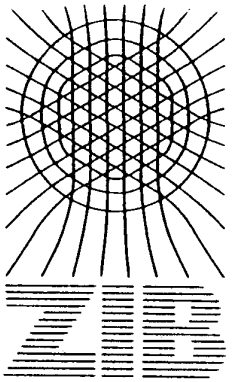
Konrad-Zuse-Zentrum für Informationstechnik Berlin



P. Deuflhard

Numerik von Anfangswertmethoden für gewöhnliche Differentialgleichungen

Skriptum zur Vorlesung, FU Berlin, WS 87/88,
überarbeitet und ergänzt von F. Bornemann

Technical Report TR 89-2 (März 1989)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Heilbronner Strasse 10
1000 Berlin 31
Verantwortlich: Dr. Klaus André
Umschlagsatz und Druck: Verwaltungsdruckerei Berlin

ISSN 0933-7911

Vorwort

Das vorliegende Skriptum entstand aus einer Vorlesung, die ich im WS 87/88 an der Freien Universität Berlin im Fachbereich Mathematik gehalten habe. Mein ursprüngliches Vorlesungsmanuskript wurde von Herrn F. Bornemann in weiten Teilen überarbeitet, reorganisiert und substantiell ergänzt. Der Inhalt stammt größtenteils aus Originalarbeiten jüngeren Datums. Darüberhinaus finden sich zahlreiche Teile, die aus meiner jahrelangen Beschäftigung mit dem Thema entstanden aber unpubliziert geblieben sind.

Das Skriptum erhebt *nicht* den Anspruch, ein *Lehrbuch* zu sein. Es war zunächst als Ausarbeitung für meinen studentischen Hörerkreis sowie als internes Arbeitspapier für das ZIB bestimmt. Die Kunde von der bloßen Existenz eines solchen Skriptums hat jedoch zu einer derart regen Nachfrage geführt, daß es hiermit als Technischer Report des ZIB einer breiteren Öffentlichkeit zugänglich gemacht werden soll.

In der vorliegenden Form richtet es sich in erster Linie an *Mathematiker*; es sollte sich jedoch auch für *Naturwissenschaftler* und *Ingenieure* eignen, die sich einen Einblick in den theoretischen und algorithmischen Hintergrund der von ihnen verwendeten wissenschaftlichen Software verschaffen wollen.

Für viele anregende Diskussionen danke ich, neben Herrn Bornemann, besonders Herrn M. Wulkow. Zahlreiche Testläufe wurden von Herrn U. Pöhle mit Sorgfalt und Geduld ausgeführt. Mein ganz spezieller Dank gilt Frau S. Wacker und Frau E. Körnig, die mit Akribie, Formgefühl und Ausdauer das vorliegende Manuskript in TEX geschrieben haben — mit Rat und Tat begleitet von Herrn R. Roitzsch. Last not least, möchte ich den Herren O. Paetsch und J. Langendorf für ihren Einsatz bei der Erstellung der Figuren sowie der Einspielung der Graphiken in das TEX–Manuskript danken.

Peter Deuflhard

Inhalt

A. Nichtsteife Anfangswertprobleme

1 Theoretische Grundlagen

Gegeben ein System von n Differentialgleichungen (DG):

$$y' = f(y),\ y(0) = y_0 \tag{1.1}$$

Falls $f = f(t,y)$: führe formal $y_{n+1} := t,\ t' = 1,\ y_{n+1}(0) = 0$ ein.

Beispiele: Mechanik (Mehrkörpersysteme, Robotik, Bahnberechnung in der Himmelsmechanik)
Elektronik (Entwurf u. Simulation von Schaltkreisen)
chem. Reaktionskinetik (n=100 - 1000)
parabol. PDG $\Longrightarrow$ Linienmethode ($n = n_x \cdot n_y \cdot n_z$)
Populationsmodelle etc.

1.1 Existenz und Eindeutigkeit

Satz 1.1 (PEANO 1890 [90])
Sei $f : D \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$ sowie $f \in C^0(D)$. Dann existiert mindestens eine Lösung des Anfangswertproblems in D.

Beweis: Explizite Euler-Diskretisierung angewendet auf Anfangswertproblem (1.1), gleichgradige Stetigkeit zeigen. Anwendung des Satzes von Arzela-Ascoli. ∎

Beispiele: Für nur C^0 stochastische Differentialgleichungen (etwa Langevin-Gleichung).

Satz 1.2 (PICARD 1890, LINDELÖF 1894 [92],Linde)
Voraussetzung wie Satz 1.1 sowie: f genüge einer Lipschitz-Bedingung

$$\| f(y) - f(z) \| \leq L_1 \| y - z \| \tag{1.2}$$

für alle $y, z \in D$, D hinreichend groß.
Dann existiert durch jeden Punkt $(t_0, y_0) \in D$ genau eine Lösung des Anfangswertproblems

$$y' = f(y),\ y(t_0) = y_0$$

d.h. D wird durch Trajektorien schlicht überdeckt.

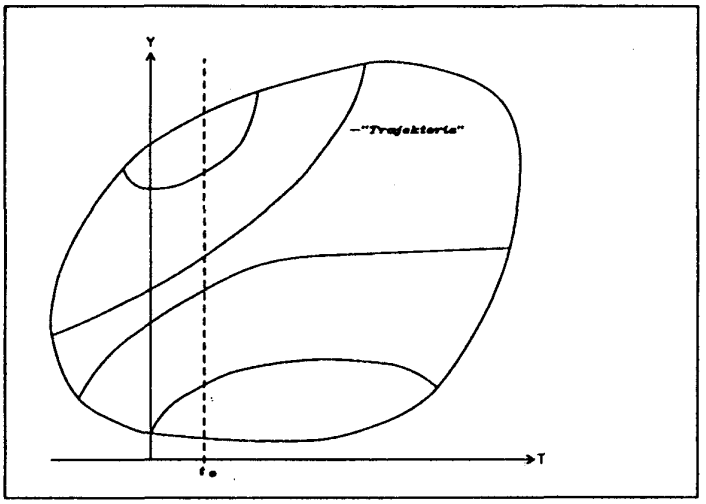


Bild A.1 "Trajektorie"

Korollar: Zwei Lösungen $u(t)$, $v(t)$, von (1.1), die in einem Punkt t_0 übereinstimmen, sind identisch.

Beweis (Skizze): Umformung in Integralgleichung (Volterra 2. Art):

$$y(\tau) = y_0 + \int_{t=0}^{\tau} f(y(t))dt \tag{1.3}$$

Konstruktion einer Fixpunkt-Iteration, der sogenannten *Picard-Iteration:*

$$\begin{aligned} y^{i+1}(\tau) &:= y_0 + \int_0^{\tau} f(y^i(t))dt \\ y^0(\tau) &\equiv y_0 \\ \hookrightarrow \quad y^1(\tau) &= y_0 + \tau \cdot f(y_0) \\ \hookrightarrow \quad &\text{Funktionenfolge } \{y^i\} \end{aligned} \tag{1.4}$$

Zur Kontraktion:

$$\begin{aligned} \| y^{i+1}(\tau) - y^i(\tau) \| &= \| \int_0^{\tau} [f(y^i(t)) - f(y^{i-1}(t))]dt \| \\ &\leq \int_0^{\tau} \| f(y^i(t)) - f(y^{i-1}(t)) \| dt \end{aligned} \tag{$*$}$$

Hier ist offensichtlich die Einführung der Lipschitz-Konstanten L_1 zwingend.

$$\begin{aligned} &\leq L_1 \cdot \int_0^\tau \underbrace{\| y^i(t) - y^{i-1}(t) \|}\, dt \\ &\leq L_1 \cdot \int_0^\tau e_{i-1}(t) dt \end{aligned}$$

Majorante eingeführt:

$$\begin{aligned} \| y^{i+1}(t) - y^i(t) \| &\leq e_i(t) \\ \hookrightarrow \| y^1(t) - y^0(t) \| &\leq t \cdot \| f(y_0) \| \end{aligned} \tag{1.5}$$

Peano-Konstante eingeführt:

$$\begin{aligned} \| f(y_0) \| &\leq L_0 \\ \hookrightarrow \| y^1(t) - y^0(t) \| &\leq L_0 \cdot t =: e_0(t) \end{aligned} \tag{1.6}$$

$$\begin{aligned} L_1 \int_0^\tau e_{i-1}(t) dt &=: e_i(\tau) \\ e_0(\tau) &= L_0 \cdot \tau \end{aligned} \tag{1.7}$$

Induktion:

$$e_i(t) = L_0 L_1^i \frac{t^{i+1}}{(i+1)!} \tag{1.7'}$$

$\Longrightarrow$ Cauchy-Konvergenz, ferner

$$\begin{aligned} \lim_{i\to\infty} \| y^{i+1}(t) - y^0(t) \| &\leq \lim_{i\to\infty} \textstyle\sum_{j=0}^{i} e_i(t) \\ &= L_0 t \cdot \lim_{i\to\infty} \underbrace{\sum_{j=0}^{i} \frac{(L_1 t)^j}{(j+1)!} \cdot \frac{L_1 t}{L_1 t}}_{\frac{1}{L_1 t} \underbrace{\sum_{j=0} \frac{(L_1 t)^{j+1}}{(j+1)!}}_{\underbrace{\sum_{j=0}^{i+1} \frac{(L_1 t)^{j+1}}{(j+1)!}}_{\to \exp(L_1 t)}} - 1} \\ &\leq L_0 t \ \varphi\ (L_1 t) \end{aligned}$$

wobei definiert wird:

$$\varphi(s) := \begin{cases} (exp(s)-1)/s & s \neq 0 \\ 1 & s = 0 \end{cases} \tag{1.8}$$

$\Longrightarrow \exists\, y^*(t)$:

$$\| y^*(t) - y_0 \| \leq L_0\, t\, \varphi(L_1 t) \tag{1.9}$$

Existenz ■

Satz 1.3 (H. KNESER 1923 [70])
Sei $f \in C^0[0,1]$. Dann gilt: die Menge der Lösungen $y(t)$ des Anfangswertproblems (1.1) bildet ein Kontinuum (abgeschlossen, zusammenhängend).

Beweis: HARTMANN [59], S. 15f.

■

Beispiel: $y' = \frac{\sqrt{1-y^2}}{y}$, $y(0) = 1$
nicht Lipschitz-stetig für $y \equiv 1$ [und $y \equiv 0$] Grenzlösungen:
$y_{max}(t) \equiv 1$, $y_{min}(t) = \sqrt{1-t^2}$

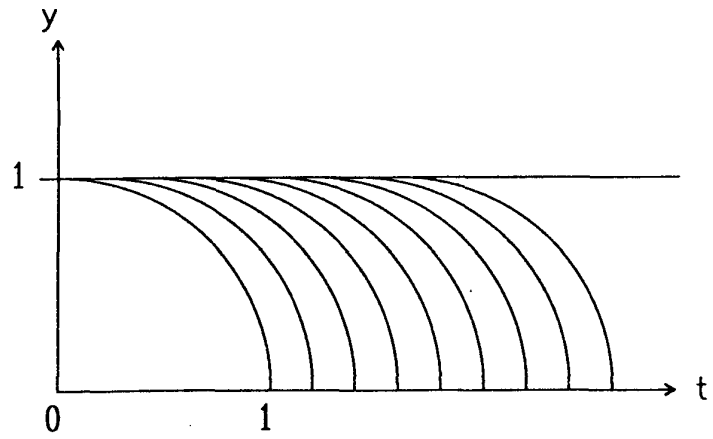


Bild A.2 Kontinuum

beliebige Lösungen zweigen von $y \equiv 1$ ab:
(Translation der Kreisbögen wegen $f(y)$ *autonom*)

Ausnahmefälle:

(I) *Singuläre Differentialgleichung*, etwa ($k \neq -1$, *i.a.* $k > 0$) : [21]

$$y'' = -k\frac{y'}{t} + g(t,y),\ y'(0) = 0\ ,\ y(0) = y_0 \tag{1.10}$$

$g(t,y)$ beschränkt, Lipschitzstetig, erster Term nicht Lipschitzstetig.

Analytische Vorbehandlung:
Übliche Transformation $y' = z$ führt auf System 1. Ordnung

$$\begin{aligned} z' &= -k\frac{z}{t} + g(t,y)\ ,\ z(0) = 0 \\ y' &= z \end{aligned}$$

Taylor-Entwicklung um $t = 0$:

$$z(t) = z(0) + t \cdot z'(0) + O(t^2)$$

Eingesetzt in (1.10):

$$\begin{aligned} z'(t) &= -k \cdot z'(0) + O(t) + g(t,y) \\ t \to 0: &\quad z'(0) = \frac{g(0,y_0)}{1+k} = y''(0)\,. \end{aligned} \tag{1.10'}$$

Hierbei muß die Lipschitz-Bedingung auf Vergleichsfunktionen mit $y'(0) = 0$ eingeschränkt werden.

Häufig vorkommender Problemtyp !
(Elastizitätstheorie, Fluidmechanik).

Beispiel: Transformation der Laplace-Gleichung im $\mathbb{R}^n$:

$$k = n - 1$$

Häufig tritt dieser Typ bei radialsymmetrischem Ansatz für die Poisson-Gleichung auf:

$$\Delta\varphi = f\ ,$$

transformiert sich dann auf

$$\varphi''(r) + \frac{n-1}{r}\varphi'(r) = f(r)\ ,$$

das heißt auf Gleichung vom Typ (1.10).

(II) *Thomas-Fermi-Differentialgleichung*

$$y''(t) = \frac{y^{\frac{3}{2}}(t)}{t^{\frac{1}{2}}}\,, \quad y(0) = y_0\,, \quad y(\infty) = 0\,.$$

Bei $t = 0$ nicht Lipschitz-stetig! Um Lipschitz-stetige Differentialgleichung zu erhalten, substituiere

$$\begin{aligned} y(t) &=: w(\sqrt{t}) \\ s &:= \sqrt{t} \\ \tfrac{d}{ds}w &=: \dot{w}\,. \end{aligned}$$

Dann erhalten wir

$$\ddot{w} = \frac{\dot{w}}{s} + 4sw^{\frac{3}{2}}\,, \quad w(0) = y_0\,, \quad w(\infty) = 0 \qquad (*)$$

Reguläre Lösung verlangt:

$$\dot{w}(0) = 0\,.$$

Wegen $k = -1$ (vergleiche (1.10)) liegt hier keine hebbare Singularität vor. Weitere Substitution des Singularitätentermes

$$u(s) := \tfrac{\dot{w}(s)}{s}$$

ergibt

$$\begin{aligned} \dot{w}(s) &= s\,u \\ \dot{u}(s) &= 4\,w^{\frac{3}{2}} \end{aligned}$$

Die rechte Seite dieses Systems ist nun Lipschitz-stetig. Als Anfangswerte erhalten wir

$$w(0) = y_0$$

und

$$u(0) = \lim_{s\to 0} \frac{\dot{w}(s)}{s} = \ddot{w}(0)$$

nach der Regel von l'Hospital.
Hiermit haben wir den weiteren Freiheitsgrad zur Erfüllung von $w(\infty) = 0$ erhalten. Beachte, daß $\ddot{w}(0)$ in (*) zunächst beliebig ist. Thomas-Fermi-Differentialgleichung ist *asymptotisches Randwertproblem:* die fehlende Anfangsbedingung $u(0)$ wird durch $w(\infty) = 0$ bestimmt (vergleiche Kapitel D.6).

Linearer Spezialfall:

$$f = Ay$$

Picard- Iteration liefert:

$$y^i(t) := \sum_{j=0}^{i} \frac{(At)^j}{j!}\ y_0 \tag{1.11}$$

Definition: Matrizen-Exponentielle (formale Darstellung)

$$y^*(t) = \lim_{i\to\infty} y^i(t) = \exp(At)\ y_0 \tag{1.12}$$

Vorsicht! $\exp(A+B) = \exp(A)\cdot\exp(B)$ nur für $AB - BA = 0$

Allgemeiner nichtlinearer Fall
Studium der Konvergenz durch Majorante:

$$\| y^i(t) - y^*(t) \| \leq \bar{e}_i(t) \tag{1.13.a}$$

Aus (1.9)

$$\| y^0(t) - y^*(t) \| \leq L_0 t\ \varphi\ (L_1 t)$$

Annahme:

$$\begin{aligned} L_1 t &\leq C_1 \\ \hookrightarrow\ \varphi(L_1 t) &\leq C_2 \\ \| y^0(t) - y^*(t) \| &\leq C_2 L_0 t =: \bar{e}_0(t) \end{aligned} \tag{1.14}$$

Induktion:

$$\bar{e}_k(t) = C_2 L_0 t \cdot \frac{C_1^k}{(k+1)!} \tag{1.13.b}$$

mindestens *superlineare* Konvergenz, falls $C_1 < 1$.

Frage: Ist die Picard-Iteration brauchbar für numerische Lösung von Differentialgleichungen?

$$\begin{aligned} y^0(\tau) &= y_0 \\ y^1(\tau) &= y_0 + \tau\cdot f(y_0) \\ \hookrightarrow\ & \text{explizites Euler-Verfahren} \\ y^2(\tau) &= y_0 + \int_0^\tau f(y_0 + t\ f(y_0))dt \end{aligned}$$

nur für spezielle f geschlossen darstellbar $\Longrightarrow$ numerische Quadratur.

$\Longrightarrow$ Diskretisierung der DG direkt, ohne Umweg über Picard-Iteration.

1.2 Sensitivität

Für jeden Algorithmus sind die Anfangswerte y_0 Eingabedaten.

Frage: Wie hängt die Lösung von Änderung der Anfangswerte ab?

Satz 1.4 (Spezialfall des sogenannten Fundamentallemmas)
Voraussetzungen des Existenz- und Eindeutigkeitssatzes 1.2.
Seien $u(t)$ und $v(t)$ Lösungen der Differentialgleichungen
$y' = f(y)$ zu den Anfangsbedingungen

$$u(0) = u_0\ ,\ \ v(0) = v_0\ ,\ \ u_0 \neq v_0$$

Dann gilt:

$$\| u(t) - v(t) \| \leq \| u_0 - v_0 \| \exp(L_1 t) \tag{1.15}$$

das heißt die Lösungen hängen stetig von den Anfangswerten ab.

Bemerkung: Für *partielle* Differentialgleichungen nicht selbstverständlich!

Beweis:

$$\begin{aligned}
u(t) &= u_0 + \int_{s=0}^{\tau} f(u(s))\, ds \\
v(t) &= v_0 + \int_{s=0}^{\tau} f(v(s))\, ds \\
\Longrightarrow u(t) - v(t) &= u_0 - v_0 + \int_{s=0}^{\tau} [f(u) - f(v)]\, ds \\
\Longrightarrow \| u(t) - v(t) \| &\leq \| u_0 - v_0 \| + \int_{s=0}^{\tau} \| f(u) - f(v) \|\, ds \\
\longrightarrow &\leq \| u_0 - v_0 \| + L_1 \int_{s=0}^{t} \underbrace{\| u(s) - v(s) \|}_{\leq m(s)}\, ds \\
\text{Definition: } m(t) &\geq \| u(t) - v(t) \| \geq 0
\end{aligned}$$

Damit gilt:

$$m(t) = m(0) + L_1 \int_{s=0}^{t} m(s)\, ds$$

Differentiation:

$$\begin{aligned} m' &= L_1 m \ , m(0) := \| u_0 - v_0 \| \\ \hookrightarrow m(t) &= m(0)\ exp(L_1 t) \end{aligned}$$

■

Formel (1.15) liefert auch *Differenzierbarkeit* bezüglich Anfangswert y_0 .
Variation von y_0 :

$$\begin{aligned} y_0 &\longrightarrow y_0 + \delta y_0 \\ \hookrightarrow y(t) &\longrightarrow y(t) + \delta y(t) \end{aligned}$$

Variationsgleichung:

$$\delta y' = f_y(y(t))\ \delta y,\ \delta y(0) = \delta y_0 \tag{1.16}$$

Definition: Wronski-Matrix

$$W(t, t_0) := \frac{\partial y(t)}{\partial y(t_0)} \tag{1.17}$$

(auch: Begleitmatrix, Übertragungsmatrix,
englisch: companion matrix, propagation matrix)

Aus dem Existenz- und Eindeutigkeitssatz ergeben sich die folgenden *Gruppeneigenschaften:*

$$\begin{aligned} &\text{a)}\ \ W(t,\ \tau)W(\tau,\ t_0) = W(t,\ t_0) \\ &\text{b)}\ \ W(t,\ \tau)^{-1} = W(\tau,\ t) \\ &\text{c)}\ \ W(t,\ t) = I \end{aligned} \tag{1.18}$$

Zugehörige Differentialgleichung:

$$\frac{dW(t, t_0)}{dt} = f_y(y(t))\ W(t, t_0),\ W(t_0, t_0) = I \tag{1.19}$$

Formale Lösung der Variationsgleichung:

$$\delta y(t) = W(t, t_0)\ \delta y_0 \tag{1.20}$$

Parameterabhängigkeit:

$f(y,p)$, $p : q$ Parameter $(p_1, \ldots, p_q)$
$y_p(t)$: $(n,\ q)$ – Matrizen

Zugehörige Variationsgleichung:

$$\begin{aligned} y_p' &= f_y(y(t),p)y_p + f_p(y(t),p) \\ y_p(0) &= \frac{\partial y_0}{\partial p} \end{aligned} \tag{1.21}$$

Definition: Kondition (bezüglich Störung δy_0):

$$\sigma(0,T) := \max_{t\in[0,T]} \| W(t,0) \| \tag{1.22}$$

$\hookrightarrow$

$$\max_{t\in[0,T]} \| \delta y(t) \| \leq \dot{\sigma}(0,T) \| \delta y_0 \| \tag{1.22'}$$

Vergleich mit (1.15) ergibt:

$$\sigma(0,T) = \exp(L_1 T) \quad \text{für nicht-steife Probleme}$$

Störungen der rechten Seite:

$$f \longrightarrow f + \delta f \Longrightarrow y \longrightarrow y + \delta\bar{y}$$

$$\delta\bar{y}(t) \doteq \int_0^t W(t,s)\ \delta f(s)\ ds \tag{1.23}$$

(Spezialfall des Satzes von ALEKSEJEW-GRÖBNER [52])

1.3 Affin-Kovarianz

Zu untersuchen ist das Transformationsverhalten von Differentialgleichung (1.1) unter Affin-Transformation. Zu fordern ist grundsätzlich, daß die numerischen Lösungsmethoden das gleiche Transformationsverhalten wie das analytische Problem aufweisen.
Sei B nichtsinguläre (n,n)-Matrix mit $B_t \equiv 0$.
Affin-Transformation der Variablen liefert:

$$\begin{aligned} &y \to By =: \bar{y} \\ &\bar{y}' = By' = Bf(y) = Bf(B^{-1}y). \end{aligned} \tag{1.24.a}$$

Somit lautet das transformierte Anfangswertproblem:

$$\begin{aligned} &\bar{y}' = \bar{f}(\bar{y}) \quad ; \bar{y}_0 = By_0 \\ &\text{mit } \bar{f}(\bar{y}) := Bf(B^{-1}\bar{y}) \end{aligned} \tag{1.24.b}$$

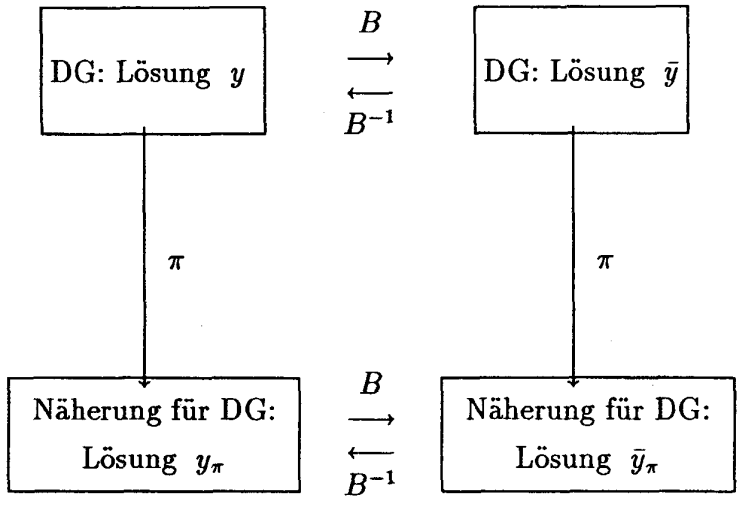
Sei $\bar{y}$ Lösung von (1.24.b) und y diejenige von (1.1). Existenz & Eindeutigkeit liefert:

$$\bar{y} = By \quad \text{Kovarianzeigenschaft} \tag{1.24.c}$$

Affin-Kovarianz bei Näherungsverfahren: Sei π die zum Näherungsverfahren gehörige (endlich-dimensionale) Projektion.
Dann muß gelten:

$$y_\pi \text{ ist Näherung für } y \Longleftrightarrow \bar{y}_\pi = B\,y_\pi \text{ ist Näherung für } \bar{y} = By \tag{1.25}$$

Also kommutiert das Diagramm:



Allgemeinste Form, welche diese Kovarianzeigenschaft hat:

"general linear methods"

(BUTCHER 1987) [16]. In etwas "gelichteter" Form, so daß die später behandelten Spezialfälle deutlich erkennbar sind,sei:

$$y_j \text{ Näherung für } y(j \cdot h)$$

Aus $y_0, \ldots, y_{k-1}$ berechne y_k gemäß:

$$\alpha_k y_k + \ldots + \alpha_0 y_0 = h \sum_{j=0}^{s} b_j \hat{u}_j$$

$$\hat{u}_j = f(u_j)$$

$$u_j = \gamma_{jk} y_k + \ldots + \gamma_{j0} y_0 + h \sum_{i=0}^{s} a_{ji} \hat{u}_i \,, \quad j = 0, \ldots, s$$

Die im weiteren behandelten Spezialfälle ergeben sich wie folgt:

Spezialfall I : Runge-Kutta-Verfahren

$$\begin{aligned} k &= 1 \\ \alpha_1 &= 1\,, \quad \alpha_0 = -1 \\ \gamma_{j1} &= 1\,, \quad \gamma_{j0} = -1 \end{aligned}$$

Speziallfall II : Mehrschrittverfahren

$$\begin{aligned} k &= s \\ a_{ji} &= 0 \qquad i,j = 0,\ldots,s \\ \gamma_{jk} &= \delta_{jk} \quad \text{(Kronecker } \delta\text{)} \end{aligned}$$

2 Einschrittverfahren

Explizites Euler-Verfahren (EULER 1768)[44]:

$$\begin{aligned} y_{k+1} &:= y_k + h \cdot f(y_k) \qquad k = 0, 1, \ldots \\ y_0 &: \text{ geg. Anfangswert} \\ h &: \text{ (gewählte) Schrittweite} \end{aligned} \tag{2.1}$$

Anstelle der kontinuierlichen Lösung erhält man eine *diskrete* Lösung über dem *Gitter* $\{t_k\}$ mit $t_k = k \cdot h$.

Bemerkung: Die Diskretisierung (2.1) ist älter als der Begriff der Differentialgleichung. Zudem ist sie das wesentliche analytische Hilfsmittel zum Existenzbeweis für Differentialgleichungen (Satz 1.1).

Verallgemeinerung: *Einschritt-Verfahren* (ESV)

$$\begin{aligned} y_{k+1} &:= y_k - h\Phi(y_k, y_{k+1}, h) \\ \Phi &: \textit{Inkrementfunktion} \end{aligned} \tag{2.2}$$

Idee: Geeignete Wahl von Φ sollte zu Verfahren führen, die besser sind als (2.1).

explizites ESV: Φ hängt *nicht* von y_{k+1} ab
implizites ESV: Φ hängt von y_{k+1} ab.

Implizite Einschrittverfahren führen (für $n > 1$) auf i.a. nichtlineare Gleichungssysteme. Bei *nichtsteifen* Problemen werden diese mit Fixpunkt-Iteration, die nur f-Auswertungen verlangt, gelöst.

Beispiele:

$$\begin{array}{lllll} \text{a)} & \Phi & = & f(y_k) & \text{expl. Euler} \\ \text{b)} & \Phi & = & f(y_{k+1}) & \text{impl. Euler} \\ \text{c)} & \Phi & = & \frac{1}{2}(f(y_k)+f(y_{k+1})) & \text{impl. Trapezregel} \\ \text{d)} & \Phi & = & f(\frac{1}{2}(y_k+y_{k+1})) & \text{impl. Mittelpunktsregel} \end{array} \tag{2.3}$$

Raffiniertere Wahl von Φ :

Extrapolationsverfahren (Kapitel 2.3)
Runge-Kutta-Verfahren (Kapitel 2.4)

Für Beweiszwecke wird i.a. Φ explizit angenommen (Satz über implizite Funktionen in Umgebung von y_k formal angewandt).

2.1 Konvergenz

Das Einschrittverfahren (2.2) stellt eine *Differenzengleichung* 1. Ordnung (D_hG) dar.

Frage: Wie verhält sich die Lösung der D_hG zur Lösung der Differentialgleichung?

1. Forderung: Konsistenz

$$\boxed{D_hG} \xrightarrow{h\to 0} \boxed{DG} \tag{2.4}$$

Umformung von (2.2)

$$\frac{y_{k+1}-y_k}{h} = \Phi\,(y_k,\ y_{k+1},\ h) \tag{2.2'}$$

Forderung (2.4) eingesetzt:

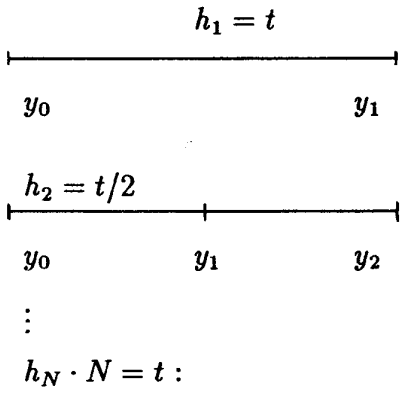
$$\frac{y_{k+1}-y_k}{h} \longrightarrow y'$$

$$\begin{array}{lll} \Phi\,(y,y,0) = f(y) & & \text{(impl.)} \\ \Phi\,(y,0) = f(y) & & \text{(expl.)} \end{array} \tag{2.4'}$$

2. Forderung: Konvergenz

$$\boxed{\text{Lösung der } D_hG} \xrightarrow{h\to 0} \boxed{\text{Lösung der } DG} \tag{2.5}$$

Struktur des Grenzprozesses:



Grenzübergang $h \to 0$ derart, daß $t = Nh$ fest, also zugleich $N \to \infty$.

$$\lim_{\substack{h\to 0 \\ Nh=t}} y_N = y(t) \tag{2.5'}$$

Frage: Impliziert Konsistenz bereits Konvergenz?

Definition: Konsistenzordnung $p > 0$

$$y(t+h) - y(t) - h\Phi(y(t), h) = \mathcal{O}(h^{p+1}) \tag{2.6}$$

(Einsetzen von DG-Lösung in D_hG)

Beispiele:

(2.3.a,b) : p = 1 (Taylor-Entwicklung um $y(t)$)
(2.3.c,d) : p = 2 (Taylor-Entwicklung um $y(t + \frac{h}{2})$)

Definition: "globaler" Diskretisierungsfehler

$$\epsilon(t,h) := y_N - y(t) \tag{2.7}$$

Damit lautet Forderung (2.5):

$$\lim_{h\to 0} \epsilon(t,h) = 0 \tag{2.5''}$$

Lemma 2.1 *Sei p Konsistenzordnung für Einschrittverfahren. Dann gilt:*

$$\epsilon(h,h) = y_1 - y(h) = \mathcal{O}(h^{p+1}). \tag{2.8}$$

Beweis: Φ explizit angenommen (s.o.).

$$\begin{aligned} \longrightarrow \quad y(h) - y_0 - h\Phi(y_0, h) &= \mathcal{O}(h^{p+1}) \quad (2.6) \\ \longrightarrow \quad y_1 - y_0 - h\Phi(y_0, h) &= 0 \qquad (2.2) \end{aligned}$$

■

Satz 2.1 ("Konsistenz $\Longrightarrow$ Konvergenz")
Sei (2.2) ein Einschrittverfahren mit

$$\| \epsilon(h,h) \| \leq C \cdot h^{p+1}, \quad p > 0. \tag{2.8'}$$

Dann gilt (mit φ nach (1.8)):

$$\| \epsilon(t,h) \| \leq C \cdot t \cdot \varphi(L_1 t) \cdot h^p. \tag{2.9}$$

Beweis: (Idee WANNER ~1970)
Man konstruiert sich N Anfangswertprobleme der folgenden Form:

$$\begin{aligned} y' &= f(y), \quad y(t_k) = y_k, \quad t_k = k \cdot h \\ &\hookrightarrow \text{Lösung} \quad y(t; t_k, y_k) \quad \text{für } t \geq t_k \end{aligned}$$

Zwischenschieben der $y(t;\cdot)$ und Anwendung der Fundamentallemmas für Differentialgleichung - (1.15):

$$\| y(t) - y_N \| \leq \underbrace{\| y(t) - y(t; t_1, y_1) \|}_{\text{Fundamentallemma } (L_1 \equiv L)} + \cdots + \| y(t; t_{N-1}, y_{N-1}) - y_N \|$$

$$\leq \underbrace{\| y(t_1) - y_1 \|}_{\text{Lemma 2.1}} exp(L \underbrace{(t - t_1)}_{(N-1)h}) + \cdots + \underbrace{\| y(t; t_{N-1}, y_{N-1}) - y_N \|}_{\text{Lemma 2.1}}$$

$$\leq C \cdot h^{p+1} [1 + exp(Lh) + \cdots + exp(L(N-1)h)]$$

$$= C \cdot h^{p+1} \cdot \frac{exp(Lt) - 1}{exp(Lh) - 1}$$

$$= C \cdot h^{p+1} \cdot Lt \cdot \varphi\,(Lt)/(Lh\varphi(Lh))$$

Zusammenfassung:

$$\| y(t) - y_N \| \leq C \cdot t \cdot h^p \frac{\varphi(L_1 t)}{\varphi(L_1 h)} \tag{2.10}$$

$$\begin{aligned} \varphi \text{ monoton wachsend, } Lh &\geq 0 \Longrightarrow \varphi \geq \varphi(0) = 1. \\ 1/\varphi(L_1 h) &\leq 1 \Longrightarrow (2.9) \end{aligned}$$

■

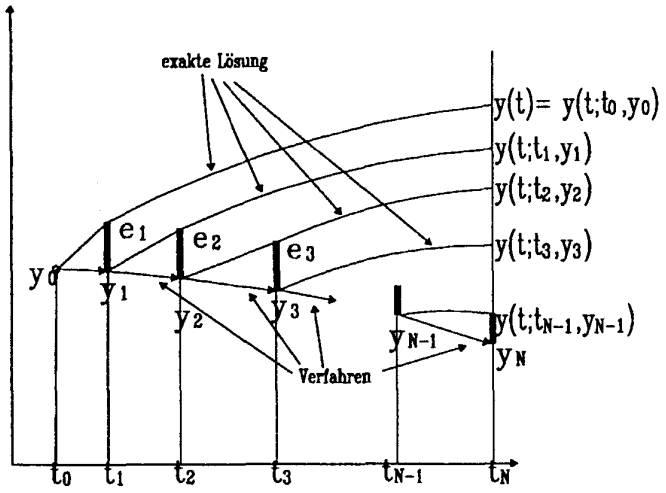


Bild A.3 (WANNER: "Lady Windermere's Fächer")

2.2 Asymptotische Entwicklung des Diskretisierungsfehlers

Nach (2.6) gilt:

$$y(t+h) - y(t) - h\Phi(y(t), h) = \mathcal{O}(h^{p+1})$$

Falls f hinreichend oft differenzierbar, so ist auch Φ (Linearkombination bezüglich f) hinreichend oft differenzierbar. Dann existiert eine *asymptotische Entwicklung des "Konsistenzfehlers"* der Form:

$$y(t+h) - y(t) - h\Phi(y(t), y(t+h), h) = d_{p+1}(t)h^{p+1} + d_{p+2}(t)h^{p+2} + \ldots \quad (2.11)$$

Frage: Besitzt der globale Fehler $\epsilon(t;h)$ ebenfalls eine asymptotische Entwicklung der Form:

$$y_N - y(t) = e_p(t)h^p + e_{p+1}(t)h^{p+1} \cdots \ ?$$

Der Beweis wurde zum ersten Mal von Gragg (1964) geführt. Im folgenden wird der wesentlich einfachere Beweis nach Hairer (1984) dargestellt. Diese Beweistechnik gestattet auch die analytische Durchdringung wesentlich komplizierterer Fragestellungen (siehe Kapitel B.4).

Satz 2.2 (HAIRER 1984 [54])
Sei $f \in C^{p+1}$, $\Phi \in C^{p+1}$ und es gelte:

$$y(t+h) - y(t) - h\Phi(y(t),h) = d_{p+1}(t)\ h^{p+1} + \mathcal{O}(h^{p+2})$$

das heißt Φ definiert ein Einschrittverfahren mit Konsistenzordnung p . Sei $e(t)$ Lösung des linearen Anfangswertproblems:

$$e' = f_y(y(t))\ e - d_{p+1}(t)\ ,\quad e(0) = 0 \tag{2.12}$$

Dann gilt: Das Einschrittverfahren definiert durch die (neue) Inkrementfunktion

$$\begin{aligned} \Phi^*(y_k^*, h) &:= \Phi(y_k, h) - (e(t_k + h) - e(t_k))h^{p-1} \\ y_k^* &:= y_k - e(t_k)h^p \end{aligned} \tag{2.13}$$

besitzt die Ordnung $(p+1)$, d. h. es gilt:

$$y(t+h) - y(t) - h\Phi^*(y(t),h) = \mathcal{O}(h^{p+2}).$$

Hierbei ist $\{y_k^*\}$ die durch Φ^* erzeugte, $\{y_k\}$ die durch Φ erzeugte Gitterfunktion.

Beweis: Ursprüngliches Einschrittverfahren: $t := t_k$, y_0 gegeben:

$$y_{k+1} = y_k + h\Phi(y_k, h)$$

"neues" Einschrittverfahren: Ziel $\mathcal{O}(h^{p+2})$

$$\begin{aligned} y_0^* &:= y_0,\ y_{k+1}^* := y_k^* + h\Phi^*(y_k^*, h) \\ \delta y_k &:= y_k^* - y_k,\ \delta y_0 = 0\ . \end{aligned}$$

Eingesetzt in "altes" Einschrittverfahren:

$$y_{k+1}^* - \delta y_{k+1} = y_k^* - \delta y_k + h\Phi(y_k^* - \delta y_k, h)$$

Also muß gelten:

$$h\Phi^*(y_k^*, h) = h\Phi(y_k^* - \delta y_k, h) + \delta y_{k+1} - \delta y_k \tag{$*$}$$

Konsistenzfehler des "neuen" Einschrittverfahrens:

$$\begin{aligned} y(t+h) - y(t) - h\Phi^*(y(t),h) &= y(t+h) - y(t) - \delta y_{k+1} + \delta y_k - \\ -h\Phi(y(t) - \delta y_k, h) + h\Phi(y(t),h) - h\Phi(y(t),h) &= d_{p+1}(t)h^{p+1} \\ +\delta y_k - \delta y_{k+1} + h \cdot \left[\frac{\partial \Phi}{\partial y}(y(t),h)\delta y_k + \mathcal{O}(\|\delta y_k\|^2)\right] &+ \mathcal{O}(h^{p+2}) \end{aligned} \tag{$**$}$$

Um $\mathcal{O}(h^{p+2})$ zu erreichen, kann man setzen:

$$\delta y_k = -e(t) \cdot h^p, e(0) = 0 \text{ (wegen: } \delta y_0 = 0)$$

Dann gilt:

$$\begin{aligned} \delta y_k - \delta y_{k+1} &= h^p \cdot [-e(t+h)e(t)] = +e'(t)h^{p+1} + \mathcal{O}(h^{p+2}) \\ \mathcal{O}(\|\delta y_k\|^2) &= \mathcal{O}(h^{2p}) \\ \Phi_y(y(t),h) &= \Phi_y(y(t),0) + \mathcal{O}(h) \end{aligned}$$

Wegen Konsistenzbedingung (2.4'):

$$\Phi_y(y(t),0) = f_y(y(t))$$

Aufsammeln der h-Potenzen:

$$= h^{p+1}\,[d_{p+1}(t) + e'(t) - f_y(y(t))e(t)] + \mathcal{O}(h^{p+2}) \qquad (**)$$

$[...] = 0$ und $e(0) = 0$ liefert (2.12). ∎

Mit dieser Vorbereitung ist asymptotische Entwicklung für Einschrittverfahren zu beweisen.

Satz 2.3 (GRAGG 1964 [49])
Sei $f \in C^{p+k}, k \geq 1, y$ Lösung des Anfangswertproblems

$$y' = f(y), y(0) = y_0.$$

Das durch Φ definierte Einschrittverfahren erfülle: $\Phi \in C^{p+k}$

$$\begin{aligned} \Phi(y,0) &= f(y) \text{ und} \\ y(t+h) - y(t) - h\Phi(y(t),h) &= d_{p+1}(t)h^{p+1} + \cdots + d_{p+k}(t)h^{p+k} + \mathcal{O}(h^{p+k+1}) \\ & \quad d_{p+i} \in C^{k-i} \end{aligned}$$

Dann gilt (für $t = N \cdot h$) :

$$y_N - y(t) = e_p(t)h^p + \cdots + e_{p+k-1}(t)h^{p+k-1} + E_{p+k}(t,h)h^{p+k} \qquad (2.14)$$

wobei $\| E(t;h) \| \leq M$ für $h \in [0,H]$. Die Koeffizientenfunktionen e_j sind definiert durch lineare Anfangswertprobleme der Form:

$$\begin{aligned} e_p' &= f_y(y(t))e_p - d_{p+1}(t)\,, \quad e_p(0) = 0 \\ e_{p+1}' &= f_y(y(t))e_{p+1} - \cdots\,, \quad e_{p+1}(0) = 0 \\ &\vdots \end{aligned} \qquad (2.15)$$

Beweis: Satz 2.1 liefert für die Gitterfunktion y_k^* mit Inkrementfunktion Φ^* (Konsistenzordnung $p+1$)

$$||y_N^* - y(t)|| \leq C^* t\varphi(L^* t)h^{p+1} ,$$

das heißt mit (2.13)

$$||y_N - e_p(t)h^p - y(t)|| \leq C^* t\varphi(L^* t)h^{p+1} .$$

Somit erhalten wir

$$y_N - y(t) = e_p(t)h^p + E_{p+1}(t;h)h^{p+1}$$

$$\text{mit } ||E_{p+1}(t;h)|| \leq C^* , \tag{2.16}$$

wobei e_p (2.15) erfüllt (vergleiche 2.12). Damit ist der Satz bewiesen für $k=1$. Für $k>1$ erhält man die Behauptung rekursiv, indem man gemäß Satz 2.2 zu Φ^* eine Inkrementfunktion Φ^{**}, dazu Φ^{***} und so weiter konstruiert. Aus Satz 2.1 erhält man

$$||y_N^{\overbrace{*\,*\,\cdots\,*}^{k}} - y(t)|| \leq C^{\overbrace{*\,*\,\cdots\,*}^{k}} t\varphi(L^{\overbrace{*\,*\,\cdots\,*}^{k}} t)h^{p+k} .$$

Setzt man

$$y_N^{\overbrace{*\,*\,\cdots\,*}^{k}} = y_N - e_p(t)h^p - \ldots - e_{p+k-1}(t)h^{p+k-1}$$

ein, erhält man den Satz. ∎

Satz 2.4 (STETTER 1970 [105])
Falls gilt (symmetrisches Einschrittverfahren)

$$\Phi(y_k, y_{k+1}, h) = \Phi(y_{k+1}, y_k, -h) , \tag{2.17}$$

so gilt in (2.14):

$$e_{2m+1}(t) \equiv 0$$

d.h. es existiert eine quadratische asymptotische Entwicklung.

Beweis: Man geht davon aus, daß eine asymptotische Entwicklung der Form (2.14) existiert. Formal definiert man:

$$\begin{aligned} \hat{t} &:= t + \frac{h}{2} \\ \Longrightarrow y_k = \eta(t;h) &= \eta(\hat{t} - \frac{h}{2}; h) \\ y_{k+1} = \eta(t+h;h) &= \eta(\hat{t} + \frac{h}{2}; h) \end{aligned}$$

$h \Longleftrightarrow -h$ in Darstellung (2.2) für Einschrittverfahren:

$$\eta(\hat{t} - \tfrac{h}{2}; -h) = \eta(\hat{t} + \tfrac{h}{2}; -h) - h \underbrace{\Phi(\eta(\hat{t} + \frac{h}{2}; -h), \eta(\hat{t} - \frac{h}{2}; -h); -h)}_{\Phi(\eta(\hat{t} - \frac{h}{2}; -h), \eta(\hat{t} + \frac{h}{2}; h); h}$$

$$\Longrightarrow \eta(\hat{t} + \tfrac{h}{2}; -h) = \eta(\hat{t} - \tfrac{h}{2}; -h) + h\Phi(\underbrace{\eta(\hat{t} - \frac{h}{2}; -h)}_{\eta(t;-h)}, \underbrace{\eta(\hat{t} + \frac{h}{2}; -h)}_{\eta(t+h;-h)}; h)$$

Eindeutigkeit des Einschrittverfahrens:

$\Longrightarrow \eta(t; h) = \eta(t; -h) \Longrightarrow$ quadratische Entwicklung ∎

2.3 Explizite Extrapolationsmethoden

Prinzip: (RICHARDSON 1910) [95]
Vorgehen wie bei ROMBERG-Quadratur (dort allerdings h^2-Entwicklung). Für eine Diskretisierung gelte *h^γ-Entwicklung* ($\gamma = 1$ *oder* $\gamma = 2$) der Form (2.14) mit $p = \gamma$. Wiederholung der Diskretisierung über *Grundschritt* $[0, H]$ mit sukzessive feinerer *interner Schrittweite*

$$h_i := H/n_i \quad n_i \in \mathcal{F}$$

liefert Approximationen

$$y_{n_i} = y(H, h_i) \doteq y(H) \,.$$

Aufbau eines *Extrapolationstableaus* (Aitken-Neville-Algorithmus)

$$\begin{array}{ccccc} T_{11} & & & & \\ & \searrow & & & \\ T_{21} & \longrightarrow & T_{22} & & \\ & \searrow & & \searrow & \\ T_{31} & \longrightarrow & T_{32} & \longrightarrow & T_{33} \\ \vdots & & \vdots & & \vdots \end{array}$$

vermöge:

$$\begin{aligned} &\text{a)} \quad T_{i1} := y(H, h_i) \quad i = 1, 2, \cdots \\ &\text{b)} \quad k = 1, \cdots i \\ &\qquad T_{ik} := T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{n_i}{n_{i-k+1}}\right)^\gamma - 1} \end{aligned} \tag{2.18}$$

Aitken-Neville-Algorithmus liefert den Diskretisierungsfehler:

$$\varepsilon_{ik} := \| T_{ik} - y(H) \| \tag{2.19}$$

Bemerkung: Für Implementierung *glatte, skalierte* Norm verwenden! (vergleiche Kapitel C.1.2)
Führender Term:

$$\begin{aligned} \varepsilon_{ik} &\doteq \| e_k(H) \| \cdot (h_{i-k+1} \cdot \ldots \cdot h_i)^\gamma \\ &\doteq \gamma_{ik} \| e_k'(0) \| \cdot H^{\gamma k+1} \\ \gamma_{ik} &= (n_{i-k+1} \ldots n_i)^{-\gamma} \end{aligned} \tag{2.20}$$

Bemerkung:

$$e_k(0) = 0 \quad \text{falls} \quad y(0,h) = y_0 \ \forall h.$$

Ordnungs- und Schrittweitensteuerung (DEUFLHARD 1983)[30] :

Aufwand A_i : bei expliziten Verfahren:
Anzahl f-Auswertungen bis inklusive Zeile i.
"Subdiagonales" Fehlerkriterium:

$$\bar{\varepsilon}_{k+1,k} := \| T_{k+1,k} - T_{k+1,k+1} \| \leq \varepsilon \tag{2.21}$$

ε : vorgeschriebene relative Genauigkeit
Es gilt:

$$\bar{\varepsilon}_{k+1,k} \doteq \varepsilon_{k+1,k}$$

Schrittweite für Konvergenz in Position $(k+1,k)$:

$$H_k := \left(\frac{\varepsilon}{\bar{\varepsilon}_{k+1,k}} \right)^{\frac{1}{\gamma k+1}} \cdot H$$

Sicherheitsfaktor: $\varepsilon \rightarrow \rho\varepsilon$, $\rho = 1/4$ (Prädiktor).
Optimale Spalte q minimiert Aufwand pro Schrittweite (work per unit step):

$$\begin{aligned} W_k &:= \frac{A_{k+1}}{H_k} \cdot H \\ W_q &:= \min_{k=1,\cdots,k_{\text{fin}}} W_k \end{aligned} \tag{2.22}$$

Weitere Details: DEUFLHARD [30]
(Shannon'sche Informationstheorie findet Verwendung)

Explizites Euler-Verfahren ($p = \gamma = 1$) : h-Extrapolation.
Programm: EULEX (DEUFLHARD)

Harmonische Unterteilungsfolge:

$$\mathcal{F}_H := \{1, 2, 3, 4, \cdots\} \quad n_i = i \tag{2.23.a}$$

$$\begin{aligned} A_1 &:= n_1 \\ A_i &:= A_{i-1} + n_i - 1 \end{aligned} \tag{2.23.b}$$

Symmetrische Einschrittverfahren ($p = \gamma = 2$):

- implizite Trapezregel (2.3.c)
- implizite Mittelpunktsregel (2.3.d)

  Lösung mit Fixpunkt-Iteration würde h^2-Entwicklung stören!
  ($\hookrightarrow$ steife Probleme)

Explizite Mittelpunktsregel ($\rho = \gamma = 2$) : h^2 - Extrapolation

Symmetrische Diskretisierung:

$$\frac{y_{k+1} - y_{k-1}}{2h} = f(y_k) \tag{2.24.b}$$

Differentialgleichung 2. Ordnung: $y_1 = ?$

Startschritt (GRAGG 1964 [49])

$$\begin{aligned} y_o &\quad \text{gegeben:} \\ y_1 &= y_o + h\, f(y_0) \end{aligned} \tag{2.24.a}$$

expl. Euler-Startschritt

Frage: Stört unsymmetrischer Startschritt a) evtl. h^2- Entwicklung für symmetrische Diskretisierung?

Satz 2.5 (GRAGG 1964 [49])
Sei $f \in C^{2N+2}$ und Lipschitz-stetig mit Konstante L. Gegeben sei das Diskretisierungsverfahren:

$$\begin{aligned} a) \quad y_1 &= y_0 + h\, f(y_0) && \textit{(Startschritt)} \\ k &= 1, ..., l \quad (l \in \mathcal{F}) \\ b) \quad y_{k+1} &= y_{k-1} + 2h\, f(y_k), \\ c) \quad S_l &:= \tfrac{1}{4}[y_{l-1} + 2y_l + y_{l+1}] && \textit{(Schlußschritt)} \end{aligned} \tag{2.25}$$

Sei $t = l \cdot h$ fest. Dann gibt es ein $H > 0$ derart, daß für alle $h = t/l, k = 1, 2, \ldots$ eine Entwicklung der folgenden Form existiert:

$$\begin{aligned} a)\quad & y_l - y(t) = \\ & = \sum_{j=1}^{N} \left[u_j(t) + (-1)^l\, v_j(t)\right] h^{2j} + \\ & + E_{N+1}(t,h)\, h^{2N+2} \end{aligned} \tag{2.26}$$

Die (u_j, v_j) sind Lösungen eines gestaffelten DG-Systems der Form $(j = 1, 2, \ldots N)$

$$\begin{aligned} b)\quad u_j &= f_y\, u_j + \ldots \quad \textit{inhomogene Terme} \\ v_j &= -f_y\, v_j + \ldots \quad \textit{inhomogene Terme} \end{aligned}$$

mit Anfangsbedingungen der Form:

$$\begin{aligned} u_j(0) &= \tfrac{1}{2}\, g_j(0) \neq 0 \\ v_j(0) &= -\tfrac{1}{2}\, g_j(0) \neq 0 \end{aligned}$$

(g hier nicht genauer angegeben)
Das Restglied ist beschränkt in [0,H]. Für $l = 2m$ ergibt sich speziell:

$$\begin{aligned} a)\qquad S_l - y(t) &= \sum_{j=1}^{N} w_j(t) h^{2j} + W_{N+1}(t;h) h^{2N+2} \\ \textit{mit} & \\ b)\qquad w_1(t) &= u_1(t) + \tfrac{1}{4}\, y''(t) \\ w_1(0) &= \ldots = w_N(0) = W_{N+1}(0,h) = 0. \end{aligned} \tag{2.27}$$

Bemerkung: Die Fehlerkomponenten v_j heißen "schwach instabil". Beispiel:

$$\begin{aligned} y(t) &= e^{-t} + e^{-1000t} \doteq e^{-t} \quad \text{für } t \text{ "groß"} \\ \Longrightarrow u_j(t) & \sim e^{-t} \quad \text{für } t \text{ "groß"} \\ v_j(t) & \sim e^{+1000t}\ ! \end{aligned}$$

Beweis (Skizze):

(I) Man formuliert (nach STETTER 1970) [105] das Zweischrittverfahren als *symmetrisches Einschrittverfahren* doppelter Dimension. Dann lassen

sich die Sätze 2.3 und 2.4 anwenden.

$$\begin{aligned}
\text{Bez.:}\quad \bar{h} &:= 2h,\ \xi_k := y_{2k}\,, \zeta_k := y_{2k+1} - \tfrac{\bar{h}}{2} f(\xi_k) \\
\Longrightarrow\quad \zeta_0 &:= y_1 - \tfrac{\bar{h}}{2} f(y_0) = y_0 \\
\xi_0 &= y_0 \text{ ohnehin.} \\
\xi_{k+1} - \xi_k &= y_{2k+2} - y_{2k} = 2h\, f(y_{2k+1}) \\
&= \bar{h}\, f(\zeta_k + \tfrac{\bar{h}}{2} f(\xi_k)) \\
\zeta_{k+1} - \zeta_k &= y_{2k+3} - \tfrac{\bar{h}}{2} f(\xi_{k+1}) - y_{2k+1} + \tfrac{\bar{h}}{2} f(\xi_k) \\
&= \tfrac{\bar{h}}{2} [f(\xi_k) - f(\xi_{k+1}) + 2f(\xi_{k+1})] = \\
&= \tfrac{\bar{h}}{2} [f(\xi_k) + f(\xi_{k+1})]
\end{aligned}$$

Damit hat man das explizite Einschrittverfahren $(k = 0, 1, \cdots)$

$$\begin{aligned}
\text{a)}\quad \frac{\xi_{k+1} - \xi_k}{\bar{h}} &= f(\zeta_k + \tfrac{\bar{h}}{2} f(\xi_k))\,, \xi_0 = y_0 \\
\text{b)}\quad \frac{\zeta_{k+1} - \zeta_k}{\bar{h}} &= \tfrac{1}{2} [f(\xi_k) + f(\xi_{k+1})]\ , \zeta_0 = y_0
\end{aligned} \tag{2.28}$$

Diese $D_h G$ geht für $\bar{h} \to 0$ über in die *Differentialgleichung*:

$$\begin{aligned}
x' &= f(z)\,, \quad x(0) = y_0 \\
z' &= f(x)\,, \quad z(0) = y_0
\end{aligned}$$

Eindeutigkeitssatz liefert:

$$x(t) \equiv z(t) \equiv y(t)$$

Konsistenzbedingung ist also in (2.28) erfüllt. Nach Satz 2.3 existiert also eine *h-Entwicklung* der Form (2.14).

(II) (2.28.b) ist bereits symmetrisch Man versucht, für (2.28.a) ebenfalls eine symmetrische Darstellung zu finden. Argument von f in a) erweitert zu:

$$\begin{aligned}
&\zeta_k + \tfrac{\bar{h}}{2} f(\xi_k) + \tfrac{1}{2} \Big[\underbrace{\zeta_{k+1} - \zeta_k - \frac{\bar{h}}{2} f(\xi_k) - \frac{\bar{h}}{2} f(\xi_{k+1})}_{0} \Big] = \qquad (*) \\
&= \tfrac{1}{2}(\zeta_k + \zeta_{k+1}) + \tfrac{\bar{h}}{4}(f(\xi_k) - f(\xi_{k+1}))
\end{aligned}$$

Damit ist Bedingung (2.17) erfüllt (*symmetrisches Einschrittverfahren*).
Nach Satz 2.4 existiert also eine h^2-Entwicklung der Form:

$$\begin{aligned} \xi(t;h) - y(t) &= \sum_{j=1}^{N} p_j(t)\bar{h}^{2j} + P_{N+1}(t;\bar{h})\bar{h}^{2n+2} \\ \zeta(t;h) - y(t) &= \sum_{j=1}^{N} q_j(t)\bar{h}^{2j} + Q_{N+1}(t;\bar{h})\bar{h}^{2n+2} \end{aligned} \tag{2.29}$$

Nach Satz 2.4 gilt Beschränktheit der Restglieder P_{N+1}, Q_{N+1} sowie:

$$\begin{aligned} p_j' &= f_y(y(t))\ \ q_j + \dots \\ & \qquad\qquad \searrow \\ & \qquad\qquad \textit{inhomogene Terme} \\ q_j' &= f_y(y(t))\ \ p_j + \dots \\ & \qquad\qquad \nearrow \\ p_j(0) &= q_j(0) = 0 \text{ für } j = 1,2,\dots,N. \end{aligned} \tag{2.30}$$

(III) *Rücktransformation*

$$h = \frac{\bar{h}}{2}, \quad y_{2k} = \xi_k, \quad [y_{2k+1} = \zeta_k + hf(\xi_k)]$$

Für y_{2k+1} benützt man die symmetrische Erweiterung (∗):

$$y_{2k+1} = \frac{1}{2}(\zeta_k + \zeta_{k+1}) + \frac{h}{2}[f(\xi_k) - f(\xi_{k+1})] \tag{**}$$

Einsetzen der quadratischen Entwicklungen (2.28) in die Ausdrücke für y_{2k} und y_{2k+1} liefert Entwicklungen der Form:

$$\begin{aligned} y_{2k} - y(t_{2k}) &= \sum_{j=1}^{N} e_j(t_{2k})h^{2j} + E_{N+1}(t;h)h^{2N+2} \\ y_{2k+1} - y(t_{2k+1}) &= \sum_{j=1}^{N} g_j\,(t_{2k+1})h^{2j} + G_{N+1}(t;h)h^{2N+2} \end{aligned}$$

Ähnlich wie in (2.30) zeigt man:

$$\begin{aligned} e_j' &= f_y(y(t))\ \ g_j + \dots\,, \quad e_j(0) = 0 \\ g_j' &= f_y(y(t))\ \ e_j + \dots\,, \quad g_j(0) \neq 0 \end{aligned} \tag{2.31}$$

Die Entwicklungen für gerade und ungerade Indizes unterscheiden sich also, sind aber gekoppelt. Standardansatz für Entkopplung:

$$u_j := \frac{1}{2}(e_j + g_j)\,, v_j := \frac{1}{2}(e_j - g_j) \tag{2.32}$$

Damit (2.26.b) gezeigt. ∎

Wegen $e_j(0) = 0$, aber $g_j(0) \neq 0$ ist $l = 2m$ vorzuziehen. Damit Basis für h^2-Extrapolation, wobei $l = n_i = 2m_i$ zu wählen ist.
Doppelt-harmonische Unterteilungsfolge:

$$\begin{array}{lll} \text{a)} & F_{2H} := \{2,4,6,8,10,\ldots\} & n_i = 2i \\ \text{b)} & \left.\begin{array}{l} A_1 := n_1 + 1 \\ A_i := A_{i-1} + n_i \end{array}\right\} & \textit{mit} \text{ Schlußschritt} \end{array} \tag{2.33}$$

Programme: DIFEX1 (DEUFLHARD) [32]
ODEX (WANNER, HAIRER)

Bemerkung:

Die explizite Mittelpunktsregel mit h^2-Entwicklung heißt auch GBS-Verfahren (GRAGG [49], BULIRSCH/STOER [12]). Bulirsch und Stoer erkannten 1965 als erste die Wichtigkeit einer adaptiven Schrittweitensteuerung. Die von ihnen entwickelten algorithmischen Vorschläge (rationale Extrapolation, Trapezform des Tableaus, Folge $2\mathcal{F}_B$ statt $2\mathcal{F}_H$, spezielle Schrittweitensteuerung) sind inzwischen veraltet.

Zusammenfassung: Extrapolationsverfahren sind eine spezielle Technik, aus einfachen Diskretisierungen niedriger Ordnung Diskretisierungen beliebig hoher Ordnung zu erzeugen. Typisch ist die *simultane* Variation von Ordnung und Schrittweite.

2.4 Explizite Runge-Kutta-Methoden

(RUNGE 1895, HEUN 1900, KUTTA 1901) [97], [62], [78]

1. Idee: Abgebrochene Taylorreihe für y :

$$\begin{aligned} y(h) &= y_0 + h\,y_0' + \tfrac{h^2}{2}y''(0) + \mathcal{O}(h^3) \\ &= y_0 + h\,f(y_0) + \tfrac{h^2}{2}f_y(y_0)f(y_0) + \mathcal{O}(h^3) \end{aligned} \tag{2.34}$$

Abbruch nach 3. Term:

$$y_1 := y_0 + h\,f(y_0) + \frac{h^2}{2}f_y(y_0)f(y_0) \tag{2.34'}$$

Nach Konstruktion gilt:

$$y_1 - y(h) = \mathcal{O}(h^3), \text{ d. h. } p = 2$$

Nachteil: Konstruktion von Verfahren höherer Ordnung verlangt symbolische Behandlung von f für *jedes* einzelne Anfangswertproblem.

2. Idee: Ansatz
Anstelle von (2.34') wählt man:

$$y_1 := y_0 + h\,[b_1 f(y_0) + b_2 f(y_0 + a_{21} h\, f(y_0))] \tag{2.35}$$

Beispiel: explizites Euler-Verfahren + 1 Extrapolation:

$$\begin{aligned}
& T_{11} &&= y_0 + h\, f(y_0) \\
& T_{21} &&= y_0 + \tfrac{h}{2} f(y_0) + \tfrac{h}{2} f(y_0 + \tfrac{h}{2} f(y_0)) \\
\text{a)}\ & T_{22} &&= T_{21} + \frac{T_{21} - T_{11}}{2-1} = \\
& && = 2 \cdot T_{21} - T_{11} \\
& && = y_0 + h\, f(y_0 + \tfrac{h}{2} f(y_0)) \\
\text{b)}\ & b_1 &&= 0,\ b_2 = 1,\ a_{21} = \tfrac{1}{2} \\
\text{c)}\ & T_{22} &- \ & y(h) = \mathcal{O}(h^3)
\end{aligned} \tag{2.36}$$

Frage: Welchen Bedingungen müssen die Parameter b_1, b_2, a_{21} genügen, damit $\mathcal{O}(h^3)$ erzielt wird?

