Numerik gewöhnlicher Differentialgleichungen

Oliver Ernst

Professur Numerische Mathematik

Wintersemester 2016/17

Mathematik!
TU Chemnitz

Organisatorisches

- Vorlesungswebseite:
  `www.tu-chemnitz.de/mathematik/numa/lehre/ode-2014`
- Vorlesung: Prof. Oliver Ernst,
  `oliver.ernst@mathematik.tu-chemnitz.de`
  Mo 9:15 (ungerade Wochen) & Do 9:15 (wöchentlich)
- Übung: Dipl.-Math. oec. Jan Blechschmidt,
  `jan.blechschmidt@mathematik.tu-chemnitz.de`
  Mo 9:15 (gerade Wochen)
- Prüfung: 30 Minuten mündlich, Termin nach Vereinbarung.
- Modul M13: 6 LP, 180 AS.

Vorbemerkungen

- Eine Differentialgleichung beschreibt eine Beziehung zwischen einer gesuchten Funktion (abhängige Variable) und ihren Ableitungen bezüglich einer oder mehreren (unabhängigen) Variablen.
- Werden nur Ableitungen bezüglich einer Variablen betrachtet, so spricht man von einer gewöhnlichen Differentialgleichung (GDG) (ordinary differential equation, ODE), andernfalls von einer partiellen Differentialgleichung (PDG) (partial differential equation, PDE).
- Da explizite Lösungen nur in wenigen Ausnahmefällen zur Verfügung stehen, ist der Einsatz numerischer Methoden zur Lösung von GDGen unvermeidbar.
- Wir befassen uns hier ausschließlich mit numerischen Verfahren zur Lösung von GDGen: Zum Einen wegen der überragenden Rolle, die sie in vielen Anwendungen spielen, was (hoffentlich) im Laufe der Vorlesung klar wird; zum Zweiten aber auch wegen ihrer Bedeutung für die Entwicklung und Analyse numerischer Methoden zur Lösung von PDGen, welche sehr viel schwieriger zu lösen sind und im Mittelpunkt einer eigenen Vorlesung stehen.

Wesentliche Inhalte

- Begriff, Beispiele, theoretische und andere Hilfsmittel
- Numerische Methoden für GDGen
- Lineare Mehrschrittverfahren
- Runge-Kutta-Verfahren
- Schrittweitensteuerung
- Steife Systeme
- Exponentielle Integratoren
- Zufällige und stochastische Differentialgleichungen (wenn Zeit erlaubt)

Literatur

- Harro Heuser. *Gewöhnliche Differentialgleichungen.* Vieweg-Teubner, 2009 (6. Auflage).
- Martin Braun. *Differentialgleichungen und ihre Anwendungen*. Springer-Verlag, 1979.
- Ernst Hairer, Syvert P. Nørsett, Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 1987.
  Ernst Hairer and Gerhard Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer, 1991.
- John D. Lambert. *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem* Wiley, 1991
- Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.
- Uri M. Ascher, Linda R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, 1998.
- David F. Griffiths and Desmond J. Higham. *Numerical Methods for Ordinary Differential Equations: Initial Value Problems*. Springer, 2010.

Siehe auch die (laufend ergänzte) Literaturliste auf der Webseite der Vorlesung.

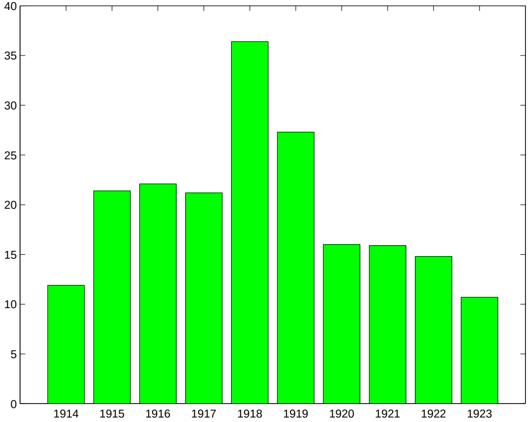
Inhalt I

Inhalt III

Inhalt

Inhalt

Der Biologe Umberto d'Ancona (1896–1964) stellte 1925 den prozentualen Anteil der Haie am Gesamtfang (Speisefische und Haie) im Hafen von Triest fest:



Benachteiligt eingeschränkter Fischfang (1. Weltkrieg) die Speisefische?

