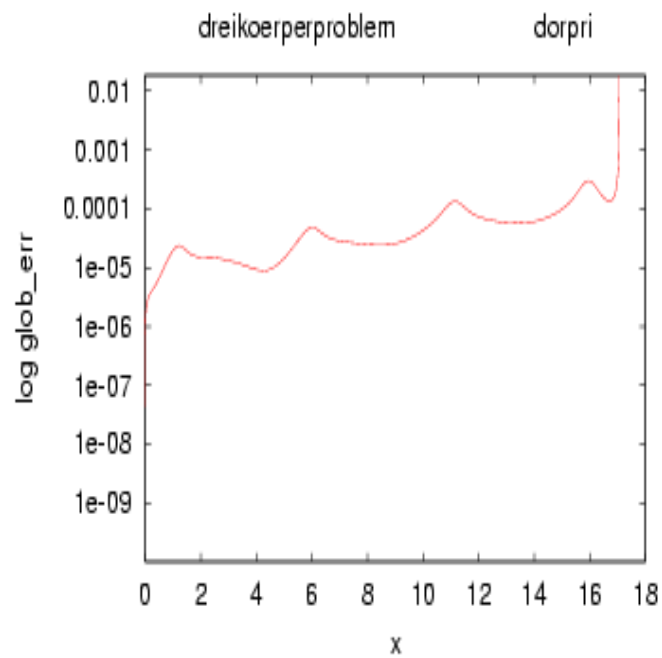
Phasendiagramm der Lösung $y_1(t), y_2(t)$



Schrittweiten über Zeitachse



lokaler Diskretisierungsfehler über Zeitachse



globaler Diskretisierungsfehler über Zeitachse

1.5.3 Kontinuierliche Näherungsformeln

In der Praxis ist es häufig erforderlich, die numerisch berechnete Lösung einer Differentialgleichung auf einem benutzerdefinierten Gitter anzugeben, während der Integrator

selbst aufgrund der Schrittweitensteuerung intern ein völlig anderes Gitter erzeugt. Mit Methoden der (lokalen oder globalen) Polynom- oder Splineinterpolation könnte man eine stetige Interpolierende auf diesem internen Gitter definieren und danach die Funktion auf dem vorgeschriebenen Gitter auswerten. Dabei muss man berücksichtigen, daß der Interpolationsfehler den globalen Diskretisierungsfehler nicht dominieren darf. Dies erfordert also ein u.U. aufwendiges Postprocessing der diskret berechneten Lösung. Es ist deshalb von grossem praktischen Interesse, daß es auch Integrationsformeln vom Runge–Kutta–Typ gibt, die direkt eine stetige stückweise definierte Approximation der Lösung auf dem gesamten Integrationsintervall liefern. Praktisch interessant sind hierbei die Formeln, bei denen die k_i wie zuvor berechnet werden und nur die Koeffizienten γ_j jetzt von einer Variablen τ abhängen mit $\tau \in [0, 1]$, wo $\gamma(1)$ den “alten“ γ entspricht, also den nächsten Gitterwert $y(t_{k+1})$ generiert (engl. “continuous Runge-Kutta methods“). Zwei Beispiele seien hier genannt:

Beispiel 1.5.2. *3-stufiges Verfahren der Ordnung 3 mit stetiger Interpolation der Ordnung 2:*

i	α_i	$\beta_{i,1}$	$\beta_{i,2}$	$\beta_{i,3}$
1	0.0	0	0	0
2	0.5	0.5	0	0
3	1.0	-1.0	2.0	0

mit den drei Gamma-Werten

$$\begin{aligned}\gamma_1(\tau) &= \tau(1 + \tau(\frac{2}{3}\tau - \frac{3}{2})) \\ \gamma_2(\tau) &= \tau^2(2 - \frac{2}{3}\tau) \\ \gamma_3(\tau) &= \tau^2(\frac{2}{3} - \frac{1}{2}\tau)\end{aligned}$$

und

Beispiel 1.5.3. *Im Zusammenhang mit der siebenstufigen Formel von Dormand und Prince der Ordnung 5(4) (s.o.) erhält man mit Hilfe der hier angegebenen Werte $\gamma_i(\tau)$ eine Interpolierende der globalen Ordnung 4:*

$$\begin{aligned}\gamma_1(\tau) &= \tau(1 + \tau(-1337/480 + \tau(1039/360 + \tau(-1163/1152)))) \\ \gamma_2(\tau) &= 0 \\ \gamma_3(\tau) &= 100\tau^2(1054/9275 + \tau(-4682/27825 + \tau(379/5565)))/3 \\ \gamma_4(\tau) &= -5\tau^2(27/40 + \tau(-9/5 + \tau(83/96)))/2 \\ \gamma_5(\tau) &= 18225\tau^2(-3/250 + \tau(22/375 + \tau(-37/600)))/848 \\ \gamma_6(\tau) &= -22\tau^2(-3/10 + \tau(29/30 + \tau(-17/24)))/7 \\ \gamma_7(\tau) &= 0\end{aligned}$$

Man beachte $\gamma_i(1) = \gamma_i$ mit den oben angegebenen Werten.

1.6 Mehrschrittverfahren. Motivation und Herleitung einiger elementarer Verfahren

Bei den ESV ist eine Steigerung der Verfahrensordnung nur möglich durch eine Steigerung der Anzahl der f -Auswertungen pro Iterationsschritt (vielleicht sogar Auswertungen der Jacobimatrix). Hat man nun schon auf den Gitterpunkten $t_{i-p}, t_{i-p+1}, \dots, t_i$ Näherungswerte $\eta_{i-p}, \dots, \eta_i$ für die Lösung y gefunden, so ist es naheliegend, diese Information zur Konstruktion von (t_{i+1}, η_{i+1}) auszunutzen (statt nur (t_i, η_i) zu benutzen), um hohe Verfahrensordnung mit möglichst wenigen zusätzlichen f -Auswertungen pro Schritt zu erzielen. Folgende Zugänge bieten sich unmittelbar an:

1.6.1 Verfahren, die auf numerischer Integration beruhen

Es gilt

$$y(t_{i+1}) = y(t_{i-j}) + \int_{t_{i-j}}^{t_{i+1}} f(\tau, y(\tau)) d\tau.$$

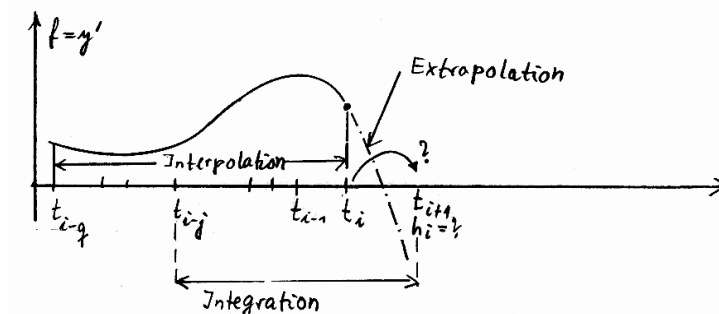
Ersetzt man nun $f(\tau, y(\tau))$ durch ein Interpolationspolynom durch die Näherungspunkte $(t_{i-q}, f_{i-q}), \dots, (t_{i+m}, f_{i+m})$, dann erhält man die Näherungsformel für η_{i+1} :

$$\eta_{i+1} = \eta_{i-j} + \sum_{l=-q}^m f_{i+l} \underbrace{\int_{t_{i-j}}^{t_{i+1}} \prod_{\substack{k=-q \\ k \neq l}}^m \frac{t - t_{k+i}}{t_{l+i} - t_{k+i}} dt}_{\beta_{i+l}(t_{i-q}, \dots, t_{i+m}, t_{i-j}, t_{i+1})}$$

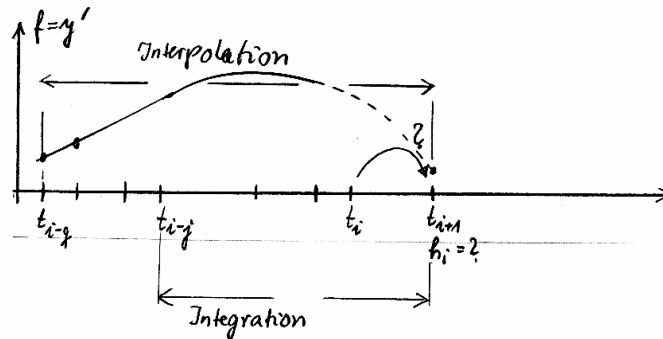
$$f_{i+l} = f(t_{i+l}, \eta_{i+l})$$

Bei diesem Zugang sind also auch nicht-äquidistante t_j zugelassen. (Bei der Theorie werden wir uns später auf den einfachen äquidistanten Fall beschränken.)

$m = 0$, explizites Verfahren, j kann beliebig ≥ 0 gewählt werden .



$m = 1$, implizites Verfahren, j kann beliebig ≥ 0 gewählt werden .



Die verschiedenen Möglichkeiten, die Parameter m, q und j zu wählen, führen zu einer Fülle von Verfahren. Von praktischem Interesse sind allerdings nur die beiden Fälle

- $m = 0, j = 0, q \in \mathbb{N}_0$: **Adams–Bashforth–Verfahren**
- $m = 1, j = 0, q \in \mathbb{N}_0$: **Adams–Moulton–Verfahren.**

Falls gilt $t_i = t_0 + ih \quad (\forall i)$, dann können die Integrale der Lagrange–Polynome explizit **im voraus** berechnet werden. Man erhält dann die Formeln

$$\eta_{i+1} = \eta_i + h \sum_{l=0}^q \beta_{ql}^* f_{i-l} \quad \text{Adams–Bashforth}$$

bzw.

$$\eta_{i+1} = \eta_i + h \sum_{l=0}^{q+1} \beta_{q+1,l} f_{i+1-l} \quad \text{Adams–Moulton}$$

mit

β_{ql}^*	$q \setminus l$	0	1	2	3	
	0	1				
	1	3/2	-1/2			
	2	23/12	-16/12	5/12		
	3	55/24	-59/24	37/24	-9/24	
$\beta_{q+1,l}$	$1 + q \setminus l$	0	1	2	3	4
	1	1/2	1/2			
	2	5/12	8/12	-1/12		
	3	9/24	19/24	-5/24	1/24	
	4	251/720	646/720	-264/720	106/720	-19/720

1.6.2 Verfahren, die auf numerischer Differentiation beruhen

Das einzige Verfahren, das in dieser Klasse von praktischem Interesse ist, entsteht auf folgende Weise:

Man interpoliert die Werte $(t_{i-q}, \eta_{i-q}), \dots, (t_{i+1}, \eta_{i+1})$ durch ein Polynom vom Grad $q + 1$,

differenziert dieses Polynom und setzt den Wert der Ableitung an der Stelle t_{i+1} gleich $f(t_{i+1}, \eta_{i+1})$. (dies entspricht $y'(t_{i+1})$): Im äquidistanten Fall wird mit

$$P_{q+1}(t) = \sum_{k=0}^{q+1} (-1)^k \binom{-\frac{t-t_{i+1}}{h}}{k} \nabla^k \eta_{i+1}$$

(Interpolationspolynom in Rückwärtsdifferenzen ausgedrückt)

$$P'_{q+1}(t_{i+1}) = \frac{1}{h} \sum_{k=1}^{q+1} \rho_k \nabla^k \eta_{i+1} = f(t_{i+1}, \eta_{i+1}),$$

das heißt man erhält ein **implizites Verfahren**. Die Koeffizienten ρ_k lauten

$$\rho_0 = 1, \quad \rho_k = \frac{1}{k}, \quad k = 1, 2, \dots$$

Von praktischem Interesse sind nur die Fälle $q = 0, \dots, 5$, d.h. ein Polynomgrad zwischen 1 und 6.

Diese Formeln sind interessant im Zusammenhang mit steifen Gleichungen (bekannt als “rückwärts genommene Differentiationsformeln” BDF oder auch “Gearsche Methode”).

1.7 Lineare Differenzgleichungen mit konstanten Koeffizienten

Bei den in Abschnitt 1.6 hergeleiteten Verfahren erhalten wir die folgende allgemeine Struktur eines linearen Mehrschrittverfahrens mit konstanten Koeffizienten:

$$\sum_{j=0}^k \alpha_j \eta_{j+i} = h \sum_{j=0}^k \beta_j f(t_{j+i}, \eta_{j+i}) \quad i = 0, \dots, N_h - k, \quad (1.20)$$

k -Schrittverfahren: $\alpha_k \neq 0, \quad |\alpha_0| + |\beta_0| \neq 0$.

Um dieses Verfahren zu starten, benötigt man noch die Startwerte

$\eta_0 \stackrel{\text{def}}{=} y_0, \quad \eta_1, \dots, \eta_{k-1}$. (In der Praxis z.B. aus einem ESV)

Falls dann

$$\frac{1}{|\alpha_k|} h |\beta_k| L < 1$$

mit L aus (V0), dann ist η_{i+k} aus (1.20) eindeutig bestimmt und durch direkte Iteration berechenbar (vgl. jedoch 1.4 a). Ist f von y unabhängig, dann erhält man die **lineare (in)homogene Differenzgleichung**

$$\sum_{j=0}^k \alpha_j \eta_{j+i} = g_{i+k} \quad i = 0, 1, 2, \dots \quad (1.21)$$

(mit $g_{i+k} \stackrel{\text{def}}{=} h \sum_{j=0}^k \beta_j f(t_{j+i})$).

In diesem Abschnitt wollen wir uns mit dem Lösungsverhalten dieser Differenzgleichung beschäftigen. Im Folgenden gelte

$$\alpha_k \alpha_0 \neq 0.$$

sodaß tatsächlich eine Kopplung von $k + 1$ konsekutiven Gitterwerten vorliegt.

Definition 1.7.1. Eine Folge $\{u_i\}_{i \in \mathbb{N}_0}$ heißt **Lösung von (1.21)**, falls mit $\eta_{j+i} \stackrel{\text{def}}{=} u_{j+i}$ (1.21) erfüllt ist für alle i .
 r Folgen $\{u_i^{(\nu)}\}_{i \in \mathbb{N}_0}$, $\nu = 1, \dots, r$ heißen linear unabhängig, wenn

$$\sum_{\nu=1}^r \beta_\nu u_i^{(\nu)} = 0 \quad \forall i \in \mathbb{N}_0 \quad \Rightarrow \quad \beta_1 = \dots = \beta_r = 0.$$

Ein System von k linear unabhängigen Lösungen der homogenen Gleichung ($g_j = 0 \quad \forall j$) heißt ein **Fundamentalsystem**. □

Im Folgenden zeigen wir, daß eine solche Differenzgleichung umgeschrieben werden kann auf eine gewöhnliche Vektoriteration für ein System und bestimmen dann die Lösungsmannigfaltigkeit mit Hilfe der Theorie der Jordannormalform einer Matrix.

<<

Wir betrachten zunächst den homogenen Fall:

$$\sum_{j=0}^k \alpha_j \eta_{j+i} = 0 \quad i = 0, 1, \dots \quad \text{o.B.d.A.} \quad \alpha_k = 1, \quad \alpha_0 \neq 0. \quad (1.22)$$

Mit

$$y_i \stackrel{\text{def}}{=} \begin{pmatrix} \eta_i \\ \vdots \\ \eta_{i+k-1} \end{pmatrix} \in \mathbb{C}^k$$

erhalten wir

$$y_{i+1} = A y_i = A^{i+1} y_0, \quad A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ & 0 & 1 & & \vdots \\ & & \ddots & \ddots & 0 \\ & & & \ddots & 1 \\ -\alpha_0 & -\alpha_1 & & & -\alpha_{k-1} \end{pmatrix}.$$

Die Matrix A ist eine sogenannte Frobeniusmatrix mit dem charakteristischen Polynom

$$\psi(\lambda) = \det(A - \lambda I) = (-1)^k (\lambda^k + \alpha_{k-1} \lambda^{k-1} + \dots + \alpha_0).$$

Zu jedem der verschiedenen Eigenwerte von A (Nullstellen von ψ) gehört nur ein Jordan-Kästchen (denn $A - \lambda I$ hat den Rang $\geq k - 1$):

$$T^{-1}AT = \begin{pmatrix} \mathcal{J}_1 & & \\ & \ddots & \\ & & \mathcal{J}_r \end{pmatrix} =: \mathcal{J}$$

mit

$$\mathcal{J}_s = \begin{pmatrix} \lambda_s & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_s \end{pmatrix} =: \lambda_s I + \hat{\mathcal{J}}_s \quad \in \mathbb{C}^{n_s \times n_s}$$

und

$$\sum_{s=1}^r n_s = k$$

Setzen wir also

$$z_i \stackrel{def}{=} Ty_i$$

dann wird

$$z_i = \mathcal{J}^i z_0 = \begin{pmatrix} \mathcal{J}_1^i & & \\ & \ddots & \\ & & \mathcal{J}_r^i \end{pmatrix} z_0 = \begin{pmatrix} \mathcal{J}_1^i z_{0,1} \\ \vdots \\ \mathcal{J}_r^i z_{0,r} \end{pmatrix}.$$

Nun ist aber

$$\mathcal{J}_s^i = (\lambda_s I + \hat{\mathcal{J}}_s)^i = \sum_{j=0}^i \binom{i}{j} \lambda_s^{i-j} \hat{\mathcal{J}}_s^j.$$

Wegen

$$\hat{\mathcal{J}}_s^k = \begin{cases} \begin{pmatrix} 0 & & & & & \\ 0 & \cdots & 0 & 1 & & 0 \\ \vdots & & & & \ddots & \\ \vdots & & & & & 1 \\ \vdots & & & & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix} & k \geq n_s \\ \leftarrow n_s - k \end{cases}$$

und

$$\binom{i}{j} = \frac{1}{j!} (i - j + 1) \cdots i = \sum_{l=0}^j \gamma_{l,j} i^l \stackrel{def}{=} p_j(i)$$

mit von i unabhängigen Koeffizienten $\gamma_{l,j}$ (die $\gamma_{l,j}$ sind die Koeffizienten des Polynoms mit den Nullstellen $0, \dots, j - 1$ und dem Höchstkoeffizienten $1/j!$) ergibt sich für $i \geq$

$$k \geq n_s, \quad s = 1, \dots, r$$

$$\begin{aligned} \mathcal{J}_s^i &= \sum_{j=0}^{n_s-1} \binom{i}{j} \lambda_s^{i-j} \hat{\mathcal{J}}_s^j = \sum_{j=0}^{n_s-1} \underbrace{\left(\sum_{l=0}^j \gamma_{l,j} i^l \right)}_{p_j(i)} \lambda_s^{i-j} \hat{\mathcal{J}}_s^j \\ & p_j \text{ Polynom vom genauen Grad } j, \quad p_0 \equiv 1 \\ &= \begin{pmatrix} p_0(i)\lambda_s^i & p_1(i)\lambda_s^{i-1} & \dots & p_{n_s-1}(i)\lambda_s^{i+1-n_s} \\ \vdots & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & p_0(i)\lambda_s^i \end{pmatrix}. \end{aligned}$$

Daher wegen

$$\begin{aligned} y_i &= T \mathcal{J}^i z_0 \\ &= \begin{pmatrix} \vdots & \vdots \\ A_1 & \vdots & A_2 & \vdots & A_r \\ \vdots & \vdots \end{pmatrix} \mathcal{J}^i z_0 \\ \text{mit } A_1 &= \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \lambda_1 & 1 & 0 & & \\ \lambda_1^2 & 2\lambda_1 & 1 & & \\ & & & \ddots & \\ \dots & \dots & \dots & \dots & 1 \\ \lambda_1^{k-1} & (k-1)\lambda_1^{k-2} & \dots & \dots & \dots \end{pmatrix} \\ & \underbrace{\hspace{10em}}_{n_1} \\ A_2 &= \begin{pmatrix} 1 & 0 & \dots \\ \lambda_2 & & \\ \vdots & & \\ \vdots & & \\ \vdots & & \\ \lambda_2^{k-1} & \dots & \dots \end{pmatrix} \quad A_r = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \lambda_r & 1 & & \\ \vdots & & \ddots & \\ \vdots & & & \\ \vdots & & & \\ \lambda_r^{k-1} & \dots & \dots & \underbrace{(\frac{k-1}{n_r-1})\lambda_r^{k-n_r}}_{n_r} \end{pmatrix} \\ \Rightarrow \eta_i &= e_1^T y_i = e_1^T T \mathcal{J}^i z_0 = \sum_{j=1}^r \underbrace{e_1^T}_{\in \mathbb{R}^{n_j \times n_j}} \mathcal{J}_j^i z_{0,i} \\ &= \sum_{j=1}^r \lambda_j^i \sum_{l=1}^{n_j} z_{0,j,l} \lambda_j^{1-l} p_{l-1}(i) \end{aligned}$$

wobei gesetzt wurde

$$z_0^T = (z_{0,1}^T, \dots, z_{0,r}^T) = (z_{0,1,1}, z_{0,1,2}, \dots, z_{0,1,n_1}, z_{0,2,1}, \dots, z_{0,r,n_r}).$$

Die Polynome $p_j(t)$ mit den Nullstellen $0, \dots, j-1$ bilden eine Basis des Π_j . Da die $z_{0,j,l}$ und damit auch $z_{0,j,l}\lambda_j^{1-l}$ beliebig wählbar sind durch geeignete Vorgabe von $y_0 = (\eta_0, \dots, \eta_{k-1})^T$ folgt

>>

Satz 1.7.1. Die Lösungsgesamtheit der linearen Differenzgleichung (1.21) ($\alpha_k \alpha_0 \neq 0$) bildet einen k -dimensionalen linearen Vektorraum über \mathbb{C} . Ein Fundamentalsystem wird gebildet durch die Folgen $\{p_j(i)\lambda_s^i\}_{i \in \mathbb{N}_0}$ mit $1 \leq s \leq r$ und $0 \leq j \leq n_s - 1$, worin λ_s die verschiedenen Nullstellen des charakteristischen Polynoms bezeichnet, n_s ihre Vielfachheit und

$$p_0 \equiv 1, \quad p_j(t) \equiv \frac{1}{j!} t(t-1) \cdots (t-j+1), \quad j \geq 1.$$

□

Bisher haben wir den Fall $\alpha_0 = 0$ ausgeschlossen. Dieser Fall ist jedoch trivial:

Satz 1.7.2. Habe das der Differenzgleichung zugeordnete charakteristische Polynom die Form

$$\psi(\lambda) = \psi_0(\lambda)\lambda^r \quad \text{mit } \psi_0(0) \neq 0.$$

Dann besitzt die Differenzgleichung neben den $k-r$ nach Satz 1.7.1 zu bestimmenden Fundamentallösungen als r übrige

$$\left\{ \sum_{s=0}^{r-1} \gamma_s \delta_{j,s} \right\}_{j \in \mathbb{N}_0}$$

mit frei wählbaren γ_s , also als Ergänzung des Fundamentalsystems die ersten r Einheitsvektoren in $\mathbb{R}^{\mathbb{N}}$. □

Die Lösung der inhomogenen Gleichung kann man mit Hilfe eines Einheitsfundamentalsystems (d.h. es ist $u_j^{(\nu)} = \delta_{j,\nu-1}$ $0 \leq j \leq k-1$, $1 \leq \nu \leq k$) geschlossen angeben. Wir benötigen jedoch nur spezielle Inhomogenitäten, bei denen sich (wie bei Differentialgleichungen) eine Partikulärlösung mittels eines geeigneten Ansatzes bestimmen läßt gemäß folgender Tabelle:

$g(n)$	Ansatzfunktion
$A\rho^n$	$c\rho^n n^\nu$
An^s	$\left(\sum_{j=0}^s c_j n^j\right) n^\nu$
$An^s \rho^n$	$n^\nu \rho^n \sum_{j=0}^s c_j n^j.$

Dabei wird der Exponent ν ganzzahlig und minimal so gewählt, daß die Ansatzfunktion keine Lösung der homogenen Gleichung mehr ist.

Beispiel 1.7.1.

$$\eta_{n+2} - 5\eta_{n+1} + 6\eta_n = 4n + 6 \cdot 2^n$$

Homogene Gleichung hat die allgemeine Lösung $c_1 2^n + c_2 3^n$.

Partikuläranatz: $\eta_n^{(p)} = a_1 + a_2 n + a_3 n 2^n$

$$\begin{aligned} a_1 + a_2(n+2) + a_3(n+2)2^{n+2} - 5a_1 - 5a_2(n+1) - 5a_3(n+1)2^{n+1} + \\ + 6a_1 + 6a_2 n + 6a_3 n 2^n \stackrel{!}{=} 4n + 6 \cdot 2^n \quad (\forall n) \end{aligned}$$

\Rightarrow Koeffizientenvergleich ergibt:

$$\begin{aligned} 2a_1 - 3a_2 &= 0 \\ 2a_2 &= 4 \\ -2a_3 &= 6 \\ (a_3 4n 2^n - a_3 10n 2^n + 6a_3 n 2^n) &= 0 \end{aligned}$$

d.h. $a_1 = 3, \quad a_2 = 2, \quad a_3 = -3.$ □

Das Wachstumsverhalten der Lösung der homogenen Differenzgleichung läßt sich mittels der oben bereits betrachteten Darstellung

$$\begin{aligned} z_i &\stackrel{def}{=} T y_i, \quad y_i \stackrel{def}{=} (\eta_i, \dots, \eta_{i+k-1}) \\ z_i &= \begin{pmatrix} \mathcal{J}_1^i z_{0,1} \\ \vdots \\ \mathcal{J}_r^i z_{0,r} \end{pmatrix} \end{aligned}$$

unmittelbar angeben:

Satz 1.7.3. Sei $\{\eta_i\}_{i \in \mathbb{N}_0}$ Lösung der homogenen Differenzgleichung

$$\sum_{j=0}^k \alpha_j \eta_{j+i} = 0 \quad . \text{ Dann gilt}$$

$$\eta_0, \dots, \eta_{k-1} \text{ bel. und } \lim_{i \rightarrow \infty} \frac{\eta_i}{i} = 0 \quad \Leftrightarrow$$

ψ erfüllt die Stabilitätsbedingung (S)

(S): λ Nullstelle von $\psi \Rightarrow |\lambda| \leq 1$ und $\psi'(\lambda) \neq 0$, falls $|\lambda| = 1$. □

Beweis: Offensichtlich gilt

$$\begin{aligned} \lim_{i \rightarrow \infty} \frac{\eta_i}{i} = 0 &\Leftrightarrow \lim_{i \rightarrow \infty} \frac{1}{i} \|y_i\| = 0 \\ &\Leftrightarrow \lim_{i \rightarrow \infty} \frac{1}{i} \|z_i\| = 0. \end{aligned}$$

Zerlegen wir z_i entsprechend der Zerlegung von \mathcal{J} , dann ergibt dies

$$\lim_{i \rightarrow \infty} \frac{\eta_i}{i} = 0 \quad \Leftrightarrow \quad \lim_{i \rightarrow \infty} \frac{1}{i} \|z_{i,j}\| = 0 \quad j = 1, \dots, r.$$

Nun ist aber

$$z_{i,j,n_j} = \lambda_j^i z_{0,j,n_j}$$

und

$$z_{i,j,n_j-1} = \lambda_j^i z_{0,j,n_j-1} + i \lambda_j^{i-1} z_{0,j,n_j} \quad \text{falls } n_j > 1.$$

Da die Komponenten von z_0 beliebig wählbar sind (weil $\eta_0, \dots, \eta_{k-1}$ frei wählbar) ergibt dies

$$\lim_{i \rightarrow \infty} \frac{\eta_i}{i} = 0 \quad \Leftrightarrow \quad \lim_{i \rightarrow \infty} \frac{\lambda_j^i}{i} = 0$$

und falls die Vielfachheit $n_j > 1$ auch

$$\lim_{i \rightarrow \infty} \lambda_j^{i-1} = 0 \quad \Leftrightarrow \quad |\lambda_j| < 1 \quad (\forall j : n_j > 1)$$

und $\psi'(\lambda_j) \neq 0$ falls $|\lambda_j| = 1$ □

1.8 Allgemeine Theorie der Mehrschrittverfahren

Wir wollen in diesem Abschnitt Mehrschrittverfahren der Form

$$\sum_{j=0}^k \alpha_j \eta_{i+j} = h \Phi(\eta_i, \dots, \eta_{i+k}, t_i, h; f) \quad \alpha_k \neq 0 \quad (1.23)$$

betrachten. Die in Abschnitt 1.6 hergeleiteten linearen MSV sind hierin enthalten mit

$$\Phi(\eta_i, \dots, \eta_{i+k}, t_i, h; f) = \sum_{j=0}^k \beta_j f(t_{j+i}, \eta_{j+i}).$$

(1.23) ist ein k -**Schritt-Verfahren**, falls $\alpha_0 \neq 0$ oder Φ von η_i tatsächlich abhängt. Falls Φ von η_{i+k} abhängt, dann ist das Verfahren implizit, sonst explizit. Φ erfülle stets folgende Voraussetzungen:

$$(V2) \left\{ \begin{array}{l} \Phi(\dots; 0) \equiv 0 \quad \Phi \in C(\mathbb{R}^{n(k+1)} \times \mathcal{J} \times [0, h_0]) \\ \|\Phi(u_1, \dots, u_{k+1}, t, h; f) - \Phi(w_1, \dots, w_{k+1}, t, h; f)\| \leq \sum_{j=1}^{k+1} L_j \|u_j - w_j\| \\ \text{mit geeigneten Konstanten } L_j \\ (\forall u_i, w_j \in \mathbb{R}^n, \quad \forall t \in \mathcal{J}_0, \quad \forall h \in [0, h_0], \quad h_0 > 0). \end{array} \right.$$

(Bei einem linearen MSV ist dies natürlich trivial gegeben mit $L_j = |\beta_j|L$ und L aus (V0).) Wir definieren nun analog zur Vorgehensweise bei den Einschrittverfahren:

$$G_h \stackrel{def}{=} \{t_0 + ih = t_i : 0 \leq i \leq N_h, \quad h = \frac{t - t_0}{N_h}\} \quad t \in \mathcal{J}_0 = [t_0, t_E]$$

$$G'_h \stackrel{def}{=} \{t_i : k \leq i \leq N_h\}$$

$$\eta(t; h) \stackrel{def}{=} \eta_{N_h}$$

$$\varepsilon(t; h) \stackrel{def}{=} \eta(t; h) - y(t) \quad y \text{ Lösung von (A).}$$

Die Größe $\varepsilon(t; h)$ heißt der **globale Diskretisierungsfehler** (auf G_h). Der lokale Abschneidefehler wird wie bei ESV definiert:

Definition 1.8.1. Sei y Lösung von (A) mit $y(t - kh) = z$.
(D.h. der Anfangswert ist an der Stelle $t - kh$ vorgeschrieben.) Dann heißt

$$\tau(t, z; h) = \begin{cases} \eta_i - y_i & \text{falls } t = t_i, \quad 0 \leq i \leq k - 1 \\ \frac{1}{h} \left(- \sum_{j=0}^k \alpha_j y(t + (j - k)h) + \right. \\ \left. + h\Phi(y(t - kh), \dots, y(t), t - kh, h; f) \right), & t \in G'_h \end{cases}$$

der lokale Abschneidefehler des Verfahrens. □

Definition 1.8.2. Das Verfahren heißt **konsistent von der Ordnung p** , falls

$$\begin{aligned} \|\tau(t, y(t - kh); h)\| &\leq Kh^p & t \in G'_h \\ \|\eta_i - y_i\| &\leq Kh^p & i = 0, \dots, k - 1 \end{aligned}$$

wobei y Lösung von (A) ist. Die Größe:

$$p^* \stackrel{def}{=} \sup\{p : \|\tau(t, y(t - kh); h)\| \leq Kh^p, \quad t \in G'_h, \quad 0 < h \leq h_0, \\ y \text{ Lsg. von (A), } f \in F^\infty(\mathcal{J} \times \mathbb{R}^n) \text{ bel.}\}$$

heißt die **Verfahrensordnung**. Das Verfahren heißt **konvergent, falls** mit $\lim_{h \rightarrow 0} \eta_i = y_0$ für $0 \leq i \leq k - 1$ gilt: $\sup\{\|\varepsilon(t; h)\| : t \in G'_h\} \rightarrow 0$ für $h \rightarrow 0$. □

Der lokale Diskretisierungsfehler beschrieb bei ESV den pro Schritt neu hinzukommenden Fehler. Wir haben ihn dort dem Zeitwert dieses Schrittes, also der "alten" Zeit, zugeordnet, während der Abschneidefehler dem neuen Zeitwert zugeordnet wurde. Dies setzen wir hier konsistent fort. Bei einem Mehrschrittverfahren kann man den lokalen Fehler nur dadurch sinnvoll definieren, daß man annimmt, die zurückliegenden $k - 1$ η -Werte seien exakt.

Definition 1.8.3. Sei y Lösung von (A) mit $y(t - kh) = z$ für $t \in G'_h$ bel. und $\eta_i \stackrel{\text{def}}{=} y(t - (k - i)h)$ $i = 0, \dots, k - 1$.
Dann heißt mit η_k Lösung von (1.23)

$$\delta(t - kh, z; h) = \frac{1}{\alpha_k} \left(- \sum_{i=0}^{k-1} \alpha_i \eta_i + h \Phi(\eta_0, \dots, \eta_{k-1}, \eta_k, t - kh, h; f) \right) - y(t)$$

der lokale Diskretisierungsfehler des Verfahrens an der Stelle $(t - kh, z)$. \square

Während bei einem ESV

$$\tau(t, y(t - h); h) = \frac{1}{h} \delta(t - h, y(t - h); h)$$

war, gilt hier nur

$$\tau(t, y(t - kh); h) = \frac{1}{h} (I + \mathcal{O}(h)) \delta(t - kh, y(t - kh); h) \quad \text{falls } \alpha_k = 1.$$

Nach der Definition des lokalen Abschneidefehlers gilt

$$\sum_{j=0}^k \alpha_j y_{j+i} = h \Phi(y_i, \dots, y_{i+k}, t_i, h; f) + h \tau(t_{i+k}, y_i; h),$$

während

$$\sum_{j=0}^k \alpha_j \eta_{j+i} = h \Phi(\eta_i, \dots, \eta_{i+k}, t_i, h; f).$$

Die entstehende Differenzgleichung für die y -Werte können wir als eine Störung der Differenzgleichung für die η -Werte interpretieren und wir können erwarten, daß $\eta_i - y_i \rightarrow 0$, wenn sich die Auswirkung dieser Störung bei "kleinem" h und $\eta_j - y_j$ "klein" für $j = 0, \dots, k - 1$ nur "wenig" bemerkbar macht. Dazu

Definition 1.8.4. Das MSV (1.23) heißt D -stabil (asymptotisch stabil) falls gilt:

$$\begin{aligned} \exists h_0 > 0 : \quad \forall \varepsilon > 0 \quad \exists \delta > 0 : \quad \forall h \in]0, h_0[\\ \sup_{t \in G_h} \|r(t; h)\| \leq \delta \quad \Rightarrow \quad \sup_{t \in G_h} \|\tilde{\eta}(t; h) - \eta(t; h)\| \leq \varepsilon, \quad \text{wo} \\ \eta(t_i; h) \stackrel{\text{def}}{=} \eta_i, \quad \tilde{\eta}(t_i, h) \stackrel{\text{def}}{=} \tilde{\eta}_i \quad \text{und} \\ \sum_{j=0}^k \alpha_j \eta_{j+i} = h \Phi(\eta_i, \dots, \eta_{i+k}, t_i, h; f), \quad i = 0, \dots, N_h - k \\ \sum_{j=0}^k \alpha_j \tilde{\eta}_{j+i} = h \Phi(\tilde{\eta}_i, \dots, \tilde{\eta}_{i+k}, t_i, h; f) + h r(t_{i+k}, h) \\ \tilde{\eta}_j = \eta_j + r(t_j, h) \quad j = 0, \dots, k - 1. \end{aligned}$$

\square

Satz 1.8.1. *Es gelte (V2). Dann ist das MSV D -stabil genau dann, wenn das dem Verfahren zugeordnete Polynom*

$$\rho(t) \stackrel{def}{=} \sum_{i=0}^k \alpha_i t^i$$

die Nullstellenbedingung (S) erfüllt. □

<<

Beweis: Seien η_i und $\tilde{\eta}_i$ gegeben aus Def. 1.8.4 und

$$\begin{aligned} s(t_{i+k}; h) &\stackrel{def}{=} \Phi(\tilde{\eta}_i, \dots, \tilde{\eta}_{i+k}, t_i, h; f) - \Phi(\eta_i, \dots, \eta_{i+k}, t_i, h; f) + r(t_{i+k}, h) \\ d_i &\stackrel{def}{=} \tilde{\eta}_i - \eta_i \quad i = 0, \dots, N_h. \end{aligned}$$

Dann gilt

$$\sum_{j=0}^k \alpha_j d_{j+i} = h s(t_{i+k}; h)$$

sowie

$$\|s(t_{i+k}; h)\| \leq \sum_{j=0}^k L_{j+1} \|d_{j+i}\| + \|r(t_{i+k}, h)\|.$$

Mit

$$D_m^{[i]} \stackrel{def}{=} (e_i^T d_m, \dots, e_i^T d_{m+k-1})^T \quad e_i = (0, \dots, 0, \underbrace{1}_{=i}, 0, \dots, 0)^T$$

$$\beta_m^{[i]} \stackrel{def}{=} e_i^T s(t_{m+k}; h) \quad 1 \leq i \leq n$$

$$\alpha_j^* \stackrel{def}{=} \alpha_j / \alpha_k$$

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -\alpha_0^* & \cdots & \cdots & \cdots & -\alpha_{k-1}^* \end{pmatrix}$$

wird

$$D_{m+1}^{[i]} = AD_m^{[i]} + h\beta_m^{[i]} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^k.$$

Setzt man

$$D_m \stackrel{def}{=} \begin{pmatrix} D_m^{[1]} \\ \vdots \\ D_m^{[n]} \end{pmatrix} \quad (\text{Permutation des Vektors } \begin{pmatrix} d_m \\ \vdots \\ d_{m+k-1} \end{pmatrix})$$

dann erhält man

$$D_{m+1} = \begin{pmatrix} A & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A \end{pmatrix} D_m + h \begin{pmatrix} \beta_m^{[1]} e_k \\ \vdots \\ \beta_m^{[n]} e_k \end{pmatrix}.$$

Sei die Nullstellenbedingung (S) erfüllt. Dann existiert eine Norm $\|\cdot\|_A$ auf \mathbb{C}^{nk} , sodaß in der zugeordneten Matrixnorm

$$\left\| \begin{pmatrix} A & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A \end{pmatrix} \right\|_A \leq 1.$$

Also

$$\|D_{m+1}\|_A \leq \|D_m\|_A + h \left\| \begin{pmatrix} \beta_m^{[1]} e_k \\ \vdots \\ \beta_m^{[n]} e_k \end{pmatrix} \right\|_A.$$

Mit einer geeigneten Konstanten γ ist (wegen der Äquivalenz der Normen)

$$\frac{1}{\gamma} \|x\|_A \leq \|x\| \leq \gamma \|x\|_A$$

und

$$\begin{aligned} \left\| \begin{pmatrix} \beta_m^{[1]} e_k \\ \vdots \\ \beta_m^{[n]} e_k \end{pmatrix} \right\|_A &\leq \gamma \max_{1 \leq i \leq n} |\beta_m^{[i]}| < \gamma \|s(t_{m+k}; h)\| \\ &\leq \gamma \left(\sum_{j=0}^k L_{j+1} \|d_{m+j}\| + \|r(t_{m+k}; h)\| \right) \\ &\leq \gamma \hat{L} (\|D_m\| + \|D_{m+1}\|) + \gamma \left\| \begin{pmatrix} r(t_{m+k}; h) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right\| \\ &\leq \gamma^2 \hat{L} (\|D_m\|_A + \|D_{m+1}\|_A) + \gamma \delta(\varepsilon) \end{aligned}$$

mit

$$\begin{aligned} \hat{L} &= \sum_{j=1}^{k+1} L_j \\ \delta(\varepsilon) &\geq \max\{\|r(t_{m+k}; h)\| : 0 \leq m \leq N_h - k\}. \end{aligned}$$

Also mit

$$\begin{aligned} \xi_m &\stackrel{def}{=} \|D_m\|_A \\ (1 - h\gamma^2 \hat{L}) \xi_{m+1} &\leq (1 + h\gamma^2 \hat{L}) \xi_m + \gamma \delta(\varepsilon) h \quad m = 0, \dots, N_h - k. \end{aligned}$$

Nach Hilfssatz 1.8.1 daher für $h \leq h_0 = h_0(\gamma, \hat{L})$

$$(F1) \quad \xi_m \leq e^{\lambda hm} \xi_0 + (e^{\lambda hm} - 1) 2\gamma/\lambda \delta(\varepsilon)$$

mit $\lambda \stackrel{def}{=} 4\gamma^2 \hat{L}$. Aber

$$(F2) \quad \xi_0 = \|D_0\|_A \leq \gamma \max_{0 \leq i \leq k-1} \|d_i\| \leq \gamma \delta(\varepsilon).$$

Wegen $hm \leq t_E - t_0$ folgt $\xi_m \leq \varepsilon$ für $k \leq m \leq N_h - k + 1$ mit $\delta(\varepsilon) = \varepsilon / (e^{\lambda(t_E - t_0)} (1 + \frac{2\gamma}{\lambda}))$.

Sei umgekehrt das MSV D -stabil. Dann betrachte man die Aufgabe $y' = 0$, $y(0) = 0$ mit den Anfangsvorgaben $\eta_0 = \dots = \eta_{k-1} = 0$ und $r(t_i; h) = 0$ für $i \geq k$. Dann genügen die d_i den Differenzgleichungen

$$\sum_{j=0}^k \alpha_j d_{m+j} = 0, \quad m \geq 0 \quad (1.24)$$

und nach Vor. folgt aus $|d_0|, \dots, |d_{k-1}| \leq \delta(\varepsilon)$, daß $|d_i| \leq \varepsilon \quad (\forall i)$. Ist also $\{d_i\}$ irgendeine Lösung von (1.24), dann mit einer geeigneten Konstanten C natürlich $|Cd_0|, \dots, |Cd_{k-1}| \leq \delta(\varepsilon)$, d.h. $|Cd_i| \leq \varepsilon \quad (\forall i)$ d.h.

$$|d_i| \leq \varepsilon/C \quad (\forall i)$$

und daher $\lim_{i \rightarrow \infty} \frac{d_i}{i} = 0 \quad (d_0, \dots, d_{k-1})$ bel. d.h. mit Satz 1.7.3 (S) □

>>

Hilfssatz 1.8.1: Sei $\{\xi_m\}_{m \in \mathbb{N}_0}$ eine Folge mit

$$\xi_m \in \mathbb{R}_+ \quad \text{und} \quad (1 - h\beta)\xi_{m+1} \leq (1 + h\gamma)\xi_m + h\mu\delta.$$

Dann gilt für $0 < h \leq 1/(2\beta) = h_0$:

$$\xi_m \leq e^{2(\beta+\gamma)mh} \xi_0 + (e^{2(\beta+\gamma)mh} - 1)\mu/(\beta + \gamma)\delta.$$

Beweis: Für $0 < h \leq h_0$ ist $1 - h\beta \geq \frac{1}{2}$ und daher

$$\xi_{m+1} \leq \left(\frac{1 + h\gamma}{1 - h\beta} \right) \xi_m + 2\mu\delta h.$$

Aber

$$\frac{1 + h\gamma}{1 - h\beta} = 1 + \frac{h(\gamma + \beta)}{1 - h\beta} \leq 1 + 2h(\gamma + \beta) \leq e^{2(\gamma+\beta)h}.$$

Also

$$\begin{aligned} \xi_m &\leq e^{2(\gamma+\beta)hm} \xi_0 + \left(\sum_{j=0}^{m-1} e^{2(\gamma+\beta)jh} \right) 2\mu\delta h \\ &= e^{2(\gamma+\beta)hm} \xi_0 + \underbrace{\frac{e^{2(\gamma+\beta)hm} - 1}{e^{2(\gamma+\beta)h} - 1}}_{\geq 2(\gamma+\beta)h} 2\mu\delta h \quad q.e.d. \end{aligned}$$

□

Satz 1.8.2. *Es sei (V2) erfüllt und das MSV konvergent für die Aufgabe $y' = 0$, $y(0) = 0$. Dann ist es D -stabil.* □

Beweis: Wir betrachten den Fall $r(t; h) \equiv 0$, $t \in G'_h$ (d.h. $\delta(\varepsilon) = 0$ in (F1).) Wegen der Konvergenz des Verfahrens gilt

$$\sup_{t \in G'_h} |\eta(t; h)| \xrightarrow{h \rightarrow 0} 0 \quad \text{mit } \eta_i \rightarrow 0 \quad i = 0, \dots, k-1.$$

Mit $\eta_i \stackrel{\text{def}}{=} c_i h$, $c_i \in \mathbb{R}$ bel., $0 \leq i \leq k-1$ folgt $\eta_i = \tilde{\eta}_i h$ wobei $\tilde{\eta}_i$ Lösung der Differenzgleichung ist mit den Anfangswerten c_i . Ist nun $t \in G'_h$ fest und $h = (t - t_0)/N$, dann gilt somit

$$\frac{\tilde{\eta}_N}{N} = \frac{\eta_N}{Nh} = \frac{\eta_N}{t - t_0} \rightarrow 0 \quad \text{für } N \rightarrow \infty,$$

d.h. nach Satz 1.7.3 die Behauptung. □

Satz 1.8.3. *Ist das MSV D -stabil und konsistent, dann ist es konvergent und es ist Konvergenzordnung gleich Konsistenzordnung.* □

Beweis: Dies ist eine unmittelbare Folge von (F1) und (F2) in Satz 1.8.1 mit den Interpretationen

$$\begin{aligned} r(t; h) &\stackrel{\text{def}}{=} \tau(t, y(t - kh); h), & d_i &= y_i - \eta_i \quad 0 \leq i \leq k-1 \\ \tilde{\eta}_i &\stackrel{\text{def}}{=} y_i. \end{aligned}$$

□

Für lineare MSV gilt folgende Verschärfung: **Äquivalenzsatz**

Satz 1.8.4. *Ein lineares MSV ist konvergent (für f mit (V0) bel.) genau dann, wenn es D -stabil und konsistent ist.* □

<<

Beweis: Die Verfahrensfunktion erfüllt hier (V2). Aus D -Stabilität und Konsistenz folgt also Konvergenz. Sei umgekehrt das Verfahren konvergent. Dann folgt mit Satz 1.8.2. die D -Stabilität. Zu zeigen bleibt

$$\sup_{t \in G'_h} \|\tau(t, y(t - kh); h)\| \xrightarrow{h \rightarrow 0} 0.$$

Wir zeigen dazu, daß

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1), \quad \text{wo } \sigma(t) = \sum_{j=0}^k \beta_j t^j$$

und daß diese Bedingung $\|\tau(t, y(t - kh); h)\| = o(1)$ nach sich zieht. Wir betrachten die beiden trivialen AWA

$$\text{A) } y' = 0, \quad y(0) = 1 \quad \text{B) } y' = 1, \quad y(0) = 0$$

A) Das MSV ergibt mit $y_i = \eta_i$ als Startwerten

$$\sum_{j=0}^k \alpha_j \eta_{m+j} = 0, \quad \eta_0 = \dots = \eta_{k-1} = 1.$$

Wegen $|\eta_m - 1| \rightarrow 0 \quad 0 \leq m \leq N_h$ folgt für beliebiges $\varepsilon > 0$

$$\left| \sum_{j=0}^k \alpha_j \right| \leq \left| \sum_{j=0}^k \alpha_j (\eta_{m+j} - 1) \right| + \left| \sum_{j=0}^k \alpha_j \eta_{m+j} \right| < \varepsilon \quad \text{falls } 0 < h \leq h_0(\varepsilon)$$

d.h. $\rho(1) = 0$.