Entwicklung nach h-Potenzen:

$$\begin{aligned}
y_1 &= y_0 + b_1\, h\, f(y_0) \ + \ b_2 h\, [f(y_0) + a_{21}\, h\, f_y(y_0) f(y_0) + O(h^2)] \\
&= y_0 + h \cdot (b_1 + b_2)\, f(y_0) \ + \ h^2 b_2 a_{21} f_y(y_0) f(y_0) + O(h^3)
\end{aligned} \tag{2.37}$$

Beobachtung: Entwicklung (2.34) und (2.37) enthalten die gleichen Terme $f(y_0)$, $f_y(y_0)\, f(y_0)$.
Koeffizientenvergleich (2.34)/(2.37) :

$$\begin{aligned}
&\text{a)}\ \ b_1 + b_2 = 1 \\
&\text{b)}\ \ b_2 \cdot a_{21} = \tfrac{1}{2}
\end{aligned} \tag{2.38}$$

Sei Parameter $\beta = b_2$ eingeführt, so gilt:

$$b_1 = 1 - \beta\,, \quad a_{21} = \frac{1}{2\beta} \tag{2.38'}$$

Bemerkung: (2.36.b) erfüllt (2.38) $\Longrightarrow$ (2.36.c). Es verbleibt 1 Freiheitsgrad, um $\mathcal{O}(h^3)$ zu erreichen:

$$y_1 := y_0 + (1-\beta)hf(y_0) + \beta hf(y_0 + \frac{1}{2\beta}hf(y_0)) \tag{2.39}$$

RUNGE 1895: $\beta := 1$ vgl. (2.36.b)
HEUN 1900: $\beta := \frac{1}{2}$

Verallgemeinerung: Nach einer ersten Idee von RUNGE schlug KUTTA (1901)[78] folgendes Einschrittverfahren vor:

$$\begin{aligned}
\text{a)}\quad k_1 &:= h\,f(y_0)\\
k_2 &:= h\,f(y_0 + a_{21}k_1)\\
k_3 &:= h\,f(y_0 + a_{31}k_1 + a_{32}k_2)\\
&\vdots\\
k_s &:= h\,f(y_0 + a_{s1}k_1 + \ldots + a_{s,s-1}k_{s-1})\\
\text{b)}\quad y_1 &:= y_0 + b_1k_1 + b_2k_2 + \ldots + b_sk_s\\
\text{c)}\quad c_i &:= \sum_{j=1}^{i-1} a_{ij}
\end{aligned} \tag{2.40}$$

Bei *Stufenzahl* s bestimmen die $\frac{s(s+1)}{2}$ Parameter (a_{ij}, b_j) ein sogenanntes *explizites Runge-Kutta-Verfahren.*

Schema (nach BUTCHER 1963) [14]

$$\begin{array}{c|cccc}
0 & & & & \\
c_2 & a_{21} & & & \\
\vdots & a_{31} & a_{32} & & \\
\vdots & \vdots & \ddots & & \\
c_s & a_{s1} & \cdots & a_{s,s-1} & \\
\hline
 & b_1 & \cdots & b_{s-1} & b_s
\end{array}$$

Ziel: Bestimme die Parameter derart, daß möglichst hohe Ordnung p erreicht wird.

Um $\mathcal{O}(h^{p+1})$ zu erzielen, benötigt man Differentiationsformeln bis $y^{(p)}(0)$

in Taylorreihe sowie h-Entwicklung von y_1 . Mit wachsendem p geht die Übersicht rasch verloren. Deshalb wurde von BUTCHER 1963 eine *Graphen-Methode* entwickelt.

Bezeichnungen:

$$(y^J)' = f^J(y^1, \ldots, y^n) \quad J = 1, \ldots, n \tag{2.41}$$

komponentenweise Darstellung der DG

Für RK-Ansatz:

$$\begin{aligned} & & k_i &=: h\, f(g_i) \\ \text{a)} \quad g_i^J &= y_0^J + h \sum_{j=1}^{i-1} a_{ij} f^J(g_j^1, \ldots, g_j^n) \\ \text{b)} \quad y_1^J &= y_0^J + h \sum_{j=1}^{s} b_j f^J(g_j^1, \ldots, g_j^n) \\ \text{c)} \quad c_i &:= \sum_{j=1}^{i=1} a_{ij} \end{aligned} \tag{2.42}$$

Hochgestellter Index: Vektorkomponente. Große Indizes zur Vermeidung von Verwechslung bei Differentiation. Korrespondenz $(j,\ J)$, etc. Konstruktionsprinzip von Runge-Kutta-Verfahren.

Differentiation von (2.41) und (2.42) $\longrightarrow$ Abgleich der Koeffizienten.

Differentiation von (2.41):

$$\begin{aligned} \text{a)} \quad (y^J)^{(1)} &= f^J(y) \\ \text{b)} \quad (y^J)^{(2)} &= f_K^J(y) f^K(y) \\ & \qquad \hookrightarrow \text{Diff. nach } y^K \\ \text{c)} \quad (y^J)^{(3)} &= f_{KL}^J f^K f^L + f_K^J f_L^K f^L \end{aligned} \tag{2.43}$$

Einstein-Konvention: $\sum$ über doppelt auftretende Indizes.

Wegen Entwicklung um $h = 0 : y = y_0$ als Argument - im folgenden weggelassen.

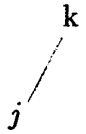
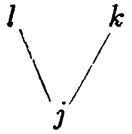
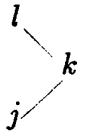
Differentiation von (2.42) nach h : Koeffizient $|_{h=0}$

$$
\begin{aligned}
\text{a)} \quad & g_i^J\,|_{h=0} = y_0^J \\
\text{b)} \quad & (g_i^J)^{(1)}\,|_{h=0} = \sum_{j=1}^{i-1} a_{ij} f^J\,|_{y=y_0} + h \ldots\,|_{h=0} \\
& = \sum_j a_{ij} f^J(y_0) \\
\text{c)} \quad & (g_i^J)^{(2)}\,|_{h=0} = \sum_{j,K} a_{ij} f_K^J a_{jk} f^K + \\
& \quad + \sum_j a_{ij} f_K^J(y_0) a_{jk} f^K + h \cdot \ldots\,|_{h=0} \\
& = 2 \sum_{j,k} a_{ij} a_{jk} f_K^J f^K(y_0)
\end{aligned}
\tag{2.44}
$$

Definition: Elementare Differentiale

$$F^J(t) := f^J_{K,\ldots}(y) f^K_{\ldots}(y) f^L(y) \cdots$$

Die genaue Abfolge der Indizes (hoch — tief) beschreibt man durch einen *(verwurzelten) indizierten Baumgraphen* (rooted labelled tree).

Graph t_{ik}:		tritt auf in:	
(t_{11})	$\bullet j$	f^J	y'
(t_{21})		$f_K^J f^K$	y''
(t_{31})		$f_{KL}^J f^K f^L$	y'''
(t_{32})		$f_K^J f_L^K f^L$	y'''

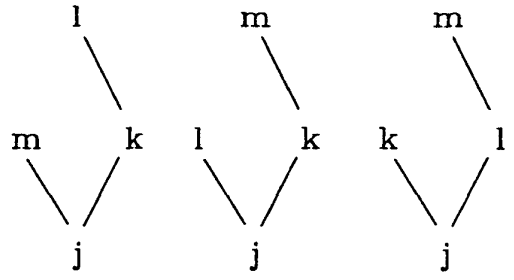
Summation über *alle* Knoten außer Wurzel (j).

Die Graphen t seien *monoton* indiziert: $j < k < l < m < \ldots$

$\varrho(t)$: Ordnung des Graphen t (=Anzahl der Knoten)

$\alpha(t)$: Anzahl verschiedener monotoner Indizierungen eines Graphen t.

Beispiel:



Satz 2.6 *Für die Lösung der DG gilt:*

$$(y^J)^{(q)}\,|_{t=0} = \sum_{t\in T_q} \alpha(t) F^J(t)(y_0) \tag{2.45}$$

wobei T_q *die Menge der Bäume* t *mit* $\varrho(t) = q$.

Beweis: Die Darstellung über t-Graphen erfaßt alle bei Differentiation auftretenden Terme genau einmal. ∎

Bei Differentiation von $y_1(h)$ nach h treten die gleichen elementaren Differentiale auf, wobei Nachdifferentiation jeweils $\sum_{j1,j2\ldots} a_{j1j2}$ bringt.

Details hier weggelassen - siehe HAIRER/NØRSETT/WANNER [58], S. 143-151. Für die Konstruktion von Runge-Kutta-Verfahren höherer Ordnung ist Graphentechnik unerläßlich (sonst zu viele Fehlerquellen!).

Satz 2.7 *Für die Lösung* y_1 *des Runge-Kutta-Ansatzes (2.40) gilt:*

$$(y_1^J)^{(q)}\,|_{h=0} = \sum_{t\in T_q} \alpha(t)\gamma(t) \sum_j b_j \Phi_j(t) F^J(t)(y_0) \tag{2.46}$$

wobei:

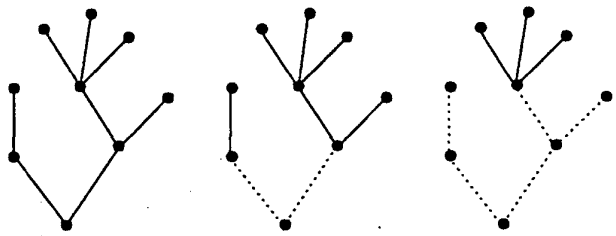
$$\Phi_j(t) := \underbrace{\sum_{k,l,\ldots}}_{q-1} \underbrace{a_{jk} a \ldots \cdot \ldots \cdot a \ldots}_{q-1} \quad \textit{Faktoren}$$

$a_{j_1 j_2}$: (j_1, j_2) *aufeinanderfolg. Indizes in* t

$\gamma(t) = \varrho(t) \cdot \varrho(t_1) \cdot \ldots \cdot \varrho(t_{j(t)})$

$t_1, \ldots, t_{j(t)}$: *Baumgraphen, die durch sukzessive Streichung aller Wurzeln entstehen*

Beispiel für Berechnung von $\gamma(t)$:



$$\varrho(t) = 9 \qquad \varrho(t_1) = 2 \qquad \varrho(t_2) = 6 \qquad \varrho(t_3) = 4$$

$\gamma(t) = 9 \cdot 2 \cdot 6 \cdot 4 = 432$

Zusammenfassung von Satz 2.7 und Satz 2.8 liefert die gewünschten *Bedingungsgleichungen.*

Satz 2.8 *Eine Runge-Kutta-Methode (2.40) ist für allgemeine f von der Ordnung p genau dann, wenn gilt:*

$$\sum_j b_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad \varrho(t) \leq p \tag{2.47}$$

Beweis: Falls alle elementaren Differentiale voneinander unabhängig sind und tatsächlich auftreten, ist (2.47) notwendig *und* hinreichend. ∎

Bemerkung: $\alpha(t)$ fällt beiderseits weg.

q	t	graph	$\gamma(t)$	$\alpha(t)$	$F^J(t)(y)$	$\Phi_j(t)$
0	$\varnothing$	$\varnothing$	1	1	y^J	
1	τ	$\bullet_j$	1	1	f^J	1
2	t_{21}		2	1	$\sum_K f^J_K f^K$	$\sum_k a_{jk}$
3	t_{31}		3	1	$\sum_{K,L} f^J_{KL} f^K f^L$	$\sum_{k,l} a_{jk} a_{jl}$
	t_{32}		6	1	$\sum_{K,L} f^J_K f^K_L f^L$	$\sum_{k,l} a_{jk} a_{kl}$
4	t_{41}		4	1	$\sum_{K,L,M} f^J_{KLM} f^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{jl} a_{jm}$
	t_{42}		8	3	$\sum_{K,L,M} f^J_{KM} f^K_L f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{jm}$
	t_{43}		12	1	$\sum_{K,L,M} f^J_K f^K_{LM} f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{km}$
	t_{44}		24	1	$\sum_{K,L,M} f^J_K f^K_L f^L_M f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{lm}$
5	t_{51}		5	1	$\sum f^J_{KLMP} f^K f^L f^M f^P$	$\sum a_{jk} a_{jl} a_{jm} a_{jp}$
	t_{52}		10	6	$\sum f^J_{KMP} f^K_L f^L f^M f^P$	$\sum a_{jk} a_{kl} a_{jm} a_{jp}$
	t_{53}		15	4	$\sum f^J_{KP} f^K_{ML} f^L f^M f^P$	$\sum a_{jk} a_{kl} a_{km} a_{jp}$
	t_{54}		30	4	$\sum f^J_{KP} f^K_L f^L_M f^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{jp}$
	t_{55}		20	3	$\sum f^J_{KM} f^K_L f^L f^M_P f^P$	$\sum a_{jk} a_{kl} a_{jm} a_{mp}$
	t_{56}		20	1	$\sum f^J_K f^K_{LMP} f^L f^M f^P$	$\sum a_{jk} a_{kl} a_{km} a_{kp}$
	t_{57}		40	3	$\sum f^J_K f^K_{LP} f^L_M f^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{kp}$
	t_{58}		60	1	$\sum f^J_K f^K_L f^L_{MP} f^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{lp}$
	t_{59}		120	1	$\sum f^J_K f^K_L f^L_M f^M_P f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{mp}$

Tabelle A.1 Baumgraphen und elementare Differentiale bis Ordnung 5

q	1	2	3	4	5	6	7	8	9	10
card(T_q)	1	21	2	4	9	20	48	115	286	719
Anzahl der Bedingungen	1	2	4	8	17	37	85	200	486	1205

Tabelle A.2 Anzahl der Baumgraphen bis Ordnung 10

Führender Fehlerterm:

$$y_1^J - y^J(h) = \frac{h^{p+1}}{(p+1)!} \sum_{t \in T_{p+1}} \alpha(t) e(t) F^J(t)(y_0) + \\ + O(h^{p+2}) \tag{2.48}$$

wobei

$$e(t) := \gamma(t) \sum_j b_j \Phi_j(t) - 1$$

Zur Konstruktion eines effektiven Runge-Kutta-Verfahrens wird man verbleibende freie Parameter derart einstellen, daß:

$$\sum_{t \in T_{p+1}} \alpha(t) \mid e(t) \mid = \min \tag{2.49}$$

Schrittweitensteuerung:

Sei y_1 Runge-Kutta-Resultat $O(h^{p+1})$

$\hat{y}_1$ Runge-Kutta-Resultat $O(h^{p+2})$

Sei angenommen, daß $\hat{y}_1$ die wesentlich genauere Lösung ist:

$$\| \hat{y}_1 - y(h) \| \ll \| y_1 - y(h) \| \tag{2.50}$$

Dann gilt:

$$\hat{\varepsilon} := \| y_1 - \hat{y}_1 \| \doteq \| y_1 - y(h) \| \tag{2.51}$$

Bemerkung: Fehlerschätzung nur für "zweitbeste" Approximation von $y(h)$.

Sei ε vom Benutzer gewünschte lokale relative Genauigkeit. Dann muß gelten:

$$\hat{\varepsilon} \leq \varepsilon \tag{2.52}$$

Andererseits gilt:

$$\hat{\varepsilon} \doteq C \cdot h^{p+1} \tag{2.53.a}$$

Für vernünftige Schrittweite $\bar{h}$ sollte gelten:

$$C \bar{h}^{p+1} \doteq \varrho \varepsilon \tag{2.53.b}$$

Mit Sicherheitsfaktor $\varrho < 1$. Division liefert:

$$\hat{h} \doteq h \cdot \left(\frac{\varrho \varepsilon}{\hat{\varepsilon}} \right)^{\frac{1}{p+1}}$$

Zur Gewinnung von $\hat{y}_1$ gibt es 2 typische Methoden:

Extrapolation:

$$\begin{array}{lll} y_1(2h) & \searrow & \\ y_2(h) & \longrightarrow & \hat{y}_1(2h) \quad \text{mit (2.18.b)} \end{array}$$

$$\begin{array}{lll} y_2(h) - y(2h) & \doteq & C \cdot 2h \cdot \overbrace{\varphi(L_1 2h)}^{\doteq 1} \cdot h^p \\ y_1(2h) - y(2h) & \doteq & C \cdot (2h)^{p+1} \end{array} \qquad (*)$$

Subtraktion:

$$y_2(h) - y_1(2h) \doteq 2C \cdot h^{p+1} \cdot (1 - 2^p)$$

Auflösung nach C, Einsetzen in $(*)$:

$$\begin{array}{rcl} \hat{y}_1(2h) & := & y_2(h) + (y_2(h) - y_1(2h))/(2^p - 1) \\ \hat{\varepsilon} & := & \| \hat{y}_1(2h) - y_2(h) \| = \mathcal{O}(h^{p+1}) \end{array} \qquad (2.54)$$

Einbettung:

y_1 Runge-Kutta-Verfahren der Ordnung p

$\hat{y}_1$ Runge-Kutta-Verfahren der Ordnung $(p+1)$

FEHLBERG (1964)[45] : Rechnung fortgesetzt mit y_1
(Widerspruch zu Fehlerschätzformel!), vgl. (2.50)!

$\Rightarrow$ Programme RKF4(5), RKF7(8)

DORMAND/PRINCE (1980)[41]:

$\longrightarrow$ DOPRI 5(4) : (2.49) erfüllt für $\hat{y}_1$
$s = 6$ Stufen

$\longrightarrow$ DOPRI 8(7) : (2.49) nicht ganz erfüllt.

($\rightarrow$ HAIRER/NØRSETT/WANNER [58], S. 194/5)

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	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
$\hat{y}_1$	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
y_1	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Tabelle A.3 Dormand-Prince 5(4) (DOPRI5)

"*Fehlberg-Trick*":

$$\begin{aligned} h\,f(y_1) &= k_s \\ (a_{si} &= b_i, b_s = 0) \\ \Rightarrow c_s &= 1 \end{aligned}$$

Interpolation: Für graphischen Output manchmal Lösung an Punkten zwischen den von der Schrittweitensteuerung gewählten verlangt:

$$y_1(\Theta h)$$

Dafür möchte man möglichst keine zusätzlichen f-Auswertungen, sondern $k_1, \ldots, k_s$ bzw. $g_1, \ldots, g_s$ verwenden. Dazu konstruiert man interpolierende Polynome $\longrightarrow$ Hermite-Interpolation auf der Basis von y *und* y'.

Beispiele:

(I) Bei Schrittweitensteuerung mittels Extrapolation hat man:

$$y_0, f(y_0),\ y_1, f(y_1),\ y_2$$

Zusätzliche Berechnung von $f(y_2)$ oder "Fehlberg-Trick" gestattet Hermite-Interpolation der Ordnung 5.

(II) DOPRI 5 gestattet direkte Interpolation der Ordnung 4:

$$\begin{aligned}
b_1(\Theta) &= \Theta(1+\Theta(-1337/480+\Theta(1039/360+\Theta(-1163/1152))))\\
b_2(\Theta) &= 0\\
b_3(\Theta) &= 100\Theta^2(1054/9275+\Theta(-4682/27825+\Theta(379/5565)))/3\\
b_4(\Theta) &= -5\Theta^2(27/40+\Theta(-9/5+\Theta(83/96)))/2\\
b_5(\Theta) &= 18225\Theta^2(-3/250+\Theta(22/375+\Theta(-37/600)))/848\\
b_6(\Theta) &= -22\Theta^2(-3/10+\Theta(29/30+\Theta(-17/24)))/7
\end{aligned} \tag{2.55.a}$$

$$y(x_0+\Theta h) \approx y_0 + h\sum_{j=1}^{6} b_j(\Theta)k_j \cdot \tag{2.55.b}$$

wobei die rechte Seite für $\Theta \to 1$ in $\hat{y}$ übergeht.

Weitere Anwendungen der Interpolation:

- on-line Steuerung von Prozessen
- Differentialgleichungen mit retardiertem Argument
- Bestimmung von Schaltpunkten bei impliziten Umschaltbedingungen oder Unstetigkeiten
  (Beispiel: Stoßdämpfer, optimale Steuerung vgl. Kap. D.5.2)

Bemerkung: Wünschenswert wären eingebettete Runge-Kutta-Verfahren variabler Ordnung (z.B. Programm VOBET, von D. G. BETTIS, dort aber falsche Koeffizienten!)

3 Mehrschrittverfahren

Alternative zur Konstruktion von Verfahren höherer Ordnung, am Beispiel erläutert:

$$y(t_{n+2}) - y(t_n) = \int_{t_n}^{t_{n+2}} f(y(t))\, dt \tag{3.1}$$

Simpson-Regel für Integral angewendet:

$$y_{n+2} - y_n = \frac{h}{3}[f(y_{n+2}) + 4f(y_{n+1}) + f(y_n)] \tag{3.2}$$

Milne-Simpson-Verfahren (Implizites Zweischrittverfahren)
Allgemein: Lineares k-Schritt-Verfahren

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \ldots + \alpha_o y_n = h\,[\beta_k f(y_{n+k}) + \ldots + \beta_0 f(y_n)] \tag{3.3}$$

$n = 0, 1, 2, \ldots$	
$\alpha_k := 1$	o.B.d.A.
$\mid \alpha_0 \mid + \mid \beta_0 \mid > 0$	o.B.d.A.
$\beta_k = 0$:	explizit
$\beta_k \neq 0$:	implizit (Lösung z.B. mit Fixpunktiteration)

Eindeutige Lösung von (3.3) verlangt Startwerte

$$y_0,\ y_1,\ \ldots,\ y_{k-1}$$

etwa erhalten aus Mehrschrittverfahren niedrigerer Schrittzahl ("Startrampe").

Bezeichnung: Verschiebeoperator E_h

$$E_h y_k = y_{k+1},\ E_h f(y(t)) = f(y(t+h)) \tag{3.4}$$

Einführung in (3.3):

$$\rho(E_h) y_n = h\sigma(E_h) f(y_n) \tag{3.3'}$$

wobei ρ, σ *charakteristische Polynome:*

$$\rho(\zeta) := \alpha_k \zeta^k + \ldots + \alpha_0\ ,\ \ \sigma(\zeta) := \beta_k \zeta^k + \ldots + \beta_0 \tag{3.5}$$

Beispiele:

(1) Milne Simpson - vergleiche (3.2):

$$\rho(\zeta) = \zeta^2 - 1, \quad \sigma(\zeta) = \frac{1}{3}[\zeta^2 + 4\zeta + 1]$$

(2) explizite Mittelpunktsregel - vergleiche (2.24.b):

$$\rho(\zeta) = \zeta^2 - 1, \; \sigma(\zeta) = 2\zeta$$

3.1 Konvergenz

(3.3) bzw. (3.3') ist D_hG k-ter Ordnung.

Konsistenz: $D_hG \Longrightarrow DG$

Lösung $y(t)$ der DG eingesetzt in D_hG liefert *Konsistenzordnung* $p > 0$.

$$\rho(E_h)y(t) - h\,\sigma(E_h)f(y(t)) =: \mathcal{O}(h^{p+1}) \tag{3.6}$$

Taylor-Entwicklung:

$$y(t + mh) = y(t) + mh\; y'(t) + \frac{1}{2}(mh)^2 y''(t) + \ldots$$

$$hf(y(t + mh)) = hy'(t + mh) = h\,[y'(t) + mh\; y''(t) + \ldots]$$

Einsetzen in (3.6):

$$\begin{aligned} &\rho(E_h)y(t) - h\sigma(E_h)y'(t) =: \\ &=: C_0 y(t) + C_1 h y'(t) + \ldots + C_q h^q y^{(q)}(t) + \ldots \end{aligned} \tag{3.7.a}$$

wobei die Koeffizienten $C_0, C_1, \ldots$ gegeben sind durch:

$$\begin{aligned} C_0 &= \alpha_k + \alpha_{k-1} + \ldots + \alpha_0 \equiv \rho(1) \\ C_1 &= k\alpha_k + \ldots + \alpha_1 - (\beta_k + \ldots + \beta_0) \equiv \rho'(1) - \sigma(1) \\ C_q &= \frac{1}{q!}[k^q\alpha_k + \ldots + \alpha_1] - \frac{1}{(q-1)!}[k^{q-1}\beta_k + \ldots + \beta_1] \end{aligned} \tag{3.7.b}$$

Konsistenzbedingung für $p > 0$ lautet also:

$$C_0 = C_1 = \ldots = C_p = 0 \tag{3.7.c}$$

Speziell muß gelten:

$$\rho(1) = 0, \rho'(1) - \sigma(1) = 0 \tag{3.8}$$

Entwicklung (3.7) gilt unabhängig von Lösung $y(t)$, also auch für die spezielle Wahl

$$y'(t) = y(t) = e^t \, .$$

Setze $t = 0$ in (3.7.a) und berücksichtige

$$E_h^k y(t) \mid_{t=0} = (e^h)^k.$$

Dann folgt unmittelbar:

$$\begin{aligned} &\rho(e^h) - h\,\sigma(e^h) = \\ &= C_0 + C_1 h + \ldots + C_q h^q + \ldots \end{aligned} \tag{3.9}$$

Konsistenzbedingung für $p > 0$:

$$\rho(e^h) - h\sigma(e^h) = \mathcal{O}(h^{p+1}) \qquad h \to 0 \tag{3.10}$$

Umformung:

$$\zeta = e^h, h = ln\zeta, \rho(\zeta) - ln\zeta\sigma(\zeta) = \mathcal{O}((\zeta - 1)^{p+1})$$

wegen $\rho(1) = 0$ und $ln\zeta = \zeta - 1 + \mathcal{O}((\zeta - 1)^2)$:

$$\frac{\rho(\zeta)}{ln\zeta} - \sigma(\zeta) = \mathcal{O}((\zeta - 1)^p). \quad \zeta \to 1 \tag{3.10'}$$

Formeln (3.10) und (3.10') für Beweiszwecke (HENRICI).

Aber (RUTISHAUSER 1952) [99]: konsistente Formeln auch hoher Ordnung sind eventuell *instabil.*

Lemma 3.1 (DAHLQUIST 1956) [19]
Seien $\zeta_1, \ldots, \zeta_k$ Wurzeln von $\rho(\zeta) = 0$. Notwendig für Stabilität des Mehrschrittverfahrens (3.3') ist:

$$| \zeta_i | \leq 1 \quad i = 1, \ldots, k \quad (\zeta_1 := 1 \;\; o.B.d.A) \tag{3.11}$$

Falls $| \zeta_j | = 1 : \zeta_j$ einfache Nullstelle ("Dahlquist'sches Wurzelkriterium").

Beweis: Man betrachtet das *Modellproblem*

$$y' = 0, \;\; y(0) = 1 \Longrightarrow y(t) \equiv 1 \tag{3.12}$$

Zugehöriges Mehrschrittverfahren

$$\rho(E_h) y_n = 0$$

Lineare D_hG k-ter Ordnung hat Fundamentallösungen:

$$\zeta^n, n\zeta^n, \ldots, n^{m-1}\zeta^n$$

ζ : Wurzel von ρ

m : zugehörige Vielfachheit ($m \leq k$)

Eine Wurzel liegt fest wegen Konsistenzbedingung (3.8):

$$\rho(1) = 0,\ \zeta_1 := 1 \quad \text{o.B.d.A} \tag{3.13.a}$$

Diese Wurzel repräsentiert die konstante Lösung $y_n = y_0$; sie muß also *einfach* sein, woraus mit (3.8) folgt:

$$\sigma(1) = \rho'(1) \neq 0 \tag{3.13.b}$$

Stabilität heißt: Nebenlösungen dürfen die Hauptlösung nicht "überwuchern". Das liefert (3.11). ∎

Obwohl das Lemma 3.1 nur auf dem trivialen Modellproblem $y' = 0$ aufbaut, hat man mit der Wurzelbedingung die wesentliche zusätzliche Bedingung für Konvergenz ($h \to 0$) gefunden.

Satz 3.1 ("Konsistenz + Stabilität $\Longleftrightarrow$ Konvergenz")
Stabilität gemäß (3.11) und Konsistenz der Ordnung $p > 0$ sind notwendig und hinreichend für Konvergenz der Ordnung p. Hierbei sind Approximationsfehler der Startwerte als $\mathcal{O}(h^p)$ vorausgesetzt.

Beweis (Skizze): Zunächst Umformung des Mehrschrittverfahrens auf Einschrittverfahren entsprechend höherer Dimension:

$$Y_n := (y_{n+k-1}, \ldots, y_n)^T \quad n \geq 0 \tag{3.14.a}$$

Zur Vereinfachung sei nur 1 Differentialgleichung behandelt. Mehrschrittverfahren als System ($\alpha_k := 1$):

$$\begin{aligned} y_{n+k} &= -[\alpha_{k-1}y_{n+k-1} + \ldots + \alpha_0 y_n] - \\ &- h\,[\beta_k f(y_{n+k}) + \ldots + \beta_0 f(y_n)] \end{aligned} \tag{$*$}$$

Definition einer Inkrementfunktion:

$$\Phi(y_{n+k-1}, \ldots, y_n, h) =: \Phi(Y_n, h)$$

für $\beta_k \neq 0$ formal nur implizit möglich:

$$\begin{aligned}\Phi(\cdot,h) = \quad & \beta_k f(-\alpha_{k-1}y_{n+k-1} - \ldots - \alpha_0 y_n - h\ \Phi(\cdot,h)) \\ & +\beta_{k-1}\ f(y_{n+k-1}) + \ldots + \beta_0\ f(y_n)\end{aligned} \tag{3.14.b}$$

Damit lautet $(*)$:

$$y_{n+k} = -[\alpha_{k-1}y_{n+k-1} + \ldots + \alpha_0 y_n] - h\Phi(\cdot,h) \qquad (**)$$

Hinzu kommen noch Identitäten durch Verschiebung des Index:

$$Y_{n+1} = \begin{bmatrix} y_{n+k} \\ y_{n+k-1} \\ \vdots \\ y_{n+1} \end{bmatrix} \begin{matrix} \swarrow \\ \vdots \\ \swarrow \end{matrix} \begin{bmatrix} y_{n+k-1} \\ \vdots \\ y_{n+1} \\ y_n \end{bmatrix} = Y_n$$

Matrix-Schreibweise:

$$A := \begin{bmatrix} -\alpha_{k-1} & -\alpha_{k-2} & \ldots & -\alpha_0 \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{bmatrix} \tag{3.14.c}$$

$$Y_{n+1} = AY_n + h\Phi(Y_n,h). \tag{3.14'}$$

Für DG-Systeme:

$$\begin{aligned} A &\rightarrow A \otimes I \\ \Phi &\rightarrow (e_1 \otimes I)\Phi \end{aligned}$$

$\hookrightarrow$ Beweistechnik etwas komplizierter.

Die Eigenwerte von A sind die Wurzeln von ρ, da ρ charakteristisches Polynom zu A (Frobenius-Matrix). Also gilt ($\bar{\rho}$: Spektralradius)

$$\bar{\rho}(A) \leq 1 \quad (3.11) \tag{3.15}$$

Im allgemeinen folgt daraus die Existenz einer Matrixnorm $\|\cdot\|_\varepsilon$ mit

$$\| A \|_\varepsilon \leq 1 + \varepsilon \ , \quad \varepsilon > 0 \text{ beliebig.}$$

Da jedoch die betragsgrößten Eigenwerte von A *einfach* sind, gilt hier sogar:

$$\| A \| \leq 1 \tag{3.15'}$$

Seien Y_{n+1}, Z_{n+1} definiert durch (3.14') aus Y_n, Z_n. Sei L^* Lipschitzkonstante zu Φ. Dann gilt mit (3.15'):

$$\| Y_{n+1} - Z_{n+1} \| \leq (1 + hL^*) \| Y_n - Z_n \| \tag{3.16}$$

Dies ist Basis für ein "diskretes Fundamentallemma". Um die Beweistechnik ("Fächer") von Satz 2.1 anwenden zu können, benötigt man noch den "lokalen Diskretisierungsfehler".

Sei

$$\hat{Y}_{n+1} := A\, Y(t_n) + h\Phi(Y(t_n), h)$$

Dann gilt (streng nur für $p > 1$):

$$\| \hat{Y}_{n+1} - Y(t_{n+1}) \| \leq M h^{p+1} \tag{3.17}$$

Fächer-Technik liefert ($p > 1, L^*$: Lipschitzkonstante zu Φ):

$$\| Y_m - Y(t_m) \| \leq \| Y_0 - Y(t_0) \| \exp(L^* t_m) + M\, h^p t_m \varphi(L^* t_m) \tag{3.18}$$

$$Y_0 - Y(t_0) = \begin{bmatrix} y_{k-1} - y(t_{k-1}) \\ \vdots \\ y_1 - y(h) \\ 0 \end{bmatrix}$$

Approximationsfehler der Startwerte. ∎

Bemerkung:

(1) Bei Verwendung einer "Startrampe" durch Mehrschrittverfahren niedrigerer Ordnung gilt (Start: expl. Euler):

$$y_j - y(jh) = \mathcal{O}(h^2) \quad j = 1, \ldots, k-1 \tag{3.19}$$

Damit $\mathcal{O}(h^p)$ - Abschätzung in (3.18) gestört!

(2) Alternativen:

- Taylor-Entwicklung für Startwerte (ADAMS [1])
- Einschrittverfahren von Ordnung p für Startwerte (u.a. GEAR, aber erst jüngst wirklich realisiert)

Für tatsächliche Konstruktion von Mehrschrittverfahren hat man nach (3.7) $(2k+1)$ freie Parameter für *impliziten* Fall, $2k$ freie Parameter für *expliziten* Fall! Man erwartet also rein algebraisch eine Konsistenzordnung $p = 2k$ (impl.) bzw. $p = 2k - 1$ (expl.). Die Stabilitätsbedingung (3.11) liefert jedoch *wesentliche* Einschränkung.

Satz 3.2 ("1.DAHLQUIST-SCHRANKE", 1956 [19])
Für die Ordnung p eines stabilen linearen k-Schritt-Verfahrens gilt:

$$\begin{aligned} &a) \quad p \leq k+2 \quad \textit{für } k \textit{ gerade} \\ &b) \quad p \leq k+1 \quad \textit{für } k \textit{ ungerade} \\ &c) \quad p \leq k \qquad \textit{für } \beta_k \leq 0 \quad (\alpha_k = 1) \end{aligned} \tag{3.20}$$

Beweis hier weggelassen: i.w. funktionentheoretische Hilfsmittel, Reihenentwicklung (⟶ HAIRER, NØRSETT, WANNER, [58], S. 332-335).

Bemerkung zu Stabilität. Falls mehrere einfache Wurzeln von ρ auf Einheitskreis, heißt Mehrschrittverfahren *schwach instabil.* So sind für $p = k+2$, k gerade, *alle* Wurzeln auf dem Einheitskreis (ohne Beweis).

$$\hookrightarrow p \leq k+1 \quad \text{falls stark stabil} \tag{3.21}$$

3.2 Adams-Verfahren

Falls $\beta_k \neq 0$:	$p \leq k+1$	nach (3.21)
	$2k+1$	freie Parameter
Falls $\beta_k = 0$:	$p \leq k$	nach (3.20.c)
	$2k$	freie Parameter

Vorgehen in beiden Fällen: Man wird über k freie Parameter derart verfügen, daß das Wurzelkriterium (3.11) erfüllt wird. Im Sinne der Stabilität ist die bestmögliche Wahl:

$$\rho_k(\zeta) := \zeta^{k-1}(\zeta - 1) \tag{3.22}$$

Zugehöriges Mehrschrittverfahren:

$$y_{n+k} - y_{n+k-1} = h\,[\beta_k\, y_{n+k} + \ldots + \beta_0\, y_n] \tag{3.22'}$$

die noch verbleibenden Parameter

$$(\beta_k),\ \beta_{k-1}, \ldots,\ \beta_0$$

sind so zu bestimmen, daß p maximal wird. Statt formaler Lösung der Gleichungen (3.7) liefert die "richtige Idee" die β's direkt:

Adams-Verfahren (ADAMS 1855, [1] BASHFORTH 1883, [4], MOULTON 1926, [87]). Mit Blick auf (3.22′) schreibt man die Differentialgleichung um in eine (formale) Integralgleichung:

$$y(t_n) = y(t_{n-1}) + \int_{t=t_{n-1}}^{t_n} f(y(t))dt \tag{3.23}$$

Approximationsidee:

$$f(y(t)) \longrightarrow p(t) \quad \text{Interpolations-Polynom} \tag{3.24}$$

Adams - Bashforth (AB): explizit

$$p(t) := P_{k-1}(t \mid t_{n-1}, \ldots, t_{n-k})$$

Adams - Moulton (AM): implizit

$$p(t) := P_k(t \mid t_n, \ldots, t_{n-k})$$

(Darstellungen für das Interpolationspolynom nach AITKEN).
Man definiert:

$$y_n := y_{n-1} + \int_{t=t_{n-1}}^{t_n} p(t)dt \tag{3.25}$$

Bezeichnung: $f_n := f(y_n)$.

(a) *AB-Verfahren über äquidistantem Gitter*
Man geht etwa aus von *Newton'schen dividierten Differenzen*:

$$\begin{aligned} p(t) &:= P_{k-1}(t \mid t_{n-1}, \ldots, t_{n-k}) = \\ &= f[t_{n-1}] + f[t_{n-1}, t_{n-2}] \cdot (t - t_{n-1}) + \ldots \\ &\quad \ldots + f[t_{n-1}, \ldots, t_{n-k}] \cdot (t - t_{n-1}) \cdot \ldots \cdot (t - t_{n-k+1}) \\ P_{k-1} &= f_{n-1} + \nabla^1 f_{n-1} \cdot \frac{t - t_{n-1}}{h} + \ldots \\ &\quad \ldots + \nabla^{k-1} f_{n-1} \cdot \frac{(t - t_{n-1}) \cdot \ldots \cdot (t - t_{n-k+1})}{(k-1)! h^{k-1}} \end{aligned} \tag{3.26}$$

wobei

$$\nabla := I - E_{-h} \quad \text{"Rückwärtsdifferenzen"-Operator}$$

Bemerkung:

$$\underbrace{\nabla^{q+1} f_j = \nabla^q f_j - \nabla^q f_{j-1}}_{\text{für Auswertung}} \equiv \nabla^q (I - E_{-h}) f_j$$

Einsetzen in (3.25):

$$y_n - y_{n-1} = \int_{t=t_{n-1}}^{t_n} P_{k-1}(t)dt \quad (\text{subst: } s := \frac{t - t_{n-1}}{h})$$

$$= h \int_{s=0}^{1} \left[f_{n-1} + s\nabla f_{n-1} + \ldots + \frac{s(s+1)\cdots(s+k-2)}{(k-1)!} \nabla^{k-1} f_{n-1} \right] ds$$

Ausführung der s-Integration:

$$\begin{aligned} y_n &:= y_{n-1} + h \sum_{i=0}^{k-1} \gamma_i \nabla^i f_{n-1} \\ \gamma_i &:= \int_{s=0}^{1} \frac{s(s+1)\cdot\ldots\cdot(s+i-1)}{i!} ds \quad i > 0 \ \ (\gamma_0 := 1) \end{aligned} \tag{3.27}$$

Umformung in Standardform (3.22'):

$$y_n := y_{n-1} + h \sum_{j=1}^{k} \beta_{k,k-j} f_{n-j} \tag{3.27'}$$

wobei: $\{\beta_{k,k-j}\}$ berechenbar aus $\{\gamma_i\}$ oder aus Lagrange-Darstellung von P_{k-1}.

Diskretisierungsfehler ϵ_n abgeschätzt über Interpolationsfehler bezüglich f:

$$\epsilon_n = \| y_n - y(t_n) \| \tag{3.28}$$

Subtraktion von (3.25) und (3.23) liefert

$$\epsilon_n \leq \epsilon_{n-1} + \int_{t=t_{n-1}}^{t_n} \| f(y(t)) - p(t) \| \, dt \tag{3.29}$$

$$\begin{aligned} \| f(y(t)) - P_{k-1}(t) \| &\leq \frac{M_k}{k!} \cdot (t - t_{n-1}) \cdot \ldots \cdot (t - t_{n-k}) \\ M_k &:= \max_{\tau \in]t_{n-1}, t_{n-k}[} \| \underbrace{f^{(k)}(y(\tau))}_{y^{(k+1)}(\tau)} \| \end{aligned} \tag{3.30}$$

Einsetzen in (3.27):

$$\epsilon_n \leq \epsilon_{n-1} + M_k \cdot \int_{t=t_{n-1}}^{t_n} \frac{(t-t_{n-1}) \cdot \ldots \cdot (t-t_{n-k})}{k!} dt$$

Substitution: $s := \dfrac{t-t_{n-1}}{h}$

$$\epsilon_n \leq \epsilon_{n-1} + M_k \cdot h^{k+1} \underbrace{\int_{s=0}^{1} \frac{s(s+1) \cdot \ldots \cdot (s+k-1)}{k!} ds}_{\text{vgl. (3.27)} := \gamma_k}$$

$$\epsilon_n \leq \epsilon_{n-1} + \gamma_k M_k h^{k+1} \tag{3.31}$$

das heißt Konsistenzordnung ist $p = k$. "Startrampe": $k = 1, 2, \ldots$
Bei Wechsel von k ist Darstellung (3.27) vorzuziehen gegenüber (3.27′).

b) *ABM-Verfahren über äquidistantem Gitter*
Die Approximation

$$p(t) := P_k(t \mid t_n, \ldots, t_{n-k})$$

liefert ein *implizites* Verfahren. Die Darstellung über diverse Differenzen eignet sich deshalb nicht. Stattdessen liefert Lagrange - Darstellung von P_k das Analogon zu (3.27′):

$$y_n = y_{n-1} + h \cdot \beta^*_{k,k} f_n + h \sum_{j=1}^{k} \beta^*_{k,k-j} f_{n-j} \tag{3.32}$$

Lösung dieses nichtlinearen Gleichungssystems durch *Fixpunktiteration* möglich, da "nichtsteife" Differentialgleichung integriert werden sollen. Man erhält das sogenannte *Prädiktor-Korrektor-Verfahren.* Wählt man als Startwert den AB-Wert, so erhält man das *ABM-Verfahren:*

$$\begin{aligned} &\text{a)} \quad y^0_n := y_{n-1} + h \sum_{j=1}^{k} \beta_{k,k-j} f_{n-j} \\ &\qquad \text{P: predictor} \\ &\text{b)} \quad y^{i+1}_n := y_{n-1} + h\beta^*_{k,k} f^i_n + h \sum_{j=1}^{k} \beta_{k,k-j} f_{n-j} \\ &\qquad \text{C: corrector} \quad \text{E: evaluate} \end{aligned} \tag{3.33}$$

Sei L die Lipschitzkonstante von f. Dann ist hinreichend für Konvergenz der Iteration:

$$h\, \beta^*_{k,k}\, L < 1 \tag{3.34}$$

Diese Bedingung ist erfüllbar, falls h "hinreichend klein". Falls keine Konvergenz für $i \leq i_{\max}$: Schrittweitenreduktion.

Varianten des ABM-Verfahrens:

(I) PECE-Verfahren:
$y_n^0, f_n^0, y_n^1, f_n^1$ berechnet
$f_n := f_n^1$ im nächsten Schritt

(II) P(EC) * oder P(EC) * E
je nachdem, ob f_n^i oder f_n^{i+1} für nächsten Schritt verwendet wird.

Schrittweiten- und Ordnungssteuerung (GEAR [48] 1971, ähnlich KROGH [72] 1969)
Bei Änderung der Schrittweite benötigt man *Zwischenwerte von f*, die man durch *Interpolation mit p* in natürlicher Weise gewinnt. Man benötigt also eine Darstellung, die für Schrittweitenänderung besonders geeignet ist. Dazu geht man aus von Taylor-Darstellung von p (hier nur für Prädiktor vorgeführt):

$$P_{k-1}(t) = P_{k-1}(t_{n-1}) + P'_{k-1}(t_{n-1}) \cdot (t - t_{n-1}) + \ldots + P_{k-1}^{(k-1)}(t_{n-1}) \frac{(t - t_{n-1})^{k-1}}{(k-1)!}$$

Einsetzen in (3.25):

$$\begin{aligned} y_n^0 &= y_{n-1} + \int_{t=tn-1}^{t_n} P_{k-1}(t)dt = \\ &= y_{n-1} + \left[hP_{k-1}(t_{n-1}) + \ldots + \tfrac{h^k}{k!} P_{k-1}^{(k-1)}(t_{n-1}) \right] \end{aligned}$$

Nordsieck - Darstellung: $h \to \vartheta \cdot h$, $0 < \vartheta < 1$

$$\begin{aligned} y^0(t_{n-1} + \vartheta h) &:= y_{n-1} + \vartheta a_{n-1,1} + \ldots + \vartheta^k a_{n-1,k} \\ a_{n-1,i} &:= \tfrac{h^i}{i!} P_{k-1}^{(i)}(t_{n-1}) \quad \text{“skalierte” Ableitung} \end{aligned} \tag{3.35}$$

Bemerkung:

(1) Auswertung mit Horner-Algorithmus

(2) Für Korrektor erhält man Polynom vom Grad $(k+1)$; man hat dann noch $a_{n-1,k+1}$.

Vorschlag GEAR:

Auch *Ordnungsänderung* in Nordsieck-Darstellung
aber: nach Änderung von k ist das Polynom nur noch *Schmiegpolynom* in $t = t_{n-1}$, es *interpoliert* nicht mehr über $\{t_{n-k}, \ldots, t_{n-k}\}$.

Für kombinierte Schrittweiten- und Ordnungssteuerung benutzt man die Beziehungen:

$$\begin{aligned} a_{n,k} &\doteq A_n h^k \\ \nabla a_{n,k} &= a_{n,k} - a_{n-1,k} \doteq B_n h^{k+1} \\ \nabla^2 a_{n,k} &\doteq C_n h^{k+2} \end{aligned} \tag{3.36}$$

Aus diesen Beziehungen verschafft man sich Schrittweitenschätzungen

$$\bar{h}^{(k)},\ \bar{h}^{(k+1)},\ \bar{h}^{(k+2)}$$

"alte" Ordnung : $k+1$
"neue" Ordnung : $q \in \{k, k+1, k+2\}$ "Ordnungsfenster"
$\bar{h}^{(q)} = \max\{\bar{h}^{(k)}, \bar{h}^{(k+1)}, \bar{h}^{(k+2)}\}$

Bemerkung: Aufwand gleich für $k, k+1, k+2$.

"Faustregeln": Schrittweitenänderung nur, falls
(a) keine Konvergenz
$\hookrightarrow$ Reduktion
oder (b) $(k+1)$ Schritte lang Ordnung oder
Schrittweite konstant

Begündung: Schrittweitenänderung erfordert hohen Aufwand.

Programme: GEAR-ABM (HINDMARSH, LLNL)[65]
DEABM (SHAMPINE, SANDIA) [101]
LSODE - nonstiff (HINDMARSH, LLNL)[64]

Adams-Verfahren über variablem Gitter:

- erfordert aufwendige Berechnung der Koeffizienten (trotz Tricks von KROGH $\longrightarrow$ Übung)
- deshalb Anlehnung an äquidistanten Fall
- Stabilität unübersichtlich

Programme: EPISODE (BYRNE/HINDMARSH)[17]
VOAS (SEDGWICK), unausgereift

4 Aufgaben

1.) Gegeben sei die *Thomas-Fermi*-Differentialgleichung:

$$y''(t) = \frac{y^{\frac{3}{2}}(t)}{t^{\frac{1}{2}}} \ , \ y(0) = y_0 \neq 0 \ , \ y'(0) = z_0 \tag{4.1}$$

Diese Form verletzt die Lipschitzbedingung und ist zudem nicht geeignet als Eingabe zur numerischen Integration.
Man transformiere (1) in ein System 1. Ordnung, welches der Lipschitzbedingung genügt.

Hinweis: Man verwende die Substitutionen:

$$s = t^{\frac{1}{2}} \ , \ y(t) = w(s) \ , \ u(s) = \frac{\dot{w}(s)}{s}$$

2.) Man zeige, daß das lineare Differentialgleichungssystem:

$$y'(x) = \frac{1}{x}(A_0 + A_1 x + A_2 x^2 + \cdots)\, y(x) \tag{4.2}$$

Lösungen der Form

$$y(x) = x^{\rho}(\nu_0 + \nu_1 x + \nu_2 x^2 + \cdots)$$

besitzt, wobei die A_i $n \times n$-Matrizen und die ν_i n-Vektoren sind.
Dazu bestimme man zuerst ρ und ν_0, dann rekursiv ν_1, $\nu_2, \cdots$.
Welche Bedingung müssen die Eigenwerte der Matrix A_0 erfüllen, damit die ν_i berechnet werden können ?
Welche Transformation $x = T(t)$ wandelt (2) in ein System ohne Singularität um ?

3.) Sei A eine konstante Matrix. Man berechne die Lösung der DG

$$y'(t) = Ay(t) + g(t) \ , \ y(0) = y_0$$

mit *Picard*-Iteration.

Hinweis: Man kann die Beziehung

$$\int_0^t \int_0^{s_1} f(s_1, s_2) g(s_2) ds_2 ds_1 = \int_0^t \left(\int_{s_2}^t f(s_1, s_2) ds_1 \right) g(s_2) ds_2$$

benutzen.

4.a.) Gegeben sei die Differentialgleichnung $y' = f(y)$, $y(0) = y_0$. Man zeige, daß für die Lösung der *adjungierten* Variationsgleichung

$$u'(t) = -f_y^T(y(t))\, u(t) \ , \ u(T) = z$$

gilt:

$$u(t) = W^T(T,t)z$$

b.) Für die parameterabhängige Differentialgleichung

$$y' = f(y,p) \ , \ y(0) = y_0 \ , \ y(t) \in R^n \ , \ p \in R^q \ ,$$

sei $y(t;p)$ Lösung. Die Abhängigkeit der Lösung von p wird durch die verallgemeinerte Variationsgleichung

$$\frac{dP(t)}{dt} = f_y(y(t;p),p)P(t) + f_p(y(t;p),p) \ , \ P(0) = 0$$

beschrieben. Man leite für $P(t) = \frac{\partial y(t)}{\partial p}$ die Beziehung

$$P(t) = \int_0^t W(t,s) f_p(y(s;p),p) ds$$

her.

5.) Gegeben sei die autonome Differentialgleichung:

$$y' = f(y) \ , \ y(0) = y_0 \ , \ f(y_0) \neq 0.$$

Man zeige: Die Existenz einer periodischen Lösung (mit zunächst unbekannter Periode T) ist äquivalent zu:
Die *Wronski*-Matrix $W(T,0)$ hat (mind.) einen Eigenwert 1.

6.a.) Man beweise den Satz: (Gröbner/Alekseev)
Sei y Lösung von $y' = f(y)$, $y(0) = y_0$, und z Lösung der gestörten DG

$$z' = f(z) + g(z) \ , \ z(0) = y_0.$$

Weiter sei die stetig partielle Differenzierbarkeit von f bzgl. y vorausgesetzt. Dann gilt:

$$z(t) = y(t) + \int_0^t \frac{\partial y}{\partial z}(t;s,z(s))\, g(z(s)) ds$$

Dabei sei $y(t;s,z(s))$ die Lösung von $y' = f(y)$ mit dem Anfangswert $y(s) = z(s)$.

b.) Man leite aus a.) die Beziehung (1.23) aus der Vorlesung für eine kleine Störung δf her.

7.) Man diskretisiere die DG $y' = \sqrt{y}$, $y(0) = 0$, mit dem expliziten Euler-Verfahren

$$y_{k+1} = y_k + hf(y_k) \ , \ k = 0, 1, 2, \ldots ,$$

und mit dem impliziten Euler-Verfahren

$$y_{k+1} = y_k + hf(y_{k+1}) \ , \ k = 0, 1, 2, \ldots ,$$

und führe in beiden Fällen für festes t den Übergang $h \to 0$ durch. Was ergibt sich im Vergleich zur theoretischen Lösung ?

8.) Gegeben sei das implizite Einschrittverfahren :

$$(*) \qquad y_{k+1} = y_k + h\Phi(y_k, y_{k+1})$$

Die Inkrementfunktion Φ genüge einer Lipschitzbedingung

$$\| \Phi(\xi, z) - \Phi(\xi, w) \| \leq L_1 \| z - w \| .$$

Man zeige : Die Bedingung $| h | < L_1^{-1}$ ist hinreichend für die eindeutige Lösung von $(*)$.

9.) Gegeben sei das Anfangswertproblem

$$y' = f(y) \ , \ y(0) = y_0, \ y(t) \in R.$$

Formale Integration und Anwendung der *Simpson*-Regel führt auf die Differenzengleichung

$$y_{n+1} = y_{n-1} + \frac{h}{3}(f(y_{n-1}) + 4f(y_n) + f(y_{n+1})).$$

a.) Man zeige : Für $f(y) = -y$ genügt y_n der 3-Term Rekursion:

$$(R) \qquad (1 + \frac{h}{3})\, y_{n+1} + \frac{4}{3} h\, y_n - (1 - \frac{h}{3})\, y_{n-1} = 0.$$

b.) Man zeige: Der Ansatz $y_n = A\, \lambda^n$, $A \in R$, $\lambda \in R$, in (R) führt auf die Lösungen

$$\begin{aligned} \lambda_1 &= 1 - h + \mathcal{O}(h^2) \\ \lambda_2 &= -(1 + \frac{h}{3}) + \mathcal{O}(h^2). \end{aligned}$$

Die allgemeine Lösung von (R) lautet :

$$y_n = A_1 \lambda_1^n + A_2 \lambda_2^n, \ A_1, A_2 \in R.$$

c.) Für kleine Schrittweiten h und $t = nh$ beweise man :

$$y_n = A_1 e^{-t}(1 + \mathcal{O}(h)) + (-1)^n A_2\, e^{t/3}(1 + \mathcal{O}(h)).$$

10.) Zur Integration eines Systems gewöhnlicher Differentialgleichungen zweiter Ordnung

$$y'' = f(y), \quad y(t_0) = y_0, \quad y'(t_0) = z_0,$$

sei das Diskretisierungsverfahren nach *Störmer* vorgegeben:

$$(S) \qquad \begin{aligned} y_1 &= y_0 + h[z_0 + \frac{h}{2} f_0], \\ y_{q+1} &= 2y_q - y_{q-1} + h^2 f_q, \ q = 1, \ldots, l-1, \\ z_l &= \frac{y_l - y_{l-1}}{h} + \frac{h}{2} f_l, \end{aligned}$$

mit $f_q := f(y_q)$.

a.) Man zeige:
Die Diskretisierung (S) besitzt eine äquivalente symmetrische Formulierung (S'):

$$\begin{aligned} \frac{z_{q+1} - z_q}{h} &= \frac{1}{2}(f_{q+1} + f_q) \\ \frac{y_{q+1} - y_q}{h} &= \frac{1}{2}[z_{q+1} + z_q - \frac{h}{2}(f_{q+1} - f_q)], \ q = 0, 1, \ldots, l-1. \end{aligned}$$

b.) Mit welchem Differentialgleichungssystem ist (S') konsistent ?

c.) Für $f \in C^{2N+2}$ beweise man für (S) die Existenz quadratischer asymptotischer Entwicklungen der Form

$$\begin{aligned} y_l - y(t_l) &= \sum_{k=1}^{N} e_k(t_l) h^{2k} + E_{N+1}(t_l; h) h^{2N+2}, \\ z_l - y'(t_l) &= \sum_{k=1}^{N} \bar{e}_k(t_l)\, h^{2k} + \bar{E}_{N+1}(t_l; h)\, h^{2N+2}, \end{aligned}$$

mit $E_{N+1}, \bar{E}_{N+1}$ gleichmäßig beschränkt in h, $h \in [0, H]$, $H > 0$.

11.) Gegeben sei ein Einschrittverfahren der Konsistenzordnung p mit der Entwicklung

$$y_N - y(t_N) = e_p(t)h^p + e_{p+1}(t)h^{p+1} \dots$$

Man überlege sich, welcher Aufwand an Funktionsaufrufen A_q in Abhängigkeit von der erreichten Ordnung q in einem Extrapolationsverfahren nötig ist. Zum Vergleich wähle man das explizite Euler-Verfahren, das klassische *Runge-Kutta*-Verfahren 4.Ordnung (RK4):

$$\begin{aligned} k_1 &= f(y_k)\,, \\ k_2 &= f(y_k + \frac{h}{2}k_1)\,, \\ k_3 &= f(y_k + \frac{h}{2}k_2)\,, \\ k_4 &= f(y_k + h\,k_3)\,, \\ y_{k+1} &= y_k + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)\,, \end{aligned}$$

und das verbesserte Euler-Verfahren ($p = 2$):

$$y_{k+1} = y_k + f(y_k + \frac{h}{2}f(y_k))$$

mit jeweils den Schrittweitenfolgen :

$$\begin{aligned} F_H &= \{1,2,3,4,\dots\}\,,\ n_i = i\,; \\ F_{2H} &= \{2,4,6,8,\dots\}\,,\ n_i = 2\,i\,. \end{aligned}$$

Welche Schlüsse lassen sich hinsichtlich der Wahl des Einschrittverfahrens und der Schrittweitenfolge für ein Extrapolationsverfahren ziehen?

12.) Zur Lösung von $y' = f(y)\,,\ y(0) = y_0\,,$ sei die explizite Mittelpunktsregel ohne Schlußschritt gegeben:

$$\begin{aligned} y_1 &= y_0 + h\,f(y_0)\,, \\ y_{k+1} &= y_{k-1} + 2h\,f(y_k)\,,\ k = 1,2,\dots \end{aligned}$$

a.) Man betrachte als Schlußschritt das implizite Euler-Verfahren:

$$y_l = y_{l-1} + h\,f(y_l)$$

Man zeige, daß damit das gesamte Verfahren *reversibel* ist, das heißt:
Die Transformation $(k\,,\,h) \rightarrow (l-k\,,\,-h)$ läßt das Verfahren unverändert.

b.) Anstelle des impliziten Euler-Schrittes betrachte man die Fixpunkt-Iteration:

$$\begin{aligned} y_l^0 &:= y_l \,, \\ y_l^{i+1} &:= y_{l-1} + h\, f(y_l^i) \,,\ i = 0, 1, \ldots \end{aligned}$$

Man zeige:

(i) y_l^1 hat eine h^2-Entwicklung. Dazu überlege man sich den Zusammenhang zwischen y_l^1 und dem Schlußschritt S_l von *Gragg* aus der Vorlesung.

(ii) y_l^2 und $y_l^* := y_{l-1} + h\, f(y_l^*)$ haben <u>keine</u> h^2-Entwicklung.

13.) Gegeben sei das Anfangswertproblem

$$\begin{aligned} y' = D(t)\, y + \varphi(t) \,,\ y(t),\, \varphi(t) &\in R^n \,,\ y(0) = y_0 \,, \\ D(t) &\in R^{n \times n}. \end{aligned}$$

Man beweise, daß der lokale Fehler $y(h) - y_1$ eines auf dieses System angewandten Runge-Kutta-Verfahrens genau dann von der Konsistenz-Ordnung p ist, wenn:

$$\begin{aligned} \sum_j b_j\, c_j^{q-1} &= \frac{1}{q} \,,\ q \le p \,, \\ \sum_{j,k} b_j\, c_j^{q-1}\, a_{jk}\, c_k^{r-1} &= \frac{1}{(q+r)r} \,,\ q + r \le p \,, \\ \sum_{j,k,l} b_j\, c_j^{q-1}\, a_{jk}\, c_k^{r-1}\, a_{kl}\, c_l^{s-1} &= \frac{1}{(q+r+s)(r+s)s} \,,\ q + r + s \le p \,, \\ &\ldots \ \textit{etc.} \end{aligned}$$

<u>Hinweis</u>: Man schreibe das System in autonomer Form und untersuche welche Bäume (und damit elementare Differentiale) bei der Herleitung der Bedingungsgleichungen wegfallen.

14.) a.) Zur Lösung von Systemen nichtautonomer Differentialgleichungen

$$y' = f(t, y) \,,\ y(t_0) = y_0 \,,\ y(t) \in R^n$$

soll ein Runge-Kutta-Verfahren RK4 (mit konstanter Schrittweite h) implementiert werden.
Das Verfahren lautet :

$$\begin{aligned} k_i &= f(t_0 + c_i\, h \,,\, y_0 + \sum_{j=1}^{i-1} a_{ij}\, k_j) \,,\ i = 1, \ldots, 4, \\ y_1 &= y_0 + \sum_{i=1}^{4} b_i\, k_i \end{aligned}$$

Dazu verwende man als Lösung der Bedingungsgleichungen den folgenden Koeffizientensatz (klassisches RK-Verfahren):

0	0			a_{ij}
$\frac{1}{2}$	$\frac{1}{2}$	0		
$\frac{1}{2}$	0	$\frac{1}{2}$	0	
1	0	0	1	0
c_i/b_i	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Mit diesen Werten schreibe man ein FORTRAN-Unterprogramm

RK4 (N, FCN, T, Y, TEND, M)

wobei:

N:	Anzahl der DG'en 1. Ordnung
T:	Anfangspunkt der Integration
Y:	Anfangswerte $y(t_0)$
TEND:	Endpunkt der Integration
M:	Anzahl der Integrationsschritte
FCN(T,Y,DY):	Unterprogramm, das die rechte Seite der DG (=: DY) an der Stelle T,Y(T) auswertet.

Mit diesem Programm integriere man die DG

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} y_2 y_3 \\ -y_1 y_3 \\ -k^2 y_1 y_2 \end{bmatrix}, \; y(t_0) = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \; k^2 = 0.51$$

(Eulersche Bewegungsgleichung für einen Körper)

von $t_0 = 0$ bis $t_{\text{END}} = 60$. Der absolute Fehler der numerischen Lösung an t_{END} sollte kleiner als $\epsilon = 10^{-6}$ sein.

b.) Man erweitere das Programm RK4 aus Aufgabe a.) durch Implementierung einer Schrittweitensteuerung und einer lokalen Fehlerkontrolle. Dabei benutze man zur Schrittweitenschätzung die Extrapolationstechnik.
Als Aufrufliste dieses Unterprogramms RKX4 verwende man die Standardliste:

RKX4 (N,FCN,T,Y,TEND,TOL,HMAX,H)

wobei (zusätzlich zu den Parametern in RK4) :

TOL : verlangte Genauigkeit der Lösung

HMAX : maximal zulässige Schrittweite

H : Anfangsschrittweite (:=10^{-3})

Mit diesem Verfahren integriere man wiederum die Eulersche Bewegungsgleichung mit der Genauigkeit $\epsilon = 10^{-6}$ und vergleiche den Aufwand (Auswertungen der rechten Seite der DG) mit dem Verfahren RK4.

15.) Ein Extrapolationsverfahren basiere auf der expliziten Mittelpunktsregel mit Schlußschritt und benutze die Schrittweitenfolge $F_{2H} = \{2, 4, \ldots\}$:

$$\begin{aligned} y_1 &= y_0 + \frac{h}{n_i} f(y_0) \,, \\ y_{n+1} &= y_{n-1} + \frac{2h}{n_i} f(y_n) \,, \; n = 1, \ldots, n_i, \\ y_l^i &= \frac{1}{4}(y_{l-1} + 2y_l + y_{l+1}) \,, \; l = n_i \,, \\ n_i &= 2i \,, \; i = 1, 2, \ldots \end{aligned}$$

Mit Hilfe des Aitken-Neville-Schemas (2.18.b) läßt sich aus $T_{11} := y_l^1$ und $T_{21} := y_l^2$ eine Näherung T_{22} für $y(h)$ der Ordnung $p = 4$ berechnen (Extrapolation).

a.) Man formuliere die Berechnung von T_{22} als ein äquivalentes 9-stufiges Runge-Kutta-Verfahren der Form (2.42) aus der Vorlesung und gebe die zugehörigen Koeffizienten a_{ij}, b_i, c_i an.

b.) Um ohne weitere Funktionsaufrufe eine Näherung von y an einer Stelle $t + \theta h \,, \theta \in [0,1]$ beliebig, angeben zu können, formuliert man ein Runge-Kutta-Verfahren zu $h^* = \theta h$ mit gleicher Stufenzahl $s^* = s = 9$ von der Ordnung $p = 3$ mit

$$a_{ij}^* = \frac{1}{\theta} a_{ij} \;\; (\Rightarrow \; c_i^* = \frac{1}{\theta} c_i).$$

(i) Man gebe die Bedingungsgleichungen für die Koeffizienten b_i^*, $i = 1, \ldots, s^*$, dieses Verfahrens an.