D'Ancona konsultierte den Mathematiker Volterra[1], der die Populationsdynamik wie folgt modellierte: Seien

$$\begin{aligned} x(t) &: \ \text{Beutepopulation zur Zeit } t \text{ (Speisefische)} \\ y(t) &: \ \text{Räuberpopulation zur Zeit } t \text{ (Haie).} \end{aligned}$$

Ohne Räuber würde sich die Beute nach dem Malthusianischen[2] Gesetz

$$x'(t) = a\,x(t) \qquad \text{(mit einer Konstanten } a > 0\text{),}$$

vermehren, d.h. der Zuwachs wäre proportional zum Bestand bzw. das Wachstum wäre exponentiell

$$x(t) = x(0)\exp(at) \qquad \text{für } t \geqslant 0$$

(eingeschränkt realistisch, falls Population nicht sehr dicht und ausreichend Nahrung vorhanden ist).

[1] Vito Volterra (1860–1940)
[2] Thomas Malthus (1766–1834)

Anzahl Räuber-Beute-Kontakte (pro Zeiteinheit):

$b\,x(t)\,y(t)$ (mit einer Konstanten $b > 0$).

Insgesamt: $x'(t) = a\,x(t) - b\,x(t)\,y(t)$.

Analog: $y'(t) = -c\,y(t) + d\,x(t)\,y(t)$, mit weiteren Konstanten $c, d > 0$.

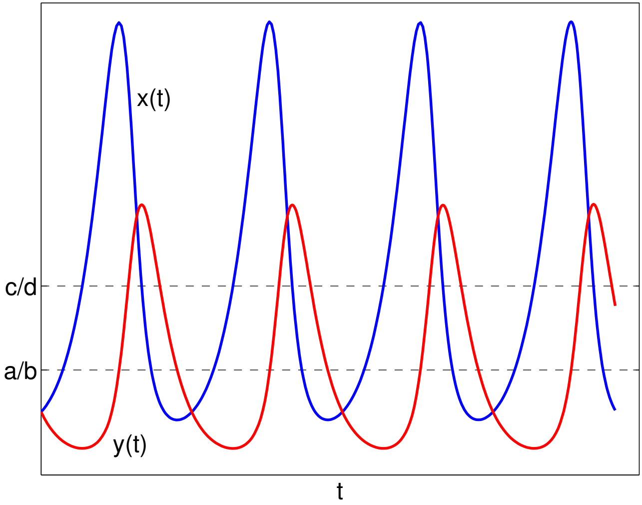
Wir erhalten ein System zweier GDGen.

Man kann zeigen: dessen Lösungen sind periodisch: d.h. $\exists\, T > 0$ sodass

$$x(t+T) = x(t), \quad y(t+T) = y(t) \qquad \text{für alle } t.$$

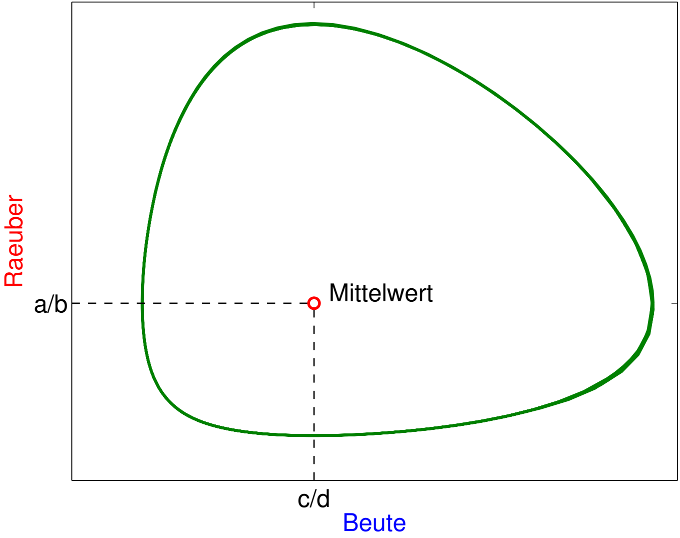
Mittelwerte: $$\overline{x} := \frac{1}{T}\int_0^T x(t)\,\mathrm{d}t = \frac{c}{d}, \qquad \overline{y} := \frac{1}{T}\int_0^T y(t)\,\mathrm{d}t = \frac{a}{b}.$$