B) Es ist nun $\rho(1) = 0$, $\rho'(1) \neq 0$ (wegen (S)). Das MSV werde nun betrachtet mit den Startwerten

$$\eta_j \stackrel{\text{def}}{=} jh\mu = jh + jh(\mu - 1) = y_j + r_j \quad 0 \leq j \leq k-1, \quad \mu \stackrel{\text{def}}{=} \frac{\sigma(1)}{\rho'(1)}.$$

Für $0 \leq i \leq N_h - k$ ist

$$\sum_{j=0}^k \alpha_j \eta_{j+i} = h \sum_{j=0}^k \beta_j = h\sigma(1) \Rightarrow \eta_i = ih\mu$$

und $\max_{0 \leq i \leq N_h} |\eta_i - ih\mu| \rightarrow 0$ d.h. $\mu = 1$.

Ist nun f bel. gegeben mit (V0), dann wird

$$\begin{aligned} -\tau(t, y(t - kh); h) &= \frac{1}{h} \left(\sum_{j=0}^k \alpha_j y(t + (j - k)h) - \right. \\ &\quad \left. - h \sum_{j=0}^k \beta_j f(t + (j - k)h, y(t + (j - k)h)) \right) \\ &= \frac{1}{h} \left(\sum_{j=0}^k \alpha_j y(t + (j - k)h) - h \sum_{j=0}^k \beta_j y'(t + (j - k)h) \right) \\ &= \frac{1}{h} \left(\underbrace{\sum_{j=0}^k \alpha_j y(t - kh)}_{0=\rho(1)} + \underbrace{\sum_{j=0}^k \alpha_j j hy'(t - kh)}_{\rho'(1)} + ho(1) - \right. \\ &\quad \left. - h \underbrace{\sum_{j=0}^k \beta_j y'(t - kh)}_{\sigma(1)} + ho(1) \right) \\ &= o(1). \end{aligned}$$

(für $f \in F^1(\mathcal{I} \times \mathbb{R}^n)$ ergibt sich an Stelle von $o(1)$ $\mathcal{O}(h)$, da dann y'' existiert und stetig ist). \square

Bei linearen MSV kann man die Bedingungen für eine genaue Verfahrensordnung p^* sehr einfach charakterisieren:

Satz 1.8.5. Ein lineares MSV hat genau dann die genaue Verfahrensordnung p^* , wenn die Funktion

$$\varphi(t) \stackrel{\text{def}}{=} \frac{\rho(t)}{\ln t} - \sigma(t)$$

die Entwicklung

$$\varphi(t) = (t-1)^{p^*} (a_0 + \mathcal{O}((t-1))) \quad \text{mit } a_0 \neq 0 \text{ besitzt.} \quad (1.25)$$

□

<<

Beweis: Wir betrachten die Anfangswertaufgabe $y' = y$, $y(0) = 1$ mit der Lösung $y(t) = e^t$. Es wird gezeigt, daß für diese Aufgabe genaue Konvergenzordnung p^* mit (1.25) äquivalent ist.

$$\begin{aligned} -\tau(t, e^{t-kh}; h) &= \left(\sum_{j=0}^k \alpha_j e^{t-(k-j)h} - h \sum_{j=0}^k \beta_j e^{t-(k-j)h} \right) \frac{1}{h} \\ &= e^{t-kh} (\rho(e^h)/h - \sigma(e^h)) \\ &= e^{t-kh} (\rho(\mu)/\ln \mu - \sigma(\mu)) \quad \mu = e^h. \end{aligned}$$

Genaue Ordnung p^* bedeutet nun, daß

$$\rho(e^h)/h - \sigma(e^h) = Ch^{p^*} (1 + \mathcal{O}(h)), \quad \text{mit } C \neq 0,$$

d.h. daß die Funktion $\rho(e^h)/h - \sigma(e^h)$ 0 als Nullstelle der genauen Vielfachheit p^* besitzt, d.h. daß $\rho(\mu)/\ln \mu - \sigma(\mu)$ $\mu = 1$ als Nullstelle der genauen Vielfachheit p^* besitzt. Die Umkehrung ist ersichtlich. Es bleibt nun noch zu zeigen, daß für jede rechte Seite f der DGL aus $F^\infty(\mathcal{J} \times \mathbb{R}^n)$ die Konsistenzordnung von τ (mindestens) p^* ist, wenn (1.25) gilt.

Dazu betrachten wir

$$\begin{aligned} \rho(e^h) - h\sigma(e^h) &= \sum_{j=0}^k \alpha_j \left(\sum_{\nu=0}^{\infty} \frac{j^\nu h^\nu}{\nu!} \right) - h \sum_{j=0}^k \beta_j \sum_{\nu=0}^{\infty} \frac{j^\nu h^\nu}{\nu!} \\ &= \sum_{\nu=1}^{\infty} \frac{h^\nu}{\nu!} \underbrace{\left(\sum_{j=0}^k (\alpha_j j^\nu - \nu j^{\nu-1} \beta_j) \right)}_{c_\nu} + \underbrace{\sum_{j=0}^k \alpha_j}_{c_0} \end{aligned}$$

und (1.25) bedeutet also $c_0 = \dots = c_{p^*} = 0$, $c_{p^*+1} \neq 0$.

Taylorentwicklung liefert für $f \in F^{p+1}(\mathcal{J} \times \mathbb{R}^n)$

$$\begin{aligned}
-h\tau(t+kh, y(t); h) &= \sum_{j=0}^k \{ \alpha_j y(t+jh) - h\beta_j y'(t+jh) \} \\
&= \sum_{j=0}^k \{ \alpha_j \sum_{s=0}^{p^*+1} \frac{j^s h^s}{s!} y^{(s)}(t) + \mathcal{O}(h^{p^*+2}) - \\
&\quad - \beta_j \sum_{s=0}^{p^*} \frac{h^{s+1} j^s}{s!} y^{(s+1)}(t) + \mathcal{O}(h^{p^*+2}) \} \\
&= \sum_{j=0}^k \alpha_j y(t) + \sum_{s=1}^{p^*+1} \frac{y^{(s)}(t)}{s!} h^s \underbrace{\left(\sum_{j=0}^k (\alpha_j j^s - s\beta_j j^{s-1}) \right)}_{c_s} + \\
&\quad + \mathcal{O}(h^{p^*+2}) \\
&= c_{p^*+1} \frac{y^{(p^*+1)}(t)}{(p^*+1)!} h^{p^*+1} + \mathcal{O}(h^{p^*+2}). \quad q.e.d.
\end{aligned}$$

□

>>

Bemerkung 1.8.1. Als Nebenprodukt des Beweises von Satz 1.8.5 erhalten wir, daß der lokale Abschneidefehler eines linearen MSV **stets** die Form

$$\tau(t, y(t-kh); h) = Cy^{(p+1)}(t)h^p + \mathcal{O}(h^{p+1})$$

hat und damit auch der lokale Diskretisierungsfehler

$$\delta(t-kh, y(t-kh); h) = Cy^{(p+1)}(t)h^{p+1} + \mathcal{O}(h^{p+2}).$$

Man kann also den lokalen Diskretisierungsfehler in diesem Fall ohne Zusatzaufwand aus zurückliegenden f -Werten mittels dividierter Differenzen schätzen, Abschnitt 1.5, (1.15) folgende. □

Die Konsistenzrelationen

$$\begin{aligned}
c_\nu &= 0 & \nu &= 0, \dots, p \\
c_{p+1} &= 1
\end{aligned}$$

stellen ein **lineares** Gleichungssystem von $p+2$ Gleichungen in den $2k+2$ Unbekannten $\alpha_0, \dots, \alpha_k, \beta_0, \dots, \beta_k$ dar. Die Koeffizientenmatrix dieses Systems ist

$$\begin{aligned}
& \left((t_i^\nu)_{\nu,i} \quad , \quad -(\nu t_i^{\nu-1})_{\nu,i} \right) & \begin{array}{l} 0 \leq \nu \leq p+1 \\ 0 \leq i \leq k \\ t_i = i. \end{array}
\end{aligned}$$

Da diese Matrix für $p \leq 2k$ stets maximalen Rang hat,⁴ **gibt es lineare k -Schritt-MSV der Ordnung $p : 1 \leq p \leq 2k$** . Nun muß aber ein MSV, um praktisch brauchbar zu sein, auch die Nullstellenbedingung (S) erfüllen. Dazu hat Dahlquist 1956 folgenden Satz bewiesen:

Satz 1.8.6. Dahlquists erste Stabilitätsschranke *Die erreichbare Höchstordnung eines D -stabilen linearen k -Schritt-Verfahrens ist*

$$p = 2(\lfloor \frac{k}{2} + 1 \rfloor).$$

(Beweis siehe z.B. bei Grigorieff, Band II Mehrschrittverfahren) □

Aufgrund ihrer Konstruktion haben die in Abschnitt 1.6 angegebenen Verfahren folgende Eigenschaften:

Verfahren	Schrittzahl	Ordnung, weil $w(t)$ polynomial interpoliert vom Grad ...	D -stabil
Adams-Bashforth	k	$k \quad w(t) = y'(t)$ Grad $k - 1$	$\forall k$
Adams-Moulton	k	$k + 1 \quad w(t) = y'(t)$ Grad k	$\forall k$
Gear-Formel	k	$k \quad w(t) = y(t)$ Grad k	$1 \leq k \leq 6$

Der Nachweis für die Stabilitätseigenschaften der Gear-Formeln ist ziemlich kompliziert. Er kann bei Grigorieff Bd. II nachgelesen werden.

Für k gerade gibt es nach Satz 1.8.6 auch noch D -stabile Verfahren der Ordnung $k + 2$, die bei den Verfahren aus der obigen Tabelle nicht erfaßt sind. Tatsächlich sind diese Verfahren, die manchmal auch als "optimale" Verfahren (bzgl. ihrer Ordnung) bezeichnet werden, nur von sehr bedingtem Wert. Dies liegt daran, daß bei diesen Verfahren das Polynom ρ neben der "Hauptwurzel" $\xi_1 = 1$ noch weitere Wurzeln ξ_ν , $\nu \geq 2$ vom Betrag 1 hat, die im globalen Diskretisierungsfehler einen oszillierenden, und bei negativen Realteilen der Eigenwerte von $f_y(t, y(t))$ exponentiell mit t wachsenden (!) Anteil

$$h^q \xi_\nu^{(t-t_0)/h} E_\nu(t) d_\nu, \quad E_\nu \in \mathbb{C}^{n \times n}, \quad d_\nu \in \mathbb{C}^n$$

hervorrufen, wobei q die Konsistenzordnung der **Startwerte** ist, d_ν durch die **Fehler in den Startwerten** bestimmt ist und die Matrix E_ν die AWA

$$E'_\nu(t) = \frac{\sigma(\xi_\nu)}{\xi_\nu \rho'(\xi_\nu)} f_y(t, y(t)) E_\nu(t), \quad E_\nu(t_0) = I$$

löst. Auf die allgemeine Herleitung der asymptotischen Entwicklung des globalen Diskretisierungsfehlers bei linearen MSV wollen wir hier verzichten (vgl. bei Grigorieff Bd. 2) und alles nur an einem Beispiel demonstrieren:

<<

⁴Die transponierte Matrix mit $p = 2k$ ist die Koeffizientenmatrix, die bei der polynomialen Hermite-Interpolation von Grad $2k + 1$ auftritt

Wir betrachten die sogenannte **Mittelpunktregel**:

$$\eta_k - \eta_{k-2} = 2hf(t_{k-1}, \eta_{k-1})$$

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

Also $\rho(\xi) = (\xi - 1)(\xi + 1), \quad p^* = 2.$

Dieses lineare MSV wollen wir auf die AWA

$$y' = -y, \quad y(0) = 1 \quad (\text{d.h. } y = e^{-t})$$

mit den speziellen Startwerten

$$\begin{aligned} \eta_0 &= 1 \\ \eta_1 &= 1 - h \quad (\text{Euler})^5 \end{aligned}$$

anwenden. Wir erhalten die Differenzgleichung

$$\eta_{j+2} + 2h\eta_{j+1} - \eta_j = 0, \quad j \geq 0, \quad \eta_0 = 1, \quad \eta_1 = 1 - h.$$

Das Polynom $\xi^2 + 2h\xi - 1$ besitzt die beiden Nullstellen

$$\xi_{1,2} = -h \pm \sqrt{1 + h^2},$$

d.h. $\xi_1 = \sqrt{1 + h^2}(1 - h/\sqrt{1 + h^2}), \quad \xi_2 = -\sqrt{1 + h^2}(1 + h/\sqrt{1 + h^2})$ also $|\xi_2| > 1$ (!).

Die allgemeine Lösung der Differenzgleichung ist $c_1\xi_1^n + c_2\xi_2^n$. Aus den Anfangswerten erhält man

$$\left. \begin{aligned} 1 &= c_1 + c_2 \\ 1 - h &= c_1\xi_1 + c_2\xi_2 \end{aligned} \right\} \Rightarrow \begin{aligned} c_1(h) &= \frac{\xi_2 - (1 - h)}{\xi_2 - \xi_1} = \frac{1 + \sqrt{1 + h^2}}{2\sqrt{1 + h^2}} \\ &= 1 + \sum_{k=1}^{\infty} c_k h^{2k} \\ c_2(h) &= \frac{\sqrt{1 + h^2} - 1}{2\sqrt{1 + h^2}} = \frac{h^2}{2} \cdot \frac{1}{1 + h^2 + \sqrt{1 + h^2}} \\ &= \frac{h^2}{4} + \sum_{k=2}^{\infty} b_k h^{2k}. \end{aligned}$$

Somit

$$\eta(t; h) = c_1(h)\xi_1(h)^{t/h} + c_2(h)\xi_2(h)^{t/h} \quad t/h \in \mathbb{N}.$$

Nun ist aber

$$\left. \begin{aligned} \xi_1^n &= \exp(n \ln(-h + \sqrt{1 + h^2})) \\ \xi_2^n &= (-1)^n \exp(n \ln(h + \sqrt{1 + h^2})) \end{aligned} \right| \begin{aligned} \ln(x + \sqrt{1 + x^2}) &= \operatorname{arsinh} x \\ &= -\operatorname{arsinh}(-x) \\ &= x(1 - \frac{1}{2 \cdot 3}x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 5}x^4 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 7}x^6 \dots) \\ &|x| < 1 \end{aligned}$$

also mit $nh = t$

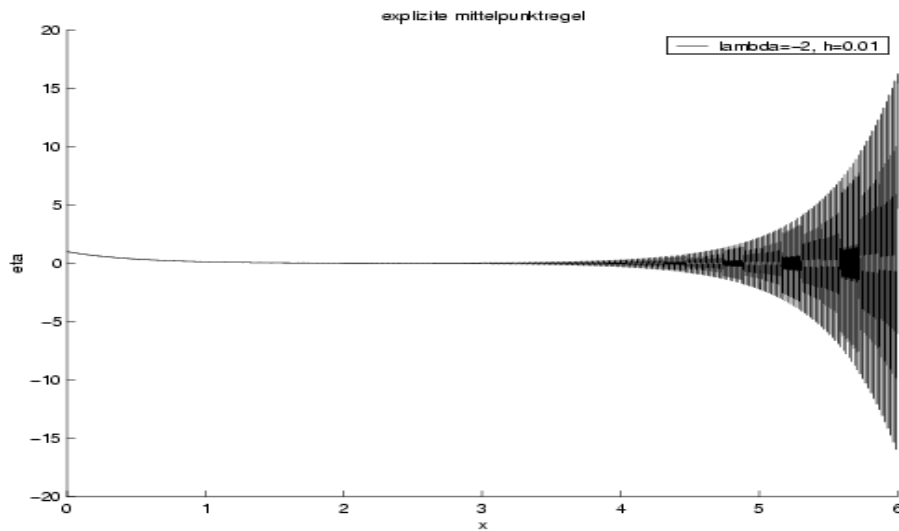
$$\begin{aligned} \xi_1^n &= \exp(-t(1 - \frac{1}{6}h^2 + \frac{3}{40}h^4 \dots)) = e^{-t}(1 + \sum_{k=1}^{\infty} v_k(t)h^{2k}) \\ \xi_2^n &= (-1)^n e^t(1 + \sum_{k=1}^{\infty} u_k(t)h^{2k}) \end{aligned}$$

⁵Beachte: $\eta_1 - y_1 = \mathcal{O}(h^2)$, d.h. diese Konstruktion paßt zur Konvergenzordnung 2

und daher

$$\eta(t; h) = e^{-t} + \sum_{k=1}^{\infty} (e^{-t} \tilde{v}_k(t) + (-1)^n e^t \tilde{u}_k(t)) h^{2k} \quad hn = t \text{ fest}$$

mit analytischen Funktionen \tilde{v}_k und \tilde{u}_k . Nur für Werte von h , die jeweils immer $n = t/h$ gerade bzw. ungerade liefern, ist dies eine asymptotische Entwicklung in geraden Potenzen von h . Man beachte, daß schon der Term in h^2 mit e^t anwächst! Die folgende Abbildung zeigt diesen Effekt für $\lambda = -2$, $h = 0.01$. Mit dieser Schrittweite würde sogar das explizite Eulerverfahren noch vernünftige Werte liefern!



Gragg hat gezeigt, daß die Mittelpunkregel auch bei allgemeiner (hinreichend oft differenzierbarer) rechten Seite f der DGL eine asymptotische Entwicklung

$$\eta(t; h) = y(t) + \sum_{k=1}^N h^{2k} (v_k(t) + (-1)^n u_k(t)) + \mathcal{O}(h^{2N+2}) \quad t - t_0 = nh$$

besitzt, falls $\eta_1 = \eta_0 + h f(t_0, \eta_0)$. Es ist naheliegend, wie bei der Romberg-Integration diese Entwicklung mit Hilfe der Richardson-Extrapolation zur Konstruktion eines Verfahrens hoher Konsistenzordnung zu nutzen. Da die Funktionen u_k sehr schnell mit t anwachsen können, ist es nützlich, daß man wenigstens den Einfluß der Funktion u_1 durch eine sogenannte **Glättungsformel** beseitigen kann:

Bildet man

$$s(t; h) = \frac{1}{2}(\eta(t; h) + \eta(t - h; h) + hf(t, \eta(t; h))) \quad (1.26)$$

dann ergibt sich nach Definition von η_i

$$s(t; h) = \frac{1}{2}(\eta(t; h) + \eta(t + h; h) - hf(t, \eta(t; h))) \quad (1.27)$$

d.h. (Addition von (1.26), (1.27))

$$s(t; h) = \frac{1}{2}(\eta(t; h) + \frac{1}{2}(\eta(t + h; h) + \eta(t - h; h))) \quad (1.28)$$

Einsetzen der asymptotischen Entwicklung für $\eta(t; h)$ ergibt dann

$$s(t; h) = y(t) + h^2(v_1(t) + \frac{1}{4}y''(t)) + \sum_{k=2}^N h^{2k}(v_k^*(t) + (-1)^k u_k^*(t)) + \mathcal{O}(h^{2N+2}).$$

In der Praxis geht man zur Ausnutzung dieser Formeln so vor, daß man zunächst $t_E \stackrel{\text{def}}{=} t_0 + H$ mit einer "kleinen" Vorschlagsschrittweite H setzt und dann für eine Folge **gerader** Zahlen n_i , gewöhnlich

$$n_i \in \{2, 4, 6, 8, 12, 16, \dots\},$$

mit

$$h_i \stackrel{\text{def}}{=} \frac{H}{n_i} \quad i = 0, \dots, m$$

die Werte

$$\sigma_i \stackrel{\text{def}}{=} s(t_0 + H; h_i) \quad i = 0, \dots, m$$

berechnet. Dann wird nach obigen Resultaten

$$P_{0,m}(0; (h_i^2, \sigma_i)) = y(t_0 + H) + \mathcal{O}(h_0^{2m+2})$$

Interpolationspolynom mit Stützstellen h_i^2 und Stützwerten σ_i vom Grad m an der Stelle 0 ausgewertet.

(Methode von Gragg–Bulirsch–Stoer)

Die Differenzen im Extrapolationsschema kann man wie beim Rombergverfahren als Schätzungen für den lokalen Diskretisierungsfehler auf den verschiedenen Extrapolationsstufen benutzen. Deshalb braucht man den Wert von m nicht von vorneherein festzulegen. Wird die gewünschte Genauigkeit im Extrapolationsschema mit einem vorgegebenen maximalen Wert für m erreicht, dann wird die Vorschlagsschrittweite H akzeptiert und der extrapolierte Wert als Näherung für $y(t_0 + H)$ genommen. Falls dies nicht der Fall ist, wird H verkleinert und die Rechnung wiederholt. War hingegen die Genauigkeit im Extrapolationsschema weit höher als erwünscht, dann wird die Vorschlagsschrittweite für den nächsten Schritt z.B. verdoppelt. Details siehe z.B. bei Stoer, J.: Extrapolation Methods for the solution of initial value problems and their practical realization. Lecture Notes Math. 362, S.1-21 (Dundee Conference 1972).

>>

Wir wollen diesen Abschnitt mit einigen Bemerkungen zum Stabilitätsverhalten der linearen MSV bei linearen DGL-Systemen und bei fester endlicher Schrittweite $h > 0$ beenden. Also

$$y' = Ay, \quad y(t_0) = y_0, \quad A = T \operatorname{diag} (\lambda_1, \dots, \lambda_n) T^{-1}, \quad \Re e \lambda_n \leq \dots \leq \Re e \lambda_1 < 0$$

$$\sum_{i=0}^k \alpha_i \eta_{m+i} = h \sum_{i=0}^k \beta_i A \eta_{m+i}$$

und mit

$$\tilde{\eta}_j \stackrel{\text{def}}{=} T^{-1}\eta_j$$

$$\sum_{i=0}^k \alpha_i \tilde{\eta}_{m+i} = h \sum_{i=0}^k \beta_i \text{diag}(\lambda_1, \dots, \lambda_n) \tilde{\eta}_{m+i}.$$

Dies ist nun ein zerfallendes System von n linearen homogenen Differenzgleichungen. Nach Satz 1.7.1 gilt

$$\tilde{\eta}_m \rightarrow 0 \quad m \rightarrow \infty \quad (\text{d.h. } \eta_m \rightarrow 0 \quad \text{mit } m \rightarrow \infty)$$

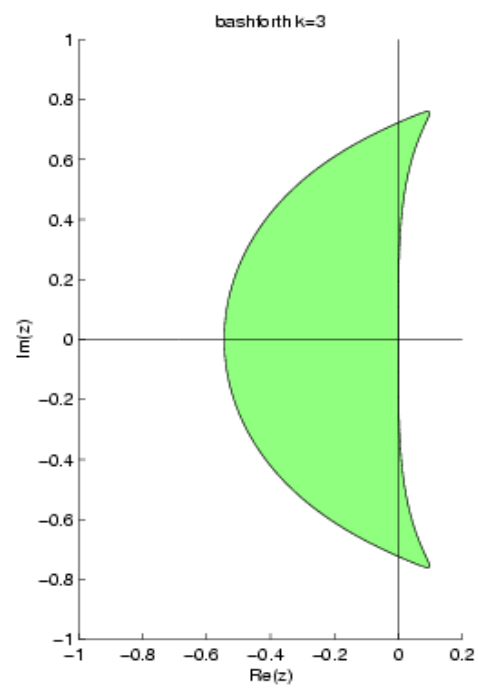
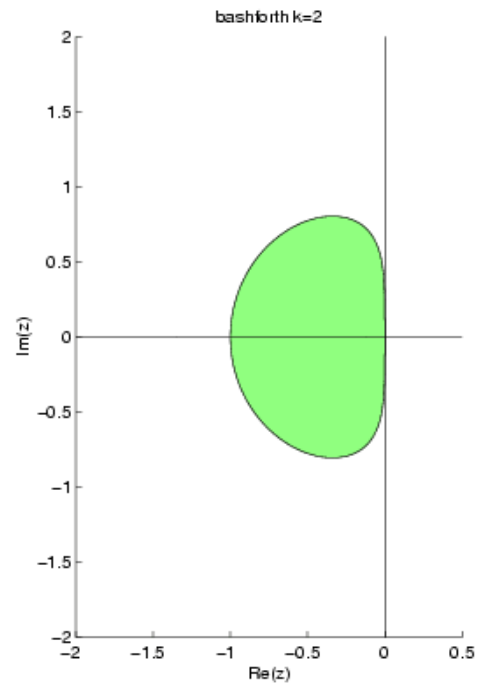
genau dann, wenn die Nullstellen des Polynoms

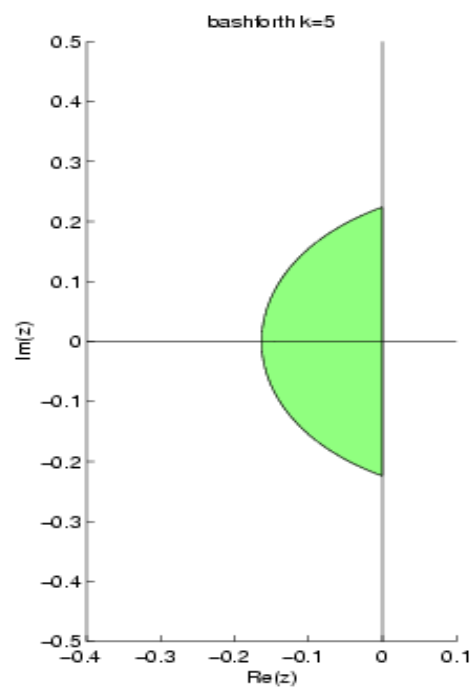
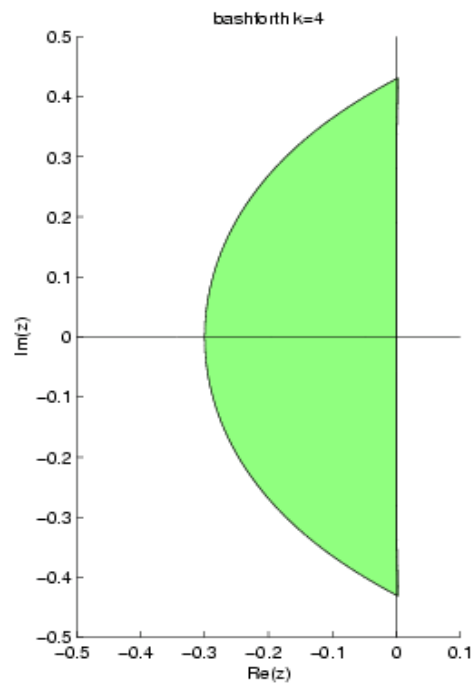
$$\pi(\zeta; q) = \rho(\zeta) - q\sigma(\zeta), \quad q \in \{h\lambda_1, \dots, h\lambda_n\}$$

alle betragsmäßig kleiner sind als 1:

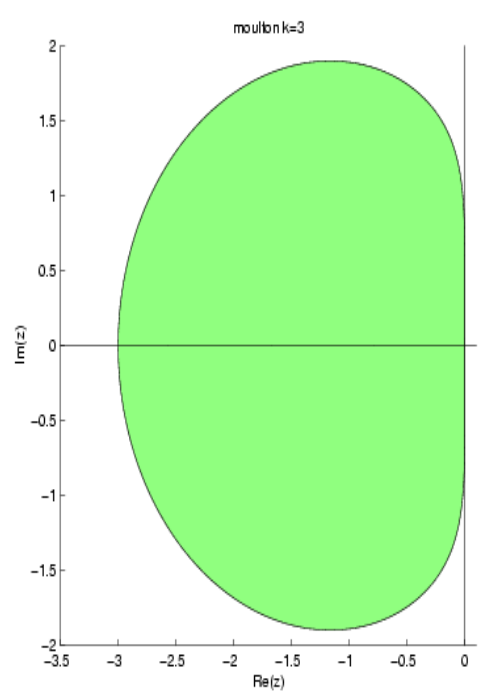
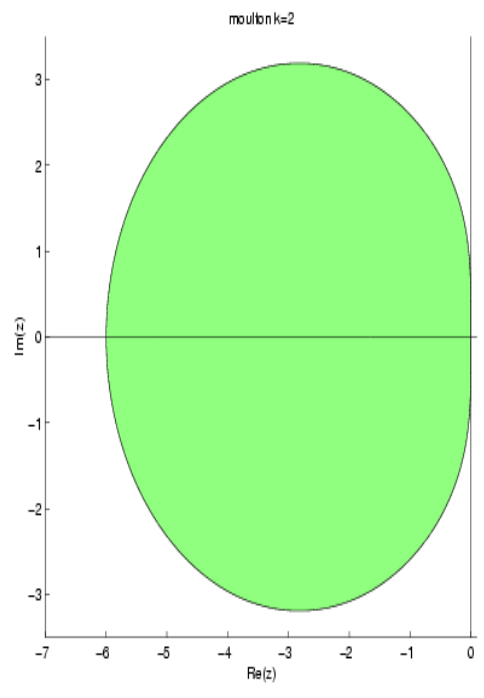
$$\pi(\zeta; q) = 0 \quad \Rightarrow \quad |\zeta| < 1. \quad (1.29)$$

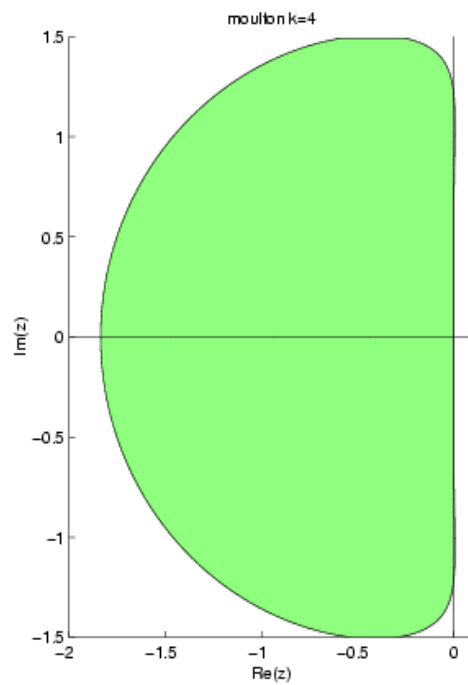
Auf den folgenden Seiten sind die Stabilitätsbereiche der Verfahren aus 1.6., d.h. die Bereiche der komplexen Ebene, in denen q liegen darf, sodaß (1.29) gilt, skizziert. Man erkennt, daß die Stabilitätsbereiche der Adams–Bashforth–Verfahren extrem klein sind (d.h. h muß sehr klein sein), weshalb man diese Verfahren niemals für sich alleine verwendet. Auch für die Adams–Moulton–Verfahren muß h eine Bedingung der Form $h < c / \max_i |\lambda_i(A)|$ erfüllen mit c in der Größenordnung von 1. Diese Verfahren sind also ungeeignet für “steife” Gleichungen. Bei den rückwärts genommenen Differentiationsformeln dagegen liegt ohne Einschränkung an h Stabilität vor, wenn die λ_j in einem Winkelbereich um die negative reelle Achse liegen. Diese Formeln eignen sich also für steife Systeme, falls keine langsam abklingenden, schnell oszillierenden Lösungskomponenten vorliegen. Zunächst die Stabilitätsbereiche der Adams-Bashforth-Verfahren: Man sieht, daß diese Stabilitätsbereiche mit wachsender Ordnung sehr klein werden. Dies ist der Hauptgrund dafür, daß diese Verfahren nie für sich alleine benutzt werden.





Das Einschritt-Adams-Moulton-Verfahren ist die Trapezregel, also A-stabil. Alle anderen Verfahren dieses Typs haben, obwohl implizit, nur einen beschränkten Stabilitätsbereich, der ebenfalls mit wachsender Ordnung kleiner wird, aber viel grösser ist als der der Bashforth-Verfahren.

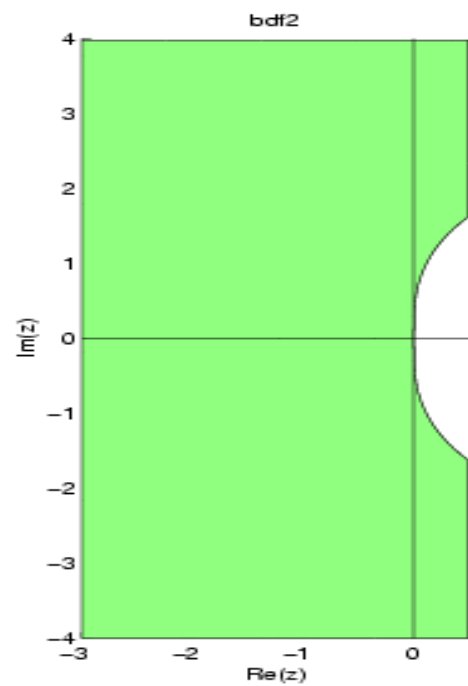


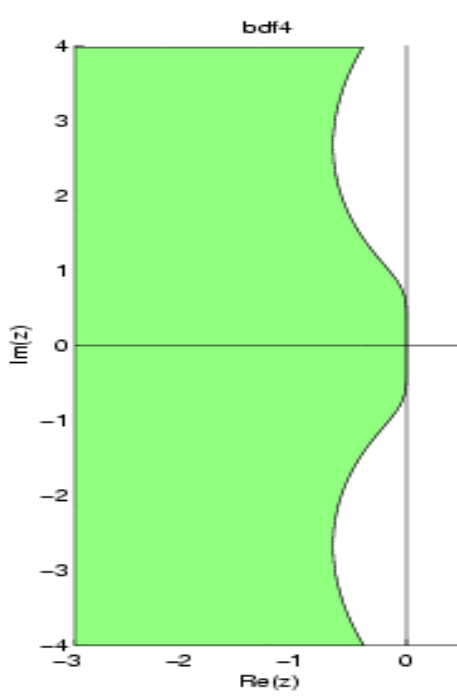
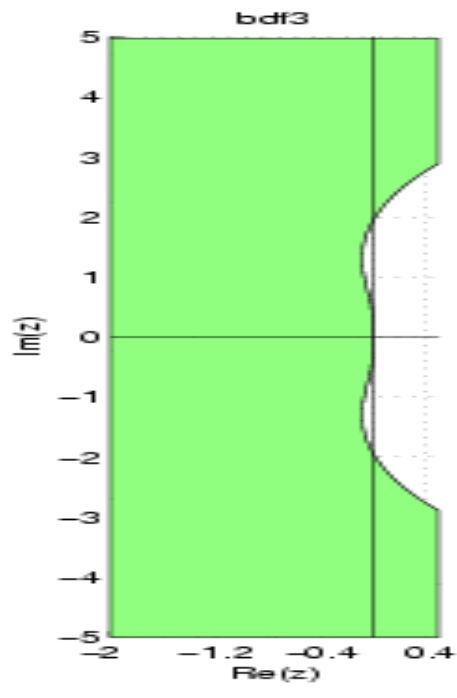


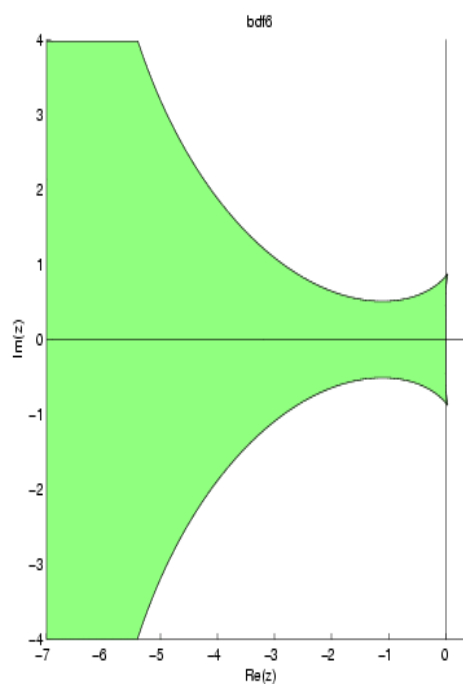
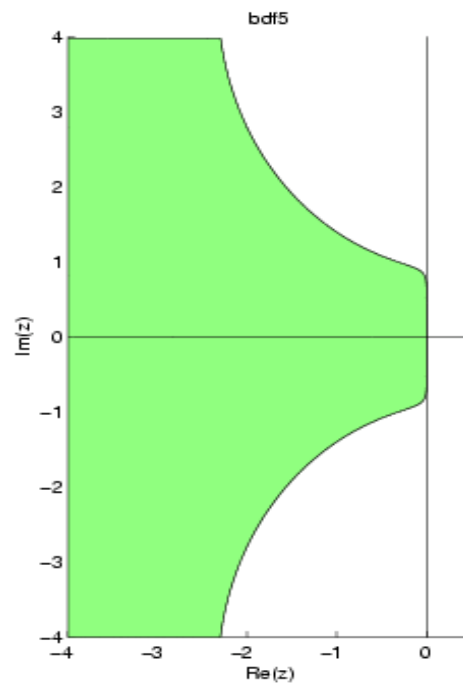
Bereiche der absoluten Stabilität der rückwärts genommenen Differentiationsformeln (Gear'sche Methode). Hier ist BDF2 noch A-stabil, alle anderen Formeln sind nur $A(\alpha)$ -stabil.

Schrittzahl = Ordnung = $k \quad 1 \leq k \leq 6$.

Stabilität im schraffierten Bereich







1.9 Prädiktor–Korrektor–Methoden

Das Ziel bei der Konstruktion der Mehrschritt–Verfahren war es, hohe Ordnung mit möglichst wenig Aufwand (gemessen in f –Auswertungen) zu erzielen. Nun erweisen sich die expliziten linearen MSV als praktisch unbrauchbar, während die impliziten Verfahren pro Schritt “im Prinzip” die Lösung einer nichtlinearen Gleichung, also unendlich

viele f -Auswertungen, erfordern. Nun ist es naheliegend, in der impliziten Formel nur eine feste Anzahl Iterationen mit einem guten Startwert aus einem expliziten Verfahren vorzusehen. Auf diese Weise entstehen Prädiktor-Korrektor-Verfahren. Von den vielen Möglichkeiten, die man bei dieser Konstruktion hat, wollen wir nur eine praktisch besonders bewährte beschreiben. Dabei benutzt man eine k -Schritt Adams-Bashforth-Formel (mit der Verfahrensordnung k) als Prädiktor und führt einen Iterationsschritt in einer k -Schritt Adams-Moulton-Formel als Korrektor aus:

$$\mathbf{P:} \quad \eta_{m+k}^{[0]} = - \sum_{i=0}^{k-1} \alpha_i^* \eta_{m+i} + h \sum_{i=0}^{k-1} \beta_i^* f_{m+i}; \quad f_j = f(t_j, \eta_j)$$

$$\mathbf{E:} \quad f_{m+k}^{[0]} \stackrel{def}{=} f(t_{m+k}, \eta_{m+k}^{[0]})$$

$$\mathbf{C:} \quad \eta_{m+k} = - \sum_{i=0}^{k-1} \alpha_i \eta_{m+i} + h \sum_{i=0}^{k-1} \beta_i f_{m+i} + h \beta_k f_{m+k}^{[0]}$$

$$\mathbf{E:} \quad f_{m+k} \stackrel{def}{=} f(t_{m+k}, \eta_{m+k})$$

α_i^*, β_i^* : Koeffizienten der k -Schritt-Adams-Bashforth - Formel mit $\alpha_k^* = 1$

α_i, β_i : Koeffizienten der k -Schritt-Adams-Moulton-Formel mit $\alpha_k = 1$ ⁶

Dieser sogenannte PECE-Modus führt somit zu einem **nichtlinearen expliziten k -Schritt-Verfahren**

$$\begin{aligned} \eta_{m+k} = & - \sum_{i=0}^{k-1} \alpha_i \eta_{m+i} + h \left(\sum_{i=0}^{k-1} \beta_i f(t_{m+i}, \eta_{m+i}) + \right. \\ & \left. + \beta_k f(t_{m+k}, - \sum_{i=0}^{k-1} \alpha_i^* \eta_{m+i} + h \sum_{i=0}^{k-1} \beta_i^* f(t_{m+i}, \eta_{m+i})) \right) \end{aligned}$$

Diese Vorgehensweise nennt man Prädiktor-Korrektor-Verfahren. Neben dem angegebenen PECE-Modus gibt es noch viele weitere Möglichkeiten, u.a. PEC, bei dem die zweite Auswertung von f unterbleibt (d.h. f wird nur auf den Prädiktorwerten ausgewertet), P(EC)^mE usw. Der PECE -Modus ist besonders günstig weil er geringen Aufwand mit einer moderaten Verkleinerung des Stabilitätsbereiches gegenüber dem reinen Korrektor verbindet.