(ii) Für welche Wahl der b_i^* sind die Bedingungsgleichungen für alle $\theta \in [0,1]$ erfüllt ?
Hinweis: Man setze $b_2^* = \sigma$, $b_3^* = \zeta$, $b_8^* = \gamma - \zeta$, (σ, ζ, γ Parameter) und überlege sich, welche b_i^* zwingend verschwinden müssen.

16.) Gegeben sei ein $k-$Schritt-Verfahren der Form

(MSV) $\alpha_k y_{n+k} + \ldots + \alpha_0 y_n = h\,[\beta_k f_{n+k} + \ldots + \beta_0 f_n] \,, \; n = 0, 1, \ldots,$

mit $f_j = f(y_j)$.

Man zeige: Zur rekursiven Auswertung von (MSV) genügt die Speicherung von $s_n^0, \ldots, s_n^{k-1}$, wobei

$$\begin{aligned} s_n^k &:= s_{n-1}^{k-1} + \alpha_k y_{n+k} - h\beta_k f_{n+k} = 0 \\ s_n^{k-1} &:= s_{n-1}^{k-2} + \alpha_{k-1} y_{n+k} - h\beta_{k-1} f_{n+k} \\ &\vdots \\ s_n^1 &:= s_{n-1}^0 + \alpha_1 y_{n+k} - h\beta_1 f_{n+k} \\ s_n^0 &:= \alpha_0 y_{n+k} - h\beta_0 f_{n+k} \end{aligned}$$

(Darstellung von *Skeel*)

17.) Sei A eine reelle $n \times n$-Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$, und $\rho(A) := \max\limits_{1 \le i \le n} |\lambda_i|$ der Spektralradius von A.

Man zeige: Für alle $\epsilon > 0$ existiert eine Vektornorm $\| \cdot \|$, so daß für die zugehörige Matrixnorm gilt:

$$\rho(A) \le \| A \| \le \rho(A) + \epsilon,$$

wobei eine Matrixnorm definiert ist durch :

$$\| A \| = \max_{\|x\| \neq 0} \frac{\| Ax \|}{\| x \|}.$$

18.) Gegeben sei die lineare DG k-ter Ordnung mit konstanten, reellen Koeffizienten α_j, $j = 1, \ldots, k$:

$$L[y] := y^{(k)} + \alpha_1 y^{(k-1)} + \ldots + \alpha_k y = 0$$

Der Ansatz $y(t) = e^{\lambda t}$ führt auf das charakteristische Polynom

$$\rho(\lambda) := \lambda^k + \alpha_1 \lambda^{k-1} + \ldots + \alpha_k = \prod_{j=1}^{k} (\lambda - \lambda_j),$$

wobei λ_j, $j = 1, \ldots, k$, die Nullstellen von ρ seien.
Zur Lösung betrachte man das Mehrschrittverfahren

$$L_h y_q := y_{q+k} + a_1(h) y_{q+k-1} + \ldots + a_k(h) y_q = 0$$

auf einem äquidistanten Gitter $\{t_q\}$ mit Schrittweite h und Koeffizienten $a_j(h)$, $j = 1, \ldots, k$, die durch

$$\begin{aligned} \omega^k + a_1(h)\omega^{k-1} + \ldots + a_k(h) &:= \prod_{j=1}^{k} (\omega - \omega_j(h)), \\ \omega_j(h) &:= e^{\lambda_j h}, \end{aligned}$$

definiert seien.
N bezeichne den Nullraum eines Operators und N_h seine Einschränkung auf das Gitter mit Schrittweite h.

a.) Man zeige:

$$\begin{aligned} N_h(L) &= N_h(L_h)\,,\ \forall h \notin \hat{H}\,, \\ N_h(L) &\subset N_h(L_h)\,,\ \forall h \in \hat{H},\ h \neq 0\,, \end{aligned}$$

mit $\hat{H} := \{h \in R^n \mid (\lambda_j - \lambda_l)h = 2\pi i\, p;\ j,l = 1,\ldots,k;$
$p = 0, \pm 1, \pm 2, \ldots\}$.

b.) Man verifiziere die Aussage von a.) am Beispiel

$$L[y] = y'' + \mu^2 y = 0$$

19.) Gegeben sei das Adams-Bashforth-Verfahren über *äquidistantem* Gitter in der Form

$$(*) \qquad y_n = y_{n-1} + h \sum_{i=0}^{k-1} \gamma_i \nabla^i f_{n-1}\,,\ t_n = nh,$$

wobei

$$\begin{aligned} \nabla &:= I - E_{-h} \qquad \text{(Rückwärtsdiff.-Operator)} \\ \gamma_i &:= \int_0^1 \frac{s(s+1)\cdot\ldots\cdot(s+i-1)}{i!} ds\,,\ i > 0\,,\ \gamma_0 := 1, \\ s &:= \frac{t - t_{n-1}}{h} \end{aligned}$$

a.) Man leite die in der Vorlesung verwendete vereinfachte Darstellung (3.27) des Interpolationspolynoms P_{k-1} her:

$$P_{k-1} = f_{n-1} + \nabla^1 f_{n-1} \cdot \frac{t - t_{n-1}}{h} + \ldots + \nabla^{k-1} f_{n-1} \frac{(t - t_{n-1})\cdot\ldots\cdot(t - t_{n-k+1})}{(k-1)!\, h^{k-1}}$$

b.) Man berechne die Koeffizienten $\{\beta_{k,k-j}\}$ der Standardform

$$y_n = y_{n-1} + h \sum_{j=1}^{k} \beta_{k,k-j}\, f_{n-j}$$

aus den Koeffizienten $\{\gamma_i\}$ der Rückwärtsdifferenzen- Formulierung $(*)$.

c.) Mit Hilfe der Lagrange-Darstellung des Interpolationspolynoms P_{k-1} gebe man eine geschlossene Darstellung der Koeffizienten $\{\beta_{k,k-j}\}$ an.

20.) Man untersuche das Adams-Moulton-Verfahren über äquidistantem Gitter:

$$y_n = y_{n-1} + h \sum_{j=0}^{k} \beta^*_{k,k-j} f_{n-j} \, , t_n = nh.$$

Man zeige:

a.) Das AM-Verfahren besitzt die maximale Konsistenzordnung $p = k+1$, also

$$\epsilon_n = \epsilon_{n-1} + \gamma^*_{k+1} M^*_{k+1} h^{k+2} \, ,$$

wobei $\{\gamma^*_i\}$ die Koeffizienten des AM-Verfahrens bei formaler Darstellung über Rückwärtsdifferenzen zu f_n seien.

b.) Die Näherung

$$y^1_n := y_{n-1} + h\, \beta^*_{k,k}\, f(y^0_n) + h \sum_{j=1}^{k} \beta^*_{k,k-j}\, f(y_{n-j}) \, ,$$

y^0_n Prädiktor des AB-Verfahrens,

des PECE-Verfahrens ist bereits von der Konsistenzordnung $p = k+1$.

c.) Für die Konvergenz der Fixpunktiteration

$$y^{i+1}_n := y_{n-1} + h\, \beta^*_{k,k}\, f(y^i_n) + h \sum_{j=1}^{k} \beta^*_{k,k-j}\, f(y_{n-j}) \, , \; i = 0, 1, \ldots ,$$

ist hinreichend:

$$h\, \beta^*_{k,k}\, L < 1 \, , \; L \text{ Lipschitzkonstante von } f \, .$$

21.) Man konstruiere spezielle Graphen, für den Fall, daß f nur maximal quadratische Terme enthält. Tabelle für Anzahl an Bedingungsgleichungen in Abhängigkeit von p (Anwendung: chemische Kinetik).

B. Steife und differentiell – algebraische Anfangswertprobleme

1 Theoretische Grundlagen

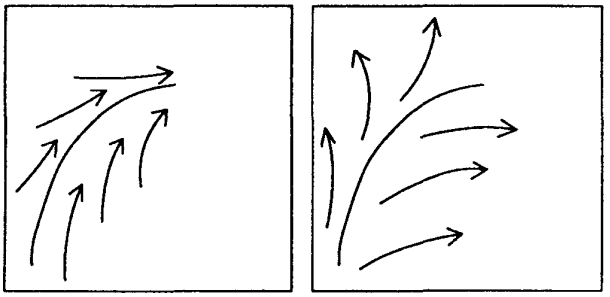
Für wichtige Klassen von Anfangswertproblemen ist die Charakterisierung durch die Lipschitzkonstante $L \equiv L_1$ theoretisch unbefriedigend. Dies zwingt zum Umdenken auch bei Diskretisierungsmethoden.

1.1 Asymptotische Stabilität von Differentialgleichungen

Beispiel:

$$\begin{aligned} &\text{a)} \quad y' = \lambda(y - g(t)) + g'(t), \lambda \in R^1 \\ &\qquad\quad y(0) := g(0) + \varepsilon_0 \\ &\text{b)} \quad \text{analytische Lösung:} \\ &\qquad\quad y(t) = g(t) + \varepsilon_0 \exp(\lambda t) \end{aligned} \tag{1.1}$$

Lösung *stabil/instabil* gegen Störung ε_0 - je nach Vorzeichen von λ.



$\lambda < 0$ stabil $\qquad$ $\lambda > 0$ instabil

Bild B.1

Übertragung auf autonome Systeme:

$$y' = f(y)$$

Sei y^* ein *kritischer Punkt* der Differentialgleichung:

$$f(y^*) = 0 \tag{1.2}$$

Taylorentwicklung der rechten Seite der Differentialgleichung um y^*

$$\underbrace{y'}_{(y-y^*)'} = \underbrace{f(y^*)}_{0} + f_y(y^*)(y-y^*) + \dots \tag{1.3}$$

Das qualitative Verhalten der Lösungen in der Umgebung eines kritischen Punktes wird also beschrieben durch die zugehörige Variationsgleichung (vgl. Kap. A. (1.16)). Dabei ist $f_y(y^*)$ eine *konstante* Matrix, da y^* Lösung von (1.2).

Satz 1.1 (LJAPUNOV 1892 [82])
Gegeben sei ein lineares homogenes DG-System

$$\varepsilon' = A\,\varepsilon\ ,\ \varepsilon(t_0) = \varepsilon_0 (\neq 0)$$

mit konstanter Matrix A. Für die Eigenwerte von A gelte:

$$\Re\{EW(A)\} < 0 \tag{1.4}$$

Dann existieren (beschränkte) Konstante $\alpha \geq 0$, $\mu \leq 0$ derart, daß

$$\| \varepsilon(t) \| \leq \alpha \exp(\mu t). \tag{1.5}$$

Ljapunov-Stabilität

Beweis: Orthogonaltransformation von A auf Normalform von Schur

$$U^H A\, U = S \Longleftrightarrow A = U\, S\, U^H\ ,U^H\, U = I = U\, U^H$$

$$S = \begin{bmatrix} \lambda_1 & * & \dots & * \\ & \lambda_2 & \ddots & \vdots \\ & & \ddots & * \\ & & & \lambda_n \end{bmatrix} \text{ komplex}$$

Transformation des DG-Systems: $\delta := U^H \varepsilon$

$$\varepsilon' = U\, S U^H\, \varepsilon \Longleftrightarrow \delta' = S\, \delta$$

$$\| \varepsilon \|_2 = \| \delta \|_2 \tag{*}$$

Explizit: $\delta^T := (\delta_1, \dots, \delta_n)$

$$\begin{bmatrix} \delta_1' \\ \vdots \\ \delta_n' \end{bmatrix} = \begin{bmatrix} \lambda_1 & \dots & * \\ & \ddots & \vdots \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix}$$

Rekursive Lösung möglich:

$$\delta_n = \delta_n(0)e^{\lambda_n t} = \delta_n(0)e^{\Re(\lambda_n)t}\, cos(Im(\lambda_n)t) + i \cdot \ldots$$

$$| \delta_n(t) | \leq | \delta_n(0) | \, e^{\Re(\lambda_n)t} \tag{1.6}$$

Eingesetzt in vorletzte Zeile:

$$\delta'_{n-1} = \lambda_{n-1}\delta_{n-1} + \ldots \text{ Inhomogenität} \Longrightarrow \delta_{n-1}(t)$$

Allgemeine Lösung liefert:

$$\delta_i(t) = \sum_{j=1}^{n} \underbrace{p_{ij}(t)}_{\text{Polynome in t}} \quad e^{\lambda_j t}$$

Abschätzung:

$$| \delta_i(t) | \leq \sum_{j=i}^{n} | p_{ij}(t) | \, e^{\Re(\lambda_j)t}$$

Sei definiert:

$$\begin{aligned} \nu &:= \max_j \Re(\lambda_j) \\ \nu &< 0 \text{ wegen (1.4).} \end{aligned} \tag{1.7}$$

Damit gilt:

$$| \delta_i(t) | \leq e^{\nu t} \underbrace{\sum_{j=i}^{n} | p_{ij}(t) |}_{\text{unbeschränkt}}$$

Abschätzung: $\exists \mu : \nu < \mu \leq 0$

$$\begin{aligned} \exists a_i > 0 : \quad | t^i e^{\nu t} | &\leq a_i \, e^{\mu t} \\ \Rightarrow \exists \alpha_i > 0 : \quad | \delta_i(t) | &\leq \alpha_i \, e^{\mu t} \\ \Rightarrow \exists \alpha > 0 : \quad \| \delta(t) \| &\leq \alpha \, e^{\mu t} \end{aligned}$$

Mit $(*)$ gilt dann (1.5). ∎

Speziell folgt aus (1.5) für $\mu < 0$:

$$\begin{gathered} \varepsilon(t) \longrightarrow 0 \quad \text{für } t \to \infty \\ \text{asymptotische Stabilität} \end{gathered} \tag{1.8}$$

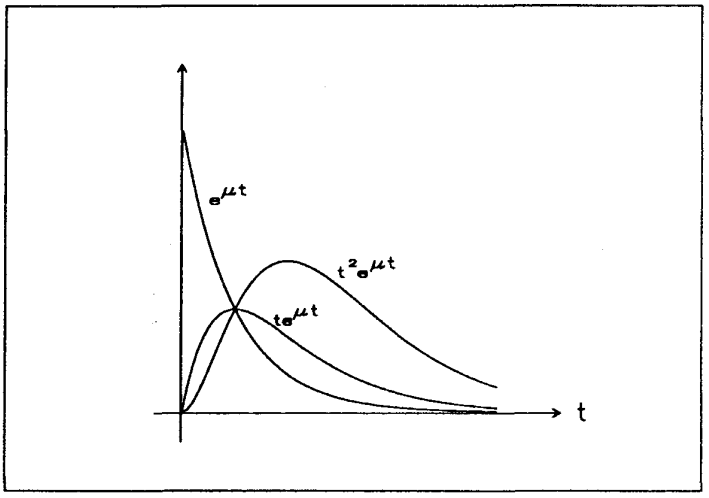


Bild B.2

Bemerkungen:

(1) Als sogenannte Ljapunov-Funktionen definiert man

$$\psi(t) := \| \varepsilon(t) \|^2,$$

wobei $\| \cdot \|$ eine geeignet gewählte Norm ist.

(2) In vielen Stabilitätsuntersuchungen werden skalare Differentialgleichungen zugrundegelegt - dann gilt (1.6).

(3) Falls $A \to A(t)$, so gilt Satz 1.1 nicht mehr!

Gegenbeispiel:

$$y' = A(t)\, y, \quad y(0) = \begin{pmatrix} 0 \\ -\varepsilon \end{pmatrix}$$

$$A(t) = \begin{bmatrix} -1 + \frac{3}{2}\cos^2 t & 1 - \frac{3}{2}\cos t \sin t \\ -1 - \frac{3}{2}\sin t \cos t & -1 + \frac{3}{2}\sin^2 t \end{bmatrix}$$

Für die Eigenwerte $\lambda_{1,2}$ zeigt man:

$$\Re(\lambda_{1,2}) = -\frac{1}{4} < 0 \text{ für alle } t$$

Analytische Lösung:

$$y(t) = \varepsilon \begin{pmatrix} -\cos t \\ \sin t \end{pmatrix} e^{+t/2} \quad \text{unbeschränkt}$$

Kontraktivität. Für *nichtlineare* Differentialgleichungen erweist sich die folgende Klassifikation als nützlich:
Für eine Differentialgleichung mit rechter Seite f existiere ein spezielles Skalarprodukt $< \cdot, \cdot >$, derart, daß gilt:

$$< f(u) - f(v), u - v > \leq \bar{\mu} < u - v, u - v > \tag{1.9}$$

$\bar{\mu}$ heißt auch *(globale) einseitige Lipschitzkonstante.*

Definition:

$$\bar{\mu} < 0 : \text{Differentialgleichung ist } \textit{global kontraktiv} \tag{1.10}$$

Satz 1.2 *Sei* $\| \cdot \|$ *durch* $< \cdot, \cdot >$ *induzierte Norm. Es gelte die Voraussetzung (1.9). Dann folgt:*

$$\| u(t) - v(t) \| \leq \| u(o) - v(o) \| \, e^{\bar{\mu} t}. \tag{1.11}$$

Beweis: Man definiert:

$$\begin{aligned} m(t) &= \| u(t) - v(t) \|^2 = < u(t) - v(t), u(t) - v(t) > \\ \Rightarrow m'(t) &= 2 < u'(t) - v'(t), u(t) - v(t) > = \\ &= 2 < f(u) - f(v), u - v > \\ (1.9) \rightarrow &\leq 2\bar{\mu} < u - v, u - v > = 2\bar{\mu}\, m(t). \end{aligned}$$

Differentialungleichung:

$$\Rightarrow m(t) \leq m(0) e^{2\bar{\mu} t} \Longrightarrow (1.11)$$

■

Offensichtlich ist Satz 1.2 eine Variante des Fundamentallemmas (Kapitel A.1.2, Satz 1.4), die jedoch zugleich die Stabilität widerspiegelt.

Beispiel: Streng parabolische Systeme: Raumdiskretisierung des streng elliptischen Teils führt auf kontraktive Systeme von gewöhnlichen Differentialgleichungen.

Linearer Spezialfall: $f(y) = Ay$

$$< x, Ax > \leq \mu < x, x > \equiv \mu \| x \|^2 \tag{1.12}$$

Sei $A := f_y(y_o)$ bei allgemein nichtlinearem Problem. Dann heißt $\mu(A)$ auch *(lokale) einseitige Lipschitzkonstante.*

Definition:

$$\begin{aligned} &\mu(A) < 0,\ A = f_y(y_o) \\ &\text{Differentialgleichung } \textit{lokal kontraktiv} \text{ in } U(y_o) \end{aligned} \tag{1.13}$$

1.2 Existenz- und Eindeutigkeitssätze

Numerische Methoden verwenden *lokale* (punktweise) Informationen von f. Stabilität erfordert zusätzliche Information $f_y(\cdot)$. Differentialgleichungen, zu deren Beschreibung $f_y(\cdot)$ wichtig ist, heißen "steife" Differentialgleichungen.

Idee: (DEUFLHARD 1987) [33]

(a) **Nichtsteife Integration:**
nur f-Information
↓
Picard-Iteration (Funktionenfolge $\{y^i\}$)
↓
theoretische Charakterisierung mit L_1
↓
Existenz + Eindeutigkeit (nichtsteife Differentialgleichung)
↓
Diskretisierungsmethoden, die nur f auswerten

(b) **Steife Integration:**
Information $(f, f_y(y_o))$
↓
Newton-Iteration (Funktionenfolge $\{y^i\}$)
↓
theoretische Charakterisierung über $\mu + \ldots$, ohne L_1
↓
Existenz + Eindeutigkeit (steife Differentialgleichung)
↓
Diskretisierungsmethoden, die $(f, f_y(y_o))$ auswerten

Newton-Iteration:

$$y^{i+1}(\tau) - A \int_{s=0}^{\tau} y^{i+1}(t)dt = y_o + \int_{s=0}^{\tau} [f(y^i(t)) - Ay^i(t)]dt \tag{1.14}$$

Satz 1.3 (DEUFLHARD 1987) *[33]*
Sei $f \in C^1(D), D \subseteq \mathbb{R}^n$. Für $A := f_y(y_o)$ gelte die einseitige Lipschitzbedingung (1.12). In der durch $< \cdot, \cdot >$ induzierten $\mathbb{R}^n$-Norm gelte ferner:

$$\| f(y_o) \| \leq L_o \tag{1.15.a}$$

$$\| f_y(u) - f_y(v) \| \leq L_2 \| u - v \| \quad \forall\, u, v \in D \tag{1.15.b}$$

Falls D hinreichend groß, so gilt Existenz und Eindeutigkeit der Lösung

$$y(t) \text{ für } t \in [o, \tau] \text{ mit}$$

$$\tau \text{ unbeschränkt, falls } \mu\bar{\tau} \leq -1 \tag{1.16.a}$$

$$\tau \leq \bar{\tau}\Psi(\mu\bar{\tau}), \text{ falls } \mu\bar{\tau} > -1 \tag{1.16.b}$$

wobei

$$\bar{\tau} := (2L_o L_2)^{1/2} \text{ und}$$

$$\Psi(s) := \begin{cases} ln(1+s)/s & s \neq 0, s > -1 \\ 1 & s = 0 \end{cases} \tag{1.16.c}$$

Beweis: (hier nicht ausgeführt). Affininvariante Form des Satzes von NEWTON-KANTOROVITCH (DEUFLHARD/HEINDL 1979 [39]). Angewendet auf Iteration (1.14).

Verbesserung vom Satz 1.3 ohne Rückgriff auf (1.14):

Satz 1.3' (W. WALTER 1987) (findet sich in [33]: Private Mitteilung.) *Voraussetzungen wie Satz 1.3. Dann existiert eine eindeutige Lösung in* $[o, \tau]$ *mit*

$$\tau \text{ unbeschränkt, falls } \mu\bar{\tau} \leq -1 \tag{1.17.a}$$

$$\tau < \bar{\tau}\bar{\Psi}(\mu\bar{\tau}), \text{ falls } \mu\bar{\tau} > -1 \tag{1.17.b}$$

wobei

$$\bar{\Psi}(s) := \begin{cases} \dfrac{1}{\sqrt{1-s^2}}(\pi - 2arctan\,(\dfrac{s}{\sqrt{1-s^2}}), & -1 < s < 1 \\ 2, & s = 1 \\ \dfrac{1}{\sqrt{s^2-1}}\, ln\left(\dfrac{s+\sqrt{s^2-1}}{s-\sqrt{s^2-1}}\right), & s > 1 \end{cases} \tag{1.17.c}$$

Beweis: Sei $L_0 > 0$ o.B.d.A. Sei definiert:

$$\rho^2 := \langle y(t) - y_o, y(t) - y_o \rangle \equiv \| y(t) - y_o \|^2 \tag{1.18}$$

Idee: Abschätzung des Intervalls, in dem ϱ beschränkt $\longrightarrow$ $y(t)$ beschränkt, sowie $y(t)$ eindeutig, da L_2 beschränkt.

Differentiation:

$$\varrho\varrho' = \frac{1}{2}(\varrho^2)' = \langle y(t) - y_o, \underbrace{f(y(t))}_{y'(t)} \rangle \tag{1.18'}$$

Charakterisierung von f:

$$f(y) = f(y_o) + A(y - y_o) + \int_{s=0}^{1} [f_y(y_o + s(y - y_o)\ ,\ -A](y - y_o)ds \tag{1.19}$$

In (1.18'):

$$\begin{aligned}
\varrho\varrho' &= \langle y(t) - y_o, f(y_o)\rangle + \langle y(t) - y_o, A(y(t) - y_o)\rangle + \\
&\quad + \langle y(t) - y_o, \int [\cdot](y(t) - y_o)ds\rangle \\
&\leq \mu L_o + \mu\varrho^2 + \varrho^2 \int_{s=0}^{1} \parallel f_y(y_o + s(y - y_o)) - A \parallel ds \\
&\leq L_o \cdot \varrho + \mu\varrho^2 + \frac{L_2}{2}\varrho^3
\end{aligned}$$

Sei $\varrho > 0$ für $t > 0$ ($\varrho(0) = 0$); abdividiert.

$$\varrho' \leq L_o + \mu\varrho + \frac{L_2}{2}\varrho^2 \tag{1.20}$$

Differentialungleichung. Zugehörige majorisierende Differentialgleichung:

$$\sigma' = L_o + \mu\sigma + \frac{L_2}{2}\sigma^2 =: g(\sigma), \sigma(0) = 0. \tag{1.21}$$

Für σ monoton wachsend, also $g(\sigma) \geq 0$, gilt:

$$\varrho(t) \leq \sigma(t)$$

Umformung von g:

$$g(\sigma) = L_o(1 - (\mu\bar{\tau})^2 + (\mu\bar{\tau} + L_2\bar{\tau}\sigma)^2)$$

Wurzeln von g:

$$\sigma_{1,2} = \frac{1}{L_2\bar{\tau}}\left[-\mu\bar{\tau} \pm \sqrt{(\mu\bar{\tau})^2 - 1}\right]$$

Es gilt: $g(0) = L_0 > 0$. Zu prüfen ist also, ob *positives* $\sigma_{1,2}$ auftreten kann.

Fälle:

(I) $\mid \mu\bar{\tau} \mid \geq 1 : \sigma_{1,2}$ beide reell

$$\mu < 0 : \sigma_{1,2} = \frac{1}{L_2\bar{\tau}} \left[\mid \mu\bar{\tau} \mid \pm \underbrace{\sqrt{\mid \mu\bar{\tau} \mid^2 - 1}}_{<|\mu\bar{\tau}|} \right]$$

Damit gilt für $t \in [0, \infty]$:

$$\varrho(t) \leq \sigma_2, \text{ falls } \mu\bar{\tau} \leq -1 \qquad (*)$$
$$\hookrightarrow (1.17.a)$$

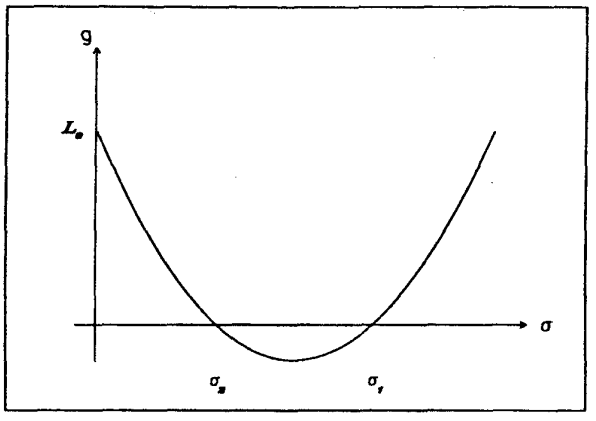


Bild B.3

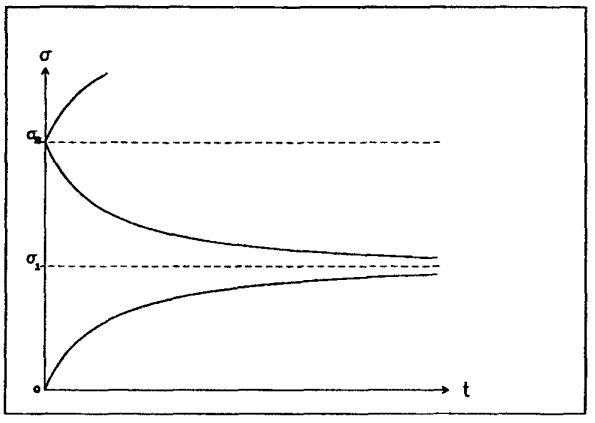


Bild B.4

(II) $\mid \mu\bar{\tau} \mid \geq 1 : \ \mu > 0 : \ \sigma_{1,2} = \frac{1}{L_2\bar{\tau}}[- \mid \mu\bar{\tau} \mid \pm\sqrt{(\mu\bar{\tau})^2 - 1}] < 0$

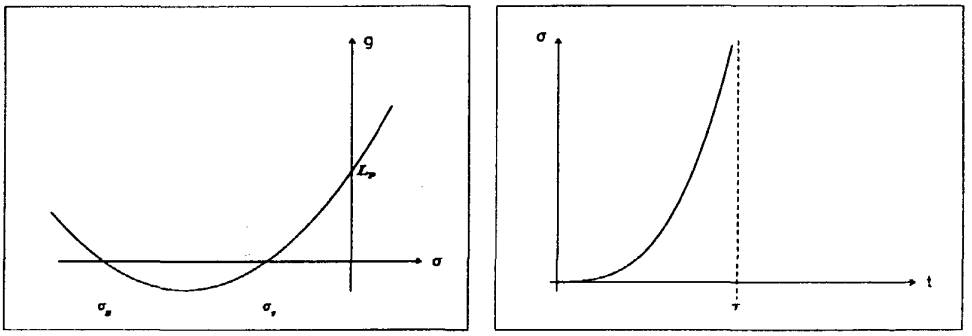


Bild B.5

(III) $-1 < \mu\bar{\tau} < 1 : \sigma_{1,2}$ konj. komplex
$\hookrightarrow \sigma > 0$ eventuell unbeschränkt.
Es bleibt zu prüfen, in welchem Intervall $[0, \tau]$ σ beschränkt ist.

Separation der Variablen in (1.21) liefert:

$$\int_{t=0}^{\tau} dt = \tau = \int_{\sigma=0}^{\infty} \frac{d\sigma}{g(\sigma)} \tag{1.22}$$

Analytische Quadratur (Formelsammlung) für (II), (III) liefert (1.17.b,c).

■

Der Beweis zu Satz 1.3 ist konstruktiv während der Beweis zu Satz 1.3′ nicht konstruktiv ist.

Frage: Eignet sich Newton-Iteration (1.14) zur Konstruktion von Verfahren?

Linearer Spezialfall: $f = Ay$

$$y^o(\tau) \equiv y_o$$

$$y^1(\tau) - A\int_0^{\tau} y^1(t)dt =$$

$$= y_0 + \int_0^{\tau} \underbrace{[f(y^o(t)) - Ay^o(t)]}_{0} dt = y_o$$

Zugehöriges Anfangswertproblem:

$$(y^1)' = Ay^1 \ , \ y^1(0) = y_0$$

$$y^1(\tau) = \exp(A\tau)y_o \tag{1.23}$$

liefert bereits in diesem Fall nur *formale* Lösung, keine brauchbare Lösungsmethode.

Bemerkung:

1. Analoge Theorie möglich, falls $A - f_y(y_o) \neq 0$.
2. Newton-Iteration für Diskretisierungsverfahren geeignet (vgl. Kap. 2 und 3).

2 Einschrittverfahren

Zu untersuchen ist das *Stabilitätsverhalten* der D_hG für $h \neq 0$ im Vergleich zum Stabilitätsverhalten der zugrundeliegenden Differentialgleichung.
Sei B nichtsinguläre (n,n)-Matrix, $\bar{y} := By$. Mit Affin-Kovarianz (Kap. A.1.3) folgt sofort:

$$\bar{f}_{\bar{y}} = B f_y B^{-1} \tag{2.1}$$

Beliebige Ähnlichkeitstransformation läßt Eigenwerte von $f_y(y)$ invariant. Diese Eigenwerte charakterisieren also die Stabilität auch bei linearen Diskretisierungsmethoden (vergleiche Kapitel A.1.3, (1.25)).

Vorgehen. Zunächst Untersuchung des linearen Spezialfalls, anschließend Erweiterung analog (Kap. B.1.2).

2.1 Lineare Stabilitätstheorie

Sei $f = Ay$ gewählt, $A_t \equiv 0$. Falls A diagonalisierbar, so existiert eine nichtsinguläre Matrix B derart, daß

$$B\,A\,B^{-1} = \Lambda = \mathrm{diag}(\lambda_1 \ldots, \lambda_n) \tag{2.2}$$

Wegen Kovarianzeigenschaft genügt also die Betrachtung der linearen Diskretisierungen für die *skalare Testgleichung* (DAHLQUIST[20] 1963) :

$$y' = \lambda y,\ y(0) = y_0 := 1,\ \lambda \in \mathbb{C} \tag{2.3}$$

Analytische Lösung:

$$y(t) = e^{\lambda t} \tag{2.4.a}$$

Stabilität:

$$\Re(\lambda) < 0 : y(t) \to 0 \quad \text{für} \quad t \to \infty \tag{2.4.b}$$

Reduktion auf Gitter mit Schrittweite $h > 0$; sei $z := \lambda h \in \mathbb{C}$:

$$y(h) = e^z y_o,\ z \in \mathbb{C} \tag{2.5}$$

Komplexe Charakterisierung des Stabilitätsverhaltens der Differentialgleichungslösung:

$$\begin{aligned} &\text{a)} \quad \Re(z) < 0 : \ |\, y(h) \,| = |\, e^z \,| < 1 \\ &\text{b)} \quad \Re(z) = 0 : \ |\, y(h) \,| = |\, e^z \,| = 1 \\ &\text{c)} \quad \Re(z) > 0 : \ |\, y(h) \,| = |\, e^z \,| > 1 \end{aligned} \tag{2.6}$$

Wesentliche Singularität in $z = \infty$:

$$\begin{aligned} &\text{a)} \quad \Re(z) < 0, \;\; z \to \infty : \;\; y(h) \to 0 \\ &\text{b)} \quad \Re(z) = 0, \;\; z \to \infty : \;\; \mid y(h) \mid = 1 \\ &\text{c)} \quad \Re(z) < 0, \;\; z \to \infty : \;\; y(h) \to \infty \end{aligned} \tag{2.7}$$

Untersuchung von Einschrittverfahren für skalare Testgleichung:
Beispiele:

(I) expliziter Euler: $y_1 = y_0 + h \cdot \lambda y_0 = 1 + z$

$z \to \infty : \; y_1 \to \infty$ unabhängig von $\Re(z)$ - vgl. (2.7.c)

(II) impliziter Euler : $y_1 = y_0 + h \cdot \lambda y_1 = \dfrac{1}{1 - z}$

$z \to \infty : \; y_1 \to 0$ unabhängig von $\Re(z)$ - vgl. (2.7.a)

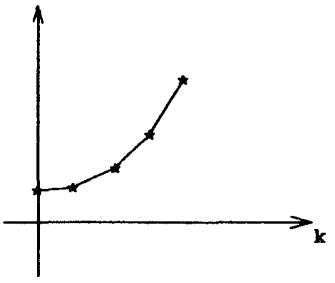
(III) implizite Trapezregel: $y_1 = y_0 + \dfrac{h}{2} \cdot \lambda(y_1 + y_0) = \left(1 + \dfrac{z}{2}\right) / \left(1 - \dfrac{z}{2}\right)$

$z \to \infty \; : \mid y_1 \mid = 1$ vgl. (2.7.b)

Qualitatives Verhalten von y_k für reelles λ:

$$\boxed{z = \lambda h > 0 :}$$

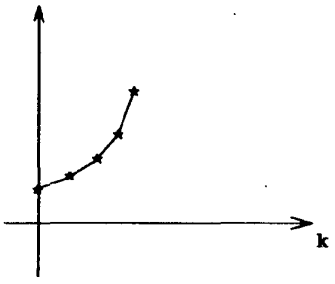
(I) $y_k = (1+z)^k$ monoton wachsend mit k:



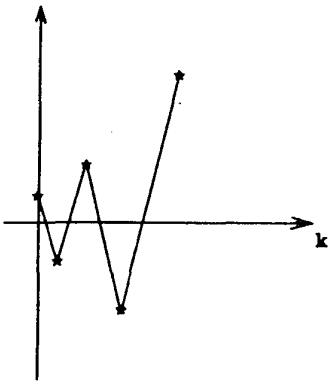
(II) $y_k = (1-z)^{-k}$

$0 < z < 1$: monoton wachsend

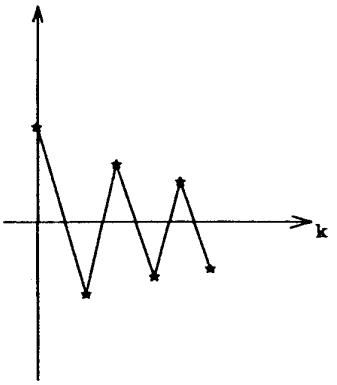
$z = 1$: Singularität !



$1 < z < 2$: anwachsend oszillatorisch



$z > 2$: beschränkt oszillatorisch



(III) $y_k = (\frac{2+z}{2-z})^k$

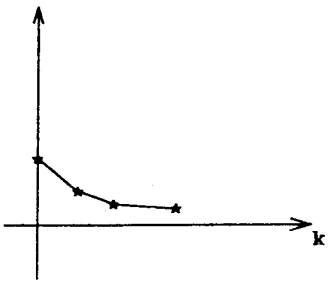
$z < 2$: monoton wachsend

$z = 2$: Singularität

$z > 2$: anwachsend oszillatorisch

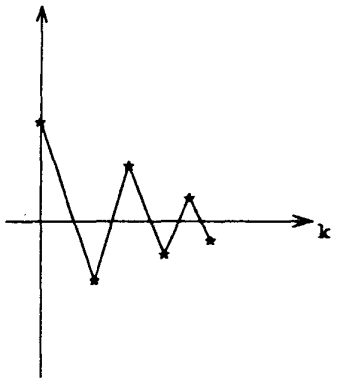
$\boxed{z = \lambda h < 0 :}$

(I) $-1 < z < 0$: monoton fallend

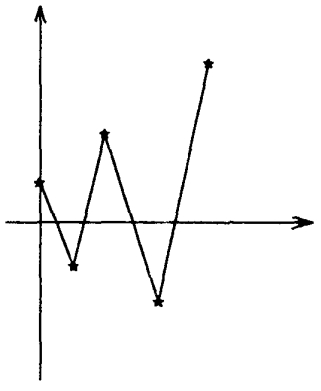


$z = -1:\quad y_k = 0 \quad k > 0$

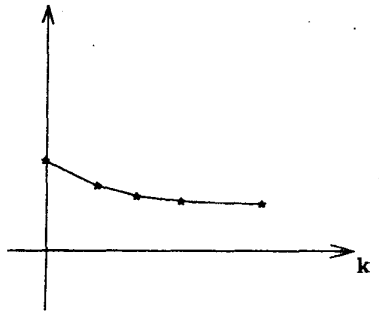
$-2 \leq z < -1:$ beschränkt oszillatorisch



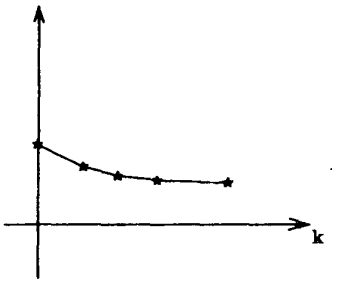
$z < -2:$ anwachsend oszillatorisch



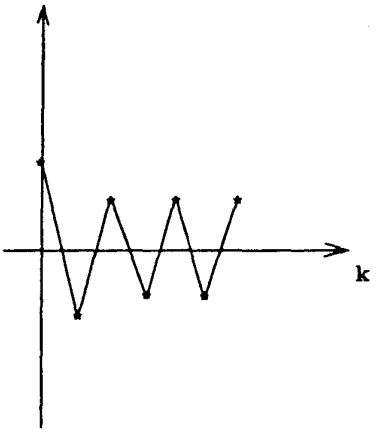
(II) $z < 0:$ monoton fallend



(III) $-2 < z < 0$: monoton fallend



$z < -2$: beschränkt oszillatorisch



$\boxed{z = \infty:}$

(I) $y_k = \infty$

(II) $y_k = 0$

(III) $y_k = (-1)^k$

Übertragung auf allgemeines Einschrittverfahren: *Explizite* Einschrittverfahren der Stufe s führen auf Polynome der Ordnung s:

$$y_1 = P_s(z)\, y_o \tag{2.8}$$

Implizite Einschrittverfahren führen auf rationale Funktionen:

$$y_1 = R(z) y_o \quad R(z) = \frac{P_l(z)}{Q_m(z)} \tag{2.9}$$

Vergleich mit (2.5) zeigt: $R(z)$ ist (komplexe) Approximation für $\exp(z)$. Konsistenzordnung p ($z \to 0$ entspricht $h \to 0$):

$$R(z) - e^z = \mathcal{O}(z^{p+1}) \Longrightarrow R(0) = 1 \tag{2.10}$$

$p = l + m$: (l, m) - Padé - Approximationen ($\hookrightarrow$ Padé - Tafel)

Zur Modellierung des Stabilitätsverhaltens diskreter Lösungen definiert man mit Blick auf (2.6):

Definition: Stabilitätsgebiet

$$G := \{z \in \mathbb{C} \mid \mid R(z) \mid \leq 1\} \tag{2.11}$$

Für die analytische Lösung gilt nach (2.6):

$$G_{anal} = \mathbb{C}^- := \{z \in \mathbb{C} \mid \Re(z) \leq 0\} \tag{2.6'}$$

Wünschenswert wäre $G = \mathbb{C}^-$ für Diskretisierungsmethoden. Das würde bedeuten:

$$\mid R(z) \mid = 1 \text{ für } \Re(z) = 0 \tag{$*$}$$

Wegen $(*)$ für $z = \infty$ muß dann gelten:

$$\begin{aligned} R(z) &= \frac{P_r(z)}{Q_r(z)} \\ P_r, Q_r &: \text{Polynome vom Grad } r \end{aligned}$$

Aus $(*)$ folgt dann:

$$1 = |R(z)|^2 = R(z)\, \bar{R}(z) = R(z)\, R(\bar{z})$$
$$\Re(z) = 0 : \bar{z} = -z$$

Das liefert:

$$P_r(z)\, P_r(-z) = Q_r(z)\, Q_r(-z) \tag{$**$}$$

Wegen $R(0) = 1$ gilt: $P_r(0) = Q_r(0)$. Bei Einschränkungen auf Padé-Approximationen (also für $p = 2r$) führt (**) zu:

$$P_r(z) = Q_r(z) \Rightarrow R(z) \equiv 1 \text{ trivial}$$

oder zu *diagonalen* Padé-Approximationen:

$$P_r(-z) = Q_r(z) \tag{2.12}$$

Beispiel:

$$P_1(z) = 1 + \frac{z}{2},\ Q_1(z) = 1 - \frac{z}{2} = P_1(-z)$$

implizite Trapezregel

$\hookrightarrow$ oszillatorisches Verhalten und $R(\infty) = (-1)^r, r = 1$ hier.

Um zumindest fallende analytische Lösungen durch betragsmäßig fallende diskrete Lösungen darzustellen, könnte man für Einschrittverfahren fordern:

$$G \supseteq \mathbb{C}^- \tag{2.13}$$

A-Stabilität (DAHLQUIST 1963)[20]

Da $R(z)$ meromorphe Funktion ist, läßt sich die wesentliche Singularität von $\exp(z)$ in $z = \infty$ *nicht* modellieren. Man kann sich bestenfalls für einen der drei Fälle von (2.7) entscheiden. (2.13) impliziert

$$|R(\infty)| \leq 1.$$

Das gestattet die Modellierung von (2.7.a) durch die verschärfte Forderung:

$$R(\infty) = 0 \tag{2.14}$$

Ein A-stabiles ESV, für das zusätzlich (2.14) gilt, heißt:

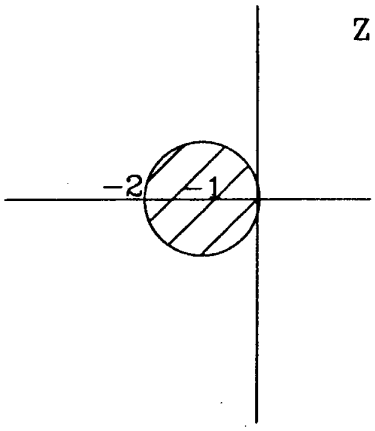
L-stabil

Bemerkung: *A-Stabilität* und *L-Stabilität* führen dazu, daß wachsende analytische Lösungen durch betragsmäßig fallende diskrete Lösungen dargestellt werden, falls $z \in G \backslash \mathbb{C}^- \neq 0$. Falls G "zu groß" spricht man von *Superstabilität.*

Ziel: möglichst gute Modellierung der imaginären Achse.

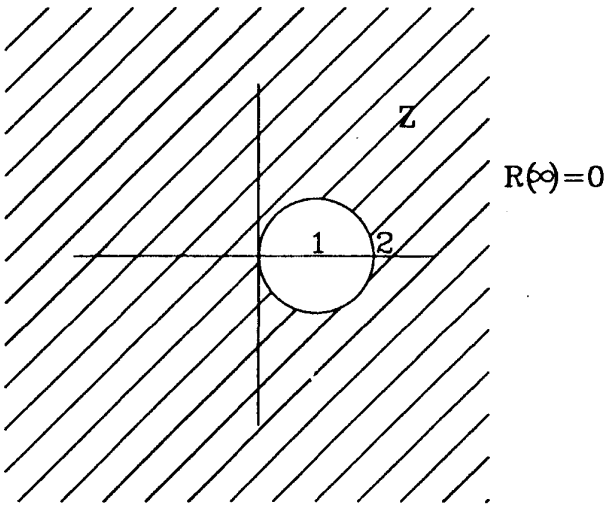
Beispiele für Stabilitätsgebiete:

(I) expliziter Euler: $R(z) = 1 + z$



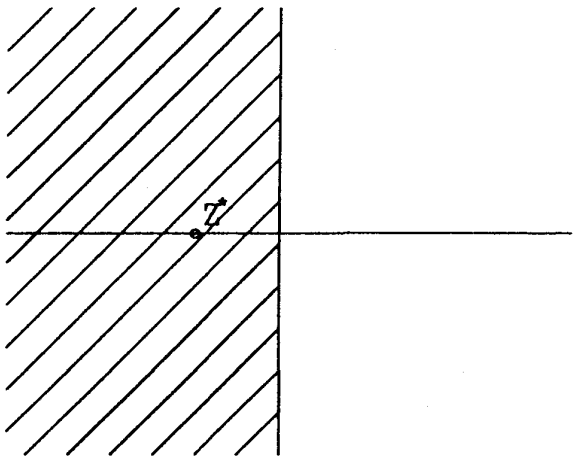
$$R(\infty) = 0$$

(II) impliziter Euler: $R(z) = (1 - z)^{-1}$



$R(\infty) = 0$ L-stabil, superstabil

(III) implizite Trapezregel: $R(z) = \dfrac{1 + z/2}{1 - z/2}$



$$R(\infty) = -1$$

A-stabil, aber nicht L-stabil

(Maximumprinzip: $|R(z)| = 1$ für $\Re(z) = 0$

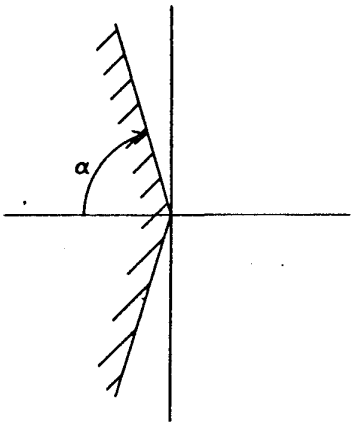
$|R(z)| < 1 \Rightarrow G = \mathbb{C}^-$)

Eine Abschwächung der *A-Stabilität* ist die sogennante:

$A(\alpha)$-Stabilitat (WIDLUND[113] 1967)

Dazu definiert man:

$$G(\alpha) := \{z \in \mathbb{C} |\, |\arg(z) - \pi| \leq \alpha\}$$



α maximal derart, daß gilt:

$$G(\alpha) \subseteq G \tag{2.15}$$

Offensichtlich ist $\alpha = \frac{\pi}{2}$ äquivalent zu *A-Stabilität*

Bemerkung: Falls $\lambda h \in G(\alpha)$ für *ein* $h > 0 \Rightarrow \lambda h \in G(\alpha) \ \forall h > 0$

Die Übertragung von skalarer Testgleichung auf lineare autonome Systeme leistet der folgende Satz.

Satz 1 (HAIRER, BADER, LUBICH 1982) [53] → J. v. NEUMANN (1951) [110]
Sei $R(z)$ rationale Funktion und analytisch in

$$G(\mu) := \{z \in \mathbb{C} \ | \Re(z) \leq \mu\}, \ \mu \in \mathbb{R}.$$

Sei A (n,n)-Matrix und sei für ein inneres Produkt $\langle \cdot, \cdot \rangle$ vorausgesetzt:

$$\Re \langle x, Ax \rangle \ \leq \ \mu \langle x, x \rangle \ \equiv \ \mu \parallel x \parallel^2 \tag{1.12'}$$

dann gilt:

$$\varphi_R(\mu) := \max_{z \in G(\mu)} |R(z)| = \sup_{t \in \mathbb{R}} |R(\mu + i\,t)| \, . \tag{2.16.a}$$

Sei $\| \cdot \|$ zugehörige Matrixnorm, so gilt:

$$\| R(A) \| \leq \varphi_R(\mu) \, . \tag{2.16.b}$$

Beweis: (LUBICH)

a) $R(z)$ analytisch in G: Maximumprinzip liefert:

$$\varphi_R(\mu) = \max_{z \in \partial G(\mu)} |R(z)|$$

b) o.B.d.A. $\langle \cdot, \cdot \rangle$ euklidisches Produkt. A habe komplexe Eigenwerte $\lambda_1, \ldots, \lambda_n$. Zerlegung von A:

$$A =: A_1 + iA_2$$
$$A_1 = (A + A^*)/2, \ \ A_2 = (A - A^*)/(2i)$$

A_1 und A_2 sind hermitesch, also diagonalisierbar durch unitäre Transformation. Diagonalisierung von A_1:

$$U A_1 U^* =: D_1 = \mathrm{diag}(z_1, \ldots, z_n)$$

Dann liefert (1.12'): $\Re\, z_i \leq \mu \quad i = 1, \ldots, n$
Außerdem:

$$\| R(U\,A\,U^*) \|_2 = \| U\,R(A)U^* \|_2 = \| R(A) \|_2$$
$$\Rightarrow A_1 = D_1 \text{ o.B.d.A.}$$

Sei $H_2 := UA_2U^*$, ebenfalls hermitesch und $\| R(D_1 + i\,H_2) \|_2 =: g(z_1, \ldots, z_n)$. Die so definierte Funktion g ist harmonisch bezüglich jedes komplexen Arguments $z_1, \ldots, z_n$ und analytisch in $(G(\mu))^n$. Maximumprinzip für jedes z_1 liefert:

$$\max_{z_i \in G(\mu)} g(z_1, \ldots, z_n) = \max_{z_i \in \partial G(\mu)} g(z_1, \ldots, z_n). \qquad (*)$$

Das heißt:

$$\| R(D_1 + iH_2) \| \leq \| R(\mu I + iH_2) \| \qquad (*')$$

Diagonalisierung von H_2:

$$VH_2V^* =: D_2 = \operatorname{diag}(w_1, \ldots, w_n)$$
$$\| R(\mu I + iH_2) \| = \| VR(\mu I + iH_2)V^* \| = \| R(V(\mu I + iH_2)\, V^*) \|_2 =$$
$$\| R(\mu I + iD_2) \|_2$$

$\mu I + iD_2$ ist Diagonalmatrix. Für Diagonalmatrizen gilt (2.16.b) trivial. ∎

In der Bezeichnung von Satz 2.1 schreibt sich *A-Stabilität* also:

$$\begin{aligned} &\text{a)} \quad \varphi_R(0) = 1 \quad (R(0) = 1) \\ &\text{b)} \quad R(z) \quad \text{analytisch in} \, \mathbb{C}^- \end{aligned} \qquad (2.17)$$

Für (l, m)-Padé-Approximationen schränkt (2.17) ein auf:

$$l \leq m \leq l + 2 \qquad (2.18)$$

("Ehle'sche Vermutung", bewiesen von WANNER, HAIRER, NØRSETT 1978)[111]

Bemerkung: $m < l + 2$: Pole in $\mathbb{C}^-$

Folgerung: Einschrittverfahren, die das Stabilitätsverhalten von Differentialgleichungen möglichst gut widerspiegeln, müssen notwendigerweise *implizit* sein. Damit ist für nichtlineares f die Lösung eines nichtlinearen Gleichungssystems verbunden. Für lineares $f = Ay$ entsteht ein lineares Gleichungssystem:

$$y_1 = R(hA)\, y_0 \qquad (2.19)$$

Es wird i.a. direkt gelöst unter Ausnutzung der Struktur von R.

Beispiele:

(I) impliziter Euler: $R(z) = (1-z)^{-1}$

$$(I - hA)\ y_1 = y_0\ ,\ \mu < 1$$

(II) implizite Trapezregel: $R(z) = \dfrac{1+\frac{z}{2}}{1-\frac{z}{2}}$

$$\left(I - \frac{h}{2}A\right) w_1 = y_0$$

$$y_1 = 2w_1 - y_0$$

$$\text{N.R.:}\ \left(1-\frac{z}{2}\right)(y_1+y_0) = \left(1+\frac{z}{2}\right) y_0 + \left(1-\frac{z}{2}\right) y_0 = 2y_0.$$

2.2 Existenz- und Eindeutigkeitssätze

Für nichtlineares f ist nichtlineares Gleichungssystem zu lösen. Lösung mit Fixpunkt-Iteration, also ohne f_y-Information, würde einschränken auf

$$hL_1 \leq C, C = \mathcal{O}(1),$$

also auf Intervall für nichtsteife Integration.

Folgerung (LINIGER, WILLOUGHBY 1970) [84]:
Steife Integration verlangt Lösung der auftretenden Gleichungssysteme mit Iterationsverfahren vom Newton-Typ, also unter Verwendung von f_y-Information.

Frage: Für welche $h > 0$ existiert eine eindeutige diskrete Lösung des gegebenen Einschrittverfahrens?

Zunächst Implizite Euler-Diskretisierung behandelt:

$$y_1 = y_0 + hf(y_1) \tag{2.20}$$

Vereinfachtes Newton-Verfahren ($i = 0, 1, \ldots$):

$$\begin{aligned} &(I - hA)\ \Delta y_1^i = -(y_1^i - y_0 - hf(y_1^i)) \\ &y_1^0 := y_0,\quad A := f_y(y_0),\quad y_1^{i+1} := y_1^i + \Delta y_1^i \end{aligned} \tag{2.21}$$

Satz 2 (DEUFLHARD 1987 [33])
Seien die Größen $L_o, \mu, L_2, \bar{\tau}$ definiert wie in Satz 1.3 (Kapitel B.1.2), dem

Existenz- und Eindeutigkeitssatz für die Lösung der Differentialgleichung. Dann existiert eine eindeutige Lösung $y_1(h)$ von (2.20) für $h\epsilon[o,\tau]$ mit

$$\begin{aligned} &a) \quad \tau \text{ unbeschränkt, falls } \mu\bar{\tau} \leq -1 \\ &b) \quad \tau \leq \bar{\tau}\Psi_{IE}(\mu\bar{\tau}), \text{ falls } \mu\bar{\tau} > -1 \\ &\text{wobei } \Psi_{IE} \text{ definiert durch :} \\ &c) \quad \Psi_{IE}(s) := 1/(1+s) \end{aligned} \tag{2.22}$$

Ferner konvergiert die vereinfachte NEWTON*-Iteration (2.21) unter der Bedingung (2.22).*

Beweis: Satz von NEWTON-KANTOROVITCH in affininvarianter Form (DEUFLHARD/HEINDL 1979)[39] angewendet auf Iteration (2.21): man benötigt Größen $\alpha(h), \omega(h)$ gemäß:

$$\begin{aligned} &\text{a)} \quad \| \Delta y_1^0 \| \leq \alpha(h) \\ &\text{b)} \quad \| (I - hA)^{-1}(F_y(u,h) - F_y(v,h) \| \leq \omega(h) \| u - v \| \end{aligned} \tag{2.23}$$

wobei

$$F(y) := y - y_0 - h\, f(y)$$

Definition von μ gemäß (1.12) und Spezifikation der Norm:

$$\| \cdot \|^2 = < \cdot, \cdot > .$$

Dann gilt:

$$\| \Delta y_1^0 \| = \| (I - hA)^{-1} h f(y_0) \| \leq \frac{hL_0}{1 - \mu h} =: \alpha(h) \tag{2.24.a}$$

unter der Voraussetzung, daß $\mu h < 1$. Sei $\| \cdot \|$ zugehörige Matrixnorm:

$$\| B \| := \max_{x \neq 0} \frac{\| Bx \|}{\| x \|}$$

Sei z Lösung von:

$$(I - hA)z = (F_y(u,h) - F_y(v,h))x$$

Dann gilt:

$$\omega(h) = \max_{x \neq 0} \frac{\| z(h) \|}{\| x \|} \Big/ \| u - v \| \tag{*}$$

Rechte Seite:

$$(F_y(u,h) - F_y(v,h))x = ((-hf_y(u)) - (-hf_y(v)))x = h(f_y(v) - f_y(u))x$$

Definition von L_2 in spezieller Norm eingeführt:

$$\| z(h) \| \leq \frac{hL_2}{1-\mu h} \| x \| \| u - v \|, \quad \text{für } \mu h < 1$$

Beziehung $(*)$ zeigt:

$$\omega(h) := \frac{hL_2}{1-\mu h} \tag{2.24.b}$$

Kantorovitch-Bedingung lautet:

$$\alpha(h)\omega(h) \leq \frac{1}{2} \tag{2.25}$$

Einsetzen liefert:

$$\frac{h^2 L_0 L_2}{(1-\mu h)^2} \leq \frac{1}{2}$$

Definition von $\bar{\tau}$ eingeführt:

$$\left(\frac{h}{1-\mu h}\right)^2 \leq \bar{\tau}^2$$

Für $\mu h < 1$ äquivalent zu:

$$\frac{h}{1-\mu h} \leq \bar{\tau} \tag{2.25'}$$

Für $\mu\bar{\tau} > -1$ gilt

$$\begin{aligned} h &\leq \bar{\tau}(1-\mu h) \quad (\text{ da } \mu h < 1) \\ h &\leq \bar{\tau}/(1+\mu\bar{\tau}) \end{aligned}$$

Ist $\mu\bar{\tau} \leq -1$, so folgt aus

$$(1+\mu\bar{\tau})h \leq \bar{\tau}$$

unbeschränkt. Beachte, daß $\mu h < 1$ unter der Bedingung (2.22.b) automatisch erfüllt ist. ∎

Bemerkung: Alternative Beweistechnik wie in Satz 1.3′ liefert hier *keine* Verbesserung.

Nichtlineare Stabilitätstheorie für allgemeine Diskretisierungsverfahren ist noch im Fluß. Im folgenden wird nur *lineare* Stabilitätstheorie verwendet. Dies führt zu folgenden Auswahlkriterien:

(I) $R(\infty) = 0$ (bei Einschrittverfahren),

(II) gute Modellierung der imaginären Achse.

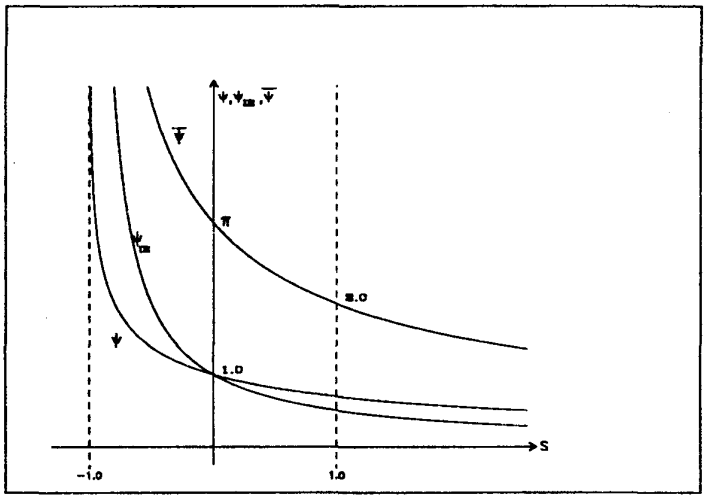


Bild B.6 Vergleich Ψ (Satz 1.3), $\bar{\Psi}$(Satz 1.3'), Ψ_{IE} (Satz 2.2)

2.3 Semi-implizite Extrapolationsverfahren

Zur Auswahl stehen zunächst *implizite* Einschrittverfahren mit Extrapolation:

(I) *implizite Trapezregel:*
Kap. A.2.2 $\longrightarrow$ h^2-Extrapolation.

(II) *implizite Euler-Diskretisierung*
Kap. A.2.2 $\longrightarrow$ h-Extrapolation.

Zu (I):

$$R(z) = \frac{1 + z/2}{1 - z/2}$$

$$| R(iy) | = 1, \; R(\infty) = -1$$

A-Stabilität

DAHLQUIST (1963)[20]:

h^2-Extrapolation mit $n_1 = 1$, zerstört bereits *A-Stabilität:*
$\longrightarrow$ $R(\infty) = 5/3$!

STETTER (1973) [106]:

$n_i = 2^{i+1}$ vorgeschlagen
$\longrightarrow | R(\infty) | = 1$

Bemerkung: Gleiche Resultate für implizite Mittelpunktsregel, da gleiche *lineare* Stabilitätstheorie

Zu (II):

$$R(z) = (1-z)^{-1}$$
$$|\, R(iy)\,| \leq 1 \quad , R(\infty) = 0$$
L-Stabilität

DEUFLHARD (1985)[32]:

h-Extrapolation mit $\mathcal{F}_H$ (vgl. Kap. A.2.3)

Lineare Stabilitätstheorie ($f = \lambda y$, $z = \lambda H$) erzeugt ein Extrapolationstableau von rationalen Funktionen:

$$\begin{array}{lll} R_{11}(z) & & \\ R_{21}(z) & R_{22}(z) & \\ R_{31}(z) & R_{32}(z) & R_{33}(z) \end{array} \tag{2.26}$$

Es gilt:

$$\begin{aligned} & R_{ik}(z) \sim \frac{l}{z^{i-k+1}} \text{ für } z \to \infty \\ & R_{i1}(z) = (1 - \frac{z}{i})^{-i} \sim \frac{1}{z^i} \\ & \frac{1}{z^i} \searrow \\ & \frac{1}{z^{i+1}} \to \frac{1}{z^i} \end{aligned} \tag{2.27}$$

$$R_{ik}(\infty) = 0 \tag{2.27'}$$

■

Zu lösen:

$$y_{k+1} - y_k - h\, f(y_{k+1}) = 0 \qquad k = 0, 1 \ldots \tag{2.28}$$

Vereinfachte Newton-Iteration:

$$\begin{aligned} (I - h\, A)\Delta y_k^i &= -(y_{k+1}^i - y_k - h\, f(y_{k+1}^i)) \\ y_{k+1}^{i+1} &:= y_{k+1}^i + \Delta y_{k+1}^i \\ \hookrightarrow y_{k+1} & \end{aligned} \tag{2.29}$$

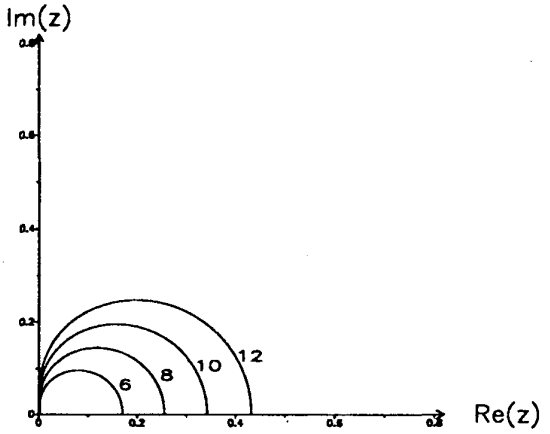


Bild B.7 Stabilitätsgebiete für implizite bzw. semi-implizite Euler-Diskretisierung mit h-Extrapolation ($z := \lambda\, H$)

Idee: (DEUFLHARD 1985) [32]:

nur 1 Iteration in (2.29)

Damit erhält man die sogenannte *semi-implizite Euler-Diskretisierung*:

$$\begin{aligned} (I - h\,A)\Delta y_k &= h\, f(y_k) \\ y_{k+1} &:= y_k + \Delta y_k \end{aligned} \tag{2.30}$$

Lineare Stabilitätstheorie identisch mit der für implizite Euler-Diskretisierung: es gilt (2.26), (2.27) und Bild B.6 (Stabilitätsgebiete).