Volterras Prinzip
Zeitlicher Verlauf der Populationen



Volterras Prinzip

Darstellung in der Phasenebene



Berücksichtige Fischfang:

$$\begin{aligned} x'(t) &= a\,x(t) - b\,x(t)y(t) - e\,x(t) &&= (a-e)\,x(t) - b\,x(t)y(t), \\ y'(t) &= -c\,y(t) + d\,x(t)y(t) - e\,y(t) &&= -(c+e)\,y(t) + d\,x(t)y(t), \end{aligned} \qquad (e > 0).$$

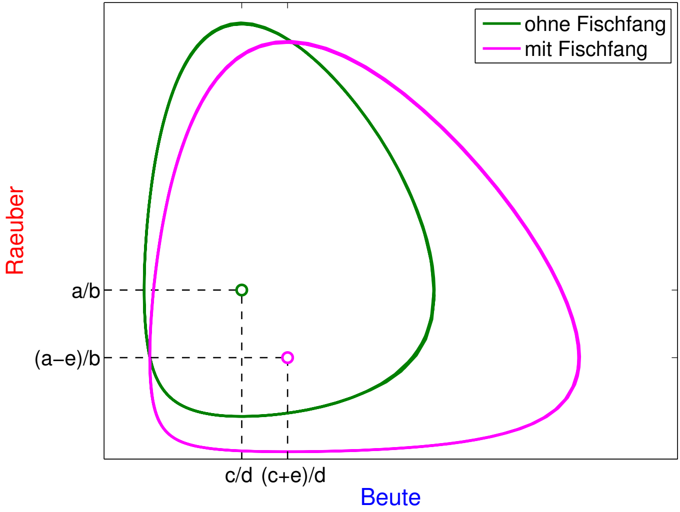
Gleiches System mit neuen Koeffizienten: $a \to a - e$ und $c \to c + e$.

Mittelwerte: $\frac{c+e}{d} > \frac{c}{d}$ (Beute), $\frac{a-e}{b} < \frac{a}{b}$ (Räuber).

Volterras Prinzip: Moderater Fischfang ($e < a$) steigert die durchschnittliche Zahl der Speisefische und reduziert die durchschnittliche Zahl der Haie.

Volterras Prinzip

Darstellung in der Phasenebene



Inhalt

Einen Ausdruck der Form

$$F\big(t, y, y', y'', \dots, y^{(n)}\big) = 0 \tag{GDG}$$

mit einer Funktion $F : \mathbb{R}^{n+2} \supset M \to \mathbb{R}$ nennen wir eine gewöhnliche Differentialgleichung (GDG) n-ter Ordnung. Eine Funktion $y : \mathbb{R} \supset I \to \mathbb{R}$ heißt Lösung von (GDG) über dem Intervall I, wenn $y \in C^n(I)$ ist und für alle $t \in I$ gilt

$$F\big(t, y(t), y'(t), y''(t), \dots, y^{(n)}(t)\big) = 0.$$

- (GDG) besitzt die Ordnung n, weil n die Ordnung der höchsten auftretenden Ableitung ist.
- Sie heißt gewöhnlich, weil nur Ableitungen der gesuchten Funktion y nach einer Variablen auftreten.
- (GDG) heißt implizit — im Gegensatz zu einer expliziten GDG n-ter Ordnung, die nach der höchsten Ableitung von y aufgelöst ist:

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

Wir werden fast ausschließlich Systeme von expliziten GDGen erster Ordnung betrachten (warum wir uns auf Systeme erster Ordnung beschränken können, wird später erklärt):

$$\begin{aligned} y_1' &= f_1(t, y_1, y_2, \dots, y_n) \\ y_2' &= f_2(t, y_1, y_2, \dots, y_n) \\ \vdots &= \vdots \\ y_n' &= f_n(t, y_1, y_2, \dots, y_n) \end{aligned} \tag{DG}$$

mit den n unbekannten Funktionen $y_1, y_2, \dots, y_n$. Jedes System von n Funktionen

$$y_1 = y_1(t), \dots, y_n = y_n(t) \in C^1(I),$$

das (DG) für alle $t \in I$ erfüllt, heißt Lösung von (DG) über I.

Das System

$$
\begin{aligned}
y_1' &= 1 \\
y_2' &= 2y_1
\end{aligned}
$$

besitzt die Lösungen

$$y_1(t) = t + \alpha, \quad y_2(t) = t^2 + 2\alpha t + \beta \qquad (\alpha, \beta \in \mathbb{R})$$

über $(-\infty, \infty)$.