Satz 1.9.1. *Die Kombination des k -Schritt Adams-Bashforth Verfahrens mit dem k -Schritt Adams-Moulton-Verfahren im Modus PECE führt zu einem D -stabilen (nicht-linearen) MSV der Verfahrensordnung $k + 1$, (d.h. konvergent von der Ordnung $k + 1$, falls die Startwerte Konsistenzordnung $k + 1$ haben).*

□

⁶also $\alpha_j = \alpha_j^* = 0 \quad j < k - 1$

Beweis: Die Aussage über die D -Stabilität ist klar. Es ist nur die Aussage über die Verfahrensordnung zu zeigen. Es ist nun aber (aufgrund einer Taylorentwicklung)

$$\begin{aligned}
 \tau_{PECE}(t_{m+k}, y(t_m); h) &= \tau_C(t_{m+k}, y(t_m); h) + \\
 &+ \beta_k \left(f(t_{m+k}, - \sum_{i=0}^{k-1} \alpha_i^* y_{m+i} + h \sum_{i=0}^{k-1} \beta_i^* f(t_{m+i}, y_{m+i})) - f(t_{m+k}, y_{m+k}) \right) \\
 &= \tau_C(t_{m+k}, y(t_m); h) + h \beta_k f_y(t_m, y_m(t_m)) \tau_P(t_{m+k}, y(t_m); h) + \mathcal{O}(h^{2k+2}) \\
 &= C_{k+1} y^{(k+2)}(t_m) h^{k+1} + \beta_k f_y(t_m, y(t_m)) C_k^* y^{(k+1)}(t_m) h^{k+1} + \mathcal{O}(h^{k+2})
 \end{aligned}$$

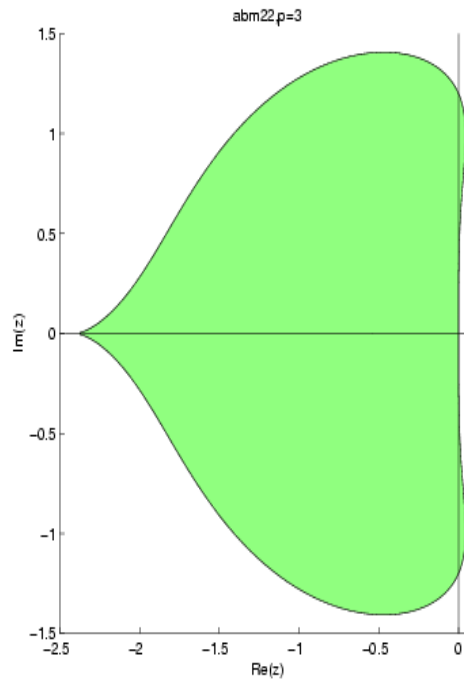
□

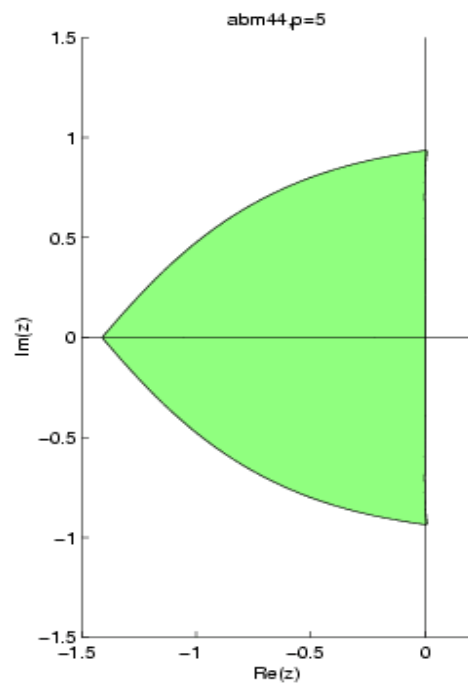
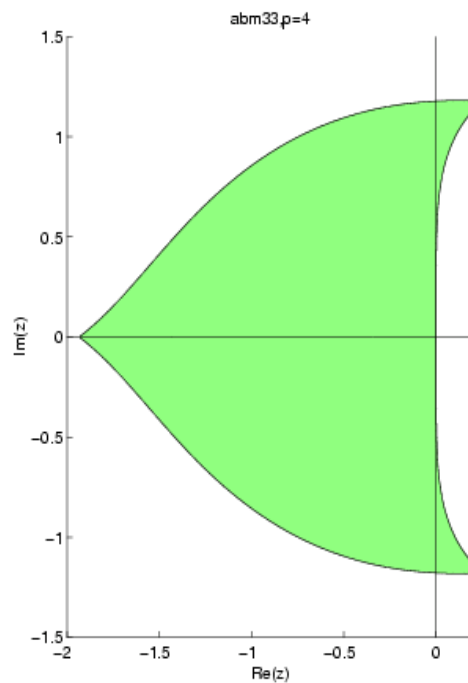
In jedem Rechenschritt hat man mit

$$\eta_{m+k} - \eta_{m+k}^{[0]}$$

eine Schätzung für den lokalen Diskretisierungsfehler im Prädiktor und weil die Ordnung des Prädiktors um 1 kleiner ist als die des Korrektors führt eine darauf aufgebaute Schrittweitensteuerung zu einer zuverlässigen Steuerung für das PECE-Verfahren.

Hier sind die Stabilitätsbereiche dieser Verfahren im PECE-Modus wiedergegeben:





1.10 Adams–Bashforth–Moulton Prädiktor–Korrektor–Verfahren mit variabler Schrittweite und Ordnung (ABMVOAS)

Bei den Prädiktor–Korrektor–Verfahren hat man die Möglichkeit, sich von der Voraussetzung eines äquidistanten (oder doch wenigstens stückweise äquidistanten) Gitters zu

befreien. Man hat dann, entsprechend der Herleitung in Abschnitt 1.6.1, Formeln des Typs

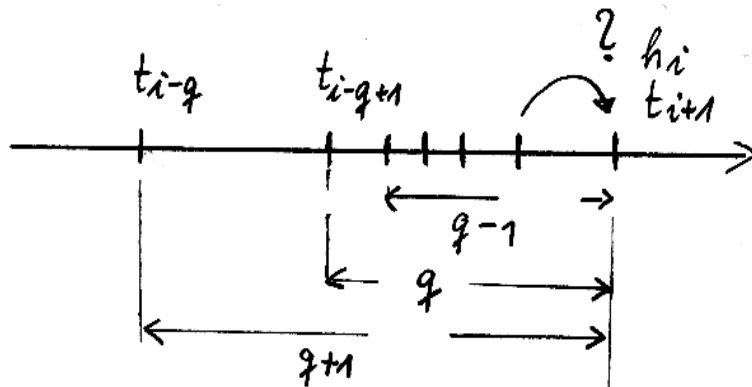
$$\mathbf{P:} \quad \eta_{m,q}^{[0]} = \eta_{m-1} + h_{m-1} \sum_{j=1}^q \beta_{q-j,q}^*(h_{m-1}, h_{m-2}, \dots, h_{m-q}) f_{m-j}$$

$q = \text{Schrittzahl} = \text{Ordnung Prädiktor}$

$$\mathbf{E:} \quad f_{m,q}^{[0]} = f(t_m, \eta_{m,q}^{[0]})$$

$$\mathbf{C:} \quad \eta_{m,q} = \eta_{m-1} + h_{m-1} \sum_{j=1}^q \beta_{q-j,q}(h_{m-1}, h_{m-2}, \dots, h_{m-q}) f_{m-j} +$$

$$+ h_{m-1} \beta_{q,q}(h_{m-1}, h_{m-2}, \dots, h_{m-q}) f_{m,q}^{[0]}$$



Simultane Änderung von h und q

Hat man bis t_{m-1} integriert, so hat man zunächst eine **Vorschlagsschrittweite** \tilde{h}_{m-1} für den **laufenden** Schritt nach t_m . Die schrittweitenabhängigen Koeffizienten β_j und β_j^* können durch numerische Integration (Polynomintegration!) exakt errechnet werden. Ist q die zur Zeit verwendete Schrittzahl, dann wertet man die Formeln aus für die Schrittzahlen $q-1, q, q+1$. In

$$\eta_{m,l} - \eta_{m,l}^{[0]}, \quad l \in \{q-1, q, q+1\}$$

hat man dann Schätzungen für den lokalen Diskretisierungsfehler der Prädiktor-Verfahren mit der Ordnung l : $\delta_{m,l}(\tilde{h}_{m-1})$. Nach der üblichen Schrittweitentechnik sollte gelten

$$\|\delta_{m,l}(h_{m-1})\| \leq \varepsilon h_{m-1} \|\eta_{m,l}\| / (t_E - t_0).$$

Es ist aber

$$\delta_{m,l}(h) = C_{m,l} h^{l+1} + \mathcal{O}(h^{l+2})$$

$$\delta_{m,l}(\tilde{h}_{m-1}) = \eta_{m,l} - \eta_{m,l}^{[0]} + \mathcal{O}(\tilde{h}_{m-1}^{l+2})$$

$$d.h. \quad C_{m,l} = (\eta_{m,l} - \eta_{m,l}^{[0]}) / \tilde{h}_{m-1}^{l+1} + \mathcal{O}(\tilde{h}_{m-1}).$$

Die tatsächlich zu verwendende Schrittweite für das Verfahren der Ordnung l errechnet sich somit aus

$$h_{m-1,l} \stackrel{\text{def}}{=} \tilde{h}_{m-1} \cdot \left(\frac{\varepsilon \tilde{h}_{m-1} \|\eta_{m,l}\|}{(t_E - t_0) \|\eta_{m,l} - \eta_{m,l}^{[0]}\|} \right)^{1/l}$$

1. Fall (Vorschlagsschrittweite, mit der gerechnet wurde, ist klein genug)

$h_{m-1,q} \geq \tilde{h}_{m-1}$, dann

$$\begin{aligned} h_{m-1} &\stackrel{\text{def}}{=} \tilde{h}_{m-1}, & \eta_m &\stackrel{\text{def}}{=} \eta_{m,q}, & \mathbf{E}: & f_m \stackrel{\text{def}}{=} f(t_m, \eta_{m,q}) \\ \tilde{h}_m &\stackrel{\text{def}}{=} \max\{h_{m-1,l} : & q-1 \leq l \leq q+1\} \end{aligned}$$

(Vorschlagsschrittweite für den nächsten Schritt)

$$q \stackrel{\text{def}}{=} \operatorname{argmax} \{h_{m-1,l} : q-1 \leq l \leq q+1\}$$

(Ordnungsvorschlag für den nächsten Schritt)

2. Fall $h_{m-1,q} < \tilde{h}_{m-1}$. Dann $\tilde{h}_{m-1} \stackrel{\text{def}}{=} \tilde{h}_{m-1}/2$ und Wiederholung dieses Schrittes.

Auf diese Art erhält man ein MSV variabler Schrittweite und Ordnung.

Gear, Watanabe und Tu haben bewiesen, daß dieses Vorgehen zu einem konvergenten Verfahren führt (mit $h = \sup h_i \rightarrow 0$), solange der Quotient $\sup h_i / \inf h_i$ beschränkt bleibt (SIAM J. Numer. Anal. 11, (1974), 1025-1058).

Diese Vorgehensweise ist dann zu empfehlen, wenn die Auswertung von f sehr aufwendig ist, hohe Genauigkeit erforderlich ist und keine "Steifheit" der DGL vorliegt. Details findet man in dem Buch: Shampine, L.F.; Gordon, M.K., Computer solution of ordinary differential equations. Freeman (1975)

1.11 Steife Differentialgleichungen

1.11.1 Einführung

Von der Behandlung der Interpolations-, Approximations- und Quadraturaufgaben her wissen wir, daß der Fehler dieser Verfahren entscheidend von einer der höheren Ableitungen der zu approximierenden Funktion bzw. des Integranden abhängt. Die Annahme, daß dies bei der Lösung einer Anfangswertaufgabe analog gilt, ist zunächst naheliegend. Tatsächlich spielt aber hier nicht nur die Glattheit der wahren Lösung eine Rolle, sondern auch die Glattheit der gesamten Lösungsmannigfaltigkeit der Differentialgleichung in der Umgebung der wahren Lösung. Dies hat schwerwiegende Folgen für die Anwendung unserer Integrationsmethoden. Wir beginnen die Diskussion mit einem besonders durchsichtigen Beispiel:

Beispiel 1.11.1. Vorgelegt sei das Anfangswertproblem

$$y' = \lambda(y - \exp(-x)) - \exp(-x), \quad y(0) = 1 + \varepsilon.$$

Die Differentialgleichung hat die allgemeine Lösung

$$y(x) = \alpha \exp(\lambda x) + \exp(-x), \quad \alpha \in \mathbb{R}.$$

Durch die Vorgabe des Anfangswertes wird α zu einer Funktion von ε :

$$1 + \varepsilon = \alpha + 1 \quad \Rightarrow \quad \alpha = \varepsilon,$$

d.h. die Lösung der Anfangswertaufgabe lautet

$$y(x) = \varepsilon \exp(\lambda x) + \exp(-x).$$

Für $\varepsilon = 0$ und $x \geq 0$ sind y und alle Ableitungen von y betragsmäßig kleiner als 1, wie auch immer λ gewählt ist!

Es ist allgemein

$$y^{(p)}(x) = \varepsilon \lambda^p \exp(\lambda x) + (-1)^p \exp(-x),$$

also

$$|y^{(p)}(x)| \geq |\lambda|^p (\varepsilon \exp(-1) - 1/|\lambda|^p)$$

für $x \in [0, -1/\lambda]$ und $\lambda < 0$. Für $\varepsilon \neq 0$ und z.B. $\lambda = -10^3$ nehmen daher die Ableitungen von y in einer kleinen Umgebung von 0 riesige Größenordnungen an.

Entsprechendes gilt, wenn wir für irgendein x_0 als Anfangswert vorschreiben

$$y(x_0) = \exp(-x_0) + \varepsilon.$$

Die folgende Abbildung zeigt die Lösungsschar und die Partikulärlösung dieser DGL für den harmlosen Fall $\lambda = -10$.

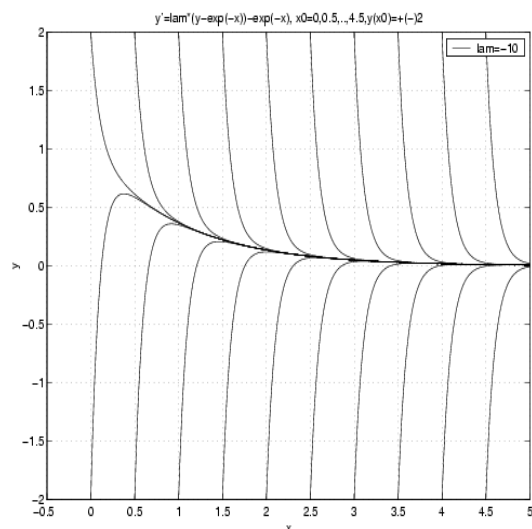


Abb. 1.11.1

Sei im folgenden $\lambda = -10^3$, $\varepsilon = 0$ gewählt. y besteht dann nur aus der glatten Komponente $\exp(-x)$. Schon für $x > \frac{1}{10}$ ist jede Lösung mit einem anderen Anfangswert bei $x = 0$ von der glatten Komponente praktisch ununterscheidbar, wenn ε in der Größenordnung von 1 bleibt. ($\exp(-100) = 3.7 \cdot 10^{-44}$).

Wahre Lösung und glatte Komponente fallen exponentiell mit x . Wir wenden nun auf dieses Problem einige unserer Diskretisierungsmethoden an und beginnen mit dem Euler-Cauchy-Verfahren.

Es gilt die Beziehung

$$\eta_{i+1} = \eta_i + h \left(-1000(y_i^h - \exp(-ih)) - \exp(-ih) \right), \quad \eta_0 = 1,$$

also

$$\eta_{i+1} = (1 - 1000h)\eta_i + 999h \exp(-ih), \quad i = 0, 1, \dots \quad \eta_0 = 1. \quad (1.30)$$

Die Ergebnisse finden sich in folgender Tabelle.

Euler-Cauchy-Näherungen für die Anfangswertaufgabe

$$y' = -1000(y - \exp(-x)) - \exp(-x), \quad y(0) = 1$$

x	$y(x)$	$h = 0.1E + 00$	$h = 0.1E - 01$	$h = 0.2E - 02$
0.1	0.9048374E + 00	0.9000000E + 00	1.7394165E + 04	9.0483751E - 01
0.5	0.6065307E + 00	-4.6047123E + 05	2.5708774E + 42	6.0653105E - 01
1.0	0.3678794E + 00	4.37904136E + 15	1.3249724E + 90	3.6788007E - 01
x	$y(x)$	$h = 0.1E - 02$	$h = 0.5E - 03$	
0.1	0.9048374E + 00	0.9048370E + 00	0.9048372E + 00	
0.5	0.6065307E + 00	0.6065304E + 00	0.6065305E + 00	
1.0	0.3678794E + 00	0.3678793E + 00	0.3678793E + 00	

Man erkennt, daß erst für $h = 0.002$ vernünftige Näherungen der Lösung errechnet werden. Offenbar liegt dies daran, daß in (1.30) $|1 - 1000h| \leq 1$ gelten sollte. Wenden wir das ebenfalls sehr einfache Verfahren „Euler-rückwärts“ an, so erhalten wir entsprechend die Rekursion

$$\eta_{i+1} = \frac{1}{1 + 1000h} \eta_i + \frac{999h}{1 + 1000h} \exp(-(i + 1)h), \quad \eta_0 = 1, \quad (1.31)$$

und schon für $h = 0.1$ ergeben sich realistische Ergebnisse:

„Euler-rückwärts“-Näherungen für die Anfangswertaufgabe

$$y' = -1000(y - \exp(-x)) - \exp(-x), \quad y(0) = 1$$

x	$y(x)$	$h = 0.1E + 00$	$h = 0.1E - 01$	$h = 0.2E - 02$
0.1	0.9048374E + 00	0.9048837E + 00	0.9048420E + 00	0.9048383E + 00
0.5	0.6065307E + 00	0.6065621E + 00	0.6065337E + 00	0.6065313E + 00
1.0	0.3678794E + 00	0.3678985E + 00	0.3678813E + 00	0.3678798E + 00
x	$y(x)$	$h = 0.1E - 02$	$h = 0.5E - 03$	
0.1	0.9048374E + 00	0.9048379E + 00	0.9048376E + 00	
0.5	0.6065307E + 00	0.6065310E + 00	0.6065308E + 00	
1.0	0.3678794E + 00	0.3678796E + 00	0.3678795E + 00	

Die mit (1.30) erzielten schlechten Ergebnisse sind offensichtlich nicht Auswirkung der geringen Konsistenzordnung, denn auch (1.31) hat ja nur die Konsistenzordnung eins. Genau die gleichen Effekte wie beim Euler-Cauchy-Verfahren würde man z.B. bei jedem expliziten Runge-Kutta-Verfahren beobachten.

Eine häufig benutzte Testgleichung ist auch die van der Pol Gleichung

$$y'' = \mu(1 - y^2)y' - y$$

Hier zeigt die Lösung lokal sehr unterschiedliches Verhalten, in gewissen Bereichen variiert sie langsam, in anderen nimmt die Ableitung extreme Werte an, falls μ gross ist :

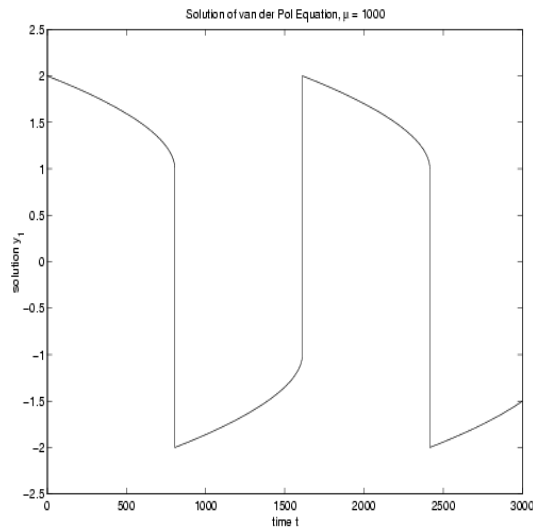


Abb 6.11.2 van der Pol : Lösung

Die in Beispiel 1.11.1 beobachteten Phänomene haben offensichtlich damit zu tun, daß sich in der Umgebung der wahren Lösung y des Anfangswertproblems Lösungen \tilde{y} der Differentialgleichung befinden, die extrem große Ableitungen besitzen. Die Werte der diskretisierten Lösung liegen ebenfalls in diesem Bereich extremer Steigungen, und bei den expliziten Verfahren entsteht, wenn h nicht klein genug ist, ein oszillierender Effekt, der schnell von der wahren Lösung wegführt. Die Anfangswertaufgabe $y' = -y$, $y(0) = 1$ mit der gleichen exakten Lösung $y = \exp(-x)$ könnte dagegen mit dem Euler-Cauchy-Verfahren und $h = 0.1$ bis $x = 1$ noch recht gut integriert werden. Differentialgleichungen, bei denen die Lösungsmannigfaltigkeit Lösungen enthält, deren Wachstumsverhalten sich wesentlich von dem der gesuchten Lösung des spezifischen Anfangswertproblems unterscheidet, nennt man *steif*. Wie der Begriff der *schlechten Kondition* einer Matrix ist auch der Begriff *Steifheit einer Differentialgleichung* eher qualitativer als quantitativer Art.

Prototyp einer steifen Differentialgleichung ist das lineare homogene System mit konstanten Koeffizienten

$$y' = Ay, \quad y(0) = y_0, \\ A \in \mathbb{R}^{n \times n} \text{ diagonalisierbar mit } \Re\lambda_1 \leq \dots \leq \Re\lambda_n \leq 0,$$

und

$$-\Re\lambda_1 \gg 1.$$

Hier könnte man

$$s \stackrel{\text{def}}{=} -\Re\lambda_1 \tag{1.32}$$

als Maß für die Steifheit wählen. Allgemeiner könnte man

$$s \stackrel{\text{def}}{=} -\inf \frac{\langle f(x, y) - f(x, z), y - z \rangle}{\langle y - z, y - z \rangle} \tag{1.33}$$

als Steifheitsmaß der Differentialgleichung definieren, wobei $\langle \cdot, \cdot \rangle$ ein Skalarprodukt auf \mathbb{R}^n ist, \inf über alle (x, y) und $(x, z) \in \mathcal{D}_f$ zu nehmen ist mit $y \neq z$ und \mathcal{D}_f den Definitionsbereich der Funktion f bezeichnet.

Ein Anfangswertproblem wird man aber nur dann als *steif* einordnen, wenn $h s \gg 1$ gilt, wobei h eine vom Aufwand her akzeptable Schrittweite ist.

Steife Anfangswertprobleme treten in vielen Anwendungen auf, so z.B. bei der Integration der Bewegungsgleichungen elastischer Körper, in der chemischen Reaktionskinetik und bei der Transientenberechnung hochintegrierter Schaltkreise.

Die in (1.33) definierte Größe s erfüllt

$$s \geq -\inf \lambda_{\min}(\partial_2 f(x, y)),$$

so daß die Spanne des Spektrums von $\partial_2 f$ in der linken komplexen Halbebene als erster Indikator für Steifheit dienen kann. Im Folgenden ist ein Beispiel eines nichtlinearen Differentialgleichungssystems zusammen mit den Eigenwerten von $\partial_2 f$ aufgeführt.

Beispiel 1.11.2.

$$\begin{aligned}Ly_1' &= E_0 - y_2 - Ry_1, \\Cy_2' &= y_1 - F(y_2), \\F(y_2) &= A_3(y_2^3 - 3A_2y_2^2 + 3A_1y_2) \\ \text{mit } L &= C = 0.01, \quad E_0 = 0.245, \quad R = 0.017, \\A_1 &= 0.0167, \quad A_2 = 0.148, \quad A_3 = 6650.\end{aligned}$$

y_2	λ_1	λ_2	$\lambda_i = \lambda(\partial_2 f(t, (y_1, y_2)))$
0.01	-27435.9	-2.06	
0.06	-4643.2	-3.81	
0.08	1465.52	4.95	
0.15	10083.44	-.7541	
0.23	-2768.55	-5.06	
0.30	-31880.9	-1.98	

Das System besitzt zwei stabile stationäre Zustände bei $(10.44, 0.0755)$ und $(1, 0.22)$. Für y_2 im Bereich von 0.08 bis 0.2 liegt eine starke Instabilität vor, die den extrem schnellen Übergang zwischen den beiden stationären Zuständen beschreibt.

Das Problem erfordert bei der Anwendung expliziter Integrationsverfahren Schrittweiten weit unter 10^{-4} . \square

1.11.2 Lineare Stabilität von Diskretisierungsverfahren für Anfangswertprobleme

Wir betrachten im Folgenden die Anwendung von Diskretisierungsverfahren auf das Testsystem, künftig als „Testgleichung“ bezeichnet:

$$y' = \lambda y, \quad y(0) = y_0 \tag{1.34}$$

mit $\Re\lambda < 0$.

Ein homogenes System mit konstanten Koeffizienten

$$\begin{aligned}y' &= A y, \quad y(0) = y_0, \\A &\in \mathbb{R}^{n \times n} \text{ diagonalisierbar mit} \\&\Re\lambda_1 \leq \dots \leq \Re\lambda_n < 0\end{aligned}$$

kann durch eine Transformation in ein zerfallendes System mit Gleichungen (1.34) überführt werden, worin λ für einen der Eigenwerte von A steht. Die hier dargestellten Überlegungen sind also auch für solche Systeme relevant.

Für $\Re\lambda < 0$ sind die Lösungen von (1.34) exponentiell fallend und das gleiche wird man auch von den durch die Diskretisierung gewonnenen Lösungen η_i erwarten, jedenfalls

für hinreichend kleines h . Zunächst gilt es, den Zusammenhang zwischen den Werten η_i zu untersuchen, wenn das Diskretisierungsverfahren auf die Anfangswertaufgabe (1.34) angewendet wird.

Definition 1.11.1. Ein *Einschritt-Verfahren* oder *Mehrschritt-Verfahren* heißt von der Klasse (AK), wenn es, angewandt auf (1.34), zu einer Differenzgleichung

$$\sum_{j=0}^k g_j(h\lambda)\eta_{m+j} = 0, \quad m = 0, 1, \dots, N - k \quad (1.35)$$

führt, worin die Funktionen g_0, \dots, g_k , $k \geq 1$, in einer Umgebung von 0 komplex analytisch sind, $g_k(0) \neq 0$ ist und das Polynom

$$\sum_{j=0}^k g_j(0)z^j$$

der Nullstellenbedingung genügt: alle Nullstellen sind betragsmäßig kleiner als 1 und Nullstellen vom Betrag 1 sind einfach. \square

Wir betrachten einige Beispiele:

Das *klassische Runge-Kutta-Verfahren vierter Ordnung* liefert

$$k = 1, \quad g_1(z) \equiv 1, \quad g_0(z) = -(1 + z + z^2/2 + z^3/6 + z^4/24).$$

Das Verfahren *Euler-rückwärts* ergibt

$$k = 1, \quad g_1(z) = 1 - z, \quad g_0(z) \equiv -1.$$

Die *Trapezregel* führt zu

$$k = 1, \quad g_1(z) = 1 - z/2, \quad g_0(z) = -(1 + z/2).$$

Die *linearen Mehrschritt-Verfahren* ergeben

$$k \geq 1, \quad g_k(z) = \alpha_k - z\beta_k.$$

Alle diese Verfahren sind also vom Typ (AK).

Wegen Satz 1.7.1 kann die allgemeine Lösung von (1.35) und damit das Wachstumsverhalten der diskretisierten Werte η_i unmittelbar aus den Nullstellen des Polynoms (Stabilitätspolynom)

$$P(z; q) \stackrel{\text{def}}{=} \sum_{j=0}^k g_j(q)z^j \quad \text{mit } q = h\lambda \quad (1.36)$$

ermittelt werden. Aus der vorangegangenen Diskussion ist klar, daß es wünschenswert ist, wenn das Polynom P die Nullstellenbedingung für einen möglichst großen Bereich

von q -Werten in der linken komplexen Halbebene erfüllt. Diese Bereiche nennt man *Stabilitätsbereiche*, ihr Inneres *Stabilitätsgebiet*.

Definition 1.11.2. Ein *Einschritt- oder Mehrschritt-Verfahren der Klasse (AK)* heißt auf einem Gebiet G der erweiterten komplexen Ebene absolut stabil, wenn die Koeffizientenfunktionen g_j aus Definition 1.11.1 auf G analytisch sind, wenn für alle $q \in G$ $g_k(q) \neq 0$ und die Nullstellen von $P(z; q)$ aus (1.36) betragsmäßig kleiner als eins sind:

$$\forall q \in G : \quad P(z; q) = \sum_{j=0}^k g_j(q) z^j = 0 \Rightarrow |z| < 1.$$

Das Verfahren heißt A -, bzw. $A(\alpha)$ -, bzw. A_0 -stabil, falls es absolut stabil ist in G mit

$$G \supset \{z : \Re z < 0\}$$

bzw.

$$G \supset \{z : \text{Arg}(-z) \in]-\alpha, \alpha[\} \quad \text{mit } \alpha > 0$$

bzw.

$$G \supset \{z : \Re z < 0, \Im z = 0\}.$$

Falls das Verfahren A -stabil ist und zusätzlich gilt

$$\limsup_{\Re q \rightarrow -\infty} (\max\{|z| : P(z; q) = 0\}) < 1,$$

dann heißt das Verfahren L -stabil. □

Ein A -stabiles Verfahren erlaubt also nach Definition die stabile (nicht genaue!) Integration von (1.34) mit beliebiger positiver Schrittweite h . Wir können also hoffen, daß durch ein solches Verfahren auch sehr steife Systeme stabil und genau integriert werden, wenn die Schrittweite nach den Ableitungen der glatten, langsam variierenden Lösungskomponenten bemessen wird, sobald die schnell abklingenden Komponenten in der Lösung keine Rolle mehr spielen. (Wenn die genaue Annäherung der schnell abklingenden Komponenten zu Beginn der Integration gefordert ist, muß man natürlich dort h weiterhin sehr klein wählen.)

Leider gibt es kein A -stabiles explizites Verfahren und auch die Konstruktion von A -stabilen impliziten Verfahren höherer Ordnung ist recht kompliziert.

Es sollen nun kurz die Stabilitätseigenschaften der bisher besprochenen Verfahren skizziert werden:

Das *Euler-Cauchy-Verfahren* ist absolut stabil genau in

$$|z + 1| < 1,$$

ist also u.a. für Schwingungsdifferentialgleichungen stets unbrauchbar, weil es die Amplituden der diskretisierten Lösung um einen Faktor $(1 + hC)$, $C =$ von h unabhängige Konstante, pro Schritt vergrößert.

Das *Verfahren Euler-rückwärts* ist absolut stabil genau in

$$|z - 1| > 1,$$

unter anderem ist es A -stabil und L -stabil. Für Schwingungsdifferentialgleichungen ist es unbrauchbar, weil es die Amplitude der diskretisierten Lösung um einen Faktor $1/(1 + hC)$ pro Schritt dämpft.

Die *Trapezregel* ist absolut stabil genau für

$$\left| \frac{1 + z/2}{1 - z/2} \right| < 1,$$

also genau auf der linken offenen Halbebene. Die Trapezregel ist also A -stabil. Zur Integration von Schwingungsdifferentialgleichungen

$$y' = \omega i y$$

ist die Trapezregel stets geeignet, da sie amplitudentreu ist:

$$\left| \frac{1 + z/2}{1 - z/2} \right| = 1 \quad \text{mit } \Re z = 0.$$

Allerdings ist die Trapezregel nicht L -stabil. Dies macht sich bei exponentiell schnell abklingenden Lösungen dadurch unangenehm bemerkbar, daß die diskretisierte Lösung von kleinen, aber beschränkt bleibenden Oszillationen überlagert wird:

Für die diskretisierte Lösung zu (1.34) erhalten wir

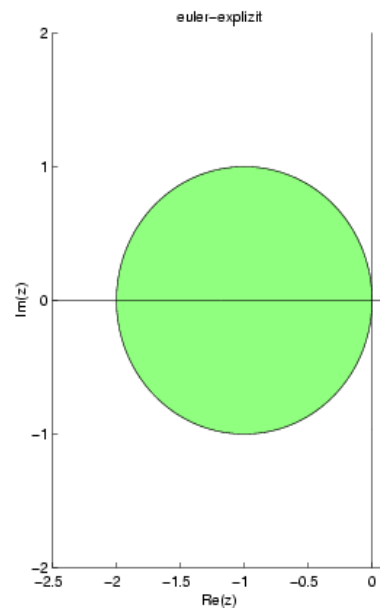
$$\begin{aligned} \eta_i &= \left(\frac{1 + h\lambda/2}{1 - h\lambda/2} \right)^i y_0 \\ &= (-1)^i \left(\frac{h\lambda + 2}{h\lambda - 2} \right)^i y_0 \\ &= (-1)^i \left(1 + \frac{4}{h\lambda - 2} \right)^i y_0 \end{aligned}$$

also für $h\lambda < -2$ eine gedämpfte, aber oszillierende Approximation einer monotonen wahren Lösung.

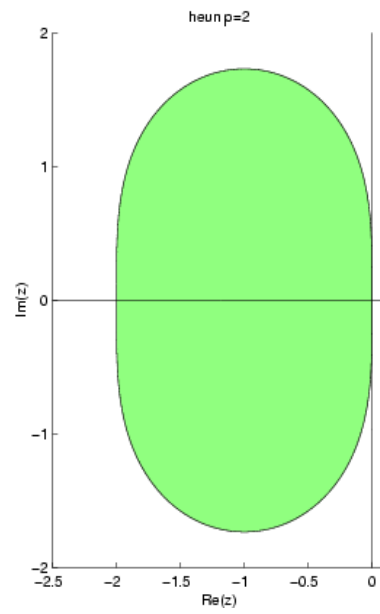
Die *BDF-Verfahren* der Ordnung und Schrittzahl k sind $A(\alpha)$ -stabil mit (vgl. Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen II, Teubner, (1977))

k	1	2	3	4	5	6	(1.37)
α	90°	90°	88°02'	73°21'	51°50'	17°50'	

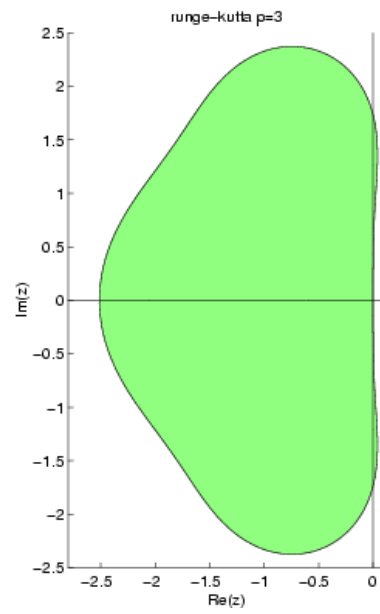
insbesondere das Verfahren der Ordnung und Schrittzahl $k = 2$ ist also A -stabil (und sogar L -stabil). Mit diesem Verfahren und der Trapezregel kennen wir zwei A -stabile lineare Mehrschritt-Verfahren. Die folgenden Abbildungen zeigen die Stabilitätsbereiche der expliziten k -stufigen Runge-Kutta-Verfahren der Stufenzahl 1 bis 4:



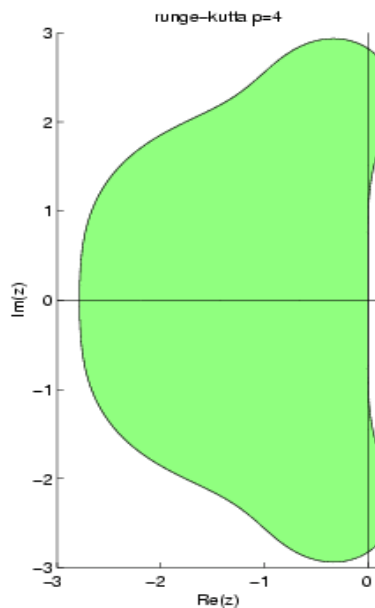
Stabilitätsgebiet des Eulerverfahrens



Stabilitätsgebiet des Heun-Verfahrens



Stabilitätsgebiet des RK3-Verfahrens



Stabilitätsgebiet des klassischen Runge-Kutta-Verfahrens

Leider gibt es keine A -stabilen linearen Mehrschritt-Verfahren höherer als zweiter Ordnung:

Satz 1.11.1. (Dahlquist's zweite Ordnungsschranke, 1963)

Jedes konsistente, lineare, A -stabile Mehrschritt-Verfahren besitzt eine Konsistenzordnung $p \leq 2$.

Zum Beweis siehe z.B. bei Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen II, Teubner, (1977) . □

Das Stabilitätsgebiet der Adams-Moulton-Verfahren der Schrittzahl $k \geq 2$, das der Adams-Bashforth-Verfahren der Schrittzahl $k \geq 1$, das der Prädiktor-Korrektor-Verfahren und das der expliziten Runge-Kutta-Verfahren ist recht klein. Für alle diese Verfahren lautet die Bedingung für absolute Stabilität

$$\left| \frac{h}{\lambda} \right| < C,$$

mit C als Konstante der Größenordnung 1. Die reellen Intervalle $]\gamma, 0[$ der absoluten Stabilität sind z.B. für die expliziten m -stufigen Runge-Kutta-Verfahren der Ordnung m , also u.a. für das klassische Runge-Kutta-Verfahren der Ordnung 4

$m = p$	1	2	3	4
γ	-2	-2	-2.51	-2.78

für die Adams-Moulton-Verfahren der Schrittzahl k

k	1	2	3	4	5
γ	$-\infty$	-6	-3	$-\frac{90}{49}$	$-\frac{45}{38}$

und für die Adams–Bashforth–Verfahren

k	1	2	3	4	5
γ	-2	-1	$-\frac{6}{11}$	$-\frac{3}{10}$	$-\frac{90}{551}$

U.a. wegen ihres kleinen Stabilitätsintervalls verwendet man die reinen Adams–Bashforth–Verfahren nicht gerne.

Von den bisher besprochenen Verfahren eignen sich also nur die Trapezregel, die implizite Mittelpunkregel und die BDF–Formeln der Schrittzahl ≤ 6 für sehr steife Systeme, letztere mit der Einschränkung, daß keine oszillierenden Lösungskomponenten mit $\arg(-\lambda) > \alpha$, α aus (1.37), auftreten dürfen.

1.11.3 Nichtlineare Stabilität

Die in Abschnitt 1.11.2 dargestellten Ergebnisse sind für die Praxis insofern unbefriedigend, als sie allenfalls grobe qualitative Rückschlüsse auf die Eigenschaften der Diskretisierungsverfahren zulassen, wenn diese auf nichtlineare steife Differentialgleichungen angewendet werden.

Schon bei linearen Systemen mit variablen Koeffizienten ändern sich die Verhältnisse grundlegend.

Beispiel 1.11.3. *Wir betrachten die Anfangswertaufgabe*

$$y' = \lambda(x)y, \quad y(0) = y_0$$

mit $\lambda : \mathbb{R}_+ \rightarrow \mathbb{R}_-$, $\lambda \in C^1(\mathbb{R}_+)$.

Hier bleibt das Verfahren Euler–rückwärts für $h > 0$ uneingeschränkt stabil. Die Trapezregel liefert die Rekursion

$$\eta_{i+1} = \frac{2 + \lambda(x_i)h}{2 - \lambda(x_{i+1})h} \eta_i,$$

und die Bedingung

$$|\eta_{i+1}| \leq |\eta_i| \quad \forall i$$

liefert für $\sup_{x \in \mathbb{R}_+} \lambda'(x) = \beta > 0$ die Einschränkung

$$h \leq 2/\sqrt{\beta}.$$

Die implizite Mittelpunkregel

$$\eta_{i+1} = \eta_i + hf\left(\frac{x_i + x_{i+1}}{2}, \frac{\eta_{i+1} + \eta_i}{2}\right)$$

ist für ein lineares System mit konstanten Koeffizienten mit der Trapezregel äquivalent, ergibt hier aber

$$\eta_{i+1} = \frac{2 + h\lambda((x_i + x_{i+1})/2)}{2 - h\lambda((x_i + x_{i+1})/2)} \eta_i,$$

ist also im vorliegenden Fall uneingeschränkt stabil.

□

Schon bei linearen Systemen mit veränderlichen Koeffizienten kann auch nicht mehr aus den Eigenwerten von $\partial_2 f(x, y)_{y=y(x)} = A(x)$ auf die Stabilität der Lösung geschlossen werden. Dies zeigt das folgende Beispiel von Kreiss (Kreiss, H.O.: Difference Methods for stiff differential equations, SINUM 15, (1978), S.21-58)

Beispiel 1.11.4.

$$y' = \frac{1}{\varepsilon} U^T(x) \begin{bmatrix} -1 & \eta \\ 0 & -1 \end{bmatrix} U(x)y = A(x)y, \quad y(0) = y_0$$

mit $\varepsilon > 0$ und

$$U(x) = \begin{bmatrix} \cos(\alpha x) & \sin(\alpha x) \\ -\sin(\alpha x) & \cos(\alpha x) \end{bmatrix}.$$

Hier sind die Eigenwerte von $A(x)$: $\lambda_1 = \lambda_2 = -1/\varepsilon$. Mit der Transformation

$$U(x) y(x) =: v(x)$$

haben wir $\|y(x)\|_2 = \|v(x)\|_2$ und

$$v' = \begin{bmatrix} -1/\varepsilon & \eta/\varepsilon + \alpha \\ -\alpha & -1/\varepsilon \end{bmatrix} v, \quad v(0) = y_0$$

mit der Lösung

$$\begin{aligned} v(x) &= v_1 \exp(\kappa_1 x) + v_2 \exp(\kappa_2 x), \quad v_1, v_2 \text{ abhängig von } y_0, \\ \kappa_{1,2} &= -\frac{1}{\varepsilon} \pm \sqrt{-\alpha(\eta/\varepsilon + \alpha)}, \end{aligned}$$

und für $0 < \varepsilon < 1$, $\alpha = -1$ und $\eta > 2$ gibt es exponentiell wachsende Lösungen. \square

Es gibt verschiedene Ansätze, die lineare Stabilitätstheorie für Diskretisierungsverfahren auf allgemeine Systeme zu übertragen. Dies ist Gegenstand gegenwärtiger intensiver Forschung. Wir beschränken unsere Darstellung auf einen speziellen Differentialgleichungstyp, für den schon recht weitreichende Resultate vorliegen.

Definition 1.11.3. Es sei $f : \mathbb{R}_+ \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ und für ein Skalarprodukt $\langle \cdot, \cdot \rangle$ auf \mathbb{R}^n gelte

$$\langle f(x, u) - f(x, v), u - v \rangle \leq m \langle u - v, u - v \rangle \quad (1.38)$$

für alle $x \in \mathbb{R}_+$ und alle $u, v \in \mathbb{R}^n$. Falls $m \leq 0$, dann heißt die Differentialgleichung $y' = f(x, y)$ dissipativ. \square

Beispiel 1.11.5. Wir betrachten die Differentialgleichung

$$\dot{x}(t) = -\nabla h(x(t)), \quad x(0) = x_0. \quad (1.39)$$

Hierbei sei $h : \mathbb{R}^n \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare, gleichmäßig konvexe Funktion, d.h. es gibt ein $\gamma > 0$, so daß

$$\gamma z^T z \leq z^T \nabla^2 h(y) z \quad \text{für alle } z \in \mathbb{R}^n, \quad y \in \mathbb{R}^n.$$

Dann wird mit den obigen Bezeichnungen $f = -\nabla h$ (nicht explizit abhängig von t) und mit $\langle x, y \rangle = x^T y$, dem gewöhnlichen euklidischen Skalarprodukt,

$$(f(x) - f(y))^T(x - y) = (y - x)^T \left(\int_0^1 \nabla^2 h(x + \tau(y - x)) d\tau \right) (x - y) \leq -\gamma(x - y)^T(x - y)$$

d.h. $m = -\gamma < 0$ in (1.38).

Die Differentialgleichung (1.39) hat als einzige stationäre Lösung die eindeutige Minimalstelle x^* von h und für $t \rightarrow \infty$ strebt die Lösung jedes Anfangswertproblems (1.39) gegen x^* . Durch numerische Integration des Anfangswertproblems kann man also x^* annähern und man wird wegen des erwünschten Übergangs $t \rightarrow \infty$ bestrebt sein, bei der Integration möglichst große Schrittweiten anzuwenden, ohne die Konvergenz der diskretisierten Lösung gegen die stationäre Lösung zu zerstören. Die Differentialgleichung (1.39) kann außerordentlich steif sein, wenn nämlich $\nabla^2 h$ schlecht konditioniert ist. \square

Man beachte, daß $\|\partial_2 f\|$ in (1.38) nicht eingeht, d.h. in die Klasse der dissipativen Differentialgleichungen fallen beliebig steife Probleme. Wir führen nun, Schneid, J.: A new approach to optimal B-convergence, Computing 38, (1987), S. 33–42 folgend, den Begriff der *optimalen B-Konvergenz* ein.

Definition 1.11.4. Ein Diskretisierungsverfahren für Anfangswertprobleme heißt optimal B-konvergent von der Ordnung p auf der Klasse aller dissipativen Anfangswertprobleme, wenn

$$\|y(x_i) - \eta_i\| \leq C(x_i)h^p \quad \text{für } h < \bar{h} \text{ und alle } i,$$

wobei $C(x_i)$ nur abhängig ist von

$$\max\{\|y^{(j)}(x)\| : x_0 \leq x \leq x_i, \quad 1 \leq j \leq \bar{p}\}$$

für ein gewisses \bar{p} (genügende Differenzierbarkeit der wahren Lösung $y(x)$ des Anfangswertproblems vorausgesetzt), für f mit (1.38) und $m \leq 0$. \square

Der große Vorteil der optimal B-konvergenten Verfahren ist es, daß die Schrittweitenrestriktion nicht von der Steifheit des Systems abhängt und daß der globale Diskretisierungsfehler ebenfalls von der Steifheit des Systems unabhängig ist. Im Gegensatz dazu haben etwa die expliziten Runge–Kutta–Verfahren einen Fehler $h^p C(x_i)$, worin $C(x_i)$ auch von $\|\partial_2 f\|$ abhängt. \bar{h} in Definition 1.11.4 hängt i.a. von m ab. In der Arbeit Schneid ist folgendes Resultat zu finden:

Satz 1.11.2. Das Verfahren Euler-rückwärts ist optimal B-konvergent von der Ordnung 1. Die Trapezregel und die implizite Mittelpunkregel sind optimal B-konvergent von der Ordnung 2, jeweils auf der Klasse der dissipativen Anfangswertprobleme. \square

Weitere optimal B-konvergente Verfahren werden wir im nächsten Abschnitt besprechen.