Interpretation:

$$\begin{aligned} & \frac{(I - hA)y_{k+1} - y_k}{h} &&= f(y_k) - Ay_k \\ h \to 0: \quad & y' - Ay &&=: f(y) - Ay \end{aligned} \tag{2.31}$$

Ordnungs- und Schrittweitenssteuerung wie in Kapitel A.2.3, wobei jedoch Aufwand A_i abzuändern ist:

$$\begin{aligned} A_1 &:= C_J + C_{LR} + n_1 \cdot (C_{subst} + C_f) \\ A_i &:= A_{i-1} + C_{LR} + n_i \cdot C_{subst} + (n_i - 1)C_f \end{aligned} \tag{2.32}$$

C_J: Auswertung Jacobi-Matrix $A \approx f_y(y)$

C_{LR}: LR-Zerlegung von $(I - h_i A) = LR$

C_{subst}: Vorwärts-/Rückwärtssubstitution bei Gauss-Elimination

C_f: Auswertung von rechter Seite f.

Programm: EULSIM (DEUFLHARD)

Frage: Existiert semi-implizites h^2-Extrapolationsverfahren?

Semi-implizite Mittelpunktsregel (BADER/DEUFLHARD 1983)[3]
In Analogie zu Kap. A.2.3 (2.24.b) *symmetrische* Diskretisierung von $y' - Ay = \bar{f}(y)$ gewählt:

$$\frac{(I - hA)y_{k+1} - (I + hA)y_{k-1}}{2h} = f(y_k) - A\, y_k \tag{2.33.b}$$

Wie im expliziten Fall erhält man notwendigerweise ein *Zweischrittverfahren.*

Startschritt: semi-impliziter Euler

$$\frac{(I - hA)y_1 - y_0}{h} = f(y_0) - A\, y_0 \tag{2.33.a}$$

Schlußschritt (BADER):

$$y_l \to \frac{1}{2}(y_{l+1} + y_{l-1}) =: \hat{y}_l \tag{2.33.c}$$

Lineare Stabilitätstheorie:

$$(f = \lambda y, \quad z := \lambda h, \quad l := 2m)$$

Man erhält:

$$\begin{aligned} &\text{a)} \quad y_{2m+1} = \frac{1}{1-z}\left(\frac{1+z}{1-z}\right)^{m-1} \\ &\text{b)} \quad y_{2m} = \left(\frac{1+z}{1-z}\right)^{m} \\ &\text{c)} \quad \hat{y}_{2m} = \frac{1}{(1-z)^2}\left(\frac{1+z}{1-z}\right)^{m-1} \end{aligned} \tag{2.34}$$

Asymptotisches Verhalten für $z \to \infty$:

$$\begin{aligned} &\text{a)} \quad y_{2m+1} \to \frac{(-1)^m}{z} \to 0 \\ &\text{b)} \quad y_{2m} \to (-1)^m \\ &\text{c)} \quad \hat{y}_{2m} \to \frac{(-1)^{m-1}}{z^2} \to 0 \end{aligned} \tag{2.34'}$$

Im Unterschied zum Gragg'schen Schlußschritt (für die *explizite* Mittelpunktsregel) ist der Bader'sche Schlußschritt (für die *semi-implizite* Mittelpunktsregel) von zentraler Bedeutung!
Zusätzliche Forderung ($x := \Re(z)$):

$$e^x > 0 \longrightarrow \hat{y}_{2m} > 0 \tag{2.35}$$

Das führt zu der Spezifikation:

$$m - 1 = 2j \tag{2.35'}$$

Einsetzen liefert die Unterteilungsfolge:

$$\mathcal{F}_\alpha := \{2, 6, 10, 14, 22, \ldots\} \tag{2.35''}$$

Zusätzlich $\frac{n_i}{n_{i+1}} \leq \alpha = \frac{5}{7}$ vorgeschrieben (ohne Begründung).

Wegen (2.34'.*c*) gilt für das gesamte Extrapolationstableau:

$$R_{ik}(z) \sim \frac{1}{z^2} \tag{2.36}$$

und somit speziell:

$$R_{ik}(\infty) = 0 \tag{2.36'}$$

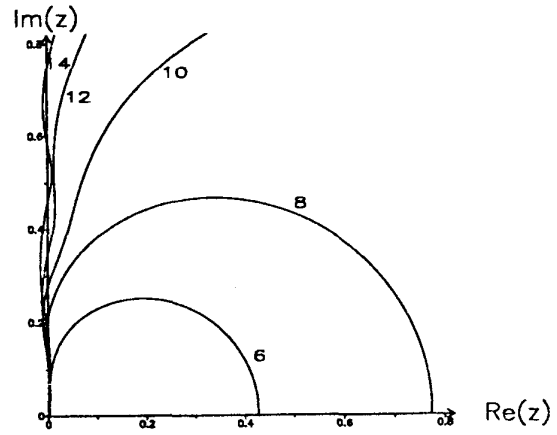


Bild B.8 Stabilitätsgebiete ($z := \lambda H$) für semi-implizite Mittelpunktsregel mit h^2-Extrapolation : Ordnung $k = 2m$

Stabilitätseigenschaften im Extrapolationstableau:

x	A-Stabilität
0	A(α)-Stabilität mit $\alpha \geq 86°$

x					
x	x				
x	x	x			
x	0	0	0		
x	0	0	0	0	
x	0	0	0	0	0

Satz 3 (BADER/DEUFLHARD 1983) [3]
Die semi-implizite Mittelpunktsregel (2.33.b) mit semi-implizitem Euler-Startschritt (2.33.a) besitzt eine asymptotische Entwicklung der Form ($lh = t$, $I - hA$ nichtsingulär):

$$y_l - y(t) = \sum_{j=l}^{N} [u_j(t) + (-1)^l v_j(t)]\ h^{2j} + E_{N+1}(t,h)h^{2N+2} \qquad (2.37.a)$$

für $f \in C^{2N+2}$. Dabei genügen die Koeffizientenfunktionen u_j, v_j dem Differentialgleichungssystem:

$$\begin{aligned} u_j' &= f_y(y)\cdot u_j + \ldots \quad (\textit{inhomogene Terme}) \\ v_j' &= (2A - f_y(y))v_j + \ldots \quad (\textit{inhomogene Terme}) \end{aligned} \qquad (2.37.b)$$

Bemerkung: Für $A \approx f_y(y_0)$ ist v_j nicht mehr "schwach instabil" $\Rightarrow$ $y(t) \sim v_j(t)$.

Beweis: Wie im Fall der expliziten Mittelpunktsregel (vgl. Kap. A.2, Satz 2.5) zunächst Umformung in symmetrisches Einschrittverfahren doppelter Dimension.

Bez.: $\bar{h} := 2h$, $\xi_k := y_{2k} \Rightarrow \xi_0 := y_0$
$y' = f(y) =: Ay + \bar{f}(y)$
$\hookrightarrow$ $\zeta_k := y_{2k+1} - \frac{\bar{h}}{2}[Ay_{2k+1} + \bar{f}(\xi_k)] \Rightarrow \zeta_0 = y_0 (2.41.a)$

$$y_{k+1} - y_{k-1} = hA(y_{k+1} + y_{k-1}) + 2h\bar{f}(y_k) \qquad (2.33'.b)$$

(*Umindizierung: gerader/ungerader Index*)

$$\xi_{k+1} - \xi_k = y_{2k+2} - y_{2k} = \qquad (*)$$
$$= \bar{h}[\tfrac{1}{2}A(y_{2k+2} + y_{2k}) + \bar{f}(y_{2k+1})] = \bar{h}[\tfrac{1}{2}A(\xi_{k+1} + \xi_k) + \bar{f}(y_{2k+1})] \quad (2.34.b)$$

$$\zeta_{k+1} - \zeta_k = y_{2k+3} - \frac{\bar{h}}{2}[Ay_{2k+3} + \bar{f}(\xi_{k+1})] - y_{2k+1} + \frac{\bar{h}}{2}[Ay_{2k+1} + \bar{f}(\xi_k)] \quad (**)$$

$$= \bar{h}[\tfrac{1}{2}A(y_{2k+1}) + \bar{f}(\overbrace{y_{2k+2}}^{\xi_{k+1}})] + \tfrac{\bar{h}}{2}[A(y_{2k+1}) + \bar{f}(\xi_k) - \bar{f}(\xi_{k+1})] \qquad (2.33.b)$$
$$= \tfrac{\bar{h}}{2}[2A\, y_{2k+1} + \bar{f}(\xi_k) + \bar{f}(\xi_{k+1})]$$

Definition von ζ_k kombiniert mit (**):

$$2\zeta_k + (\zeta_{k+1} - \zeta_k) = \zeta_{k+1} + \zeta_k = 2\{y_{2k+1} - \tfrac{\bar{h}}{2}Ay_{2k+1} - \tfrac{\bar{h}}{2}\bar{f}(\xi_k)\} +$$

$$+\tfrac{\bar{h}}{2}[2Ay_{2k+1} + \bar{f}(\xi_k) + \bar{f}(\xi_{k+1})] = 2y_{2k+1} + \tfrac{\bar{h}}{2}[\bar{f}(\xi_{k+1}) - \bar{f}(\xi_k)]$$

Damit symmetrische Darstellung für y_{2k+1}:

$$y_{2k+1} = \frac{1}{2}(\zeta_{k+1} + \zeta_k) - \frac{\bar{h}}{4}[\bar{f}(\xi_{k+1}) - \bar{f}(\xi_k)] \tag{2.38}$$

In (*):

$$\frac{\xi_{k+1} - \xi_k}{\bar{h}} = \frac{1}{2}A(\xi_{k+1} + \xi_k) + \bar{f}(y_{2k+1}) \tag{2.39.a}$$

In (**):

$$\frac{\zeta_{k+1} - \zeta_k}{\bar{h}} = A\ y_{2k+1} + \frac{1}{2}(\bar{f}(\xi_k) + \bar{f}(\xi_{k+1})) \tag{2.39.b}$$

Das Einschrittverfahren (2.39) ist konsistent mit dem Differentialgleichungssystem ($\xi \to x(t), \zeta \to z(t)$):

$$\begin{aligned} x' &= Ax + \bar{f}(z) \quad x(0) = y_0 \\ z' &= Az + \bar{f}(x) \quad z(0) = y_0 \end{aligned} \tag{2.40}$$

Also besitzt (2.39) eine asymptotische h-Entwicklung. Darüberhinaus ist das Einschrittverfahren (2.39) + (2.38) *symmetrisch* gegen Transformation:

$$(\xi_k,\ \xi_{k+1},\ \zeta_k,\ \zeta_{k+1},\ h) \Longleftrightarrow (\xi_{k+1},\ \xi_k,\ \zeta_{k+1},\ \zeta_k,\ -h)$$

Also existiert h^2-*Entwicklung.*

Rücktransformation:

$$y_{2k} = \xi_k\ ,\ y_{2k+1} = \dots \quad (2.38)$$

liefert zwei getrennte h^2-Entwicklungen (analog zum Beweis von Satz 2.5, Kap. A). ∎

Der Bader'sche Schlußschritt (2.33.c) ist symmetrisch

$$\hat{y}_{2m} = \frac{1}{2}(y_{2m+1} + y_{2m-1})$$

und hat damit eine asymptotische Entwicklung der Form:

$$\hat{y}_{2m} - y(t_{2m}) = \sum_{j=1}^{N} g_j(t)h^{2j} + \hat{E}_{N+1}(t,h)h^{2N+1} \tag{2.41.a}$$

Hier gilt jedoch i.a.:

$$g_j(0) \neq 0 \quad \text{für} \quad A \neq 0 \tag{2.41.b}$$

Beweis:

$$\begin{aligned}\hat{y}_0 - y_0 &= \tfrac{1}{2}(y_1(h) + y_1(-h)) - y_0 = \\ &= \tfrac{1}{2}[(y_1(h) - y_0) + (y_1(-h) - y_0)] = \\ &= \tfrac{1}{2}[(I - hA)^{-1} h f(y_0) + (I + hA)^{-1}(-h) f(y_0)] \neq 0 \text{ für } A \neq 0\end{aligned}$$

■

Zur Restgliedabschätzung für beide Diskretisierungen (2.30) und (2.33):

(1) Natürlich gilt Abschätzung über Lipschitzkonstante L_1, die jedoch für *steife* Probleme nicht sachgerecht ist, da $hL_1 \leq C = O(1)$ zu einschränkend.

(2) Ersetzt man das Fundamentallemma Satz 1.4 (Kapitel A.1.2)

$$\| u(t) - v(t) \| \leq \| u(0) - v(0) \| e^{L_1 t}$$

durch das Fundamentallemma Satz 1.2 (Kapitel B.1.1)

$$\| u(t) - v(t) \| \leq \| u(0) - v(0) \| e^{\bar{\mu} t}$$

mit *globaler* Kontraktivitätskonstante $\bar{\mu} \gtrless 0$, so läßt sich die "Fächertechnik" im Beweis zu Satz 2.1 (Kap. A.2.1) direkt übertragen. Meist läßt sich jedoch keine bessere Konstante als $\bar{\mu} = L_1$ abschätzen.

(3) Wünschenswert wäre Abschätzung mit der Charaktisierung $\mu, \bar{\tau}$ - fehlt jedoch (eventuell nicht möglich, sicher schwieriger).

Programmierung der Diskretisierung (2.33) erfolgt über das *kompakte Schema* ($\Delta_k := y_{k+1} - y_k, l \in \mathcal{F}_\infty$):

$$\begin{aligned}\Delta_0 &:= (I - hA)^{-1} h f(y_0) \\ y_1 &:= y_0 + \Delta_0\end{aligned} \tag{2.42.a}$$

$$\begin{aligned}k &= 1, \ldots, l-1 : \\ \Delta_k &:= \Delta_{k-1} + 2(I - hA)^{-1}[h f(y_k) - \Delta_{k-1}] \\ y_{k+1} &:= y_k + \Delta_k\end{aligned} \tag{2.42.b}$$

$$\begin{aligned}\hat{\Delta}_l &:= (I - hA)^{-1}[h f(y_l) - \Delta_{l-1}] \\ \hat{y}_l &:= y_l + \hat{\Delta}_l\end{aligned} \tag{2.42.c}$$

Vermeidet Auswertung von $\bar{f}$ (Gefahr der Auslöschung!). Ordnungs- und Schrittweitensteuerung wie in Kapitel A.2.3, wobei Aufwand abzuändern gemäß:

$$\begin{aligned}A_1 &:= C_J + C_{LR} + (n_1 + 1)(C_{\text{subst}} + C_f) \\ A_i &:= A_{i-1} + C_{LR} + C_{\text{subst}} + n_i(C_{\text{subst}} + C_f)\end{aligned} \tag{2.43}$$

(⟶ "interne Uhr" oder Approximation)

Programm METAN1 (BADER/DEUFLHARD [3])

Numerische Schätzung der *lokalen* Kontraktionskonstante μ

(DEUFLHARD 1987) Beide Diskretisierungen in EULSIM und METAN1 beginnen mit einem semi-impliziten Euler-Schritt

$$\Delta_0(h) = (I - hA)^{-1} hf(y_0) \ , \ A \approx f_y(y_0).$$

Man hat mindestens für $i = 1, 2$:

$$d_i := \| \Delta_0(h_i) \| \quad (\| \cdot \| \text{ skalierte Norm}) \tag{2.44}$$

(2.24.a) liefert

$$d_i \leq h_i L_0/(1 - \mu h_i) \ , \ \text{falls } \mu h_i < 1. \tag{2.45}$$

Definiere

$$\begin{aligned} &\text{a) } \kappa_i &:=& \ d_i/d_{i-1} \quad \text{falls } d_{i-1} \neq 0 \\ &\text{b) } \bar{\kappa}_i &:=& \ h_i/h_{i-1} = n_{i-1}/n_i < 1 \end{aligned} \tag{2.46}$$

Nimmt man die Schrittweitenreduzierung als vernünftig an, muß für die Korrekturen gelten:

$$d_i < d_{i-1}, \text{ d.h. } \kappa_i < 1 \ .$$

Nimmt man in (2.45) Gleichheit an und dividiert für $i, i-1$, so gilt etwa:

$$\begin{aligned} \kappa_i &\doteq \bar{\kappa}_i \frac{1 - \mu h_{i-1}}{1 - \mu h_i} \\ &= \bar{\kappa}_i \frac{1 - \mu h_i/\bar{\kappa}_i}{1 - \mu h_i} \end{aligned}$$

Auflösen nach μh_i liefert

$$\mu h_i \doteq \frac{\bar{\kappa}_i - \kappa_i}{1 - \kappa_i} =: [\mu]_i h_i \tag{2.47}$$

Aufgrund der Korrekturen (2.44) erfolgt die Schätzung in Richtung der Trajektorien → angemessen für steife Probleme.

2.4 Implizite und semi-implizite Runge-Kutta-Verfahren

In Kapitel A.2.4 wurden *explizite* Runge-Kutta-Methoden vorgestellt. Einsetzen der skalaren Testgleichung in die allgemeine Form (2.32) führt auf *Polynome* $P_s(z)$ für explizite Runge-Kutta-Verfahren der Stufe s. Wegen $P_s(\infty) = \infty$ eignen sich solche Verfahren *nicht* für steife Probleme.

Idee: (BUTCHER 1964)[15]:
Erweiterung des Ansatzes zu impliziten Runge-Kutta-Methoden (IRK):

$$\begin{aligned} k_1 &= h\, f(y_0 + \sum_{j=1}^{s} a_{1j} k_j) \\ &\vdots \quad\quad \vdots \\ k_s &= h\, f(y_0 + \sum_{j=1}^{s} a_{sj} k_j) \end{aligned} \tag{2.48.a}$$

$$y_1 = y_0 + b_1\, k_1 + \ldots + b_s\, k_s \tag{2.48.b}$$

$$C_i := \sum_{j=1}^{s} a_{ij} \tag{2.48.c}$$

Schema:

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & & \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array} \qquad \begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}$$

Für die $k_i, \ldots, k_s$ hat man ein i.a. nichtlineares Gleichungssystem der *Dimension* $s \cdot n$ zu lösen. Mit der Bezeichnung:

$$k_1 = h\, f(g_i)$$

erhält man:

$$\begin{aligned} g_i &= y_0 + h \sum_{j=1}^{s} a_{ij}\, f(g_j) \\ i &= 1, \ldots, s \end{aligned} \tag{2.48'.a}$$

Die Graphenmethode von BUTCHER ist direkt erweiterbar: einzige Änderung ist die Erweiterung sämtlicher Summen:

$$\sum_{j=1}^{i=1} \longrightarrow \sum_{j=1}^{s}$$

Herleitung der Bedingungsgleichungen damit prinzipiell klar.

Lineare Stabilitätstheorie: Skalare Testgleichung eingesetzt in (2.48'.a)

$$\begin{bmatrix} g_1 \\ \vdots \\ g_s \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} y_0 + z \cdot \mathcal{A} \begin{bmatrix} g_1 \\ \vdots \\ g_s \end{bmatrix} \tag{2.49}$$

Bezeichnung: $g^T := (g_1, \ldots, g_s)$, $k^T := (k_1, \ldots, k_s)$

$e^T := (1, \ldots, 1) \in \mathbb{R}^s$

$\hookrightarrow \quad g = e\, y_0 + z\, \mathcal{A} g$

$$(I_s - z\mathcal{A})\, g = e\, y_0 \ , \quad k = z\, g \tag{2.49'}$$

Sei $I_s - z\mathcal{A}$ nichtsingulär, so gilt:

$$y_1 = y_0 + b^T k = [1 + z b^T (I_s - z\mathcal{A})^{-1} e]\ y_0 \tag{2.50}$$

$\hookrightarrow$ Stabilitätsfunktion:

$$R_s(z) = 1 + b^T (\frac{1}{z} I_s - \mathcal{A})^{-1} e \tag{2.50'}$$

$\hookrightarrow$ natürliche Forderung:

$$\mathcal{A} \text{ nichtsingulär} \tag{$*$}$$

Dann gilt:

$$R_s(\infty) = 1 - b^T \mathcal{A}^{-1} e \tag{2.51}$$

Dieser Ausdruck sollte also für effizientes IRK-Verfahren verschwinden.

Beispiel: Übertragung des "Fehlberg-Tricks" auf IRK

$$a_{sj} = b_j \quad j = 1, \ldots, s \tag{2.51.a}$$

Damit gilt:

$$\begin{aligned} c_s &= 1 \\ y_1 &= g_s = y_0 + h \sum_{j=1}^{s} b_j f(g_j) \end{aligned}$$

Umformung:

$$\begin{aligned} b^T &= e_s^T \mathcal{A} \\ e_s^T &= (0, \ldots, 0, 1) \end{aligned} \tag{2.51'.a}$$

In (2.50):

$$R_s(\infty) = 1 - e_s^T \mathcal{A}\mathcal{A}^{-1} e = 1 - \underbrace{e_s^T e}_{1} = 0 \tag{2.51.b}$$

Bemerkung:

1. (2.51.a) ist ohnehin sinnvollste Erweiterung von ERK $\rightarrow$ IRK (nach Struktur des Ansatzes).

2. Gleiche Stabilitätseigenschaft für jeden internen Zwischenschritt: hierzu ersetzt man lediglich $y_1 \rightarrow g_i$ sowie $e_s \rightarrow e_i$, $i = 1, \ldots, s$.

Zugehöriges $s \cdot n$-Gleichungssystem:

$$G(g_1, \ldots, g_s\, h) := \begin{cases} g_1 - y_0 - h\sum_{j=1}^{s} a_{1j}\, f(g_j) \\ \vdots \qquad\qquad = 0 \\ g_s - y_0 - h\sum_{j=1}^{s} a_{sj}\, f(g_j) \end{cases}$$

$$h = 0 : g_i = y_0 \quad i = 1, \ldots, s$$

Jacobi-Matrix: $J = \dfrac{\partial G}{\partial g}$

wobei:

$$G := (G_1, \ldots, G_s)^T \;, A := f_y(y_0)$$

$$\frac{\partial G_i}{\partial g_l}\,|_{g_l = y_0} = \delta_{li}\, I_n - h \cdot a_{il}\, f_y(g_e)\,|_{g_l = y_0}$$

$$= \delta_{li}\, I_n - h\, a_{il} \underbrace{f_y(y_0)}_{A}$$

Die Matrix $\dfrac{\partial G}{\partial g}$ ist also (s,s)-Blockmatrix bestehend aus (n,n)-Blöcken:

$$J = \begin{bmatrix} I_n - ha_{11}A, & -ha_{12}A, & \ldots & -ha_{1s}A \\ -ha_{21}A & I_n - ha_{22}A & \ddots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ \vdots & & & -ha_{s-1,s}A \\ \vdots & & & \\ -ha_{s,1}A \ldots & \ldots & -ha_{s,s-1}A\;, & I_n - ha_{ss}A \end{bmatrix} \qquad (2.52.a)$$

Für $h = 0 : J = I$ nichtsingulär
$\hookrightarrow$ Lösung $y_1(h) = g_s(h)$ lokal eindeutig fortsetzbar von $g_s(0) = y_0$ aus (Homotopie).
Vereinfachtes Newton-Verfahren wäre:

$$J\Delta g^k = -G(g^k) \qquad (2.52.b)$$

Vereinfachungen von IRK:

DIRK: diagonal-implizite RK

$$a_{1j} = 0 \quad \text{für } i < j \tag{2.53.a}$$

$$\begin{array}{c|ccc} * & * & & \\ \vdots & & \ddots & 0 \\ * & * & \dots & * \\ \hline & * & \dots & * \end{array}$$

$\hookrightarrow$ Auflösung des Gleichungssystems (2.52) erfordert nur Zerlegung von s Matrizen der Form:

$$\begin{gathered} I_n - h a_{ii} A = L_i R_i \\ \hookrightarrow \quad \| (I_n - h\, a_{ii} A)^{-1} \| \leq \frac{1}{1 - h\, a_{ii} \mu} \end{gathered} \tag{2.53.b}$$

$$a_{ii} > 0 \quad \text{notwendig} \tag{2.53.c}$$

SDIRK: singly diagonally implicit **R**unge-**K**utta

$$\begin{gathered} a_{ii} = \gamma \quad i = 1, \dots, s \\ \text{zusätzlich zu (2.53).} \\ \hookrightarrow \quad \text{nur noch 1 } (n,n) - \text{Matrix-Zerlegung :} \end{gathered} \tag{2.54}$$

$$I - \gamma h A = LR$$

$$R_s(z) = \frac{P_s(z)}{(1 - \gamma z)^s}$$

SIRK: singly implicit **R**unge-**K**utta
(BURRAGE/BUTCHER/CHIPMAN(1979)[13]$\longrightarrow$ Programm STRIDE)
Hierbei hat $\mathcal{A}$ den s-fachen Eigenwert γ

Semi-implizite Runge-Kutta-Verfahren.

Nur ein *lineares* Gleichungssystem zu lösen, keine Newton-Iteration.

Rosenbrock-Verfahren

(ROSENBROCK, WANNER) Ansatz (2.47) erweitert durch Hinzunahme von $f_y(y_0)$ in Ordnungsbedingungen.

Nachteil: bei großen Beispielen liefert Rückwärtsanalyse der auftretenden linearen Gleichungssystemen eine Störung von $f_y(y_0)$.
$\longrightarrow$ Störung der Ordnungsbedingungen.
(Vergleiche KAPS/RENTROP[68]: 1979 GRK4T, $R(\infty) \neq 0$)

Es entstehen *mehr* Bedingungsgleichungen als bei IRK.

W-Methoden

(WOLFBRANDT, WANNER)
Hierbei genügt $A \approx f_y(y_0)$ bei Auflösung linearer Gleichungssysteme

$\hookrightarrow$ Erweiterung der Butcher-Graphen:
$\hookrightarrow$ noch mehr Bedingungsgleichungen

Bisher wurden sowohl Rosenbrock- als auch W-Methoden nur bis Ordnung 4-5 konstruiert. Die einzigen W-Methoden *höherer* Ordnung sind bisher die semi-impliziten Extrapolationsverfahren (Kapitel B.2.3).

3 Mehrschrittverfahren

Lineares k-Schritt-Verfahren, (vergleiche Kapitel A.3.3),

$$\begin{aligned} &\text{a)} \quad \rho(E_h)y_n = h\sigma(E_h)f(y_n) \\ &\text{b)} \quad \rho(\zeta) := \alpha_k\zeta^k + \ldots + \alpha_0\,, \;\; \sigma(\zeta) := \beta_k\zeta^k + \ldots + \beta_0 \end{aligned} \tag{3.1}$$

3.1 Lineare Stabilitätstheorie

Einsetzen der skalaren Testgleichung in (3.1) liefert ($z := \zeta h \in \mathbb{C}$):

$$(\alpha_k - \beta_k z)\, y_{n+k} + \ldots + (\alpha_0 - \beta_0 z)y_n = 0 \tag{3.2}$$

Ansatz $y_n = \zeta^n$ liefert *charakteristische Gleichung*:

$$\rho(\zeta) - z\sigma(\zeta) = 0 \tag{3.3}$$

Seien $\zeta_j(z)$ die Wurzeln dieser Gleichung.
Beschränktheit der Lösungen von (3.1) unabhängig von Startwerten $y_0, \ldots, y_{k-1}$ verlangt Wurzelkriterium für $\zeta_j(z)$, vergleiche (3.11), Lemma 3.1, Kapitel A.3.1 (dort nur für $\zeta_j(0)$ verlangt).

Stabilitätsgebiet:

$$\begin{aligned} G \;:=\; \{z \in \mathbb{C} \;||\; &|\zeta_j(z)| \leq 1, j = 1, \ldots, k, \\ &\zeta_j(z) \text{ einfach, falls } |\zeta_j(z)| = 1\} \end{aligned} \tag{3.4}$$

Damit lautet Lemma 3.1, Kap. A.3.1:

$$0 \in G \tag{3.5}$$

(Dahlquist'sches Wurzelkriterium)

Parametrisierung von ∂G:

$$|\zeta| = 1 \longrightarrow \zeta = e^{i\varphi}$$

liefert die sogenannte *Wurzelortskurve*

$$\Gamma := \{z \in \mathbb{C} \mid z = \frac{\rho(e^{i\varphi})}{\sigma(e^{i\varphi})},\ \varphi \in [0, 2\pi]\} \tag{3.6}$$

Γ berücksichtigt nicht mögliche mehrfache Wurzeln ζ_j, also gilt nur:

$$\partial G \subset \Gamma \tag{3.7}$$

A-Stabilität:

$$\mathbb{C}^{-} \subseteq G \tag{3.8}$$

Beispiele: explizite Mittelpunktsregel

$$y_{n+2} - y_n = 2h\zeta\, y_{n+1} = 2z\, y_{n+1}$$
$$\hookrightarrow \quad \rho(\zeta) = \zeta^2 - 1,\ \sigma(\zeta) = 2\zeta$$

charakteristische Gleichung:

$$\zeta^2 - 1 - 2z\,\zeta = 0$$
$$\hookrightarrow \quad \zeta_1 = z + \sqrt{1+z^2},\ \zeta_2 = z - \sqrt{1+z^2}$$

Wurzelortskurve Γ:

$$z = \frac{\rho(e^{i\varphi})}{\sigma(e^{i\varphi})} = \frac{e^{2i\varphi} - 1}{2e^{i\varphi}} = \frac{1}{2}(e^{i\varphi} - e^{-i\varphi}) = i\sin(\varphi)$$
$$\hookrightarrow \quad \Gamma = \{z \epsilon \mathbb{C} \mid z = i\tau,\ \tau \epsilon [-1, +1]\}$$

mehrfache Wurzeln:

$$\zeta_1 = \zeta_2 \Longleftrightarrow 1 + z^2 = 0 \Longleftrightarrow z = \pm i$$

Stabilitätsgebiet:

$$G = \partial G = \{z \in \mathbb{C} \mid z = i\tau,\ \tau \in]-1, +1[\}$$
$$\partial G \subset \Gamma$$

keine A-Stabilität

Satz 1 ("2. Dahlquist-Schranke", DAHLQUIST 1963)
A-Stabilität für lineare k-Schritt-Verfahren der Form (3.1) impliziert notwendig:

$$\begin{aligned} &a) \quad \beta_k \neq 0 \text{ (implizites Mehrschrittverfahren)} \\ &b) \quad \text{Konsistenzordnung } p \leq 2 \end{aligned} \tag{3.9}$$

Sei $\alpha_k := 1$ ohne Beschränkung der Allgemeinheit. Unter den A-stabilen Verfahren mit $p = 2$ hat gewiß die implizite Trapezregel die betragskleinste (normierte) Fehlerkonstante:

$$C_2^* = C_2/\sigma(1) \; = \; -\frac{1}{12}$$

Beweis: Der ursprüngliche Dahlquist'sche Beweis von 1962 ist lang und kompliziert. Verbesserung durch Theorie der "Ordnungssterne" von WANNER/HAIRER/NØRSETT (1978) [111]. Hier wird der kurze elementare Beweis von GRIGORIEFF (1977) [50] dargestellt.
Man betrachtet die komplexe Funktion:

$$g(\zeta) := \frac{\sigma(\zeta)}{\rho(\zeta)} - \frac{1}{2}\,\frac{\zeta+1}{\zeta-1} \tag{3.10.a}$$

Entwicklung von ρ um $\zeta = 1$ ergibt

$$\rho(\zeta) = \rho(1) + \rho'(1)(\zeta-1) + \mathcal{O}((\zeta-1)^2)\,,$$

so daß man mit Konsistenzbedingung (3.8), Kapitel A.3:

$$\rho(1) = 0\,, \quad \rho'(1) = \sigma(1)$$

erhält:

$$\rho(\zeta) = \sigma(1)(\zeta-1) + \mathcal{O}((\zeta-1)^2)\,, \text{ also}$$

$$\begin{aligned} g(\zeta) \; &= \left(\frac{\sigma(1)}{\sigma(1)(\zeta-1)} - \frac{1}{2}\frac{2}{\zeta-1}\right) + \mathcal{O}(\zeta-1) \text{ für } \zeta \neq 1\,. \\ &= \mathcal{O}(\zeta-1) \end{aligned} \tag{$*$}$$

Nach (3.3) gilt für $|\zeta| > 1$ und

$$\rho(\zeta) - z\sigma(\zeta) = 0\,, \text{ daß}$$

$$z \notin G\,, \text{ insbesondere}$$

$$\Re z > 0\,, \text{ da wir A-Stabilität voraussetzen.}$$

Damit ist $\frac{1}{z} = \frac{\sigma(\zeta)}{\rho(\zeta)}$ analytisch in $|\zeta| > 1$ (σ, ρ teilerfremd!), so daß

$$\Re \frac{\sigma(\zeta)}{\rho(\zeta)} > 0 \quad \text{in } |\zeta| > 1 .$$

Weiter gilt wegen:

$$\Re\left(\frac{\zeta+1}{\zeta-1}\right) = \frac{|\zeta|^2 - 1}{|\zeta|^2 + 1 - 2\Re\zeta} \text{ , daß}$$

$$\Re\left(\frac{\zeta+1}{\zeta-1}\right) = 0 \quad \text{für } |\zeta| = 1 , \ \zeta \neq 1 . \qquad (**)$$

Sei nun $\varepsilon > 0$ gegeben. Wegen (*) gibt es eine δ-Umgebung $\mathcal{U}$ von 1 mit

$$|g(\zeta)| < \varepsilon \text{ für } |\zeta| > 1 , \ \zeta \in \mathcal{U} .$$

Ferner ist $-\frac{1}{2}\frac{\zeta+1}{\zeta-1}$ in $\{\zeta \in \mathbb{C} \| \zeta| \leq 2\} \cap \mathcal{U}$ gleichmäßig stetig, so daß es ein $1 < \vartheta_\varepsilon < 1 + \varepsilon$ mit

$$\Re\left(-\frac{1}{2}\frac{\zeta-1}{\zeta+1}\right) \geq -\varepsilon \text{ für } |\zeta| = \vartheta_\varepsilon \text{ und } \zeta \notin \mathcal{U}$$

gibt (vergleiche (**)).
Insgesamt folgt

$$\Re g(\zeta) \geq -\varepsilon \text{ für } |\zeta| = \vartheta_\varepsilon .$$

Aus dem Maximumsprinzip (hier für Minimum!!) angewendet auf $\{|\zeta| \geq \vartheta_\varepsilon\}$ und $\varepsilon \downarrow 0$ folgt

$$\Re g(\zeta) \geq 0 \text{ für } |\zeta| > 1 . \qquad (3.10.b)$$

Sei nun $p \geq 2$, Entwicklung (3.10'), Kapitel A.3 herangezogen ergibt

$$\frac{\rho(\zeta)}{\ln \zeta} - \sigma(\zeta) = C_p(\zeta - 1)^p + \mathcal{O}((\zeta-1)^{p+1}) ,$$

das heißt

$$\begin{aligned} \frac{\rho(\zeta)}{\sigma(\zeta)} &= \ln\zeta + \tfrac{C_p}{\sigma(\zeta)} \ln\zeta (\zeta-1)^p + \mathcal{O}((\zeta-1)^{p+1}) , \ \zeta \to 1 \\ &= (\zeta-1) + \frac{1}{2}(\zeta-1)^2 + (\frac{1}{3} + C)(\zeta-1)^3 + \mathcal{O}((\zeta-1)^4) , \ \zeta \to 1 , \end{aligned}$$

$$\text{mit } C = \begin{cases} \dfrac{C_2}{\sigma(1)} & p = 2 \\ 0 & p > 2 \end{cases} \qquad (3.11)$$

Man rechnet sofort nach, daß damit

$$\begin{aligned} \frac{\sigma(\zeta)}{\rho(\zeta)} &= \frac{1}{\zeta-1}+\frac{1}{2}-(\frac{1}{12}+C)(\zeta-1)+\mathcal{O}((\zeta-1)^2) \\ \text{und } -\frac{1}{2}\frac{\zeta+1}{\zeta-1} &= -\tfrac{1}{\zeta-1}-\frac{1}{2} \text{ für } \zeta\to 1\,, \text{ also} \\ g(\zeta) &= -(\tfrac{1}{12}+C)(\zeta-1)+\mathcal{O}((\zeta-1)^2) \text{ für } \zeta\to 1\,. \end{aligned}$$

Somit erhalten wir aus (3.10.b) , wenn wir $\zeta = 1+\varepsilon$, $\varepsilon > 0$ setzen:

$$\begin{aligned} \tfrac{1}{12}+C &\leq 0\,, \quad \text{das heißt} \\ C &\leq -\tfrac{1}{12}\,. \end{aligned}$$

Nach (3.11) scheidet daher $p > 2$ aus, und wir erhalten

$$C_2^* = \frac{C_2}{\sigma(1)} \leq -\frac{1}{12}$$

Für $p = 2$ und $C_2^* = -\frac{1}{12}$ folgt aus der Entwicklung von $g(\zeta)$, daß $g(\zeta)$ bei 1 eine mehrfache Nullstelle hat. Wie man sich leicht überzeugt, ist dies mit (3.10.b) nur für $g \equiv 0$ möglich. Aus der Teilerfremdheit von ρ, σ und $\alpha_k = 1$ folgt daher

$$\begin{aligned} \rho &= \zeta - 1 \\ \sigma &= \tfrac{1}{2}(\zeta+1)\,, \end{aligned}$$

das heißt die implizite Trapezregel. ∎

3.2 BDF - Verfahren

(CURTISS/HIRSCHFELDER 1951 [18], GEAR 1971 [48])

Einschrittverfahren: Wichtiger als *A-Stabilität* ist $R(\infty) = 0$

Übertragung dieser Forderung auf Mehrschrittverfahren:

$$\zeta_j(\infty) = 0,\ j = 1,\ \ldots,\ k$$

Anwendung auf Mehrschrittverfahren:

$$(\alpha_k - \beta_k z)\zeta^k + (\alpha_{k-1} - \beta_{k-1}z)\zeta^{k-1} + \ldots + (\alpha_0 - \beta_0 z) = 0 \qquad (3.3')$$

Abdividieren:

$$\zeta^k + \frac{\alpha_{k-1} - \beta_{k-1}z}{\alpha_k - \beta_k z}\zeta^{k-1} + \ldots + \frac{\alpha_0 - \beta_0 z}{\alpha_k - \beta_k z} = 0$$

Forderung:

$$z \longrightarrow \infty : \quad \zeta^k = 0$$

$$\Longrightarrow \beta_{k-1} = \ldots = \beta_0 = 0 \tag{3.13}$$

zugehöriges Mehrschrittverfahren:

$$\alpha_k y_{n+k} + \ldots + \alpha_0 y_0 = h\beta_k f(y_{n+k}) \tag{3.13'}$$

BDF-Verfahren (**B**ackward **D**ifferentiation **F**ormula)

Approximationsidee:

$$\begin{array}{rcl} y(t) & \longrightarrow & p(t) \\ p(t) & := & P_k(t \mid t_{n+k}, \ldots, t_n) \end{array} \tag{3.14}$$

Differentialgleichung ist damit zu ersetzen durch:

$$p'(t_{n+k}) = f(p(t_{n+k})) \tag{3.15}$$

Darstellung von P_k durch *Rückwärtsdifferenzen* (analog zu Adams-Verfahren, vergleiche Kapitel A.3.2).

BDF über äquidistantem Gitter

Man erhält ein Mehrschrittverfahren vom Typ (3.13').
Für $k > 6$: Stablilitätsbedingung (Wurzelkriterium für $\zeta_j(0)$) verletzt!
(3.13') verlangt die Lösung von:

$$F(y) := y - h\,\beta_k\, f(y) - \varphi(y_{n+k-1}, \ldots, y_n) = 0 \tag{3.16}$$

wobei $\alpha_k := 1$ o.B.d.A.

$$\hookrightarrow y = y_{n+k}$$

Newton-ähnliche Iteration:

$$F_y = I - \beta_k\, h\, f_y(y) \rightarrow I - \beta_k\, hA$$
$$A \approx f_y(y)$$

$$\begin{array}{rcl} y^{i+1} & := & y^i - (I - \beta_k\, hA)^{-1}\, F(y^i) \\ \hookrightarrow \quad y^* & =: & y_{n+k} \end{array} \tag{3.17}$$

Konvergenz dieser Iteration gesichert, falls Approximation A "hinreichend gut" und h "hinreichend klein". Algorithmische Überprüfung durch Überwachung der Kontraktion. Man beachte:

$$\| (I - \beta_k\, hA)^{-1} \| \leq \frac{1}{1 - \beta_k\, h\mu} \tag{3.18}$$

wobei:

$$\langle u, Au\rangle \leq \mu\langle u, u\rangle \equiv \mu \parallel u \parallel^2$$

Tatsächlich liefert (3.15) $\beta_k > 0$, d.h. Iterationsmatrix regulär für $\mu < 0$.
Startwert für (3.17):

$$y^0 = P_{\text{alt}}(t_{n+k})$$

(Stabilitätsprobleme für $k > 2$)

Schrittweiten- und Ordnungssteuerung ähnlich wie bei ABM (Kapitel A.3.2) beschrieben (meist Nordsieckform + "Faustformeln"). Regeneration von A an Kontraktionsverhalten gekoppelt.

Lineare Stabilitätseigenschaften: *A-Stabilität* für $k = 1, 2$
($k = 1$: impliziter Euler)
$A(\alpha)$-Stabilität:

Tab. k	3	4	5	6
α	$\sim 86°$	$\sim 76°$	$\sim 50°$	$\sim 16°$

Programme: LSODE (HINDMARSH, LLNL)
DASSL (PETZOLD, LLNL)

Bemerkung: Semi-implizite Variante (nur 1 - 2 Iterationen) versucht von KROGH/STEWART (1981) [75]- noch nicht ausgereift.

BDF über variablem Gitter

Man geht zurück auf Darstellung (3.15) und dividierte Differenzen.
Programm EPISODE (BYRNE/HINDMARSH) :

Anlehnung an äquidistanten Fall soweit möglich
$\hookrightarrow$ verläßlicher, aber langsamer als LSODE.

Stabilitätstheorie GRIGORIEFF (1983)[51]: Schrittweitenvariation asymptotisch (für $n \to \infty$) erlaubt für:

$$\omega \leq \frac{h_n}{h_{n+1}} \leq \Omega$$

	k	2	3	4	5
Tab.	ω	0	0.84	0.98	0.997
	Ω	2.4	1.13	1.02	1.003

Für endliche n sind stärkere Variationen zulässig (BOCK/EICH 1987).

4 Implizite Differentialgleichungen und differentiell-algebraische Systeme

4.1 Theoretische Grundlagen

Die allgemeinste Form impliziter Differentialgleichungen wäre

$$F(y, y', t) = 0 \ , \ y_0 \text{ gegeben.}$$

Hierbei könnte, wie bisher, die explizite t-Abhängigkeit als wegtransformiert angenommen werden (Autonomisierung). In den Anwendungen tritt jedoch fast ausnahmslos die *quasilineare Form impliziter Differentialgleichungen* auf:

$$B(t,y)\ y' = f(t,y) \tag{4.1}$$

Autonome Variante:

$$B(y)y' = f(y)\ , \qquad y_0 \text{ gegeben.} \tag{4.1'}$$

Anwendungen:

Verfahrenstechnik	(Entwurf von chemischen Anlagen)
Mehrkörperdynamik	(Straße/Fahrwerk, Schiene/Zug)
diskretisierte partielle DG	(bewegte räumliche Gitter bzw. bewegte Finite Elemente)

Linearer Spezialfall
Sei zunächst betrachtet:

$$By' - Ay = f(t) \ , \ y_0 \text{ gegeben.} \tag{4.2}$$

Falls B regulär, ist (4.2) äquivalent zu einem expliziten DG-System. Es existieren jedoch auch Lösungen für B *singulär*. Zur Untersuchung dieses Falls transformiert man auf die Form:

$$\bar{B}z' - \bar{A}z = \bar{f}(t) \tag{4.3}$$

wobei definiert:

$$\begin{aligned} \bar{B} &:= PBQ\ ,\ \bar{A} := PAQ \\ \bar{f} &:= Pf\ ,\ y = Qz \end{aligned}$$

für *nichtsinguläre* Matrizen P, Q.

Bemerkung: Algorithmus zur Konstruktion von $P, Q, \bar{B}, \bar{A}$: AB-Algorithmus von (KUBLANOVSKAJA [76], (Leningrad)).

Nach KRONECKER (1890) [77] existieren Matrizen P, Q derart, daß (4.2) zerfällt in Subsysteme der folgenden Form mit der Bezeichnung: $z^T := (z_1, \ldots, z_n)$, $\bar{f}(t)^T = \left(\bar{f}_1(t), \ldots, \bar{f}_n(t)\right)$.

$$
\begin{array}{lrcll}
a) & 0 \cdot z_i & = & \bar{f}_i(t) & i = 1, \ldots, n_0 \\
b) & z_i' + z_{i+1} & = & \bar{f}_i(t) & i = n_0 + 1, \ldots, n_1 - 1 \\
c) & z_{n_1+1}' & = & \bar{f}_{n_1+1}(t) & \\
 & z_{i+1}' + z_i & = & \bar{f}_{i+1}(t) & i = n_1 + 1, \ldots, n_2 - 1 \\
 & z_{n_2} & = & \bar{f}_{n_2+1}(t) & \\
d) & z_{i+1}' + z_i & = & \bar{f}_i(t) & i = n_2 + 1, \ldots, n_3 - 1 \\
 & z_{n_3} & = & \bar{f}_{n_3}(t) & \\
e) & z_{\text{red}}' + J z_{\text{red}} & = & \bar{f}_{\text{red}}(t) &
\end{array}
\qquad (4.3')
$$

J: Jordankästchen, z_{red}: reduzierter Vektor.

Jedes Subsystem kann mehrfach, auch mit verschiedener Dimension, auftreten.

Beweis: GANTMACHER [47], Matrizentheorie, (II) Kapitel 12: Singuläre Matrizenbüschel (singular matrix pencils). ∎

Das Gesamtsystem (4.3') heißt

Kronecker'sche Normalform
(Kronecker canonical form = KCF)

Untersuchung der Subsysteme:

(4.3′.a) : $\bar{f}_i(t) \equiv 0\,, \quad i = 1, \ldots, n_0$ (4.4.a)
notwendig und hinreichend für Widerspruchsfreiheit
$\hookrightarrow z_1, \ldots, z_{n_0}$ beliebige Funktionen

(4.3′.b) : stets lösbares System: z_{n_1} beliebige Funktion
$\hookrightarrow z_{n_1-1}, \ldots, z_{n_0+1}$ durch Quadratur

(4.3′.c) : bedingt lösbares System:

$$z_{n_2} = \bar{f}_{n_2+1}(t)$$
$$z_{n_2-1} = \bar{f}_{n_2}(t) - \bar{f}'_{n_2+1}(t)$$
$$\vdots$$
$$z_{n_1+1} = \bar{f}_{n_1+2}(t) - \bar{f}_{n_1+3}(t) \pm \ldots$$
$$+(-1)^{n_2-n_1-1} \bar{f}^{(n_2-n_1-1)}_{n_2+1}(t)$$

Widerspruchsfreiheit verlangt:

$$\bar{f}_{n_1+1} - \bar{f}'_{n_1+2}(t) + \bar{f}''_{n_1+3}(t) \pm \ldots + (-1)^{n_2-n_1} \bar{f}^{(n_2-n_1)}_{n_2+1}(t) \equiv 0 \qquad (4.4.b)$$

(4.3′.*d*): stets lösbares System:

$$\begin{aligned} z_{n_3} &= \bar{f}_{n_3}(t) \\ z_{n_3-1} &= \bar{f}_{n_3-1} - \bar{f}'_{n_3}(t) \\ &\vdots \\ z_{n_2+1} &= \bar{f}_{n_2+1}(t) - \bar{f}'_{n_2+2}(t) \pm \ldots + (-1)^{n_3-n_2-1} \bar{f}^{(n_3-n_2-1)}_{n_3}(t) \end{aligned}$$

(4.3′.e): Standardsystem

$$z_{\text{red}} = \exp(-Jt) z_{\text{red}}(0) + \int_0^t \exp(-J(t-s)) \bar{f}_{\text{red}}(s) ds$$

Zusammenfassung für System (4.2):

1. *Existenz* von Lösungen verlangt Bedingungen (4.4), die wiederum von Transformationsmatrizen P, Q abhängen
2. *Eindeutigkeit* der Lösung verlangt, bei gegebenem $z_0 = Q^{-1} y_0$:

$$
\begin{aligned}
&n_0 = 0 \\
&n_1 = 0 \\
&\text{falls } n_2 > 0 \ : \ z_0 \textit{ konsistent } \text{mit} \\
&\qquad\qquad \text{rechter Seite } \bar{f} + (4.4.b) \\
&\text{falls } n_3 > 0 \ : \ z_0 \textit{ konsistent } \text{mit} \\
&\qquad\qquad \text{rechter Seite } \bar{f}
\end{aligned}
$$

Falls (4.4.b) erfüllt ist, genügt Betrachtung des Falles (4.3'.d) und (4.3'.e). Für numerische Behandlung von Systemen (4.2) muß also nach Transformation gelten:

$$u' + Cu = \tilde{f}(t) \tag{4.5.a}$$

$$Ev' + v = g(t) \tag{4.5.b}$$

$z = (u, v)$ partitioniert , $v \in \mathbb{R}^{\nu}$

$$E = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{bmatrix} \qquad (\nu, \nu) - \text{ Matrix}$$

E hat die Eigenschaft:

$$\begin{aligned} &E^{\nu} = 0, \quad E^{\nu-1} \neq 0 \\ &\nu : \text{ Nilpotenz, Index} \end{aligned} \tag{4.5.c}$$

Explizite Form von (4.5.b):

$$
\begin{aligned}
v_1 &= g_1(t) \\
v_2 &= g_2(t) - v_1' = g_2(t) - g_1'(t) \\
&\vdots \\
v_{\nu} &= g_{\nu}(t) - v_{\nu-1}' = \\
&= g_{\nu}(t) - g_{\nu-1}'(t) \pm \ldots + (-1)^{\nu-1} g_1^{(\nu-1)}(t)
\end{aligned}
\tag{4.5'.b}
$$

Bei numerischer Integration gibt man die rechte Seite $g(t)$ sowie $v(0)$ vor. Für $\nu > 1$: Überprüfung der Konsistenz der Anfangswerte $v_2(0), \ldots, v_{\nu}(0)$ verlangt innerhalb der numerischen Integration die *numerische Differentiation der rechten Seite* g - was bekanntlich *schlechtkonditioniert* ist!

Fazit: Numerische Integration von allgemeinen impliziten Differentialgleichungssystemen vom Typ (4.5) nur für $\nu = 0, 1$ gut konditioniert.

Bemerkung: Erweiterung auf $\nu > 1$ im linearen Fall möglich, falls höhere Ableitungen von $g(t)$ explizit angegeben.

Matrizenbüschel:

$$\{E - \lambda I_\nu\} \text{ regulär für } \lambda \neq 0 \tag{4.6}$$

Rücktransformation mit P, Q:

$$\{B - \lambda A\} \text{ regulär für } \lambda \neq 0 \tag{4.6'}$$

Interpretation von ν: ν-malige Differentiation der rechten Seite liefert gewöhnliches implizites Differentialgleichungs-System (mit Index 0).

Allgemeiner nichtlinearer Fall (RHEINBOLDT 1985) [94]
Betrachtet wird der *separierte* Spezialfall:

$$\begin{aligned} \text{a)} \quad B(y)y' &= g(y) \\ \text{b)} \quad F(y) &= 0 \end{aligned} \tag{4.7}$$

$$\left.\begin{aligned} &F : D \subseteq \mathbb{R}^n \to \mathbb{R}^{m_2} \\ &B : (m_1, n) - \text{ Matrix} \\ &g : D \to \mathbb{R}^{m_1} \\ &F, B, g \in C^r(D), r \geq 2\,. \end{aligned}\right\} m_1 \leq n \leq m_1 + m_2$$

Die Abbildung $F = 0$ definiert eine Mannigfaltigkeit $\mathcal{M}$, so daß (4.7.b) äquivalent zu

$$y \in \mathcal{M} \tag{4.7'.b}$$

Richtungsfeld auf $\mathcal{M}$ durch Differentiation:

$$\begin{aligned} &F_y(y)v(y) = 0 \\ &v(y) \in \mathcal{T}_y\mathcal{M} \text{ (Tangentialbündel)} \\ &F_y : (m_2, n) - \text{ Matrix} \end{aligned} \tag{$*$}$$

Zusammenfassung von (4.7.a) und (*):

$$\begin{bmatrix} B(y) \\ F_y(y) \end{bmatrix} y' = \begin{bmatrix} g(y) \\ 0 \end{bmatrix} \tag{4.8}$$

Dies ist ein (im allgemeinen überbestimmtes) DG-System. Eindeutigkeitsbedingung:

$$\operatorname{rg}\begin{bmatrix} B \\ F_y \end{bmatrix} = \operatorname{rg}\begin{bmatrix} B, g \\ F_y, 0 \end{bmatrix} = n \tag{4.9}$$

Definition: Falls (4.9) automatisch erfüllt, ist (4.7) algebraisch vollständig. Falls (4.9) zusätzlich für Eindeutigkeit gefordert, ist (4.7) *algebraisch unvollständig*, das heißt zu den algebraischen Bedingungen (4.7.b) müssen weitere algebraische Bedingungen aus (4.9) hinzugenommen werden. Für $\nu = 1$ gilt (4.9) nach Definition automatisch, das heißt algebraisch unvollständig ist äquivalent zu $\nu > 1$.

Bemerkung: Verallgemeinerung der Beobachtung aus (4.5'.b). Der Prozeß der algebraischen Vervollständigung läuft gegebenenfalls über mehrere Stufen (für $\nu \geq 2$).

Übertragung auf den allgemeinen Fall (4.1') (DEUFLHARD, NOWAK 1987)[40]:

$$\begin{aligned} &\{B(y) - \lambda[f_y(y) - \Gamma(y, y')]\} \text{ regulär für } \lambda \neq 0 \\ &\Gamma(y, y')\delta y = (B_y(y)\delta y)y' \\ &\Gamma_{ik} := \sum_{j=1}^{n} \frac{\partial B_{ij}(y)}{\partial y_k} \cdot y'_j \end{aligned} \tag{4.10}$$

4.2 Anpassung steifer Integratoren

Anpassung von BDF-Verfahren

(GEAR 1971 , [48], PETZOLD 1981 , [91]).
Die Mehrschrittverfahren vom BDF-Typ eignen sich sogar für den allgemeinen *impliziten DG-Typ*:

$$F(t, y, y') = 0 \tag{4.11}$$

Rückgriff auf BDF-Darstellung (Kap. B.3.2) liefert:

$$\begin{aligned} &F\big(t_{n+k}, y_{n+k}, p'(t_{n+k})\big) = 0 \\ &p(t) := P_k(t|t_{n+k}, \ldots, t_n) \end{aligned} \tag{4.12}$$

Das Iterationsverfahren (3.17) für $F = 0$ nach (3.16) wird lediglich erweitert:

$\hookrightarrow$ Newton-ähnliches Verfahren für (4.11).

Programme: LSODI (HINDMARSH [64]), DASSL (PETZOLD [91])

Anpassung von semi-impliziten Extrapolationsverfahren

Eine genaue Analyse der Verfahren zeigt, daß sich im weiteren nur die semi-implizite Euler-Diskretisierung zur Erweiterung auf den Fall $\nu = 1$ eignet. Die Diskretisierung von (4.1') führt auf:

$$\begin{aligned} y_{k+1} &:= y_k + h\big(B(y_k) - hA\big)^{-1} f(y_k) \\ A &:= f_y(y_0) - \Gamma(y_0, y_0') \end{aligned} \tag{4.13}$$

Bemerkung:

1. $A = \dfrac{\partial}{\partial y}\big(f(y) - B(y)y'\big)\Big|_{y_0}$
2. Falls y_0' nicht gegeben:
   anfangs $A_0 := f_y(y_0)$, später y' durch Extrapolation aus $\left\{\dfrac{y_l - y_{l-1}}{h}\right\}$, $l = n_1, n_2, \ldots$ gewonnen.

Spezialfall:

$$\begin{aligned} &\text{a)} \quad y' = f(y,z), \quad y(0) = y_0 \\ &\text{b)} \quad \varepsilon z' = g(y,z), \quad z(0) = z_0 \end{aligned} \tag{4.14}$$

$\varepsilon \to 0^+$ liefert differentiell-algebraisches System.

Für den Fall $\nu = 1$ muß g_z *nichtsingulär* gelten.

Beweis: 1-mal differenzieren $(\varepsilon = 0)$.

$$g_z z' = -g_y(y,z) f(y,z)$$

■

Satz 1 (DEUFLHARD, HAIRER, ZUGCK 1987) [38]

Seien $f, g \in C^{N+1}(D)$ in (4.14). Anwendung der semi-impliziten Euler-Diskretisierung (4.13) auf System (4.14) mit $\varepsilon = 0$. Man erhält die gestörte asymptotische Entwicklung $(t_n = t = nh)$.

$$\begin{aligned} a) \quad y_n - y(t_n) &= b_1(t)h + (b_2(t) + \beta_n^2)h^2 + \ldots \\ &\quad \ldots + (b_N(t) + \beta_n^N)h^N + B(n,h)h^{N+1} \\ b) \quad z_n - z(t_n) &= (c_1(t) + \gamma_n^1)h + \ldots \\ &\quad \ldots + (c_N(t) + \gamma_n^N)h^{N+1} + C(n,h)h^{N+1} \end{aligned} \tag{4.15}$$

wobei gilt:

$$\begin{aligned} &a) \quad \beta_n^2 = 0\,, \ \beta_n^3 = 0\,, \ \beta_n^4 = 0\,, \ \gamma_n^1 = 0\,, && n \geq 0 \\ &b) \quad \gamma_n^2 = 0\,, \ \gamma_n^3 = 0\,, && n \geq 1 \\ &c) \quad \beta_n^{j+1} = 0\,, \ \gamma_n^j = 0\,, && n \geq j-2 \ \textit{und} \ j \geq 4 \end{aligned} \tag{4.16}$$

Beweis: [38] ■

Die Funktionen $b_j(t), c_j(t)$ repräsentieren glatte Fehleranteile, die bei h-Extrapolation verschwinden. Die Funktionen β_n^j , γ_n^j repräsentieren nichtglatte Fehleranteile, die bei Extrapolation nur mit Vorfaktoren multipliziert werden, ohne zu verschwinden.

Folgerung:

$$\begin{aligned} &\text{Falls} \quad n_1 = 1 \;:\; \text{Ordnungsbeschränkung} \\ &\qquad\qquad\quad k_{\max} = 4 \\ &\text{Falls} \quad n_1 = 2 \;:\; k_{\max} = 5 \;. \\ &\vdots \end{aligned}$$

Programm: LIMEX (DEUFLHARD, NOWAK 1985 [40])

Bemerkung: keine Verbesserung durch *implizite* Euler-Diskretisierung!

Anpassung von IRK-Verfahren

Beschränkung wie in Kapitel B.2.4 auf Verfahren vom Typ (2.58):

$$a_{sj} = b_j \quad j = 1, \ldots, s$$

Anwendung auf System (4.14) (DEUFLHARD, HAIRER, ZUGCK) [38]:

$$\begin{aligned} &\text{a)} \quad Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} f(Y_j, Z_j) \;, \quad i = 1, \ldots, s \\ &\text{b)} \quad g(Y_i, Z_i) = 0 \;, \qquad\qquad i = 1, \ldots, s \\ &\text{c)} \quad y_1 = Y_s \;, \quad z_1 = Z_s \end{aligned} \tag{4.17}$$

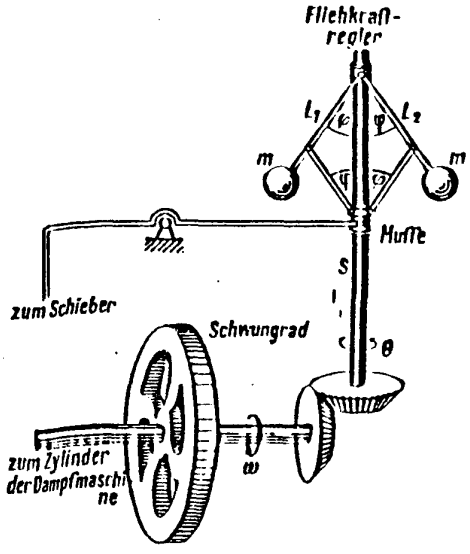
Anwendung auf allgemeines System (4.1') (HAIRER 1988) [57]:

$$\begin{aligned} &\text{a)} \quad Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} U_j, \quad i = 1, \ldots, s \\ &\text{b)} \quad B(Y_i)\; U_i = f(Y_i) \end{aligned}$$

Da $\mathcal{A}^{-1}$ existiert (vergleiche Kapitel 2.4), ist Elimination von U_i mittels a) möglich. Zur Lösung der nichtlinearen Gleichung b) erforderliches Newton-Verfahren erfordert Vorsicht. Programm: RADAU5 (HAIRER) [57]

5 Aufgaben

21.) (Stabilität des Fliehkraftreglers bei Dampfmaschinen, nach *Wyschnegradski* 1877/78)



Nebenstehend skizziert ist das mechanische System *Maschine-Regler* einer Dampfmaschine. Der Fliehkraftregler steuert in Abhängigkeit von der Drehzahl der Maschine die Dampfzufuhr.
Die Dynamik dieses Systems wird mathematisch beschrieben durch das DG-System

$$\begin{aligned} \dot{\varphi} &= \psi \\ \dot{\psi} &= (\theta^2 \cos\varphi - g)\sin\varphi - \frac{b}{m}\psi \\ \dot{\omega} &= (k\cos\varphi - F)/J \end{aligned}$$

wobei die Massen m, der Winkel φ und die Winkelgeschwindigkeiten θ, ω gemäß Skizze zu interpretieren sind.
Ferner seien die folgenden Konstanten gegeben:

g : Erdbeschleunigung
b : Reibungskoeffizient
k : positiver Proportionalitätsfaktor
F : äußere Belastung
J : Trägheitsmoment des Schwungrades

Da Regler und Maschine über ein Zahnrad verbunden sind, gilt zusätzlich:

$\theta = n\omega$
n : Übersetzungsverhältnis des Zahnrades

Der mechanische Regler soll den "Gleichlauf" $\omega = \omega_0$ des Schwungrades sichern (d.h. Drosselung der Dampfzufuhr, falls $\omega > \omega_0$, sowie Erhöhung der Dampfzufuhr, falls $\omega < \omega_0$).
Im Gleichlaufpunkt sei demnach

$$(*) \qquad \varphi = \varphi_0\,,\ \psi = 0\,,\ \omega = \omega_0.$$

a.) Man bestimme φ_0 , ω_0.

b.) Mit den Bezeichnungen

$$\delta\varphi := \varphi - \varphi_0 \ , \ \delta\psi := \psi - \psi_0 \ , \ \delta\omega := \omega - \omega_0$$

leite man die zugehörige Variationsgleichung her.

c.) Man zeige:
Die Gleichgewichtslösung (*) ist stabil (im Sinne von Ljapunov) genau dann, wenn gilt:

$$\frac{bJ}{m} > \frac{2F}{\omega_0}$$

Hinweis: Seien λ_1, λ_2, λ_3 Nullstellen des Polynoms

$$p(\lambda) = \lambda^3 + a\lambda^2 + b\lambda + c \ ,$$

so gilt nach *Routh-Hurwitz*:

$$a > 0 \ , \ b > 0 \ , \ c > 0 \ , \ ab > c \Leftrightarrow \Re(\lambda_i) < 0 \ , \ i = 1, 2, 3.$$

(Nicht zu beweisen.)