Für eine eindeutige Lösung: Anfangsbedingungen, z.B.

$$y_1(0) = 1, \quad y_2(0) = 2.$$

Dann ist

$$y_1(t) = t + 1, \quad y_2(t) = t^2 + 2t + 2$$

die einzige Lösung.

Allgemein: Die Aufgabenstellung, eine Lösung von (DG) zu finden, die die Anfangsbedingung

$$y_1(t_0) = y_{0,1}, \ \ldots, \ y_n(t_0) = y_{0,n} \tag{AB}$$

erfüllt, heißt Anfangswertproblem (AWP) oder Anfangswertaufgabe für die gewöhnliche Differentialgleichung (DG).

Mit der Vektornotation

$$\boldsymbol{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{f} := \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}, \qquad \boldsymbol{y}_0 := \begin{bmatrix} y_{0,1} \\ \vdots \\ y_{0,n} \end{bmatrix},$$

ergibt sich die Kurzschreibweise

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \tag{DG'}$$

$$\boldsymbol{y}(t_0) = \boldsymbol{y}_0. \tag{AB'}$$

Bemerkung. GDGen höherer Ordnung lassen sich in (äquivalente) Systeme von GDGen erster Ordnung umschreiben:
Aus

$$y''' + 3y'' + y' = \sin(t)$$

wird etwa

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ -3y_3 - y_2 + \sin(t) \end{bmatrix}$$

mit den neuen Variablen

$$y_1 = y, \quad y_2 = y_1' = y', \quad y_3 = y_2' = y''.$$

Die explizite Abhängigkeit der rechten Seite von der unabhängigen Variable (hier t) kann durch Hinzunahme einer zusätzlichen Gleichung bzw. Komponente des Lösungsfunktionsvektors $\boldsymbol{y}$ beseitigt werden:

$$y_4(t) = t \quad \text{(d.h. } y_4'(t) \equiv 1), \qquad y_4(t_0) = t_0.$$

Im obigen Beispiel resultiert dies in der autonomen Differentialgleichung $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$, oder genauer:

$$\boldsymbol{y}'(t) = \boldsymbol{f}(\boldsymbol{y}(t)), \qquad \boldsymbol{f}(\boldsymbol{y}) = \begin{bmatrix} y_2 \\ y_3 \\ -3y_3 - y_2 + \sin(y_4) \\ 1 \end{bmatrix}.$$

Satz 1.1 (Picard-Lindelöf)

Gegeben ist die Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad \boldsymbol{y}(t_0) = \boldsymbol{y}_0. \tag{AWP}$$

Die rechte Seite $\boldsymbol{f}$ sei stetig im 'Quader'

$$Q := \{(t, \boldsymbol{y}) \,:\, |t - t_0| \leqslant a,\ \|\boldsymbol{y} - \boldsymbol{y}_0\| \leqslant b\}, \quad Q \subset \mathbb{R}^{n+1},$$

und es sei $M := \max\{\|\boldsymbol{f}(t, \boldsymbol{y})\| \,:\, (t, \boldsymbol{y}) \in Q\}$.
Außerdem erfülle $\boldsymbol{f}$ in Q die Lipschitz-Bedingung

$$\|\boldsymbol{f}(t, \boldsymbol{y}) - \boldsymbol{f}(t, \tilde{\boldsymbol{y}})\| \leqslant L\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| \quad \forall\, (t, \boldsymbol{y}), (t, \tilde{\boldsymbol{y}}) \in Q. \tag{Lip}$$

Dann besitzt das Problem (AWP) genau eine Lösung über $I := [t_0 - \alpha,\ t_0 + \alpha]$, wobei $\alpha = \min\{a, b/M\}$.

Emile Picard (1856–1941), Ernst Lindelöf (1870–1946), Rufolf Lipschitz (1832–1903).

Bemerkungen.

(1) In der gesamten Vorlesung wird vorausgesetzt, dass die fundamentale Bedingung (Lip) erfüllt ist.

(2) (AWP) besitzt in $[t_0 - a, t_0 + a]$ eine eindeutige Lösung, wenn $\boldsymbol{f}$ die Bedingung (Lip) in $\tilde{Q} = \{(t, \boldsymbol{y}) : |t - t_0| \leqslant a, \|\boldsymbol{y}\| < \infty\}$ erfüllt.