1.11.4 Implizite Runge–Kutta–Verfahren und Rosenbrock–Verfahren

Die allgemeinen Runge–Kutta–Verfahren sind definiert durch

$$k_i(t, y) = f\left(t + \alpha_i h, y + h \sum_{j=1}^m \beta_{ij} k_j(t, y)\right)$$

$$i = 1, \dots, m,$$

$$\tilde{\Phi}(y, t, h; f) = \sum_{j=1}^m \gamma_j k_j(t, y).$$

Die k_i sind also nur implizit (als Lösungen eines Gleichungssystems) definiert, wenn $\beta_{ij} \neq 0$ für ein $j \geq i$. Implizite Verfahren liegen in dieser Klasse vor, wenn einer der Koeffizienten β_{ij} mit $j \geq i$ ungleich null ist.

Wegen des hohen Aufwandes, der notwendig ist, um die auftretenden impliziten Gleichungen zu lösen, wäre die Anwendung solcher Verfahren bei nichtsteifen Differentialgleichungen sehr unökonomisch. Bei sehr steifen Problemen bietet sie aber große Vorteile, da es unter ihnen optimal B -konvergente Verfahren beliebig hoher Ordnung gibt.

Wir zeigen zunächst, daß auch die impliziten Runge–Kutta–Verfahren zur Klasse (AK) gehören, vgl. Definition 1.11.1.

Wir haben

$$k_j = f(x + h\alpha_j, \eta_i + h \sum_{l=1}^m \beta_{jl} k_l), \quad j = 1, \dots, m,$$

für $f(x, y) = \lambda y$ und $n = 1$ also

$$(I - \lambda h B)k = \lambda \eta_i e \quad \text{mit} \quad \begin{aligned} k &= [k_1, \dots, k_m]^T, \\ e &= [1, \dots, 1]^T \in \mathbb{R}^m, \\ B &= [\beta_{jl}]. \end{aligned}$$

Für hinreichend kleines h ist $(I - \lambda h B)$ stets invertierbar und somit k wohldefiniert. Nach der Cramerschen Regel ist dann

$$k_j = \lambda \eta_i R_j(\lambda h), \quad j = 1, \dots, m,$$

worin R_j eine rationale Funktion in λh ist:

$$R_j(z) = P_{j,m-1}(z) / \det(I - zB) = P_{j,m-1}(z) \left(\prod_{i=1}^m (1 - z\mu_i) \right)^{-1},$$

wobei μ_i die Eigenwerte von B sind.

Der Zählergrad (Grad des Polynoms $P_{j,m-1}$) ist höchstens $m - 1$, der Nennergrad ist höchstens m .

Wegen

$$\eta_{i+1} = \eta_i + h \sum_{j=1}^m \gamma_j k_j$$

wird

$$\begin{aligned} \eta_{i+1} &= \eta_i + h \lambda y_i^h \sum_{j=1}^m \gamma_j R_j(\lambda h) \\ &= \tilde{R}_m(\lambda h) \eta_i, \end{aligned}$$

worin \tilde{R}_m eine rationale Funktion mit Zählergrad und Nennergrad $\leq m$ ist.

Beispiel 1.11.6. Wir betrachten das zum Schema mit $m = 2$

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

gehörende Verfahren. Mit $f(x, y) = \lambda y$ wird

$$\begin{aligned} k_1 &= \lambda(\eta_i + \frac{h}{4}k_1), \\ k_2 &= \lambda(\eta_i + h(\frac{1}{2}k_1 + \frac{1}{4}k_2)), \end{aligned}$$

also

$$\begin{aligned} k_1 &= \frac{\lambda \eta_i}{1 - h\lambda/4}, \\ k_2 &= \frac{\lambda \eta_i}{1 - h\lambda/4} + \frac{(\lambda h/2)\lambda \eta_i}{(1 - h\lambda/4)^2}, \end{aligned}$$

d.h.

$$\begin{aligned} \eta_{i+1} &= \eta_i \left(\frac{(1 - h\lambda/4)^2 + h\lambda(1 - h\lambda/4) + (h\lambda/2)^2}{(1 - h\lambda/4)^2} \right) \\ &= \eta_i \left(1 + \frac{h\lambda}{(1 - h\lambda/4)^2} \right). \end{aligned}$$

Wegen

$$\left| 1 + \frac{h\lambda}{(1 - h\lambda/4)^2} \right| = \frac{(1 + (\Re\lambda)h/4)^2 + (\Im\lambda)^2 h^2/16}{(1 - (\Re\lambda)h/4)^2 + (\Im\lambda)^2 h^2/16}$$

ist dieses Verfahren A-stabil. Es ist konsistent von der Ordnung 2.

Nach Burrage, K.; Hundsdorfer, W.H.: The order of algebraically stable Runge-Kutta-methods, BIT27, (1987), S.62–71 ist dieses Verfahren auch optimal B-konvergent von der Ordnung 2. \square

Die linearen Stabilitätseigenschaften eines impliziten Runge–Kutta–Verfahrens hängen also gemäß Definition 1.11.2 davon ab, für welche $z \in \mathbb{C}$

$$|\tilde{R}_m(z)| < 1$$

gilt. In diesen Verfahrensklassen sind A –stabile Verfahren beliebig hoher Ordnung möglich.

Wenn man als α_i die Gauß–Knoten zur Gewichtsfunktion 1 auf $[0, 1]$ wählt, als γ_i die zugehörigen Gauß–Gewichte und die β_{ij} (die inneren Quadraturgewichte) so, daß die Formel

$$\sum_{j=1}^m \beta_{ij} g(\alpha_j) = \int_0^{\alpha_i} g(x) dx \quad \text{für } g \in \Pi_{m-1}, \quad i = 1, \dots, m$$

gilt, erhält man die sogenannten *Gauß–Runge–Kutta–Verfahren*. Für $m = 2$ ergibt sich z.B. das Koeffizientenschema

$(3 - \sqrt{3})/6$	$\frac{1}{4}$	$(3 - 2\sqrt{3})/12$
$(3 + \sqrt{3})/6$	$(3 + 2\sqrt{3})/12$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

Es gilt

Satz 1.11.3.

1. Die m –stufigen Gauß–Runge–Kutta–Verfahren sind konsistent von der Ordnung $2m$.

Zum Beweis siehe z.B. bei Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen I, Teubner, (1972).

2. Die m –stufigen Gauß–Runge–Kutta–Verfahren sind A –stabil und optimal B –konvergent von der Ordnung m auf der Klasse der dissipativen Probleme.

Zum Beweis siehe z.B. bei Burrage, K. und Hundsdorfer, W.H. □

Die Ordnung der B –Konvergenz kann also wesentlich kleiner sein als die klassische Konsistenzordnung. Dies liegt darin begründet, daß bei der optimalen B –Konvergenz verlangt wird, daß das Restglied unabhängig ist von der Steifheit des Systems.

Man kann zeigen, daß bei dissipativem f die Steigungen k_j eines Gauß–Runge–Kutta–Verfahrens für ein festes $\bar{h} > 0$ für $h \leq \bar{h}$ eindeutig aus der Verfahrensgleichung bestimmt sind, unabhängig von der Steifheit des Systems. Dennoch ist die Implementierung dieser Verfahren mit erheblichen Schwierigkeiten belastet, da u.a. bei jedem Integrationsschritt für ein Differentialgleichungssystem der Ordnung n ein nichtlineares Gleichungssystem der Ordnung nm gelöst werden muß. Außerdem ist eine Schrittweitenkontrolle nur auf dem Weg über Schrittweithalbung möglich, was den Aufwand weiter erhöht.

Für die Praxis einfacher sind die Verhältnisse, wenn

$$\beta_{ij} = 0 \quad \text{für } j > i.$$

Dies sind die sogenannten *diagonal-impliziten Runge-Kutta-Verfahren*. Von diesen Verfahren haben wir eines bereits im Beispiel 1.11.6 kennengelernt, sie sind aber meist nur von erster Ordnung optimal B -konvergent Frank, R. und Schneid, J. und Überhuber, C. W.: Stability properties of implicit Runge-Kutta methods, J. SIAM Numer. Anal. 22, (1985), S. 497–514.

Wenn man steife Probleme löst, bei denen nur $\|\partial_2 f\|$ groß ist, während die Größenordnung aller anderen partiellen Ableitungen von f klein ist gegen $\|\partial_2 f\|$, also z.B.

$$y' = A y + g(x, y)$$

mit einer Funktion g , deren sämtliche partielle Ableitungen eine Norm in der Größenordnung von 1 haben, während die Eigenwerte von A in der linken komplexen Halbebene sehr weit gestreut liegen, kann man mit gutem Erfolg die sehr viel einfacher handhabbaren Rosenbrock-Verfahren und ihre Modifikationen anwenden.

Diese Verfahren kann man sich auf folgende Weise entstanden denken. Wir gehen aus von den Bestimmungsgleichungen eines diagonal-impliziten Runge-Kutta-Verfahrens:

$$\begin{aligned} k_1 &= f(x + \alpha_1 h, \eta + \beta_{11} h k_1), \\ k_2 &= f(x + \alpha_2 h, \eta + \beta_{21} h k_1 + \beta_{22} h k_2), \\ \dots &\quad \dots \dots \dots \\ k_m &= f(x + \alpha_m h, \eta + \beta_{m1} h k_1 + \dots + \beta_{mm} h k_m). \end{aligned}$$

Diese Gleichungen werden sukzessiv nur näherungsweise gelöst, indem jeweils ein Schritt des Newton-Verfahrens mit $k_i^{[0]} = 0$ als erster Näherung durchgeführt wird. Dies liefert dann die expliziten Bestimmungsgleichungen

$$\left(I - h \beta_{ii} \partial_2 f(x + \alpha_i h, \eta + h \sum_{j=1}^{i-1} \beta_{ij} k_j) \right) k_i = f\left(x + \alpha_i h, \eta + h \sum_{j=1}^{i-1} \beta_{ij} k_j\right),$$

$i = 1, \dots, m$.

Hier hat man also m lineare Gleichungssysteme sukzessiv zu lösen. Der Nachteil dieses Zugangs besteht noch darin, daß man m Jacobi-Matrizen $\partial_2 f$ berechnen und m Dreieckszerlegungen ausführen muß. Mit einer festen Matrix

$$I - h \beta \partial_2 f(x, y)$$

auf der linken Seite und einer entsprechenden Korrektur auf der rechten Seite erhält man die modifizierte Ansatzformel

$$\left(I - h \beta \partial_2 f(x, \eta) \right) k_i = f\left(x + \alpha_i h, \eta + h \sum_{j=1}^{i-1} \beta_{ij} k_j\right) + h \partial_2 f(x, \eta) \sum_{j=1}^{i-1} \sigma_{ij} k_j, \quad (1.40)$$

$i = 1, \dots, m$.

Dies sind die sogenannten *Rosenbrock-Wanner-Formeln*. In dieser Verfahrensklasse ist eine Fülle A - bzw. $A(\alpha)$ -stabiler Verfahren hoher Ordnung bekannt, vgl. etwa bei Kaps,

P. und Rentrop, P.: Generalized Runge Kutta methods of order 4 with stepsize control for stiff ODE's , Numer. Math. 33, (1979), S. 55–68 und Kaps, P. und Wanner, G.: A study of Rosenbrock methods of high order , Numer. Math. 38, (1982), S.279–298.

In Kaps / Rentrop sind folgende Koeffizienten eines A -stabilen Verfahrens der Ordnung 4 mit einem eingebetteten A -stabilen Verfahren der Ordnung 3 angegeben:

$$\beta = 0.395$$

$i \setminus j$	1	2	3
2	0.438		
3	0.796920457938	0.073079542062	
4	0	0	0

$i \setminus j$	1	2	3	
2	-0.767672395484			
3	-0.851675323742	0.522967289188		
4	0.288463109545	0.08802142734	-0.337389840627	(1.41)

i	1	2	3	4	Ordnung
γ_i	0.199293275701	0.482645235674	0.0680614886256	0.25	4
$\hat{\gamma}_i$	0.346325833758	0.285693175712	0.367980990530	0	3

Hier kann man also ohne wesentlichen Zusatzaufwand den lokalen Diskretisierungsfehler schätzen durch $h \|\sum_{i=1}^4 (\gamma_i - \hat{\gamma}_i) k_i\|$ und darauf auch eine Schrittweitensteuerung aufbauen.

Gottwald und Wanner haben in Gottwald, B.A. und Wanner, G.: A reliable Rosenbrock Integrator for stiff Differential Equations, Computing 26, (1981), S. 355–360 ein auf dieser Formel basierendes Programm beschrieben, mit dem sehr gute Resultate erzielt werden.

Beispiel 1.11.7. *Wir betrachten das Anfangswertproblem*

$$\begin{aligned} \dot{y}_1 &= c_6 y_1 / y_2^3 - p(t) c_2, \\ \dot{y}_2 &= y_1, \\ y_1(0) &= 0, \\ y_2(0) &= \frac{1}{2} 10^{-4}, \\ p(t) &= \frac{1}{0.5 + (t - 0.2)^2} + \frac{1}{0.02 + (t - 0.7)^2} + \frac{1}{0.05 + (t - 1.1)^2} \end{aligned}$$

für $t \in [0, 1.5]$.

Dabei ist $c_1 = \pi$, $c_2 = 0.013$, $c_3 = 0.088$, $c_4 = 0.035$, $c_5 = 4 \cdot 10^{-3}$, $c_6 = -c_1 c_3 c_4 c_5^3 / c_2 = -4.7636 \cdot 10^{-8}$.

Die Differentialgleichung beschreibt in vereinfachter Form die Bewegung eines Metallringes, der in einer viskosen Flüssigkeit durch eine äußere Kraft p gegen eine ebene

starre Unterlage gedrückt wird. y_2 ist der Abstand und y_1 die Geschwindigkeit des Ringes. Mit dem durch (1.40) und (1.41) gegebenen Verfahren und einer Schrittweitenstrategie wie in Abschnitt 14.4 beschrieben sowie folgender Forderung an den lokalen Diskretisierungsfehler:

$$\left(\frac{1}{n} \sum_{j=1}^n \left(\frac{\eta_{i+1,j}^{[1]} - \eta_{i+1,j}^{[2]}}{\max\{|\eta_{i+1,j}^{[2]}|, |\eta_{i,j}^{[2]}|, \varrho\}} \right)^2 \right)^{1/2} \leq \varepsilon,$$

$$\varepsilon = 10^{-2}, \quad \varrho = 10^{-12}, \quad h_{\max} = 0.1, \quad h_{\min} = 10^{-12},$$

ergab sich folgendes Resultat

```

VORGEGEBENE FEHLERTOLERANZ= 1.0000000000000000E-02
HMAX= 0.1000000000000000
HMIN= 1.0000000000000000E-12
REYNOLDS DGL
      T=          Y(1)=          Y(2)=          HCUR=          NFE=
0.30000D-01  -.32958D-03  0.34963D-04  0.27023D-02          180
0.60000D-01  -.17454D-03  0.27721D-04  0.29907D-04          759
0.90000D-01  -.11262D-03  0.23498D-04  0.12785D-03         1415
0.12000D+00  -.80590D-04  0.20615D-04  0.30231D-04         2259
0.15000D+00  -.61794D-04  0.18476D-04  0.13872D-04         3157
0.18000D+00  -.49629D-04  0.16798D-04  0.66212D-05         4112
0.21000D+00  -.41244D-04  0.15430D-04  0.64712D-04         5214
0.24000D+00  -.35139D-04  0.14278D-04  0.65740D-04         6710
0.27000D+00  -.30659D-04  0.13282D-04  0.45303D-04         8462
0.30000D+00  -.27264D-04  0.12405D-04  0.39787D-04        10446
0.33000D+00  -.24643D-04  0.11618D-04  0.21271D-04        12698
0.36000D+00  -.22607D-04  0.10902D-04  0.30489D-04        15206
0.39000D+00  -.21013D-04  0.10240D-04  0.26610D-04        18106
0.42000D+00  -.19776D-04  0.96215D-05  0.22232D-04        21402
0.45000D+00  -.18821D-04  0.90358D-05  0.24097D-04        25142
0.48000D+00  -.18090D-04  0.84754D-05  0.13601D-04        29402
0.51000D+00  -.17518D-04  0.79348D-05  0.23521D-04        34182
0.54000D+00  -.17047D-04  0.74096D-05  0.12051D-04        39602
0.57000D+00  -.16575D-04  0.68985D-05  0.13213D-04        45578
0.60000D+00  -.15970D-04  0.64034D-05  0.11317D-04        52030
0.63000D+00  -.15058D-04  0.59308D-05  0.14627D-04        58578
0.66000D+00  -.13678D-04  0.54928D-05  0.13341D-04        64494
0.69000D+00  -.11802D-04  0.51050D-05  0.48203D-04        68430
0.72000D+00  -.98244D-05  0.47888D-05  0.14235D-02        69318
0.75000D+00  -.78071D-05  0.45255D-05  0.40656D-06        70616
0.78000D+00  -.60268D-05  0.43218D-05  0.45795D-06        75272
0.81000D+00  -.46985D-05  0.41641D-05  0.15810D-05        81514

```

0.84000D+00	-.37613D-05	0.40398D-05	0.24088D-05	88142
0.87000D+00	-.31225D-05	0.39386D-05	0.21628D-04	94002
0.90000D+00	-.26957D-05	0.38530D-05	0.21585D-04	98722
0.93000D+00	-.24150D-05	0.37778D-05	0.52761D-04	101866
0.96000D+00	-.22315D-05	0.37094D-05	0.97044D-04	103498
0.99000D+00	-.21075D-05	0.36465D-05	0.57444D-03	104010
0.10200D+01	-.20108D-05	0.35897D-05	0.25160D-02	104110
0.10500D+01	-.19290D-05	0.35399D-05	0.70829D-02	104138
0.10800D+01	-.18723D-05	0.35047D-05	0.15000D-01	104150
0.11100D+01	-.18295D-05	0.34772D-05	0.15000D-01	104158
0.11400D+01	-.16313D-05	0.34265D-05	0.99661D-07	105499
0.11700D+01	-.14584D-05	0.33807D-05	0.17561D-05	107994
0.12000D+01	-.12831D-05	0.33401D-05	0.10508D-05	113170
0.12300D+01	-.11172D-05	0.33046D-05	0.11391D-05	119006
0.12600D+01	-.96735D-06	0.32738D-05	0.31511D-05	125402
0.12900D+01	-.83643D-06	0.32472D-05	0.90146D-05	132290
0.13200D+01	-.72457D-06	0.32242D-05	0.89689D-05	139354
0.13500D+01	-.62995D-06	0.32042D-05	0.10132D-04	146378
0.13800D+01	-.55046D-06	0.31867D-05	0.14650D-04	153270
0.14100D+01	-.48365D-06	0.31715D-05	0.53374D-05	159986
0.14400D+01	-.42752D-06	0.31580D-05	0.19304D-04	166430
0.14700D+01	-.38006D-06	0.31461D-05	0.84341D-05	172690
0.15000D+01	-.33979D-06	0.31355D-05	0.36816D-05	178694

ENDRESULTAT AUS GRK4A

Y(1)=-0.3397934D-06 Y(2)= 0.3135464D-05

ANZAHL DER ERFOLGREICHEN INTEGRATIONSSCHRITTE	44147
ANZAHL DER WIEDERHOLTEN SCHRITTE	702
ANZAHL FUNKTIONSAUFRUFE	178694
ANZAHL DER AUSWERTUNGEN JF	44849

CPU-ZEIT 47 SEC (VAX 8530) .

In der Spalte H_{CUR} ist die laufende Schrittweite angegeben. Man erkennt, in wie hohem Maß die Schrittweitensteuerung die Schrittweite variiert.

NFE ist die Zahl der insgesamt benutzten Funktionsauswertungen.

Das Problem ist außerordentlich steif und mit $y_2 \rightarrow 0$ nimmt die Steifheit zu. Es ist ja

$$\partial_2 f = \begin{bmatrix} c_6/y_2^3 & -3c_6y_1/y_2^4 \\ 1 & 0 \end{bmatrix},$$

also für $t = 0.6$, $y_1 = -1.56 \cdot 10^{-5}$, $y_2 = 6.45 \cdot 10^{-6}$

$$\partial_2 f = \begin{bmatrix} -1.775 \cdot 10^8 & 1.288 \cdot 10^9 \\ 1 & 0 \end{bmatrix}$$

mit den Eigenwerten

$$\lambda_1 = -7.256, \quad \lambda_2 = -1.775 \cdot 10^8.$$

Dennoch bereitet die Integration dieses Systems keinerlei Schwierigkeiten. Mit dem expliziten eingebetteten Runge-Kutta-Verfahren der Ordnung 5 und 4 von Dormand und Prince ergab sich bei gleichen Verfahrensparametern das Resultat

REYNOLDS DGL

VORGEGEBENE TOLERANZ= 1.0000000000000000E-02

HMAX= 0.1000000000000000

HMIN= 1.0000000000000000E-12

T=	Y(1)=	Y(2)=	HCUR=	NFE=
0.3000D-01	-0.3116D-03	0.3413D-04	0.3448D-05	48625
0.6000D-01	-0.1683D-03	0.2729D-04	0.1535D-05	158438
0.9000D-01	-0.1098D-03	0.2322D-04	0.1031D-05	349773
0.1200D+00	-0.7905D-04	0.2042D-04	0.6990D-06	642664
0.1500D+00	-0.6048D-04	0.1834D-04	0.4300D-06	1058387
0.1800D+00	-0.4912D-04	0.1669D-04	0.3391D-06	1619652
0.2100D+00	-0.4084D-04	0.1535D-04	0.2779D-06	2351755
0.2400D+00	-0.3485D-04	0.1421D-04	0.2223D-06	3283142
0.2700D+00	-0.3065D-04	0.1322D-04	0.1869D-06	4446615
0.3000D+00	-0.2728D-04	0.1236D-04	0.1521D-06	5880280
0.3300D+00	-0.2462D-04	0.1158D-04	0.1058D-06	7629755
0.3600D+00	-0.2262D-04	0.1087D-04	0.1061D-06	9750228
0.3900D+00	-0.2089D-04	0.1021D-04	0.2204D-07	12309607
0.4200D+00	-0.1978D-04	0.9597D-05	0.2030D-07	15393122
0.4500D+00	-0.1891D-04	0.9015D-05	0.5217D-07	19109919
0.4800D+00	-0.1818D-04	0.8458D-05	0.4838D-07	23601802
0.5100D+00	-0.1760D-04	0.7920D-05	0.3754D-07	29056265
0.5400D+00	-0.1710D-04	0.7397D-05	0.3236D-07	35724864
0.5700D+00	-0.1670D-04	0.6888D-05	0.2767D-07	43947175
0.6000D+00	-0.1596D-04	0.6395D-05	0.1702D-07	54178850
0.6300D+00	-0.1518D-04	0.5924D-05	0.1550D-07	67014291
0.6600D+00	-0.1380D-04	0.5487D-05	0.1332D-07	83175334
0.6900D+00	-0.1177D-04	0.5100D-05	0.8614D-08	103424309
0.7200D+00	-0.9681D-05	0.4774D-05	0.8012D-08	128385348
0.7500D+00	-0.7604D-05	0.4514D-05	0.6441D-08	158356225
0.7691D+00	-0.6414D-05	0.4380D-05	0.6080D-08	180000032

INTEGRATION KANN NICHT FORTGESETZT WERDEN , DA MEHR ALS 300000000
FUNKTIONSAUSWERTUNGEN !

VERBRAUCHTE CPU-ZEIT 7 STD 32 MIN (VAX 8530) .

Hier ist neben der laufenden Schrittweite H_{CUR} , die wegen der zunehmenden Steifheit des Systems erwartungsgemäß immer kleiner wird, auch die akkumulierte Anzahl der Funktionsauswertungen N_{FE} angegeben. Man sieht, daß der Aufwand völlig unvertretbar wird. Die Schrittweite liegt stets an der Stabilitätsgrenze des Verfahrens. \square

1.12 Andere Problemstellungen bzw. Methoden ERG

In der Praxis treten Differentialgleichungen in der einfachen, bisher benutzten Form

$$y' = F(t, y)$$

meist nicht auf. Oft ist es zwar möglich, durch einfache algebraische Operationen diese Form herzustellen, aber ob dies empfehlenswert ist, ist fraglich. Die Transformation von Aufgaben mit höheren Ableitungen in Systeme erster Ordnung ist in der Regel zweckmässig, es gibt nur wenige Ausnahmen, wo man direkt z.B. mit Differentialgleichungen zweiter Ordnung arbeitet. Deshalb gehen wir hier zunächst von der allgemeinen Form

$$F(t, y, y') = 0$$

aus. (Die Anfangswerte lassen wir zunächst beiseite). Hier hat man nun Fälle verschiedenen Schwierigkeitsgrades: Ist

$$F(t, y, y') = A(t, y)y' + B(t, y) + f(t) = 0$$

mit invertierbarer Matrix A , dann kann man im Prinzip nach y' auflösen und erhält die explizite Form. Dies ist aber jedenfalls dann unzuweckmässig, wenn die Dimension hoch und A dünn besetzt ist oder wenn A schlecht konditioniert ist. Man schreibt dann zunächst das Diskretisierungsverfahren für die explizite Form auf und multipliziert dann mit A zurück und erhält so eine Diskretisierungsform, bei der zusätzlich pro Schritt noch ein Gleichungssystem zu lösen ist. Wenn das Verfahren schon implizit ist, hat man hier gar nichts verloren.

Beispiel 1.12.1. *Differentialgleichung*

$$A(t, y)y' = G(t, y) .$$

Diskretisierung Euler vorwärts:

$$\eta_{m+1} = \eta_m + h(A(t_m, \eta_m))^{-1}G(t_m, \eta_m)$$

und dies ist äquivalent zu

$$A(t_m, \eta_m)(\eta_{m+1} - \eta_m) = hG(t_m, \eta_m) .$$

Für eine BDF-Diskretisierung

$$\sum_{i=0}^k \alpha_i \eta_{m-i} = h\beta_k f(t_m, \eta_m)$$

ist die Form

$$A(t_m, \eta_m) \frac{1}{h\beta_k} \sum_{i=0}^k \alpha_i \eta_{m-i} = G(t_m, \eta_m)$$

als Gleichung für η_m zu benutzen. □

Genauso kann man in Verbindung mit dem Newtonverfahren vorgehen, wenn man eine vollimplizite DGL

$$F(t, y, y') = 0$$

vorliegen hat, aber die Jacobi-Matrix von F bezüglich der dritten Variablen (y') invertierbar ist. Ein solches System, bei dem y' im Prinzip explizit darstellbar ist, nennt man vom Index 0. Hier ist auch die Beschaffung eines konsistenten Anfangswertes für y'_0 , also eines Wertes y'_0 mit

$$F(t_0, y_0, y'_0) = 0$$

kein Problem.

Das System habe jetzt die Gestalt

$$\begin{aligned} F(t, x, x', y) &= 0, \\ G(t, x, y) &= 0. \end{aligned} \tag{1.42}$$

mit den unbekannt Funktionen $x(t)$, $y(t)$, und die Jacobimatrix von F bezüglich der dritten Variablen (x') sei invertierbar. Die Ableitung von y taucht also hier gar nicht auf, y ist aber eine gesuchte (Vektor)funktion. Die Schwierigkeit dieses Falles wird im wesentlichen bestimmt durch den sogenannten "Index" des Systems. Das ist, etwas locker formuliert, die Anzahl der Differentiationsschritte, die man auf Teile des Systems anwenden muss, bis man die obige Situation "Index=0" erreicht hat. Dieser Fall kommt in den Anwendungen sehr häufig vor, etwa in der chemischen Reaktionskinetik, der Schaltkreisanalyse, bei restringierten Variationsproblemen etc. Ein einfaches, aber zugleich recht schwieriges Beispiel ist das Kreispendel

$$x'' = \lambda x, \tag{1.43}$$

$$y'' = \lambda y - g, \tag{1.44}$$

$$0 = x^2 + y^2 - L. \tag{1.45}$$

Hier ist $\lambda(t)$ die im Pendelarm wirkende Kraft, g die Erdbeschleunigung, und $x(t)$, $y(t)$ sind die Koordinaten der Pendelspitze. Dieses Beispiel besitzt Index 3. Einführende Literatur dazu: Brenan, K.E.; Campbell, S.L.; Petzold, L.R.: *Numerical solution of initial value problems in differential- algebraic equations*. Amsterdam, Elsevier North-Holland, 1989.

Ist F noch von allgemeinerer Struktur nichtlinear, also nicht in der vorstehenden Form schreibbar und zugleich ist die Ableitung von F bezüglich y' singulär, dann muss das System dann als eine Differentialgleichung auf einer Mannigfaltigkeit untersuchen. Auch dazu gibt es Methoden, Literatur:

Rabier, P.J.; Rheinboldt, W.C.: *Theoretical and numerical analysis of differential algebraic equations*. Handbook of numerical analysis VIII 183–540, (Ciarlet und Lions, eds.) Elsevier , Amsterdam 2002.

In der Praxis trifft man häufig auch auf sogenannte retardierte Differentialgleichungen. Dies sind Gleichungen der Form

$$F(t, y(t), y'(t), y(t - \tau), y'(t - \tau)) = 0$$

wobei die Retardierung τ u.U. sogar selbst noch von t abhängt. Solche Gleichungen kann man nicht mit den hier besprochenen Methoden behandeln. Es gibt dazu spezielle Methoden, siehe z.B. im Band I von Hairer, Norsett und Wanner oder bei Bellen, Alfredo; Zennaro, Marino: *Numerical methods for delay differential equations*. Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press (2003)

Alle in diesem Kapitel behandelten Methoden konzentrieren sich darauf, die Lösung der Differentialgleichung auf einem (kleinen) Zeitintervall möglichst effizient und mit möglichst hoher Genauigkeit zu approximieren. Spezielle Eigenschaften der Lösung qualitativer Art und die Möglichkeit einer Langzeitintegration bleiben dabei ausser Betracht. Dies ist aber häufig unangemessen. Eine relativ junge Forschungsdisziplin ist die der stukturerehaltenden Methoden für gewöhnliche Differentialgleichungen: symplektische Integratoren für Hamilton'sche Systeme, das sind Differentialgleichungen der Form

$$\begin{aligned} p' &= -\frac{\partial}{\partial q} H(p, q) \\ q' &= \frac{\partial}{\partial p} H(p, q) \end{aligned}$$

mit der Eigenschaft

$$H(p(t), q(t)) = \text{const} \quad \text{entlang einer Lösung,}$$

symmetrische Integratoren für reversible Systeme, Methoden, die erste Integrale erhalten , Methoden für Differentialgleichungen auf Mannigfaltigkeiten (u.a. Liegruppen) sowie spezielle Verfahren für restringierte mechanische Systeme und Gleichungen mit hochoszillatorischen Lösungen. Eine Einführung in diese Gebiete findet man bei Hairer,E.; Lubich, C.;Wanner, G.: *Geometric numerical integration*. Springer, Berlin–Heidelberg-etc (2002).

Kapitel 2

Rand- und Eigenwertaufgaben gewöhnlicher Differentialgleichungen

Aufgaben dieses Typs treten sowohl als solche auf etwa in der Regelungs- und Steuertechnik, aber auch als Subprobleme bei der Lösung partieller Differentialgleichungen im Zusammenhang mit der sogenannten (horizontalen) Linienmethode. Wir beschränken uns hier auf die Grundverfahren, diskutieren jedoch Einiges auch im Detail, weil dies das Verständnis der Methoden bei partiellen Differentialgleichungen sehr erleichtert.

2.1 Einführungsbeispiel. Einige theoretische Grundlagen. ERG

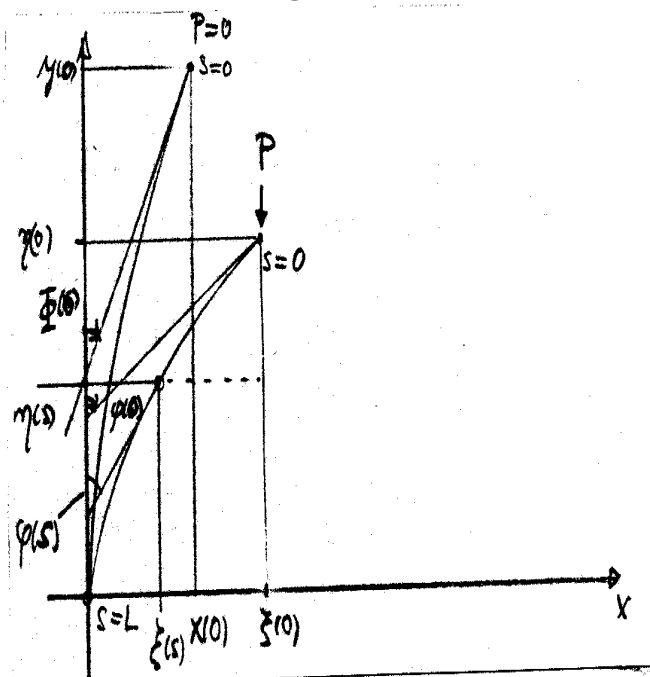
Beispiel 2.1.1. *Wir betrachten einen aufrecht stehenden (nicht notwendig geraden) Stab der Länge L , der am Fußpunkt eingespannt und am oberen Ende frei ist. In rechtwinkligen Koordinaten ist die Lage des Stabes durch die Koordinatendarstellung $(x(s), y(s))$, $0 \leq s \leq L$ ($s =$ Bogenlänge, vom freien Ende des Stabes aus gerechnet) gegeben. $\Phi(s)$ bezeichnet den Winkel der Tangente im Punkt $(x(s), y(s))$ mit der y -Achse im unbelasteten Zustand. Am freien Ende des Stabes greife eine konstante Kraft P stets senkrecht nach unten an. In der ausgelenkten Lage sei der Stab entsprechend durch $\xi(s), \eta(s), \varphi(s)$ beschrieben. $\varphi'(s)$ beschreibt die Krümmung des verformten Stabes. Das Biegemoment an der Stelle s wird proportional zur Änderung der Krümmung angenommen:*

$$B(s) = \beta(s)(\varphi - \Phi)'(s)$$

Die Biegefestigkeit β nehmen wir als ortsabhängig und strikt positiv an.
 Die Dimensionen sind:

Biegemoment : [Kraft \times Weg]
 Biegesteifigkeit : [Kraft \times Weg²]
 Änderung der Krümmung : [Weg⁻¹].

Abbildung 2.1.1.



In der Gleichgewichtslage gilt dann:

Summe der angreifenden Momente in jedem Punkt $(\xi(s), \eta(s))$ ist null.

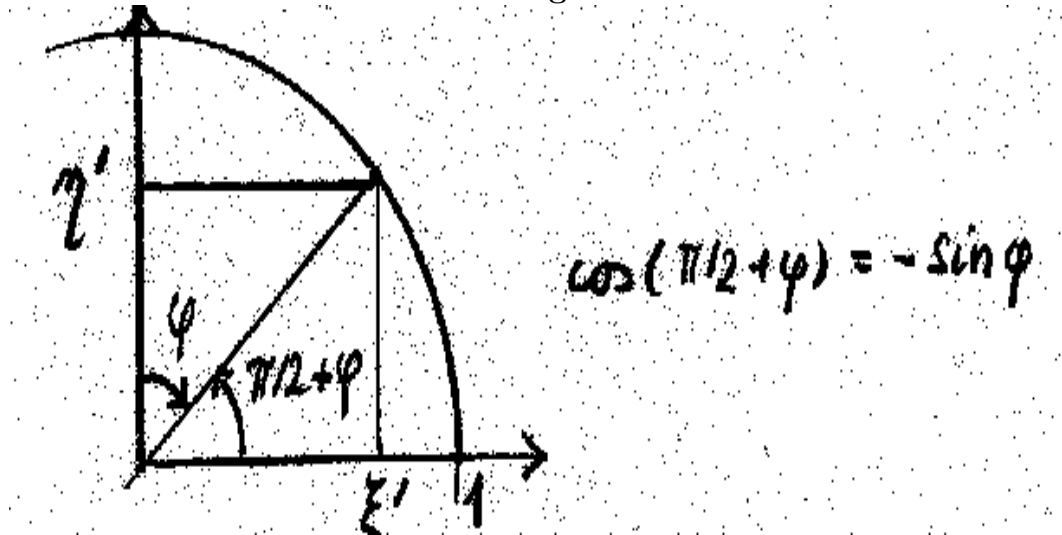
$$B(s) = P(\xi(s) - \xi(0)) \quad 0 \leq s \leq L, \quad (2.1)$$

insbesondere $\varphi'(0) = \Phi'(0)$.

Aufgrund unserer Annahme ist stets

$$\varphi(L) = 0 \quad (\text{senkrechte Einspannung})$$

Abbildung 2.1.2.



Differentiation von (2.1) liefert

$$(\beta\varphi)' - (\beta\Phi)' - P\xi'(s) = 0$$

Aber $(\xi'(s))^2 + (\eta'(s))^2 = 1$, $\xi'(s) = -\sin\varphi(s)$,
d.h. wir erhalten die DGL 2. Ordnung

$$-(\beta\varphi)' = P \sin\varphi - (\beta\Phi)' \quad 0 \leq s \leq L \quad (2.2)$$

mit den Randbedingungen $\varphi'(0) = \Phi'(0)$, $\varphi(L) = 0$.

Beim senkrecht stehenden Stab ($\Phi(s) \equiv 0$) ist natürlich $\varphi \equiv 0$ eine Lösung des Problems. In Abhängigkeit von P können jedoch auch in diesem Fall Lösungen $\varphi \neq 0$ auftreten. Die kleinste Last P , unter der dies auftritt, heißt Euler'sche Knicklast des Stabes. (Dies ist also ein Verzweigungsproblem, bei dem bei einem bestimmten Parameterwert zu einer zunächst eindeutigen Lösung eine zweite hinzukommt.)

Die hergeleitete Aufgabe (2.2) ist ein Spezialfall der allgemeinen nichtlinearen Randwertaufgabe 2. Ordnung

$$\begin{aligned} y'' &= f(x, y, y') & a \leq x \leq b \\ r_1(y(a), y(b), y'(a), y'(b)) &= 0 \\ r_2(y(a), y(b), y'(a), y'(b)) &= 0 \end{aligned}$$

Transformieren wir die DGL auf ein System 1. Ordnung, dann erhalten wir

$$\begin{aligned} u &\stackrel{\text{def}}{=} \begin{pmatrix} y \\ y' \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} & R &\stackrel{\text{def}}{=} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \\ u' &= F(x, u) & a \leq x \leq b \\ R(u(a), u(b)) &= 0 \end{aligned}$$

Die allgemeine Form einer 2-Punkt Randwertaufgabe für ein DGL-System 1. Ordnung ist

$$\begin{aligned} y' &= F(x, y), & a \leq x \leq b, \\ R(y(a), y(b)) &= 0, \end{aligned}$$

mit

$$F: \mathcal{I} \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad R: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n.$$

Selbst unter starken Voraussetzungen an F und R sind im nichtlinearen Fall keine der Situationen bei Anfangswertaufgaben entsprechenden Existenz- und Eindeutigkeitsätze bekannt. Selbst bei einfachsten linearen Randwertaufgaben kann es Probleme geben, wie man am folgenden Beispiel erkennt:

Beispiel 2.1.2.

$$y'' = -y \quad \text{d.h.} \quad y(x) = A \sin x + B \cos x$$

a) $y(0) = 0, \quad y(\pi/2) = 1 \Rightarrow y(x) = \sin x$ *eindeutige Lösung*

b) $y(0) = 0, \quad y(\pi) = 0 \Rightarrow y(x) = A \sin x, \quad A \in \mathbb{R}$ *bel.*
unendl. viele Lösungen

c) $y(0) = 0, \quad y(\pi) = 1 \Rightarrow$ **keine Lösung** □

In einigen Spezialfällen sind jedoch Aussagen bekannt, insbesondere bei Zweipunkt-Randwertaufgaben 2. Ordnung für eine skalare Funktion y :

Satz 2.1.1. *Sei*

$$S \stackrel{\text{def}}{=} [a, b] \times \mathbb{R} \times \mathbb{R} \quad \text{und} \quad f \in C^0(S).$$

Ferner gelte: f ist nach der 2. und der 3. Variablen stetig partiell differenzierbar und

$$0 < \partial_2 f(w) \leq L_1, \quad -L_2 \leq \partial_3 f(w) \leq L_2 \quad (\forall w \in S)$$

sowie

$$a_0 a_1 \geq 0, \quad b_0 b_1 \geq 0, \quad |a_0| + |b_0| \neq 0, \quad |a_0| + |a_1| > 0, \quad |b_0| + |b_1| > 0$$

Dann hat die RWA

$$y'' = f(x, y, y') \quad a \leq x \leq b$$

$$a_0 y(a) - a_1 y'(a) = \alpha$$

$$b_0 y(b) + b_1 y'(b) = \beta$$

genau eine Lösung.

Beweis: *siehe z.B. bei Keller.* □

Für lineares f :

$$f(x, y, y') = p(x)y' + q(x)y - r(x)$$

bedeuten die Voraussetzungen von Satz 2.1.1

$$p, q, r \in C[a, b] \quad \text{und} \quad q(x) > 0 \quad \text{für} \quad a \leq x \leq b.$$

Die **Anfangswertaufgabe**

$$\begin{aligned} y' &= F(x, y) \\ y(a) &= y_0 \end{aligned}$$

ist für bzgl. y global lipschitzstetiges F stets eindeutig lösbar. Dies legt zur Lösung der RWA

$$\begin{aligned} y' &= F(x, y) \\ R(y(a), y(b)) &= 0 \end{aligned}$$

folgende Vorgehensweise nahe: Man definiere für $s \in \mathbb{R}^n$ $y(\cdot; s)$ als Lösung der AWA

$$y' = F(x, y) \quad a \leq x \leq b, \quad y(a; s) \stackrel{\text{def}}{=} s.$$

Dann ist die Lösung der RWA gegeben durch die Lösung des **nichtlinearen Gleichungssystems**

$$G(s) = R(s, y(b; s)) = 0 \tag{2.3}$$

in Form der Funktion $y(\cdot; s^*)$, wobei s^* eine Lösung von (2.3) ist. Offenbar hat die RWA genauso viele verschiedene Lösungen, wie das nichtlineare Gleichungssystem (2.3) verschiedene Lösungen hat (immer globale Lipschitzstetigkeit vorausgesetzt). Wenn F und R stetig differenzierbare Funktionen sind, dann ist $y(b; s)$ stetig nach s differenzierbar und damit auch G . Zur Lösung von (2.3) bietet sich dann das Newton-Verfahren an (oder eine seiner effizienteren Modifikationen).

Beispiel 2.1.3.

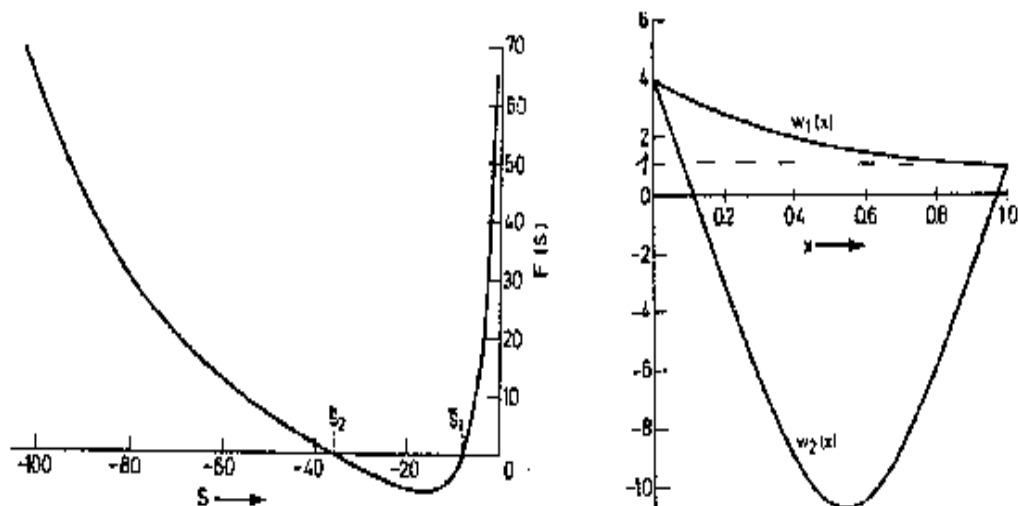
$$w'' = \frac{3}{2}w^2, \quad w(0; s) = 4. \quad w'(0; s) = s$$

Gesucht s mit $w(1; s) = 1$.