22.) Man integriere mit dem Verfahren RKX4 aus Aufgabe 14 die Differentialgleichung

$$y'(t) = \lambda(y(t) - g(t)) + g'(t) \ , \ y(0) = \epsilon_0 + g(0) \ , \ t \in [0,5],$$

für $g(t) = t^2 + 1$ und alle Kombinationen von $\epsilon_0 = 0, 1, 10, 100,$ und $\lambda = -100, -50, -10, -1, 1, 10,$ und gebe jeweils die Anzahl der Funktionsaufrufe an. Für ϵ wähle man einen dem verwendeten Rechner und dem eigenen Programm angepaßten Wert ($\epsilon = 10^{-6} - 10^{-4}$).

23) Gegeben sei eine $n \times n$-Matrix A, die für ein gegebenes Skalarprodukt $< \cdot, \cdot >$ der folgenden Ungleichung genüge:

$$< x, Ax > \ \leq \ \mu < x, x > \ \equiv \ \mu \parallel x \parallel^2 \ , \ \forall x \in R^n \ ,$$

mit $\parallel \cdot \parallel^2 = < \cdot, \cdot >$.
Man zeige: Für $\mu h < 1$ gilt in der zugehörigen Matrixnorm:

a) $\parallel (I - hA)^{-1} \parallel \leq \frac{1}{1 - \mu h}$,

b) $\parallel (I - hA)^{-1}(I + hA) \parallel \leq \frac{1 + \mu h}{1 - \mu h}$, falls zusätzlich $\mu h \geq 0$.

24) Man zeige: Auch für eine nicht-diagonalisierbare Matrix A läßt sich das DG-System:

$$y' = Ay, \quad A \text{ constant},$$

so transformieren, daß zur Stabilitätsuntersuchung die Betrachtung der skalaren Testgleichung

$$y' = \lambda y, \quad y(0) = y_0 \, , \; \lambda \in C \, ,$$

genügt.

25) Sei A eine normale $n \times n$-Matrix , d.h. es gilt $A^H A = A A^H$, und seien $\lambda_1, \lambda_2, \cdots, \lambda_n$ die Eigenwerte von A. Für den Wertebereich

$$G(A) := \left\{ \frac{x^H A x}{x^H x} \mid x \neq 0 \right\}$$

von A zeige man:

a) $G(A)$ ist die konvexe Hülle der Eigenwerte, also:

$$G(A) = \left\{ \mu \mid \mu = \sum_{i=1}^{n} \tau_i \lambda_i, \; \tau_i \geq 0, \; \sum_{i=1}^{n} \tau_i = 1 \right\}$$

b) Für reelles A und alle $x \in R^n$ gilt:

$$x^T A \, x \leq \mu \, x^T x \, ,$$

mit $\mu := \max\{ \Re(y) \mid y \in G(A) \}$.

26.) Gegeben sei für $u(x,t) \in R$ die partielle DG:

$$\frac{\partial u(x,t)}{\partial t} = \sigma \frac{\partial^2 u(x,t)}{\partial x^2} \, , \; x \in [0,\pi] \, , \; t \geq 0 \, , \sigma > 0 \, ,$$

mit den Randbedingungen:

$$\begin{aligned} u(x,0) &= \varphi(x) \, , \; x \in [0,\pi] \, , \; \varphi(0) = \varphi(\pi) = 0 \, , \; (\varphi \text{ gegeben}) \, , \\ u(0,t) &= u(\pi,t) = 0 \, , \; t \geq 0 \, . \end{aligned}$$

Für festes N ersetze man auf dem Gitter $\{ j\Delta x \mid \Delta x = \frac{\pi}{N} \, , \; j = 0, \ldots, N \}$ in der rechten Seite der Differentialgleichung die Ableitung bezüglich x durch einen symmetrischen Differenzenquotienten und zeige, daß das resultierende DG-System für

$$U(t) := (U_1(t), \ldots, U_{n-1}(t))^T$$

kontraktiv ist.
Dabei ist $U_j(t)$ eine Näherung für $u(j\Delta x, t)$, $j = 1, \ldots, N-1$.

27.) a.) Man beweise die A-Stabilität der impliziten Trapezregel:

$$y_{k+1} = y_k + \frac{h}{2}\left(f(y_k) + f(y_{k+1})\right)\,,\ k = 0,\,1,\ldots$$

b.) Die impl. Trapezregel besitzt eine h^2-Entwicklung. Man extrapoliere einmal mit $n_1 = 1, n_2 = 2$, und zeige, daß für T_{22} die A-Stabilität verlorengeht.

c.) Man extrapoliere mit $n_1 = 2, n_2 = 4$ und zeige, daß bei Anwendung auf die skalare Testgleichung gilt:

$$\lim_{z\to\infty} \mid T_{22}(z) \mid = 1\,,\ z := \lambda\, h.$$

28.) Gegeben sei für $y(t) \in \Re$ die DG zweiter Ordnung:

$$y''(t) + \mu^2\, y(t) = f(y(t))\,,\ y(0) = y_0\,,\ y'(0) = z_0\,,$$

mit $\mu > 0$, μ konstant.

a.) Man untersuche das qualitative Verhalten von $y(t)$ für $\| f_y(y) \| \ll \mu^2$.
In welcher Größenordnung würde eine von einem numerischen Integrator vorgeschlagene Schrittweite liegen ?

b.) Zur Vorbereitung für die numerische Behandlung sei definiert :

$$\begin{aligned} g(t) &:= \frac{\sin(\mu t)}{\mu}\,, \\ \bar{y}(t) &:= \frac{1}{T}\int_{t-T/2}^{t+T/2} y(\tau)\,d\tau\,,\ T := \frac{2\pi}{\mu}\,. \end{aligned}$$

Man zeige:

(i) $y(t) = y_0\cos(\mu t) + z_0\, g(t) + \int_{s=0}^{t} f(y(s))\, g(t-s)\, ds$

(ii) $\bar{y}(t) = \frac{2}{\mu^2 T}\int_{t-T/2}^{t+T/2} f(y(s))\,\cos^2(\frac{\mu}{2}(t-s))\, ds$

(iii) Unter der Voraussetzung:

$$\lim_{\mu\to\infty} \frac{f(y(s))}{\mu^2} = \alpha$$

gilt:

$$\lim_{\mu\to\infty} \bar{y}(t) = \alpha.$$

29.) Das skalare Modellproblem

$$y' = \lambda\,(y - g(t)) + g'(t)\ ,\ y(0) = g(0)\ ,$$

hat für $\Re(\lambda) < 0$ die asymptotische Lösung $y(t) \equiv g(t)$
(vgl. Kap B.1).

a.) Der Ansatz

$$\frac{E(-h\,A)\,y(t+h) - y(t)}{h} = \bar{f}(t, y(t)) := f(t, y(t)) - A\,y(t)\ ,\ A := f_y(y_0)\ ,$$

zur Lösung einer DG $y' = f(t, y(t))$, $y(0) = y_0$, stellt ein Verfahren vom Typ des semi-impliziten Euler-Verfahrens dar, wobei aus Konsistenzgründen gelten muß:

$$E(0) = I\ ,\ \left.\frac{dE}{dh}\right|_{h=0} = A.$$

Sei $g(t) := g_0$, $g_0 \in \Re$. Für welche Wahl von $E(h\,A)$ ist bei Anwendung des Verfahrens (mit $A := \lambda$) auf das Modellproblem $y(t) \equiv g(t)$ stationäre Lösung auch der entstehenden Differenzengleichung ?

b.) Man führe die analoge Rechnung für den Ansatz

$$\frac{E(-h\,A)\,y(t+h) - E(h\,A)\,y(t-h)}{2h} = \bar{f}(t, y(t))$$

durch, wobei jetzt $g(t) := g_0 + g_1 t$, $g_0, g_1 \in \Re$.
Schränkt ein Startschritt gemäß Teil a.) die Wahl von $g(t)$ wieder auf $g(t) = g_0$ ein ?

30.) Untersuchung der Stabilität des semi-impliziten Euler-Verfahrens und der semi-impliziten Mittelpunktsregel:

a.) Das semi-implizite Euler-Verfahren lautet:

$$y_{k+1} = y_k + (I - h\,A)^{-1}\,h\,f(y_k)\ ,\ k = 0,\,1,\ldots$$

Man wende das Verfahren auf die Testgleichung

$$y' = \lambda\,y\ ,\ y(0) = y_0\ ,$$

an und setze : $z := \lambda\,h$, $z_0 := A\,h$, $\Re(z_0) \leq 0$.

(i) Man zeige die A-Stabilität des Verfahrens für $z = z_0$.

(ii) Für $z_0 \neq z$ berechne man den Rand ∂G des Stabilitätsgebietes

$$G(z_0) = \{z \in C \mid \mid R(z, z_0) \mid \leq 1\}$$

b.) Man wende die semi-implizite Mittelpunktsregel (ohne Start- und Schluß-schritt)

$$y_{k+1} = (I - h\,A)^{-1}[(I + h\,A)\,y_{k-1} + 2h\,\bar{f}(y_k)]$$

auf die skalare Testgleichung an und setze z, z_0 wie in a.).

(i) Man zeige die A-Stabilität des Verfahrens für $z = z_0$.

(ii) Für $z_0 \neq z$ leite man mit dem Ansatz $y_k = \zeta^k$ die zugehörige charakteristische Gleichung her. Seien $\zeta_1(z, z_0)$ und $\zeta_2(z, z_0)$ die Wurzeln dieser Gleichung. Man gebe den Rand ∂G des Stabilitätsgebietes

$$G(z_0) = \{z \in C \mid \mid \zeta_i(z, z_0) \mid \leq 1\ ,\ i = 1, 2\}$$

an und untersuche die Spezialfälle:

α) z_0 reell, $z_0 \to \infty$.

β) z_0 nähert sich der imaginären Achse.

31.) Anwendung der semi-impliziten Mittelpunktsregel mit Start- und Schlußschritt auf die skalare Testgleichung

$$y' = \lambda\,y\ ,\ y(0) = 1\ ,$$

liefert für die erste Spalte eines Extrapolationstableaus: $(z := \lambda\,h)$

$$R_{l,1}(z) = \frac{1}{(1 - z/n_l)^2}\left(\frac{1 + z/n_l}{1 - z/n_l}\right)^{\frac{n_l}{2}-1}\ ,\ n_l = 2, 6, 10, 14, 22, \ldots,$$
$$l = 1, 2, \ldots$$

Man berechne $R(z) := R_{2,2}(z)$ und weise hierfür die A-Stabilität nach.

Hinweis: Für $R(z) = \frac{P(z)}{Q(z)}$, P, Q Polynome, gilt:

$$E(i\,y) := \mid Q(i\,y) \mid^2 - \mid P(i\,y) \mid^2 \geq 0\ ,\ y \in R \Longrightarrow \mid R(iy) \mid \leq 1\ , y \in R.$$

($E(i\,y)$ heißt *Ehle*-Polynom.)

32.) a.) Man zeige, daß sich für das i.a. nicht-äquidistante Gitter

$$\{t_n,\ t_n + h\,\gamma,\ t_{n+1}\},\ h := t_{n+1} - t_n\ ,\ \gamma \in]0, 1[\ ,$$

das folgende BDF2-Verfahren ergibt:

$$y_{n+1} + \frac{1}{\gamma(\gamma - 2)}y_{n+\gamma} - \frac{(\gamma - 1)^2}{\gamma(\gamma - 2)}y_n = \frac{\gamma - 1}{\gamma - 2}\,h\,f(y_{n+1})\ .$$

b.) Das zusammengesetzte Verfahren TR-BDF2 bestehe aus einem Schritt mit der impliziten Trapezregel von t_n nach $t_n + h\gamma$ und einem anschließenden Schritt nach t_{n+1} mit dem in a.) hergeleiteten Verfahren.
Man berechne den führenden Fehlerterm dieses Verfahrens und optimiere ihn bzgl. γ.

c.) Bei beiden Schritten von TR-BDF2 ist ein nichtlineares Gleichungssystem der Form $F_{TR}(y_{n+\gamma}) = 0$ bzw. $F_{BDF2}(y_{n+1}) = 0$ zu lösen. F_{TR} und F_{BDF2} ergeben sich aus den Verfahren.
Man berechne die Jacobi-Matrix von F_{TR} und F_{BDF2} bzgl. $y_{n+\gamma}$ bzw. y_{n+1} und überlege sich, für welches $\gamma \in]0,1[$ die Jacobi-Matrizen die gleiche Gestalt haben (unter der Voraussetzung $f_y(y_{n+\gamma}) \approx f_y(y_{n+1})$).

33.) Bei Problemen der Parameteridentifizierung in DG-Systemen ergibt sich die Aufgabe, zusätzlich zur Integration einer parameterabhängigen DG :

$$y'(t;p) = f(y(t;p),p)\ ,\ y(0;p) = y_0\ ,\ p \in R\ ,\ t \in [0,T]\ ,$$

auch den Wert von $y_p(T;p) := \frac{\partial y(T;p)}{\partial p}$ zu bestimmen. Dabei gibt es zwei Möglichkeiten:

a.) Man löst (numerisch) die zugehörige Variationsgleichung

$$y'_p(t;p) = f_y(y(t;p),p)\,y_p(t;p) + f_p(y(t;p),p)\ ,\ y_p(0,p) = 0\ ,\ t \in [0,T]\ ,$$

und erhält damit eine Näherung für $y_p(T;p)$.

b.) Man berechnet numerisch Näherungen $\eta(p)$ und $\eta(p+\Delta p)$ für $y(T;p)$ und $y(T;p+\Delta p)$ und bildet den Differenzquotienten

$$\frac{\eta(p+\Delta p) - \eta(p)}{\Delta p}.$$

Man überlege sich Vorgehensweise, sowie Vor- und Nachteile bei beiden Möglichkeiten anhand des semi-impliziten Euler-Verfahrens (EULSIM). Dabei beachte man die Aspekte: Implementierung, Speicherplatz, Fehlerkontrolle.

34.) Folgende Gleichungen beschreiben ein Pendel:
(y_1 , y_2 : Abstand vom Drehpunkt, y_5 : Federspannung)

$$\begin{aligned} y'_1 &= y_3 \\ y'_2 &= y_4 \\ y'_3 &= -y_1\,y_5 \\ y'_4 &= -y_2\,y_5 + 1 \\ y_1^2 + y_2^2 &= 1 \end{aligned}$$

Man zeige, daß man gemäß (4.9) der Vorlesung die Bedingungen

$$\begin{aligned} y_1\, y_3 + y_2\, y_4 &= 0\,, \\ y_3^2 + y_4^2 + y_2 - y_5 &= 0\,, \end{aligned}$$

hinzufügen muß, damit man ein algebraisch vollständiges System erhält.

Man gebe den Index ν des urspünglichen und des vervollständigten Systems an.

35.) Zur Bestimmung des Index und zur Untersuchung der Konsistenz der Anfangswerte wird im Programm LIMEX ein sogenannter *Indexmonitor* benutzt.
Sei $y_1(h) - y_0 = \Delta_0(h)$ der Startschritt des modifizierten semi-impliziten Eulerverfahrens angewandt auf ein System der Form

$E\,y' = f(y)\,,\ y(0) = y_0\,,$ ($E\,(\nu\times\nu)$ −Matrix wie in der Vorlesung (1.5.c).)

Man zeige:
Für $h_{j+1} < h_j\,,\ j = 0, 1, \ldots,$ gilt:

$$\frac{\| \Delta_0(h_{j+1}) \|}{\| \Delta_0(h_j) \|} \doteq \left(\frac{h_{j+1}}{h_j}\right)^p\,,$$

wobei:

$p = 1$: $\nu = 1\,,$ y_0 konsistent
$p = 0$: $\nu = 1\,,$ y_0 nicht konsistent, oder
$\nu = 2\,,$ y_0 konsistent
$p < 0$: $\nu \geq 2$

Warum sind für eine solche Abfrage Extrapolationsverfahren besonders gut geeignet?

C. Implementierung und Vergleich numerischer Integratoren

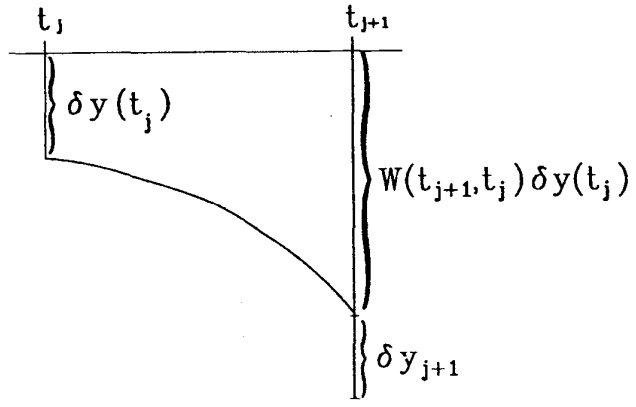
1 Genauigkeit und Skalierung

1.1 Lokale/globale Genauigkeit

TOL: lokal verlangte Genauigkeit (local error **tol**erance)

ERR: global erzielte Genauigkeit (global **err**or)

Zusammenhang ERR/TOL: Seien $[0,T]$ Integrationsintervall und $0 = t_0 < \ldots < t_m = T$ die vom Integrator automatisch gewählten Zwischenpunkte. Sei δy_j in t_j entstandener lokaler *absoluter* Diskretisierungsfehler und sei $\delta y(t_j)$ der *gesamte* Fehler im Punkt t_j. Dieser Fehler wird sich mittels der Propagationsmatrizen fortpflanzen:



Linearisierte Näherung:

$$\delta y(t_{j+1}) \doteq \delta y_{j+1} + W(t_{j+1}, t_j)\ \delta y(t_j) \tag{1.1}$$

Rekursiv erhält man mit $\delta y_0 = 0$ und der Gruppeneigenschaft (1.18.a) (vergleiche Kapitel A.1.2).

$$\delta y(T) \doteq \sum_{j=1}^{m} W(T, t_j) \delta y_j \tag{1.1'}$$

Ist $\| \cdot \|$ gerade diejenige Norm, für die die lokale Fehlerabfrage durchgeführt wird, dann gilt:

$$\| \delta y_j \| \leq \text{TOL} \quad , \quad \| \delta y(T) \| = \text{ERR} \tag{1.2}$$

In (1.1′) eingesetzt:

$$\text{ERR} \dot{\leq} \text{TOL} \sum_{j=1}^{m} \underbrace{\| W(T, t_j) \|}_{\text{problemabhängiger Verstärkungsfaktor}} \tag{1.3}$$

Genauere Untersuchung an einem Modellproblem:

Annahmen:
- globale Lipschitz-Konstante $L \equiv L_1$ über [0,T] liefert vernünftige Beschreibung
- äquidistantes Gitter $t_j = j \cdot h.$
- fixe Ordnung $p.$

Nach (1.15 – 1.17), Kapitel A.1.2 gilt:

$$\begin{aligned} \| W(T, t_j) \| &\leq \exp\left(L(T - t_j)\right), \text{ somit} \\ \sum_{j=1}^{m} \| W(T, t_j) \| &\leq \sum_{j=1}^{m} e^{Lh(m-j)} = \frac{e^{Lmh} - 1}{e^{Lh} - 1} = \\ &= \frac{T}{h} \frac{\varphi(LT)}{\varphi(Lh)} \leq m\varphi(LT), \quad \text{da} \quad \varphi(Lh) \geq 1. \end{aligned}$$

Somit:

$$\text{ERR} \leq m \, \text{TOL} \, \varphi(LT) \tag{1.4}$$

Man beachte, daß m von TOL abhängt und daher keine Linearität zwischen ERR und TOL vorzuliegen braucht.
Wegen der fixen Ordnung gilt:

$$\begin{aligned} \text{TOL} &\doteq C_p h^{p+1} \\ h &\doteq \sqrt[p+1]{\frac{\text{TOL}}{C_p}} \end{aligned}$$

Setzt man dies mit $m = T/h$ in (1.4) ein, erhält man

$$\text{ERR} \doteq C_p h^p T \, \varphi(LT). \tag{1.4′}$$

Die Rundungsfehler führen auf den zusätzlichen Fehler

$$\overline{\mathrm{ERR}} \doteq \gamma_p \cdot m \cdot eps = \frac{\gamma_p epsT}{h}$$

(eps relative Maschinengenauigkeit, (1.5)

γ_p abhängig von Arithmetik und Verfahren)

Jetzt soll h so optimiert werden, daß

$$\mathrm{ERR}_{\mathrm{ges}} = \overline{\mathrm{ERR}} + \mathrm{ERR} = \min!$$

Da $\mathrm{ERR}_{\mathrm{ges}}$ für $h \to 0$ und $h \to \infty$ unbeschränkt wächst, erhält man das Minimum aus

$$O = \frac{d}{dh}\mathrm{ERR}_{\mathrm{ges}} = p\,h^{p-1} C_p T \varphi(LT) - \gamma\ eps\ T/h^2, \text{ also}$$

$$h_{\mathrm{opt}} = \sqrt[p+1]{\frac{\gamma\ eps}{pC_p\ \varphi(LT)}} \tag{1.6.a}$$

$$\mathrm{ERR}_{\min} = \frac{p+1}{p}\gamma\ eps\ T/h_{\mathrm{opt}} \tag{1.6.b}$$

Für $\mathrm{ERR}_{\min} = \min$ muß $h_{\mathrm{opt}} = \max$ gelten. Bei tatsächlicher Integration bedeutet das:

p variabel, so daß h *lokal* maximal.

Fazit: Bestmögliche globale Genauigkeit erzielbar, falls p variabel und möglichst wenige Schritte. Das ergibt einen Konstruktionsvorteil für Extrapolationsverfahren - vergleiche Kapitel 3. und 4..

1.2 Skalierung

Affinkovarianz der Diskretisierung (vergleiche Kapitel A.1.3) wird im allgemeinen durch Schrittweitensteuerung *global* zerstört, da hierbei Normen eingehen. Integrationsverlauf sollte aber zumindestens von der *Skalierung* der Differentialgleichung *unabhängig* sein.

↪ Zusätzliche Vorsorge für *Skalierungsinvarianz* : Man geht über zu intern skalierten Größen

$$y_i \longrightarrow y_i/s_i \ , \ s_i > 0. \tag{1.7}$$

Invarianzforderung:

$$\begin{aligned} y_i &\longrightarrow \alpha_i y_i \quad \alpha_i > 0 \\ \Rightarrow\ s_i &\longrightarrow \alpha_i s_i \end{aligned} \tag{1.8}$$

Die *lokale* Fehlerabfrage muß jetzt skaliert durchgeführt werden. Dabei hängt die Genauigkeit und Effizienz stark von der internen Skalierung ab. Sie sollte deshalb möglichst optimal gewählt werden:
Bezeichnung:

$$D_j = \text{diag}\,(s_1(t_j), \ldots, s_n(t_j)). \tag{1.9}$$

Aus (1.1) wird in skalierter Form:

$$D_{j+1}^{-1}\delta y(t_{j+1}) = D_{j+1}^{-1}\delta y_{j+1} + D_{j+1}^{-1}W(t_{j+1}t_j)D_jD_j^{-1}\delta y(t_j) \tag{1.1''}$$

Mit den Bezeichnungen:

$$\left.\begin{array}{l} \delta\bar{y}(t_j) := D_j^{-1}\delta y(t_j) \\ \delta\bar{y}_j := D_j^{-1}\delta y_j \\ \| \delta\bar{y}(t_j) \| := \epsilon_j \\ \| D_{j+1}^{-1}W(t_{j+1},t_j)D_j \| =: \bar{\sigma}(t_{j+1},t_j) \\ \text{und der lokalen Fehlerabfrage} \\ \| \delta\bar{y}_j \| \leq \text{TOL} \end{array}\right\} \begin{array}{l} j = 1, \ldots, m \\ \text{(in Anlehnung an (Kap. A.1.2))} \end{array} \tag{1.10}$$

erhält man

$$\epsilon_{j+1} \leq \text{TOL} + \bar{\sigma}(t_{j+1}, t_j)\,\epsilon_j \tag{1.11}$$

Da $D_{j+1}^{-1}W(t_{j+1},\, t_j)D_j$ die Rolle der Propagationsmatrix übernimmt, ist:

$$\lim_{t_{j+1}\to t_j} D_{j+1}^{-1}W(t_{j+1},\, t_j)\,D_j = I \tag{1.12}$$

zu fordern, d.h. mit $h_j := t_{j+1} - t_j$

$$\lim_{h_j\to 0} \frac{s_i(t_j)}{s_i(t_{j+1})}(1 + \mathcal{O}(h_j)) = 1, \tag{1.13}$$

das heißt die s_i sind *stetig zu definieren*,

$$s_i(t) \in C^\circ \quad i = 1, \ldots, n. \tag{1.13'}$$

Bemerkung: Häufig falsch in Software, z.B. für $s_{\min} < 1$:

$$\text{falls } |\, y \,| < s_{\min} : s = 1.$$

Als nächstes ist $\bar{\sigma}(t_{j+1},\, t_j)$ abzuschätzen. Nach Satz 1.2, Kapitel B.1.1 gilt

$$\| W(t_{j+1},\, t_j) \|_s \doteq e^{\mu h_j}, \tag{1.14}$$

wobei $\| \cdot \|_s$ die skalierte Norm ist, bezüglich derer μ berechnet wird.

Bemerkung: Die Definition (1.14) verlangt ein μ, das zum Teilintervall $[t_j,\ t_{j+1}]$ gehört. In der algorithmischen Realisierung wird man auf die Schätztechnik $\mu \to [\mu]$ gemäß Kapitel B.2.3, (2.47) zurückgreifen – mangels besserer Alternativen.

Weitere Abschätzung:

$$\begin{aligned} \bar{\sigma}(t_{j+1}, t_j) &\leq \| D_{j+1}^{-1} D_j \| \, \| D_j^{-1} W(t_{j+1}, t_j) D_j \| \\ &= \| D_{j+1}^{-1} D_j \| \, \| W(t_{j+1}, t_j \|_s \\ &\dot{\leq} \| D_{j+1}^{-1} D_j \| \, e^{\mu h_j} \end{aligned} \tag{1.15}$$

Für jede monotone Norm gilt:

$$\| D_{j+1}^{-1} D_j \| = \max_i \frac{s_i(t_j)}{s_i(t_{j+1})} \quad (s_i > 0!), \tag{1.16}$$

somit

$$\epsilon_{j+1} \dot{\leq} \max_i \frac{s_i(t_j)}{s_i(t_{j+1})} e^{\mu h_j} \epsilon_j + \mathrm{TOL}. \tag{1.17}$$

Ist D_j gegeben, so enstehen bei der Wahl von D_{j+1} zwei gegenläufige Interessen. Einerseits möchte man keine Verstärkung des globalen Fehlers ϵ_j, das heißt

$$\bar{\sigma}(t_{j+1}, t_j) \leq 1, \tag{1.18}$$

Dies führt auf:

$$s_i(t_{j+1}) \geq s_i(t_j) \cdot e^{\mu h_j} \tag{1.18'}$$

Andererseits möchte man nicht genauer als bei rein relativem Fehlerkonzept rechnen. Dies führt auf:

$$s_i(t_{j+1}) \geq | y_i(t_{j+1}) | \tag{1.19}$$

Nimmt man noch eine Umgebung der Null von diesem Konzept aus, um aus Aufwandsgründen auf absoluten Fehler umzuschalten, so erhält man als *interne Skalierungsstrategie*

$$\begin{aligned} &s_i(t_{j+1}) := \max \left\{ s_i(t_j) e^{\mu h_j}, | y_i(t_{j+1}) |, s_i^{th} \right\} \\ &s_i^{th} = \text{“threshold”-Wert notfalls vom Benutzer zu wählen,} \\ &\text{möglichst skalierungsinvariant.} \end{aligned} \tag{1.20}$$

Vorschlag (falls Komponenten y_i vergleichbar, zum Beispiel Chemie):

$$s_i^{th} := \max_{i=1,\dots,n} \{ \mid y_i(0) \mid, \text{ epmach/TOL} \}$$

epmach : relative Maschinengenauigkeit

Mit (1.20) gilt dann (unabhängig von s_i^{th}) in etwa:

$$\epsilon_{j+1} \dot{\leq} \epsilon_j + \text{ TOL} \tag{1.21}$$

$$\epsilon_{\text{ges}} \dot{\leq} m \cdot \text{TOL} \tag{1.21'}$$

Interpretation von (1.20). Durch das Auftreten von μ findet eine Ausrichtung aller Komponenten auf die dominante Komponente statt: für diese wird möglichst der relative Fehler berechnet. (1.21) bedeutet zudem, daß lokale Genauigkeit an globale Genauigkeit angepaßt wird vermöge der Skalierung (1.20).

In den meisten öffentlich verfügbaren (public domain) Integratoren ist die Skalierungsfrage unbefriedigend behandelt. Folgende Varianten sind in Gebrauch:

$$s_i(t_{j+1}) := \max \{ s_i(t_j), \mid y(t_{j+1}) \mid, s_{\text{rel},i} \} \tag{1.22}$$

Diese Version (in etwa) findet sich in nahezu allen *steifen* Integratoren. Bei BDF-Verfahren (LSODE, LSODI, DASSL, etc.) findet sich zusätzlich eine vom Benutzer vorzugebende, also *externe* Skalierung, die zu folgender komponentenweisen Abfrage führt:

$$\begin{gathered} \mid \delta y_i \mid \leq s_{\text{rel},i} \text{ TOL} + s_{\text{abs},i} \\ (s_{\text{rel}} : \text{ RTOL}, \; s_{\text{abs}} : \text{ ATOL}) \end{gathered} \tag{1.23}$$

Bei *nichtsteifen* Integratoren wird im allgemeinen ein relatives Fehlerkonzept verwendet:

$$s_i(t_{j+1}) := \max \{ \mid y(t_{j+1}) \mid, s_{\text{rel},i} \} \tag{1.24}$$

Auch hier realisieren die Mehrschrittverfahren im wesentlichen die *externe* Skalierung (1.22). Die in (1.20) vorgeschlagene *interne* Skalierung vermeidet Fehler durch den Benutzer und erhöht die Effizienz der Codes gerade in realistischen Anwendungsproblemen. Sie ist bisher nur in steifen Extrapolationsintegratoren implementiert (Kapitel B.2.3).

Illustrationsbeispiele:

1. ZHABOTINSKY–BELOUSOV–REAKTION

Die Zhabotinski–Belousov–Reaktion zweier sich abwechselnder chemischer Prozesse führt auf das *allgemeine kinetische Schema* des OREGONATORS (FIELD/NOYES) (1974) [46].

$$\begin{aligned} BrO_3^- + Br^- &\rightarrow HBrO_2 \\ HBrO_2 + Br^- &\rightarrow P \\ BrO_3^- + HBrO_2 &\rightarrow 2HBrO_2 + Ce(IV) \\ 2HBrO_2 &\rightarrow P \\ Ce(IV) &\rightarrow Br^- \end{aligned}$$

Dieses Schema wird durch folgendes Differentialgleichungssystem beschrieben:

$$\begin{aligned} y_1' &= -c_1 y_1 y_2 - c_3 y_1 y_3 \\ y_2' &= -c_1 y_1 y_2 - c_2 y_3 y_2 + c_5 y_5 \\ y_3' &= c_1 y_1 y_2 - c_2 y_3 y_2 + c_3 y_1 y_3 - 2c_4 y_3^2 \\ y_4' &= c_2 y_3 y_2 + c_4 y_3^2 \\ y_5' &= c_3 y_1 y_3 - c_5 y_5 \end{aligned}$$

Dabei entsprechen sich

$$\begin{aligned} y_1 &\text{ und } BrO_3^- \\ y_2 &\text{ und } Br^- \\ y_3 &\text{ und } HBrO_2 \\ y_4 &\text{ und } P \\ y_5 &\text{ und } Ce(IV)\,. \end{aligned}$$

Dieses System besitzt eine oszillatorische Lösung und Übergänge von steifem zu nichtsteifem Verhalten in scharfen "Spitzen". Es stellt sich als besonders anfällig gegen ungeeignete Skalierungen heraus, so daß es zum herausforderndem Testbeispiel für Skalierungsstrategien wird. Um die Wirkung der Skalierung (1.20) zu verstehen, wurde zusätzlich zu der interessanten Lösungskomponente $HBrO_2$ die zugehörige Skalierung $s_i(t_j)$ (gepunktet.....) und der gemäß (1.21) globale Fehler

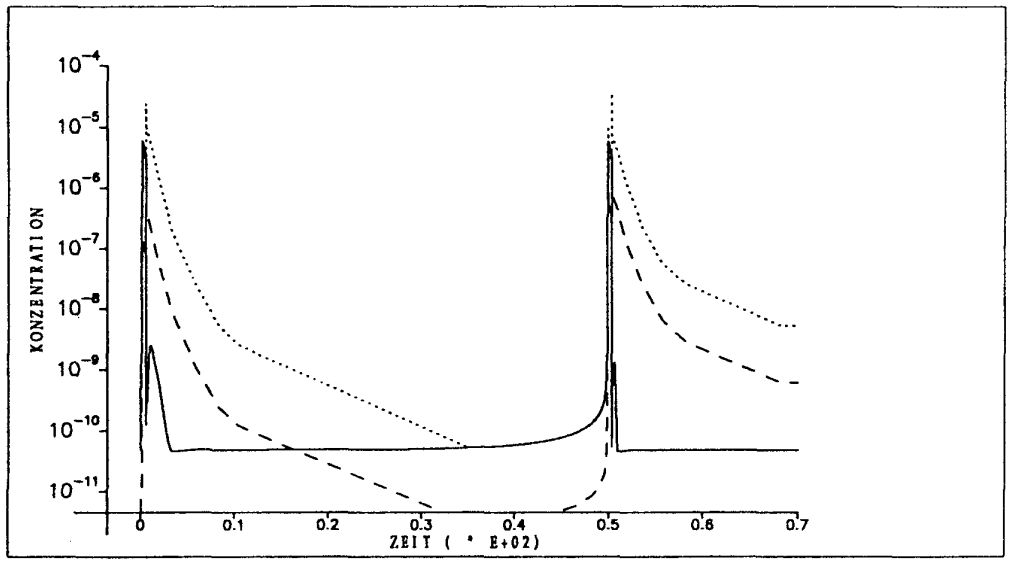
$$\delta y(t_j) \doteq j \cdot TOL \cdot s_i(t_j) \tag{1.25}$$

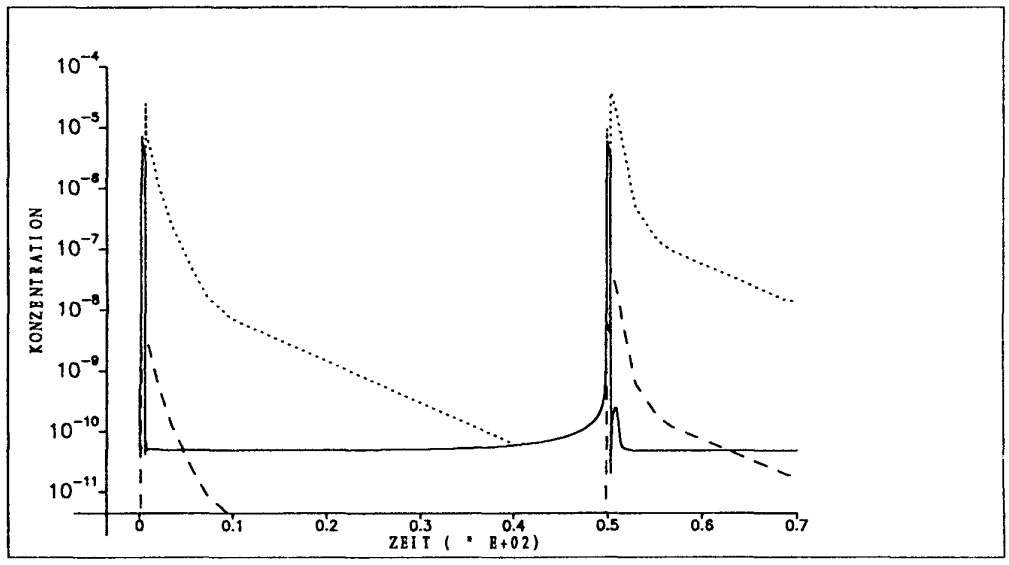
(gestrichelt -----) eingezeichnet.

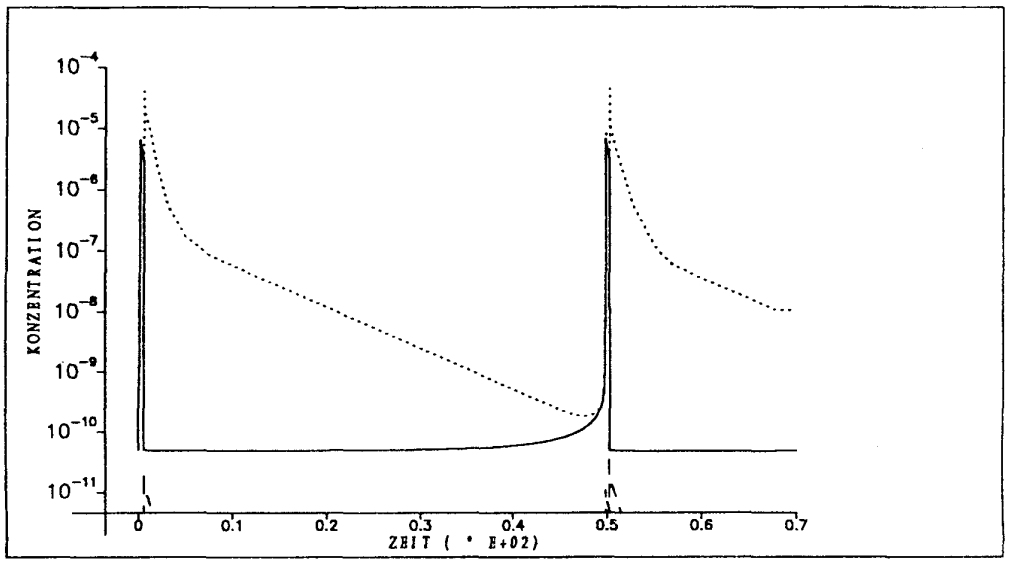
Liegt der globale Fehler über der Komponente, so besitzen wir für die t_j keine Information über diese Komponente, der qualitative Verlauf kann fehlerhaft sein (vergleiche Vergrößerung im Bild C.2). Andererseits ist die Differentialgleichung durch die Steifheit außerhalb der Flanken selbstkorrigierend, so daß die Skalierung mit negativem μ fällt, daß heißt der globale Fehler wieder aus der Komponente gedämpft wird, um so doch zu einem global richtigen Integrationsverlauf zu gelangen.

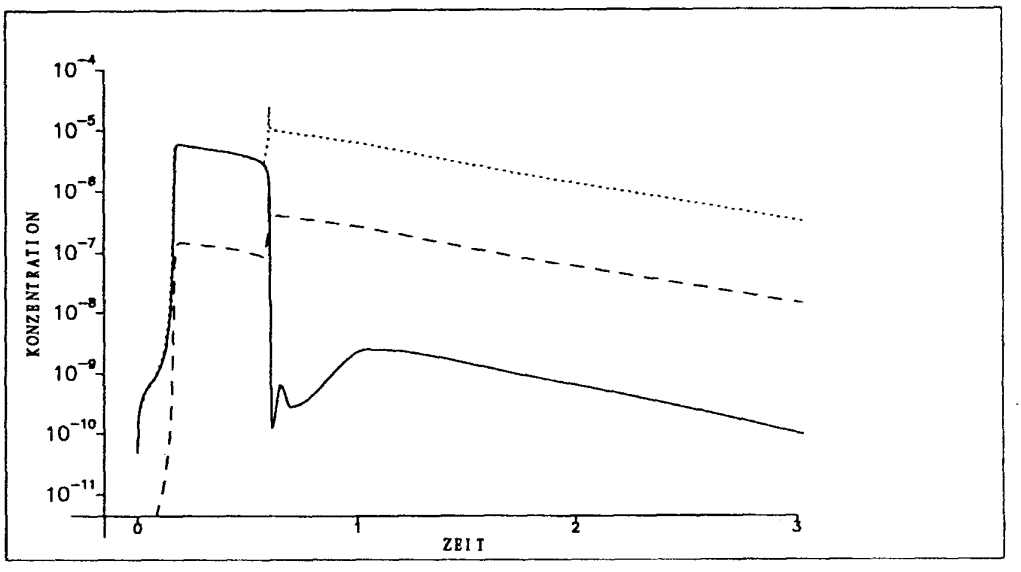
Die folgenden Bilder stellen die Resultate für TOL=10^{-3}, 10^{-5}, 10^{-8} untereinander vergleichsweise dar – mit Spreizung auf der anschließenden Seite.

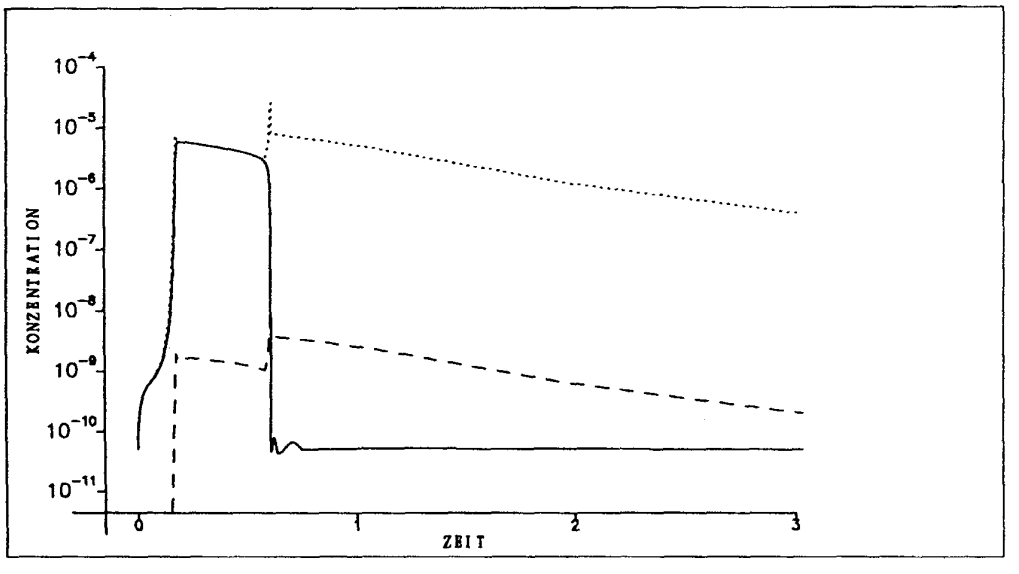
Bemerkung. In logarithmischer Skala läßt sich am Abstand des globalen Fehlers zur Komponente der relative Fehler ablesen.

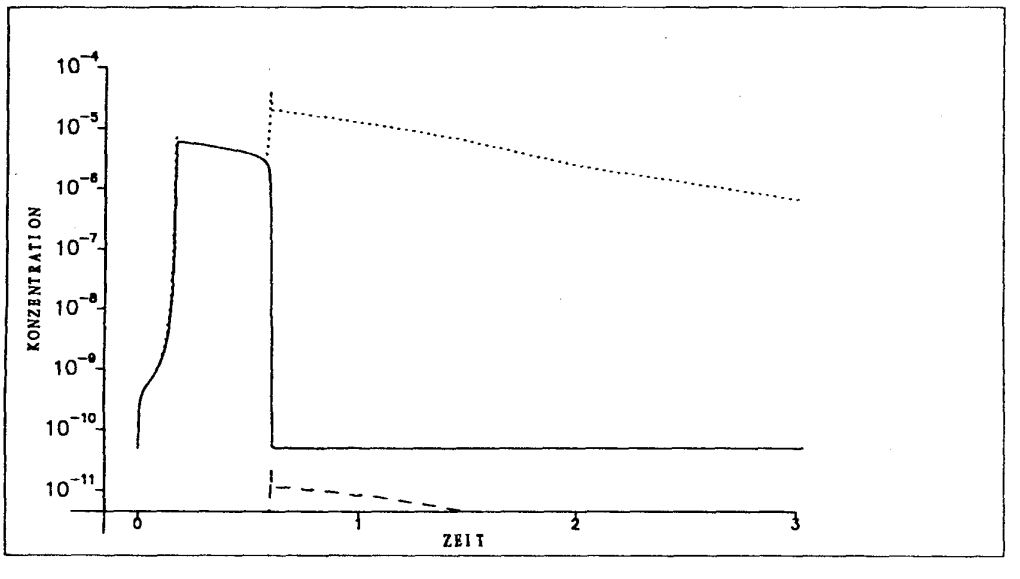

KONZENTRATION
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
ZEIT (* E+02)


KONZENTRATION
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
ZEIT (* E+02)

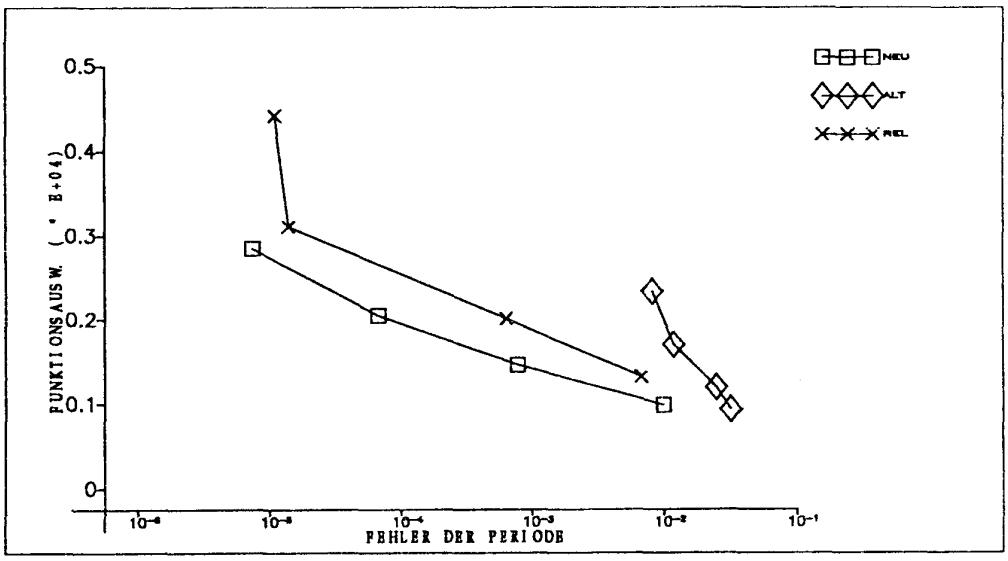

KONZENTRATION
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
ZEIT (* E+02)


KONZENTRATION
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
0
1
2
3
ZEIT


KONZENTRATION
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
0
1
2
3
ZEIT


KONZENTRATION
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
0
1
2
3
ZEIT

Ein Vergleich der Skalierungen fällt besonders eindrucksvoll aus, wenn man Aufwand und relativen Fehler des Gesamtverlaufs anhand der Periode vergleicht:



Die neue Skalierung liefert eine Genauigkeit der Periode im Rahmen der vorgegebenen Toleranz bei linearem Anwachsen (in logarithmischer Skala!) des Aufwandes. Das relative Fehlerkonzept in der Skalierung führt anfänglich zu übergenauen Perioden, dann aber beim Übergang $TOL = 10^{-4}$ zu $TOL = 10^{-5}$ zu keiner Verbesserung, da durch die hohe Anzahl der Schritte ein Moment der Ungenauigkeit zu überwiegen beginnt (vergleiche(1.1')). Das Skalierungskonzept (1.22) stellt sich als absolut ungenau heraus bei vergleichbarem Aufwand zu neuen Skalierung.

Fazit. Die Skalierung (1.20) erreicht höhere Genauigkeit, ist stark mit TOL verkoppelt und ist effizient.

2. Nieren-Modell-Problem (SCOTT/WATTS [100])

Dieses für bestimmte Anfangswerte extrem steife Beispiel macht die höhere Effizienz besonders deutlich. Hier führt das relative Fehlerkonzept (1.24) zu horrenden Rechenzeiten sowie zu einer drastischen Verschlechterung des globalen Fehlers (zu viele Schritte!). Das übliche Skalierungskonzept (1.22) für steife Integratoren bei mäßiger verlangter Genauigkeit TOL liefert abgebrochene bzw. inkorrekte Läufe. Mit interner Skalierung (1.20) dagegen wird das Problem für Toleranzen TOL $\leq 10^{-4}$ schnell und verläßlich gelöst (siehe Tabelle C.1).

Aus Stabilitäts- und hierbei auch Effizienzgründen empfiehlt sich in diesem Beispiel die Benutzung von EULSIM gegenüber METAN 1 (siehe Tabelle C.2).

Tabelle C.1 Aufwand (NFCN) mit Skalierung (1.20)

TOL	METAN 1	EULSIM
10^{-3}	bricht ab	1459
10^{-4}	3185	2002
10^{-5}	3985	2837
10^{-6}	5686	3701

Tabelle C.2 Für Anfangswert $y_5(0) = 0.990268835$ (Lösung eines Randwertproblems) ist die Differentialgleichung "nichtsteif".

F-Aufrufe:

	METAN 1 (1.22)	METAN 1 (1.20)	EULSIM (1.22)	EULSIM (1.20)
$y_5(0) := 0.990268835$	363	363	498	493
$y_5(0) := 0.99$	1034	763	901	544
$y_5(0) := 0.9$	2818	2106	918	834

2 Beurteilungskriterien für den Vergleich von Verfahren

Für jedes einigermaßen ausgereifte Verfahren gibt es Einzelprobleme, für die es "optimal" läuft.
Frage: Was heißt "optimal"?

Ziel: Herausarbeitung von *Problem-Klassen*, für die jeweils ein Verfahren brauchbar und effizient ist ("domain of applicability").

Integratoren sind zu vergleichen bezüglich:

- Rechenzeit (computing time)
- Speicherplatz (array storage)
- Genauigkeit (accuracy, precision)
- Verläßlichkeit (reliability)
- Robustheit (robustness)
- Output
- Interpolationseigenschaft
- Einfachheit

1. Rechenzeit

Bezeichnungen:

NFEV : Anzahl an f – Auswertungen (number f-evaluations)
OVHD : Overhead-Zeit für reinen Ablauf des Programmes
TIME : Gesamtrechenzeit

Zusätzlich bei steifen Integratoren
NDEC : Anzahl LR-Zerlegungen (LU *dec*ompositions)
NJAC : Anzahl Auswertungen der *Jac*obi-Matrix
NSOL : Anzahl Vorwärts-/Rückwärtssubiterationen (number of *sol*ves)

Im allgemeinen gilt:

$$\begin{aligned} \text{OVHD} &= \text{A(TOL)} \cdot \text{N} \\ \text{NSOL} &\sim \text{NFEV} \end{aligned} \tag{2.1}$$

Für *nichtsteife* Integratoren gilt:

$$\text{TIME} = \text{COSTF} * \text{NFEV} + \text{OVHD} \tag{2.2}$$

Damit wird bei nichtsteifen Integratoren der Wert

$$\text{C:=COSTF/N} \tag{2.3}$$

die unterscheidende Größe. Für *steife* Integratoren gilt:

$$\begin{aligned} \text{TIME} = {} & \text{COSTF} * \text{NFEV} + \text{COSTJ} * \text{NJAC} \\ & + \text{NDEC} * \text{COSTLR} + \text{NSOL} * \text{COSTS} \\ & + \text{OVHD} \end{aligned} \tag{2.4}$$

Kontrolle der verschiedenen Aufwandsanteile mit "interner Uhr" ermöglicht verfeinerte Anpassung an das zu lösende Problem.

2. Speicherplatz

Auch bei sehr großen Rechnern ist der aktuelle Arbeitsspeicher in der Regel nicht groß. Um häufiges Umspeichern (Zeitverlust) auf den Hintergrundspeicher zu vermeiden, sollte daher der Integrator bei großen DG-Systemen die Struktur nutzen. Das heißt zum Beispiel bei steifen Gleichungen und dünn besetzten Jacobi-Matrizen: Einsatz von Sparse- bzw. Bandsolvern.

3. Genauigkeit

Ausführliche Diskussion siehe Kapitel C.1. Vergleiche von Integratoren nur sinnvoll bei vergleichbarer Skalierung!

4. Verläßlichkeit

"A program may fail, but it *must not* lie" (B. PARLETT).
Steuerung des Verfahrens bzgl. Schrittweite (und Ordnung) soll möglichst sensitiv gegen lokale Änderungen von f sein.

→ Einschrittverfahren :	"neueste" f-Information
Mehrschrittverfahren :	"Geschichte" von f wesentlich benützt für Effizienz; häufig höheres TOL nötig aus Gründen der Verläßlichkeit.

5. Robustheit

Unempfindlichkeit auch bei kritischen Beispielen, abzufangen durch möglichst flexible Anpassung an Einzelprobleme. Dies verlangt im

wesentlichen: Flexibilität der Schrittweitensteuerung, zum Beispiel effektive Anpassung der Startschrittweite oder Reduktion der Schrittweite (und Ordnung) bei Unstetigkeitsstellen (eventuell höherer Ableitungen).

(1) Einschrittverfahren im allgemeinen robuster als Mehrschrittverfahren (Auswertung dividierter Differenzen numerisch manchmal empfindlich)

(2) Extrapolations-Verfahren empfindlich gegen instabile Programmierung von f.

Bemerkung: Übergang zu *einfacher* Genauigkeit (etwa sechs Dezimalziffern bei IBM) beeinträchtigt im allgemeinen die Effizienz von Mehrschrittverfahren, nicht jedoch von Einschrittverfahren.

6. Output

Meist graphische Ausgabe (Plot).

Einschrittverfahren :	stark nichtuniforme Gitter, besonders bei Extrapolation
	↪ geringer Speicherbedarf (wichtig bei *großen* Systemen)
	↪ benötigt spezielle Interpolationsroutinen für Plot
Mehrschrittverfahren :	meist quasi-uniforme Gitter
	↪ hoher Speicherbedarf
	↪ Standard-Interpolationsroutinen können verwendet werden.

Bemerkung: "zu viele" Zwischenpunkte würgen Codes ohne Interpolationsroutine ab.

7. Interpolationseigenschaft

Adams-Verfahren: Nordsieck-Darstellung (3.25) gestattet globale Darstellung der Lösung, die billig auszuwerten ist ($\rightarrow$ LSODE)

Runge-Kutta-Verfahren: DOPRI5 gestattet $O(h^5)$-Darstellung (2.55), allgemeine RK unterschiedlich bezüglich Interpolation.

Extrapolationsverfahren: DIFEX1 - für niedrige Ordnung Interpolation einfach (SHAMPINE/BACA/BAUER [102]), noch offene Entwicklungsmöglichkeiten.

Die Interpolationseigenschaft wird benötigt bei:

- Anwendung von Standard-Plotroutinen
- retardierten Differentialgleichungen:

  Beispiel:

  $$\begin{aligned} y' &= f(y(t), y(t-\tau)) \\ \tau &\geq 0 : \text{ Retardierung} \end{aligned}$$

  (Kreisläufe in chemischen Reaktoren, in biologischen Systemen)

- Bestimmung von Umschaltpunkten durch implizite Bedingungsgleichungen (Schaltbedingungen):

  $$f(y) = \begin{cases} f_1(y) \text{ für } S(y) < 0 \\ f_2(y) \text{ für } S(y) < 0 \end{cases} \quad S(y)(t)) : \text{ Schaltfunktion} \qquad (2.5)$$

  Beispiel: Zustandsbeschränkungen, Schaltpunkte bei Problemen der optimalen Steuerung (vgl. Kap. D.5.2).

8. Einfachheit

Wichtig bei Einbau in kompliziertere Fragestellungen des Scientific Computing.

3 Vergleich nichtsteifer Integratoren

Folgende nichtsteife Integratoren werden verglichen:

a) Extrapolationsmethoden

EULEX	explizites Euler-Verfahren mit h-Extrapolation (Kapitel A.2.3)
DIFEX1	explizite Mittelpunktsregel mit h^2-Extrapolation (Kapitel A.2.3)
ODEX	explizite Mittelpunktsregel mit h^2-Extrapolation, Programm von HAIRER/NØRSETT/WANNER 1987 [58]
ODEXS	ODEX mit derselben Skalierung wie DIFEX1

b) Explizite Runge-Kutta-Methoden

DOPRI5	Verfahren der Ordnung 5(4) nach Dormand-Prince, Programm von HAIRER/NØRSETT/WANNER 1987[58]
DOPRI8	Verfahren der Ordnung 8(7) nach Dormand-Prince, Programm von HAIRER/NØRSETT/WANNER 1987[58]
DOPRI8S	DOPRI8 mit derselben Skalierung wie DIFEX1

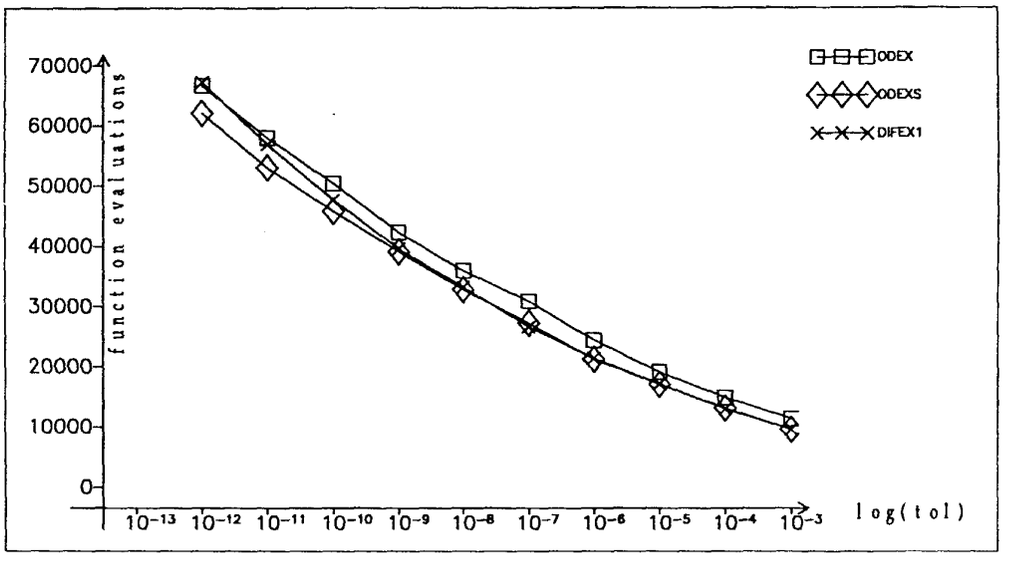
c) Mehrschrittmethoden

EPISODE	Adams-Methode, Programm von (HINDMARSH/BYRNE 1977) [66]
LSODE	Adams-Methode, Programm von (HINDMARSH 1981) [64]

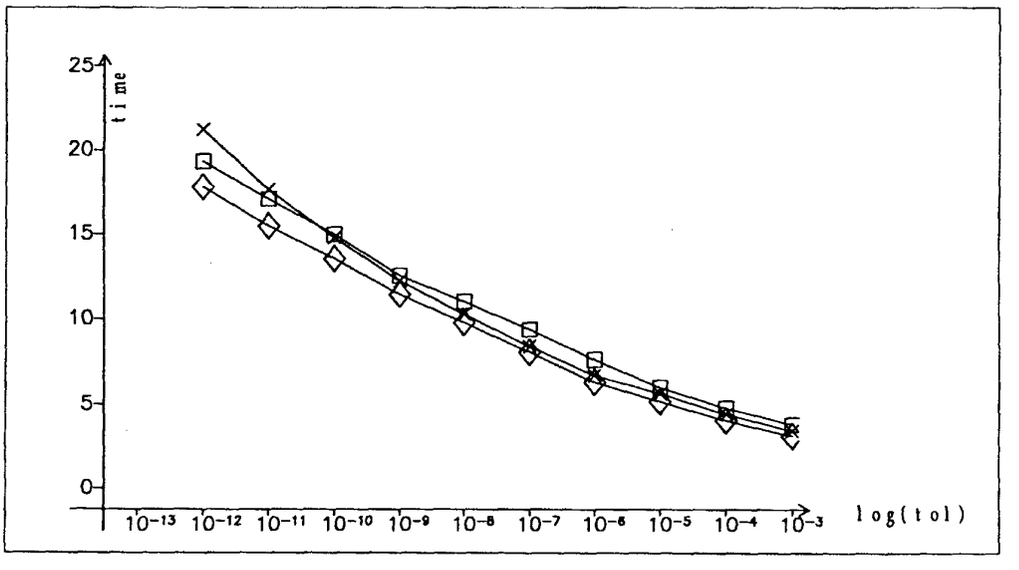
Als Testset wurde die Beispielsammlung von HULL et. al. [67] genommen, angereichert um einige etwas schwierigere Beispiele. In diesem Testset ist die Größe C nach (2.3) "klein" bis "mittel" , die Dynamik im wesentlichen ziemlich "regulär". Skalierung (1.22) durchgängig verwendet, falls nicht anders aufgeführt, um Vergleiche zu gestatten.

Alle Beispiele wurden mit den Toleranzen TOL=$10^{-3}, 10^{-4}, \ldots, 10^{-12}$ gerechnet und jeweils die Aufwandszahlen der einzelnen Beispiele addiert.

1a : Es ist deutlich der Effekt der Skalierung bei Übergang ODEX → ODEXS zu sehen. Ansonsten geringe Unterschiede.



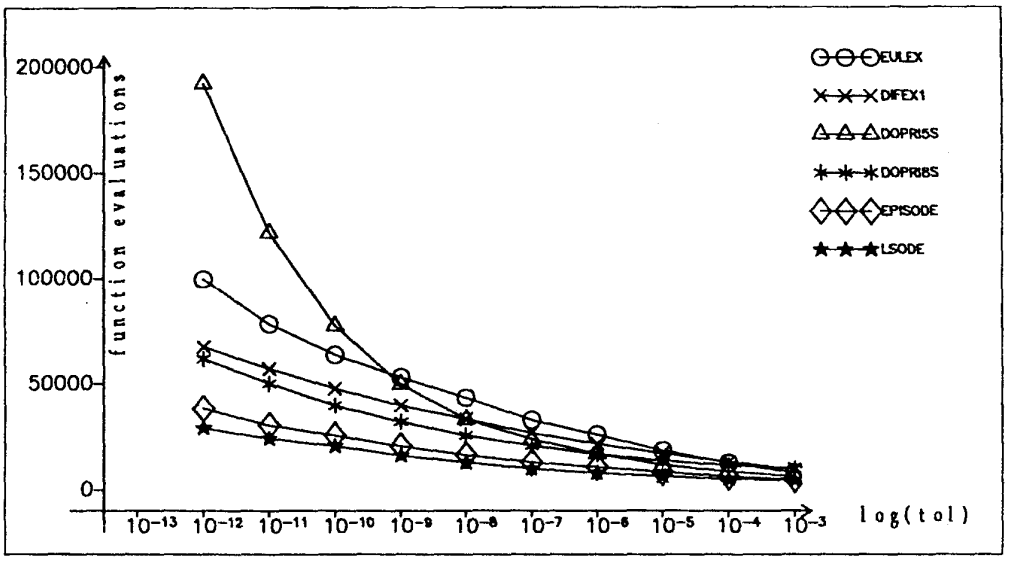
1b : Im wesentlichen wie 1a) : Extrapolationsverfahren haben geringen Overhead.