(3) Ist $\boldsymbol{f}$ auf Q bez. $\boldsymbol{y}$ stetig differenzierbar und bezeichnet $\boldsymbol{f_y} = [\partial f_i / \partial y_j]_{1 \leqslant i,j \leqslant n}$ die zugehörige Jacobi-Matrix, dann folgt aus dem Mittelwertsatz, dass die Voraussetzungen von Satz 1.1 mit

$$L = \sup_{(t,\boldsymbol{y}) \in Q} \|\boldsymbol{f_y}(t, \boldsymbol{y})\| < \infty$$

erfüllt sind.

(4) (AWP) besitzt auch dann noch Lösungen, wenn $\boldsymbol{f}$ nur als stetig auf Q vorausgesetzt wird (Existenzsatz von Peano[3]). Deren Eindeutigkeit ist aber nicht mehr gesichert.

[3] Giuseppe Peano (1858–1932)

Beispiel:

$$y' = f(t,y) = \sqrt{y}, \quad y(0) = 0, \qquad Q = \mathbb{R} \times [0,\infty),$$

mit den Lösungen

$$y_\lambda(t) = \begin{cases} 0, & 0 \leqslant t \leqslant \lambda, \\ (t-\lambda)^2/4, & t \geqslant \lambda, \end{cases} \qquad (\lambda \geqslant 0).$$

Satz 1.2

Die Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad \boldsymbol{y}(t_0) = \boldsymbol{y}_0,$$

erfülle die Voraussetzungen von Satz 1.1. Über eine weitere Anfangswertaufgabe

$$\boldsymbol{y}' = \tilde{\boldsymbol{f}}(t, \boldsymbol{y}), \quad \boldsymbol{y}(t_0) = \tilde{\boldsymbol{y}}_0,$$

setzen wir nur voraus, dass $\tilde{\boldsymbol{f}}$ stetig in Q ist. Sind dann $\boldsymbol{y}$ und $\tilde{\boldsymbol{y}}$ Lösungen dieser Anfangswertaufgaben über dem Intervall I und gilt

$$\|\boldsymbol{y}_0 - \tilde{\boldsymbol{y}}_0\| \leqslant \gamma \quad \text{sowie} \quad \|\boldsymbol{f}(t, \boldsymbol{y}) - \tilde{\boldsymbol{f}}(t, \boldsymbol{y})\| \leqslant \delta \quad \forall\, (t, \boldsymbol{y}) \in Q,$$

so folgt für $t \in I$

$$\|\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t)\| \leqslant \gamma\, e^{L(t-t_0)} + \frac{\delta}{L}\left(e^{L(t-t_0)} - 1\right).$$

(vgl. [Heuser, Satz 13.1])

Inhalt

Eine wichtige Rolle werden lineare Differenzengleichungen

$$y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = \beta_{n+k} \qquad (n = 0, 1, 2, \dots) \tag{DzG}$$

spielen.

- Genauer spricht man hier von einer linearen Differenzengleichung der Ordnung k mit konstanten Koeffizienten (die α's hängen nicht von n ab). (O.B.d.A. sei $\alpha_0 \neq 0$).
- Die Gleichung heißt homogen, wenn $\beta_{n+k} = 0$ für alle n, andernfalls inhomogen.
- Jede Folge $\{y_n\}_n$, die (DzG) erfüllt, heißt eine Lösung von (DzG).
- Gibt man sich k Startwerte y_0, y_1, ..., y_{k-1} (beliebig) vor, kann man sich mit (DzG) rekursiv eine solche Lösung berechnen.

Lemma 1.3

Die Lösungsmenge einer homogenen linearen Differenzengleichung der Ordnung k ist ein Vektorraum der Dimension k.
Besitzt die Differenzengleichung darüber hinaus konstante Koeffizienten, so kann man eine Basis dieses Lösungsraums mit Hilfe der Nullstellen des zugehörigen charakteristischen Polynoms

$$p_k(\zeta) = \zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_1\zeta + \alpha_0$$

angeben: Bezeichnen λ_j $(1 \leqslant j \leqslant \ell)$ die Nullstellen von p (mit Vielfachheiten m_j, $\sum_{j=1}^{\ell} m_j = k$), so bilden die k Folgen

$$(\lambda_j^n)_n,\ (n\lambda_j^{n-1})_n,\ \ldots,\ (n(n-1)\ldots(n-m_j+2)\lambda_j^{n-m_j+1})_n \quad (j = 1, 2, \ldots, \ell)$$

eine solche Basis.