Es gibt zwei Lösungen: $s_1 = -8$, $s_2 = -35.8585487278\dots$

Den Verlauf von $F(s) = w(1; s) - 1$ und der beiden Lösungen der RWA zeigt

Abbildung 2.1.3.



Es ist

$$w(x; s_1) = \frac{4}{(1+x)^2}$$

während $w(x; s_2)$ wesentlich komplizierter durch höher transzendente Funktionen dargestellt werden kann.

Ein weiteres einfaches Beispiel für diese Problematik ist

Beispiel 2.1.4.

$$y'' + \exp(y+1) = 0, \quad x \in [0, 1], \quad y(0) = y(1) = 0$$

mit der Lösung

$$y(x) = -2 \ln \left(\frac{\cosh((x-1/2)\theta/2)}{\cosh(\theta/4)} \right)$$

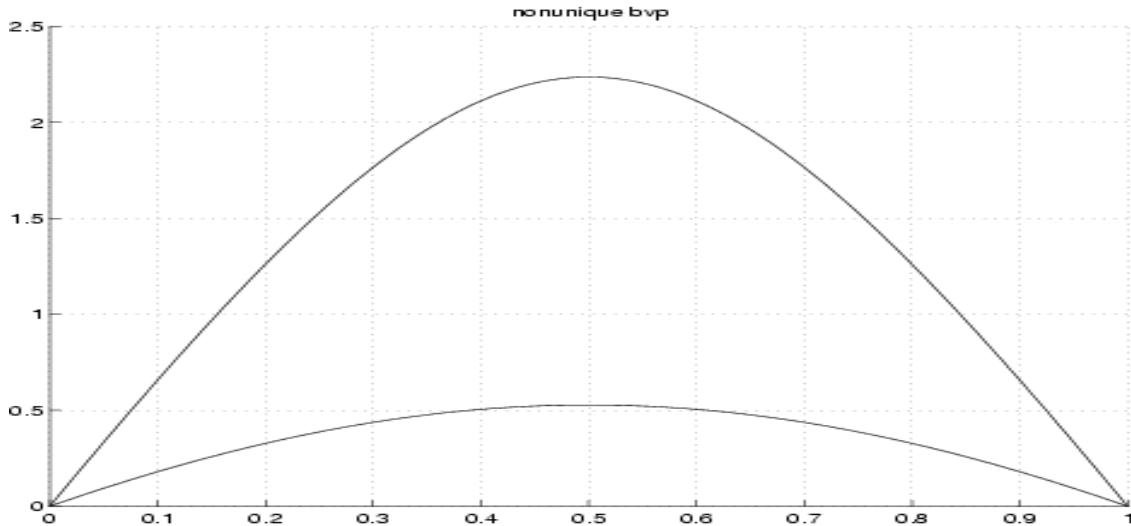
wo θ die beiden Lösungen von

$$\theta = \sqrt{2e} \cosh(\theta/4)$$

bezeichnet.

□

Abbildung 2.1.4.



Eine Reihe anderer Aufgaben kann man auf RWA zurückführen, so z.B.

Eigenwertprobleme gewöhnlicher Differentialgleichungen:

$$y' = f(x, y, \lambda) \quad a \leq x \leq b, \quad f : [a, b] \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$$

$$r(y(a), y(b), \lambda) = 0 \quad r : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^{n+1}$$

Gesucht sind Werte λ , für die Lösungen ungleich null existieren nebst den zugehörigen Lösungen y .

Beispiel 2.1.5.

$$-y'' = \lambda y, \quad y'(0) = 0, \quad y(0) = 1, \quad y(1) = 0,$$

Euler'scher Knickstab $\Phi = 0$, linearisiert, mit den Eigenwerten

$$\lambda \in \left\{ \left(\frac{(2k+1)\pi}{2} \right)^2 : k \in \mathbb{Z} \right\}.$$

und den Lösungen

$$y(x) = \cos(\sqrt{\lambda}x).$$

□

Mit der Einführung einer weiteren Variablen y_{n+1} und der Substitution

$$y_{n+1}(x) \stackrel{def}{=} \lambda \quad \bar{y} = (y, y_{n+1})$$

$$f_{n+1} \equiv 0 \quad \bar{f} = (f, f_{n+1})$$

$$\bar{r}(\underbrace{u_1, \dots, u_n}_{\hat{=} y(a)}, \underbrace{u_{n+1}}_{\hat{=} \lambda}, \underbrace{v_1, \dots, v_n}_{\hat{=} y(b)}, \underbrace{v_{n+1}}_{\hat{=} \lambda}) \stackrel{def}{=} r(u_1, \dots, u_n, v_1, \dots, v_n, v_{n+1})$$

entsteht eine RWA.

Randwertprobleme mit freiem Rand

Hier hat man neben der DGL

$$y' = f(x, y) \quad a \leq x \quad f: \mathcal{I} \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad a \in \mathcal{I}$$

$n + 1$ Randbedingungen

$$r(y(a), y(b)) = 0 \quad r: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^{n+1},$$

und neben der Lösung y ist noch die Randabszisse $b > a$ so zu bestimmen, daß die Randbedingung erfüllt ist.

Man substituiert $z_{n+1} \stackrel{\text{def}}{=} b - a$, d.h. $\dot{z}_{n+1} = 0$,

$$\begin{aligned} x - a &= \tau z_{n+1} & 0 \leq \tau \leq 1 \\ z_i(\tau) &= y_i(a + \tau z_{n+1}) & i = 1, \dots, n \\ \hat{z} &\stackrel{\text{def}}{=} (z_1, \dots, z_n)^T \end{aligned}$$

und erhält die RWA mit $z = (z_1, \dots, z_{n+1})$

$$\begin{aligned} \dot{z} &= \tilde{f}(\tau, z) & 0 \leq \tau \leq 1, \quad \tilde{f} = \begin{pmatrix} z_{n+1} f(a + \tau z_{n+1}, \hat{z}) \\ 0 \end{pmatrix} \\ r(\hat{z}(0), \hat{z}(1)) &= 0 \end{aligned}$$

□

Mehrpunktbedingungen

In den Anwendungen treten häufig auch Probleme auf, bei denen Bedingungen an die Lösung in Zwischenstellen im Intervall $[a, b]$ gestellt werden, also etwa

$$\begin{aligned} y' &= F(x, y), \quad F: [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n \\ R(y(x_0), y(x_1), \dots, y(x_m)) &= 0, \quad R: (\mathbb{R}^n)^{m+1} \rightarrow \mathbb{R}^n \end{aligned}$$

mit

$$a = x_0 < x_1 < \dots < x_m = b.$$

Diese Aufgabe kann man durch eine Substitution in eine Zweipunkttrandwertaufgabe umwandeln: Die Funktionen

$$s = \phi_i(x) = \frac{x - x_i}{x_{i+1} - x_i} \quad i = 0, \dots, m - 1$$

transformieren das i -te Teilintervall auf $[0, 1]$ und

$$z_i(s) \stackrel{\text{def}}{=} y((x_{i+1} - x_i)s + x_i), \quad i = 0, \dots, m - 1$$

stellt die Lösung y auf dem i -ten Teilintervall dar. Damit erhalten wir m (vektorwertige) Differentialgleichungen auf $[0, 1]$

$$z'_i = (x_{i+1} - x_i) F((x_{i+1} - x_i)s + x_i, z_i(s)), \quad i = 0, \dots, m - 1$$

mit den m (vektorwertigen) Randbedingungen

$$\begin{aligned} R(z_0(0), z_0(1), z_1(1), \dots, z_{m-1}(1)) &= 0, \\ z_i(0) &= z_{i-1}(1), \quad i = 1, \dots, m-1 \end{aligned}$$

die wieder die Standardform bilden für die Vektorfunktion

$$z(s) = (z_0(s), \dots, z_{m-1}(s))^T, \quad [0, 1] \rightarrow \mathbb{R}^{nm}.$$

Weiterführende Literatur zu diesem Kapitel:

1. Ascher, Uri M.; Mattheij, Robert M.M.; Russell, Robert D.: *Numerical solution of boundary value problems for ordinary differential equations* Classics in Applied Mathematics. 13. Philadelphia, PA: SIAM 1995
2. Bohl, E.: *Finite Modelle gewöhnlicher Randwertaufgaben* Teubner Studienbücher Mathematik 51 (1981)
3. Keller, H.B.: *Numerical methods for two point boundary-value problems* Blaisdell 1968
4. Quarteroni, A.; Sacco, R.; Saleri, F.: *Numerische Mathematik 2* Springer 2002. Kapitel 12.
5. Stoer, J., Bulirsch, R.: *Numerische Mathematik II*, Abschnitte 7.3 bis 7.6

2.2 Das Schießverfahren und das Mehrfachschießverfahren

Diese Verfahren knüpfen an die bereits im vorigen Abschnitt geschilderte Zurückführung der Lösung der RWA auf ein (nichtlineares) Gleichungssystem

$$r(s, y(b; s)) = 0$$

wo $y(x; s)$ die Lösung des Anfangswertproblems

$$y'(x; s) = f(x, y(x; s)), \quad y(a) = s$$

bezeichnet. Am einfachsten ist die Situation bei einer DGL 2. Ordnung mit separierten Dirichlet-Daten

$$y'' = f(x, y, y'), \quad y(a) = y_0, \quad y(b) = y_1$$

Hier braucht man nur $y'(a) = s$ zu nehmen und $s = s^*$ so zu bestimmen, daß $y(b; s^*) = y_1$ wird, wo $y(x; s)$ die DGL $y'' = f(x, y, y')$ erfüllt mit den Anfangsbedingungen $y(a) = y_0$, $y'(a) = s$.

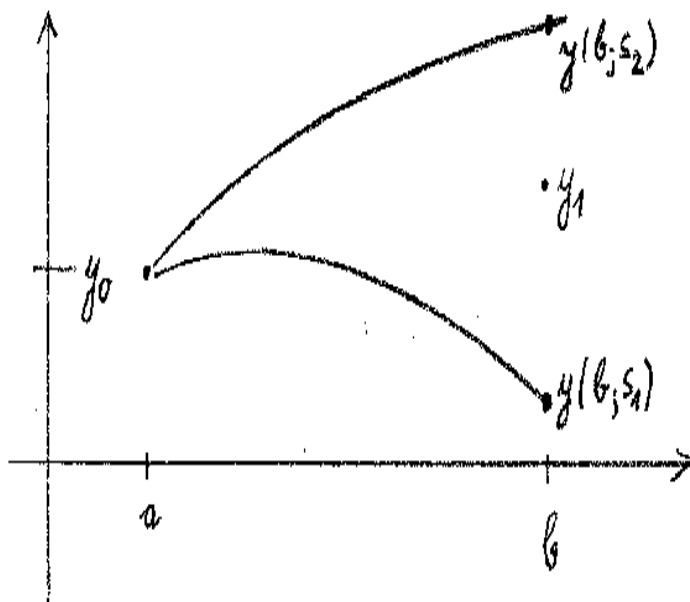


Abbildung 2.2.1.

Hat man Werte s_1, s_2 gefunden mit $y(b; s_1) < y_1 < y(b; s_2)$, dann kann man s^* z.B. durch Intervallhalbierung oder auch effizientere global konvergente Einschachtelungsmethoden wie das Brent-Decker-Verfahren finden. Der Vorteil dieser Vorgehensweise ist der, daß man dann nur Funktionswerte $y(b; s)$ (für jeden s -Wert also die Lösung einer AWA) zu berechnen braucht, also keine Ableitungen von $y(b; s)$ bezüglich s .

Falls f differenzierbar ist, dann ist auch $y(b; s)$ bezüglich s differenzierbar. Durch Differentiation der Gleichung

$$y''(x; s) = f(x, y(x; s), y'(x; s))$$

nach s erhalten wir nach der (erlaubten) Vertauschung der Differentiationsreihenfolge bzgl. x und s die Differentialgleichung

$$y''_s(x; s) = \partial_2 f(x, y(x; s), y'(x; s)) y_s(x; s) + \partial_3 f(x, y(x; s), y'(x; s)) y'_s(x; s)$$

(der Index s bezeichnet Differentiation bzgl. s) mit den Anfangsvorgaben

$$y_s(a) = 0, \quad y'_s(a) = 1$$

Die Ableitung $y_s(b; s)$ berechnet sich also als Lösung einer linearen AWA 2. Ordnung, wobei die Koeffizienten von der Lösung $y(x; s)$ abhängen. Bei der numerischen Lösung kann man diese AWA simultan mit der AWA für $y(b; s)$ behandeln, benötigt aber natürlich formelmäßig die partiellen Ableitungen $\partial_2 f = f_y$ und $\partial_3 f = f_{y'}$.

Bei einer linearen RWA liefert trivialerweise ein Schritt des Newton-Verfahrens die exakte Lösung s^* .

Beispiel 2.2.1. Gegeben sei das lineare Randwertproblem

$$y'' = -y \quad \text{mit} \quad y(0) = 0 \quad \text{und} \quad y(0.4) = \sin(0.4).$$

mit der Lösung $y(x) = \sin(x)$. Wir wollen eine Approximation der Lösung im Punkt $x = 0.2$ unter Verwendung des einfachen Schießverfahrens bestimmen. Zunächst überführen wir durch die Transformation

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} y \\ y' \end{pmatrix},$$

das RWP in ein zweidimensionales AWP und führen vier Schritte des Euler-Verfahrens ($h = 0.1$) durch, unter expliziter Verwendung des Schießparameters s . Dann schätzen wir mit dem Näherungswert im Punkt $x = 0.4$ den Parameter s . Das lineare RWP

$$y'' = -y \quad \text{mit} \quad y(0) = 0, \quad y(0.4) = \sin(0.4)$$

kann mit der Transformation

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} y \\ y' \end{pmatrix}$$

in das lineare AWP

$$u' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} u \quad \text{mit} \quad u(0) = \begin{pmatrix} 0 \\ s \end{pmatrix}$$

überführt werden.

Vier Schritte des Euler-Verfahrens ergeben

$$\begin{aligned} u_0 &= s \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ u_1 &= s \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix} + 0.1 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = s \begin{pmatrix} 0.1 \\ 1 \end{pmatrix} \\ u_2 &= s \left(\begin{pmatrix} 0.1 \\ 1 \end{pmatrix} + 0.1 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0.1 \\ 1 \end{pmatrix} \right) = s \begin{pmatrix} 0.2 \\ 0.99 \end{pmatrix} \\ u_3 &= s \left(\begin{pmatrix} 0.2 \\ 0.99 \end{pmatrix} + 0.1 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0.2 \\ 0.99 \end{pmatrix} \right) = s \begin{pmatrix} 0.299 \\ 0.97 \end{pmatrix} \\ u_4 &= s \left(\begin{pmatrix} 0.299 \\ 0.97 \end{pmatrix} + 0.1 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0.299 \\ 0.97 \end{pmatrix} \right) = s \begin{pmatrix} 0.396 \\ 0.9401 \end{pmatrix} \end{aligned}$$

Die Komponente $(u_4)_1 = 0.396s$ soll dem Funktionswert $y(0.4) = \sin(0.4)$ der Lösung entsprechen. Damit ergibt sich für s der Wert

$$s = \frac{\sin(0.4)}{0.396} = 0.9833797.$$

Die Näherung im Punkt $x = 0.2$ ist also $(u_2)_1 = 0.9833797 \cdot 0.2 = 0.196676$. Das exakte s ist in diesem Falle 1 und der exakte Funktionswert $\sin(0.2) = 0.198669$.

□

Bei einer RWA für ein allgemeines System von Differentialgleichungen

$$\begin{aligned} y' &= f(x, y) & a \leq x \leq b & & f : [a, b] \times \mathbb{R}^n &\rightarrow \mathbb{R}^n \\ r(y(a), y(b)) &= 0 & r : \mathbb{R}^n \times \mathbb{R}^n &\rightarrow \mathbb{R}^n \end{aligned}$$

scheidet die Anwendung der Intervallhalbierungsmethode zur Nullstellenbestimmung aus und man ist ganz auf das Newton–Verfahren und seine Varianten angewiesen. Die Anfangswertaufgabe für $y_s(x; s)$ lautet jetzt

$$y'_s(x; s) = f_y(x, y(x; s))y_s(x; s), \quad y_s(a; s) = I$$

und dies ist ein gekoppeltes System von n^2 linearen DGLen (mit variablen Koeffizienten). Aus Aufwandsgründen scheidet die numerische Behandlung dieses Systems praktisch immer aus. Stattdessen ist man darauf angewiesen, die Jacobi–Matrix des Systems

$$G(s) = r(s, y(b; s)) = 0$$

$$\mathcal{J}_G(s) = \partial_1 r(s, y(b; s)) + \partial_2 r(s, y(b; s))y_s(b; s)$$

durch Differenzenquotienten anzunähern:

$$\mathcal{J}_G(s) = \frac{1}{\Theta} \underbrace{(G(s + \Theta e_1) - G(s), G(s + \Theta e_2) - G(s), \dots, G(s + \Theta e_n) - G(s))}_{=: A(s, \Theta)} + \mathcal{O}(\Theta)$$

Die Berechnung dieser Näherung und damit die Durchführung eines Schrittes des diskretisierten Newton–Verfahrens für die Gleichung

$$G(s) = r(s, y(b; s)) = 0$$

erfordert also die Lösung von $n + 1$ AWA mit der rechten Seite f . Deshalb führt man das Verfahren in der Regel als ein vereinfachtes diskretisiertes Newton–Verfahren durch. Da der Fehler in den Differenzenquotienten die Größenordnung globaler Diskretisierungsfehler in $y(b; s)$ geteilt durch Θ_k aufweist, darf dieser globale Diskretisierungsfehler bei der Lösung der Anfangswertaufgaben die Größenordnung $\mathcal{O}(\Theta_k^2)$ nicht übersteigen. Eine untere sinnvolle Schranke für Θ_{\min} bei t -stelliger dezimaler Rechnung ist also $\|s\|10^{-t/2}$. (Dann müssen die AWA’s auf Maschinengenauigkeit gelöst werden).

Damit erscheint die Lösung auch komplizierter nichtlinearer Randwertaufgaben auf die Lösung von (nichtlinearen) AWA (was auf effiziente Weise möglich ist) und die Anwendung des Newton–Verfahrens zurückgeführt. Es tritt jedoch neben die grundsätzliche Problematik bei der Lösung nichtlinearer Gleichungssysteme eine weitere eigentümliche Schwierigkeit hinzu:

Während die in den Anwendungen auftretenden AWA in der Regel stabile Prozesse beschreiben (die Lösungsmannigfaltigkeit besteht aus abklingenden oder periodischen Funktionen), besitzen die Lösungsmannigfaltigkeiten der DGLen, die bei RWAen auftreten, oft instabile Lösungskomponenten (exponentielles Wachstum, bewegliche Singularitäten) und die stabile Lösung der RWA wird durch die Vorgabe der rechten Randbedingung ausgewählt. Bei der Behandlung als AWA führen geringste Abweichungen von der “richtigen” Vorgabe s^* dann auf instabile Lösungszweige (auch die lokalen Diskretisierungsfehler haben diese Wirkung), sodaß die numerische Lösung der AWA unmöglich wird.

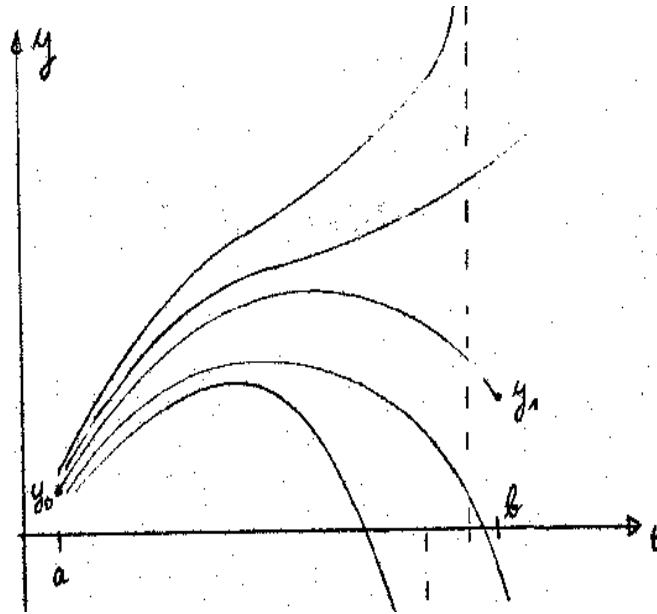


Abbildung 2.2.2.

Beispiel 2.2.2. Wir betrachten die DGL

$$y'' + (y')^2 = 0$$

mit den Randwerten

$$y(0) = 1 \quad \text{und} \quad y(1) = b$$

mit einem freien Parameter b . Wir bestimmen zunächst in Abhängigkeit von b die möglichen Lösungen. Dann untersuchen wir das einfache Schiessverfahren auf seine Durchführbarkeit. Mit der Substitution $u = y'$ ergibt sich :

$$u' + u^2 = 0$$

d.h. $u \equiv 0$ oder $\frac{du}{u^2} = -dx$, also

$u \equiv 0$ oder $u(x) = \frac{1}{x+c_1}$, also

$y(x) \equiv \text{const}$ oder $y(x) = \ln|x+c_1| + c_2$ für $0 \leq x \leq 1$.

(i) $y(x) \equiv \text{const}$: Nur für $b = 1$ gibt es in diesem Fall die Lösung $y(x) = 1$.

(ii) $y(x) = \ln|x+c_1| + c_2$. Aus den Randbedingungen folgt: $b = 1 + \ln|1+c_1| - \ln|c_1|$ und somit

$$c_1 = \frac{e}{\pm e^b - e}$$

$$c_2 = 1 - \ln\left|\frac{e}{\pm e^b - e}\right|$$

$$y(x) = \ln|\pm e^b x - ex + e|$$

$$s = y'(0) = u(0) = \frac{1}{c_1}$$

$$y(x; s) = \ln\left|x + \frac{1}{s}\right| + 1 - \ln\left|\frac{1}{s}\right|$$

Speziell erhalten wir hieraus:

$$y(1; -\frac{99}{100}) = 1 - \ln 100 = -3.6052$$

$$y(1; -\frac{999}{1000}) = 1 - \ln 1000 = -5.9078$$

Für $s = -\frac{100}{99}$ ergibt sich eine Singularität an der Stelle $t = \frac{99}{100}$. Die exakte Lösung $y(x; s)$ reagiert also sehr sensibel auf Änderungen des Parameters s , daher ist das einfache Schießverfahren nicht geeignet zur Lösung des RWP's. \square

Diesen Effekt kann man durch eine Unterteilung des Ausgangsintervalls $[a, b]$ abmildern. Sei $a = x_1 < \dots < x_m = b$ eine Zerlegung von $[a, b]$. Dann ist die RWA

$$\begin{aligned} y' &= f(x, y) & a \leq x \leq b \\ r(y(a), y(b)) &= 0 \end{aligned}$$

äquivalent zu folgendem System von Anfangswertaufgaben und (nicht-)linearen Gleichungen:

Mehrzielmethode multiple shooting

$$\left. \begin{aligned} y'_k(x; s_k) &= f(x, y_k(x; s_k)), & x_k \leq x \leq x_{k+1} \\ y_k(x_k) &= s_k \end{aligned} \right\} k = 1, \dots, m-1$$

$$\begin{pmatrix} y_1(x_2) - s_2 \\ y_2(x_3) - s_3 \\ \vdots \\ y_{m-1}(b) - s_m \\ r(s_1, s_m) \end{pmatrix} \equiv G(s_1, \dots, s_m) = 0$$

Die Aufgabe besteht also darin, die Lösung von $m-1$ AWA durch geeignete Vorgabe der Anfangswerte stetig aneinander anzusetzen und außerdem noch die ursprüngliche Randbedingung zu erfüllen. Man erhält somit ein im allgemeinen nichtlineares Gleichungssystem der Dimension nm . Die folgende Abbildung zeigt die Situation für die Startwerte.

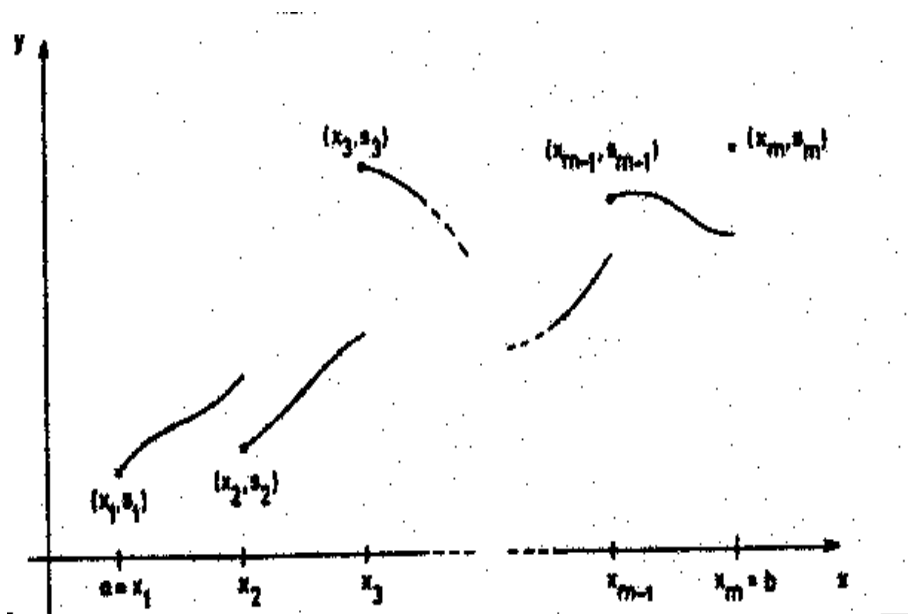


Abbildung 2.2.3. .

Dieses System kann man wieder mit dem Newton-Verfahren bzw. den oben besprochenen Modifikationen behandeln. Die spezielle Form von G erlaubt dabei erhebliche technische Vereinfachungen. Eine geeignete Einteilung des Intervalls $[a, b]$ kann man oft adaptiv finden (vgl. bei Stoer & Bulirsch).

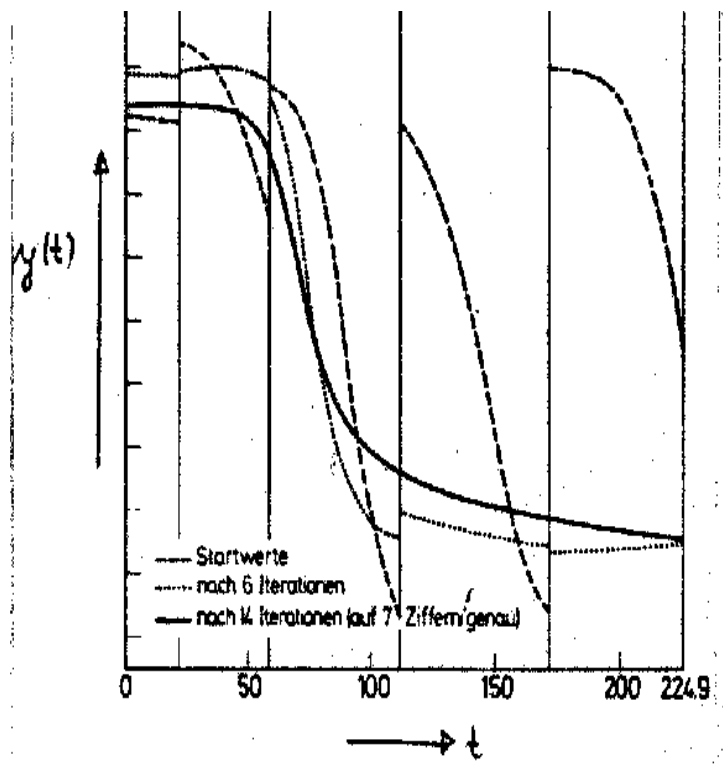


Abbildung 2.2.4.

Die Stärke der Mehrzielmethode besteht darin, daß sehr allgemeine und auch außerordentlich empfindliche Systeme damit erfolgreich behandelt werden können. Die Haupt-

nachteile sind in der nur lokalen Konvergenz der Nullstellenverfahren und der hohen Genauigkeit, mit der die AWA's selbst gelöst werden müssen (insbesondere wenn man die Jacobimatrix von G durch numerisches Differenzieren bestimmen will), zu sehen. Ausserdem kann man nicht erwarten, daß man ohne vernünftige Startwerte mit dem Newtonverfahren Erfolg haben wird.

2.3 Elementare Differenzenverfahren

Die in der Praxis auftretenden Randwertprobleme haben oft eine sehr spezielle Struktur (und besitzen eine global eindeutige Lösung). Außerdem ist die Lösung in der Regel mit geringer Genauigkeit gesucht (2 bis 3 Stellen). Dann ist die Anwendung der Mehrzielmethode nicht sinnvoll. Vielmehr gelangt man oftmals schon mit einfachsten Differenzenformeln zum Ziel.

Beispiel 2.3.1. *Wir betrachten die lineare RWA*

$$y'' - y + 1 = 0, \quad y(0) = y(10) = 0$$

mit der Lösung $y(x) = (e^{10} - e^x)(1 - e^{-x})/(e^{10} + 1)$.

(Die Lösungsmannigfaltigkeit der Anfangswertaufgabe besteht aus den Linearkombinationen von e^x und e^{-x} .) Wir unterteilen das Intervall $[0, 10]$ in N Teilintervalle der Länge $h = \frac{10}{N}$ und ersetzen an den inneren Teilpunkten $x_i = ih$, $1 \leq i \leq N - 1$ die zweite Ableitung durch den symmetrischen Differenzenquotienten 2. Ordnung.

$$y''(x_j) = \frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1}))}{h^2} + \mathcal{O}(h^2)$$

und lösen das entstehende lineare Gleichungssystem für die Näherungswerte

$$\begin{aligned} \eta_j &\approx y(x_j) \\ -\frac{1}{h^2}\eta_{j-1} + \left(\frac{2}{h^2} + 1\right)\eta_j - \frac{1}{h^2}\eta_{j+1} - 1 &= 0 & j = 1, \dots, N-1 \\ \eta_0 = \eta_N &= 0 \\ N = 5 &\Rightarrow h = 2 \end{aligned}$$

1.5	-0.25	0	0	= 1
-0.25	1.5	-0.25	0	= 1
0	-0.25	1.5	-0.25	= 1
0	0	-0.25	1.5	= 1
0.827586207	0.965517241	0.965517241	0.827586207	
= η_1	= η_2	= η_3		
$\approx y(2)$	$\approx y(4)$	$\approx y(6)$	$\approx y(8)$	

$N = 10 \Rightarrow h = 1$

3	-1				...	= 1
-1	3	-1			...	= 1
	-1	3	-1		...	= 1
		-1	3	-1	...	= 1
			-1	3	...	= 1
					...	⋮
					...	⋮

0.617886179	0.853658537	0.943089431	0.975609756	0.983739837	
$= \eta_1$	$= \eta_2$	$= \eta_3$	$= \eta_4$	$= \eta_5$	$\eta_{10-j} = \eta_j$
$\approx y(1)$	$\approx y(2)$	$\approx y(3)$	$\approx y(4)$		

Wir werden in Satz 2.3.3 zeigen, daß

$$\eta(x; h) = y(x) + e(x)h^2 + \mathcal{O}(h^4)$$

wird, so daß Richardsonextrapolation anwendbar ist:

$(4\eta(x; h) - \eta(x; 2h))/3$ eine $\mathcal{O}(h^4)$ -Näherung für $y(x)$ ist, z.B.

$(4 \cdot 0.975609756 - 0.96551724)/3 = 0.97897392$, $y(4) = 0.97920655$

mit einem Fehler von nur $3 \cdot 10^{-4}$!

□

Wir betrachten jetzt die spezielle Randwertaufgabe mit homogenen Dirichletdaten

$$\begin{aligned} y'' &= f(x, y, y') \quad a \leq x \leq b \quad f : [a, b] \times \mathbb{R}^2 \rightarrow \mathbb{R} \\ y(a) &= 0, \quad y(b) = 0 \end{aligned} \tag{2.4}$$

Bemerkung 2.3.1. Falls statt (2.4) allgemeine lineare Randbedingungen vorgeschrieben sind und eine lineare Funktion $l(x)$ existiert mit

$$r(l(a), l(b)) = 0,$$

dann kann man durch den Ansatz $u(x) := y(x) - l(x)$ den Fall der homogenen Randbedingungen herstellen. Dies ist insbesondere bei separierten Dirichlet - Randbedingungen $y(a) = y_0$, $y(b) = y_1$ mit

$$l(x) = y_0 + \frac{x - a}{b - a}(y_1 - y_0)$$

der Fall.

□

Wir diskretisieren das Problem nun wie in Beispiel 2.3.1. Für die erste Ableitung verwenden wir dabei den symmetrischen Differenzenquotienten 1. Ordnung:

$$y'(x_j) = \frac{y(x_{j+1}) - y(x_{j-1}))}{2h} + \mathcal{O}(h^2)$$

Wir erhalten mit $x_j = a + jh$, $h = (b - a)/N$ das Gleichungssystem für die Näherungswerte $\eta_j = \eta(x_j; h)$:

$$-\eta_{j-1} + 2\eta_j - \eta_{j+1} + h^2 f(x_j, \eta_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h}) = 0$$

$$j = 1, \dots, N - 1, \quad \eta_0 = \eta_N = 0.$$

Dieses Gleichungssystem ist nichtlinear, wenn f nichtlinear von y oder y' abhängt. Wir wollen zunächst zeigen, daß dieses Gleichungssystem für hinreichend kleines h stets eindeutig lösbar ist, wobei wir die Voraussetzungen von Satz 2.1.1 noch etwas abschwächen können. Dies geschieht in Satz 2.3.1. Sodann werden wir beweisen, daß

$$\eta(x_j; h) - y(x_j) = \mathcal{O}(h^2)$$

und, wenn $f \in C^4([a, b] \times \mathbb{R}^2)$, daß sogar

$$\eta(x_j; h) = y(x_j) + e(x_j)h^2 + \mathcal{O}(h^4),$$

so daß Richardson–Extrapolation zur Genauigkeitssteigerung verwendet werden kann. Sei also zunächst

$$F : \mathbb{R}^M \rightarrow \mathbb{R}^M \quad (M := N - 1)$$

definiert durch

$$F_j(\eta) = -\eta_{j-1} + 2\eta_j - \eta_{j+1} + h^2 f(x_j, \eta_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h}) \quad (2.5)$$

$$j = 1, \dots, M, \quad \eta_0 = \eta_{M+1} = 0.$$

Für die Jacobimatrix von F ergibt sich eine Dreibandmatrix:

$$\mathcal{J}_F(\eta) = A + h^2 \begin{pmatrix} \ddots & & & & & & \\ & \ddots & & & & & \\ 0 \cdots 0, & -\frac{1}{2h} \partial_3 f(w_j), & \partial_2 f(w_j), & +\frac{1}{2h} \partial_3 f(w_j), & 0 \cdots 0 & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \end{pmatrix}$$

mit $w_j := (x_j, \eta_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h})$

$$A = \begin{pmatrix} \ddots & & & & & & \\ & \ddots & & & & & \\ 0 \cdots 0 & -1 & 2 & -1 & 0 \cdots 0 & & \\ & & \ddots & & & & \\ & & & \ddots & & & \end{pmatrix}$$

Der folgende Satz zeigt, daß unter den Voraussetzungen von Satz 2.1.1 (sogar in etwas abgeschwächter Form) die Inverse dieser Matrix gleichmässig in η beschränkt ist.

Satz 2.3.1. Es gelte $f \in C^2(\mathcal{S})$ mit

$$\mathcal{S} \stackrel{\text{def}}{=} [a, b] \times \mathbb{R}^2.$$

Ferner gelte mit geeigneten Konstanten $C > 0$, $B > 0$ und γ

$$\gamma_0 \leq \gamma \leq \partial_2 f(w) \leq C, \quad |\partial_3 f(w)| \leq B$$

wo

$$\gamma_0 = -\frac{B+1}{2(\exp((B+1)(b-a)) - 1)}.$$

Sei

$$h = \frac{b-a}{N} \quad \text{und} \quad 0 < h \leq \frac{2}{(B+4)^2}.$$

Dann ist $\mathcal{J}_F(\eta) = (g_{ij}(\eta))$ für alle η definiert, eine M-Matrix und es gilt

$$\|(\mathcal{J}_F(\eta))^{-1}\|_\infty \leq \frac{2}{h^2 |\gamma_0|}.$$

<<

Beweis: Wir erläutern zunächst die wesentlichen Schritte des nachfolgenden Beweises. Für eine M-Matrix B gilt

$$\|B^{-1}\|_\infty = \|B^{-1}e\|_\infty \quad \text{mit} \quad e = (1, \dots, 1)^T.$$

Zunächst wird der Fall

$$\partial_2 f(x, y, y') = 0$$

behandelt. Dann wird der Fall

$$\partial_2 f(x, y, y') = \gamma_0$$

behandelt. Der Fall

$$\partial_2 f(x, y, y') \geq \gamma_0$$

ist dann bereits erledigt, da für eine M-Matrix gilt: Ist A M-Matrix und D diagonal nichtnegativ, dann ist auch $A + D$ eine M-Matrix und

$$(A + D)^{-1} \leq A^{-1} \quad \text{komponentenweise}.$$

(Der Beweis dieser Behauptung findet sich im folgenden Lemma) Wir konstruieren nun ein lineares Randwertproblem mit analytisch bekannter Lösung y , bei dem die beschriebene Diskretisierung als Jacobimatrix genau die in Frage stehende Matrix $\mathcal{J}_F(\eta)$ ist. Setzt man statt η , dessen Abschätzung ja noch nicht bekannt ist, die wahre Lösung y in das System ein, so erhält man ein Residuum r , das man mit Hilfe der Taylorentwicklung und der bekannten Ableitungen von y abschätzen kann. Etwa

$$\|r\|_\infty \leq h^4 \beta \quad \text{oder} \quad r \geq -h^4 \beta e \quad \text{komponentenweise}$$

Die Inhomogenität des Problems ist so gewählt, daß sie sich ebenfalls durch ein positives Vielfaches von $h^2 e$ nach unten abschätzen lässt. Für genügend kleines h bekommt man dann

$$\mathcal{J}_F(\eta)y \geq h^2 \alpha e$$

und damit

$$\frac{\|y\|_\infty}{h^2 \alpha} \geq \|(\mathcal{J}_F(\eta))^{-1} e\|_\infty = \|(\mathcal{J}_F(\eta))^{-1}\|_\infty .$$

Nun zum eigentlichen Beweis. Die Außerdiagonalelemente von $\mathcal{J}_F(\eta)$ haben die Form $g_{j+1,j} = -(1 + \frac{h}{2} \partial_3 f(w_{j+1}))$ bzw. $g_{j,j+1} = -(1 - \frac{h}{2} \partial_3 f(w_j))$. Falls nun $0 < h \leq \frac{2}{(B+4)^2}$, dann ist wegen $|\partial_3 f(w)| \leq B$

$$-\frac{B}{(B+4)^2} \leq \frac{h}{2} \partial_3 f(w_j) \leq \frac{B}{(B+4)^2}$$

und somit für alle j

$$\begin{aligned} g_{j,j-1} \leq 0, \quad g_{j,j+1} \leq 0, \quad \frac{3}{4} \leq |g_{j,j-1}|, |g_{j,j+1}| \leq \frac{5}{4} \\ |g_{j,j-1}| + |g_{j,j+1}| = 2. \end{aligned}$$

Andererseits gilt für die Diagonalelemente

$$g_{jj} = 2 + h^2 \partial_2 f(w_j) \geq 2 + h^2 \gamma_0.$$

Wir betrachten zunächst den Fall $\partial_2 f(w_j) = 0 \quad (\forall j)$.

$\mathcal{J}_F(\eta)$ ist dann also eine Tridiagonalmatrix mit positiver Diagonale und strikt negativen Nebendiagonalelementen. Ausserdem ist sie schwach diagonaldominant und in mindestens einer Zeile (hier in zwei, der ersten und der letzten) strikt diagonaldominant. Eine solche Matrix bezeichnet man als eine irreduzibel diagonaldominante L -Matrix, und eine solche Matrix ist eine M -Matrix (vergl. Num. Math. II, Kapitel 6). Insbesondere ist also mit

$$E(\eta) := \frac{1}{2} h(0 \cdots 0, -\partial_3 f(w_j), 0, \partial_3 f(w_j), 0 \cdots 0)$$

$A + E(\eta)$ eine M -Matrix. Wir schätzen zunächst $(A + E(\eta))^{-1}$ ab. Wir betrachten dazu ein lineares, analytisch lösbares Problem, das nach Diskretisierung die gleiche Jacobi-Matrix wie die vorliegende besitzt. Sei dazu $p : [0, b-a] \rightarrow \mathbb{R}$ eine beliebige beschränkte Funktion mit

$$|p(x)| \leq B \quad \forall x \in [0, b-a] .$$

Das lineare Randwertproblem

$$\begin{aligned} -y'' + p(x)y' &= (B+1)(B+1-p(x)) \exp((B+1)x) \\ y(0) &= \exp((B+1)(b-a)) - 1 > 0, \quad y(b-a) = 0 \end{aligned}$$

hat die Lösung

$$y(x) = \exp((B+1)(b-a)) - \exp((B+1)x) \quad (\in C^\infty \text{ unabhängig von } p(x))$$

Nun ist mit $x_j = jh$

$$\begin{aligned} -y''(x_j) &= \frac{-y(x_{j+1}) + 2y(x_j) - y(x_{j-1}))}{h^2} + \frac{1}{12}h^2y^{(4)}(\xi_j), \\ &\quad \xi_j \in [x_{j-1}, x_{j+1}] \\ p(x_j)y'(x_j) &= p(x_j) \cdot \frac{y(x_{j+1}) - y(x_{j-1}))}{2h} - p(x_j)\frac{h^2}{6}y^{(3)}(\tilde{\xi}_j), \\ &\quad \tilde{\xi}_j \in [x_{j-1}, x_{j+1}]. \end{aligned}$$

Man beachte, daß für $j = 1$ auf der rechten Seite ein Term der Form $(h/2)y_0p(x_1) + y_0 > 0$ auftritt wegen $y_0 > 0$. Also gilt

$$\begin{aligned} (A + \frac{1}{2}h(0, \dots, 0, \underbrace{-p(x_j)}_{=-\partial_3 f(w_j)}, 0, \underbrace{p(x_j)}_{=\partial_3 f(w_j)}, 0, \dots, 0)) \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix} &\geq \quad (2.6) \\ = (h^2(B+1)(B+1-p(x_j)) \exp((B+1)x_j) - \frac{1}{12}h^4y^{(4)}(\xi_j) + p(x_j)\frac{h^4}{6}y^{(3)}(\tilde{\xi}_j))_j \end{aligned}$$

Nun ist aber

$$\begin{aligned} y^{(4)}(x) &= -(B+1)^4 \exp((B+1)x) < 0 \\ y^{(3)}(x) &= -(B+1)^3 \exp((B+1)x) < 0 \end{aligned}$$

und somit die rechte Seite in (2.6) komponentenweise \geq

$$h^2(B+1) \exp((B+1)x_j) (1 - B\frac{h^2}{6}(B+1)^2 \exp((B+1)h))$$

Wegen

$$h \leq \frac{2}{(B+4)^2}$$

erhalten wir

$$\begin{aligned} h(B+1) &\leq \frac{1}{6} \\ \frac{h^2}{6}B(B+1)^2 &\leq \frac{1}{6} \end{aligned}$$

Also ist die rechte Seite in (2.6) komponentenweise \geq

$$h^2(B+1)(1 - \frac{1}{6} \exp(1/6)) \geq \frac{3}{4}(B+1)h^2.$$

Dies ergibt nach Multiplikation mit der positiven Matrix $(A + E(\eta))^{-1}$

$$\begin{aligned} \frac{3}{4}(B+1)h^2 \|(A + E(\eta))^{-1}\|_\infty &\leq \max_{1 \leq j \leq M} |y(x_j)| \\ &\leq \exp((B+1)(b-a)) - 1 \end{aligned}$$

Also

$$\|(A + E(\eta))^{-1}\|_\infty \leq \frac{4}{3}h^{-2} \cdot \frac{\exp((B+1)(b-a)) - 1}{B+1} \stackrel{def}{=} \frac{1}{h^2\kappa}.$$

Wir betrachten nun den Fall, daß γ auch negativ sein darf. Wir betrachten nun zunächst die Matrix

$$A + E(\eta) + \gamma_0 h^2 I = (A + E(\eta))(I - |\gamma_0| h^2 (A + E(\eta))^{-1})$$

Wegen

$$\|(A + E(\eta))^{-1}\|_\infty |\gamma_0| h^2 \leq \frac{1}{h^2 \kappa} \cdot \frac{2}{3} \kappa h^2 = \frac{2}{3}$$

ist $\| |\gamma_0| h^2 (A + E(\eta))^{-1} \| \leq \frac{2}{3}$ und daher $A + E(\eta) + \gamma_0 h^2 I$ invertierbar. (Zur Erinnerung: Ist $\|H\| < 1$ für eine einer Vektornorm zugeordnete Matrixnorm, dann ist $I + H$ invertierbar und $(I + H)^{-1}$ in eine Potenzreihe entwickelbar (Neumannreihe)). Also

$$(A + E(\eta) + \gamma_0 h^2 I)^{-1} = \sum_{\nu=0}^{\infty} (|\gamma_0| h^2)^\nu ((A + E(\eta))^{-1})^{\nu+1},$$

d.h. $A + E(\eta) + \gamma_0 h^2 I$ ist eine M -Matrix, denn für die Außerdiagonalelemente gilt die Abschätzung $g_{j,j+1}, g_{j+1,j} \leq 0$. Da

$$\mathcal{J}_F(\eta) = A + E(\eta) + \gamma_0 h^2 I + \text{diag}(\partial_2 f(w_j) - \gamma_0) h^2$$

und $\partial_2 f(w_j) - \gamma_0 \geq 0$, gilt wie oben benutzt

$$\begin{aligned} \|\mathcal{J}_F(\eta)^{-1}\|_\infty &\leq \|(A + E(\eta) + \gamma_0 h^2 I)^{-1}\|_\infty \\ &\leq \frac{1}{h^2 \kappa} \cdot \frac{1}{1 - |\gamma_0|/\kappa} \leq \frac{2}{|\gamma_0| h^2}. \end{aligned}$$

Der allgemeine Fall

$$\partial_2 f(w) \geq \gamma_0$$

folgt nun mit dem folgenden Lemma. □

>>

Lemma 2.3.1. *Es sei A eine M -Matrix und D eine nichtnegative Diagonalmatrix. Dann ist auch $A + D$ eine M -Matrix und es gilt*

$$(A + D)^{-1} \leq A^{-1} \quad \text{komponentenweise.}$$

<<

Beweis: Der Beweis zerfällt in zwei Teile. Zunächst zeigen wir

$$|C| \leq B \Rightarrow \varrho(C) \leq \varrho(B).$$

Sei

$$\sigma = \varrho(B) \quad \text{und} \quad \varepsilon > 0 \quad \text{beliebig.}$$

Man setze

$$B_1 = (\sigma + \varepsilon)^{-1} B \quad \text{und} \quad C_1 = (\sigma + \varepsilon)^{-1} C.$$

Dann ist

$$|C_1| \leq B_1 \text{ und } \varrho(B_1) < 1 .$$

Aber

$$|C_1^k| \leq |C_1|^k \leq B_1^k \rightarrow O \text{ für } k \rightarrow \infty$$

und daher auch

$$\varrho(C_1) < 1 \Rightarrow \varrho(C) < \sigma + \varepsilon$$

und mit $\varepsilon \rightarrow 0$ folgt die Teilbehauptung. Diese impliziert auch noch

$$\varrho(C) \leq \varrho(|C|) .$$

Sei nun

$$A = -L_A + D_A - U_A$$

die übliche Zerlegung von A in striktes unteres Dreieck, Diagonale und striktes oberes Dreieck. Dann ist

$$O \leq (D + D_A)^{-1}(L_A + U_A) \leq D_A^{-1}(L_A + U_A)$$

da A eine M -, also auch eine L -Matrix ist. Für eine M -Matrix ist aber

$$\varrho(D_A^{-1}(L_A + U_A)) < 1$$

also ist nach dem ersten Teil des Beweises auch

$$\varrho(D + D_A)^{-1}(L_A + U_A) < 1$$

und da $D + A$ auch L -Matrix ist folgt, daß $D + A$ eine M -Matrix ist. Sei nun

$$(A + D)^{-1}e_i = z_i (\geq 0) .$$

Dann ist

$$e_i = (A + D)z_i \geq Az_i$$

und daher

$$A^{-1}e_i \geq A^{-1}Az_i = z_i .$$

□

>>

Wir erinnern nun an den Satz von Hadamard:

Sei $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $F \in C^1(\mathbb{R}^n)$, $\mathcal{J}_F(x)$ invertierbar für alle $x \in \mathbb{R}^n$ und $\|\mathcal{J}_F(x)^{-1}\| \leq \gamma \quad \forall x \in \mathbb{R}^n$ mit einer geeigneten Konstanten γ . Dann ist F ein Diffeomorphismus von \mathbb{R}^n auf \mathbb{R}^n .