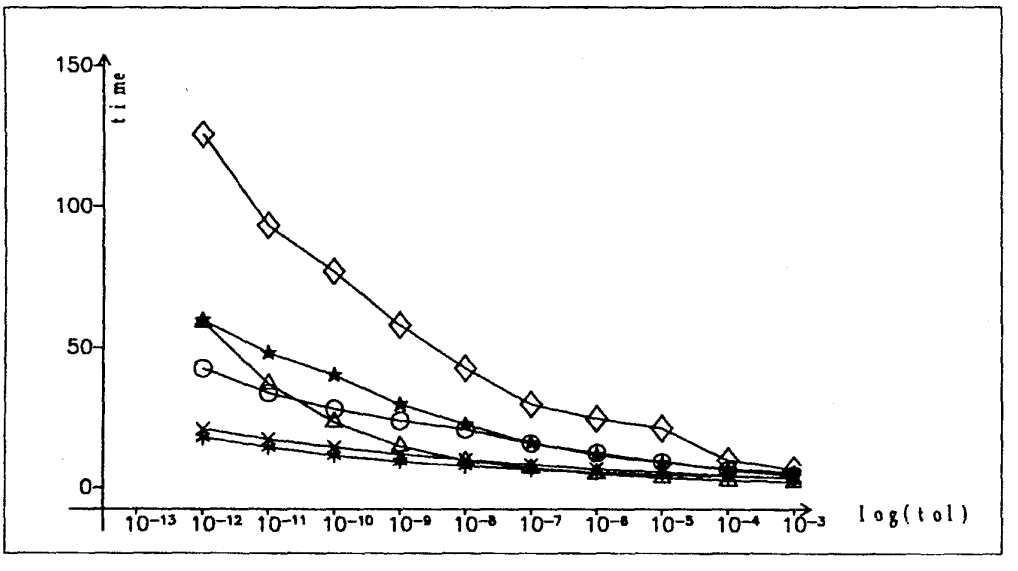


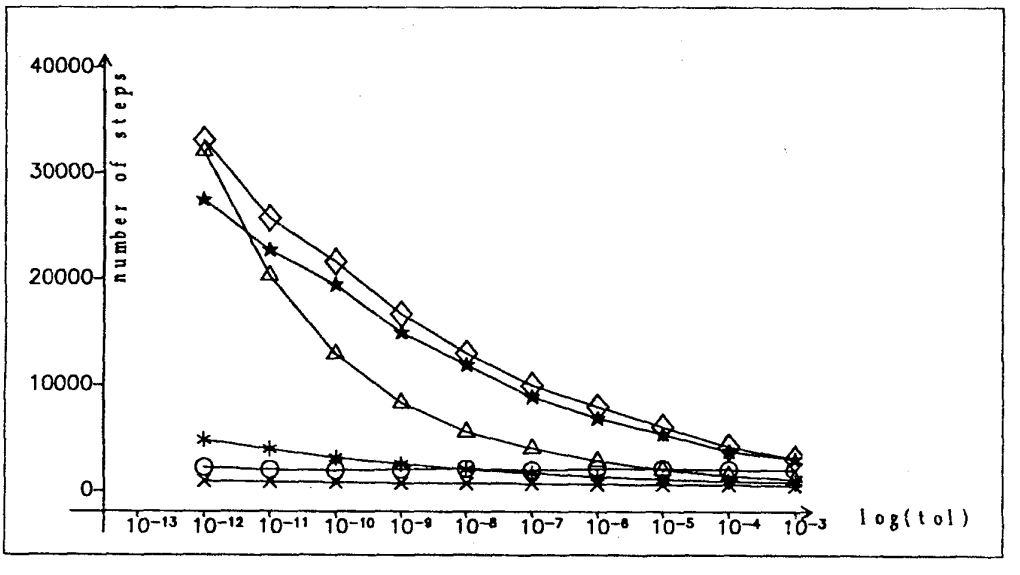
2a, 2b, 2c :

- Verfahren fixer Ordnung DOPRI5 nur für technische Genauigkeiten $10^{-3} - 10^{-5}$ verwenden,
- wenig Funktionsauswertungen bei Mehrschrittverfahren: Vorteil bei komplizierten ($\hat{=}$ teuren) rechten Seiten,
- deutlich ist der große Overhead bei Mehrschrittverfahren, insbesondere bei dem nicht-äquidistanten Verfahren EPISODE,
- für hohe Genauigkeit DOPRI8 und DIFEX1 gleichwertig,
- für technische Genauigkeit DOPRI5, DOPRI8 und DIFEX1 gleichwertig,

• Einschrittverfahren benötigen signifikant weniger Schritte als Mehrschrittverfahren, am wenigsten DIFEX1 (wichtig für graphischen Output bei *großen* Systemen).


EULEX
DIFEX1
DOPRI5S
DOPRI8S
EPISODE
LSODE
function evaluations
200000
150000
100000
50000
0
10^-13 10^-12 10^-11 10^-10 10^-9 10^-8 10^-7 10^-6 10^-5 10^-4 10^-3
log(tol)


time
150
100
50
0
10^-13 10^-12 10^-11 10^-10 10^-9 10^-8 10^-7 10^-6 10^-5 10^-4 10^-3
log(tol)


number of steps
40000
30000
20000
10000
0
10^-13 10^-12 10^-11 10^-10 10^-9 10^-8 10^-7 10^-6 10^-5 10^-4 10^-3
log(tol)

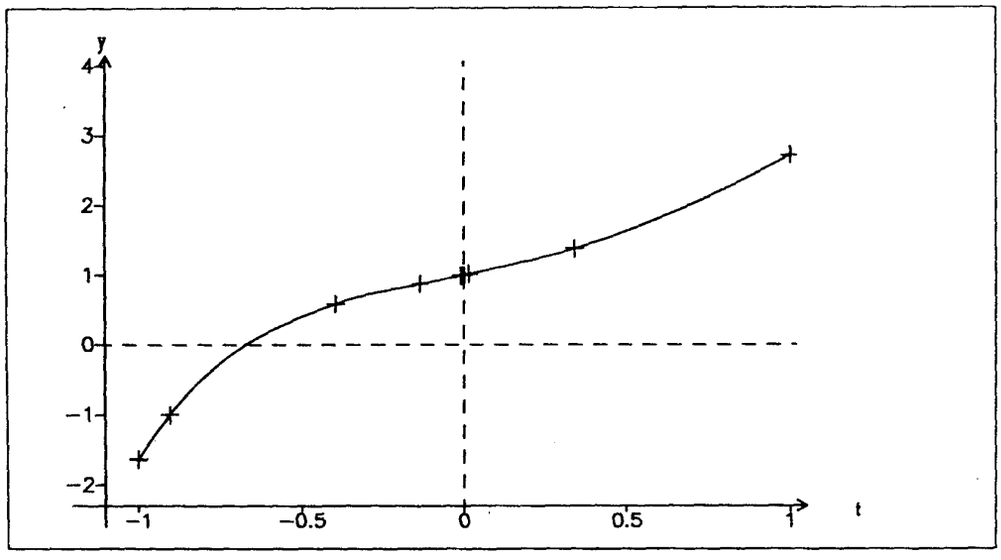
3a, 3b : Für die Bilder 3 bis 5 wurde ein Randwertproblem aus der Arbeit von RUSSELL/SHAMPINE [98] in ein Anfangswertproblem umgeschrieben:

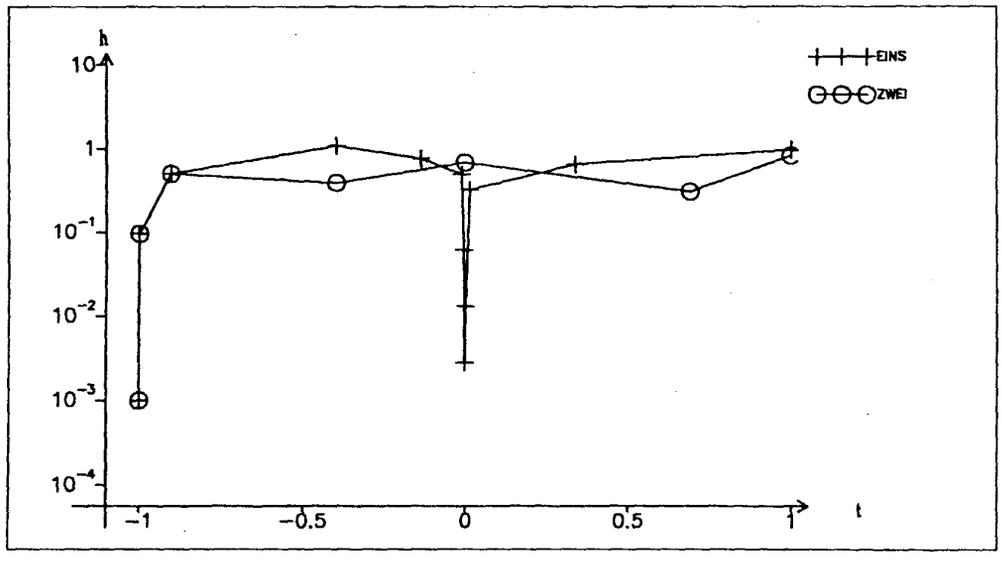
$$y'' = y - ty' + te^t - |t|(6 - 12t + 2t^2 - 3t^3)$$
$$y(-1) = e^{-1} - 2\,, \quad y'(-1) = e^{-1} + 7$$
$$t_{\text{end}} = 1$$

Die Lösungskurve $y(t)$ (Bild 3a) ist zweimal stetig differenzierbar, hat aber in der 3. Ableitung im Punkt 0 eine Sprungstelle. Um die Auswirkung auf die Schrittweitensteuerung zu sehen, wurde das Problem zweimal mit DIFEX1 (Toleranz 10^{-8}) gelöst und die Schrittweite H dargestellt (Bild 3b).

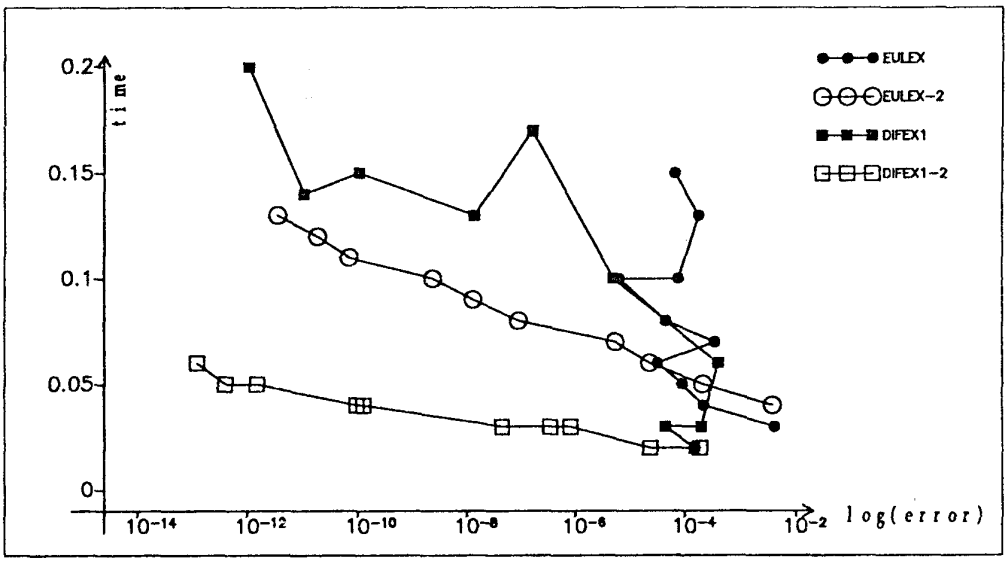
- einmal mit "Hinwegintegrieren" über die Singularität (EINS),
- einmal mit Zerlegung des Intervalls (ZWEI)

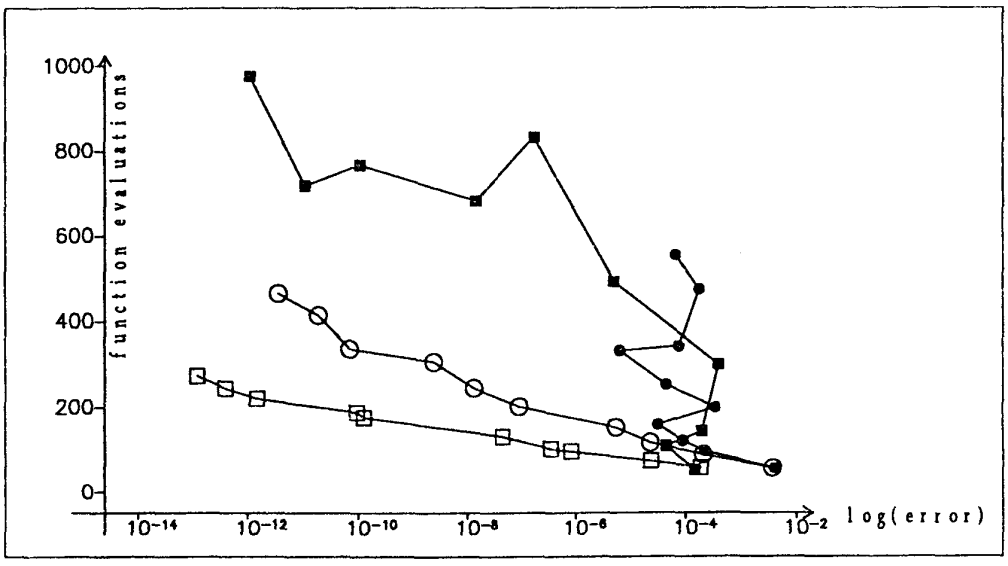
Fazit: Sind Singularitäten in den Ableitungen bekannt: Zerlegung des Intervalls.


y
4
3
2
1
0
−1
−2
−1
−0.5
0
0.5
1
t


h
10
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
EINS
ZWEI
−1
−0.5
0
0.5
1
t

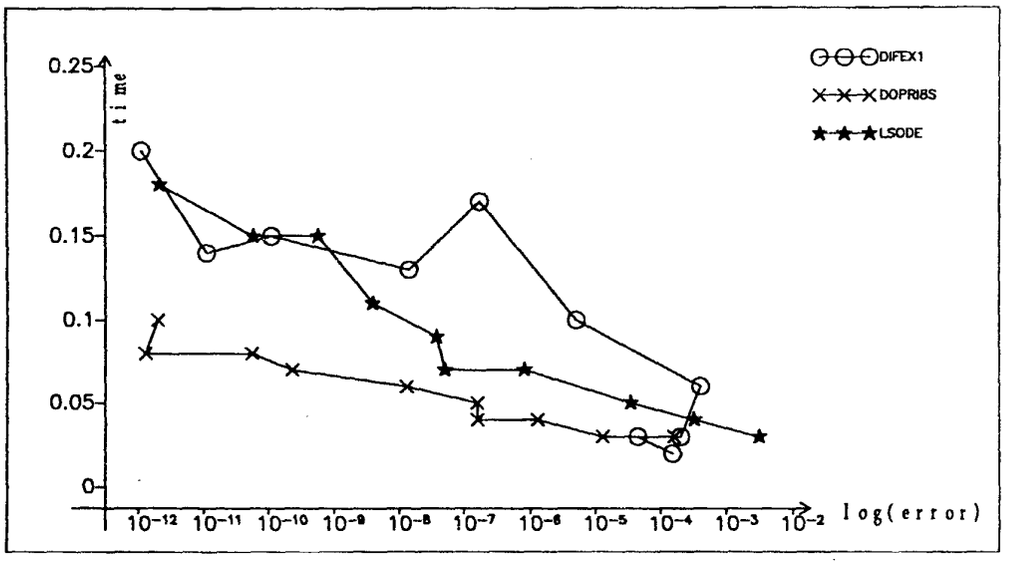
4a, 4b : Hier ist der Effekt speziell bei den Extrapolationsverfahren zu sehen: beim "Hinwegintegrieren" ist die interne Fehlerschätzung in EULEX wegen h-Entwicklung nicht mehr zuverlässig, im Gegensatz zur h^2-Entwicklung bei DIFEX1.

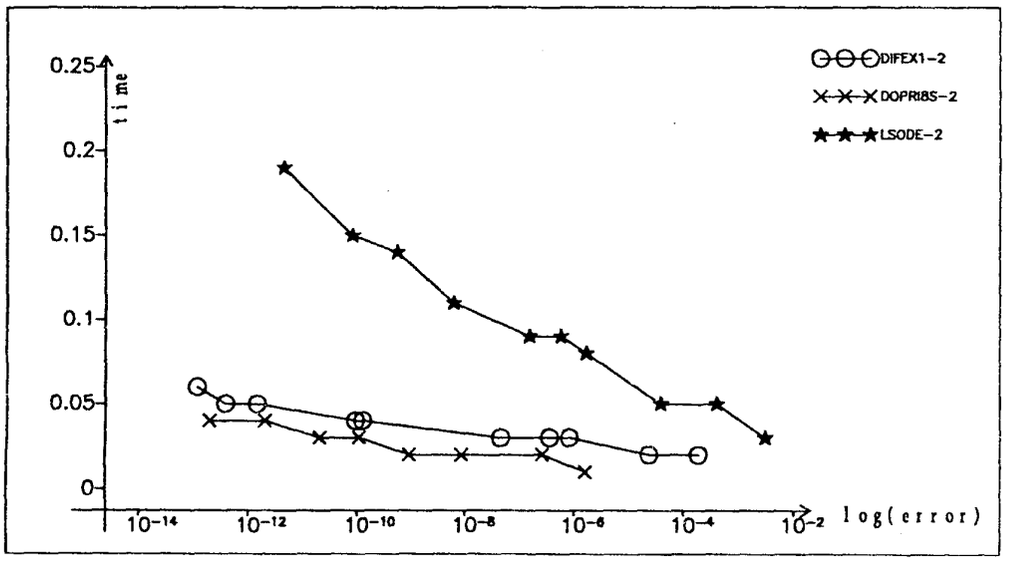




5a, 5b :

- Beim "Hinwegintegrieren" (Bild 5a) schlägt gewisse Robustheit des Verfahrens DOPRI8 mit fixer Ordnung zu Buche, es kann auf Unstetigkeiten in den Ableitungen nicht so empfindlich reagieren wie zum Beispiel das Extrapolationsverfahren DIFEX1. Bei unkritischen Unstetigkeiten wie hier von Vorteil.
- Auch das Mehrschrittverfahren besitzt ein gewisse Trägheit gegenüber Unstetigkeiten.





6.: Die folgenden Differentialgleichungen beschreiben die Bahnkurve eines künstlichen Erdsatelliten (nach STIEFEL/BETTIS) [107]:

$$
\begin{aligned}
y_1'' &= -\frac{y_1}{r^3} + k\left(\frac{5y_1y_3^2}{r^7} - \frac{y_1}{r^5}\right) \\
y_2'' &= -\frac{y_2}{r^3} + k\left(\frac{5y_2y_3^2}{r^7} - \frac{y_2}{r^5}\right) \\
y_3'' &= -\frac{y_3}{r^3} + k\left(\frac{5y_3y_3^3}{r^7} - \frac{3y_3}{r^5}\right) \\
\text{mit} \quad r^2 &= y_1^2 + y_2^2 + y_3^2\,, \quad k = 1.4 \cdot 10^{-3}
\end{aligned}
$$

und den Anfangsbedingungen

$$
\begin{aligned}
y_1(0) &= 1 \\
y_2(0) &= y_3(0) = y_1'(0) = 0 \\
y_2'(0) &= y_3'(0) = \tfrac{1}{2}\sqrt{2}
\end{aligned}
$$

Dieses Beispiel zeigt erneut die Unterschiede von Extrapolations-, Mehrschritt- und Runge-Kutta-Verfahren:

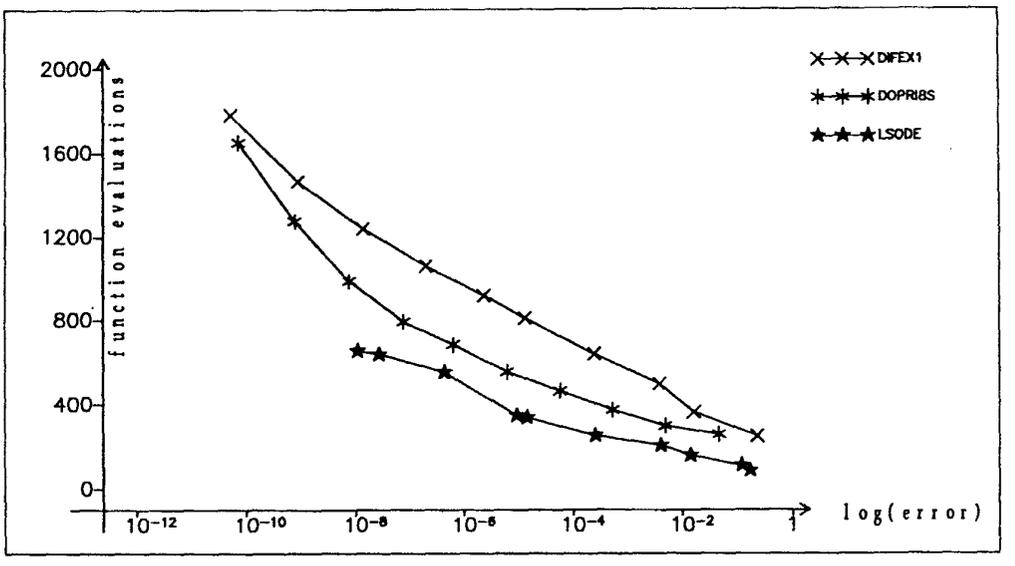
a) Auswertungen der rechten Seite

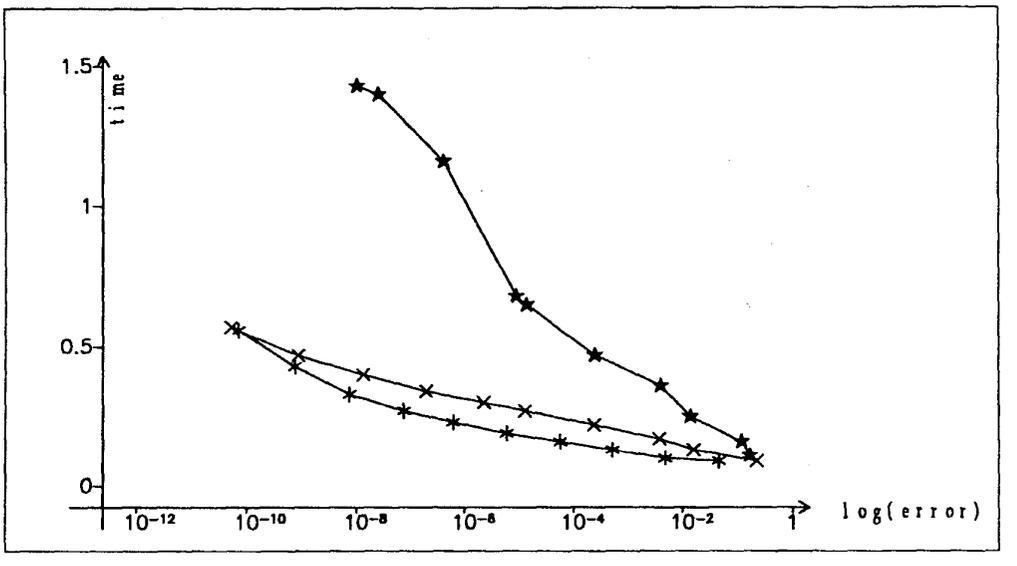
b) Rechenzeit

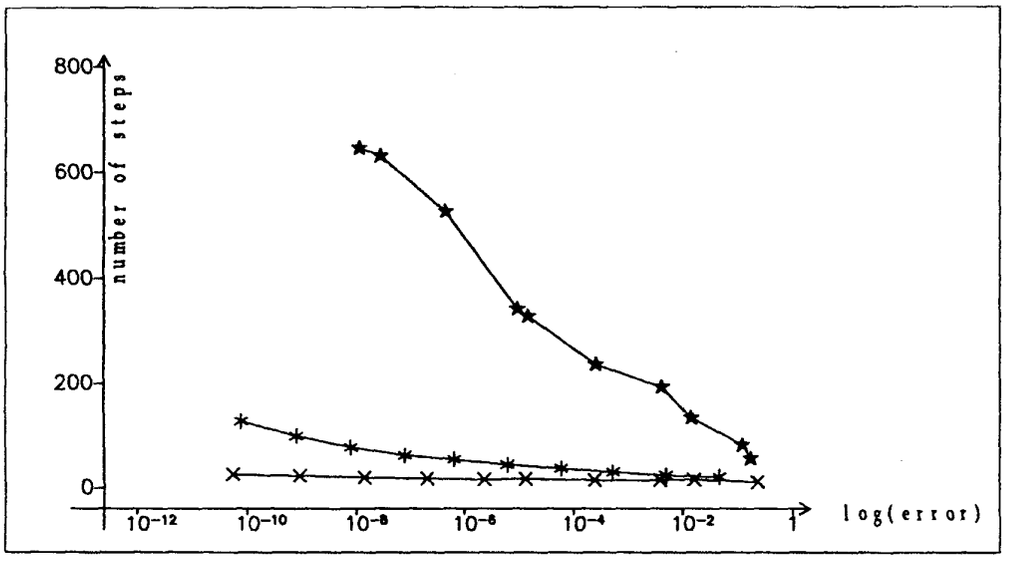
c) Integrationsschritte

Fazit:

- Mehrschrittverfahren braucht am wenigsten f-Aufrufe,
- hoher Overhead des Mehrschrittverfahrens,
- hohe Schrittzahl des Mehrschrittverfahrens: hieraus ergibt sich die geringe erzielte Genauigkeit,
- gewisse Vorteile in Rechenzeit und f-Aufrufen bei Verfahren hoher fixer Ordnung,
- Charakteristiken für Extrapolationsverfahren:
  Nahezu konstante Schrittzahl, sehr wichtig $\Longrightarrow$ resultieren hohe Genauigkeiten. Außerdem spart dies Speicherplatz für graphischen Output. Speicheraufwand bei TOL=10^{-3} um Faktor 5 und bei TOL=10^{-8} um den Faktor 21 kleiner in diesem Beispiel.


DIFEX1
DOPRI8S
LSODE
2000
1600
1200
800
400
0
function evaluations
10-12
10-10
10-8
10-6
10-4
10-2
1
log(error)


1.5
1
0.5
0
time
10-12
10-10
10-8
10-6
10-4
10-2
1
log(error)


800
600
400
200
0
number of steps
10-12
10-10
10-8
10-6
10-4
10-2
1
log(error)

Auf der Basis der Erfahrung sind folgende Faustregeln gerechtfertigt:

(I) Falls Dynamik des Problems "irregulär":
Extrapolationsverfahren
(aus Gründen von Robustheit und Verläßlichkeit)

(II) Falls Dynamik des Problems eher "regulär" : C:=COSTF/N

C "groß" : LSODE
C "mittel" : DIFEX1, DOPRI8

(III) Falls "dichter" Output gewünscht:

hohe Genauigkeit : LSODE
mittlere Genauigkeit : EULEX, DOPRI5 hohe Genauigkeit

Bemerkung: "Expertensystem" muß Benutzerdialog mit obigem Entscheidungsbaum aufbauen, zusätzlich intern COSTF/N durch "interne Uhr" prüfen und Entscheidung des Benutzers eventuell korrigieren.

4 Vergleich steifer Integratoren

Folgende steife Integratoren werden verglichen:

a) Extrapolationsmethoden

EULSIM	semi-implizites Euler-Verfahren mit h-Extrapolation (Kap. B.2.3)
METAN1	semi-implizite Mittelpunktsregel mit h^2-Extrapolation (Kap. B.2.3)

b) Mehrschrittmethoden

EPISODE	BDF-Verfahren, Programm von HINDMARSH/BYRNE 1977 [66]
LSODE	BDF-Verfahren mit maximaler Ordnung 5, Programm von HINDMARSH 1981 [64]
LSODE-3	LSODE mit auf 3 beschränkter Ordnung

Als Testset wurde die Beispielsammlung von HULL et. al. [67] ohne das Beispiel B5 gewählt. Zusätzlich die Van der Pol Gleichung :

$$\begin{aligned} y_1' &= y_2 \\ \varepsilon y_2' &= (1-y_1^2)y_2 - y_1 \end{aligned} \tag{4.1.a}$$

mit $\varepsilon = 10^{-2}$, den Anfangsbedingungen

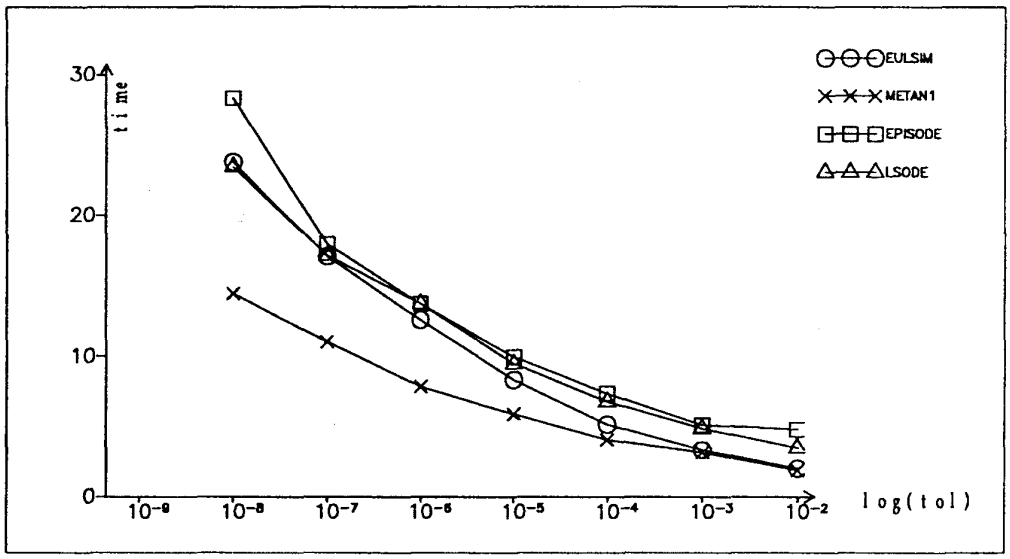
$$\begin{aligned} y_1(0) &= 1.693213222307211 \\ y_2(0) &= -0.906925252881142 \end{aligned} \tag{4.1.b}$$

Integrationsende:

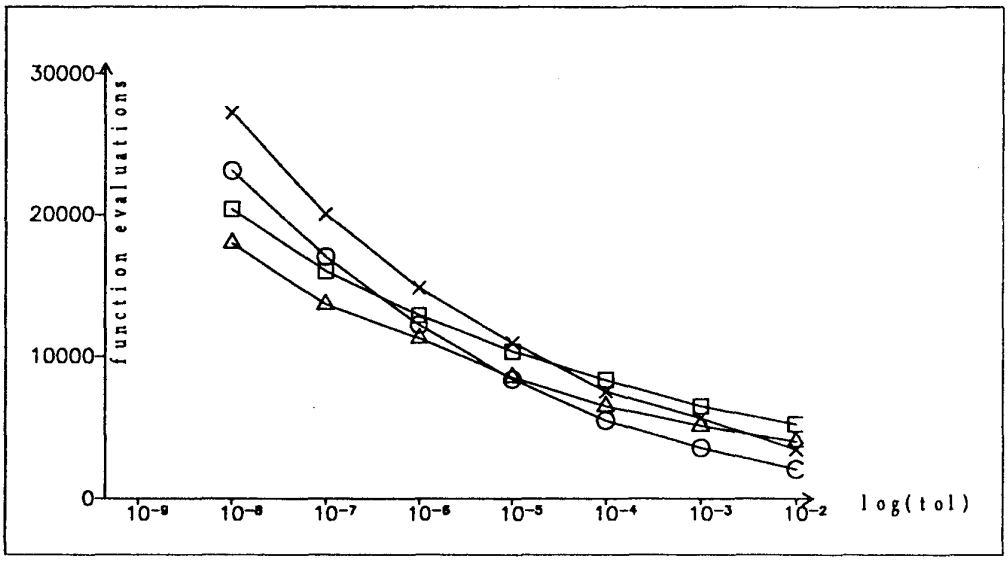
$$t_{\text{end}} = 2(3 - \ln 2)$$

Rechenzeit, Funktionsauswertungen und Schrittzahl wurden jeweils für die Toleranzen TOL=10^{-2}, 10^{-3}, ..., 10^{-8} über summiert.

1a :



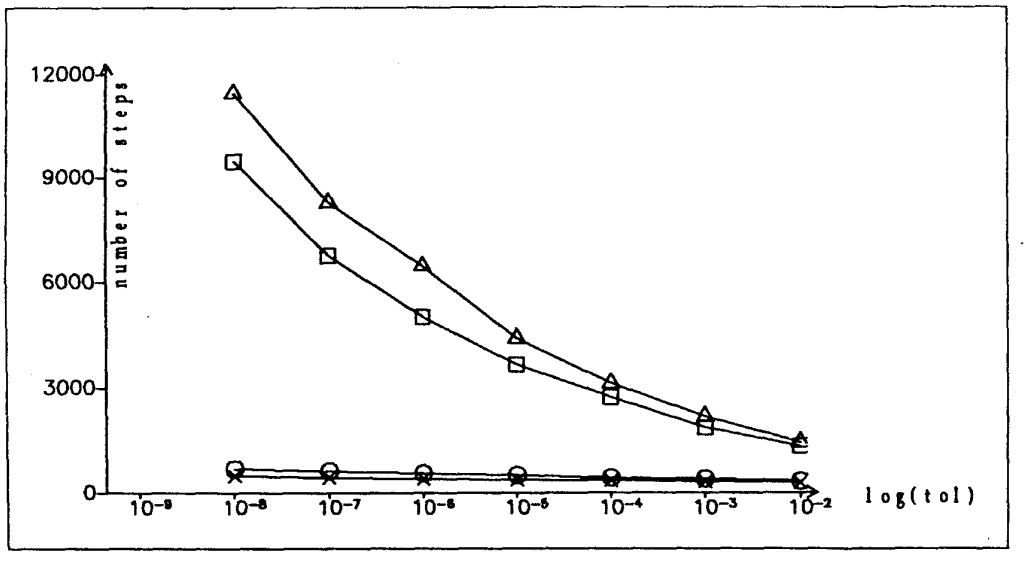
- Deutliche Tendenz der Mehrschrittverfahren zu weniger f-Aufrufen; man beachte, daß diese für große Beispiele dominieren, hier nur relativ klein.



1b:

- Bei steifen Problemen in der Regel nur geringe Genauigkeit erforderlich, dort EULSIM von Vorteil. **Bemerkung:** Zwei Fail-Läufe von MSV einfach in der Statistik weggelassen.

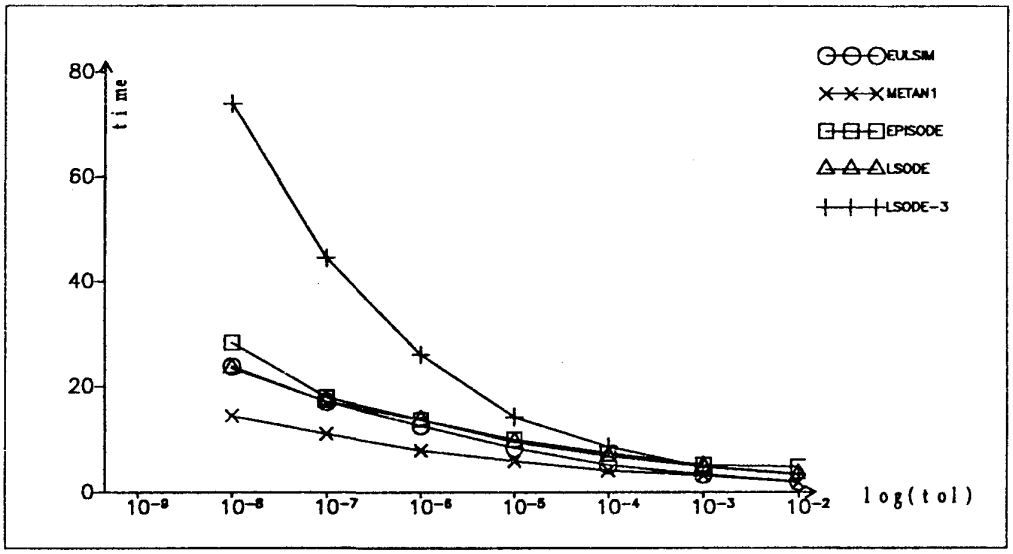
1c :



- Wie im nichtsteifen Fall nahezu Konstanz der Schrittzahlen bei Extrapolationsverfahren
  - ↪ a) hohe erzielbare Genauigkeit vgl. 1.1′
  - b) geringer Speicherbedarf für graphischen Output

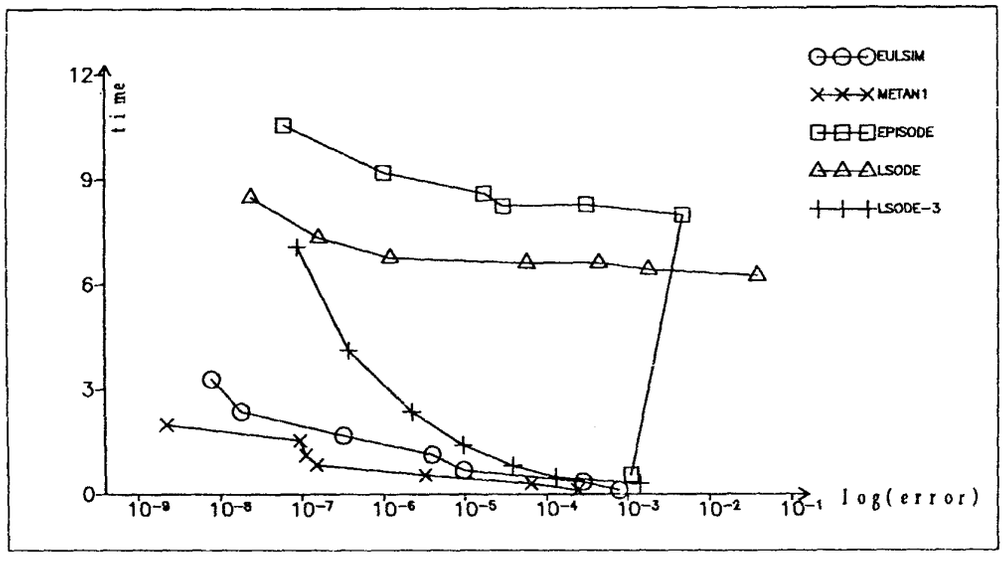
- großes Anwachsen der Schrittzahlen bei Mehrschrittverfahren
  ↪ Tendenz: ungenauer als Extrapolationsverfahren

1d :



Schränkt man die BDF-Verfahren grundsätzlich auf Ordnung 3 ein (Grund siehe Kap. B.3.2, Illustration siehe Beispiel B5, Fig. 2), so steigen die Rechenzeiten erheblich.

2 :



Beispiel B5 aus HULL et. al [67]

$$\begin{aligned} y_1' &= -10y_1 - \beta y_2\,, \quad y_1(0) = 1\,, \\ y_2' &= -\beta y_1 - 10y_2;, \quad y_2(0) = 1\,, \\ y_3' &= -4y_3;, \quad y_3(0) = 1\,, \\ y_4' &= -y_4;, \quad y_4(0) = 1\,, \\ y_5' &= -0.5y_5;, \quad y_5(0) = 1\,, \\ y_6' &= -0.1y_6;, \quad y_6(0) = 1\,, \end{aligned} \tag{4.2}$$

$$\beta = 100\,, \quad t_{\text{end}} = 20$$

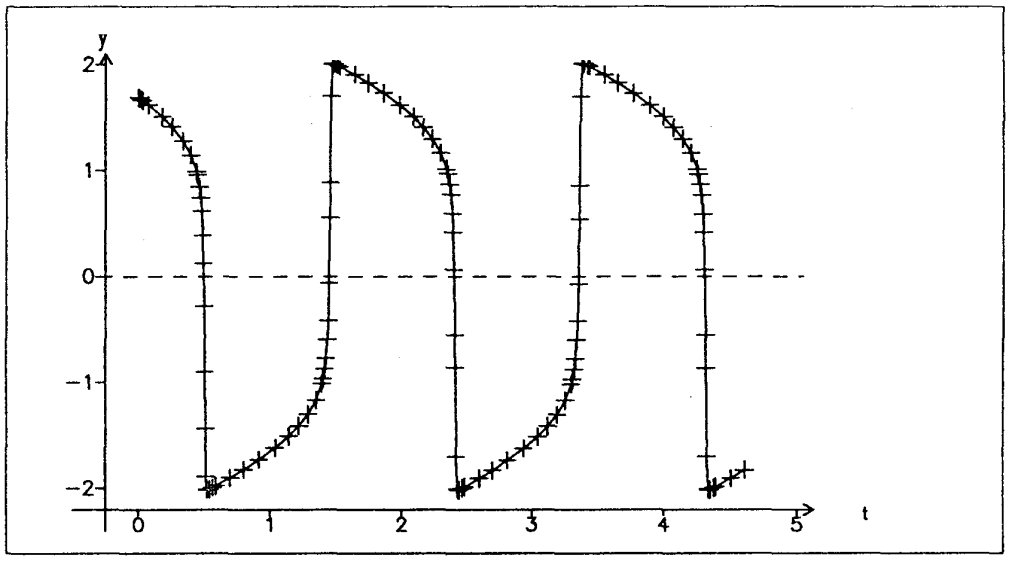
benötigt die $A(\alpha)$- Stabilität von 84°. Hier scheitern die BDF-Verfahren hoher Ordnung (vergleiche Tabelle 1 in B.3.2)

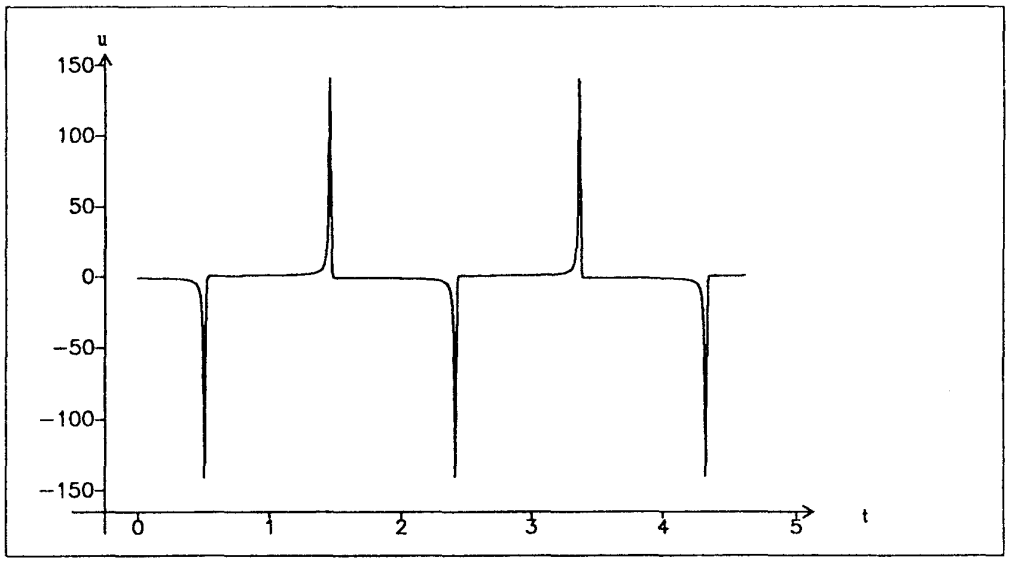
Deshalb:

- Für Vergleich auf Vergleichbarkeit der Stabilitätsgebiete bezüglich des Beispiels achten!
- Da in Praxis die nötigen Stabilitätsgebiete kaum bekannt: BDF nur bis Ordnung 3: Nachteil (vergleiche Fig. 1d)

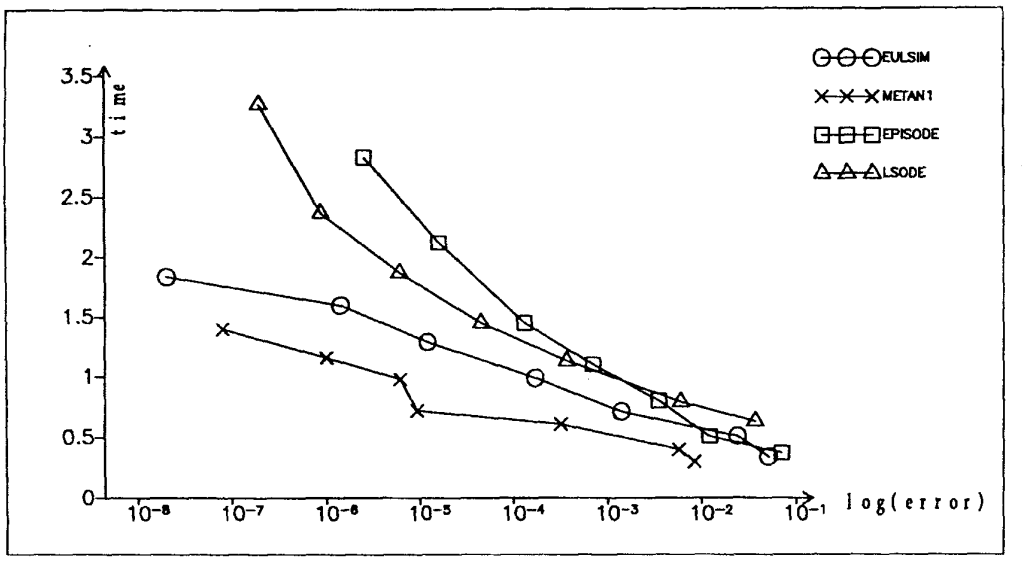
Bemerkung: große α für Chemie wichtig, falls Konvektiondominiert.

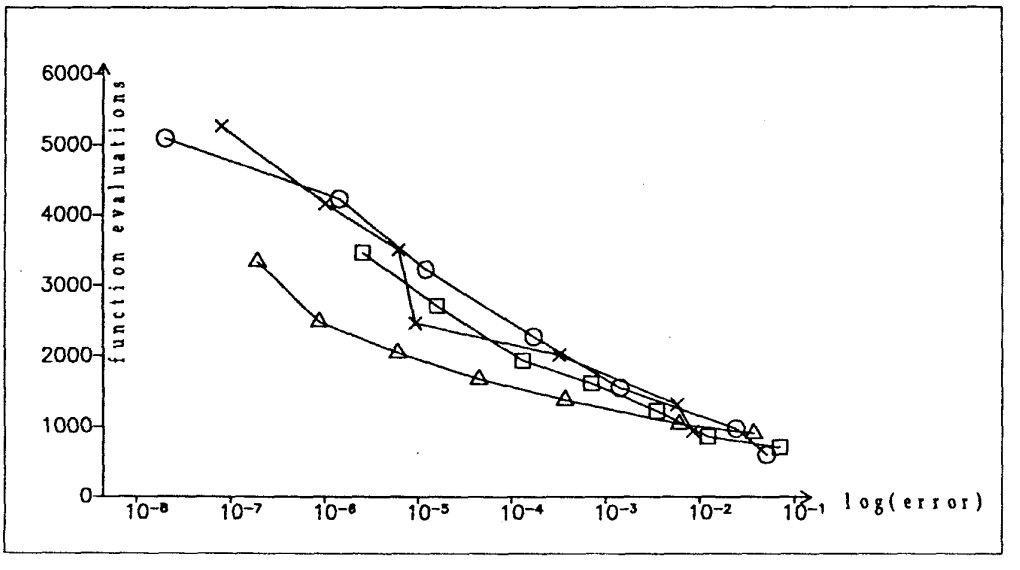
3 : Lösung $y(t)$ und Ableitung $u(t) = y'(t)$ der **Van der Pol Gleichung** (4.1) berechnet durch EULSIM mit Toleranz TOL=10^{-6}. Der Effekt der Schrittweitensteuerung ist deutlich sichtbar.





4 : Aufwand zur Lösung der obigen Van der Pol Gleichung gegenüber dem globalen Fehler bei lokalen Toleranzen von TOL=10^{-2}, 10^{-3}, ..., 10^{-8}.





- Extrapolationsverfahren: Tendenz genauer zu sein (siehe 1c).
- EULSIM erzielt höchste Genauigkeiten.
- Overhead bei Mehrschrittverfahen.
- Weniger f-Aufrufe bei Mehrschrittverfahren.

Fazit: Es ergeben sich keine einfachen Faustregeln. Für technisch relevante Genauigkeiten sind semi-implizite Extrapolationsverfahren vorzuziehen – auch und gerade bei großen und kritischen Problemen. Wie bei nichtsteifen Problemen ergibt sich als *Charakteristik der Extrapolationsverfahren*:
Nahezu konstante Schrittzahl, sehr wichtig $\Longrightarrow$ resultieren hohe Genauigkeiten. Außerdem spart dies Speicherplatz für graphischen Output. Speicheraufwand bei TOL=10^{-3} um Faktor 7 und bei TOL=10^{-8} um den Faktor 24 kleiner für diesen Testset.

D. Mehrzielmethode zur Lösung von Randwertproblemen

Randwertprobleme für gewöhnliche Differentialgleichungen lassen sich im Prinzip auf zwei Arten lösen:

(A) durch Rückführung auf eine Kette von Anfangswertproblemen,

(B) durch sogenannte globale Diskretisierung (etwa finite Differenzen oder Kollokation).

Ein Überblick über eine Reihe von Lösungsmethoden findet sich in dem jüngst erschienenen Buch [2] von ASCHER/MATTHEIJ/RUSSELL. Im folgenden werden nur Methoden vom Typ (A) behandelt, da sie in natürlicher Weise in den Kontext der Kapitel A und B passen.

1 Theoretische Grundlagen

Betrachtet werden zunächst sogenannte *Zweipunkt-Randwertprobleme:*

$$\begin{aligned} &\text{a)} \quad y' = f(y), && t \in [a,b] \\ &\text{b)} \quad r(y(a),\, y(b)) = 0 \end{aligned} \tag{1.1}$$

Brauchbare Existenzsätze zur Lösung von Randwertproblemen gibt es nur bei Zusatzstruktur (Linearität, Elliptizität, ...). Deshalb sei im folgenden stets *Existenz* einer Lösung y^* vorausgesetzt.

Satz 1 (Lokale Eindeutigkeit) *(*WEISS 1973 [112], DEUFLHARD 1980 [29]).
Sei y^* eine Lösung des Randwertproblems (1.1). *Sei $W^*(t,a)$ die Wronski-Matrix der Variationsgleichung*

$$\delta y' = f_y(y^*(t))\delta y \ .$$

Mit den Bezeichnungen

$$\begin{aligned} A^* &:= \left.\frac{\partial r}{\partial y_a}\right|_{y^*}, \quad B^* := \left.\frac{\partial r}{\partial y_b}\right|_{y^*} \\ E^*(t) &:= A^* W^*(a,t) + B^* W^*(b,t) \end{aligned}$$

Sensitivitätsmatrix in $t \in [a,b]$

folgt dann: Falls

$$\begin{aligned} & E^*(t) \quad \textit{nichtsingulär für ein } t \in [a,b] \\ \Rightarrow\ & E^*(t) \quad \textit{nichtsingulär für alle } t \in [a,b] \end{aligned} \tag{1.2}$$

und y^* ist lokal eindeutige Lösung von (1.1).

Beweis: (I) Zur Verifikation zeigt man mit der Gruppeneigenschaft (1.18), Kapitel A.1.2.

$$E^*(t) = E^*(a)W^*(a,t), \quad \forall t \in [a,b] \ . \tag{1.2'}$$

Da W^* nichtsingulär, folgt (1.2). Sei also o.B.d.A. $E^*(a)$ als nichtsingulär angenommen.

(II) (Skizze:) Man konstruiert den zum Randwertproblem (1.1) gehörigen Operator:

$$T(y) := \begin{cases} y(t) - y(a) - \int_{s=a}^{t} f(y(s))ds \ , & t \in [a,b] \\ r(y(a), y(b)) \end{cases}$$

Dann gilt für ein $B \subseteq C^1$:

$$\begin{aligned} T : B &\rightarrow B \times \mathbb{R}^n \\ T(y^*) &= 0 \end{aligned}$$

Es liegt eine lokal eindeutige Lösung sicher dann vor, wenn gilt (hinreichende Bedingung):

$$T(y) \quad \text{injektiv in } \mathcal{U}(y^*),$$

oder äquivalent:

$$T_y(y^*)\delta y = 0 \Rightarrow \delta y = 0$$

Berechnung mit Frechét-Ableitung:

$$0 = T_y(y^*)\delta y = \begin{cases} \delta y(t) - \delta y(a) - \int_{s=a}^{t} f_y(y^*(s))\delta y(s)ds \\ A^*\delta y(a) + B^*\delta y(b) \end{cases}$$

Explizit:

$$\begin{aligned} & \delta y(t) = \delta y(a) + \int_{s=a}^{t} f_y(y^*(s))\delta y(s)ds \\ \Rightarrow\ & \delta y(t) = W^*(t,a)\delta y(a) \end{aligned} \tag{1.3}$$

Da W^* nichtsingulär (Eindeutigkeit bei Anfangswertproblemen!), folgt:

$$\delta y(a) = 0 \iff \delta y(t) = 0 \ , \ \forall t \in [a,b] \tag{$*$}$$

Für $\delta y(a)$ erhält man das lineare Gleichungssystem

$$\underbrace{(A^* + B^* W^*(b,a))}_{E^*(a)} \delta y(a) = 0$$

Damit gilt: $E^*(a)$ nichtsingulär $\Rightarrow \delta y(a) = 0$. Mit $(*)$ ist der Beweis vollständig.

■

Approximationen von Sensitivitätsmatrizen werden demnach bei der numerischen Lösung von Randwertproblemen auftreten. Probleme wird man erwarten bei *fastsingulärer* Sensitivitätsmatrix.

Beispiel: Künstliches Grenzschichtproblem (LENTINI/PEREYRA [81]):

$$\begin{aligned} &\text{a) } y'' = -\frac{3\tau}{(\tau + t^2)^2} \cdot y, \quad \tau: \text{ Scharparameter} \\ &\text{b) } y(0.1) = -y(-0.1) = \frac{0.1}{\sqrt{\tau + 0.01}} \end{aligned} \tag{1.4}$$

Nach Symmetrie in a) und b) erwartet man:

$$\begin{aligned} & y^*(-t) = -y^*(t) \\ \Rightarrow \quad & y^*(0) = 0 \end{aligned}$$

Für $\tau \neq 0.01$ gilt dies auf verlangte Genauigkeit. Für $\tau = 0.01$ ergibt sich jedoch eine zusammenhängende Lösungsschar – siehe das Bild D.1.
Die Lösungsschar wurde mit dem Anfangswertlöser DIFEX1 (vgl. Kap. A.2.3) ausgehend von den folgenden Anfangswerten berechnet:

$$y(-0.01) \quad = -7.0710678119$$

und nacheinander

$$y'(-0.01) \quad = 140.0, 100.0, 60.0, 0.0, -40.0 \text{ und } -80.0.$$

Bemerkung: Der Randwertlöser BVPSOL (siehe Kap. D.3.2) liefert mit den Randbedingungen

$$y(0.0) = 0.0\,, \quad y(0.01) = 7.0710678119$$

die ungerade Lösung. Mit den Randbedingungen

$$y(-0.01) = -7.0710678119\,, \quad y(0.01) = -y(-0.01)$$

liefert BVPSOL, abhängig von der gewählten Anfangsnäherung, jedoch *irgendeine* Lösung aus der Lösungsschar.

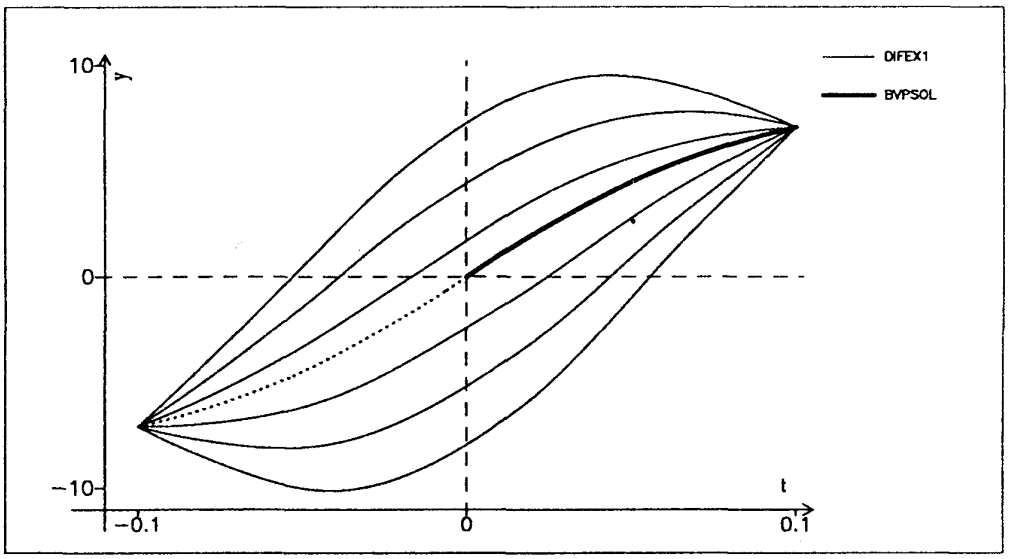


Bild D.1 "Numerische Hysteresis"

Begründung: Die Sensitivitätsmatrix hat die Gestalt:

$$E(a) = \begin{bmatrix} 1 & 0 \\ * & \alpha(\tau) \end{bmatrix}, \quad \alpha(\tau) = \frac{\partial y(+0.1;\tau)}{\partial y'(-0.1)} \tag{1.4.c}$$

Für $\tau = 0.01$ gilt: $\alpha = 0$ (vgl. Aufgabe 37).

Sensitivität gegen Störung der Randbedingungen

Eingabedaten : Komponenten von $y(a)$ und $y(b)$, zusammengefaßt:

$$y(a,b) \ .$$

Gestörte Randbedingungen : $r - \delta r$.
Zugehörige Störung der Lösung sei δy_r .
Störung der Eingabedaten: $\delta y(a,b)$ (* weggelassen).
Einsetzen von (1.2') und (1.3):

$$\delta r = A\delta y(a) + B\delta y(b) = E(a)\delta y(a)$$

$$\delta y_r(t) = W(t,a)E(a)^{-1}\delta r = E(t)^{-1}\delta r \tag{1.5}$$

Äquivalent zu:

$$\delta y_r(t) = E(t)^{-1}A\ \delta y(a) + E(t)^{-1}B\ \delta y(b) \tag{1.5'}$$

Übergang zu Norm ($\|\cdot\|_\infty$) :

$$\begin{aligned} \|\delta y_r(t)\| &\leq \|E(t)^{-1}A\| \cdot \|\delta y(a)\| + \\ &+ \|E(t)^{-1}B\| \cdot \|\delta y(b)\| \\ \text{Für } \|\cdot\|_\infty \text{ gilt} &: \|\delta y(a)\| \leq \|\delta y(a,b)\| \\ &\|\delta y(b)\| \leq \|\delta y(a,b)\| \end{aligned}$$

Eingesetzt:

$$\|\delta y(t)\|_\infty \leq \left[\|E(t)^{-1}A\| + \|E(t)^{-1}B\|\right] \|\delta y(a,b)\|_\infty$$

Konditionszahlen bezüglich δr
(MATTHEIJ 1982 [85], DEUFLHARD/BADER 1983 [35])

$$\begin{aligned} \rho_a(t) &:= \|E(t)^{-1}A\|, \quad \rho_b(t) := \|E(t)^{-1}B\| \\ \bar{\rho} &:= \max_{t\in[a,b]} \rho(t) \equiv \max_{t\in[a,b]} [\rho_a(t) + \rho_b(t)] \end{aligned} \tag{1.6}$$

$$\begin{aligned} &\text{a)} \quad \|\delta y_r(t)\|_\infty \leq \rho(t)\, \|\delta y(a,b)\|_\infty \\ &\text{b)} \quad \max_{t\in[a,b]} \|\delta y_r(t)\|_\infty \leq \bar{\rho}\, \|\delta y(a,b)\|_\infty \end{aligned} \tag{1.7}$$

Sensitivität gegen Störung der Differentialgleichungen

Sei $\delta f(s)$ lokale Störung der rechten Seite der Differentialgleichung. Durch Lösung der Variationsgleichung erhält man die dadurch bedingte Störung der Lösung:

$$\begin{aligned} \delta y_f(t) &= \int_a^t E(t)^{-1}AW(a,s)\delta f(s)ds - \\ &\quad - \int_t^b E(t)^{-1}BW(b,s)\delta f(s)ds \end{aligned} \tag{1.8}$$

(Vergleiche (1.23) in Kapitel A.1.2.)

Diese Formel definiert zugleich die *Green'sche Funktion* $G(t,s)$ gemäß:

$$G(t,s) := \begin{cases} E(t)^{-1}AW(a,s) & a \leq s \leq t \\ -E(t)^{-1}BW(b,s) & t \leq s \leq b \end{cases} \tag{1.9}$$

Damit (1.8) äquivalent zu:

$$\delta y_f(t) = \int_a^b G(t,s)\delta f(s)ds \tag{1.8'}$$

Konditionszahlen bezüglich δf (KREISS 1972 [71], OSBORNE 1979 [89]):

$$|\cdot|_\infty = \max_{t\in[a,b]} \|\cdot\|$$

$$\begin{aligned} &\text{a)} \quad |\delta y_f|_\infty \le \kappa_f \cdot |\delta f|_\infty \\ &\text{wobei} \\ &\text{b)} \quad \kappa_f := |b-a| \cdot \max_{s,t\in[a,b]} \|G(t,s)\| \end{aligned} \tag{1.10}$$

Verfeinerung (DEUFLHARD, BADER 1983 [35]):
Man geht aus von (1.8) und der Gruppeneigenschaft für W:

$$\begin{aligned} \delta y_f(t) \;=\; & G(t,t)^- \int_a^t W(t,s)\delta f(s)ds + \\ & + G(t,t)^+ \int_t^b W(t,s)\delta f(s)ds \end{aligned}$$

Interpretation nach Kapitel A.1.2. (1.23):

$$\begin{aligned} &\text{a)} \quad \delta y_a(t) := \int_a^t W(t,s)\delta f(s)ds \\ &\qquad \text{Störung bei Integration von } a \to t \text{(vorwärts)} \\ &\text{b)} \quad \delta y_b(t) := \int_t^b W(t,s)\delta f(s)ds \\ &\qquad \text{Störung bei Integration von } b \to t \text{(rückwärts)} \end{aligned} \tag{1.11}$$

Außerdem folgt aus (1.9):

$$\begin{aligned} &\text{a)} \quad G(t,t)^- - G(t,t)^+ = I \\ &\text{b)} \quad \|G(t,t)^-\| + \|G(t,t)^+\| \ge 1 \end{aligned} \tag{1.12}$$

Beweis:

$$E(t)^{-1} \underbrace{\left(AW(a,t) - (-BW(b,t))\right)}_{E(t)} = I$$

$$1 = \|I\| = \|G(t,t)^- - G(t,t)^+\| \le \|G(t,t)^-\| + \|G(t,t)^+\|$$

■

Dies führt auf:

$$\begin{aligned} &\text{a)} \quad \|\delta y_f(t)\| \le \bar{\kappa}_f(t) \max\{\|\delta y_a(t)\|, \|\delta y_b(t)\|\} \\ &\qquad \text{wobei} \\ &\text{b)} \quad \bar{\kappa}_f(t) := \|G(t,t)^-\| + \|G(t,t)^+\| \ge 1 \end{aligned} \tag{1.13}$$

Die Hoffnung wäre, mit dieser punktweisen Definition feiner abschätzen zu können. Für $|\cdot|_\infty$ erhält man

$$\begin{aligned} &\text{a)} \quad |\delta y_f|_\infty \leq \bar{\kappa}_f \max_{t\in[a,b]} \{\|\delta y_a(t)\|, \|\delta y_b(t)\|\} \\ &\text{b)} \quad \bar{\kappa}_f := \max_{t\in[a,b]} \left[\|G(t,t)^-\| + \|G(t,t)^+\|\right] \geq 1 \end{aligned} \tag{1.13'}$$

2 Schießverfahren

Die hier beschriebenen Methoden führen die Lösung des Randwertproblems zurück auf die Lösung einer Folge von Anfangswertproblemen. Dabei kann der vergleichsweise hohe Entwicklungsstand von Anfangswert-Lösern ausgenutzt werden (vgl. Kap. A, B).