Lemma 1.4

Für die homogene Differenzengleichung

$$y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = 0 \qquad (n = 0, 1, 2, \ldots) \tag{$*$}$$

sind die folgenden drei Aussagen äquivalent :

(1) Jede Lösung $\{y_n\}_n$ von $(*)$ ist beschränkt.

(2) Für jede Lösung $\{y_n\}_n$ von $(*)$ ist $\{y_n/n\}_n$ eine Nullfolge.

(3) Das zugehörige charakteristische Polynom p erfüllt die sogenannte Stabilitätsbedingung:

$$\begin{aligned} p(\lambda) = 0 &\;\Rightarrow\; |\lambda| \leqslant 1, \\ p(\lambda) = 0 \text{ und } |\lambda| = 1 &\;\Rightarrow\; \lambda \text{ ist einfach.} \end{aligned} \tag{Stab}$$

Lemma 1.5

Es gebe Konstanten $M, K \geqslant 0$, so dass die ersten Glieder der Vektorfolge $(\boldsymbol{y}_n)_n$ die Ungleichung

$$\|\boldsymbol{y}_{n+1}\| \leqslant K\|\boldsymbol{y}_n\| + M \qquad (n = 0, 1, \ldots, n_0)$$

erfüllen. Dann gilt die Abschätzung

$$\|\boldsymbol{y}_{n+1}\| \leqslant K^{n+1}\|\boldsymbol{y}_0\| + \begin{cases} M\dfrac{K^{n+1}-1}{K-1}, & \text{für } K \neq 1, \\ (n+1)M, & \text{für } K = 1, \end{cases} \qquad (n = 0, 1, \ldots, n_0).$$

($\|\cdot\|$ bezeichnet eine beliebige Norm.)

Gesucht ist eine explizite Darstellung der Lösung $(z_n)_n$ der inhomogenen Differenzengleichung

$$y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_1 y_{n+1} + \alpha_0 y_n = \beta_{n+k}, \qquad n \in \mathbb{N}_0,$$

die die k Anfangsbedingungen $z_n = y_n$ $(n = 0, 1, \ldots, k-1)$ erfüllt.

Antwort: Bezeichnen $(y_n^{(j)})_n$, $j = 0, 1, \ldots, k-1$, die Lösungen der homogenen Gleichung

$$y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_1 y_{n+1} + \alpha_0 y_n = 0 \qquad (n = 0, 1, \ldots),$$

die die Anfangsbedingungen $y_n^{(j)} = \delta_{n,j}$ (Kronecker-Symbol) $(n, j = 0, 1, \ldots, k-1)$ erfüllen, so ist

$$z_n = \sum_{j=0}^{k-1} y_j y_n^{(j)} + \sum_{j=0}^{n-k} \beta_{j+k} y_{n-j-1}^{(k-1)} \qquad (n = 0, 1, \ldots),$$

wobei $\beta_{n+k} = 0$ und $y_n^{(k-1)} = 0$ für $n < 0$ gesetzt wird.

Inhalt

In diesem Abschnitt sei $A \in \mathbb{C}^{n\times n}$ stets eine quadratische Matrix.
Außerdem sei eine Funktion

$$f : D \to \mathbb{C}, \qquad D \subset \mathbb{C},$$

gegeben. Wir klären hier, wann und wie die Matrix

$$f(A) \in \mathbb{C}^{n\times n}$$

definiert ist, und wiederholen einige ihrer Eigenschaften. Im Zusammenhang mit GDGen von Interesse ist besonders $\exp(A)$, die Exponentialfunktion angewandt auf A.

Für einige elementare Funktionen f ist $f(A)$ kanonisch gegeben. Ist z.B. $f \in \mathscr{P}_m$ ein Polynom vom Grad m,

$$f(\lambda) = \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2 + \cdots + \alpha_m\lambda^m \qquad (\alpha_j \in \mathbb{C},\ j = 0, 1, \dots, m),$$

so ist

$$f(A) = \alpha_0 I + \alpha_1 A + \alpha_2 A^2 + \cdots + \alpha_m A^m \in \mathbb{C}^{n\times n}.$$

Lemma 1.6

Sei $f \in \mathscr{P}_m$.