Dies angewendet mit dem vorausgegangenen Satz ergibt: Die Gleichung $F(\eta) = 0$ besitzt genau eine Lösung.

Bemerkung 2.3.2. Man kann zeigen, daß auch für negative $\gamma \geq \gamma_0$ auch die **RWA** noch eindeutig lösbar ist (nicht nur das diskretisierte System) □

Bemerkung 2.3.3. Die Schrittweitenrestriktion an h kann im Falle $\gamma \geq 0$ und f linear auf $h < 2/B$ abgeschwächt werden. Eine Schrittweitenrestriktion dieser Art ist aber keineswegs akademisch, wie man an den schlechten Ergebnissen des folgenden Beispiels erkennt:

$$-y'' = 60 - 60y' \quad 0 \leq x \leq 1, \quad x(0) = x(1) = 0 \quad (B = 60, \quad \gamma = C = 0)$$

$$\text{Lösung } y(x) = x - \frac{e^{60x} - 1}{e^{60} - 1}, \quad h = 0.1$$

x	0.0	0.1	0.2	0.3	0.4	0.5
$y(x)$	0.0	0.1	0.2	0.3	0.3999	0.4999
$\eta(x; h)$	0.0	0.1029	0.1970	0.3087	0.3853	0.5322
x	0.6	0.7	0.8	0.9	1.0	
$y(x)$	0.5999	0.6999	0.7999	0.8975	0.0	
$\eta(x; h)$	0.5384	0.8260	0.5507	1.4014	0.0	

Die diskrete Lösung gibt hier den qualitativen Verlauf der wahren Lösung nicht wieder und zeigt eine unsinnige Oszillation. Dies ist in diesem Zusammenhang eine typische Erscheinung bei zu groß gewähltem h . Bei $h = 1/30$ ergibt sich dagegen

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
$\eta(x; h)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
x	0.933333		0.966667							
$y(x)$	0.915018		0.831331							
$\eta(x; h)$	0.933333		0.966667							

und erst mit $h = 1/100$ ist auch die Genauigkeit am rechten Intervallende einigermaßen akzeptabel.

x	0.95	0.96	0.97	0.98	0.99
$y(x)$	0.900213	0.869282	0.804701	0.678806	0.441188
$\eta(x; h)$	0.904734	0.875934	0.813878	0.690059	0.451539

□

Probleme dieses Typs bezeichnet man als "konvektionsdominiert". Es gibt abgeänderte Diskretisierungsschemata für solche Gleichungen, die eine solch strenge Schrittweitenbeschränkung nicht erfordern, s.h.

Als Folgerung aus Satz 2.3.1 erhalten wir, daß das gedämpfte Newtonverfahren zur Lösung von $F(\eta) = 0$, also das durch

$$\begin{aligned} \mathcal{J}_F(\eta^{[i]})d_i &= -F(\eta^{[i]}) & \eta^{[i]} &:= (\eta_1^{[i]}, \dots, \eta_M^{[i]})^T \\ \sigma_i &:= \max\{2^{-j} : \|J^{-1}F(\eta^{[i]} + 2^{-j}d_i)\|_2^2 \leq (1 - 2^{-j-2})\|J^{-1}F(\eta^{[i]})\|_2^2\} \\ \eta^{[i+1]} &= \eta^{[i]} + \sigma_i d_i \end{aligned}$$

(mit einer festen regulären Matrix J) definierte Verfahren, **global konvergiert**. In der Praxis ist es oft günstiger, J nicht fest zu wählen, sondern $J = \mathcal{J}_F(\eta^{[i]})$ zu setzen.

Bisher haben wir nur die Lösbarkeit der diskretisierten Gleichungen geklärt. Wir wenden uns der Frage nach der Konvergenz der Differenzenapproximationen zu.

Satz 2.3.2. *Es gelte $f \in C^2(\mathcal{S})$ mit $\mathcal{S} = [a, b] \times \mathbb{R}^2$, $\gamma \leq \partial_2 f(w) \leq C$, $|\partial_3 f(w)| \leq B \forall w \in \mathcal{S}$ und*

$$\gamma \geq -\frac{B+1}{2(\exp((B+1)(b-a)) - 1)}$$

Dann gilt für die eindeutig bestimmte Lösung $y \in C^4[a, b]$ von $y'' = f(x, y, y')$, $y(a) = y(b) = 0$:

$$|y(x_j) - \eta(x_j; h)| \leq Kh^2 \quad j = 1, \dots, M$$

mit einer geeigneten Konstanten K . Falls sogar $f \in C^4([a, b] \times \mathbb{R}^2)$, dann gilt mit einer Funktion $e \in C^2[a, b]$:

$$\eta(x_j; h) = y(x_j) + e(x_j)h^2 + \mathcal{O}(h^4).$$

e löst die lineare RWA

$$\begin{aligned} e''(x) &= \partial_3 f(x, y(x), y'(x))(e'(x) - \frac{1}{6}y^{(3)}(x)) + \partial_2 f(x, y(x), y'(x))e(x) + \frac{1}{12}y^{(4)}(x) \\ e(a) &= e(b) = 0. \end{aligned}$$

Beweis: Es gilt wegen der Taylorformel

$$\begin{aligned} -y_{j-1} + 2y_j - y_{j+1} &= -h^2 y''(x_j) - \frac{h^4}{12} y^{(4)}(\xi_j) \\ &= -h^2 f(x_j, y_j, y'(x_j)) - \frac{h^4}{12} y^{(4)}(\xi_j) \\ &= -h^2 f(x_j, y_j, \frac{y(x_{j+1}) - y(x_{j-1}))}{2h}) - \frac{h^2}{6} y^{(3)}(\tilde{\xi}_j) \\ &\quad - \frac{h^4}{12} y^{(4)}(\xi_j) \\ &= -h^2 f(x_j, y_j, \frac{y(x_{j+1}) - y(x_{j-1}))}{2h}) \\ &\quad + \frac{h^4}{6} \partial_3 f(x_j, y_j, \tilde{y}_j^{(1)}) y^{(3)}(\tilde{\xi}_j) - \frac{h^4}{12} y^{(4)}(\xi_j) \\ -\eta_{j-1} + 2\eta_j - \eta_{j+1} &= -h^2 f(x_j, \eta_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h}) \end{aligned}$$

und mit

$$\varepsilon_j := \eta_j - y_j$$

nach Subtraktion dieser Gleichungen somit

$$\begin{aligned}
-\varepsilon_{j-1} + 2\varepsilon_j - \varepsilon_{j+1} &= -h^2 \partial_2 f(x_j, \eta_j + \Theta_j \varepsilon_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h} + \Theta_j \frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h}) \varepsilon_j \\
&\quad - h^2 \partial_3 f(x_j, \eta_j + \Theta_j \varepsilon_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h} + \Theta_j \frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h}) \frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} \\
&\quad - \frac{h^4}{12} (2\partial_3 f(x_j, y_j, \tilde{y}_j^{(1)}) y^{(3)}(\tilde{\xi}_j) - y^{(4)}(\xi_j))
\end{aligned}$$

d.h. mit $\varepsilon := (\varepsilon_1, \dots, \varepsilon_M)^T$

$$G(\eta, \varepsilon)\varepsilon = -\frac{h^4}{12} (2\partial_3 f(x_j, y_j, \tilde{y}_j^{(1)}) y^{(3)}(\tilde{\xi}_j) - y^{(4)}(\xi_j))_{j=1}^M.$$

Dabei hat die Matrix G die gleichen Eigenschaften wie die Matrix $\mathcal{J}_F(\eta)$, die im Beweis von Satz 2.3.1 diskutiert wurde, d.h.

$$\|G^{-1}(\eta, \varepsilon)\|_\infty \leq \frac{3}{h^2 \kappa}, \quad \text{so da\ss} \quad \|\varepsilon\|_\infty \leq \frac{h^2}{4\kappa} (2BM_3 + M_4) \text{ mit}$$

$$M_j := \max\{|y^{(j)}(x)| : a \leq t \leq b\}.$$

Im Folgenden wird nun die asymptotische Entwicklung des globalen Fehlers bewiesen. Dazu benutzen wir eine verfeinerte Abschätzung des Fehlers und eine Variante des Banachschen Fixpunktsatzes.

<<

Sei nun $f \in C^4([a, b] \times \mathbb{R}^2)$. Dann wird

$$\begin{aligned}
-y_{j-1} + 2y_j - y_{j+1} &= -h^2 y_j'' - \frac{h^4}{12} y_j^{(4)} - \frac{h^6}{360} y^{(6)}(\xi_j) \\
\frac{y_{j+1} - y_{j-1}}{2h} &= y_j' + \frac{h^2}{6} y_j^{(3)} + \frac{h^4}{120} y^{(5)}(\tilde{\xi}_j)
\end{aligned}$$

Wir setzen zur Abkürzung

$$w_j = (x_j, y_j, y_j')$$

Taylorentwicklung ergibt:

$$\begin{aligned}
f(x_j, \eta_j, \frac{\eta_{j+1} - \eta_{j-1}}{2h}) &= f(w_j) + \partial_3 f(w_j) (\frac{\eta_{j+1} - \eta_{j-1}}{2h} - y_j') + \partial_2 f(w_j) \varepsilon_j \\
&\quad + \frac{1}{2} \partial_2^2 f(w_j) \varepsilon_j^2 + \partial_2 \partial_3 f(w_j) \varepsilon_j (\frac{\eta_{j+1} - \eta_{j-1}}{2h} - y_j') \\
&\quad + \frac{1}{2} \partial_3^2 f(w_j) (\frac{\eta_{j+1} - \eta_{j-1}}{2h} - y_j')^2 + \mathcal{O}(h^3)
\end{aligned}$$

(wegen $\frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} = \mathcal{O}(h)$).

$$\begin{aligned}
-y_{j-1} + 2y_j - y_{j+1} &= -h^2 f(w_j) - \frac{h^4}{12} y_j^{(4)} + \mathcal{O}(h^6) \\
-\eta_{j-1} + 2\eta_j - \eta_{j+1} &= -h^2 f(w_j) - h^2 \partial_3 f(w_j) \left(\frac{\eta_{j+1} - \eta_{j-1}}{2h} - y'_j \right) + h^2 \partial_2 f(w_j) \varepsilon_j \\
&\quad - \frac{h^2}{2} \partial_3^2 f(w_j) \left(\frac{\eta_{j+1} - \eta_{j-1}}{2h} - y'_j \right)^2 + \mathcal{O}(h^5) \\
-\varepsilon_{j-1} + 2\varepsilon_j - \varepsilon_{j+1} &= -\frac{h^4}{12} y_j^{(4)} + h^2 \partial_3 f(w_j) \left(-\frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} + \frac{y_{j+1} - y_{j-1}}{2h} - y'_j \right) \\
&\quad - h^2 \partial_2 f(w_j) \varepsilon_j \\
&\quad + \frac{h^2}{2} \partial_3^2 f(w_j) \cdot \left(-\frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} + \underbrace{\frac{y_{j+1} - y_{j-1}}{2h} - y'_j}_{\frac{h^2}{6} y_j^{(3)} + \mathcal{O}(h^4)} \right)^2 + \mathcal{O}(h^5) \\
-\varepsilon_{j-1} + 2\varepsilon_j - \varepsilon_{j+1} &= -\frac{h^4}{12} y_j^{(4)} + h^2 \partial_3 f(w_j) \left(-\frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} \right) \\
&\quad + \frac{h^4}{6} \partial_3 f(w_j) y_j^{(3)} - h^2 \partial_2 f(w_j) \varepsilon_j \\
&\quad + \frac{h^2}{2} \partial_3^2 f(w_j) \left(\frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} \right)^2 + \mathcal{O}(h^5)
\end{aligned}$$

d.h.

$$\begin{aligned}
&-(1 + \partial_3 f(w_j) \frac{h}{2}) \frac{\varepsilon_{j-1}}{h^2} + (2 + h^2 \partial_2 f(w_j)) \frac{\varepsilon_j}{h^2} - (1 - \frac{h}{2} \partial_3 f(w_j)) \frac{\varepsilon_{j+1}}{h^2} = \\
&-\frac{h^2}{12} y_j^{(4)} + \frac{h^2}{6} \partial_3 f(w_j) y_j^{(3)} + \frac{1}{2} h^4 \partial_3^2 f(w_j) \left(\frac{\varepsilon_{j+1}}{h^2} - \frac{\varepsilon_{j-1}}{h^2} \right)^2 + \mathcal{O}(h^3) \quad (2.7)
\end{aligned}$$

Definiere ε_j^* mit $\varepsilon_0^* = \varepsilon_N^* = 0$ als Lösung von

$$\begin{aligned}
-(1 + \partial_3 f(w_j) \frac{h}{2}) \varepsilon_{j-1}^* + (2 + h^2 \partial_2 f(w_j)) \varepsilon_j^* - (1 - \frac{h}{2} \partial_3 f(w_j)) \varepsilon_{j+1}^* = \\
\frac{h^2}{6} (\partial_3 f(w_j) y_j^{(3)} - \frac{1}{2} y_j^{(4)}) \quad j = 1, \dots, N-1
\end{aligned}$$

Dann gilt nach dem ersten Teil des Beweises:

ε^* ist die mit dem Diskretisierungsverfahren angenäherte Lösung der linearen RWA

$$\begin{aligned}
e''(x) &= p(x)e'(x) + q(x)e(x) + r(x), \quad e(a) = e(b) = 0 \quad \text{mit} \\
p(x) &:= \partial_3 f(x, y(x), y'(x)), \\
q(x) &:= \partial_2 f(x, y(x), y'(x)) \\
r(x) &:= -\frac{1}{6} (p(x)y^{(3)}(x) - \frac{1}{2} y^{(4)}(x))
\end{aligned}$$

also

$$\varepsilon_j^* - e(x_j) = \mathcal{O}(h^2), \quad \frac{\varepsilon_{j+1}^* - \varepsilon_{j-1}^*}{2h} = e'(x_j) + \mathcal{O}(h) \quad (2.8)$$

Definiere nun ε_j^{**} mit $\varepsilon_0^{**} = \varepsilon_N^{**} = 0$ als Lösung von

$$\begin{aligned} -(1 + p_j \frac{h}{2})\varepsilon_{j-1}^{**} + (2 + h^2 q_j)\varepsilon_j^{**} - (1 - \frac{h}{2} p_j)\varepsilon_{j+1}^{**} &= -\frac{h^2}{12} y_j^{(4)} + \frac{h^2}{6} p_j y_j^{(3)} \\ &+ \frac{1}{2} h^4 \partial_3^2 f(w_j) \left(\frac{\varepsilon_{j+1}^{**} - \varepsilon_{j-1}^{**}}{2h} \right)^2 \end{aligned} \quad (2.9)$$

Wir haben zunächst zu zeigen, daß dieses nichtlineare Gleichungssystem eine (lokal eindeutige) Lösung in einer Umgebung von $(e(x_j))_j$ besitzt. Dazu soll der Banach'sche Fixpunktsatz benutzt werden. Mit

$$\begin{aligned} A(h) &:= \text{tridiag}(-1 + p_j \frac{h}{2}, (2 + h^2 q_j), -(1 - p_j \frac{h}{2})) \\ g(h) &:= (-\frac{1}{12} y_j^{(4)} + \frac{1}{6} p_j y_j^{(3)})_j \\ F(h, \varepsilon^{**}) &= (\frac{1}{8} \partial_3^2 f(w_j) (\varepsilon_{j+1}^{**} - \varepsilon_{j-1}^{**})^2)_j \end{aligned}$$

hat das System (2.9) die Form

$$A(h)\varepsilon^{**} = h^2(g(h) + F(h, \varepsilon^{**}))$$

oder

$$\varepsilon^{**} = h^2(A(h)^{-1})(g(h) + F(h, \varepsilon^{**})) = \Phi(h, \varepsilon^{**})$$

Dabei gilt

$$\|h^2(A(h))^{-1}\| = \mathcal{O}(1), \quad \|g(h)\| = \mathcal{O}(1) \quad \text{für } 0 < h \leq h_0$$

mit $h_0 > 0$ geeignet gewählt. (vgl. oben)

Genauer gilt nach Satz 2.3.1 und dem soeben bereits bewiesenen mit geeigneten Konstanten M_3, M_4, B und $\kappa > 0$

$$\begin{aligned} \|h^2 A(h)^{-1}\|_\infty &\leq 3/\kappa \\ |\varepsilon_j^* - e(x_j)| &\leq h^2(2BM_3 + M_4)/(4\kappa) =: h^2\delta \end{aligned}$$

für $0 < h \leq h_0$, $h_0 > 0$ geeignet.

Die Vektorfunktion $F(h, \varepsilon^{**})$ ist bezüglich ε^{**} differenzierbar und es gilt

$$\mathcal{J}_F(h, \varepsilon^{**}) = \frac{1}{4} \text{tridiag}(-\partial_3^2 f(w_j)(\varepsilon_{j+1}^{**} - \varepsilon_{j-1}^{**}), 0, +\partial_3^2 f(w_j)(\varepsilon_{j+1}^{**} - \varepsilon_{j-1}^{**})).$$

Wir betrachten nun einen Streifen der Breite $2h_1^2 K$ um den Graphen der Funktion $e(x)$. (mit einer noch festzulegenden Konstanten K)

$$S := \{(x, y) \in \mathbb{R}^2 : |y - e(x)| \leq h_1^2 K\}$$

mit $0 < h_1 < h_0$. Dann gilt für $K > \delta$

- (i) $(x_j, \varepsilon_j^*) \in S \quad j = 1, \dots, N-1$
- (ii) $(x_j, \varepsilon_j^{**}) \in S \quad j = 1, \dots, N-1 \Rightarrow$

$$\|\mathcal{J}_\Phi(\varepsilon^{**})\|_\infty \leq 1.5\kappa^{-1}C \cdot (2h_1^2 K + 2h_1 M_1)$$

$$\text{wo } M_1 = \|e'\|_{\infty, [a, b]}, \quad C := \|\partial_3^2 f(x, y(x), y'(x))\|_{\infty, [a, b]}$$

Somit ist $\Phi(h, \varepsilon^{**})$ kontrahierend auf \mathcal{D} :

$$\mathcal{D} := \{\varepsilon^{**} : (x_j, \varepsilon_j^{**}) \in S\}$$

für hinreichend kleines h_1 . Wir können h_1 so klein wählen, daß die Kontraktionskonstante $L \leq \frac{1}{2}$ wird.

Wir setzen nun

$$\varepsilon_j^{[0]} := \varepsilon_j^* \quad \text{mit } \varepsilon_j^* \quad \text{aus (2.8)}$$

und wollen zeigen, daß für genügend kleines h_1 eine Umgebung von $\varepsilon^{[0]}$ unter Φ auf sich abgebildet wird.

Dazu benutzen wir das Selbstabbildungskriterium

$$\mathcal{D}_0 := \{\varepsilon^{**} : \|\varepsilon^{**} - \varepsilon^{[0]}\|_\infty \leq \frac{1}{1-L} \|\varepsilon^{[1]} - \varepsilon^{[0]}\|_\infty\} \subset \mathcal{D}$$

mit $L = \frac{1}{2}$ und

$$\varepsilon^{[1]} := \Phi(h, \varepsilon^{[0]}).$$

Nach Setzung von $\varepsilon^{[0]}$ gilt

$$\begin{aligned} A(h)\varepsilon^{[0]} &= h^2 g(h) \\ A(h)\varepsilon^{[1]} &= h^2 g(h) + h^2 F(h, \varepsilon^{[0]}) \end{aligned}$$

Also

$$\|\varepsilon^{[1]} - \varepsilon^{[0]}\|_\infty \leq 3\kappa^{-1} \|F(h, \varepsilon^{[0]})\|_\infty.$$

Aber

$$\begin{aligned} |(F(h, \varepsilon^{[0]}))_j| &\leq \frac{1}{8} C(2h_1^2 \delta + 2h_1 M_1)^2 \\ &= \frac{1}{2} C h_1^2 M_1^2 (1 + h^2 \delta / M_1) \\ &\leq C h_1^2 M_1^2 \end{aligned}$$

für genügend kleines h_1 und daher

$$\frac{1}{1-L} \|\varepsilon^{[1]} - \varepsilon^{[0]}\|_\infty \leq 6\kappa^{-1} C h_1^2 M_1^2$$

Für $\varepsilon^{**} \in \mathcal{D}_0$ ergibt sich also

$$|\varepsilon_j^{**} - e_j| \leq 6\kappa^{-1} C h_1^2 M_1^2 + h_1^2 \delta$$

d.h. mit $K := \delta + 6\kappa^{-1} C M_1^2$

$$(x_j, \varepsilon_j^{**}) \in S \quad \text{d.h.} \quad \varepsilon^{**} \in \mathcal{D}$$

d.h. die Selbstabbildung auf \mathcal{D}_0 . Somit existiert auf \mathcal{D}_0 genau ein Fixpunkt von Φ . Dies ist zugleich der einzige Fixpunkt in \mathcal{D} . Somit hat das nichtlineare System (2.9) genau eine Lösung ε^{**} mit $\|\varepsilon^{**} - e\| = \mathcal{O}(h^2)$ für $0 < h \leq h_1$.

Nun betrachten wir mit

$$\tilde{\varepsilon}_j := \varepsilon_j/h^2$$

das ursprüngliche System (2.7).

Danach ist

$$\begin{aligned}\tilde{\varepsilon} &= \Phi(h, \tilde{\varepsilon}) + \mathcal{O}(h) \\ \varepsilon^{**} &= \Phi(h, \varepsilon^{**})\end{aligned}$$

d.h.

$$\|\tilde{\varepsilon} - \varepsilon^{**}\| = \mathcal{O}(h)$$

(Störungssatz für das System $\varepsilon - \Phi(h, \varepsilon) = 0$) d.h.

$$\varepsilon_j = h^2 e_j + \mathcal{O}(h^3)$$

Damit aber wird

$$\frac{\varepsilon_{j+1} - \varepsilon_{j-1}}{2h} = \mathcal{O}(h^2).$$

Setzt man dies statt der ursprünglichen Abschätzung

$(\varepsilon_{j+1} - \varepsilon_{j-1})/(2h) = \mathcal{O}(h)$ in der obigen Herleitung wieder ein, bekommt man in (2.7) einen Term $\mathcal{O}(h^4)$ statt $\mathcal{O}(h^2)$ und damit

$$\tilde{\varepsilon} = \Phi(h, \tilde{\varepsilon}) + \mathcal{O}(h^2),$$

d.h.

$$\begin{aligned}\|\tilde{\varepsilon} - \varepsilon^{**}\| &= \mathcal{O}(h^2) \\ \varepsilon_j &= h^2 e_j + \mathcal{O}(h^4)\end{aligned}$$

□

>>

Beispiel 2.3.2. *Wir lösen das lineare Randwertproblem*

$$\begin{cases} y'' &= 10y - 9y' + x \\ y(0) &= 0 \\ y(1) &= 0 \end{cases}$$

und bestimmen gemäß Satz 2.3.2 den Hauptteil des globalen Diskretisierungsfehlers.

a) *Als homogene Lösung der linearen Dgl. 2. Ordnung erhalten wir:*

$$\eta(x; h) = C_1 e^x + C_2 e^{-10x}$$

Mit einem "Ansatz vom Typ der rechten Seite" erhalten wir als partikuläre Lösung:

$$y_p(x) = -\frac{t}{10} - \frac{9}{100}$$

Insgesamt erhalten wir:

$$y(x) = C_1 e^x + C_2 e^{-10x} - \frac{x}{10} - \frac{9}{100}$$

Einarbeitung der Randdaten liefert:

$$C_1 = \frac{9}{100} - \frac{19 - 9e}{100(e^{-10} - e)} \doteq 0.0699$$

$$C_2 = \frac{19 - 9e}{100(e^{-10} - e)} \doteq 0.0201$$

b) Nach Satz 2.3.2 lautet die Differentialgleichung zur Bestimmung des Hauptteils des lokalen Diskretisierungsfehlers:

$$\begin{cases} e''(x) = p(x)e'(x) + q(x)e(x) + r(x) \\ e(0) = 0 \\ e(1) = 0 \end{cases}$$

mit

$$f(x, y, y') = -9y' + 10y + x$$

$$p(x) = \partial_3 f(x, y(x), y'(x)) = -9$$

$$q(x) = \partial_2 f(x, y(x), y'(x)) = 10$$

$$r(x) = -\frac{1}{6} \{ p(x)y'''(x) - \frac{1}{2}y^{(4)}(x) \}$$

$$r(x) = -\frac{1}{6} \left\{ -9C_1 e^x + 9000C_2 e^{-10x} - \frac{1}{2}C_1 e^x - \frac{10000}{2}C_2 e^{-10x} \right\}$$

$$r(x) = Ae^x + Be^{-10x}$$

wobei $A := \frac{19}{12}C_1$ und $B := -\frac{2}{3}C_2 10^3$ sei.

Als homogene Lösung erhalten wir somit wiederum:

$$e_h(x) = D_1 e^x + D_2 e^{-10x}$$

Als partikuläre Lösung ergibt sich nach dem "Ansatz vom Typ der rechten Seite":

$$e_p(x) = \frac{19}{132}C_1 x e^x + \frac{2}{33}C_2 10^3 x e^{-10x}$$

Aus den Randbedingungen folgt nun:

$$D_1 = -D_2 \doteq -0.0103$$

$$e(x) = -0.0103e^x + 0.0103e^{-10x} + 0.0101xe^x + 1.218xe^{-10x}$$

□

Bemerkung 2.3.4. Bei entsprechend höherer Differenzierbarkeitsordnung von f hat auch die asymptotische Entwicklung von $\eta(x; h) - y(x)$ entsprechend mehr Terme. Im Prinzip ist also auf diesem Wege unter Benutzung der Richardsonextrapolation eine beliebige Steigerung der Ordnung möglich. \square

Bemerkung 2.3.5. Wenn man für die einzelnen Ableitungen in der Differentialgleichung direkt Differenzenapproximationen höherer Ordnung einsetzt, was ja leicht machbar ist, dann erhält man direkt eine höhere Konsistenzordnung. Man verliert dann aber die M -Matrizeigenschaft der Jacobimatrix des diskretisierten Systems und damit die Stabilität des Verfahrens. Deshalb wählt man diesen Weg nicht. Es gibt aber Ansätze, bei denen man die Jacobi-Matrix durch ein Produkt von M -Matrizen majorisieren kann, das die gleichen Abschätzungstechniken wie oben erlaubt. Details siehe z.B. bei Bohl. \square

Wenn die Differentialgleichung von y' abhängt, dann ist die Matrix bzw. Jacobimatrix des Systems bei der hier besprochenen Vorgehensweise zwar eine M -Matrix, aber nicht symmetrisch. In einem speziellen Fall kann man jedoch direkt eine symmetrieerhaltende Diskretisierung angeben:

Definition 2.3.1. Sei $\langle \cdot, \cdot \rangle$ das L_2 -Skalarprodukt auf einem Intervall $[a, b]$ und L ein Differentialoperator für reelle Funktionen auf diesem Intervall mit Definitionsbereich \mathcal{D}_L . Dann heißt L selbstadjungiert auf \mathcal{D}_L wenn

$$\langle u, L(v) \rangle = \langle L(u), v \rangle \quad \forall u, v \in \mathcal{D}_L.$$

Bemerkung 2.3.6. Ein selbstadjungierter Differentialoperator ist notwendig von gerader Ordnung. Ist

$$(Ly)(x) = \sum_{i=0}^n f_i(x)y^{(i)}(x),$$

dann ist L selbstadjungiert genau dann, wenn er mit seinem formal adjungierten Operator L^* mit

$$(L^*y)(x) = \sum_{i=0}^n (-1)^i \left(\frac{d}{dx} \right)^i \{f_i(x)y(x)\}$$

übereinstimmt.

Für einen selbstadjungierten Differentialoperator zweiter Ordnung auf $C^2[a, b]$ bedeutet dies, daß er stets die Form

$$(Ly)(x) = -\frac{d}{dx}\left(p(x)\frac{d}{dx}y(x)\right) + q(x)y(x)$$

hat. Diskretisiert man hier

$$\frac{d}{dx}\left(p(x)\frac{d}{dx}y(x)\right)|_{x=x_i} = \frac{1}{h^2}(p(x_i + h/2)(y_{i+1} - y_i) - p(x_i - h/2)(y_i - y_{i-1}))$$

dann erzielt man eine symmetrische Koeffizientenmatrix für das zugehörige Gleichungssystem (bei Dirichletranddaten). Auch diese Diskretisierung ist von zweiter Ordnung konsistent und stabil, also konvergent von der Ordnung h^2 .

Beispiel 2.3.3. Gegeben sei die Randwertaufgabe in selbstadjungierter Form

$$-((1+x^3)y')' + xy = 0, \quad x \in [0, 1] \quad \text{mit} \quad y(0) = y(1) = 1.$$

Wir diskretisieren das RWP nach der symmetrischen Diskretisierung für selbstadjungierte Probleme mit der Schrittweite $h = \frac{1}{4}$. Bei der symm. Diskretisierung für selbstadjungierte Probleme wird $-((1+x^3)y')'$ in zwei Schritten symmetrisch approximiert. Das Resultat ist

$$-\frac{1}{h^2} \left((1 + (x_i + \frac{h}{2})^3)(y_{i+1} - y_i) - (1 + (x_i - \frac{h}{2})^3)(y_i - y_{i-1}) \right) + x_i y_i = 0$$

mit $i = 1, 2, 3$ und $y_0 = y_4 = 1$.

Benötigt werden die Zahlenwerte

x_i	$x_i \pm \frac{h}{2}$	$\frac{1+(x_i \pm \frac{h}{2})^3}{h^2}$
0.25	0.125	16.03125
0.5	0.375	16.84375
0.75	0.625	19.90625
	0.875	26.71875

Sortiert man in den einzelnen Gleichungen nach y_i und bringt die Randwerte auf die rechte Seite, so ergibt sich das lineare Gleichungssystem

$$\begin{pmatrix} 33.125 & -16.84375 & 0 \\ -16.84375 & 37.25 & -19.90625 \\ 0 & -19.90625 & 47.375 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 16.03125 \\ 0 \\ 26.71875 \end{pmatrix}.$$

Es handelt sich offensichtlich um eine irreduzibel diagonaldominante tridiagonale L-Matrix, also eine tridiagonale M-Matrix. Da diese zusätzlich symmetrisch ist folgt die positive Definitheit.

2.4 Kompakte Differenzenschemata für spezielle Randwertaufgaben zweiter Ordnung. ERG

Definition 2.4.1. Ein Differenzenschema für eine gewöhnliche Differentialgleichung der Ordnung m heisst kompakt, wenn es nur $m + 1$ Gitterpunkte zur Diskretisierung jedes Ableitungswertes benutzt.

Wir betrachten hier nur das einfachere Problem

$$-y'' = f(x, y), \quad x \in]a, b[, \quad y(a) = y_0, \quad y(b) = y_1$$

mit einer Vektorfunktion y und f . Anstatt nun nur y'' durch eine zentrierte Differenz zu ersetzen, machen wir den Ansatz

$$\alpha_0 \eta_{i+1} + \alpha_1 \eta_i + \alpha_2 \eta_{i-1} = h^2 (\beta_0 f_{i+1} + \beta_1 f_i + \beta_2 f_{i-1}), \quad 1 \leq i \leq N-1$$

mit der Setzung

$$f_i = f(x_i, \eta_i).$$

Dies beschreibt also mit den Randwerten wieder ein Gleichungssystem zur Bestimmung von $\eta_1, \dots, \eta_{N-1}$. Ziel wird es sein, die Koeffizienten α_0, \dots, β_2 so zu bestimmen, daß die Konsistenzordnung maximal wird. Offenbar kann man eine Konstante frei wählen, etwa

$$\alpha_0 = -1.$$

Dann ergibt sich notwendig

$$\alpha_1 = 2, \quad \alpha_2 = -1, \quad \beta_0 + \beta_1 + \beta_2 = 1$$

und für die Konsistenz zweiter Ordnung auch noch

$$\beta_0 = \beta_2.$$

$\beta_0 = \beta_2 = 0$ ergibt unsere zuerst betrachtete Methode und z.B.

$$\beta_0 = \beta_2 = \frac{1}{12}, \quad \beta_1 = \frac{10}{12}$$

die Methode von Cowell, die von vierter Ordnung konsistent ist. Das entstehende Gleichungssystem für die Näherungen η_i hat dann wieder eine (Block-)tridiagonale (Jacobi-)Matrix. Man kann diesen Ansatz weitertreiben: Mit jeweils fünf Gitterpunkten erhalten wir den Ansatz

$$\sum_{k=-2}^2 \alpha_k \eta_{i+k} = h^2 \sum_{k=-2}^2 \beta_k f_{i+k} \quad (2.10)$$

Dieser ist allerdings in den randnahen Punkten $a+h$ und $b-h$ nicht möglich. Für Lösungen y aus $C^8([a, b])$ ist (2.10) von sechster Ordnung konsistent, wenn man die Koeffizienten wählt wie folgt:

$$(\alpha_{-2}, \dots, \alpha_2) = \frac{1}{20}(-1, -16, 34, -16, -1)$$

und

$$(\beta_{-2}, \dots, \beta_2) = \frac{1}{15}(0, 2, 11, 2, 0).$$

In den beiden randnahen Punkten kann man nur die Formel von Cowell benutzen. Dennoch kann man zeigen, daß die Konvergenzordnung dieses Schemas 6 ist (siehe bei Bohl). Solche Schemata kann man auch für allgemeinere Fälle angeben.

2.5 Behandlung von Grenzschichten

Wir betrachten hier nur das einfache Modell

$$-\varepsilon y'' + \beta y' = 0, \quad y(0) = 0, \quad y(1) = 1$$

mit der Lösung

$$y(x) = (\exp(\beta x/\varepsilon) - 1)/(\exp(\beta/\varepsilon) - 1).$$

Für $\beta/\varepsilon \gg 1$ weist diese Lösung eine sehr schmale Grenzschicht bei $x = 1$ auf und schon für $x = 1 - 10\varepsilon/\beta$ ist der Wert der Lösung praktisch null:

$$x \leq 1 - 10\varepsilon/\beta, \quad \beta/\varepsilon \geq 10 \Rightarrow y(x) \leq 2/(\exp(10) - 1) = .000090804.$$

Bei Anwendung unserer zentralen Differenzenquotienten mit einem äquidistanten Gitter müsste dann die Gitterweite h

$$h \ll 2\varepsilon/\beta$$

erfüllen, um zu vernünftigen Ergebnissen zu gelangen. In dieser Situation ist es viel besser, mit nichtäquidistanten Gittern zu arbeiten, in diesem Fall hier mit einem geometrisch progressiven Gitter

$$h_i = ch_{i+1}, \quad N - 2 \geq i \geq 0, \quad h_{N-1} = \varepsilon/10 \text{ z.B. und } c = 2.$$

Dazu benutzt man dann die Differenzenapproximationen

$$\begin{aligned} y''(x_j) &= \frac{2}{h_{j-1}(h_j + h_{j-1})}y_{j-1} - \frac{2}{h_{j-1}h_j}y_j + \frac{2}{h_j(h_j + h_{j-1})}y_{j+1} + \mathcal{O}(h_{\max}) \\ y'(x_0) &= \frac{y_1 - y_0}{h_0} + \mathcal{O}(h_{\max}) \\ y'(x_N) &= \frac{y_N - y_{N-1}}{h_{N-1}} + \mathcal{O}(h_{\max}) \\ y'(x_j) &= \frac{\frac{h_{j-1}}{h_j}(y_{j+1} - y_j) - \frac{h_j}{h_{j-1}}(y_{j-1} - y_j)}{h_j + h_{j-1}} + \mathcal{O}(h_{\max}^2) \end{aligned}$$

und erhält wiederum eine tridiagonales System, allerdings nur mit der Konvergenzordnung 1. Wenn keine Randschicht vorliegt, die man fein auflösen muss, kann man sich bei solchen Problemen anders behelfen, ohne die Äquidistanz des Gitters aufzugeben: Wir betrachten sogleich die allgemeinere Aufgabe

$$-\varepsilon y'' + a(x)y' + b(x)y = q(x), \quad x \in]a, b[$$

und den Fall

$$|a(x)| \gg \varepsilon > 0.$$

Wir wollen nun so diskretisieren, daß die M-Matrix Eigenschaft der Matrix der diskretisierten Gleichung erhalten bleibt. Für hy'_i benutzen wir eine Formel

$$hy'_i = \alpha_{-1}y_{i-1} + \alpha_0y_i + \alpha_1y_{i+1}.$$

Nach Multiplikation mit h^2 und unter Benutzung des zentralen Differenzenquotienten für y_i'' lauten die wesentlichen Koeffizienten in der Matrix des Systems

$$-\varepsilon + h\alpha_{-1}a(x_i), 2\varepsilon + h\alpha_0a(x_i) + h^2b(x_i), -\varepsilon + h\alpha_1a(x_i)$$

Setzen wir für

$$a(x_i) \geq 0 : \alpha_0 = 1, \alpha_{-1} = -1, \alpha_1 = 0,$$

also den Rückwärtsdifferenzenquotienten für y_i' und für

$$a(x_i) < 0 : \alpha_{-1} = 0, \alpha_0 = -1, \alpha_1 = 1,$$

also den Vorwärtsdifferenzenquotienten für y_i' , dann ist die Koeffizientenmatrix eine irreduzibel diagonaldominante L-Matrix, also eine M-Matrix. Diese Vorgehensweise nennt man die "upwind"-Diskretisierung. Sie liefert natürlich nur Konsistenz und Konvergenz von der Ordnung eins. Kompakt geschrieben lautet sie

$$\begin{aligned} \frac{\varepsilon}{h^2}(-\eta_{i-1} + 2\eta_i - \eta_{i+1}) + \frac{a(x_i)}{h}\phi_i + b(x_i)\eta_i &= q(x_i), \quad 1 \leq i \leq N-1 \\ \phi_i &= \begin{cases} \eta_{i+1} - \eta_i & \text{falls } a(x_i) < 0 \\ \eta_i - \eta_{i-1} & \text{falls } a(x_i) \geq 0 \end{cases} \end{aligned}$$

Es ist

$$\frac{a(x_i)}{h}\phi_i = a(x_i)\frac{\eta_{i+1} - \eta_{i-1}}{2h} + h|a(x_i)|\frac{-\eta_{i-1} + 2\eta_i - \eta_{i+1}}{2h^2}$$

der Übergang vom symmetrischen Differenzenquotienten zu dieser upwind Diskretisierung wirkt also wie die

$$\text{Addition von } -\frac{h|a(x)|y''}{2} \text{ zur Differentialgleichung.}$$

also wie die Addition eines künstlichen Diffusionsterms. Die physikalisch unsinnigen Oszillationen für $h > 2\varepsilon/|a(x)|$ treten dann nicht auf.

2.6 Kollokationsmethoden

Differenzenverfahren haben wir bisher nur für spezielle skalare Gleichungen zweiter Ordnung betrachtet. Wir kehren nun zurück zur allgemeinen Zweipunkttrandwertaufgabe

$$y' = F(x, y), \quad x \in]a, b[, \quad R(y(a), y(b)) = 0.$$

Auch hier können wir sinnvolle Differenzenformeln direkt ansetzen. Ein Beispiel ist die implizite Mittelpunkregel

$$\frac{\eta_{i+1} - \eta_i}{h} = F\left(x_i + \frac{h}{2}, \frac{\eta_i + \eta_{i+1}}{2}\right), \quad i = 0, \dots, N-1$$

mit $h = (b - a)/N$ oder die Trapezregel

$$\frac{\eta_{i+1} - \eta_i}{h} = \frac{1}{2}(F(x_i, \eta_i) + F(x_{i+1}, \eta_{i+1})).$$

Zusammen mit der Randbedingung entsteht also ein

Gleichungssystem für η_0, \dots, η_N .

Dieses ist nichtlinear, wenn F oder R nichtlinear in y sind.

Beispiel 2.6.1. Anwendung der Trapezregel auf ein lineares Randwertproblem:

$$y' = Ay + f(x) \quad \text{mit} \quad \alpha y(0) - y(1) = 0.$$

$N = 3$ ergibt als Unbekannte die Vektoren η_i , $i = 0, 1, 2, 3$ und das gekoppelte Gleichungssystem

$$\begin{pmatrix} \alpha I & O & O & -I \\ -(I + \frac{h}{2}A) & (I - \frac{h}{2}A) & O & O \\ O & -(I + \frac{h}{2}A) & (I - \frac{h}{2}A) & O \\ O & O & -(I + \frac{h}{2}A) & (I - \frac{h}{2}A) \end{pmatrix} \begin{pmatrix} \eta_0 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{f_0+f_1}{2} \\ \frac{f_1+f_2}{2} \\ \frac{f_2+f_3}{2} \end{pmatrix}.$$

□

In Abhängigkeit von der Jacobi-Matrix von F bezüglich y bzw. der von R können wieder die gleichen Probleme auftreten, die wir schon bei den Gleichungen zweiter Ordnung diskutiert haben. Diese beiden Formeln ergeben sich als Spezialfälle einer Vorgehensweise, die als (lokale) Kollokationsmethode bezeichnet wird.