2.1 Einfachschießen

(single shooting)

Man löst die Differentialgleichung

$$y' = f(y)\,, \quad y(a) := x \tag{2.1}$$

zu geschätzten Anfangswerten (soweit nicht bekannt). Sei $y(t|x)$ die Lösung dieses Anfangswertproblems.

Beispiel: "Artillerie-Problem" $\rightarrow$ Name

$$y'' = \ldots, \quad y(a) := y_a\,, \quad y(b) := y_b \text{ zu erzielen}$$
$$x := y'(a) \in \mathbb{R}^1 \text{ unbekannter Anstellwinkel}$$

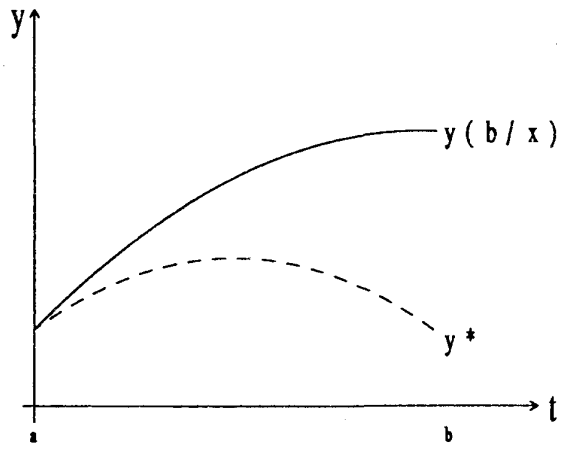


Bild D.2 Variation der Anfangssteigung bei Einfachschießen

Im Beispiel ist x derart zu bestimmen, daß

$$y(b|x) - y_b = 0 \text{ (eine nichtlineare Gleichung)}$$

Allgemeine Bedingung:

$$r(x, y(b|x)) = 0 \tag{2.2}$$

Funktionalmatrix dieses im allgemeinen nichtlinearen Gleichungssystems am Lösungspunkt $x^* = y^*(a)$:

$$\frac{\partial r}{\partial x}\Big|_{x^*} = A^* + B^* W^*(b,a) = E^*(a)\,. \tag{2.3}$$

Falls für die Lösung y^* des Randwertproblems die Bedingungen von Satz 1.1 (lokale Eindeutigkeit) gelten, so hat auch das nichtlineare Gleichungssystem (2.2) eine lokal eindeutige Lösung x^*.

Newton-Verfahren für (2.2):

Man benötigt eine Approximation der Jacobi-Matrix

$$\frac{\partial r}{\partial x}\Big|_{x^k} = A + BW(b,a)\Big|_{y=y(t|x^k)} =: E^k(a) \tag{2.3'}$$

(I) *semi-analytische Differentiation:*

A, B durch analytische Differentiation von r , $W(b,a)$ durch numerische Lösung der Variationsgleichung: benötigt analytischen Ausdruck $f_y(y)$.

(II) *externe numerische Differentiation:*

häufig in älteren Codes, etwas empfindlich.

(III) *interne numerische Differentiation:*

vgl. (2.11.b) - schneller, verläßlicher.

Newton-Iteration:

$$E^k(a)\Delta x^k = -r(x^k, y(b|x^k)) \tag{2.4}$$

Falls x^0 nicht "hinreichend nahe" an x^*: Dämpfungsstrategie oder/und Homotopiemethode (vgl. DEUFLHARD [34]).

Nachteile des Einfachschießens:

(I) Es ist häufig unmöglich, die Trajektorie $y(t;x)$ über das ganze Intervall zu berechnen. Grund: Bei nichtlinearer Differentialgleichung hängt die Position von Singularitäten ab von den Anfangswerten (*bewegliche Singularitäten*); falls also der Anfangswert "schlecht geschätzt" wird, ist y evtl. unbeschränkt in $[a,b]$.

(II) **Stabilitätsprobleme:**
Auch bei gutkonditionierten Randwertproblemen kann das zugehörige Anfangswertproblem schlechtkonditioniert sein.

Beispiel:

$y'' - \lambda^2 y = 0\ ,\ \ y(a) = y_a,\ y(b) = y_b$

Kondition (1.6) des Randwertproblems: $\bar{\rho} \doteq \lambda$ für $\lambda \to \infty$

(vgl. Aufgabe 36)

Kondition (1.22), Kap. A.1.2., des Anfangswertproblems:

$$\sigma(a,b) = \max_{t\in[a,b]} \|W^*(t,a)\| \sim \lambda e^{\lambda(b-a)} \text{ für } \lambda \to \infty$$

Zusammenfassung: Schießverfahren ist häufig äußerst empfindlich gegenüber der Wahl der Anfangswerte.

Bemerkungen :

1. In manchen Beispielen gilt

$$\sigma(b,a) \ll \sigma(a,b) \tag{2.5}$$

   Dann empfiehlt sich "Rückwärtsschießen".

2. Für *steife* Differentialgleichungen ist "Vorwärtsschießen" klar ausgezeichnet, da $\sigma(a,b) \ll \sigma(b,a)$ (vgl. Kap. B.).

3. Für *asymptotische* Differentialgleichungen ($T \to \infty$) ist im allgemeinen "Rückwärtsschießen" ausgezeichnet (vgl. Kap. D.6.).

4. Bei sogenannten *Grenzschichtproblemen* gilt häufig

$$\sigma(b,a) \approx \sigma(a,b) \gg 1\ ,$$

   also *keine* brauchbare Richtung zunächst (vgl. Kap. D.7.).

2.2 Mehrzielmethode für 2-Punkt-Randwertprobleme

(Mehrfachschießen, multiple shooting method)
(MORRISON/RILEY /ZANCANARO 1962 [86], H.B. KELLER 1968 [69], BULIRSCH 1971 [10], STOER/BULIRSCH 1973 [108], DEUFLHARD 1972-76 [25,26,28] DEUFLHARD/BADER 1983 [35])
Hierbei wird $[a,b]$ unterteilt:

$$\Delta := \{a = t_1 < t_2 < \ldots < t_m = b\}\ ,\quad m > 2$$

Man löst nun $(m-1)$ *unabhängige* Anfangswertprobleme über den Teilintervallen $[t_j, t_{j+1}]$. Als Anfangswert wählt man Schätzungen $x_j \in \mathbb{R}^n$ für die unbekannten Werte $y(t_j)$.

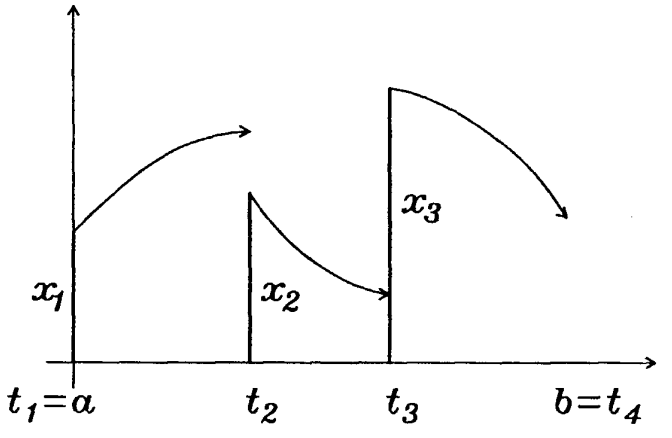


Bild D.3 Prinzip der Mehrzielmethode

Man erhält $(m-1)$ *Teiltrajektorien.*

$$\begin{aligned} &y(t|x_j), && t \in [t_j, t_{j+1}] \\ &y(t_j|x_j) &:= x_j\ , & \text{Anfangswert} \end{aligned} \tag{2.6}$$

Neben den Randbedingungen müssen zusätzlich Verheftungsbedingungen gelten.

Verheftungsbedingungen: $j = 1, \ldots, m-1$
(=Stetigkeitsbedingungen)

$$F_j(x_j, x_{j+1}) := y(t_{j+1}|x_j) - x_{j+1} = 0 \tag{2.7.a}$$

Randbedingungen:

$$F_m(x_1, x_m) := r(x_1, x_m) = 0 \tag{2.7.b}$$

(2.7) heißt *zyklisches* (nichtlineares) Gleichungssystem - siehe Bild D.4:

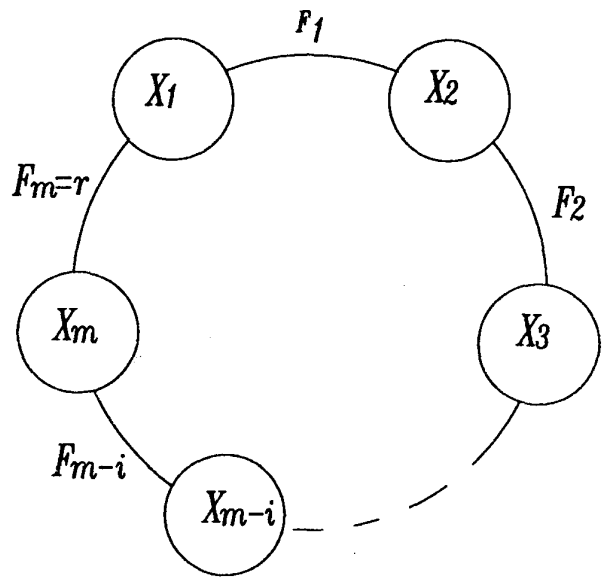


Bild D.4 Zyklisches Gleichungssystem

Bemerkung: Aus technischen Gründen der Implementierung (und der Newton-Iteration) empfiehlt es sich *nicht*, die Ersetzung $x_m \equiv y(b|x_{m-1})$ vorzunehmen. Kurzschreibweise:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \in \mathbb{R}^{n \cdot m} \quad , \quad F(x) = \begin{pmatrix} F_1(x_1, x_2) \\ \vdots \\ F_m(x_1, x_m) \end{pmatrix}$$

Durch dieses Vorgehen lassen sich gewisse Nachteile des Schießverfahrens (single shooting) vermeiden.

"Bewegliche Singularitäten" : sie lassen sich bei der Mehrzielmethode durch geeignete Knotenwahl abfangen.

Beispiel:

$y'' = \lambda \sin(\lambda y)$
$y(0) = 0, \quad y(1) = 1$
Singularität bei Anfangswertproblemen: $t_\infty \doteq \frac{1}{\lambda} \ln(8/|y_0'|)$
$\Rightarrow |y_0'| \dot{\leq} 8e^{-\lambda}$ um rechten Rand 1 zu erreichen! $(\lambda = 5 : |y_0'| \dot{\leq} 0.05)$

Stabilitätsverbesserung:

Kondition des Anfangswertproblems in $[t_j, t_{j+1}]$:

$$\sigma(t_j, t_{j+1})$$

Kondition für Mehrzielmethode:

$$\sigma_\Delta := \max_{t_j \in \Delta} \sigma(t_j, t_{j+1}) \tag{2.8}$$

Beispiel:

$y'' - \lambda^2 y = 0$, $y(a)$, $y(b)$ vorgegeben
$\kappa(a,b) \doteq e^{\lambda(b-a)}$
Sei $t_{j+1} - t_j = \dfrac{b-a}{10}$: $\sigma_\Delta \doteq \sqrt[10]{\sigma(a,b)}$
äußerst wirkungsvolle Reduktion!

Allgemein gilt nach Definition

$$\lim_{h \to 0} \sigma(t, t+h) = 1$$

Die Knoten t_j sind in etwa derart zu wählen, daß gilt:

$$\sigma(t_j, t_{j+1}) \leq \bar{\sigma} \quad (\bar{\sigma} \approx 10 - 10^2) \tag{2.9}$$

Eine *automatische* Knotenwahl sollte unter anderem die Bedingung (2.9) erfüllen.

Bemerkungen:

(1) Für *steife* Differentialgleichungen gilt $\sigma_\Delta \approx 1$
für *nichtsteife* Differentialgleichungen gilt $\sigma_\Delta = e^{L|\Delta|}$,
wobei $|\Delta| = \max_j |t_{j+1} - t_j|$,
L : Lipschitz-Konstante von f.

(2) Bei singulären Störungsproblemen ist $L \sim \dfrac{1}{\varepsilon}$,
damit zu *viele* Knoten!

(3) Integration immer in stabiler Richtung (bezüglich Anfangswerten):
(a) Steife Differentialgleichung in positiver t-Richtung
(b) Asymptotische Randwertprobleme in negativer t-Richtung

Newton-Verfahren für Mehrzielmethode

Zu lösen ist das lineare mn-System:

$$\begin{aligned} J(x^k)\Delta x^k &= -F(x^k) \\ x^{k-1} &:= x^k + \Delta x^k \end{aligned}$$

Blockzeilenweise:

$$\begin{aligned} G_1\Delta x_1 \quad -\Delta x_2 \qquad\qquad\qquad\qquad &= -F_1 \\ \ddots \qquad\qquad\qquad\qquad & \\ G_{m-1}\Delta x_{m-1} - \Delta x_m \quad &= -F_{m-1} \\ A\Delta x_1 \qquad\qquad +B\Delta x_m \quad &= -F_m = -r \end{aligned} \tag{2.10}$$

wobei

$$G_j := W(t_{j+1}, t_j)|_{y(t|x_j)} \text{ Wronski-Matrix}$$

$$A := \frac{\partial r}{\partial x_1} = \frac{\partial r}{\partial y(a)}\,, \quad B := \frac{\partial r}{\partial x_m} = \frac{\partial r}{\partial y(b)}$$

Berechnung der Matrizen:
A, B durch numerische Differentiation, Berechnung von G_j:

a) *semi-analytisch*
numerische Lösung der Variationsgleichung

$$G_j' = f_y(y)G_j\,, \quad G_j(t_j) = I \tag{2.11}$$

(Lösung von n Anfangswertproblem!)

b) *interne numerische Differentiation* (BOCK) [6] :
$f_y \longrightarrow \dfrac{\Delta f}{\Delta y}$
in Lösung der Variationsgleichung
(ebenfalls n Anfangswertproblem!)

c) *externe numerische Differentiation*:
in älteren Codes, etwas empfindlich.

d) *Rang-1-Approximationen* nach BROYDEN
Newton-Verfahren → quasi-Newton-Verfahren

Bei "schlechten" Startdaten x^0: Dämpfungsstrategie in Newton-Verfahren oder Homotopiemethoden (vgl. DEUFLHARD [26,27,28,34]).

3 Lösung der zyklischen linearen Gleichungssysteme

Zu lösen ist das Blocksystem (2.10). Dabei möchte man die Struktur der Nullblöcke möglichst erhalten (Speicherplatz!).

3.1 Block-Gauss–Elimination (="Condensing")

(STOER/BULIRSCH 1973)[108]

Illustration für $m = 3$:

$$\begin{array}{lrcl} (1) & G_1\Delta x_1 \quad -\Delta x_2 & = & -F_1 \;/\; G_2 \\ (2) & G_2\Delta x_2 - \Delta x_3 & = & -F_2 \\ (3) & A\Delta x_1 \; + B\Delta x_3 & = & -r \end{array}$$

$$\begin{array}{rcrcl} G_2\cdot(1)\to(1') & : & G_2G_1\Delta x_1 - G_2\Delta x_2 & = & -G_2F_1 \\ (1')+(2)\to(2') & : & G_2G_1\Delta x_1 - \Delta x_3 & = & -(F_2 - G_2F_1) \\ B(2')\to(2'') & : & BG_2G_1\Delta x_1 - B\Delta x_3 & = & -B(f_2 - G_2F_1) \\ (2'')+(3)\to(3') & : & \underbrace{(A+BG_2G_1)}_{=:E}\Delta x_1 & = & \underbrace{-r - B(F_2 - G_2F_1)}_{=:-u} \end{array}$$

$$G_2 \doteq W(t_3,t_2) = W(b,t_2)$$
$$G_1 \doteq W(t_2,t_1) = W(t_2,a)$$
$$\Rightarrow E \doteq A + BW(b,t_2)W(t_2,a)$$
für glatte Lösung y : $E \doteq A + BW(b,a)$

Formalisierung der Block-Elimination:

Satz 1 (DEUFLHARD 1972) [25]
Sei die Jacobi-Matrix der Mehrzielmethode

$$J = \begin{bmatrix} G_1 - I & & \\ \ddots & & \\ & G_{m-1} - I & \\ A & & B \end{bmatrix} \quad (mn,mn) - Matrix$$

und sei definiert

$$E := A + BG_{m-1}\ldots G_1 \qquad (n,n) - Matrix$$

Dann gilt:

$$\begin{aligned} a)\ \det(J) &= \det(E) \\ b)\ LJR &= S \\ J^{-1} &= RS^{-1}L \end{aligned} \tag{3.1}$$

mit

$$L := \begin{bmatrix} BG_{m-1}\dots G_2, & B & I \\ -I & & \\ \ddots & & \\ & -I & 0 \end{bmatrix}$$

$$R^{-1} = \begin{bmatrix} I & & \\ -G_1, I & & \\ \ddots & & \\ & -G_{m-1}, & I \end{bmatrix}, S := \begin{bmatrix} E & & \\ & I & \\ & & \ddots \\ & & I \end{bmatrix}$$

Beweis: Zu a) $\det(S) := \det(E), \det(L) = \det(R) = 1$, zu b) Nachrechnen. ∎

Interpretation

$$E|_{x^*} = E^*(a) \qquad \text{vergleiche Satz 1.1} \tag{3.2}$$

Sensitivitätsmatrix

Falls Randwertproblem lokal eindeutig lösbar, so ist das Gleichungssystem (2.10) ebenfalls in Umgebung von x^* eindeutig lösbar.

Zugehöriger Algorithmus:

$$
\begin{aligned}
&\text{a)} \quad E := A + BG_{m-1} \ldots G_1 \\
&\qquad u := r + B\left[F_{m-1} + G_{m-1}F_{m-2} + \ldots + G_{m-1} \ldots G_2 F_1\right] \\
&\qquad \text{rekursive Berechnung} \\
&\text{b)} \quad E\Delta x_1 = -u \\
&\qquad \text{lineares } (n,n) - \text{Gleichungssystem} \\
&\qquad \text{(QR-Zerlegung mit Rangentscheidung)} \\
&\qquad \rightarrow \text{Bezeichnung "condensing"} : (N,N) \rightarrow (n,n) \\
&\text{c)} \quad \Delta x_{j+1} = G_j \Delta x_j + F_j \quad j = 1, \ldots, m-1 \\
&\qquad \text{explizite Rekursion}
\end{aligned}
\tag{3.3}
$$

Speicherplatz: $\sim m \cdot n^2$

3.2 Nachiteration bei Block-Gauss-Elimination

(iterative refinement sweeps: DEUFLHARD/BADER 1983 [35])

Anstelle der Korrekturen Δx_j erhält man fehlerbehaftete Korrekturen $\Delta\tilde{x}_j \equiv \Delta\tilde{x}_j^0$. Sei $\nu = 0, 1, \ldots$ Iterationsindex. Man berechnet mit *gleicher* Mantissenlänge die Residuen:

$$
\begin{aligned}
&\text{a)} \quad dr^\nu := fl\left\{r + A\Delta\tilde{x}_1^\nu + B\Delta\tilde{x}_m^\nu\right\} \\
&\text{b)} \quad dF_j^\nu := fl\left\{G_j\Delta\tilde{x}_j^\nu + F_j - \Delta\tilde{x}_{j+1}\right\} \quad j = 1, \ldots, m-1
\end{aligned}
\tag{3.4}
$$

Gleichungssystem für Korrekturen könnte wiederum mit Algorithmus (3.3) gelöst werden. Standard-Nachiteration wäre:

$$
\begin{aligned}
&\text{a)} \quad E \text{ bleibt wie bisher} \\
&\qquad du^\nu := dr^\nu + B\left[dF_{m-1}^\nu + G_{m-1}dF_{m-2}^\nu + \ldots G_{m-1}G_2 dF_1^\nu\right] \\
&\text{b)} \quad Edx_1^\nu = -du^\nu \\
&\qquad \text{lineares } (n,n)\text{-Gleichungssystem} \\
&\qquad \text{mit vorhandener QR-Zerlegung von } E \\
&\text{c)} \quad dx_{j+1}^\nu = G_j dx_j^\nu + dF_j^\nu \quad j = 1, \ldots, m
\end{aligned}
\tag{3.5}
$$

$$
\begin{aligned}
\Delta\tilde{x}_j^{\nu+1} &:= \Delta\tilde{x}_j^\nu + d\tilde{x}_{j+1}^\nu \quad j = 1, \ldots, m \\
d\tilde{x}_j^\nu &:= fl(dx_j^\nu)
\end{aligned}
\tag{3.6}
$$

Eine detaillierte (elementweise) Rundungsfehleranalyse [35] zeigt, daß die Standard-Nachiteration konvergiert unter der Bedingung

$$\begin{aligned} &\text{a)} \quad \varepsilon\sigma(a,b)g(n,m) \ll 1 \\ &\qquad \varepsilon : \text{ relative Maschinengenauigkeit} \\ &\text{b)} \quad g(n,m) := (m-1)(2n+m-1) \end{aligned} \tag{3.7}$$

$\hookrightarrow$ ungeeignet : $\sigma(a,b)$ gehört zu Einfachschießen

Abhilfe: "iterative refinement sweeps"
Sei für ein $\bar{\varepsilon} \leq$ eps (vorgeschriebene relative Genauigkeit) der *Sweep-Index* j_ν definiert über die Beziehung

$$\begin{aligned} &\text{a)} \quad \| d\tilde{x}_j^\nu \| \leq \bar{\varepsilon} \quad j = 1, \ldots, j_\nu \\ &\qquad \text{Dann setzt man:} \\ &\text{b)} \quad dF_j^\nu := 0 \quad j = 1, \ldots, j_\nu - 1 \\ &\qquad \text{(Maschinen-Null!)} \end{aligned} \tag{3.8}$$

Wahl von $\bar{\varepsilon}$: Eine Nachiteration auf linearem n-System

$$E\Delta\tilde{x}_1 - u \doteq 0$$

liefert

$$\bar{\varepsilon} := \| d\tilde{x}_1^0 \| \tag{3.8.c}$$

Dieser Wert ist zu vergleichen mit eps, der vom Benutzer verlangten relativen Genauigkeit.
Falls $j_0 > 1$ (also $\varepsilon \dot{\leq} \bar{\varepsilon} \dot{<}$ eps): unter der Bedingung

$$\varepsilon\sigma_\Delta g(n,m) < 1 \,, \tag{3.9.a}$$

wobei σ_Δ aus (2.8), läßt sich zeigen:

$$j_{\nu+1} \geq j_\nu + 1 \tag{3.9.b}$$

Daraus folgt:

$$\nu \leq m - 1 \tag{3.9.c}$$

Falls $j_0 = 0$ ($\bar{\varepsilon} \dot{\geq}$ eps):
Umschaltung auf "globale" Lösung des linearen Blocksystems (vgl. Kap.3.3).

Grenzen der Nachiteration (ASCHER)[2] erläutert am Beispiel:

$$A := \begin{bmatrix} A_1 \\ 0 \end{bmatrix}, \quad B := \begin{bmatrix} 0 \\ B_2 \end{bmatrix}$$
$$A_1 : (n - \bar{n}, n), \quad B_2 : (\bar{n}, n)$$
$$\mathrm{rg}\,(A_1) = n - \bar{n}, \quad \mathrm{rg}\,(B_2) = \bar{n}$$

Falls *dominante* Lösung $(a \to b)$ existiert, gilt:

$$\mathrm{rg}\,(W(b,a)) \doteq k < n\,.$$

Damit folgt für $E = A + BW(b,a)$:

$$\mathrm{rg}\,(E) \doteq \min(n, n - \bar{n} + k)\,.$$

Falls

$$\begin{aligned} \bar{n} > k \quad &\Longrightarrow \quad \mathrm{rg}\,(E) < n \\ &\Longrightarrow \quad \bar{\varepsilon} > \mathrm{eps} \Longrightarrow j_0 := 0\,. \end{aligned}$$

Bemerkung: Nachiteration liefert zusätzliche Schätzungen:

$$\mathrm{cond}(E) \doteq \frac{||d\tilde{x}_1^0||}{||\Delta\tilde{x}_1||}/epmach \tag{3.10.a}$$

$$\sigma_\Delta \doteq \max_j \frac{||d\tilde{x}_{j+1}^0||}{||d\tilde{x}_j^0||} \tag{3.10.b}$$
$$\hookrightarrow \quad \mathrm{TOL} \doteq \mathrm{eps}/\sigma_\Delta$$

Programm: BVPSOL (DEUFLHARD/BADER 1983[35])

3.3 Globale Lösung

Bemerkung: Für Spezialfälle läßt sich zeigen (MATTHEIJ [85])

$$\text{cond}(J) \longrightarrow m \cdot \bar{\rho} \text{ für } m \gg 1 \tag{3.11}$$

Programme:

RWPM (HERMANN/BERNDT 1982 [61])

Gauss-Elimination mit Spaltenpivoting und einer Nachiteration (SKEEL[104])

$\hookrightarrow$ Speicherplatz $\sim 3m \cdot n^2$

BOUNDSCO (OBERLE 1987 [88])

Householder-Transformationen von rechts, keine Nachiteration

$\hookrightarrow$ Speicherplatz $\sim 5 \cdot m \cdot n^2$

BVPSOG (DEUFLHARD/KUNKEL 1985)

Sparse-Elimination mit Programm MA28 (DUFF, Harwell)
angewendet in Block-Variante
$\hookrightarrow$ Speicherplatz $\sim s \cdot m \cdot n^2$, $s \approx 2 - 3$ (abhängig von fill-in)
Dieser Code wird angesprochen, falls BVPSOL seine eigenen
Anwendungsgrenzen in einem Beispiel erkannt hat ($j_0 = 0$ in (3.8)).

4 Varianten für allgemeinere Randwertprobleme

Die Mehrzielmethode gestattet einfache Anpassung an relativ allgemeine Randwertprobleme.

4.1 Mehrpunkt Randwertprobleme

Am Beispiel 3-Punkt-Randwertproblem vorgeführt:

$$r(y(a)\ ,\ y(\tau)\ ,\ y(b)) = 0 \quad \tau \in]a,b[\tag{4.1.a}$$

Zugehörige Sensitivitätsmatrix (vgl. Aufgabe 38):

$$\begin{aligned} E(t) &= AW(a,t) + R_\tau W(\tau,t) + BW(b,t) \\ R_\tau &= \frac{\partial r}{\partial y(\tau)} \end{aligned} \tag{4.1.b}$$

Sei τ fest: man wählt τ als Knoten der Mehrzielmethode. Dadurch wird nur die *letzte Blockzeile* der Jacobi-Matrix verändert zu

$$[A, 0, \ldots, 0, R_\tau, 0, \ldots, 0, B] \tag{4.1.c}$$

Übertragung auf mehr als eine innere Punkt-Bedingung trivial. Auflösung des linearen Systems für Newton-Korrektur entsprechend Kapitel 3.

4.2 Parameterabhängige Randwertprobleme

(ENGLAND 1976 [43])

In technischen Anwendungen tritt häufig folgender Problemtyp auf (z. B. nichtlineare Eigenwertprobleme oder Randwertprobleme mit "freiem" Ende):

$$y' = f(y;p) \quad p \in \mathbb{R}^q\ ,\ \ y \in \mathbb{R}^n \tag{4.2.a}$$

$$r(y(a), \ldots, y(b); p) = 0 \quad r \in \mathbb{R}^{n+q} \tag{4.2.b}$$

Die q "überzähligen" Randbedingungen dienen dazu, die q unbekannten Parameter $p^T := (p_1, \ldots, p_q)$ zu bestimmen. Entsprechende Erweiterung der

Jacobi-Matrix (gegenüber 4.1.c):

$$J = \begin{bmatrix} G_1 & - & I & \cdot & \cdot & P_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & G_{m-1} & -I & P_{m-1} \\ R_1 & & & R_{m-1} & R_m & P_m \end{bmatrix} (mn+q\ ,\ mn+q) \qquad (4.2.c)$$

Bezeichnung:

$$\begin{aligned} G_j &:= W(t_{j+1},t_j)\big|_{y(t|x_j,p)} && (n,n)\ \ j=1,\ldots,m-1 \\ R_j &:= \left.\frac{\partial r}{\partial y(t_j)}\right|_{\ldots} && (n+q,n)\ \ j=1,\ldots,m \\ P_j &:= P(t_{j+1},t_j)\big|_{y(t|x_j,p)} && (n,q)\ \ j=1,\ldots,m-1 \\ P_m &:= \left.\frac{\partial r}{\partial p}\right|_{\ldots} && (n+q,q) \end{aligned}$$

Berechnung der G_j, R_j wie gehabt. Berechnung der (n,q)-Matrizen P_j:
(I) *Semi-analytische Differentiation:*
$P(t,t_j)$ ist Lösung der verallgemeinerten Variationsgleichung zu speziellem Anfangswert:

$$\frac{dP(t,t_j)}{dt} = f_y(y(t|x_j,p);p)P(t,t_j) + f_p(y(t|x_j,p);p) \qquad (4.3.a)$$

$$P(t_j,t_j) = 0 \qquad (4.3.b)$$

(II) *Interne Differentiation* (BOCK [6]) : $f_p \longrightarrow \dfrac{\Delta f}{\Delta p}$.

4.3 Parameteridentifizierung bei Differentialgleichungen

(BOCK 1981, 1985 [6,7])

Inverses Problem: Zu vorgegebenen Meßpunkten

$$(\tau_i, z_i)\quad i=1,\ldots,\bar{m}, \bar{m} > n+q$$

bestimme man eine Trajektorie $y(t|p)$ derart, daß y Lösung von

$$y' = f(y;p)\ ,\ \ p \in \mathbb{R}^q \qquad (4.4.a)$$

und daß gilt:

$$\frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} ||z_i - y(\tau_i|p)||_2^2 = \min \ . \tag{4.4.b}$$

Anpassung der Mehrzielmethode:
man wählt Knoten $\{t_1, \ldots, t_m\} \subseteq \{\tau_1, \ldots, \tau_{\bar{m}}\}$ mit im allgemeinen $m \ll \bar{m}$.

Beispiel:

Unbekannte: x_1, x_2, x_3, x_4, p

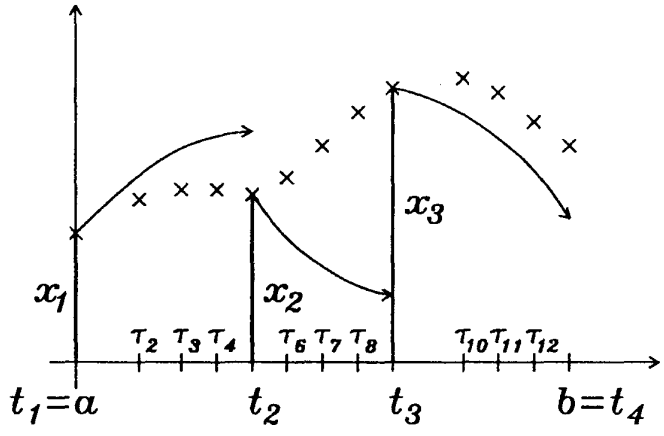


Bild D.5 Mehrzielmethode bei inversem Problem

Mit der Bezeichnung:

$$r(x_1, ..., x_m; p) := \begin{bmatrix} z_1 - y(\tau_1|x_1, p) \\ \vdots \\ z_{\bar{m}-1} - y(\tau_{\bar{m}-1}\ x_{m-1}, p) \\ z_{\bar{m}} - x_m \end{bmatrix} \tag{4.5}$$

erhält man ein Ausgleichsrandwertproblem:
Verheftungsbedingungen: $j = 1, \ldots, m-1$

$$F_j(x_j, x_{j+1}; p) := y(t_{j+1}|x_j, p) - x_{j+1} = 0 \tag{4.6.a}$$

Randbedingungen:

$$||r(x_1, \ldots, x_m; p)||_2 = \min \ . \tag{4.6.b}$$

Sei wieder definiert:

$$F(x) := \begin{bmatrix} F_1(x_1, x_2; p) \\ \vdots \\ F_{m-1}(x_{m-1}, x_m; p) \\ r(x_1, \ldots, x_m; p) \end{bmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \\ p \end{pmatrix} \in \mathbb{R}^{n \cdot m + q}$$

Jacobi-Matrix hat die Gestalt (4.2.c), wobei nun die Blöcke $R_1, \ldots, R_m, P_m$ jeweils $\bar{m} \gg m$ Zeilen haben.

Zur Lösung konstruiert man die spezielle Gauss-Newton-Iteration

$$\Delta x^k := -J(x^k)^- F(x^k) \tag{4.7}$$

wobei J^- verallgemeinerte Inverse (vgl. DEUFLHARD 1972 [25]), für die gilt:

$$J^- J J^- = J^- \text{ (äußere Inverse)}$$

$\hookrightarrow$ lokale Konvergenz, falls $\kappa(x^*) < 1$.

Zugehörige Block-Gauss-Elimination:

Der Condensing-Algorithmus kann auf elementare Weise verallgemeinert werden. Sei allgemein zu lösen:

$$\begin{aligned} G_1 \Delta x_1 - \quad \Delta x_2 + \qquad\qquad & P_1 \Delta p = -F_1 \\ G_2 \Delta x_2 - \Delta x_3 + \qquad\qquad & P_2 \Delta p = -F_2 \\ \ddots \qquad\qquad & \\ G_{m-1} \Delta x_{m-1} - \Delta x_m + & P_{m-1} \Delta p = -F_{m-1} \end{aligned} \tag{4.8.a}$$

$$||R_1 \Delta x_1 + R_2 \Delta x_2 + R_m \Delta x_m + P_m \Delta p + r||_2 = \min \ . \tag{4.8.b}$$

Rekursiv berechnet man:

$$\begin{aligned} \bar{P}_m &:= P_m, \bar{R}_m := R_m \\ j &= 1, \ldots, m-1: \\ \bar{P}_j &:= \bar{P}_{j+1} + \bar{R}_{j+1} P_j \\ \bar{P}_j &:= R_j + \bar{R}_{j+1} G_j \\ \Longrightarrow \quad P &:= \bar{P}_1, E := \bar{R}_1 \end{aligned} \tag{4.9.a}$$

u wie (3.3.a)

$$||E \Delta x_1 + P \Delta p + u||_2 = \min \ . \tag{4.9.b}$$

Dies ist äquivalent zu:

$$\begin{pmatrix} \Delta x_1 \\ \Delta p \end{pmatrix} = -[E, P]^+ u \qquad (4.9.'b)$$

Moore-Penrose-Pseudoinverse

Explizite Rekursion:

$$\Delta x_{j+1} = G_j \Delta x_j + P_j \Delta p + F_j \qquad j = 1, \ldots, m-1 \,. \qquad (4.9.c)$$

Bemerkung: Refinement-Sweeps möglich.

Programm: PARFIT (BOCK)

4.4 Periodische Lösungen autonomer Differentialgleichungen

(DEUFLHARD 1984 [31])

Idee: Im autonomen Fall hat man zusätzlich zu den üblichen Variablen $(x_1, \ldots, x_m)$ noch die Variable T. Die Vielfachheit der Lösungen bezüglich Zeittranslation interessiert nicht: es reicht eine beliebige Lösung.

Einfachschießen:

Trajektorie	:	$y(t\|x,T)\,,\quad z := (x,T) \in \mathbb{R}^{n+1}$
Anfangswertproblem	:	$y' = f(y)\,,\quad y(0) = x$
Randbedingung	:	$r(x, y(T\|x,T)) := y(T\|x,T) - x = 0\,.$

Funktionalmatrix:

$$r_z = [r_x, r_T] = [E(0)\,,\quad f(y(T))]\,. \tag{4.10}$$

Falls Randbedingung erfüllt:

$$r_z|_{z^*} = [E^*(0)\,,\quad f(y^*(0))]$$

Es gilt (vgl. Kap. A.4, Aufgabe 5: $E(0) = W(T,0) - I$):

$$E(0)f(y(0)) = 0\,. \tag{4.11}$$

Falls $E(0)$ nur einen *einfachen* Eigenwert 0 hat und $f(y(0)) \neq 0$, dann gilt:

$$\operatorname{rg}\,[E(0)\,,\quad f(y(0))] = n\,. \tag{4.12}$$

Anstelle der Newton-Iteration benutzt man eine Gauss-Newton-Iteration:

$$\Delta z^k := -(r_z|_{z^k})^+ r(z) \tag{4.13.a}$$

$$z^{k+1} := z^k + \Delta z^k\,. \tag{4.13.b}$$

Die Beziehung (4.13.a) ist äquivalent zu:

$$\begin{pmatrix} \Delta x^k \\ \Delta T^k \end{pmatrix} = -[E^k(0), f(x^k)]^+ r(x^k, y(T^k|x^k, T^k)) \tag{4.13'.a}$$

Dieses Iterationsverfahren konvergiert *lokal quadratisch* wie das gewöhnliche Newton-Verfahren.

Mehrzielmethode: Zeittransformation:

$$\tau := \frac{t}{T} \in [0,1]$$

Knoten: $0 = \tau_1 < \tau_2 < \ldots < \tau_m = 1$, $\Delta\tau_j = \tau_{j+1} - \tau_j$.
Stetigkeitsbedingungen:

$$F_j(x_j, x_{j+1}; T) := y(\tau_{j+1}|x_j, T) - x_{j+1} = 0 \qquad (4.14.a)$$

Randbedingungen:

$$r(x_1, x_m) := x_m - x_1 = 0 \qquad (4.14.b)$$

Man fixiert während der gesamten Iteration

$$x_m := x_1$$

Damit formal $r \equiv 0$ gesichert.
Bezeichnung:

$$G_j \text{ wie bisher}$$
$$g_j := \frac{\partial F_j}{\partial T} = \frac{\partial y(\tau_{j+1}|x_j, T)}{\partial T}$$

Es gilt wegen Autonomie der Differentialgleichung:

$$y(\tau_{j+1}|x_j, T) = x_j + \int_{t=0}^{T\Delta\tau_j} f(y(t))dt \qquad (4.15.a)$$

Daraus folgt:

$$g_j = \Delta\tau_j \cdot f(y(\tau_{j+1}|x_j, T)) \qquad (4.15.b)$$

Man hat also analog zu (4.13) das folgende unterbestimmte Gleichungssystem zu lösen:

$$\begin{array}{lllll} \text{a)} & G_1\Delta x_1 - & \Delta x_2 & +g_1 \cdot \Delta T & = -F_1 \\ & \ddots & & & \\ & G_{m-1}\Delta x_{m-1} - \Delta x_m & +g_{m-1}\Delta T & & = -F_{m-1} \\ \text{b)} & -\Delta x_1 \qquad + \Delta x_m & = 0 & & \end{array} \qquad (4.16)$$

Bemerkung. b) nur für "iterative refinement sweeps".
Programm: PERIOD (DEUFLHARD 1984 [31])

5 Probleme der optimalen Steuerung

(optimal control problems, BOLZA [8], BRYSON/HO [9], BULIRSCH [10,11])

Anwendungen:

Optimale Steuerung zeitabhängiger Prozesse (unter anderem Wirtschaftsmodelle, Raumfahrt, Sonnenhäuser, Optimierung chemischer Indudstrieprozesse) läßt sich darstellen als spezielles Randwertproblem für Differentialgleichungen.

5.1 Grundaufgabe der Variationsrechnung

(JOHANN BERNOULLI 1696 [5])

Zu minimieren sei ein Funktional

$$I[\varphi] := \int_a^b f(t, \varphi, \varphi')dt \tag{5.1}$$

über einer vorgegebenen Klasse von Vergleichsfunktionen, etwa

$$\varphi \in K := \{y \in C^1[a,b] \mid y(a) = y_a, y(b) = y_b\} . \tag{5.2}$$

Im klassischen Variationsproblem gilt:

$$K \text{ offen}$$

Idee von LAGRANGE (1755): Sei φ_0 die Lösung des Variationsproblems:

$$I[\varphi_0] \leq I[\varphi] , \quad \varphi \in K . \tag{5.3}$$

Man betrachte die folgende 1-parametrige Einbettung:

$$\varphi := \varphi_0 + \varepsilon\delta\varphi$$

wobei gilt:

$$\forall \varepsilon : |\varepsilon| \leq \bar{\varepsilon} \Rightarrow \varphi \in K(\text{offen!})$$

Bezeichnung: $J(\varepsilon) := I[\varphi_0 + \varepsilon\delta\varphi]$ Dann ist notwendige Bedingung für (5.3), daß $J(0)$ inneres Minimum bezüglich ε ist. Es muß also gelten:

$$\begin{array}{lll} \text{a)} & J'(0) = 0 & \text{“1. Variation”} \\ \text{b)} & J''(0) > 0 & \text{“2. Variation”} \end{array} \tag{5.4}$$

Satz 1 *Falls eine Lösung φ_0 des Variationsproblems (5.3) existiert, so muß gelten:*

$$\varphi_0 \in C^2[a,b] \tag{5.5.a}$$

Euler-Lagrange-Gleichungen:

$$f_\varphi - \frac{d}{dt} f_{\varphi'} = 0 \qquad (1.\ Variation) \tag{5.5.b}$$

Legendre-Clebsch-Bedingung:

$$0 < f_{\varphi'\varphi'}(t, \varphi_0(t), \varphi_0'(t)) \ \ \forall t \in [a,b] \qquad (2.\ Variation) \tag{5.5.c}$$

Beweisskizze: Nur Herleitung von (5.5.b) skizziert hier:

$$\begin{aligned} J'(\varepsilon) &= \int_a^b \Big[\frac{\partial f}{\partial \varphi}(t, \varphi_0 + \varepsilon\delta\varphi\ ,\ \varphi_0' + \varepsilon\delta\varphi')\delta\varphi + \\ &\quad + \frac{\partial f}{\partial \varphi'}(t, \varphi_0 + \varepsilon\delta\varphi'\ ,\ \varphi_0' + \varepsilon\delta\varphi')\delta\varphi' \Big] dt \\ J'(0) &= \int_a^b [f_\varphi(t, \varphi_0, \varphi_0')\delta\varphi + f_{\varphi'}(t, \varphi_0, \varphi_0')\delta\varphi']dt \end{aligned}$$

Partielle Integration des 2. Terms:

$$\begin{aligned} &\textstyle\int_a^b \ f_{\varphi'}(t, \varphi_0, \varphi_0')\delta\varphi' dt = \\ &= \ f_{\varphi'}(t, \varphi_0, \varphi_0')\delta\varphi\big|_a^b - \int_a^b \frac{d}{dt} f_{\varphi'}(t, \varphi_0, \varphi_0') \cdot \delta\varphi dt \end{aligned}$$

Da $\varphi, \varphi_0 \in K$, muß gelten:

$$\delta\varphi(a) = \delta\varphi(b) = 0$$

Daraus folgt:

$$0 = J'(0) = \int_a^b [f_\varphi(t, \varphi_0, \varphi_0') - \frac{d}{dt} f_{\varphi'}(t, \varphi_0, \varphi_0')]\delta\varphi(a) dt \tag{5.6}$$

Das Integral muß verschwinden für alle $\delta\varphi$ derart, daß $\varphi \in K$. Mit dem sogenannten *Fundamental-Lemma* der Variationsrechnung zeigt man dann, daß $[\ldots] \equiv 0$. Die Bedingung (5.5.a) wird ebenfalls damit gezeigt: sie ist wichtig, um φ'' in Euler-Lagrange-Gleichungen definieren zu können: man erhält nämlich

$$\frac{d}{dt} f_{\varphi'} = f_{\varphi' t} + f_{\varphi'\varphi}\varphi_0' + f_{\varphi'\varphi'} \cdot \varphi_0''$$

Auflösung nach φ'' ist also nur möglich, falls $f_{\varphi'\varphi'}(t, \varphi_0, \varphi_0') \neq 0$ für alle $t \in [a,b]$. Vorzeichen für Minimum (5.5.c). ∎

Erweiterung:

(I) Falls *mehrere* Funktionen, etwa $\varphi_1, \ldots, \varphi_k$, im Funktional vorkommen, so gilt die Euler-Lagrange-Gleichung für jede einzelne von ihnen:

$$f_{\varphi_i} - \frac{d}{dt} f_{\varphi_i'} = 0;\ i = 1, \ldots, k \tag{5.5'.b}$$

Entsprechend ist (5.5.b) zu ersetzen durch:

$$f_{\varphi'\varphi'} \text{ positiv–definite Matrix} \tag{5.5'.c}$$

(II) Falls eine Randbedingung, etwa $\varphi(b)$, *nicht* vorgeschrieben, so gilt in Beweis:

$$\delta\varphi(b) \text{ beliebig} \tag{5.7.a}$$

Trotzdem muß (5.6) gelten, das heißt

$$f_{\varphi'}(t, \varphi_0, \varphi_0')|_{t=b} = 0 \tag{5.7.b}$$

natürliche Randbedingung (auch: Transversalitätsbedingung)

Unabhängig von K erhält man also immer ein Randwertproblem.

Allgemeineres Variationsproblem

Anstelle von (5.1) hat man zu *minimieren*

$$\begin{aligned} I[\varphi] := \ & g(a, \tau, b, \varphi(a), \varphi(\tau^-), \varphi(\tau^+), \varphi(b)) + \\ & + \int_a^b f(t, \varphi, \varphi')dt \end{aligned} \tag{5.8}$$

wobei gilt:

$$\alpha \leq \tau \leq b\ ,\ \ a, \tau, b \text{ frei}\ ,\ \varphi(\tau^-) \neq \varphi(\tau^+) \text{ möglich}$$

g enthält zusätzliche Bedingungen.

Beispiel: Stufentrennung einer Trägerrakete

Man geht ähnlich vor wie bisher und definiert

$$J(\varepsilon) = I[\varphi_{\text{opt}} + \varepsilon\delta\varphi]\ ,\ a := a_{\text{opt}} + \varepsilon\delta a\ ,\ \ \tau := \ldots, b \ldots$$

Definition:

$$H(t, \varphi, \varphi') := -f + \varphi' f_{\varphi'} \tag{5.9}$$

Eine etwas umfangreichere Rechnung liefert (BULIRSCH 1971 [11]):

$$\begin{aligned} J'(0) &= \left[\frac{\partial g}{\partial a} + H^+\right]_a \delta a + \left[\frac{\partial g}{\partial b} - H^-\right]_b \delta b + \\ &+ \left[\frac{\partial g}{\partial \tau} - H^- + H^+\right]_\tau \delta\tau + \\ &+ \left[\frac{\partial g}{\partial \varphi(a)} - f^+_{\varphi'}\right]_a \delta\varphi(a) + \left[\frac{\partial g}{\partial \varphi(b)} + f^-_{\varphi'}\right]_b \delta\varphi(b) + \\ &+ \left[\frac{\partial g}{\partial \varphi(\tau^-)} + f^-_{\varphi'}\right] \delta\varphi(\tau^-) + \left[\frac{\partial g}{\partial \varphi(\tau^+)} - f^+_{\varphi'}\right] \delta\varphi(\tau^+) + \\ &+ \int_a^b \left[f_\varphi - \frac{d}{dt} f_{\varphi'}\right] \delta\varphi(t)dt = 0 \end{aligned} \tag{5.10}$$

Ähnlich wie im einfacheren Spezialfall erhält man daraus *natürliche* Randbedingungen, falls keine anderen Randbedingungen vorgeschrieben, etwa

$$\frac{\partial g}{\partial \varphi(\tau^-)} + f_{\varphi'}\Big(\tau^-, \varphi(\tau^-),\ \varphi'(\tau^-)\Big) = 0\,. \tag{5.11}$$

Falls *mehrere* Variable $\varphi_1, \ldots, \varphi_k$ vorliegen:

$$H := -f + \sum_{i=1}^{k} \varphi_i' f'_{\varphi_i} \tag{5.9'}$$

Alle übrigen Terme werden komponentenweise definiert, etwa $f_{\varphi_i'}$, $\dfrac{\partial g}{\partial \varphi_i(a)}$, ...

Beispiele:

(1) $a = 0$, τ kommt nicht vor, $b = T$ freie Endzeit. Dann muß gelten:

$$H(T, \varphi(T)\,,\ \varphi'(T)) = 0$$

(2) *Zweistufenrakete:*

$a = 0$, τ freier Trennungspunkt, T frei, φ: Masse der Rakete.

Zu minimieren: Treibstoffverbrauch $\varphi(0) - \varphi(T)$.

Nebenbedingungen:

$$\begin{aligned} h &:= \varphi(\tau^+) - \varphi(\tau^-) - K_1(\varphi(\tau^-) - \varphi(a)) - K_2 = 0 \\ &\quad \text{Stufenabtrennung} \\ \varphi'(t) &= -\beta \qquad (\beta > 0 : \text{ Schub}) \end{aligned}$$

Man koppelt die Nebenbedingungen an mittels Lagrange-Multiplikatoren l, λ, so daß man folgendes Variationsproblem erhält: Minimiere

$$\begin{aligned} I[\varphi] := \ & \varphi(0) - \varphi(T) + l \cdot h(\varphi(a), \varphi(\tau^-), \varphi(\tau^+)) + \\ & + \int_0^T [\lambda(t)(\varphi' + \beta) + F(\varphi)]dt \end{aligned} \tag{5.12}$$

Im Vergleich mit (5.8) erhält man:

$$g(\varphi(\tau^-),\varphi(\tau^+),\varphi(T)) := \varphi(0) - \varphi(T) + lh$$
$$f(\varphi,\lambda,\varphi') := \lambda(\varphi' + \beta) + F(\varphi)$$

Unter anderem sind folgende Variationen frei: $\delta\tau$, $\delta\varphi(\tau^-)$, $\delta\varphi(\tau^+)$, $\delta\varphi(T)$

Nebenrechnung:
$$f_{\varphi'} = \lambda(t)\,, \quad f_{\lambda'} = 0$$
$$H = -f + \varphi' f_{\varphi'} = -\beta\lambda(t) - F(\varphi)$$
$$\frac{\partial g}{\partial\varphi(\tau^-)} = l, \quad \frac{\partial h}{\partial\varphi(\tau^-)} = -l, \quad (1 + K_1)$$
$$\frac{\partial g}{\partial\varphi(\tau^+)} = l, \quad \frac{\partial g}{\partial\varphi(T)} = -1$$

Man erhält folgende Bedingung aus (5.10):

$$\begin{aligned} -H(\tau^-) + H(\tau^+) &= 0 && \delta\tau \neq 0 \\ -l(1+K_1) + \lambda(\tau^-) &= 0 && \delta\varphi(\tau^-) \neq 0 \\ l - \lambda(\tau^+) &= 0 && \delta\varphi(\tau^+) \neq 0 \\ -1 + \lambda(T) &= 0 && \delta\varphi(T) \neq 0 \end{aligned} \tag{5.13}$$

Elimination des Lagrange-Parameters l:

$$\lambda(T) - 1 = 0$$
$$\lambda(\tau^-) - (1 + K_1)\lambda(\tau^+) = 0$$

Man erhält also ein 3-Punkt-Randwertproblem, wobei τ zu bestimmen ist.

5.2 Steuerungsprobleme

Sei u Steuervariable, $u_0(t)$ die gesuchte optimale Steuerung, wobei im allgemeinen

$$u \in C^0 \text{ stückweise}\,, \quad u : [a,b] \to \mathbb{R}^k\,.$$

Typisches Problem der optimalen Steuerung:
Zu *minimieren* ist das Funktional über K

$$I[u] := \Phi(t, y(b)) \tag{5.14.a}$$

(*Mayer'sches Problem*)

unter den Nebenbedingungen (*"dynamisches System"*):

$$\begin{aligned} &y' = f(t,y,u) \quad y \in C^1 \text{ Zustandsvariable} \\ &y : [a,b] \to \mathbb{R}^n \end{aligned} \tag{5.14.b}$$

mit den *Anfangsbedingungen*

$$y(a) = y_a \tag{5.14.c}$$

und den *Endbedingungen* (hier separiert):

$$r(b, y(b)) = 0 \quad r : \mathbb{R}^{n+1} \to \mathbb{R}^p, p \leq n \tag{5.14.d}$$

Bemerkung: Häufig treten noch sogenannte *Zustandsbeschränkungen* auf in der Form

$$s_i(t, y(t)) \leq 0 \quad \forall t \in [a,b] .$$

Sie komplizieren die mathematische Behandlung außerordentlich; deswegen wurden sie hier weggelassen.
Im allgemeinen wird u partitioniert in der Form

$$u = \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} , \quad u^1 \in \mathbb{R}^m , \quad u^2 \in \mathbb{R}^q , \quad m + q = k$$

derart, daß u^1 nichtlinear, u^2 linear auftritt:

$$f(t,y,u) =: g(t,y,u^1) + h(t,y,u^1)u^2 . \tag{5.15}$$

Für u^2 müssen *Steuerbeschränkungen* gelten, etwa:

$$\begin{aligned} &\alpha_i \leq u_i \leq \beta_i \quad i = m+1, \ldots, k \\ &\alpha_i, \beta_i \in \mathbb{R} \end{aligned} \tag{5.14.e}$$

Zur Behandlung dieses Problems koppelt man die Differentialgleichung und die Randbedingungen an das Funktional:

$$\begin{aligned} \hat{I}[u,y,y',\lambda] \;&:=\; \Phi(t,y(b)) + \nu^T r \\ &+\; \int_a^b \left[\sum_{i=1}^n \lambda_i [y_i' - f_i(t,y,u)]\right] dt \end{aligned} \tag{5.16}$$

Zugehörige Hamiltonfunktion:

$$H(t,y,\lambda,u) := \sum_{i=1}^{n} \lambda_i f_i(t,y,u) \tag{5.17}$$

$\lambda(t) = (\lambda(t)_1, \ldots, \lambda(t)_n)$ adjungierte Variable
$\nu = (\nu_1, \ldots, \nu_p)$ Lagrange-Multiplikatoren
$(\nu_i' = 0)$

Minimumprinzip von PONTRJAGIN (1959) [93]

Sei v beliebige Steuerung unter den Nebenbedingungen (5.14.e), so muß gelten:

$$H(t,y,u_0,\lambda) = \min_v H(t,y,v,\lambda) \tag{5.18}$$

Aus der 1. Variation erhält man die sogenannten *kanonischen* Gleichungen.

$$y_i' = H_{\lambda_i} = f_i(t,y,u) \quad i = 1,\ldots,n \tag{5.19.a}$$

$$\lambda_i' = -H_{y_i} = -\sum_{j=1}^{n} \lambda_j \frac{\partial f_j}{\partial y_i}(t,y,u) \quad i = 1,\ldots,n \tag{5.19.b}$$

Die Maximierung bezüglich u zerfällt in zwei Teile:

(I) Bestimmung von u^1:

$$\begin{array}{ll} H_{u^1} = 0 & \text{“1. Variation”} \\ H_{u^1u^1} \text{ positiv (semi-) definit} & \text{“2. Variation”} \end{array} \tag{5.19.c}$$

In der Regel erhält man hieraus einen analytischen Ausdruck.

$$u^1 = u^1(t,y,\lambda) \tag{5.19'.c}$$

Einsetzen in (5.19.a,b).

(II) Bestimmung von u^2:
Die Partitionierung (5.15) führt auch zu einer Aufspaltung der Hamiltonfunktion:

$$H(t,y,\lambda,u^1(t,y,\lambda),u^2) =: H_0(t,y,\lambda) + \sum_{i=1}^{q} S_i(t,y,\lambda) u_{m+i} \tag{5.20}$$

bang-bang-Steuerung

Sei nun

$$S_i(t,y,\lambda) \not\equiv 0\,, \quad i = 1,\ldots,q \tag{5.21.a}$$

Dann folgt aus dem Minimumprinzip mit Steuerbeschränkung:

$$u_{m+i} = \begin{cases} \alpha_{m+i} & \text{für } S_i > 0 \\ \beta_{m+i} & \text{für } S_i < 0 \end{cases} \tag{5.21.b}$$

Die Funktionen S_i heißen *Schaltfunktionen.* In *Schaltpunkten* τ^i muß gelten:

$$S_i(\tau^i, y(\tau^i), \lambda(\tau^i)) = 0 \tag{5.21}$$

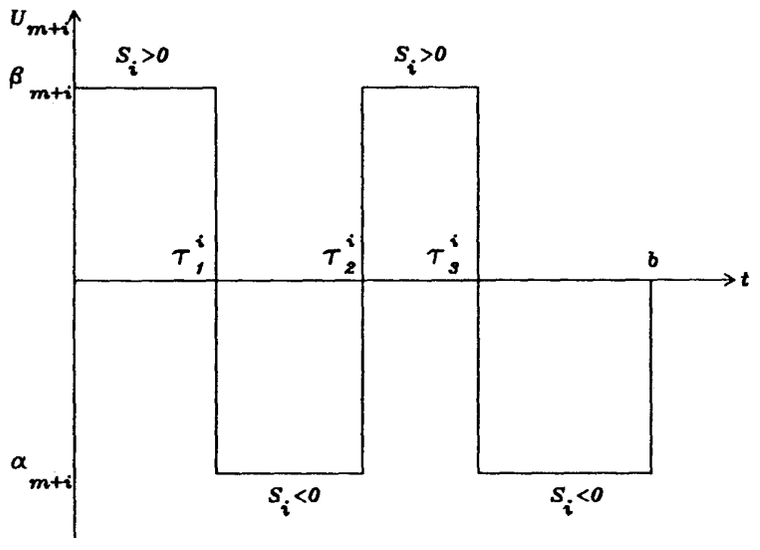


Bild D.6 bang-bang-Steuerung

Bemerkung:

Falls $S_i(t, y, \lambda) \equiv 0 \rightarrow$ singuläre Steuerungen
$2r$–malige Differentiation liefert expliziten Ausdruck für u_{m+i}
(singuläre Steuerung der Ordnung r)

Herleitung der Randbedingung für $\lambda_i(b)$ analog zu Kapitel 5.1 liefert schließlich noch:

$$\lambda_i(b) = +(\Phi_{y_i(b)} + \nu^T r_{y_i(b)}) \tag{5.22}$$

Formulierung von Steuerungsproblemen als Mehrpunkt-Randwertprobleme (Multiplexing)

Steuerungsprobleme ohne Zustandsbeschränkungen und ohne u^2-Anteil liefern Randwertprobleme vom Standardtyp (eventuell mit Parametern).
Zusammenfassung von Kapitel 5.2:

Differentialgleichungen: $2n$

$$\begin{array}{lll} \text{a)} & y_i' = f_i(y,u) & i=1,\ldots,n \\ \text{b)} & \lambda_i' = -\sum_{j=1}^{n} \lambda_j \frac{\partial f_j}{\partial y_i}(t,y,u) & i=1,\ldots,n \end{array} \tag{5.23}$$

Randbedingung: $2n+p$

$$\begin{array}{lll} \text{a)} & y_i(a) = y_{i,a} & i=1,\ldots,n \\ \text{b)} & r_j(t,y(b)) = 0 & j=1,\ldots,p \\ \text{c)} & \lambda_i(b) = \Phi_{y_i(b)} + \nu^T r_{y_i(b)} & p \text{ Parameter } : \nu_1,\ldots,\nu_p \end{array} \tag{5.24}$$

Steuerung: $u = (u^1, u^2)$

$$\begin{array}{ll} \text{a)} & u^1 = u^1(y,\lambda) \text{ vergleiche (5.19'.c)} \\ \text{b)} & u^2 : \quad \text{bang-bang} \end{array} \tag{5.25}$$

Schaltbedingungen:

$$\begin{array}{l} S(\tau_j, y(\tau_j), \lambda(\tau_j)) = 0 \;\; j=1,\ldots,m_s \\ m_s = ? \end{array} \tag{5.26}$$

$\hookrightarrow$ Die Lösung des Problems (5.23) – (5.26) setzt Vorabkenntnis *der Struktur der optimalen Steuerung* voraus, speziell

Anzahl der Schaltpunkte m_s

Vorzeichenstruktur der Schaltfunktionen

Multiplexing-Technik Man führt die Schaltpunkte als zusätzliche Parameter ein:

$$\tau_1 < \tau_2 < \ldots < \tau_{m_s}$$

und transformiert jedes Teilintervall auf konstante Intervall-Länge (ähnlich wie bei "freiem" Randwertproblem). Schematische Erläuterung:

$$\begin{aligned} [a, \tau_1] \to [0,1] \quad &: \bar{t} \;:=\; \frac{t-1}{\tau_1 - a} \\ [\tau_1, \tau_2] \to [1,2] \quad &: \bar{t} \;:=\; \frac{t-\tau_1}{\tau_2 - \tau_1} + 1 \\ &\vdots \\ [\tau_m, b] \to [m, m+1] \quad &: \bar{t} \;:=\; \frac{t-\tau_m}{b - \tau_m} + m \end{aligned} \tag{5.27}$$

$$\text{Differentiation: } y(t) \to \bar{y}(\bar{t}),\;\; \lambda(t) \to \bar{\lambda}(\bar{t}),\;\; \frac{d}{dt} \longrightarrow \frac{d}{d\bar{t}} = \ldots$$

Dies liefert Vorfaktoren in rechter Seite der Differentialgleichung (5.23), welche damit zusäzlich von einem Parameter (Randintervalle) beziehungsweise von zwei Parametern (Zwischenintervalle) abhängen. Die Bestimmung der m_s zusätzlichen Parameter $\tau_1, \ldots, \tau_{m_s}$ erfolgt durch die *inneren Punktbedingungen.*

$$\bar{S}(j, \bar{y}(j), \bar{\lambda}(j)) = 0 \quad j = 1, \ldots, m_s \tag{5.28}$$

Damit ist die formale Rückführung auf parameterabhängiges Mehrpunkt-Randwertproblem vollständig – vergleiche (4.2). Wegen der speziellen Struktur von (5.28) sind Vereinfachungen der Speicherung möglich.

Newton-Iteration benötigt Startwerte für die τ_j, etwa aus Anfangstrajektorie mit Integrator, der mit Nullstellenlöser kombiniert ist:

z. B.: LSODAR (root finder) (HINDMARSH [63])

bisher keine gute Extrapolationsvariante (eventuell SHAMPINE/BACA/BAUER [102]).

Sollte sich die Schaltstruktur im Lauf der Iteration verändern: Neustart mit verändertem Multiplexing.

Bemerkung: Bei Zustandsbeschränkungen treten differentiell-algebraische Gleichungen auf (wegen Lagrange–Ankopplung der Beschränkungen), i.a. vom Index 1 - vgl. Kap. B.4.