Diese Vorgehensweise gilt gegenwärtig als die beste universelle Lösungsmethode für solche Randwertaufgaben. Die Idee dieser Methoden ist die folgende: man ersetzt die gesuchte Lösung der DGL lokal (d.h. zwischen zwei Gitterpunkten x_i und x_{i+1}) durch eine "einfache" Funktion $\varphi(x)$ und legt die Parameter, die diese Funktion beschreiben, dadurch fest, daß man fordert, daß diese an gewissen Zwischenpunkten die Differentialgleichung (exakt) erfüllt. In der Regel ist φ ein Polynom vom Grad s , das man durch die Forderungen

$$\begin{aligned} \varphi(x_i) &= \eta_i \\ \varphi'(x_i + \alpha_j h) &= F(x_i + \alpha_j h, \varphi(x_i + \alpha_j h)), \quad j = 1, \dots, s \end{aligned}$$

festlegt. Man setzt dann

$$\eta_{i+1} = \varphi(x_i + h).$$

Die Koeffizienten α_j erfüllen dabei

$$0 \leq \alpha_1 < \alpha_2 < \dots < \alpha_s \leq 1.$$

In den obigen Beispielen ist $s = 1$ und $\alpha_1 = 1/2$ bzw. $s = 2$ und $\alpha_1 = 0$, $\alpha_2 = 1$. Durch die Interpolationsforderungen ist das Polynom φ eindeutig bestimmt. Nach der Formel von Lagrange ergibt sich

$$\varphi'(x_i + \tau h) = \sum_{j=1}^s L_j(x_i + \tau h) K_j$$

mit

$$K_j \stackrel{\text{def}}{=} \varphi'(x_i + \alpha_j h)$$

und

$$L_j(x_i + \tau h) = \prod_{l=1, l \neq j}^s \frac{\tau - \alpha_l}{\alpha_j - \alpha_l} .$$

Man beachte, daß L_j unabhängig von h und x_i ist. Integration von φ' ergibt dann

$$\begin{aligned} \varphi(x_i + \alpha_j h) - \varphi(x_i) &= h \sum_{l=1}^s \beta_{j,l} K_l \\ \varphi(x_{i+1}) - \varphi(x_i) &= h \sum_{l=1}^s \gamma_l K_l \end{aligned}$$

wobei

$$\begin{aligned} \beta_{j,l} &= \int_0^{\alpha_j} L_l(\tau) d\tau , \\ \gamma_l &= \int_0^1 L_l(\tau) d\tau . \end{aligned}$$

Setzen wir dies und die Definition der K_l in die Bestimmungsgleichungen wieder ein, so erhalten wir die Gleichungen

$$\begin{aligned} K_j &= F(x_i + \alpha_j h, \eta_i + h \sum_{l=1}^s \beta_{j,l} K_l), \quad j = 1, \dots, s \\ \eta_{i+1} &= \eta_i + h \sum_{l=1}^s \gamma_l K_l . \end{aligned}$$

Bei gegebenem η_i sind also zunächst die Gleichungen für die K_l zu lösen, danach kann man η_{i+1} unmittelbar angeben. Dies entspricht genau der Situation bei den Runge–Kutta–Verfahren für Anfangswertprobleme. Aber etwas ist hier grundlegend anders: Man beachte, daß die $\beta_{k,l}$ und die γ_l nur von den gewählten Knoten α_j abhängen, durch diese aber bereits eindeutig festgelegt sind. Durch die obigen Festlegungen der $\beta_{i,j}$ und γ_j hat das Verfahren automatisch die Mindestordnung s .

Im Zusammenhang mit einer Randwertaufgabe erhalten wir also simultane Gleichungen für alle η_i und die Werte $K_l(i)$, die ja auch noch von der Gitterstelle abhängen. Sinnvolle Wahlen der α_j sind hier solche, die symmetrisch zur Intervallmitte liegen und eine möglichst hohe Ordnung besitzen. Es bieten sich dazu also die Gaussknoten (bezogen auf $[0, 1]$) (dies ergibt die Ordnung $2s$) und die sogenannten Lobatto-Knoten an. Diese haben stets $\alpha_1 = 0, \alpha_s = 1$ und die noch freien Knoten werden so gewählt, daß die Ordnung maximal wird, also $2s - 2$.

Beispiel 2.6.2. Gauss - Runge- Kutta - Verfahren mit 2 Stufen der Ordnung 4:

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Beispiel 2.6.3. Gauss - Lobatto - Runge- Kutta - Verfahren mit 3 Stufen der Ordnung 4:

$$\begin{array}{c|ccc}
 0 & 0 & 0 & 0 \\
 \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\
 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\
 \hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
 \end{array}$$

Die implizite Mittelpunkregel ist zugleich die einfachste Gauss-Regel und die Trapezregel die einfachste Lobatto-Quadratur. Es gilt zur Konvergenz dieser Verfahren der

Satz 2.6.1. Die Randwertaufgabe sei lokal eindeutig lösbar und F und R hinreichend oft differenzierbar. Dann konvergiert die Kollokationsmethode mit den Gaussknoten und s Stufen von der Ordnung $2s$ und die mit s Stufen und den Lobatto-Knoten von der Ordnung $2s - 2$.

Für den Beweis siehe z.B. bei Ascher und Petzold. Diese Kollokationsmethoden kann man auch bei viel allgemeineren Problemen (Mehrpunktprobleme, wo auch an Zwischenstellen Vorschriften für die Funktion gegeben sind, Gleichungen höherer Ordnung, Gleichungen mit algebraischen Nebenbedingungen) einsetzen und es gibt gute Software dafür (z.B. COLDAE, COLNEW von Ascher und Bader in der NETLIB).

Bemerkung 2.6.1. Man kann Kollokation auch mit einem globalen Ansatz durchführen. Man wählt dazu eine geeigneten endlich dimensionalen Unterraum von differenzierbaren Funktionen, die die Randbedingungen erfüllen, ersetzt formal die Lösung y durch eine Linearkombination einer Basis dieses Unterraums, etwa

$$\eta(x) = \sum_{i=1}^n \alpha_i \varphi_i(x)$$

und legt die Parameter α_i dieses Ansatzes durch die Forderung

$$\eta'(x_j) = F(x_j, \eta(x_j)) \quad , \quad j = 1, \dots, n$$

mit geeignet gewählten Kollokationspunkten x_j fest. Dieser globale Ansatz ist jedoch recht problematisch und wird heute kaum noch verfolgt.

Beispiel 2.6.4. Wir betrachten die Randwertaufgabe

$$y''(x) + y(x) = -x, \quad y(0) = y(1) = 0$$

mit der exakten Lösung $y(x) = \frac{\sin x}{\sin 1} - x$ und berechnen eine Näherungslösung dieses Problems mit dem Kollokationsverfahren mit den Kollokationsstellen $x_i = i/3$, $i = 1, 2$ und den Ansatzfunktionen

$$\Phi_i(x) = x^i(1-x), \quad i = 1, 2.$$

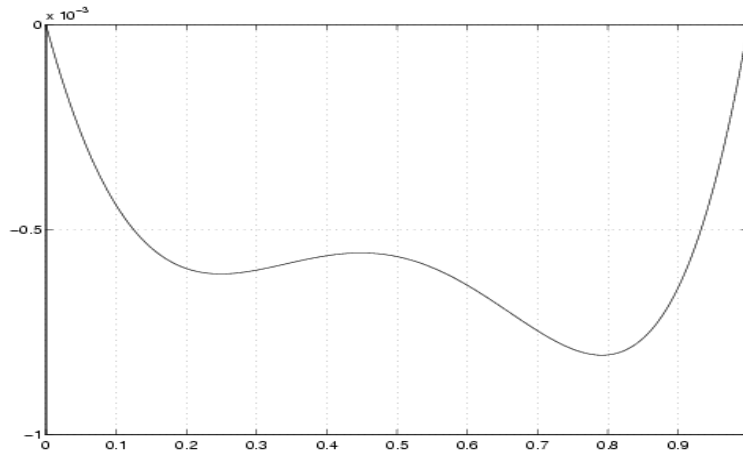
Wir setzen eine Linearkombination der Ansatzfunktionen in die DGL ein. Dies führt auf das lineare Gleichungssystem $Ax = b$ mit:

$$\begin{aligned} a_{11} &= \Phi_1''(x_1) + \Phi_1(x_1) = -2 + x - x^2|_{x=x_1} = -\frac{16}{9} \\ a_{21} &= \Phi_1''(x_2) + \Phi_1(x_2) = -2 + x - x^2|_{x=x_2} = -\frac{16}{9} \\ a_{12} &= \Phi_2''(x_1) + \Phi_2(x_1) = 2 - 6x + x^2 - x^3|_{x=x_1} = \frac{2}{27} \\ a_{22} &= \Phi_2''(x_2) + \Phi_2(x_2) = 2 - 6x + x^2 - x^3|_{x=x_2} = -\frac{50}{27} \\ b_1 &= -x|_{x=x_1} = -\frac{1}{3} \\ b_2 &= -x|_{x=x_2} = -\frac{2}{3} \end{aligned}$$

Als Lösung ergibt sich

$$u_1(x) = x(1-x) \left(\frac{81}{416} + \frac{9}{52}x \right).$$

Diese Lösung stellt hier schon eine sehr gute Näherung dar, der Fehler ist unten aufgetragen.



2.7 Variationsmethoden

Finite Elemente in einer Dimension

Die bei den Differenzenapproximationen auftretenden Gleichungssysteme haben den Nachteil, nicht immer symmetrisch zu sein (wenn $\partial_3 f \neq 0$ oder Ableitungen in den Randwerten auftreten). Einige besonders attraktive Lösungsverfahren für Gleichungssysteme mit positiv definiten Matrizen (eine symmetrische M -Matrix ist positiv definit!) sind dann nicht anwendbar. Die im Folgenden beschriebene Vorgehensweise vermeidet

diesen Nachteil und erlaubt zugleich einen einfachen Zugang zu Verfahren höherer Ordnung. Wir beginnen mit einem Einführungsbeispiel:

Beispiel 2.7.1. Ein Balken mit der Biegesteifigkeit $\mathcal{J}E$ liegt auf einem elastischen Untergrund (mit der Federkonstanten k pro Längeneinheit) und wird mit dem Gewicht $p(x)$ pro Längeneinheit belastet. Die dabei auftretende Durchbiegung hat die Eigenschaft, die Summe der auftretenden Energien zu minimieren:

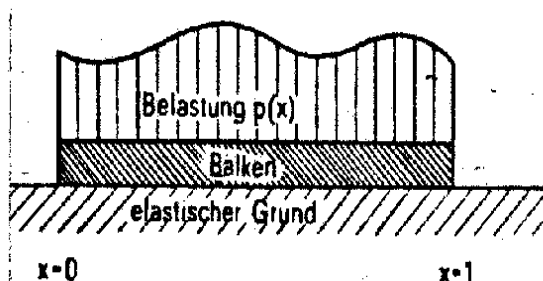


Abbildung 2.7.1.

Folgende Energien sind zu berücksichtigen:

$$\begin{aligned} \text{Die Biegeenergie} & : \frac{1}{2} \int_0^1 \mathcal{J}E(x)(y''(x))^2 dx \\ \text{Die Formänderungsenergie des Untergrundes} & : \frac{1}{2} \int_0^1 k(x)(y(x))^2 dx \\ \text{Die Arbeit der äußeren Kräfte} & : \int_0^1 p(x)y(x) dx. \end{aligned}$$

Also gilt für die Durchbiegung $y(x)$:

$$\frac{1}{2} \int_0^1 (\mathcal{J}E(x)(y''(x))^2 + k(x)(y(x))^2) dx + \int_0^1 p(x)y(x) dx = \min_y$$

Mit Hilfe der Variationsrechnung folgt hieraus das RWP

$$\begin{aligned} (\mathcal{J}E(x)y'')'' + k(x)y + p(x) &= 0 \quad 0 \leq x \leq 1, \\ y''(0) = y''(1) &= 0. \end{aligned}$$

Dabei ist \mathcal{J} das Trägheitsmoment des Balkenquerschnitts und E der Elastizitätsmodul. Beide dürfen von x abhängen. Wir setzen $(\mathcal{J}E)(x) \geq \alpha > 0$ voraus. Anstatt nun diese Differentialgleichung mit dem Differenzenverfahren (das auch direkt auf Gleichungen vierter Ordnung übertragen werden kann) zu lösen, diskretisieren wir hier direkt das Integral:

Sei speziell $\mathcal{J}E(x) \equiv 1$, $k(x) \equiv 1$. Wir benutzen die Formeln

$$\begin{aligned} y''(x_k) &= \frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1}))}{h^2} + \mathcal{O}(h^2) \quad 1 \leq k \leq M \\ y''(x_0) &= \frac{2y(x_0) - 5y(x_1) + 4y(x_2) - y(x_3))}{h^2} + \mathcal{O}(h^2) \\ y''(x_N) &= \frac{2y(x_N) - 5y(x_{N-1}) + 4y(x_{N-2}) - y(x_{N-3}))}{h^2} + \mathcal{O}(h^2) \end{aligned}$$

Die Integrale diskretisieren wir mit gleicher Genauigkeit, also z.B. mit der zusammengesetzten Trapezregel. Mit $h = 0.2$ ergibt sich

$$\begin{aligned} \frac{1}{2} \int_0^1 (y'')^2 dx &\approx \frac{1}{10} \sum_{k=1}^4 (25y_{k+1} - 50y_k + 25y_{k-1})^2 + \frac{h}{4} (50y_0 - 125y_1 + \\ &\quad + 100y_2 - 25y_3)^2 + \frac{h}{4} (50y_5 - 125y_4 + 100y_3 - 25y_2)^2, \\ \frac{1}{2} \int_0^1 y^2 dx &\approx \frac{1}{10} \left[\frac{1}{2}y_0^2 + y_1^2 + y_2^2 + y_3^2 + y_4^2 + \frac{1}{2}y_5^2 \right], \\ \int_0^1 py dx &\approx 0.2 \left[\frac{1}{2}p_0y_0 + \sum_{k=1}^4 p_k y_k + \frac{1}{2}p_4y_5 \right]. \end{aligned}$$

Die Summe der rechts auftretenden Approximationen ist eine Approximation für die Gesamtenergie des Systems. Dieser Ausdruck wird nun bezüglich der Werte y_0, \dots, y_5 minimiert, d.h. der Gradient wird Null gesetzt. Dies liefert ein lineares Gleichungssystem für Näherungswerte η_0, \dots, η_5 für $y(0), y(0.2), \dots, y(1)$:

η_0	η_1	η_2	η_3	η_4	η_5	1	
1875.5	-4375	3125	-625			$p_0/2$	= 0
-4375	10938.5	-8750	2187.5			p_1	= 0
3125	-8750	9063.5	-5000	2187.5	-625	p_2	= 0
-625	2187.5	-5000	9063.5	-8750	3125	p_3	= 0
		2187.5	-8750	10938.5	-4375	p_4	= 0
		-625	3125	-4375	1875.5	$p_5/2$	= 0

Die **Koeffizientenmatrix** dieses Systems ist als Jacobi-Matrix eines Gradienten, also als Hessematrix einer reellen Funktion, automatisch symmetrisch. Wenn die zu minimierende Funktion gleichmäßig konvex ist, d.h.

$$f(x) \geq f(y) + \nabla f(y)^T(x - y) + \gamma \|x - y\|^2 \text{ mit } \gamma > 0, \quad \forall x, y \in \mathbb{R}^n,$$

dann ist sie auch **automatisch positiv definit**. Im vorliegenden Fall einer definiten quadratischen Form ist diese Bedingung erfüllt. Daß hier eine Energie minimiert wird, hat der Methode auch den Namen **Energie-Methode** gegeben. \square

Diese Methode der direkten Diskretisierung eines Integralfunktionals und der anschließenden Lösung des endlichdimensionalen Minimierungsproblems ist auch in viel allgemeineren Fällen möglich und häufig nützlich.

Im obigen Beispiel stellt die Randwertaufgabe die notwendige Extremalbedingung für das Minimierungsproblem dar. Umgekehrt kann man häufig einer Randwertaufgabe ein Minimierungsproblem (oder allgemeiner Stationaritätsproblem) für ein Integralfunktional auf einem (unendlich dimensionalen) Funktionenraum zuordnen. Statt nun das Integralfunktional zuerst durch eine Quadraturformel zu diskretisieren kann man das

Problem auch dadurch angehen, daß man den Funktionenraum durch einen endlich dimensionalen Unterraum ersetzt. Wir wollen nun diese Vorgehensweise an dem allgemeinen semilinearen Problem

$$\begin{aligned} -(p(x)y'(x))' + q(x)y(x) &= g(x, y(x)), & a \leq x \leq b \\ y(a) = y(b) &= 0 \end{aligned}$$

erläutern. (Bezüglich anderer Randbedingungen, z.B. $y(a) = \alpha$, $y(b) = \beta$ oder $y(a) = \alpha$, $y'(b) = \beta$, vgl. Bem. 2.3.1.)

Die Koeffizienten p, q und g sollen folgende Voraussetzungen erfüllen:

$$\begin{aligned} p &\in C^1[a, b], & p(x) &\geq p_0 > 0. \\ q &\in C[a, b], & q(x) &\geq 0 \\ g &\in C([a, b] \times \mathbb{R}), & \exists \partial_2 g \text{ und } \partial_2 g &\in C([a, b] \times \mathbb{R}) \\ & & \text{und } -C &\leq \partial_2 g(x, u) \leq \lambda_0 - \varepsilon, \end{aligned}$$

wobei λ_0 der kleinste Eigenwert der Aufgabe

$$EWP : \quad -(pz')' - (\lambda - q)z = 0, \quad z(a) = z(b) = 0$$

ist. (Es gilt $\lambda_0 > 0$.) Man weiß dann, daß die RWA eine eindeutige Lösung $y \in C_0^2[a, b]$ besitzt. Dabei ist

$$\begin{aligned} C_0^k[a, b] &= \{f : f \in C^k[a, b], \\ & \quad f(k) \text{ ist in } a \text{ und } b \text{ beschränkt rechts- bzw. linksseitig stetig fortsetzbar und} \\ & \quad f(a) = f(b) = 0\}. \end{aligned}$$

Bemerkung 2.7.1. In der Form $y'' = -f(x, y, y')$ geschrieben lautet die DGL

$$y'' = -\frac{p'}{p}(x)y' + \frac{q}{p}(x)y - \frac{g(x, y)}{p(x)}$$

d.h. in Satz 2.3.1 wäre dann zu setzen

$$\begin{aligned} \partial_3 f(x, u, v) &= -\frac{p'}{p}(x) & B &= \max_{x \in [a, b]} |p'(x)|/p_0, \\ \partial_2 f(x, u, v) &= \frac{q}{p}(x) - \frac{\partial_2 g(x, u)}{p(x)}, & d.h. & \quad \gamma \geq -\lambda_0/p_0 \\ C &= \max_{a \leq z \leq b} |q(x)|/p_0 + \sup_{a \leq x \leq b, u \in \mathbb{R}} |\partial_2 g(x, u)|/p_0. \end{aligned}$$

Durch geeignete Abänderung von g außerhalb des interessierenden Bereichs kann man stets erreichen, daß die Konstante C wohldefiniert ist. \square

Sei nun der Differentialoperator $L : D_L = C_0^2[a, b] \rightarrow C[a, b]$ definiert durch

$$L(v) \stackrel{\text{def}}{=} -(pv')' + qv$$

Wir suchen offenbar eine Lösung $y \in D_L$ der Gleichung $L(v) = g(\cdot, v)$ (eine solche Lösung nennt man "klassische" Lösung, weil ihre Glattheit der Ordnung der Differentialgleichung entspricht).

$C_0^2[a, b]$ ist mit der durch

$$\|u\|_{2,m} \stackrel{\text{def}}{=} \left(\int_a^b \sum_{\nu=0}^m (u^{(\nu)}(x))^2 dx \right)^{\frac{1}{2}} \quad \text{mit } m = 2$$

definierten Norm bekanntlich nicht vollständig. Seine Vervollständigung bezüglich dieser Norm ist der **Hilbertraum** (Sobolev-Raum) $\mathcal{K}_0^2[a, b]$ wo

$$\mathcal{K}_0^s[a, b] = \left\{ f : f \in C_0^{s-1}[a, b], f^{(s-1)} \text{ absolutstetig auf } [a, b] \text{ und } \|f^{(s)}\|_{2,0} < \infty \right\}$$

Bemerkung 2.7.2. *Ist f absolutstetig auf $[a, b]$, dann ist f fast überall differenzierbar auf $[a, b]$. Eine Funktion $f : [a, b] \mapsto \mathbb{R}$ heisst absolut stetig, wenn*

$$\forall \varepsilon > 0 \exists \delta > 0 : \forall n \forall x_i, y_i \in [a, b], 1 \leq i \leq n \text{ mit } \sum_{i=1}^n |x_i - y_i| < \delta \Rightarrow \sum_{i=1}^n |f(x_i) - f(y_i)| < \varepsilon .$$

Auf $L_2[a, b]$ und damit auch auf dem Teilraum

$$D_L = \mathcal{K}_0^2[a, b]$$

ist als Skalarprodukt definiert

$$(u, v) \stackrel{\text{def}}{=} \int_a^b u(x)v(x)dx.$$

Der Differentialoperator L hat bezüglich dieses Skalarprodukts folgende Eigenschaft:

Satz 2.7.1. : L ist ein symmetrischer Operator auf D_L :

$$(u, L(v)) = (L(u), v) \quad \forall u, v \in D_L$$

Beweis: Partielle Integration ergibt

$$\begin{aligned} (u, L(v)) &= \int_a^b u(x) (-p(x)v'(x))' + q(x)v(x) dx & (2.11) \\ &= \underbrace{-u(x)p(x)v'(x)}_{=0} \Big|_a^b + \int_a^b (u'(x)p(x)v'(x) + u(x)q(x)v(x)) dx \end{aligned}$$

Man beachte, daß in allen diesen Überlegungen vorausgesetzt (und benutzt) wird, daß die eingehenden Funktionen die beiden Randbedingungen erfüllen. Die rechte Seite dieses Ausdrucks ist aber symmetrisch in u, v , also

$$(u, L(v)) = (v, L(u)) = (L(u), v). \quad \square$$

Die rechte Seite in (2.11) ist aber offenbar auch auf $\mathcal{K}_0^1[a, b]$ wohldefiniert. Auf $\mathcal{K}_0^1[a, b] \supset \mathcal{K}_0^2[a, b] \supset D_L$ führen wir durch

$$[u, v] \stackrel{\text{def}}{=} \int_a^b (p(x)u'(x)v'(x) + q(x)u(x)v(x))dx$$

eine symmetrische Bilinearform ein. Es ist für $v \in C_0^2([a, b])$ nach Obigem offensichtlich

$$[u, v] = (u, L(v))$$

Diese Bilinearform ist auch positiv definit auf $\mathcal{K}_0^1[a, b]$:

Satz 2.7.2. *Es gilt mit geeigneten Konstanten $\gamma_\infty, \gamma_{2,1}, \Gamma_\infty, \Gamma_{2,1} > 0$*

$$\begin{aligned} \gamma_\infty \|u\|_\infty^2 &\leq [u, u] \leq \Gamma_\infty \|u'\|_\infty^2 & \forall u \in \mathcal{K}_0^2[a, b] \\ \gamma_{2,1} \|u\|_{2,1}^2 &\leq [u, u] \leq \Gamma_{2,1} \|u\|_{2,1}^2 & \forall u \in \mathcal{K}_0^1[a, b] \end{aligned}$$

insbesondere also $[u, u] > 0$ für $u \neq 0 \in \mathcal{K}_0^1[a, b]$.

Beweis: Für $u \in \mathcal{K}_0^1[a, b]$ gilt

$$u(x) = \int_a^x u'(\xi) d\xi.$$

Also nach der Cauchy–Schwarzschen Ungleichung $|(u, v)|^2 \leq \|u\|_2^2 \|v\|_2^2$

$$\begin{aligned} (u(x))^2 &= \left(\int_a^x 1 \cdot u'(\xi) d\xi \right)^2 \leq \int_a^x 1^2 \cdot d\xi \int_a^x (u'(\xi))^2 d\xi \\ &\leq (b-a) \int_a^b (u'(\xi))^2 d\xi \end{aligned}$$

Ist also $u \in \mathcal{K}_0^2[a, b]$, dann

$$\|u\|_\infty^2 \leq (b-a)^2 \|u'\|_\infty^2 \quad \text{und} \quad \|u\|_\infty^2 \leq (b-a) \int_a^b (u'(\xi))^2 d\xi$$

Bemerkung 2.7.3. *Aus der vorausgegangenen Abschätzung folgt auch*

$$\int_a^b u(x)^2 dx \leq (b-a)^2 \int_a^b (u'(x))^2 dx \quad \text{für } u \in \mathcal{K}_0^1[a, b].$$

In dieser Form ist die Abschätzung als Poincaré'sche Ungleichung bekannt.

Für $u \in \mathcal{K}_0^1$ kann man nur bezüglich der L_2 -Norm abschätzen

$$\int_a^b (u(x))^2 dx \leq (b-a)^2 \int_a^b (u'(x))^2 dx.$$

Nun ist

$$\begin{aligned} [u, u] &= \int_a^b (p(x)(u'(x))^2 + q(x)(u(x))^2) dx \\ &\leq (\|p\|_\infty + \|q\|_\infty) \int_a^b ((u'(x))^2 + (u(x))^2) dx \end{aligned}$$

bzw. für $u \in \mathcal{K}_0^2[a, b]$

$$\begin{aligned} [u, u] &\leq (b-a)(\|p\|_\infty + \|q\|_\infty)(\|u'\|_\infty^2 + \|u\|_\infty^2) \\ &\leq (b-a)(1 + (b-a)^2)(\|p\|_\infty + \|q\|_\infty)\|u'\|_\infty^2 \end{aligned}$$

und

$$[u, u] \geq p_0 \int_a^b (u'(x))^2 dx \geq \frac{p_0}{b-a} \|u\|_\infty^2$$

bzw.

$$\begin{aligned} [u, u] &\geq p_0 \int_a^b (u'(x))^2 dx \\ &= \frac{p_0}{1 + (b-a)^2} (1 + (b-a)^2) \int_a^b (u'(x))^2 dx \\ &\geq \frac{p_0}{1 + (b-a)^2} \left(\int_a^b (u'(x))^2 dx + \int_a^b (u(x))^2 dx \right) \end{aligned}$$

d.h.

$$\begin{aligned} \gamma_\infty &= p_0/(b-a), & \Gamma_\infty &= (b-a)(1 + (b-a)^2)(\|p\|_\infty + \|q\|_\infty) \\ \gamma_{2,1} &= p_0/(1 + (b-a)^2), & \Gamma_{2,1} &= \|p\|_\infty + \|q\|_\infty \end{aligned}$$

□

Dieser Satz besagt also, daß die sogenannte

Energienorm $[u, u]^{1/2}$

äquivalent zur Sobolevnorm auf $\mathcal{K}_0^1[a, b]$ ist.

Im Falle $g(x, y) \equiv f(x)$, also bei einer linearen RWA, ergibt sich die Eindeutigkeit einer Lösung daraus unmittelbar über

$$\begin{aligned} L(y_1) = f, \quad L(y_2) = f &\Rightarrow L(y_1 - y_2) = 0 \Rightarrow \\ (y_1 - y_2, L(y_1 - y_2)) = 0 &\geq \gamma_\infty \|y_1 - y_2\|_\infty^2. \end{aligned}$$

Im nichtlinearen Fall erkennt man die Bedeutung der Voraussetzung an $\partial_2 g$ aus dem

Widerspruchsbeweis für die Eindeutigkeit einer Lösung:

$$\begin{aligned}
 \lambda_0(y_1 - y_2, y_1 - y_2) &\leq (y_1 - y_2, L(y_1 - y_2)) \\
 &= (y_1 - y_2, g(\cdot, y_1) - g(\cdot, y_2)) \\
 &= \int_a^b (y_1 - y_2)(x)(g(x, y_1(x)) - g(x, y_2(x))) dx \\
 &= \int_a^b (y_1 - y_2)(x) \underbrace{(\partial_2 g(x, y_1(x)) + \theta(y_1, y_2, x)(y_1(x) - y_2(x)))}_{\text{stetig in } x} \\
 &\quad \cdot (y_1 - y_2)(x) dx \\
 &= \partial_2 g(\tilde{x}, \tilde{y}) \int_a^b (y_1 - y_2)^2(x) dx \\
 &< \lambda_0(y_1 - y_2, y_1 - y_2) \quad \text{falls } y_1 - y_2 \neq 0 \quad \text{Widerspruch!}
 \end{aligned}$$

Man beachte, daß $\lambda_0 = \inf_{z \in D_L} (z, L(z)) / (z, z)$.

Für $u \in \mathcal{K}_0^1[a, b]$ definieren wir nun durch

$$F(u) \stackrel{\text{def}}{=} [u, u] - 2 \int_a^b \underbrace{\int_0^{u(x)} g(x, \eta) d\eta}_{=: h(x, u(x))} dx \quad F : \mathcal{K}_0^1[a, b] \rightarrow \mathbb{R}.$$

ein **nichtlineares Funktional**. Dieses Funktional hat folgende bemerkenswerte Eigenschaft:

Satz 2.7.3. *Sei y Lösung der Randwertaufgabe und $u \in \mathcal{K}_0^1[a, b]$ beliebig. Dann gilt*

$$F(u) > F(y) \quad \text{falls } u \neq y.$$

<<

Beweis:

$$\begin{aligned}
F(u) &= [u, u] - 2 \int_a^b \int_0^{u(x)} g(x, \eta) d\eta dx \\
&= [u, u] - 2[u, y] + [y, y] - [y, y] + 2[u, y] - 2 \int_a^b \int_0^{u(x)} g(x, \eta) d\eta dx \\
&= [u - y, u - y] - [y, y] + 2 \int_a^b \left(u(x)g(x, y(x)) - \int_0^{u(x)} g(x, \eta) d\eta \right) dx \\
&= [u - y, u - y] + F(y) - 2[y, y] + 2 \int_a^b \int_0^{y(x)} g(x, \eta) d\eta dx \\
&\quad + 2 \int_a^b \int_0^{u(x)} (g(x, y(x)) - g(x, \eta)) d\eta dx \\
&= [u - y, u - y] + F(y) - 2 \int_a^b \int_0^{y(x)} g(x, y(x)) d\eta dx + \\
&\quad + 2 \int_a^b \int_0^{y(x)} g(x, \eta) d\eta dx + 2 \int_a^b \int_0^{u(x)} (g(x, y(x)) - g(x, \eta)) d\eta dx \\
&= [u - y, u - y] + F(y) - 2 \int_a^b \int_{y(x)}^{u(x)} \int_{y(x)}^{\eta} \partial_2 g(x, w) dw d\eta dx \\
&\geq F(y) + \lambda_0(u - y, u - y) - 2 \int_a^b \int_{y(x)}^{u(x)} \int_{y(x)}^{\eta} (\partial_2 g(x, w))^+ dw d\eta dx
\end{aligned}$$

In dieser Beweiskette wurde die Definition

$$z^+(x) = \begin{cases} z(x) & \text{falls } z(x) \geq 0 \\ 0 & \text{sonst} \end{cases}$$

benutzt. Weil

$$\max_{x \in [a, b], w \in \text{int}(y(x), u(x))} (\partial_2 g(x, w))^+ \leq \lambda_0 - \varepsilon$$

kann man die Ungleichungskette fortsetzen zu

$$\geq F(y) + \lambda_0(u - y, u - y) - (\lambda_0 - \varepsilon)(u - y, u - y) > F(y) \text{ falls } u \neq y .$$

□

>>

Hier und an anderer Stelle bedeutet $\text{int}(u, v)$ das abgeschlossene Intervall mit den Grenzen u und v .

Bemerkung 2.7.4. Man kann natürlich zur Definition von F jede Funktion $h(x, u)$ mit $\partial_2 h(x, u) \equiv g(x, u)$ nehmen, da dies F nur um eine Konstante ändert. □

Bemerkung 2.7.5. Im linearen Fall $g(x, u) \equiv f(x)$ vereinfacht sich die Definition von F zu $F(u) = [u, u] - 2(f, u)$. □

Das Resultat von Satz 2.7.3 legt nun folgende Vorgehensweise nahe:

Man bestimme eine **Näherung η für y , indem man F auf einem endlichdimensionalen linearen Teilraum von $\mathcal{K}_0^1[a, b]$ minimiert.**

Sei also $S_h \subset \mathcal{K}_0^1[a, b]$ von der Dimension n und $\varphi_1, \dots, \varphi_n$ eine Basis von S_h .

Ferner sei

$$h(x, u) \quad \text{gegeben mit} \quad \partial_2 h(x, u) = g(x, u).$$

Dann lautet unser **Näherungsproblem**: $y(x) \approx \sum_{i=1}^n \gamma_i^* \varphi_i(x) = \eta(x)$:

Bestimme $\gamma_1^*, \dots, \gamma_n^*$ so, daß

$$\begin{aligned} \Psi(\gamma_1^*, \dots, \gamma_n^*) &\stackrel{\text{def}}{=} \int_a^b (p(x) (\sum_{i=1}^n \gamma_i^* \varphi_i'(x))^2 + \\ &\quad + q(x) (\sum_{i=1}^n \gamma_i^* \varphi_i(x))^2 - 2h(x, \sum_{i=1}^n \gamma_i^* \varphi_i(x))) dx \end{aligned}$$

minimiert wird bzgl. $\gamma_1, \dots, \gamma_n$. Dies ist nun ein unrestringiertes Minimierungsproblem in \mathbb{R}^n . (Man beachte, daß alle Elemente von S_h automatisch die Randbedingungen erfüllen). Differentiation von $\Psi(\gamma_1, \dots, \gamma_n)$ bzgl. $\gamma_1, \dots, \gamma_n$ und Nullsetzen des Gradienten liefert nun ein **(nicht)lineares Gleichungssystem** für die γ_i^* :

$$\begin{aligned} \int_a^b (2p(x) (\sum_{i=1}^n \gamma_i^* \varphi_i'(x)) \varphi_j'(x) + 2q(x) (\sum_{i=1}^n \gamma_i^* \varphi_i(x)) \varphi_j(x) \\ - 2g(x, \sum_{i=1}^n \gamma_i^* \varphi_i(x)) \varphi_j(x)) dx = 0, \quad j = 1, \dots, n \end{aligned}$$

oder in Matrix-Schreibweise mit

$$\begin{aligned} A &= ([\varphi_i, \varphi_j])_{1 \leq i, j \leq n} \quad G(c) = \left((\varphi_j, g(\cdot, \sum_{i=1}^n \gamma_i^* \varphi_i))_{j=1}^n \right) \\ Ac &= G(c), \quad c = (\gamma_1, \dots, \gamma_n)^T. \end{aligned}$$

Die Matrix A entsteht also aus den Bilinearformen der Basisfunktionen in S_h . Sie ist automatisch symmetrisch und auch positiv definit. Denn

$$c^T A c = \sum_{i=1}^n \sum_{j=1}^n \gamma_i \gamma_j [\varphi_i, \varphi_j] = [\sum_{i=1}^n \gamma_i \varphi_i, \sum_{j=1}^n \gamma_j \varphi_j] > 0$$

für $c = (\gamma_1, \dots, \gamma_n)^T \neq 0$.

Die Jacobi-Matrix der nichtlinearen Vektorfunktion G hat folgende Gestalt:

$$\mathcal{J}_G(\gamma_1, \dots, \gamma_n) = (\varphi_j, \partial_2 g(\cdot, \sum_{i=1}^n \gamma_i \varphi_i) \varphi_i)_{j,i}$$

Es ist dann die Hessematrix von Ψ

$$H_{\Psi}(\gamma_1, \dots, \gamma_n) = A - \mathcal{J}_G(\gamma_1, \dots, \gamma_n)$$

und

$$\begin{aligned} c^T H_{\Psi}(\gamma_1, \dots, \gamma_n) c &= c^T A c - c^T \mathcal{J}_G(c) c \\ &= \left[\sum_{i=1}^n \gamma_i \varphi_i, \sum_{j=1}^n \gamma_j \varphi_j \right] - \int_a^b \left(\sum_{i=1}^n \gamma_i \varphi_i(x) \right) \cdot \\ &\quad \cdot \left(\sum_{j=1}^n \partial_2 g(x, \sum_{k=1}^n \gamma_k \varphi_k(x)) \gamma_j \varphi_j(x) \right) dx \\ &\geq \int_a^b \left(\sum_{i=1}^n \gamma_i \varphi_i(x) \right) \left(\lambda_0 - \partial_2 g(x, \sum_{k=1}^n \gamma_k \varphi_k(x)) \right) \left(\sum_{j=1}^n \gamma_j \varphi_j(x) \right) dx \\ &\geq \varepsilon \left\| \sum_{j=1}^n \gamma_j \varphi_j(x) \right\|_{2,0}^2 \\ &\geq \delta \sum_{i=1}^n \gamma_i^2 = \delta \|c\|_2^2 \text{ mit einer geeigneten Konstanten } \delta > 0 \end{aligned}$$

d.h. die Hessematrix ist gleichmäßig positiv definit (der kleinste Eigenwert ist durch $\delta > 0$ nach unten abschätzbar unabhängig vom Wert des Parametervektors c). Das nichtlineare Funktional Ψ ist also gleichmäßig konvex, d.h. das formulierte Minimierungsproblem hat eine global eindeutige Lösung. Als endlichdimensionale Unterräume kann man nun sowohl Funktionen mit globalem Träger $[a, b]$ als auch solche mit lokalem Träger (dessen Breite mit $n \rightarrow \infty$ gegen null geht) benutzen. In ersterem Fall spricht man vom klassischen Ritzansatz, im zweiten gelangen wir zur Methode der finiten Elemente. Der klassische Ritzansatz erfordert eine sehr sorgfältige Auswahl der Basisfunktionen, weil sonst die Gleichungssysteme so schlecht konditioniert werden, daß in der Praxis Rundungsfehler das Verfahren zusammenbrechen lassen. Deshalb wird meistens der Ansatz mit finiten Elementen benutzt. In diesem Fall wird die Hessematrix H_{Ψ} (d.h. die Jacobi-Matrix des Gradienten von Ψ):

$$\nabla^2 \Psi(c) = H_{\Psi} = A - \mathcal{J}_G(c)$$

eine Bandmatrix **geringer** Bandbreite (Dreiband-, Siebenband-Matrix), also wird man auch hier zur Lösung die Anwendung des gedämpften Newton-Verfahrens (bzw. im linearen Fall die Lösung mittels der Cholesky-Zerlegung) in Betracht ziehen. Man beachte jedoch, daß \mathcal{J}_D hier **nicht** die speziellen Eigenschaften besitzt, wie sie zum Nachweis der globalen und monotonen Konvergenz des ungedämpften Newtonverfahrens in Abschnitt 3 benutzt wurden.

Mittels der in Satz 2.7.2 geleisteten Normabschätzung kann man nun mühelos zu Konvergenzaussagen gelangen:

Satz 2.7.4. (Lemma von C ea) Sei y die exakte L osung der RWA und $S_h \subset \mathcal{K}_0^1[a, b]$ ein endlichdimensionaler Teilraum. $\eta \in S_h$ sei definiert durch

$$F(\eta) = \min_{u \in S_h} F(u).$$

Die Nichtlinearit at g erf ulle die Voraussetzung

$$-C \leq \partial_2 g(x, u) \leq \lambda_0 - \varepsilon \quad \forall (x, u) \in [a, b] \times \mathbb{R}.$$

Dann gilt mit einer geeigneten Konstanten $C_{2,1}$

$$\|\eta - y\|_{2,1} \leq C_{2,1} \inf_{u \in S_h} \|u' - y'\|_{2,0}$$

und f ur $S_h \subset \mathcal{K}_0^2[a, b]$ sogar

$$\|\eta - y\|_\infty \leq C_\infty \inf_{u \in S_h} \|u' - y'\|_\infty$$

<<

Beweis: Es ist f ur $u \in \mathcal{K}_0^1$ beliebig

$$F(u) = F(y) + [u - y, u - y] - 2 \int_a^b \int_{y(x)}^{u(x)} \int_{y(x)}^w \partial_2 g(x, z) dz dw dx.$$

Also wegen $F(u) \geq F(\eta)$ f ur $u \in S_h$

$$\begin{aligned} F(u) - F(y) \geq F(\eta) - F(y) &\geq [\eta - y, \eta - y] - (\lambda_0 - \varepsilon)(\eta - y, \eta - y) \\ &\geq \left(1 - \frac{\lambda_0 - \varepsilon}{\lambda_0}\right) [\eta - y, \eta - y] \\ &\geq (\varepsilon \gamma_{2,1} / \lambda_0) \|\eta - y\|_{2,1}^2. \end{aligned}$$

Andererseits ist mit

$$\begin{aligned} F(u) - F(y) &\leq \Gamma_{2,1} \|u - y\|_{2,1}^2 + C \|u - y\|_{2,0}^2 \\ &\leq \Gamma_{2,1} (1 + (b - a)^2) \|u' - y'\|_{2,0}^2 + C (b - a)^2 \|u' - y'\|_{2,0}^2 \end{aligned}$$

d.h.

$$\|\eta - y\|_{2,1}^2 \leq \frac{\lambda_0}{\varepsilon \gamma_{2,1}} (\Gamma_{2,1} + (\Gamma_{2,1} + C)(b - a)^2) \|u' - y'\|_{2,0}^2. \quad (\forall u \in S_h)$$

Analog ergibt sich f ur $S_h \subset \mathcal{K}_0^2[a, b]$ und $u \in S_h$

$$F(u) - F(y) \geq F(\eta) - F(y) \geq (\varepsilon / \lambda_0) [\eta - y, \eta - y] \geq -(\varepsilon / \lambda_0) \gamma_\infty \|\eta - y\|_\infty^2$$

und

$$\begin{aligned} F(u) - F(y) &\leq \Gamma_\infty \|u' - y'\|_\infty^2 + C \int_a^b (y(x) - u(x))^2 dx \\ &\leq \Gamma_\infty \|u' - y'\|_\infty^2 + C (b - a)^2 \int_a^b (y'(x) - u'(x))^2 dx \\ &\leq (\Gamma_\infty + C (b - a)^3) \|u' - y'\|_\infty^2. \end{aligned}$$

Also

$$\|\eta - y\|_\infty^2 \leq \frac{\lambda_0}{\varepsilon \gamma_\infty} (\Gamma_\infty + C(b-a)^3) \|u' - y'\|_\infty^2 \quad \forall u \in S_h \subset \mathcal{K}_0^2[a, b].$$

□

>>

Der Satz besagt also, daß die konstruierte Approximation eine Approximationsgüte besitzt, die der Bestapproximation von y' durch Ableitungen der Funktionen aus dem Ansatzraum entspricht, d.h. diese Approximationsgüte ist nur abhängig von der Glätte der Lösung bei geeigneter Wahl des Approximationsraumes.

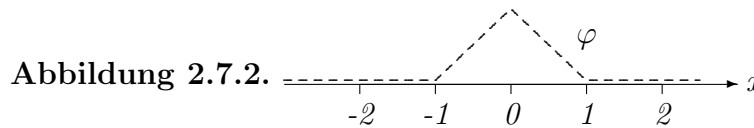
Beispiele:

(Man beachte die Randbedingungen $u(a) = u(b) = 0$ für $u \in S_h$)

Ansatzfunktionen mit kompaktem Träger:

Finite-Elemente-Räume

1. $h = \frac{b-a}{N}$, $x_i \stackrel{\text{def}}{=} a + ih$, $i = 0, \dots, N$, $n = N - 1$
 $\varphi_i(x) = \varphi\left(\frac{x-x_i}{h}\right)$ $1 \leq i \leq N - 1$



$$\varphi(x) = \begin{cases} 0 & x < -1 \\ x + 1 & -1 \leq x \leq 0 \\ 1 - x & 0 \leq x \leq 1 \\ 0 & x > 1 \end{cases}$$

$S_h \subset \mathcal{K}_0^1[a, b]$. **stückweise lineare Interpolation**

Es ist (siehe Einführung in die Numerische Mathematik I)

$$\inf_{u \in S_h} \|u' - y'\|_{2,0} \leq \frac{1}{2} h M_2 \sqrt{b-a}$$

mit $M_2 \stackrel{\text{def}}{=} \max_{x \in [a,b]} |y''(x)|$.

$H_\Psi(\gamma_1, \dots, \gamma_n)$ **wird hier zu einer Dreibandmatrix**, im linearen Fall mit $y'' = f(x)$ identisch mit der Matrix aus dem Differenzenverfahren!

Beispiel 2.7.2. *Wir betrachten die Randwertaufgabe*

$$y'' + y = -x, \quad y(0) = y(1) = 0.$$

und stellen das Gleichungssystem für die Finite Element Methode (FEM) unter Benutzung der linearen Elemente

$$\phi_i(x) = \begin{cases} \frac{1}{h}(x - x_{i-1}) & x \in [x_{i-1}, x_i] \\ \frac{1}{h}(x_{i+1} - x) & x \in [x_i, x_{i+1}] \\ 0 & \text{sonst} \end{cases}$$

und äquidistanter Stützstellen auf. Um die schwache Form der Ausgangsgleichung zu bekommen, multiplizieren wir sie mit einer Testfunktion $\phi(x)$ und integrieren über das gegebene Intervall. Wir bekommen

$$\int_0^1 -y'(x)\phi'(x) + y(x)\phi(x) dx = - \int_0^1 x\phi(x) dx.$$

Mit dem Ansatz

$$u_h(x) = \sum_{j=1}^n \gamma_j \phi_j(x)$$

führt dies auf ein lineares Gleichungssystem der Form $A\gamma = b$ mit

$$a_{ij} = \int_0^1 -\phi_j'(x)\phi_i'(x) + \phi_j(x)\phi_i(x) dx, \quad b_i = - \int_0^1 x\phi_i(x) dx.$$

Aufgrund der Lokalität der $\phi_i(x)$ gilt $a_{ij} = 0$ für $|i - j| \geq 2$. Es bleiben drei Integrale zu berechnen:

$$a_{ii} = \int_{x_{i-1}}^{x_i} -\frac{1}{h^2} + \frac{1}{h^2}(x - x_{i-1})^2 dx + \int_{x_i}^{x_{i+1}} -\frac{1}{h^2} + \frac{1}{h^2}(x_{i+1} - x)^2 dx = \frac{2}{3}h - \frac{2}{h}$$

$$a_{i,i-1} = \int_{x_{i-1}}^{x_i} \frac{1}{h^2} + \frac{1}{h^2}(x_i - x)(x - x_{i-1}) dx = \frac{1}{6}h + \frac{1}{h}$$

$$b_i = \int_{x_{i-1}}^{x_i} -\frac{x}{h}(x - x_{i-1}) dx + \int_{x_i}^{x_{i+1}} -\frac{x}{h}(x_{i+1} - x) dx = -hx_i = -ih^2.$$

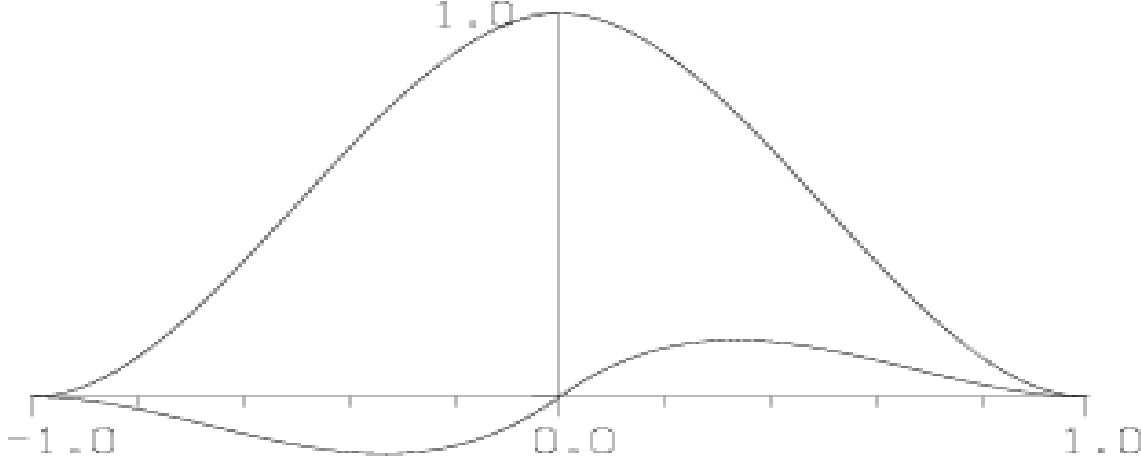
Wegen der Symmetrie gilt $a_{i,i-1} = a_{i,i+1}$. Das zusammengesetzte LGS lautet

$$-\frac{1}{h} \underbrace{\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & \end{pmatrix}}_{\text{Steifigkeitsmatrix}} + \frac{h}{6} \underbrace{\begin{pmatrix} 4 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 4 & 1 & \\ & & & 1 & 4 & \end{pmatrix}}_{\text{Massematrix}} = -h^2 \begin{pmatrix} 1 \\ 2 \\ \vdots \\ \vdots \\ n \end{pmatrix}.$$

2. $h = \frac{b-a}{N}$, $x_i \stackrel{\text{def}}{=} a + ih$, $i = 0, \dots, N$, $n = 2N$

$$\begin{aligned} \varphi_{2i}(x) &= \psi\left(\frac{x - x_i}{h}\right) & 1 \leq i \leq N - 1 \\ \varphi_{2i+1}(x) &= \tilde{\psi}\left(\frac{x - x_i}{h}\right) & 0 \leq i \leq N - 1 \\ \varphi_{2N}(x) &= \tilde{\psi}\left(\frac{x - b}{h}\right) \end{aligned}$$

Abbildung 2.7.3.



mit

$$\psi(x) = \begin{cases} (1-2x)(1+x)^2 & -1 \leq x \leq 0 \\ (1+2x)(1-x)^2 & 0 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

$$\tilde{\psi}(x) = \begin{cases} x(1+x)^2 & -1 \leq x \leq 0 \\ x(1-x)^2 & 0 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

(stückweise kubische Hermite-Interpolation), $S_h \subset \mathcal{K}_0^2[a, b]$.

$H_\Psi(\gamma_1, \dots, \gamma_n)$ ist bei der angegebenen Numerierung eine 6-Band-Matrix.

Ist $p \in C^3[a, b]$, $q \in C^2[a, b]$, $g \in C^2([a, b] \times \mathbb{R})$, dann ist $y \in C_0^4[a, b]$.

In diesem Fall gilt für die durch

$$u_I(x_i) = y(x_i), \quad u'_I(x_i) = y'(x_i), \quad 0 \leq i \leq N$$

definierte stückweise kubische Hermite-Interpolierende u von y

$$\|y' - u'_I\|_\infty \leq \frac{\sqrt{3}}{216} h^3 \|y^{(4)}\|_\infty,$$

(vgl. bei M.H. Schulz, Spline Analysis, Th. 3.6.), so daß wir aus Satz 2.7.4 direkt die Abschätzung

$$\|y - \eta\|_\infty \leq C_\infty \frac{\sqrt{3}}{216} h^3 \|y^{(4)}\|_\infty$$

erhalten.

Bemerkung 2.7.6. Für die Norm $\|\cdot\|_{2,0}$ ergibt sich eine analoge Abschätzung, nämlich

$$\|u'_I - y'\|_{2,0} \leq 4 \frac{h^3}{\pi^3} \|y^{(4)}\|_{2,0},$$

vgl. bei von Finckenstein, Einführung in die Numerische Mathematik, Band I

□

3. $h = \frac{b-a}{N}$, $x_i \stackrel{\text{def}}{=} a + ih$, $i = 0, \dots, N$, $n = N + 1$

$$\begin{aligned} \hat{\varphi}_i(x) &= \psi\left(\frac{x - x_i}{h}\right), \quad -1 \leq i \leq N + 1, \\ \varphi_{i+1} &= \hat{\varphi}_i, \quad 2 \leq i \leq N - 2 \\ \varphi_1 &= \hat{\varphi}_0 - 4\hat{\varphi}_{-1}, \quad \varphi_2 = \hat{\varphi}_1 - \hat{\varphi}_{-1} \\ \varphi_N &= \hat{\varphi}_{N-1} - \hat{\varphi}_{N+1}, \quad \varphi_{N+1} = \hat{\varphi}_N - 4\hat{\varphi}_{N+1} \end{aligned}$$

mit

$$\psi(x) = \begin{cases} \psi_1(x) & \text{für } -2 \leq x \leq -1 & \psi_1(x) = \frac{1}{6}[x^3 + 6x^2 + 12x + 8] \\ \psi_2(x) & \text{für } -1 \leq x \leq 0 & \psi_2(x) = \frac{1}{6}[-3x^3 - 6x^2 + 4] \\ \psi_3(x) & \text{für } 0 \leq x \leq 1 & \psi_3(x) = \frac{1}{6}[3x^3 - 6x^2 + 4] \\ \psi_4(x) & \text{für } 1 \leq x \leq 2 & \psi_4(x) = \frac{1}{6}[-x^3 + 6x^2 - 12x + 8] \\ 0 & \text{sonst} \end{cases}$$

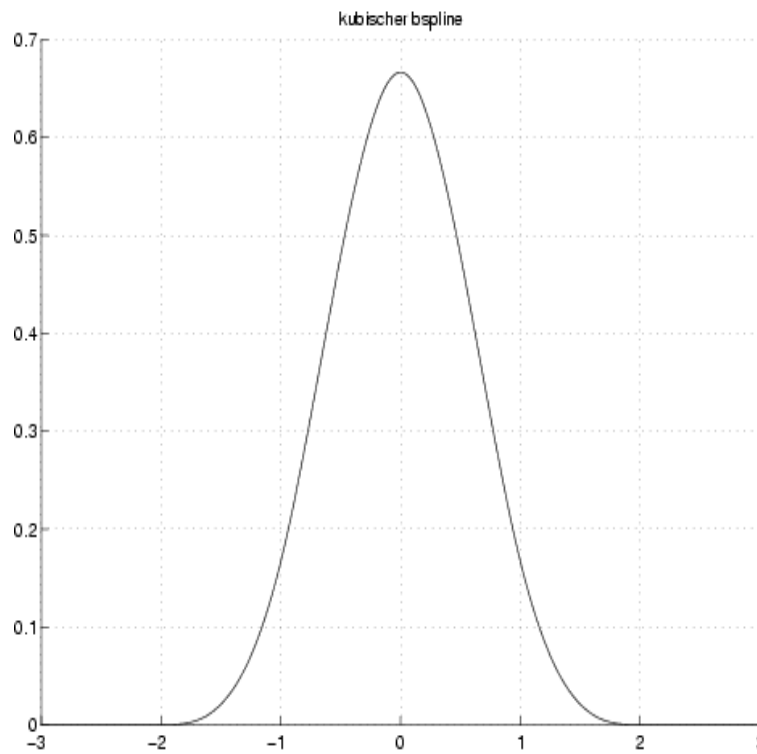


Abbildung 2.7.4.

(kubische Splinefunktion zur Zerlegung

$a = x_0 < x_1 < \dots < x_{N-1} < x_N = b$)

$S_h \subset \mathcal{K}_0^3$. $H_\Psi(\gamma_1, \dots, \gamma_n)$ ist hier eine Siebenbandmatrix. Aufgrund der bekannten Fehlerabschätzung für den durch

$$\begin{aligned} u'_I(a) &= y'(a) & u_I(a) &= 0 = y(a) \\ u'_I(b) &= y'(b) & u_I(b) &= 0 = y(b) \\ u_I(x_i) &= y(x_i) & 1 \leq i &\leq N - 1 \end{aligned}$$

definierten hermiteschen kubischen Spline erhalten wir sofort

$$\|u'_I - y'\|_\infty \leq 2\|y^{(4)}\|_\infty h^3.$$

Die so hergeleiteten Fehlerabschätzungen haben den Nachteil, daß ihre h - Ordnung um 1 kleiner ist, als man eigentlich erwarten würde. (Man beachte aber, daß sie sich auf die Sobolevnorm beziehen, also auch auf die Approxiamtionsgüte für die erste Ableitung!) Bei einer Abschätzung von $\|\eta - y\|_{2,0}$ kann man tatsächlich noch einen Faktor h gewinnen, und zwar mit Hilfe des sogenannten **Nitsche-Tricks** (Nitsche 1968) Dies führen wir nun im Folgenden durch: Die Approximationslösung η ist definiert durch das (nicht-)lineare Gleichungssystem

$$[\eta, \varphi_j] = (g(\cdot, \eta), \varphi_j) \quad j = 1, \dots, n,$$

d.h. (da $\varphi_1, \dots, \varphi_n$ Basis von S_h) es ist

$$[\eta, v] = (g(\cdot, \eta), v) \quad \forall v \in S_h. \quad (2.12)$$

Ist y Lösung der RWA, d.h.

$$L(y) = g(\cdot, y),$$

dann ist trivialerweise

$$[y, v] = (L(y), v) = (g(\cdot, y), v) \quad \forall v \in \mathcal{K}_0^1[a, b] \quad (\supset S_h) \quad (2.13)$$

Bemerkung 2.7.7. Jede Lösung $y \in \mathcal{K}_0^1[a, b]$ von (2.13) nennt man eine **schwache Lösung der RWA**. Man kann zeigen, daß unter den hier vorliegenden Voraussetzungen die schwache Lösung eindeutig bestimmt ist und eine Lösung im klassischen Sinn ist (d.h. $\in C^2[a, b]$).

Falls die Inhomogenität $g \in L_2[a, b]$ erfüllt, dann ist auch noch $y'' \in L_2[a, b]$. Den Ansatz (2.12) nennt man **Galerkin-Ansatz**, und die Approximation η eine **Galerkin-Approximation** für y . Man kann diesen Ansatz auch dann formulieren, wenn **L nicht symmetrisch und positiv definit ist**. In unserem Fall ist (2.12) aus der **Minimierungsforderung** für das nichtlineare Funktional F auf einem endlichdimensionalen Teilraum S_h entstanden. In diesem Fall nennt man η eine **Rayleigh-Ritz-Galerkin-Approximation** für η . Falls die Inhomogenität g von y unabhängig ist, wird

$$[y - \eta, v] = 0 \quad \forall v \in S_h,$$

d.h. der Approximationsfehler ist orthogonal zu S_h (bzgl. des Skalarproduktes $[\cdot, \cdot]$), d.h. η ist die $[\cdot, \cdot]$ -orthogonale Projektion von y auf S_h . Man spricht deshalb in diesem Zusammenhang auch von **Projektionsverfahren**. \square

Aus (2.12) und (2.13) folgt nun

$$\begin{aligned} [y - \eta, v] &= (g(\cdot, y) - g(\cdot, \eta), v) \quad \forall v \in S_h, \\ &= \underbrace{\int_a^b \int_0^1 \partial_2 g(x, \eta(x) + \tau(y(x) - \eta(x))) d\tau (y(x) - \eta(x)) v(x) dx}_{r(x) \text{ stetig in } x}. \end{aligned}$$

Sei

$$\psi(x) \stackrel{\text{def}}{=} \frac{y(x) - \eta(x)}{\|y - \eta\|_{2,0}}, \quad (\text{d.h. } \|\psi\|_{2,0} = 1)$$

wobei η eine der oben definierten Rayleigh–Ritz–Galerkin Approximationen für y ist. In jedem Fall ist $\psi \in \mathcal{K}_0^1[a, b]$.

Wir definieren ϕ als eindeutige Lösung des Problems

$$[\phi, v] - (r\phi, v) = (\psi, v), \quad \phi \in \mathcal{K}_0^1[a, b], \quad \forall v \in \mathcal{K}_0^1[a, b]$$

also als schwache Lösung von $L(\phi) - r\phi = \psi$, $\phi(a) = \phi(b) = 0$.

Mit $v \stackrel{def}{=} \psi$ folgt

$$[\phi, \psi] - (r\phi, \psi) = 1.$$

Das Skalarprodukt

$$b(u, v) \stackrel{def}{=} [u, v] - (ru, v)$$

ist auf $\mathcal{K}_0^1[a, b] \times \mathcal{K}_0^1[a, b]$ ebenfalls symmetrisch und positiv definit:

$$\begin{aligned} [u, u] - (ru, u) &\geq [u, u] - (\lambda_0 - \varepsilon)(u, u) \\ &\geq [u, u] - (\lambda_0 - \varepsilon)/\lambda_0 [u, u] = \varepsilon/\lambda_0 [u, u] \\ &\geq \varepsilon/\lambda_0 \gamma_{2,1} \|u\|_{2,1}^2 \geq \varepsilon/\lambda_0 \gamma_{2,1} \|u'\|_{2,0}^2 \\ [u, u] - (ru, u) &\leq \Gamma_{2,1} \|u\|_{2,1}^2 + C \|u\|_{2,0}^2 \\ &\leq (\Gamma_{2,1}(1 + (b-a)^2) + C(b-a)^2) \|u'\|_{2,0}^2 \end{aligned}$$

Nun gilt

$$b(\phi, \phi) = (\phi, \psi) \leq \|\phi\|_{2,0} \underbrace{\|\psi\|_{2,0}}_{=1} \leq \sqrt{b-a} \|\phi'\|_{2,0}$$

und

$$b(\phi, \phi) \geq \varepsilon/\lambda_0 \gamma_{2,1} \|\phi'\|_{2,0}^2$$

d.h. $\|\phi'\|_{2,0} \leq \frac{\lambda_0 \sqrt{b-a}}{\varepsilon \cdot \gamma_{2,1}}$ und wegen der Poincaré-Abschätzung $\|\phi\|_{2,0} \leq \frac{(b-a)\lambda_0}{\varepsilon \cdot \gamma_{2,1}}$

Damit kann man schließlich zeigen, daß sogar gilt

$$\phi \in \mathcal{K}_0^2[a, b], \quad \|\phi''\|_{2,0} \leq C^*,$$

wobei C^* unabhängig von y, h ist. (vgl. von Finckenstein, Einführung in die Numerische Mathematik, Bd II, Satz 6.1).

Nun haben wir

$$\begin{aligned} [y - \eta, v] - (r(y - \eta), v) &= 0 \quad \forall v \in S_h \\ [\psi, \phi] - (r\psi, \phi) &= (\psi, \psi) = 1 \quad | * \|y - \eta\|_{2,0} \end{aligned}$$

also

$$[y - \eta, \phi - v] - (r(y - \eta), \phi - v) = b(y - \eta, \phi - v) = \|y - \eta\|_{2,0}$$

d.h.

$$\|y - \eta\|_{2,0} \leq (\Gamma_{2,1}(1 + (b-a)^2) + C(b-a)^2) \|y' - \eta'\|_{2,0} \|\phi' - v'\|_{2,0} \quad \forall v \in S_h.$$

Wegen $\phi \in \mathcal{K}_0^2[a, b]$ aber ist (vgl. von Finckenstein, Einführung in die Numerische Mathematik, Bd I, S6.2)

$$\|\phi' - v'\|_{2,0} \leq Ch\|\phi''\|_{2,0} \leq CC^*h$$

für $v \in S_h$ **geeignet** im Falle der stückweise linearen Interpolation. Die gleiche Abschätzung gilt auch für die Hermite- und die Spline-Elemente (vgl. bei M.H. Schulz, Spline-Analysis, Th. 3.4 und Th. 4.5).

Wegen

$$\|\phi^{(4)}\|_{2,0} \leq \frac{\tilde{C}\|y'' - \eta''\|_{2,0}}{\|y - \eta\|_{2,0}}$$

bringt die Ausnutzung eventuell vorhandener höherer Regularität von ϕ keine Verbesserung mehr. Dies ergibt schliesslich

Satz 2.7.5. *Sei y die Lösung des Randwertproblems und η_L bzw. η_S bzw. η_H die Rayleigh-Ritz-Galerkin Näherung für y im Raum der stückweise linearen stetigen bzw. stückweise kubischen zweimal stetig differenzierbaren bzw. stückweise kubischen einmal stetig differenzierbaren Funktionen.*

Dann gilt, falls $p \in C^3[a, b]$, $q \in C^2[a, b]$, $g \in C^2([a, b] \times \mathbb{R})$:

$$\begin{aligned} \|y - \eta_L\|_{2,0} &\leq C_1 h^2 \tilde{M}_2 & C_i \text{ unabh. von } y \\ \|y - \eta_S\|_{2,0} &\leq C_2 h^4 \tilde{M}_4 & \tilde{M}_j \stackrel{\text{def}}{=} \|y^{(j)}\|_{2,0} \\ \|y - \eta_H\|_{2,0} &\leq C_3 h^4 \tilde{M}_4 \end{aligned}$$

□

2.8 Sturm Liouville'sche Eigenwertprobleme (ERG)

Im Zusammenhang mit Schwingungsproblemen treten Eigenwertprobleme gewöhnlicher Differentialgleichungen folgender Gestalt auf

$$\left. \begin{aligned} (Ly)(x) &\equiv -(p(x)y'(x))' + q(x)y(x) = \lambda y(x), & a \leq x \leq b, \\ y(a) &= y(b) = 0. \end{aligned} \right\} \quad (2.14)$$

(Bei der Formulierung der Voraussetzungen an das semilineare Randwertproblem in Abschnitt 2.6 ist uns bereits dieses Modell begegnet).

Im Folgenden gelte:

$$\left. \begin{aligned} p(x) &\geq p_0 > 0 & \forall x \in [a, b], & q(x) &\geq 0 & \forall x \in [a, b] \\ p &\in C^3[a, b], & q &\in C^2[a, b]. \end{aligned} \right\} \quad (2.15)$$

Dann gilt:

Satz 2.8.1. *Unter den Voraussetzungen (2.15) besitzt das Problem (2.14) abzählbar unendlich viele Eigenwerte*

$$0 < \lambda_0 < \lambda_1 < \dots$$

die keinen endlichen Häufungswert besitzen. Zu jedem Eigenwert λ_i gehört eine bis auf Normierung eindeutige Eigenfunktion $y_{[i]} \in C_0^4[a, b]$ mit genau i einfachen Nullstellen in $]a, b[$, $i = 0, 1, 2, \dots$. Die $y_{[i]}$ erfüllen bei der Normierung

$$\int_a^b y_{[i]}^2(x) dx = 1$$

die Orthogonalitätsrelation

$$\int_a^b y_{[i]}(x)y_{[j]}(x) dx = (y_{[i]}, y_{[j]}) = \delta_{ij}$$

und bilden ein vollständiges Orthogonalsystem in $L_2[a, b]$. Ferner gilt

$$[y, y] = (Ly, y) \geq \lambda_0(y, y) \quad \forall y \in \mathcal{K}_0^1[a, b].$$

Beweis: siehe z.B. bei Courant & Hilbert. Mathematische Methoden der Physik Bd. 1, V §3, §14. □

Die Eigenwerte und Eigenfunktionen sollen nun ebenfalls durch ein Diskretisierungsverfahren angenähert werden. Wir beschränken uns dabei auf die einfachste Vorgehensweise.

Sei dazu

$$h = \frac{b-a}{N+1}, \quad x_j = a + jh \quad 0 \leq j \leq N+1.$$

Es gilt

$$\begin{aligned} \tilde{L}_h(y_j) &\stackrel{def}{=} -\frac{1}{h}\left(p\left(x_j + \frac{h}{2}\right)\frac{y(x_{j+1}) - y(x_j)}{h} - p\left(x_j - \frac{h}{2}\right)\frac{y(x_j) - y(x_{j-1}))}{h}\right) \\ &= -\frac{1}{h}\left(\left(p(x_j) + \frac{h}{2}p'(x_j) + \frac{h^2}{8}p''(x_j) + \frac{h^3}{48}p'''(\xi_j)\right)(y'(x_j) + \frac{h}{2}y''(x_j) + \right. \\ &\quad \left. + \frac{h^2}{6}y'''(x_j) + \frac{h^3}{24}y^{(4)}(\Theta_j)) - \left(p(x_j) - \frac{h}{2}p'(x_j) + \frac{h^2}{8}p''(x_j) - \right. \right. \\ &\quad \left. \left. - \frac{h^3}{48}p'''(\tilde{\xi}_j)\right)(y'(x_j) - \frac{h}{2}y''(x_j) + \frac{h^2}{6}y'''(x_j) - \frac{h^3}{24}y^{(4)}(\tilde{\Theta}_j))\right) \\ &= -p(x_j)y''(x_j) - \frac{h^2}{12}p(x_j)y^{(4)}(\Theta_j^*) - p'(x_j)y'(x_j) - \frac{h^2}{6}p'(x_j)y'''(x_j) + \mathcal{O}(h^3) \\ &\quad - \frac{h^2}{8}p''(x_j)y''(x_j) - \mathcal{O}(h^3) - \frac{h^3}{48}p'''(\xi_j)y'(x_j) - \frac{h^2}{48}p'''(\tilde{\xi}_j)y'(x_j) \\ &= -(p(x_j)y'(x_j))'_{x=x_j} + h^2\tau_j \end{aligned}$$

Wir definieren nun als Näherungslösung η für y und $\tilde{\lambda}$ für λ ein durch

$$-\frac{1}{h}\left(p\left(x_j + \frac{h}{2}\right)\frac{\eta_{j+1} - \eta_j}{h} - p\left(x_j - \frac{h}{2}\right)\frac{\eta_j - \eta_{j-1}}{h}\right) + q(x_j)\eta_j - \tilde{\lambda}\eta_j = 0,$$

$$j = 1, \dots, N \quad \eta_0 = \eta_{N+1} = 0$$

gegebenes Eigenwert- und Eigenvektorpaar. (vgl. Fußnote).

Es ist aber

$$-\frac{1}{h}\left(p\left(x_j + \frac{h}{2}\right)\frac{y_{j+1} - y_j}{h} - p\left(x_j - \frac{h}{2}\right)\frac{y_j - y_{j-1}}{h}\right) + q(x_j)y_j - \lambda y_j = -h^2\tau_j$$

Sei $\vec{\eta} := (\eta_1, \dots, \eta_N)$, und A die irreduzibel diagonaldominante, tridiagonale, **symmetrische** L -Matrix ¹

$$A := \begin{pmatrix} 0, \dots, 0 & -p\left(x_j - \frac{h}{2}\right), p\left(x_j - \frac{h}{2}\right) + p\left(x_j + \frac{h}{2}\right) + h^2q(x_j), -p\left(x_j + \frac{h}{2}\right), & 0, \dots, 0 \end{pmatrix}$$

Dann können wir schreiben

$$\begin{aligned} (A - \tilde{\lambda}h^2I) \vec{\eta} &= 0 & (h^2\tilde{\lambda} \text{ Eigenwert von } A) \\ (A - \lambda h^2I) \vec{y} &= -h^4 \vec{\tau} \end{aligned}$$

$$\text{mit } \vec{y} = (y_1, \dots, y_N)^T, \quad \vec{\tau} = (\tau_1, \dots, \tau_N)^T.$$

Sei nun λ kein Eigenwert von A . Dann wird

$$\vec{y} = -(A - \lambda h^2I)^{-1} h^4 \vec{\tau}$$

und daher

$$\|\vec{y}\|_2 \leq \underbrace{\|(A - \lambda h^2I)^{-1}\|_2}_{\text{symmetrisch}} h^4 \|\vec{\tau}\|_2 = \frac{1}{\min_{\substack{\text{EW von } A \\ h^2\tilde{\lambda} \\ -h^2\lambda}} |h^2\tilde{\lambda} - h^2\lambda|} h^4 \|\vec{\tau}\|_2$$

d.h.

$$\min |\tilde{\lambda} - \lambda| \leq h^2 \frac{\|\vec{\tau}\|_2}{\|\vec{y}\|_2}.$$

$\tilde{\lambda}$ EW von $\frac{1}{h^2}A$.

Wenn λ selbst Eigenwert von $\frac{1}{h^2}A$ ist, gilt dies erst recht. Weiterhin ist

$$\sqrt{h} \|\vec{\tau}\|_2 = \sqrt{h} \left(\sum_{i=1}^N \tau_i^2 \right)^{1/2} \leq \sqrt{h} \sqrt{N} \|\vec{\tau}\|_\infty \leq \sqrt{b-a} \|\vec{\tau}\|_\infty$$

¹Wegen $p(x) \geq p_0 > 0$ besitzt A **nur einfache Eigenwerte**. A ist positiv definit und eine M -Matrix.

und

$$\begin{aligned}
h \|\vec{y}\|_2^2 &= \left(h \underbrace{\left(\frac{1}{2} y_0^2 \right)}_{=0} + \sum_{i=1}^N y_i^2 + \frac{1}{2} \underbrace{y_{N+1}^2}_{=0} \right) \\
&= \int_a^b y^2(x) dx + \frac{1}{12} h^2 (b-a) y''(\xi) \\
&= 1 + \frac{1}{12} h^2 (b-a) y''(\xi), \\
&(\geq c > 0, \quad \text{falls } h \text{ hinreichend klein})
\end{aligned}$$

d.h. es folgt

$$|\tilde{\lambda}_i - \lambda_i| \leq \frac{\sqrt{b-a}}{\sqrt{1 + \frac{b-a}{12} h^2 y''(\xi)}} h^2 C(p, p', p'', p''', y'_{[i]}, \dots, y_{[i]}^{(4)})$$

Im Folgenden schreiben wir einfach C_i für die Konstante auf der rechten Seite dieser Abschätzung.

<<

Wir wenden uns nun der Abschätzung des Fehlers in den Eigenvektoren zu. Wir haben bereits bewiesen, daß

$$h \|\vec{y}_{[i]}\|_2^2 = 1 + \frac{1}{12} h^2 (b-a) y''_{[i]}(\xi).$$

Es ist nun

$$\begin{aligned}
(A - \tilde{\lambda} h^2 I) \sqrt{h} \vec{\eta} &= 0 & (2.16) \\
(A - \tilde{\lambda} h^2 I) \sqrt{h} \vec{y}_{[i]} &= -h^4 \sqrt{h} \vec{\tau} + (\lambda_i - \tilde{\lambda}) h^2 \sqrt{h} \vec{y}_{[i]} \\
&= -h^4 \sqrt{h} \vec{\tau} + \Theta h^4 \sqrt{b-a} C_i \frac{\sqrt{h} \vec{y}_{[i]}}{\sqrt{h} \|\vec{y}_{[i]}\|_2} \quad \text{mit } \Theta \in [-1, 1]
\end{aligned}$$

Also

$$(A - \tilde{\lambda} h^2 I) \sqrt{h} (\vec{\eta} - \vec{y}_{[i]}) = +h^4 \sqrt{h} \vec{\tau} - \Theta h^4 \sqrt{b-a} C_i \frac{\sqrt{h} \vec{y}_{[i]}}{\sqrt{h} \|\vec{y}_{[i]}\|_2}$$

Sei

$$\vec{\eta}_1, \dots, \vec{\eta}_N, \quad \text{o.B.d.A.} \quad \vec{\eta} = \vec{\eta}_i, \quad \tilde{\lambda} = \tilde{\lambda}_i$$

ein vollständiges orthogonales Eigenvektorsystem von $\frac{1}{h^2} A$ mit der Normierung $h \|\vec{\eta}_k\|_2^2 = 1$ und den zugehörigen Eigenwerten $\tilde{\lambda}_k$. Wir entwickeln $\vec{y}_{[i]}$ nach den $\vec{\eta}_{[k]}$:

$$\vec{y}_{[i]} = \sum_{k=1}^N \gamma_k \vec{\eta}_{[k]}.$$

Dann

$$h \sum_{k=1}^N \gamma_k^2 = 1 + \frac{1}{12} h^2 (b-a) y''_{[i]}(\xi).$$

Aus (2.16) folgt

$$- \sum_{\substack{k=1 \\ k \neq i}}^N (\tilde{\lambda}_k - \tilde{\lambda}_i) \gamma_k \sqrt{h} \vec{\eta}_{[k]} = h^2 (\sqrt{h} \vec{\tau} - \Theta \sqrt{b-a} C_i \frac{\sqrt{h} \vec{y}_{[i]}}{\sqrt{h} \|\vec{y}_{[i]}\|_2})$$

Also

$$\sum_{\substack{k=1 \\ k \neq i}}^N |\tilde{\lambda}_k - \tilde{\lambda}_i|^2 \gamma_k^2 \leq 4h^4 (b-a) \|C_i\|^2$$

Daher

$$\sum_{\substack{k=1 \\ k \neq i}}^N \gamma_k^2 \leq \frac{4(b-a)C_i^2}{\min_{k \neq i} |\tilde{\lambda}_k - \tilde{\lambda}_i|^2} h^4$$

$$0 < \gamma_i^2 = 1 + \left(\frac{1}{12} h^2 (b-a) \|y''_{[i]}\|_\infty + \frac{4(b-a)}{\min_{k \neq i} |\tilde{\lambda}_k - \tilde{\lambda}_i|^2} C_i^2 h^4 \right) \Theta,$$

$\Theta \in [-1, 1]$ für hinreichend kleines h .

Da das Vorzeichen von $y_{[i]}$ frei gewählt werden kann, kann man o.B.d.A. annehmen, daß

$$\begin{aligned} \gamma_i &= 1 + \Theta \tilde{C}_i(h) h^2, & \Theta &\in [-1, 1], \\ \tilde{C}_i(h) &= \frac{1}{12} (b-a) \|y''_{[i]}\|_\infty + \frac{4(b-a)}{\min_{k \neq i} |\tilde{\lambda}_k - \tilde{\lambda}_i|^2} C_i^2 h^2 \\ &\leq \frac{1}{6} (b-a) \|y''_{[i]}\|_\infty && \text{für } h \leq h_0. \end{aligned}$$

Somit

$$\begin{aligned} h \|\vec{y}_{[i]} - \vec{\eta}\|_2^2 &= h \sum_{\substack{k=1 \\ k \neq i}}^n \gamma_k^2 + \tilde{\Theta} \tilde{C}_i^2 h^4 \\ &\leq \left(\frac{4(b-a)C_i^2}{\min_{k \neq i} |\tilde{\lambda}_k - \tilde{\lambda}_i|^2} h^4 + \frac{1}{36} (b-a)^2 \|\vec{y}_{[i]}\|_\infty^2 h^4 \right) \text{ für } h \leq h_0 \end{aligned}$$

Wegen $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ folgt

>>

Satz 2.8.2. Sei λ_i ein beliebiger Eigenwert des Sturm–Liouville–Problems (2.14) und $y_{[i]}$ eine zugehörige, auf $\int_a^b y_{[i]}^2(x) dx = 1$ normierte Eigenfunktion. Ist dann N hinreichend groß, dann gibt es ein Eigenwert–Eigenvektorkpaar $(\tilde{\lambda}_i, \vec{\eta}_{[i]})$ des diskreten Eigenwertproblems

$$\frac{1}{h^2} A \vec{\eta} = \tilde{\lambda} \vec{\eta} \quad \text{mit } h \|\vec{\eta}\|_2^2 = 1,$$

so daß

$$|\tilde{\lambda}_i - \lambda_i| \leq \frac{\sqrt{b-a}}{\sqrt{1 + \frac{b-a}{12} h^2 y_{[i]}''(\xi)}} h^2 \underbrace{C(p, p', \dots, p''', y'_{[i]}, \dots, y_{[i]}^{(4)})}_{=: C_i},$$

$$\sqrt{h} \|\vec{y}_{[i]} - \vec{\eta}_{[i]}\|_2 \leq \left(\frac{2(b-a)C_i}{\min_{k \neq i} |\tilde{\lambda}_k - \tilde{\lambda}_i|} + \frac{1}{6}(b-a) \|y_{[i]}''\|_\infty \right) h^2$$

Die numerische Lösung des diskreten Eigenwertproblems geschieht am zweckmäßigsten über das z.B. bei Stoer&Bulirsch beschriebene Intervallhalbierungsverfahren, die Bestimmung der Eigenvektoren über das Wielandtverfahren. In der Praxis interessieren stets nur wenige (m) der kleinsten Eigenwerte mit den zugehörigen Eigenfunktionen. Man wählt dann $N \gg m$ und nimmt die m kleinsten Eigenwerte des zugehörigen diskreten Problems mit den zugehörigen Eigenvektoren als Näherung.

Beispiel 2.8.1. $-y'' = \lambda y$, $y(0) = y(1) = 0$ die einfachste Modellierung eines gelenkig eingespannten Stabes unter konstanter Belastung P , mit konstantem Elastizitätsmodul E und konstantem Flächenträgheitsmoment \mathcal{J}

$\lambda = \frac{P}{\mathcal{J}E}$ (linearisiertes Problem mit $\sin y \mapsto y$)

Das diskrete Problem lautet mit $h = 1/(N+1)$

$$\frac{1}{h^2} A \vec{\eta} = \tilde{\lambda} \vec{\eta}, \quad A = \begin{pmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}$$

mit den Eigenwerten

$$\tilde{\lambda}_{j-1} = \frac{2}{h^2} (1 - \cos(jh\pi)) \quad j = 1, \dots, N$$

und den Eigenvektoren $\vec{\eta}_{j-1} = (\sin(jh\pi), \dots, \sin(jNh\pi))^T$.

Die exakten Eigenwerte sind $\lambda_{j-1} = j^2 \pi^2 \quad j = 1, 2, 3, \dots$

mit den Eigenfunktionen $y_{[j-1]}(x) = \sin(j\pi x) \quad j = 1, 2, 3, \dots$

Hier ist also $\vec{y}_{[j-1]} = \vec{\eta}_{j-1}$ und

$$\begin{aligned}
 |\lambda_{j-1} - \tilde{\lambda}_{j-1}| &= \left| j^2 \pi^2 - \frac{2}{h^2} \left(1 - \sum_{k=0}^{\infty} (-1)^k \frac{(jh\pi)^{2k}}{(2k)!} \right) \right| \\
 &= \left| j^2 \pi^2 - \frac{2}{h^2} \left(1 - 1 + \frac{j^2 h^2 \pi^2}{2} - \sum_{k=2}^{\infty} (-1)^k \frac{(jh\pi)^{2k}}{(2k)!} \right) \right| \\
 &= \left| \frac{2}{h^2} \sum_{k=2}^{\infty} (-1)^k \frac{(jh\pi)^{2k}}{(2k)!} \right| \leq \frac{1}{12} h^2 j^4 \pi^4 \cosh(hj\pi) \quad 1 \leq j \leq N
 \end{aligned}$$

entsprechend den theoretischen Resultaten.

Man beachte, daß man für große j keine brauchbaren Resultate erhält, wenn nicht zugleich $N \gg j$ extrem groß wird! \square

2.9 Zusammenfassung

In diesem Kapitel haben wir eine Reihe von Diskretisierungsverfahren für Zweipunkttrandwertaufgaben besprochen. Allgemeine Zweipunkttrandwertprobleme von Systemen erster Ordnung werden in der Praxis entweder mit dem Mehrfachschieß-Verfahren oder mit der (lokalen) Kollokationsmethode behandelt. Ist das zugeordnete Anfangswertproblem nicht steif, wird man bevorzugt das Mehrfachschieß-Verfahren einsetzen. Beide Verfahrenstypen erfordern bei nichtlinearen Problemstellungen gute Anfangsnäherungen für die Lösung, was aber eine natürliche Einschränkung ist. Die einfachen Differenzenverfahren spielen eine Rolle bei der Lösung von Subproblemen bei der Behandlung partieller Differentialgleichungen. Die Methode der finiten Elemente haben wir hier nur deshalb so ausführlich dargestellt, weil sie hier besonders leicht zu analysieren ist. Sie spielt ihre Hauptrolle bei partiellen Differentialgleichungen. Es wird sich zeigen, daß viele Beweise dort formal mit den hier angegebenen fast identisch sind.

Kapitel 3

Software

Einige der in dieser Vorlesung besprochenen Zeitintegratoren

1. Euler explizit und Heun (als Einbettung)
2. Runge-Kutta in der Version von Dormand und Prince mit der Ordnung 4 und 5,
3. das vollimplizite Runge-Kutta-Verfahren Radau IIa mit 3 Stufen der Ordnung 5
4. ein Rosenbrock-Wanner-Verfahren der Ordnung 4 und 5

können interaktiv auf unserem Server

<http://numawww.mathematik.tu-darmstadt.de:8080>

erprobt werden.

Die MATLAB-ODE-Suite enthält die folgenden Codes :

1. ode45: Ein eingebettetes explizites Runge-Kutta-Verfahren der Ordnung 4 und 5 mit Koeffizienten von Shampine
2. ode23: Ein eingebettetes explizites Runge-Kutta-Verfahren der Ordnung 2 und 3
3. ode113: Das Verfahren ABMVOAS (Ordnungen 1 bis 13)
4. ode15s: Die BDF-Verfahren der Ordnung 1 bis 5
5. ode23s: ein eingebettetes Rosenbrock-Wanner-Verfahren mit den Ordnungen 2 und 3
6. ode23t: Trapezregel (Ordnung 2) mit Fehlerkontrolle
7. ode23tb: Implizites Runge-Kutta-Verfahren mit den Ordnungen 2 und 3.

8. bvp4c: Randwertprobleme mittels lokaler Kollokation

Eine grosse Anzahl von Codes ist über das Netz herunterladbar, unter anderem die Entwicklungen von Hairer und Wanner, (u.a. alle Codes aus den Lehrbüchern, aber auch neuere Entwicklungen)

<http://www.unige.ch/hairer/software.html>

sowie aus der Codelib der elib (z.B. das Verfahren von Gragg-Bulirsch-Stoer und Entwicklungen aus der Arbeitsgruppe von P. Deuffhard):

<http://elib.zib.de>

Ferner findet man viele weitere gute Codes hier:

<http://www.netlib.ornl.gov/ode>
<http://www.netlib.ornl.gov/odepack>

u.a. die Fortran-Version von **ABMVOAS**, den Code **DASSL** zur Integration von Algebra-Differentialgleichungen von Linda Petzold, einen für mild steife Gleichungen geeigneten expliziten Runge-Kutta-Code von Sommeier **RKC**.

Index

- $A(\alpha)$ -stabil, 85
- A_0 -stabil, 85
- Äquivalenzsatz, 59
- Index 0, 102
- A-stabil, 85
- Abschneidefehler, 28
- Abschneidefehler, lokaler, 54
- absolut stabil, 85
- absolut stetig, 147
- Adams-Bashforth, 46
- Adams-Moulton, 46
- amplitudentreu, 86
- asymptotisch stabil, 55
- asymptotische Entwicklung, 63
- BDF, 47
- Cowell, 137
- Céa, 154
- D-stabil, 55
- Dahlquist's zweite Ordnungsschranke, 89
- Dahlquists erste Stabilitätsschranke, 63
- Defektkorrektur, 36
- diagonal-implizite Runge-Kutta-Verfahren, 96
- Differenzgleichung, inhomogene, 51
- Diskretisierungsfehler, globaler, 12, 54
- Diskretisierungsfehler, lokaler, 12
- Diskretisierungsfehler, lokaler (MSV), 54
- dissipative Differentialgleichung, 91
- Dormand-Prince, 34
- Dreikörperproblem, 37
- Einschrittverfahren, 14
- Energienorm, 149
- Euler-Verfahren, modifiziertes, 13
- Euler-Verfahren, 9
- explizite Runge-Kutta Verfahren, 24
- Fehlerschätzung, 27
- finite Elemente, 153
- Fundamentalsystem, 51
- Gaus-Runge-Kutta-Verfahren, 95
- Gear-Methode, 47
- Glättung, 65
- Gragg, 65
- Gragg-Bulirsch-Stoer, 66
- Gronwall Lemma, 7
- Hamiltonsche Systeme, 103
- implizite Runge-Kutta-Verfahren, 93
- implizites Verfahren, direkte Iteration, 27
- implizites Verfahren, Newtonlöser, 27
- irreduzibel diagonaldominant, 123
- Kollokation, 140
- konsistent, 16
- konsistent (MSV), 54
- Konsistenzrelationen, 25, 62
- konvektionsdominiert, 127
- konvergent, 15
- konvergent (MSV), 54
- L-stabil, 85
- lineare Stabilität, 83
- logarithmische Matrixnorm, 29
- logarithmische Norm, 29
- Mehrschrittverfahren, allgemeine, 54
- Mehrzielmethode, 117
- Mittelpunktregel, 63
- multiple shooting, 117
- nichtäquidistantes Gitter, 29
- nichtlineare Stabilität, 9

Nitsche, 159
Norm, logarithmische, 29
Nullstellenbedingung, 56

Obreschkov, 24
optimal B -konvergent, 92
Ordnung, variable, 78

PECE, 74
Poincaré, 148
Prädicator-Korrektor, 74

randnah, 137
retardierte DGL, 103
Richardsonextrapolation, 13, 36
Ritzansatz, 153
Rosenbrock-Verfahren, 93
Rosenbrock-Wanner-Formel, 96
Runge-Kutta, 24
Runge-Kutta-Fehlberg, 26
Runge-Kutta-Verfahren, eingebettete, 33

Schrittfunktion, 14
Schrittweite, variable, 78
Schrittweitensteuerung, 27
Sobolevraum, 147
stabil, 15
Stabilitätsbereich, 85
Stabilitätsgebiet, 85
Stabilitätspolynom, 84
steif, 82
Steife Differentialgleichungen, 78
Steifheitsmas, 82

Taylorverfahren, 23
Trapezregel, 13

upwind, 139

Verfahren von Heun, 13
Verfahrensordnung (MSV), 54
Vorschlagsschrittweite, 32