6 Asymptotische Randwertprobleme

(DE HOOG/WEISS [23,24], LENTINI [79], LENTINI/KELLER [80])

Problem:

$$\begin{aligned} &\text{a)} \quad y' = f(t,y)\,, \quad y \in \mathbb{R}^n \\ &\text{b)} \quad r(y(0)) = 0\,, \quad \text{Randbedingung links} \\ &\text{c)} \quad \lim_{t\to\infty} y(t) \text{ existiert; asymptotische Randbedingung} \end{aligned} \tag{6.1}$$

Beispiele:

- Symmetrieansätze für äußere Randwertprobleme elliptischer Differentialgleichungen, etwa aus Elastizitätstheorie:

$$\begin{aligned} \Delta u &= \varphi(r) \quad \text{in } \Omega = \{x \,|\, \|x\| > r_0\} \\ u(r_0) &= u_0 \\ \lim_{r\to\infty} u(r) &= 0 \end{aligned}$$

- Selbstähnlichkeitsansätze bei partiellen Differentialgleichungen
- Grenzschichtprobleme (vgl. Kap. D.7)
- Eigenwertaufgaben der Quantenphysik

6.1 Theoretische Vorbereitungen

Zuerst *skalarer* Fall ($n = 1$) und linear in y:

$$\begin{aligned} &\text{a)} \quad y' - t^\alpha \lambda y = t^\alpha g(t)\,, \quad t \in [0,\infty[\\ &\text{b)} \quad \lim_{t\to\infty} y(t) \text{ existiert} \end{aligned} \tag{6.2}$$

Voraussetzungen:

$$\begin{aligned} &\text{a)} \quad \alpha > -1 \quad \text{(Singularität 2. Art)} \\ &\text{b)} \quad \Re\lambda \neq 0\,. \end{aligned} \tag{6.3}$$

Außerdem sei g stetig und $\lim_{t\to\infty} g(t)$ existiere. Variation der Konstanten liefert als allgemeine Lösung von (6.2.a):

$$\begin{aligned} &\text{a)} \quad y(t) = W(t_1 0)y(0) + \int_0^t W(t,s)s^\alpha g(s)ds\,, \\ &\text{wobei} \\ &\text{b)} \quad W(t,s) = \exp\left\{\frac{\lambda}{\alpha+1}(t^{\alpha+1} - s^{\alpha+1})\right\} \end{aligned} \tag{6.4}$$

Lösung der homogenen Differentialgleichung ist. Elementare Abschätzungen zeigen, daß durch

$$\text{a)} \quad \tilde{y}(t) = (H^\lambda g)(t) = \begin{cases} -\int_t^\infty W(t,s)s^\alpha g(s)ds \text{ für } \Re\lambda > 0 \\ \int_0^t W(t,s)s^\alpha g(s)ds \text{ für } \Re\lambda < 0 \end{cases} \tag{6.5}$$

eine spezielle Lösung von (6.2.a,b) gegeben ist, für die

$$\text{b)} \quad \lim_{t\to\infty} \tilde{y}(t) = -g(\infty)/\lambda \text{ gilt.}$$

Für den Operator H^λ gilt nun

Lemma 6.1 (DE HOOG/WEISS [24])

$$H^\lambda : C[0,\infty] \to C[0,\infty] \textit{ ist stetig und linear, wobei}$$
$$\|H^\lambda\| \le C \textit{ für } \lambda \in K \subset \mathbb{C} \textit{ kompakt.} \tag{6.6}$$

Beweis: [24] aufwendig, aber elementar. ∎

Sei nun umgekehrt y Lösung von (6.2.a,b).
$\Re\lambda > 0$: (6.3) umgeschrieben lautet

$$\begin{aligned} y(t) &= -\int_t^\infty W(t,s)s^\alpha g(s)ds \\ &\quad + W(t,0)\left[y(0) + \int_0^\infty W(0,s)s^\alpha g(s)ds\right] \\ &=: (H^\lambda g)(t) + W(t,0)[\ldots] \end{aligned} \tag{6.7}$$

Da $\lim_{t\to\infty}(H^\lambda g)(t)$ existiert und wegen $\Re\lambda > 0$

$$\lim_{t\to\infty} W(t,0) = \infty \ , \tag{6.8}$$

muß aus (6.7) folgen: $[\ldots] = 0$, das heißt

$$y \equiv H^\lambda g \tag{6.9}$$

II. $\Re\lambda < 0$: (6.3) ergibt hier

$$y(t) = W(t,0)y(0) + (H^\lambda g)(t) \ , \tag{6.10}$$

da wegen $\Re\lambda < 0$

$$\lim_{t\to\infty} W(t,0) = 0 \ , \text{ kann } y(0) \text{ beliebig gewählt werden.} \tag{6.11}$$

Zusammengefaßt ist damit bewiesen:

Satz 1 (DE HOOG/WEISS) [24]
Unter den Voraussetzungen (6.2) *sind die Lösungen von* (6.1) *genau:*

$$\begin{aligned} &a) \quad y(t) = (H^\lambda g)(t) && \textit{für } \Re\lambda > 0 \ . \\ &b) \quad y(t) = W(t,0)\xi + (H^\lambda g)(t) && \textit{für } \Re\lambda < 0 \ . \end{aligned} \tag{6.12}$$

Bemerkung: Der Fall $\lambda = 0$ führt auf die allgemeine Lösung

$$y(t) = y(0) + \int_0^t s^\alpha g(s) ds \tag{6.13}$$

Hier muß für

$$\lim_{t\to\infty} y(t) \text{ existiert}$$

beispielsweise

$$g(t) = \mathcal{O}\left(t^{-(\alpha+1+\varepsilon)}\right),\ t \to \infty,\ \varepsilon > 0 \tag{6.14}$$

vorausgesetzt werden.

Dieser Fall, der außerdem nicht "generisch" , das heißt invariant gegen beliebige Störungen ist, führt auf zusätzliche Voraussetzungen technischer Natur.

$\hookrightarrow$ im folgenden fortgelassen.

Zum n-dimensionalen Fall: Modellproblem:

$$\begin{aligned} &\text{a)} \quad y' - t^\alpha M y = t^\alpha g(t)\ ,\ \ t \in [0, \infty[\\ &\text{b)} \quad Ry(0) = \rho \\ &\text{c)} \quad \lim_{t\to\infty} y(t) \text{ existiert} \end{aligned} \tag{6.2'}$$

Hierbei sei M eine reelle $(n \times n)$-Matrix und R eine reelle $(m \times n)$-Matrix.
Voraussetzungen:

$$\begin{aligned} &\text{a)} \quad \alpha > -1 \\ &\text{b)} \quad M \text{ hat } \textit{keine} \text{ Eigenwerte } \lambda \text{ mit } \Re\lambda = 0\ . \end{aligned} \tag{6.3'}$$

Bezeichnung:

$$\mathbb{C}^n = E^+ \oplus E^-\ ,$$

wobei $E^+(E^-)$ der $p_+(p_-)$-dimensionale invariante Teilraum bezüglich M ist, wo M nur Eigenwerte λ mit $\Re\lambda > 0$ $(\Re\lambda < 0)$ hat. P_+ und P_- seien die zugehörigen Projektoren. Die allgemeine Lösung von (6.2'.a) lautet:

$$\begin{aligned} &\text{a)} \quad y(t) \quad = W(t,0)y(0) + \tilde{y}(t)\ , \\ &\text{wobei} \qquad\quad \tilde{y}(\cdot) \text{ spezielle Lösung von (6.2'.a) und} \\ &\text{b)} \quad W(t,s) \ = \exp\{M(t^{1+\alpha} - s^{1+\alpha})/(1+\alpha)\} \end{aligned} \tag{6.15}$$

Mit Jordan-Zerlegung o.ä. sieht man, daß gilt

$$\lim_{t\to\infty} W(t,0)y(0) \text{ existiert } \iff P_+ y(0) = 0 \tag{6.16}$$

Nun ist $\tilde{y}(t)$ mit $\lim_{t\to\infty} \tilde{y}(t)$ existiert gesucht: Dies soll jetzt mittels des *Dunford'schen Funktionalkalküls* aus der skalaren Theorie hergeleitet werden: Sei $\Gamma(\Gamma_+, \Gamma_-)$ Weg in $\mathbb{C}(\mathbb{C}^+, \mathbb{C}^-)$, der alle (die mit $\Re > 0, \Re < 0$) Eigenwerte von M enthält. Dann gilt für $f : \mathbb{C} \to \mathbb{C}$:

$$f(M) = \frac{1}{2\pi i} \int_\Gamma f(\lambda)(\lambda I - M)^{-1} d\lambda \tag{6.17}$$

(Vergleiche Cauchy'sche Integralformel, "Residuensatz" für Operatoren.)
Nun gilt:

$$P_\pm = \frac{1}{2\pi i} \int_{\Gamma_\pm} (\lambda I - M)^{-1} d\lambda \,, \tag{6.18}$$

Ansatz für $\tilde{y}$:

$$\begin{aligned} &\text{a)} \quad \tilde{y}(t) = \frac{1}{2\pi i} \int_\Gamma \tilde{y}_\lambda(t)(\lambda I - M)^{-1} d\lambda \\ &\text{mit} \\ &\text{b)} \quad \lim_{t\to\infty} \tilde{y}_\lambda(t) \text{ existiert für } \lambda \in \Gamma. \end{aligned} \tag{6.19}$$

Funktionalkalkül und (6.2'.a):

$$0 = \frac{1}{2\pi i} \int_\Gamma (\tilde{y}'_\lambda(t) - t^\alpha \lambda \tilde{y}_\lambda - t^\alpha g(t))(\lambda I - M)^{-1} d\lambda \tag{6.20}$$

Dies erfüllt man mit

$$\tilde{y}'_\lambda(t) - t^\alpha \lambda \tilde{y}_\lambda(t) = t^\alpha g(t) \,,$$

das heißt wegen (6.19.b) und (6.12):

$$\tilde{y}_\lambda(t)_i = H^\lambda g_i(t) \qquad i = 1, \ldots, n$$

kurz:

$$\tilde{y}_\lambda(t) = H^\lambda g(t) \,. \tag{6.21}$$

Eingesetzt in (6.19.a) mit (6.5) ergibt:

$$\begin{aligned} \tilde{y} \;&= \frac{1}{2\pi i} \int_\Gamma H^\lambda g(t)(\lambda I - M)^{-1} d\lambda \\ &= -\int_t^\infty \left[\frac{1}{2\pi i} \int_{\Gamma_+} \exp(\frac{\lambda}{\alpha+1} t^{\alpha+1}) \exp(-\frac{\lambda}{\alpha+1} s^{\alpha+1})(\lambda I - M)^{-1} \right] s^\alpha g(s) ds \\ &+ \int_0^t \left[\frac{1}{2\pi i} \int_{\Gamma_-} \exp(\frac{\lambda}{\alpha+1} t^{\alpha+1}) \exp(-\frac{\lambda}{\alpha+1} s^{\alpha+1})(\lambda I - M)^{-1} \right] s^\alpha g(s) ds \end{aligned} \tag{6.19'}$$

Funktionalkalkül:

$$\tilde{y}(t) = \int_0^t P_- W(t,s) s^\alpha g(s) ds - \int_t^\infty P_+ W(t,s) s^\alpha g(s) ds \,. \tag{6.22}$$

Aus (6.19') folgt mit Lemma 6.1 und Einführung des Operators

$$Hg(t) := \tilde{y}(t).$$

Lemma 6.2

$H : C[0,\infty] \to C[0,\infty]$ *ist linear und stetig, wobei*

$$\|H\| \leq C_1 \tag{6.5'}$$

Zusammenfassend haben wir:
$y(t)$ genau dann Lösung von (6.2') a) und b) wenn

$$y(t) = W(t,0)\xi + (Hg)(t) \qquad \text{mit } \xi \in E^- \,. \tag{$*$}$$

Somit bleiben p_- Freiheitsgrade um (6.2'.c) zu erfüllen $\hookrightarrow$

$$m = p_- \,. \tag{6.23}$$

Bezeichnungen:

$$V_\pm : \mathbb{C}^{p_\pm} \longrightarrow E^\pm \text{ Isomorphismen.}$$

Setzt man (*) in (6.2'c) ein:

$$RV_-\xi = \rho - RHg(0)\,, \quad \text{wegen } W(0,0) = I \tag{6.24}$$

legt ξ eindeutig fest, wenn RV_- invertierbar. Somit erhalten wir den

Satz 2 *(*DE HOOG/WEISS *[23])*
(6.2') hat unter den gemachten Voraussetzungen genau dann eine eindeutige reelle Lösung für alle $g \in C[0,\infty]$ und $\rho \in \mathbb{R}^{p_-}$, wenn für ein $T \in [0,\infty[$ (und damit alle)

$$RW(0,T)V_- \textit{ invertierbar}\,. \tag{6.25}$$

Beweis: Es muß nur noch "reell" begründet werden. Dies folgt aus (6.18) und dem Residuensatz. $W(0,T)$ propagiert den Startwert gegen den die bisherige Theorie invariant ist. ∎

Bemerkung: $RW(0,T)V_-$ ist hier die Sensitivitätsmatrix E – vgl. Satz 1.1.

6.2 Approximation auf endlichem Intervall

Idee: Ersetzung der asymptotischen Bedingung (6.2'c)

$$\lim_{t\to\infty} y(t) \text{ existiert}$$

durch eine Bedingung im Endlichen

$$R_T y(T) = \rho_T \text{ für ein gewisses } T \tag{6.2'.c'}$$

und R_T eine (p_+, n)-Matrix (vgl. (6.23)).

Forderungen: Die Lösung von (6.2') a, b, c' sei x_T.

1. $x_T \longrightarrow y$ für $T \to \infty$
   in irgendeinem Sinn.
2. Problem (6.2') a, b, c' soll "gute" Eigenschaften haben.

Forderung 2 läßt sich im Sinne von Kapitel 6.1 sofort präzisieren. Die Sensitivitätsmatrix E_T zu (6.2) a, b, c' soll nichtsingulär sein. Nun gilt:

$$\begin{aligned} E_T &= \begin{bmatrix} R \\ 0 \end{bmatrix} W(0,T) + \begin{bmatrix} 0 \\ R_T \end{bmatrix} \\ &= \begin{bmatrix} RW(0,T) \\ R_T \end{bmatrix} \\ &= S \begin{bmatrix} RW(0,T)V_- & RW(0,T)V_+ \\ R_T V_- & R_T V_+ \end{bmatrix} S^{-1}, \end{aligned} \tag{6.26}$$

bei entsprechender Partitionierung und Basiswahl.
Da nach (6.25) $RW(0,T)V_-$ invertierbar, ist E_T nicht-singulär, falls

$$\begin{aligned} &\text{a)} \quad R_T V_- = 0 \\ &\text{b)} \quad R_T V_+ \text{ invertierbar.} \end{aligned} \tag{6.27}$$

Dies führt sofort zur Wahl

$$R_T = V_+^{-1} P_+ \tag{6.28}$$

Für Forderung 1: Stabilität benötigt:

Satz 3 (DE HOOG/WEISS [23])
Bei der Wahl (6.28) *gilt für hinreichend große* T*:*

$$\|x_T\|_{[0,T]} \leq K(\|g\|_{[0,T]} + \|\rho\| + \|\rho_T\|) \tag{6.29}$$

Hierbei sei $\|x_T\|_{[0,T]} = \sup_{t \in [0,T]} \|x_T(t)\|$.

Beweis:

Bezeichnungen:

$$g_T(t) = \begin{cases} g(t) & 0 \leq t \leq T \\ g(T) & t \geq T \end{cases} \tag{6.30}$$

Zugehörige Lösung:

$$z_T(t) = (H g_T)(t) \tag{6.31}$$

Nach Lemma 6.2 gilt:

$$\|z_T\|_{[0,T]} \leq \|z_T\|_{[0,\infty]} \leq C_1 \|g_T\|_{[0,\infty]} = C_1 \|g\|_{[0,T]} ,$$

also

$$\|z_T\|_{[0,T]} \leq C_1 \|g\|_{[0,T]} \tag{6.32}$$

Die allgemeine Lösung für (6.2') a) im Intervall $[0,T]$ läßt sich ansetzen als

$$\text{a)} \quad z(t) = X_+(t)\xi_+ + X_-(t)\xi_- + z_T(t), 0 \leq t \leq T$$

mit

$$\begin{aligned} &\text{b)} \quad X_+(t) = W(t,T)V_+ \\ &\text{c)} \quad X_-(t) = W(t,0)V_- \\ &\qquad \xi_\pm \in \mathbb{C}^{p_\pm} \end{aligned} \tag{6.33}$$

Beachtet man Bild $V_+ = E_+$, so folgt mit einem von T unabhängigen C_2:

$$\|X_-\|_{[0,T]},\ \|X_+\|_{[0,T]} \leq C_2 \tag{6.34}$$

und es gilt zusätzlich:

$$\begin{aligned} &\text{a)} \quad \lim_{T\to\infty} X_+(0) = 0 \\ &\text{b)} \quad X_+(T) = V_+ \end{aligned} \tag{6.35}$$

Einsetzen von (6.33) in die Randbedingungen bei $t = 0$ und $t = T$ liefert mit (6.35.b) und (6.28)

$$\begin{aligned} &\text{a)} \quad RV_-\xi_- + RX_+(0)\xi_+ &&= \rho - Rz_T(0) \\ &\text{b)} \quad \xi_+ &&= \rho_T - V_+^{-1}P_+z_T(T) \end{aligned} \tag{6.36}$$

Wegen Satz 6.2 ist RV_- invertierbar, so daß eine eindeutige Lösung $(\bar{\xi}_-, \bar{\xi}_+)$ von (6.36) existiert, und damit die eindeutige Lösung X_T des endlichen Randwertproblems.

Für die Norm gilt nach (6.32), (6.34), (6.35.a) und (6.36) für große T:

$$\begin{aligned} \|x_T\|_{[0,T]} &= \sup_{t\in[0,T]} \|X_+(t)\bar{\xi}_+ + X_-(t)\bar{\xi}_- - z_T(t)\| \\ &\leq C_2\,(|\rho_T| + C_3\{|z_T(0)| + |z_T(T)|\} + C_4|\rho|) + C_1\|g\|_{[0,T]} \end{aligned}$$

das heißt (6.29). ∎

Damit Konvergenzsatz als Präzisierung von Forderung 1:

Satz 4 (DE HOOG/WEISS) [23]

Bezeichnung wie bisher. Wählt man als Randbedingung im Endlichen

$$V_+^{-1}P_+x(T) = -V_+^{-1}P_+M^{-1}g(\infty) \quad (\textit{Projektionsbedingung}), \tag{6.2'c''}$$

so gilt:

$$\lim_{T\to\infty} \|x_T - y\|_{[0,T]} = 0\,. \tag{6.37}$$

Beweis.

$u = x_T - y$ ist Lösung des endlichen Randwertproblems

$$\begin{aligned} &\text{a)} \quad u' - t^\alpha M u &&= 0 \\ &\text{b)} \quad Ru(0) &&= 0 \\ &\text{c)} \quad V_+^{-1}P_+u(T) &&= V_+^{-1}P_+\left(-M^{-1}g(\infty) - y(T)\right) \end{aligned} \tag{6.38}$$

das heißt von (6.2') a, b, c' mit

$$\begin{aligned} &\text{a)} \quad g \equiv 0 \\ &\text{b)} \quad \rho = 0 \\ &\text{c)} \quad \rho_T = V_+^{-1}P_+\left(-M^{-1}g(\infty) - y(T)\right)\,. \end{aligned} \tag{6.38'}$$

Stabilität (6.29) liefert:

$$\|x_T - y\|_{[0,T]} \le K\|V_+^{-1}P_+\|\,\|y(T) - \left(-M^{-1}g(\infty)\right)\|\,.$$

Nun gilt $\lim\limits_{T\to\infty} y(T) = -M^{-1}g(\infty)$, so daß die Behauptung folgt. ∎

Bemerkung:

Die Wahl von (6.2'c") ist in mehrfacher Hinsicht "optimal" :

(1) Bei homogenen Problemen exponentielle Konvergenz

(2) Besitzt M verschiedene Eigenwerte und ist $\alpha \in \mathbb{N}_0$, dann liegt "bestmögliche" Konvergenz vor (→ DE HOOG/WEISS [23])

(3) K aus Satz 6.3 wird minimal, das heißt kleinstmögliche Kondition des Randwertproblems (6.2' a,b).

Allgemeiner nichtlinearer Fall

Problem: f hinreichend gutartig:

$$\begin{aligned} &\text{a)} \quad y' = t^\alpha f(t,y) \quad 0 \le t < \infty \\ &\text{b)} \quad r(y(0)) = 0 \\ &\text{c)} \quad \lim_{t\to\infty} y(t) \text{ existiert.} \end{aligned} \tag{6.2''}$$

Es muß eine Wurzel $y(\infty)$ von $f(\infty, y(\infty)) = 0$ geben. Stabilität muß jetzt gelten bezüglich Störungen $\delta y \hookrightarrow$ also Übergang zur Variationsgleichung. Die Rolle von M übernimmt

$$M := f_y(\infty, y(\infty)).$$

Endliche Randbedingung hier

$$V_+^{-1} P_+ M^{-1} f(T, x(T)) = 0\,. \tag{6.39}$$

Bemerkung: Für $f(t,y) = g(t) + My$ ist dies

$$V_+^{-1} P_+ M^{-1}(g(T) + Mx(T)) = 0\,, \text{ das heißt}$$

$$V_+^{-1} P_+ x(T) = -V_+^{-1} P_+ M^{-1} g(T)\,,$$

wegen $g(T) \longrightarrow g(\infty)$ auch hier Satz 6.4 richtig.

6.3 Numerische Umsetzung und Preprocessing

Hinweise zur numerischen Lösung des approximativen endlichen Problems

Homogenisierung der einen Randbedingung: Erreichbar durch Transformation $y \longrightarrow y + M^{-1} g(\infty)$, so daß

$$\lim_{t\to\infty} y(t) = 0\,.$$

Gründe:

(1) (BADER) $y' = t^\alpha(My + g(t)) \longrightarrow 0$ für $t \to \infty$
↪ Auslöschung bei Diskretisierung, außer neue Randbedingung ist homogen
↪ Aufrauhung der rechten Seite
↪ kleine Ordnungen und Schrittweiten

(2) In der Regel Fehler in P_+, das heißt $P_+ + F$ in Projektionsbedingung:

$$(P_+ + F)x(T) = -(P_+ + F)M^{-1}g(\infty)$$

wird gelöst: Nach Stabilität

$$\|x_T - y\|_T \xrightarrow{\approx} \|FM^{-1}g(\infty)\| = 0\,,$$

nur wenn homogenisiert.

Inhärent instabile Anfangswertprobleme in Vorwärtsrichtung:

↪ Rückwärtsintegration bei Mehrzielmethode

↪ Bei Homogenisierung ist das Problem dann zusätzlich stabil und nicht-steif.

Wahl von T :

In der Regel: T aus Zusatzkenntnissen des Anwenders (Ingenieur, Wissenschaftler mit Einsicht in das Problem) oder durch Lösung einer Kette von Problemen mit Folge $\{T_\nu\}$.

Preprocessing: Berechnung von P_+ für reelles M (BORNEMANN 1988)
Rückgriff auf (6.18):

$$P_+ = \frac{1}{2\pi i}\int_{\Gamma_+}(\lambda I - M)^{-1}d\lambda$$

Idee:

- Ausnutzen der immanent reellen Struktur
- Verbiege Γ_+ auf imaginäre Achse ⟶ reelle numerische Integration.

1. Schritt: Nach der Cramer'schen Regel ist (Berechnung zum Beispiel mit REDUCE)

$$(\lambda I - M)^{-1} = \frac{1}{\Delta(\lambda)}\begin{bmatrix} P_{11}(\lambda) & \dots & P_{1n}(\lambda) \\ \vdots & & \vdots \\ P_{n1}(\lambda) & \dots & P_{nn}(\lambda) \end{bmatrix}, \qquad (6.40)$$

wobei $\Delta(\lambda)$ das charakteristische Polynom von M ist und $P_{ij}(\lambda)$ reelles Polynom in λ vom Grad $< n$. Somit ist das Problem reduziert auf die Berechnung von

$$\alpha_j = \frac{1}{2\pi i}\int_{\Gamma_+} \frac{\lambda^j}{\Delta(\lambda)} d\lambda\,, \qquad j = 0,\ldots,n-1\,. \tag{6.41}$$

Diese α_j müssen reell sein (Residuensatz).

2. Schritt: Verbiegen des Integrationsweges ergibt mit Residuensatz

$$\begin{aligned} &\text{a)} \quad \alpha_j = \frac{1}{2\pi}\int_{+\infty}^{-\infty} R_j(\lambda)d\lambda + \frac{1}{2}\delta_{j,n-1} \\ &\text{b)} \quad R_j(\lambda) = \Re\frac{(i\lambda)^j}{\Delta(i\lambda)} \quad j = 0,1,\ldots,n-1 \end{aligned} \tag{6.42}$$

3. Schritt: Umformung der unbeschränkten Integrationsgrenzen in beschränkte:

$$\alpha_j = -\frac{1}{2\pi}\int_0^1 \Bigg[\underbrace{\frac{R_j(\frac{1-x}{x}) + R_j(\frac{x-1}{x})}{x^2}}_{\in C^\infty[0,1]\ !!}\Bigg] dx + \frac{1}{2}\delta_{j,n-1} \tag{6.43}$$

wegen Voraussetzungen an die Eigenwerte von M. Einsetzen der α_j ergibt mit (6.18) P_+, Streichung linear abhängiger Zeilen $V_+^{-1}P_+$.

Beispiel zur Berechnung von P_+

$$\begin{aligned} &\text{a)} \quad y' = x^r A(x) y \quad \text{Plattengleichung} \\ \text{mit} \quad &\text{b)} \quad A(x) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -rx^{-(r+1)} & 1 & 0 \\ 0 & 0 & -2rx^{-(1+r)} & 1 \\ -1 & 0 & 0 & -3rx^{-(1+r)} \end{bmatrix} \end{aligned} \tag{6.44}$$

Hier ist

$$M = \lim_{x\to\infty} A(x) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \end{bmatrix} \tag{6.45}$$

REDUCE liefert

$$(\lambda I - M)^{-1} = \frac{1}{\lambda^4+1}\begin{bmatrix} \lambda^3 & \lambda^2 & \lambda & 1 \\ -1 & \lambda^3 & \lambda^2 & \lambda \\ -\lambda & -1 & \lambda^3 & \lambda^2 \\ -\lambda^2 & -\lambda & -1 & \lambda^3 \end{bmatrix} \tag{6.46}$$

Damit

$$\begin{aligned} &\text{a)} \quad R_0(\lambda) = \frac{1}{\lambda^4 + 1} \\ &\text{b)} \quad R_1(\lambda) = 0 \\ &\text{c)} \quad R_2(\lambda) = -\frac{\lambda^2}{\lambda^4 + 1} \\ &\text{d)} \quad R_3(\lambda) = 0 \end{aligned} \tag{6.47}$$

Also

$$\begin{aligned} &\text{a)} \quad \alpha_0 = -\frac{1}{2\pi} \int_0^1 \frac{2x^2}{(1-x)^4 + x^4} dx \\ &\text{b)} \quad \alpha_1 = 0 \\ &\text{c)} \quad \alpha_2 = +\frac{1}{2\pi} \int_0^1 \frac{2(1-x)^2}{(1-x)^4 + x^4} dx = -\alpha_0 \\ &\text{d)} \quad \alpha_3 = \frac{1}{2} \,. \end{aligned} \tag{6.48}$$

Man berechnet:

$$\alpha_2 = \frac{1}{4}\sqrt{2} \doteq 0.353552 \tag{6.49}$$

Es ergibt sich:

$$P_+ = \begin{bmatrix} \frac{1}{2} & \alpha_2 & 0 & -\alpha_2 \\ \alpha_2 & \frac{1}{2} & \alpha_2 & 0 \\ 0 & \alpha_2 & \frac{1}{2} & \alpha_2 \\ -\alpha_2 & 0 & \alpha_2 & \frac{1}{2} \end{bmatrix} \tag{6.50}$$

Wegen Rang $(P_+) = p_+ = 2$ können offensichtlich 1. und 3. Zeile gewählt werden:

$$V_+^{-1} P_+ = \begin{bmatrix} 1/2 & \alpha_2 & 0 & -\alpha_2 \\ 0 & \alpha_2 & 1/2 & \alpha_2 \end{bmatrix} , \tag{6.51}$$

das heißt, die zusätzlichen Randbedingungen lauten

$$\begin{aligned} &\text{a)} \quad \frac{1}{2} y_1(T) + \alpha_2 y_2(T) - \alpha_2 y_4(T) = 0 \\ &\text{b)} \quad \alpha_2 y_2(T) + \frac{1}{2} y_3(T) + \alpha_2 y_4(T) = 0 \,. \end{aligned} \tag{6.52}$$

Mit dem Wert α_2 nach (6.49) sind dies die Beziehungen von LENTINI [79].

Bemerkung: I.a. wird man die Integrale nicht analytisch auswerten können. Dies entspricht dem Satz von ABEL, daß die Eigenwerte i.a. nicht geschlossen-analytisch berechenbar sind. Hier empfiehlt sich die numerische Auswertung der Ausdrücke (6.43), was schnell und aufgrund der Struktur der Integrale

stabil z.B. mit dem Code TRAPEX von DEUFLHARD/BAUER [36] geschieht. Dies ist der Bestimmung der Jordan-Form o.ä. vorzuziehen.

7 Singuläre Störungsprobleme

(Grenzschichtprobleme, boundary layer problems)

Anwendungen: Transistoren, Membranen (Grenzschichtdicke $\sim \sqrt{\varepsilon}$)

Beispiel:

$$\begin{aligned} &\text{a)} \quad \varepsilon y'' + f(t)y' + g(t)y = h(t) \\ &\text{b)} \quad y(-1) = y_a\,, \quad y(+1) = y_b \end{aligned} \tag{7.1}$$

Lösung für $\varepsilon \to 0^+$: nichtgleichmäßige Konvergenz.
Charakteristische Gleichung:

$$\varepsilon^2\lambda^2 + f(t)\lambda + g(t) = 0 \tag{7.2.a}$$

$$\Rightarrow \lambda_{1,2}(t) = \frac{1}{\varepsilon}[-\frac{f(t)}{2} \pm \sqrt{\frac{f^2(t)}{4} - \varepsilon g(t)}]\,. \tag{7.2.b}$$

Sei $\tau \in]a,b[$ definiert durch:

$$f(\tau) = 0\,. \tag{7.3}$$

$$\begin{aligned} &\text{a)} \quad f(t) \;< 0 \qquad\qquad \text{für } t \in [a,\tau[\\ &\qquad\; f(t) \;> 0 \qquad\qquad \text{für } t \in]\tau,b] \end{aligned} \tag{7.3'}$$

Sei zusätzlich:

$$\text{b)} \quad g(t) \;\leq 0\;.t \in [a,b]$$

Dann gilt für $t \in [a,\tau[$:

$$\lambda_{1,2}(t) = \frac{1}{\varepsilon}\Big[\frac{|f(t)|}{2} \pm \sqrt{\frac{|f(t)|^2}{4} - \varepsilon g(t)}\Big]$$

sowie für $|f(t)|^2 \gg \varepsilon|g(t)|$:

$$\lambda_1(t) \doteq \frac{1}{\varepsilon}|f(t)| > 0\,, \quad \lambda_2(t) \doteq 0 \tag{7.4}$$

Anfangswertproblem stabil in Rückwärtsrichtung ab $t = \tau$ bis $t = a$. (steifes Anfangswertproblem!)

$$\lambda_{1,2}(\tau) = \pm\frac{1}{\varepsilon}\sqrt{\varepsilon|g(t)|} = \pm\frac{1}{\sqrt{\varepsilon}}|g(t)|^{1/2}$$

Anfangswertproblem inhärent instabil in Grenzschicht $\hookrightarrow$ viele Gitterpunkte.
Analog gilt für $t \in [\tau, b]$ und wieder $|f(t)|^2 \gg \varepsilon|g(t)|$:

$$\lambda_1(t) \doteq 0\,, \quad \lambda_2(t) \doteq -\frac{1}{\varepsilon}\, f(t)| < 0 \tag{7.5}$$

Anfangswertproblem stabil in Vorwärtsrichtung ab $t = \tau$ bis $t = b$.

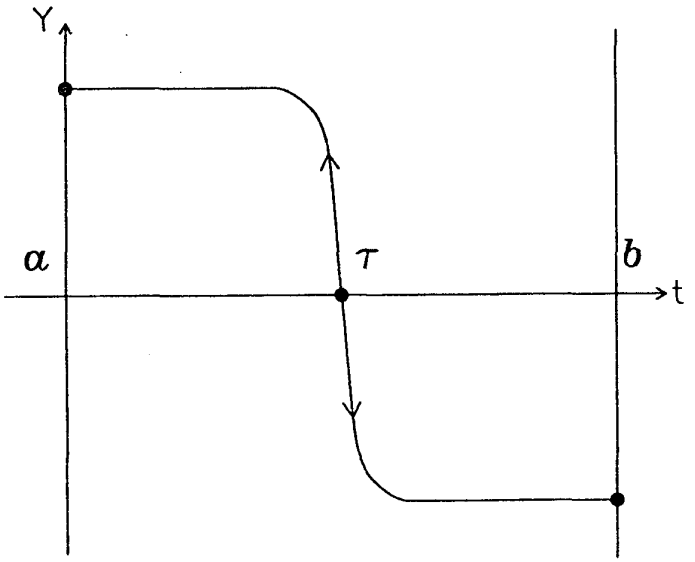


Bild D.7 Stabilität der Anfangswertprobleme gemäß (7.4) und (7.5).

$\hookrightarrow$ Multiplexing bezüglich τ, falls *nichtlineare* Differentialgleichung, Formulierung als 3-Punkt-Randwertproblem

- Differentialgleichung (7.1.a)
- Randbedingung (7.1.b)
- innere Punktbedingung (7.3)

Alternative: Parametrisierung durch Wendepunkt

$$y''(\tau) := 0$$

Dann muß gelten:

$$f(\tau)y'(\tau) + g(\tau)y(\tau) = h(\tau) \tag{7.6}$$

ebenfalls brauchbar als innere Punktbedingung anstelle von (7.3).

Wahl des Integrators:

Man muß auf jeden Fall einen Integrator verwenden, der differentiell-algebraische Gleichungen mit Index= 1 verarbeiten kann!

8 Aufgaben

36.) Betrachtet wird das Randwertproblem

$$\begin{aligned} y''(t) - \lambda^2 y(t) &= 0\,, \\ y(0) = 0 \quad &, \quad y(1) = 1\,. \end{aligned}$$

Man berechne die Kondition $\bar{\rho}$ des Randwertproblems (vgl. (1.6)) und die Kondition

$$\sigma(0,1) = \max_{[t\in 0,1]} \|W(t,0)\|_\infty\,, \quad W \text{ Wronskimatrix}\,,$$

des zugehörigen Anfangswertproblems in Abhängingkeit vom Parameter λ, für $\lambda \gg 1$.

37.) Gegeben sei das Randwertproblem (künstliches Grenzschichtproblem)

$$\begin{aligned} y''(t) &= -\frac{3\tau}{(\tau + t^2)^2} \cdot y(t)\,, \\ y(0.1) &= -y(-0.1) = \frac{0.1}{\sqrt{\tau + 0.01}}\,, \\ (\tau &> 0\,, \quad \text{Scharparameter})\,. \end{aligned}$$

Die zugehörige Sensitivitätsmatrix E^* hat folgende Gestalt:

$$E^*(a) = \begin{bmatrix} 1 & 0 \\ * & \alpha(\tau) \end{bmatrix}\,, \quad \alpha(\tau) = \frac{\partial y(0.1;\tau)}{\partial y'(-0.1)}\,.$$

Man berechne $|\alpha(\tau)|$ und zeige insbesondere:

$$\alpha(\tau) = 0 \text{ für } \tau = 0.01\,.$$

38.) Gegeben sei ein 3-Punkt-Randwertproblem der Form

$$\begin{aligned} y' &= f(y) \\ r\big(y(a), y(\tau), y(b)\big) &= 0\,. \end{aligned}$$

Sei y^* eine Lösung. Man gebe eine hinreichende Bedingung dafür an, daß y^* lokal eindeutig ist.

Hinweis: Man gehe ähnlich vor wie bei Satz 1.1 (Kap. D) der Vorlesung.

39.) Man zeige, daß für die Lösung der Mehrzielgleichungen (2.10) gilt:

$$\Delta x_j = -E_j^{-1} u_j$$

mit

$$\begin{aligned} E_j &:= AG_1^{-1} \cdot \ldots \cdot G_{j-1}^{-1} + BG_{m-1} \cdot \ldots \cdot G_j \\ u_j &:= r + B \sum_{l=j}^{m-1} G_{m-1} \cdot \ldots \cdot G_{l+1} F_l - A \sum_{l=1}^{j} G_1^{-1} \cdot \ldots \cdot G_l^{-1} F_l \,, \end{aligned}$$

wobei gesetzt wird

$$G_{j-1} \cdot \ldots \cdot G_{l+1} := I \text{ für } l = j-1 \,.$$

40.) Gegeben sei das Funktional

$$I(z, \dot{x}, \dot{y}, \dot{z}) := \int_{t=0}^{T} \left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2 - 2gz \right) dt$$

T : Endzeit (unbekannt)

g : Gravitationskonstante

a) Man transformiere I auf die Form $I(\varphi, \dot{\varphi}, \dot{\Theta})$ mittels Toroidkoordinaten

$$\begin{aligned} x &= (R + r \, \sin\varphi) \cos\Theta \\ y &= (R + r \, \sin\varphi) \sin\Theta \\ z &= -r \, \cos\varphi \\ 0 &< r < R \quad R, r \text{ vorgegeben, fest.} \end{aligned}$$

b) Man leite die zugehörigen Euler-Lagrange-Gleichungen her.

c) Die Berechnung der *Ideallinie* eines Bobs in einer 180°–Kurve im Eiskanal (= "Bayernkurve") führt auf das Problem

$$I(\varphi, \dot{\varphi}, \dot{\Theta}) = \min \,.$$

unter den Nebenbedingungen

$$\begin{aligned} &\varphi(0) = \varphi(T) = 0 \,, \\ &\Theta(0) = 0 \,, \quad \Theta(T) = \pi \,, \quad \dot{\Theta}(0) = \omega \,. \end{aligned}$$

Man formuliere das zugehörige Randwertproblem in Standardform (zur weiteren Behandlung mit einem Randwertlöser).

d) Man berechne durch analytische Behandlung die Ableitungen $\dot{\varphi}(0), \ddot{\varphi}(0), \ldots$ der Ideallinie: wie viele Ableitungen verschwinden?

41.) Gegeben sei die Thomas-Fermi-Differentialgleichung

$$y''(t) = \frac{y(t)^{3/2}}{t^{1/2}}$$

mit den Randbedingungen

$$y(0) = 1\,, \quad y(25) = 0$$

Nach Transformation (vgl. Aufgabe 1)

$$s = t^{1/2}\,, \quad w(s) = y(t)\,, \quad u(s) = \dot{w}(s)/s$$

erhält man die Differentialgleichungen

$$\dot{w}(s) = su,\ \dot{u}(s) = 4w^{3/2}$$

mit den Randbedingungen

$$w(0) = 1\,, \quad w(5) = 0\,.$$

Man löse das so gegebene Randwertproblem mit der Routine BVPSOL.
Als Startwerte verwende man z.B.

$$\begin{aligned} w(s_i) &= u(s_i) = 1 \qquad\qquad i = 1, \ldots, 20 \\ w(s_{21}) &= 0\,, \quad u(s_{21}) = 1 \end{aligned}$$

bei äquidistanten Mehrziel-Knoten.

Literatur

[1] Adams, J.C.: (1855) siehe Bashforth, F.

[2] Ascher, U.M.; Mattheij, R.M.M.; Russell, R.D.: *Numerical solution of boundary value problems for ordinary differential equations.* Prentice Hall 1988.

[3] Bader, G.; Deuflhard, P.: *A semi-implicit midpoint rule for stiff systems of ordinary differential equations.* Num. Math. 41, p. 373-398 (1983).

[4] Bashforth, F.: *An attempt to test the theories of capillary action by comparing the theoretical and measured forms of drops of fluid. With an explanation of the method of integration employed in constructing the tables which give the theoretical form of such drops, by J. C. Adams.* Cambridge Univ. Press (1883).

[5] Bernoulli, Johann: *Problema novum ad cujus solutionem mathematici invitantur.* Act. Erud. Leipzig, June 1696, p. 269.

[6] Bock, H.G.: *Numerical treatment of inverse problems in chemical reaction kinetics.* In: Ebert, Deuflhard (Eds.): Modelling of Chemical Reaction Systems; Springer 1981.

[7] Bock, H.G.: *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen.* Universität Bonn: Dissertation 1985.

[8] Bolza, O.: *Vorlesung über Variationsrechnung.* Koehler und Amelang, Leipzig 1949.

[9] Bryson, A.E; Ho, Y.C.: *Applied optimal control.* Ginn and Company, Waltham 1969.

[10] Bulirsch, R.: *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung.* Carl-Cranz-Gesellschaft: Tech. Rep. (Oct. 1971).

[11] Bulirsch, R.: *Variationsrechnung und optimale Steuerung.* Vorlesung gehalten an der Universität zu Köln 1971.

[12] Bulirsch, R., Stoer, J.: *Numerical treatment of ordinary differential equations by extrapolation methods.* Num. Math. 8., p. 1-13 (1966).

[13] Burrage, K.; Butcher, J.C.; Chipman, F.H.: *An implementation of singly-implicit Runge-Kutta Mmethods.* University of Auckland, Report Series No. 149, Computational Mathematics No. 19 (1979).

[14] Butcher, J.C.: *Coefficients for the study of Runge–Kutta integration processes.* J. Austral.. Math. Soc., vol. 3, p. 185–201 (1963).

[15] Butcher, J.C.: *Implicit Runge–Kutta processes.* Math. Comput., Vol. 18, p. 50–64 (1964).

[16] Butcher, J.C.: *The numerical analysis of ordinary differential equations.* John Wiley & Sons. (1987).

[17] Byrne, G.E.; Hindmarsh, A.C.: *A polyalgorithm for numerical solution of ordinary differential equations.* ACM Trans. Math. Software, p. 79–96 (1975).

[18] Curtiss, C.F.; Hirschfelder, J.O.: *Integration of stiff equations.* Proc. of the National Academy of Sciences of U.S., vol. 38, p. 235–243 (1952).

[19] Dahlquist, G.: *Convergence and stability in the numerical integration of ordinary differential equations.* Math. Scand., vol. 4, p. 33–53 (1956).

[20] Dahlquist, G.: *A special stability problem for linear multistep methods.* BIT 3, p. 27–43 (1963).

[21] De Hoog, F.R.; Weiss, R.: *The application of linear multistep methods to Singular Initial Value Problems.* Math. Comp. 31, (1977) 676–690.

[22] De Hoog, F.R.; Weiss, R.: *The numerical solution of boundary value problems with an essential singularity.* SIAM J. Numer. Anal. 16 (1979), 637–669.

[23] De Hoog, F.R.; Weiss, R.: *An Approximation theory for boundary value problems on infinite intervals.* Computing 24 (1980), 227–239.

[24] De Hoog, F.R.; Weiss, R.: *On the boundary value problem for systems of ordinary differential equations with singularity of the second kind.* SIAM J. Math. Anal. 11 (1980), 41–60.

[25] Deuflhard, P.: *Ein Newton–Verfahren bei fast–singulärer Funktionalmatrix zur Lösung von nichtlinearen Randwertaufgaben mit der Mehrzielmethode.* Universität zu Köln, Mathematisches Institut: Dissertation (1972).

[26] Deuflhard, P.: *A modified Newton method for the solution of ill–conditioned systems of nonlinear equations with applications to multiple shooting.* Num. Math. 22 (1974), 289–315.

[27] Deuflhard, P.: *A relaxation strategy for the modified Newton method.* In: Springer Lecture Notes in Math. 447 (1975), R. Bulirsch, W. Oettli, and J. Stoer, eds.

[28] Deuflhard, P.: *A stepsize control for continuatuion methods and its special application to multiple shooting techniques.* Num. Math. 33 (1979), 115–146.

[29] Deuflhard, P.: *Recent advances in multiple shooting techniques.* In: Computational Techniques for Ordinary Differential Equations. NewYork: Academic Press. 1980.

[30] Deuflhard, P.: *Order and stepsize control in extrapolation methods.* Num. Math., vol. 41, p. 399–422 (1983).

[31] Deuflhard, P.: *Computation of periodic solutions of nonlinear ODE's.* BIT 24 (1984), p. 456–466.

[32] Deuflhard, P.: *Recent progress in extrapolation methods for ordinary differential equations.* SIAM Rev., vol. 27, p. 505–535 (1985).

[33] Deuflhard, P.: *Uniqueness theorems for stiff ODE initial value problems.* Preprint SC–87–3, ZIB (1987).

[34] Deuflhard, P.: *Newton techniques for highly nonlinear problems — Theory, Algorithms, Codes.* Academic Press (erscheint 1989/90).

[35] Deuflhard, P.; Bader, G.: *Multiple–shooting techniques revisited.* In: [37].

[36] Deuflhard, P.; Bauer, H.J: *A note on Romberg quadrature.* Universität Heidelberg, SFB 123: Tech. Rep. 169 (1982).

[37] Deuflhard, P.; Hairer, E.: *Numerical treatment of inverse problems in differential and Integral Equations.* Boston: Birkhäuser, 1983.

[38] Deuflhard, P.; Hairer, E.; Zugck, J.: *One–step and extrapolation methods for differential–algebraic systems.* Num. Math. 51, p. 501–516 (1987).

[39] Deuflhard, P.; Heindl, G.: *Affine invariant convergence theorems for Newton's method and extension to related methods.* SIAM J. Numer. Anal. 16., p. 1–10 (1979).

[40] Deuflhard, P.; Nowak, U.: *Extrapolation integrators for quasilinear implicit ODE's.* In: Deuflhard, P.; Engquist, B. (Eds.): Large Scale Scientific Computing, p. 37–50, Birkhäuser Boston, Basel, Stuttgart (1987).

[41] Dormand, J.R.; Prince, P.J.: *A family of embedded Runge–Kutta formulae.* J. Comp. Appl. Math., vol. 6, p. 19–26 (1980).

[42] Ehle, B.L.: *High–order A-stable methods for the numerical solution of systems of D.E.'s.* BIT, vol. 8, p. 276–278 (1968).

[43] England, R.: *A program for the solution of boundary value problems for systems of ordinary differential equations.* Culham Laboratory Report, PDN 3/73, 1976.

[44] Euler, L.: *Institutionum Calculi Integralis.* Volumen Primum, Opera Omnia, vol. XI. (1768).

[45] Fehlberg, E.: *New high-order Runge-Kutta formulas with step size control for systems of first and second order differential equations.* ZAMM, vol. 44, Sonderheft, T17–T19 (1964).

[46] Field, J.R.; Noyes, R.M.: *Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction.* J. Chem. Physics, vol. 60, p. 1877–1884 (1974).

[47] Gantmacher, F.R.: *Matrizentheorie.* Springer (1986).

[48] Gear, C.W.: *Numerical initial value problems in ordinary differential equations.* Prentice–Hall, 253 pp. (1971).

[49] Gragg, W.B.: *Repeated extrapolation to the limit in the numerical solution of ordinary differential equations.* Thesis, Univ. of California (1964); see also SIAM J. Numer. Anal., vol. 2, p. 384–403 (1965).

[50] Grigorieff, R.D.: *Numerik gewöhnlicher Differentialgleichungen.* 2 Bände, Teubner, Stuttgart (1977).

[51] Grigorieff, R.D.: *Stability of multistep-methods on variable grids.* Num. Math. 42, p. 359–377 (1983).

[52] Gröbner, W.: *Die Liereihen und ihre Anwendungen.* D. Verl. d. Wiss. Berlin, 2nd ed. (1967).

[53] Hairer, E.; Bader, G.; Lubich, C.: *On the stability of semi-implicit methods for ordinary differential equations.* BIT 22, p. 211–232 (1982).

[54] Hairer, E.; Lubich, C.: *Asymptotic expansions of the global error of fixed-stepsize methods.* Num. Math., vol. 45, p. 345–360 (1984).

[55] Hairer, E.; Lubich, C.: *Extrapolation at stiff differential equations.* Preprint Université de Genève (1987).

[56] Hairer, E.; Wanner, G.: *On the instability of the BDF formulas.* SIAM J. Numer. Anal., vol. 20, No. 6, p. 1206–1209 (1983).

[57] Hairer, E.; Wanner, G.: *RADAU5 – an implicit Runge-Kutta Code.* Preprint Université de Genève (1988).

[58] Hairer, E.; Nørsett, S.P.; Wanner, G.: *Solving ordinary differential equations I. Nonstiff problems.* Springer–Verlag Berlin, Heidelberg, New York (1987).

[59] Hartman, P.: *Ordinary differential equations.* Reprint of the Second Edition, Birkhäuser Boston, Basel, Stuttgart(1982).

[60] Henrici, P.: *Discrete variable methods in ordinary differential equations.* John Wiley & Sons, Inc., New York, London, Sydney (1962).

[61] Hermann, M.; Berndt, H.: *RWPM, a multiple shooting code for nonlinear two-point boundary value problems.* Preprint 67, 68, 69, FSU Jena, 1982.

[62] Heun, K.: *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen.* Zeitschrift für Mathematik und Physik, vol. 45, p. 23–38 (1900).

[63] Hiebert, K.L.; Shampine, L.F.: *Implicitly defined output points for solutions of ODE's.* Sandia Report SAND80–0180, Feb. 1980.

[64] Hindmarsh, A.C.: *LSODE and LSODI, two new initial value ordinary differential equation solvers.* ACM Signum Newsletter 15, 4 (1980).

[65] Hindmarsh, A.C.: *GEAR–ordinary differential equation system solver.* Tech. Rep. UCID–30001, Rev. 3, Lawrence Livermore National Laboratory, Livermore, CA, (December 1974).

[66] Hindmarsh, A.C.; Byrne G.D.: *EPISODE — An effective package for the integration of systems of ordinary differential equations.* UCID–30112, Rev. 1, Lawrence Livermore National Laboratory, Livermore, CA, (April 1977).

[67] Hull, T.E.; Enright, W.H.; Fellen, B.M.; Sedgwick, A.E.: *Comparing numerical methods for ordinary differential equations* SIAM J. Numer. Anal. 9, p. 603–637 (1972)

[68] Kaps, P.; Rentrop, P.: *Generalized Runge–Kutta methods of order four with stepsize control for stiff ODE's.* Num. Math. 33, p. 55–68 (1979).

[69] Keller, H.B.: *Numerical methods for two–point boundary problems.* Waltham: Blaisdell, 1968.

[70] Kneser, H.: *Über die Lösungen eines Systems gewöhnlicher Differentialgleichungen, das der Lipschitz'schen Bedingung nicht genügt.* S.–B. Preuss. Akad. Wiss. Phys.–Math. Kl., p. 171–174 (1923).

[71] Kreiss H.–O.: *Difference approximations for boundary and eigenvalue problems for ordinary differential equations.* Math. Comp. 26 (1972), 605–624.

[72] Krogh, F.T.: *A Variable step variable order multistep method for the numerical solution of ordinary differential equations.* Information Processing 68, p. 194–199, Amsterdam (1969).

[73] Krogh, F.T.: *Algorithms for changing the step size.* SIAM J. Num. Anal. 10, p. 949–965 (1973).

[74] Krogh, F.T.: *Changing step size in the integration of differential equations using modified devided differences.* Proceedings of the Conference on the Num. Sol. of ODE, Lecture Notes in Math., No. 362, Springer Verlag New York, p. 22–71 (1974).

[75] Krogh, F.T.; Stewart, K.: *Implementation of variable step BDF–methods for stiff ODE's.* Pasadena (1981).

[76] Kublanovskaja, V.N.: *AB-algorithm and its modification for the spectral problem of linear pencils of matrices.* LOMI Preprints E-10-81, Leningrad (1981).

[77] Kronecker, L.: *Algebraische Reduktion der Scharen bilinearer Formen.* Sitz.-Ber. Akad. Wiss., p. 763-776 (1890).

[78] Kutta, W.: *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen.* Zeitschr. für Mathematik und Physik, vol. 46, p. 435-453 (1901).

[79] Lentini, M.: *Boundary value problems over semi-infinite intervals.* Doctoral thesis, California Institute of Technology, 1978.

[80] Lentini, M.; Keller, H.B.: *Boundary value problems on semi-infinite intervals and their numerical solution.* SIAM J. Numer. Anal. 17 (1980), 577-604.

[81] Lentini, M.; Pereyra, V.: *An adaptive finite difference solver for nonlinear two-point boundary value problems with mild boundary layers.* SIAM J. Numer. Anal. 14 (1977), 91-111.

[82] Ljapunov, A.M.: *Problème général de la stabilité du mouvements.* russ. (1892), trad. en fraçaise 1907. (Annales de la Faculté des Sciences de Toulouse), reprinted Princeton University Press, 474 pp. (1947).

[83] Lindelöf, E.: *Sur l'application des méthodes d'approximation successives à l'étude des intégrales réelles des équations différentielles ordinaires.* J. de Math., 4e série, vol. 10, p. 117-128 (1894).

[84] Liniger, W.; Willoughby, R.A.: *Efficient integration methods for stiff systems of ordinary differential equations.* SIAM J. Numer. Anal. 7 (1), p. 47-66 (1970).

[85] Mattheij, R.M.M.: *The conditioning of linear boundary value problems.* SIAM J. Numer. Anal. 19 (1982), 963-978.

[86] Morrison, D.D., Riley, J.D.; Zancanaro, J.F.: *Multiple shooting method for two-point boundary value problems.* Comm. ACM 5 (1962), 613-614.

[87] Moulton, F.R.: *New methods in exterior ballistics.* Univ. Chicago Press (1926).

[88] Oberle, H.J.: *BOUNDSCO, Hinweise zur Benutzung des Mehrzielverfahrens für die numerische Lösung von Randwertproblemen mit Schaltbedingungen.* Hamburger Beiträge zur angewandten Mathematik, Reihe B, Bericht 6, November 1987.

[89] Osborne, M.R. *The stabilized march is stable.* SIAM J. Numer. Anal. 16 (1979), 923-933.

[90] Peano, G.: *Dèmonstration de l'intégrabilité des équations différentielles ordinaires.* Math. Annalen, vol. 37, p. 182–228; see also the german translation and commentation: G. Mie, Math. Annalen, vol. 43, (1893) p. 553–568 (1890).

[91] Petzold, L.: *A Description of DASSL: a differential–algebraic system solver.* Proc. IMACS World Congress (1982).

[92] Picard, E.: *Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives.* J. de Math. pures et appl., 4e série, vol. 6, p. 145–210 (1890).

[93] Pontrjagin, L.: *Optimal control processes.* Uspehi Mat. Nauk. 14 (1959) 1, 3–20 (Russian).

[94] Rheinboldt, W.C.: *Differential–algebraic systems as differential equations on manifolds.* Math. Comp. vol. 43, p. 473–482 (1985).

[95] Richardson, L.F.: *The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam.* Phil. Trans., A, vol. 210, p. 307–357 (1910).

[96] Romberg, W.: *Vereinfachte numerische Integration.* Norske Vid. Selsk. Forhdl., vol. 28, p. 30–36 (1955).

[97] Runge, C.: *Über die numerische Auflösung von Differentialgleichungen.* Math. Ann., vol. 46, p. 167–178 (1895).

[98] Russell, R.D.; Shampine, L.F.: *A collocation method for boundary value problems.* Num. Math. 19, p. 1–28 (1972).

[99] Rutishauser, H.: *Über die Instabilität von Methoden zur Integration gewöhnlicher Differentialgleichungen.* ZAMP, vol. 3, p. 65–74 (1952).

[100] Scott, M.R.; Watts, H.A.: *A Systematized collection of codes for solving two–point boundary–value problems.* SANDIA Lab., Albuquerque: Tech. Rep. SAND 75–0539, 1975.

[101] Shampine, L.F.; Gordon, M.K.: *Computer solution of ordinary differential equations, the initial value problem.* Freeman and Company, San Francisco, p. 318 (1975).

[102] Shampine, L.F.; Baca L.S.; Bauer, H.J: *Output in extrapolation codes.* Comp. Math. Appl. 9. No.2 (1983) p. 245–255.

[103] Skeel, R.D.: *Equivalent forms of multistep formulas.* Report R–78–940, Dept. of Comp. Sci., Univ. of Illinois at Urbana–Champaign (1978).

[104] Skeel, R.D.: *Iterative refinement implies numerical stability for Gaussian elimination.* Math. Comp. 35 , (1980), 817-832.

[105] Stetter, H.J.: *Symmetric two-step algorithms for ordinary differential equations.* Computing 5, p. 267–280 (1970).

[106] Stetter, H.J.: *Analysis of discretization methods for ordinary differential equations.* Springer Verlag, Berlin, Heidelberg, New York (1973).

[107] Stiefel, E.; Bettis, D.G.: *Stabilization of Cowell's method..* Num. Math. 13, p. 154–175 (1969).

[108] Stoer, J.; Bulirsch, R.: *Einführung in die Numerische Mathematik II.* Springer Verlag, Berlin, Heidelberg, New York (1973).

[109] Störmer, C.: *Méthodes d'intégration numérique des équations différentielles ordinaires.* C.R. congr. intern. math., Strasbourg, p. 243–257 (1921).

[110] von Neumann, J.: *Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes.* Math. Nachrichten 4, p. 258–281 (1951).

[111] Wanner, G.; Hairer, E; Nørsett, S.P.: *Order stars and stability theorems.* BIT 18, p. 475–489 (1978).

[112] Weiss, R.: *The convergence of shooting methods.* BIT 13 (1973), 470–475.

[113] Widlund, O.B.: *A note on unconditionally stable linear multistep methods.* BIT 7, p. 65–70 (1967).

[114] Wilde, Oscar: *Lady Windermere's Fan, Comedy in four acts.* (1892).

Veröffentlichungen des Konrad-Zuse-Zentrums für Informationstechnik Berlin
Preprints März 1989

SC 86-1. P. Deuflhard; U. Nowak. *Efficient Numerical Simulation and Identification of Large Chemical Reaction Systems.* (vergriffen)

SC 86-2. H. Melenk; W. Neun. *Portable Standard LISP for CRAY X-MP Computers.*

SC 87-1. J. Anderson; W. Galway; R. Kessler; H. Melenk; W. Neun. *The Implementation and Optimization of Portable Standard LISP for the CRAY.*

SC 87-2. Randolph E. Bank; Todd F. Dupont; Harry Yserentant. *The Hierarchical Basis Multigrid Method.* (vergriffen)

SC 87-3. Peter Deuflhard. *Uniqueness Theorems for Stiff ODE Initial Value Problems.*

SC 87-4. Rainer Buhtz. *CGM-Concepts and their Realization.*

SC 87-5. P. Deuflhard. *A Note on Extrapolation Methods for Second Order ODE Systems.*

SC 87-6. Harry Yserentant. *Preconditioning Indefinite Discretization Matrices.*

SC 88-1. Winfried Neun; Herbert Melenk. *Implementation of the LISP-Arbitrary Precision Arithmetic for a Vector Processor.*

SC 88-2. H. Melenk; H.M. Möller; W. Neun. *On Gröbner Bases Computation on a Supercomputer Using REDUCE.* (vergriffen)

SC 88-3. J. C. Alexander; B. Fiedler. *Global Decoupling of Coupled Symmetric Oscillators.*

SC 88-4. Herbert Melenk; Winfried Neun. *Parallel Polynomial Operations in the Buchberger Algorithm.*

SC 88-5. P. Deuflhard; P. Leinen; H. Yserentant. *Concepts of an Adaptive Hierarchical Finite Element Code.*

SC 88-6. P. Deuflhard; M. Wulkow. *Computational Treatment of Polyreaction Kinetics by Orthogonal Polynomials of a Discrete Variable.* (vergriffen)

SC 88-7. H. Melenk; H. M. Möller; W. Neun. *Symbolic Solution of Large Stationary Chemical Kinetics Problems.*

SC 88-8. Ronald H. W. Hoppe; Ralf Kornhuber. *Multi-Grid Solution of Two Coupled Stefan Equations Arising in Induction Heating of Large Steel Slabs.*

SC 88-9. Ralf Kornhuber; Rainer Roitzsch. *Adaptive Finite-Element-Methoden für konvektionsdominierte Randwertprobleme bei partiellen Differentialgleichungen.*

SC 88-10. C. -N. Chow; B. Deng; B. Fiedler. *Homoclinic Bifurcation at Resonant Eigenvalues.*

Veröffentlichungen des Konrad-Zuse-Zentrums für Informationstechnik Berlin
Technical Reports März 1989

TR 86-1. H.J. Schuster. *Tätigkeitsbericht 1985.* (vergriffen)

TR 87-1. Hubert Busch; Uwe Pöhle; Wolfgang Stech. *CRAY-Handbuch. - Einführung in die Benutzung der CRAY..*

TR 87-2. Herbert Melenk; Winfried Neun. *Portable Standard LISP Implementation for CRAY X-MP Computers. Release of PSL 3.4 for COS.*

TR 87-3. Herbert Melenk; Winfried Neun. *Portable Common LISP Subset Implementation for CRAY X-MP Computers.*

TR 87-4. Herbert Melenk; Winfried Neun. *REDUCE Installation Guide for CRAY 1 / X-MP Systems Running COS Version 3.2.*

TR 87-5. Herbert Melenk; Winfried Neun. *REDUCE Users Guide for the CRAY 1 / X-MP Series Running COS. Version 3.2.*

TR 87-6. Rainer Buhtz; Jens Langendorf; Olaf Paetsch; Danuta Anna Buhtz. *ZUGRIFF - Eine vereinheitlichte Datenspezifikation für graphische Darstellungen und ihre graphische Aufbereitung.*

TR 87-7. J. Langendorf; O. Paetsch. *GRAZIL (Graphical ZIB Language).*

TR 88-1. Rainer Buhtz; Danuta Anna Buhtz. *TDLG 3.1 - Ein interaktives Programm zur Darstellung dreidimensionaler Modelle auf Rastergraphikgeräten.*

TR 88-2. Herbert Melenk; Winfried Neun. *REDUCE User's Guide for the CRAY 1 / CRAY X-MP Series Running UNICOS. Version 3.3.*

TR 88-3. Herbert Melenk; Winfried Neun. *REDUCE Installation Guide for CRAY 1 / X-MP Systems Running UNICOS. Version 3.3.*

TR 88-4. Danuta Anna Buhtz; Jens Langendorf; OLaf Paetsch. *GRAZIL-3D. Ein graphisches Anwendungsprogramm zur Darstellung von Kurven- und Funktionsverläufen im räumlichen Koordinatensystem.*

TR 88-5. Gerhard Maierhöfer; Georg Skorobohatyj. *Ein paralleler, adaptiver Algorithmus zur numerischen Integration ; seine Implementierung für SUPRENUM-artige Architekturen mit SUSI.*

TR 89-1. *CRAY-HANDBUCH. Einführung in die Benutzung der CRAY X-MP unter UNICOS.*

TR 89-2. P. Deuflhard. *Numerik von Anfangswertmethoden für gewöhnliche Differentialgleichungen.*

TR 89-3. Artur Rudolf Walter. *Ein Finite-Element-Verfahren zur numerischen Lösung von Erhaltungsgleichungen.*