(a) Hat $A = \operatorname{diag}(A_{1,1}, A_{2,2}, \dots, A_{k,k})$ Blockdiagonalstruktur mit quadratischen Diagonalblöcken

$$A_{j,j} \in \mathbb{C}^{n_j \times n_j}, \qquad (j = 1, 2, \dots, k), \quad n_1 + n_2 + \dots + n_k = n,$$

dann gilt

$$f(A) = \operatorname{diag}(f(A_{1,1}), f(A_{2,2}), \dots, f(A_{k,k})).$$

(b) Ist $T \in \mathbb{C}^{n \times n}$ invertierbar und $B := TAT^{-1}$, dann gilt

$$f(B) = Tf(A)T^{-1}.$$

(c) Ist λ ein Eigenwert von A mit zugehörigem Eigenvektor $\boldsymbol{v}$, so ist $f(\lambda)$ ein Eigenwert von $f(A)$ mit zugehörigem Eigenvektor $\boldsymbol{v}$:

$$A\boldsymbol{v} = \lambda \boldsymbol{v} \qquad \Longrightarrow \qquad f(A)\boldsymbol{v} = f(\lambda)\boldsymbol{v}.$$

Wir bestimmen $m_k(J)$ für das k-te Monom $m_k(\lambda) = \lambda^k$ und einen Jordan-Block[4]

$$J = J(\lambda) = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{bmatrix} \in \mathbb{C}^{n\times n}.$$

Eine elementare Rechnung zeigt, dass $m_k(J) = J^k$ eine obere Dreiecksmatrix mit Toeplitz-Struktur[5] ist. Der Eintrag in der j-ten Diagonale ist

$$\binom{k}{j}\lambda^{k-j} = \frac{k(k-1)\cdots(k-j+1)\lambda^{k-j}}{j!} = \frac{m_k^{(j)}(\lambda)}{j!} \qquad (j = 0, 1, \ldots, n-1).$$

[4] Camille Jordan (1838–1922)
[5] Otto Toeplitz (1881–1940)

Mit anderen Worten:

$$m_k(J) = J^k = \begin{bmatrix} m_k(\lambda) & m_k'(\lambda) & \cdots & \frac{m_k^{(n-2)}(\lambda)}{(n-2)!} & \frac{m_k^{(n-1)}(\lambda)}{(n-1)!} \\ & m_k(\lambda) & \cdots & \frac{m_k^{(n-3)}(\lambda)}{(n-3)!} & \frac{m_k^{(n-2)}(\lambda)}{(n-2)!} \\ & & \ddots & \vdots & \vdots \\ & & & m_k(\lambda) & m_k'(\lambda) \\ & & & & m_k(\lambda) \end{bmatrix}.$$

Jetzt sind wir in der Lage, $f(A)$ für beliebiges f zu definieren: Sei dazu $J_A = \operatorname{diag}(J_1, J_2, \dots, J_k)$ die Jordansche Normalform von A, $A = TJ_AT^{-1}$. Die einzelnen Jordan-Blöcke $J_j = J_j(\lambda_j)$ seien $(n_j \times n_j)$-Matrizen. Das charakteristische Polynom c_A von A hat dann die Form

$$c_A(\lambda) = \prod_{j=1}^{k} (\lambda - \lambda_j)^{n_j}.$$

Wir sagen *f ist auf A definiert*, wenn f auf einer offenen Menge D definiert ist, die das Spektrum $\Lambda(A) = \{\lambda_1, \dots, \lambda_k\}$ von A enthält, und außerdem f in λ_j $(n_j - 1)$-mal differenzierbar ist.

In diesem Fall setzen wir für $j = 1, 2, \dots, k$

$$f(J_j(\lambda_j)) := \begin{bmatrix} f(\lambda_j) & f'(\lambda_j) & \cdots & \frac{f^{(n_j-2)}(\lambda_j)}{(n_j-2)!} & \frac{f^{(n_j-1)}(\lambda_j)}{(n_j-1)!} \\ & f(\lambda_j) & \cdots & \frac{f^{(n_j-3)}(\lambda_j)}{(n_j-3)!} & \frac{f^{(n_j-2)}(\lambda_j)}{(n_j-2)!} \\ & & \ddots & \vdots & \vdots \\ & & & f(\lambda_j) & f'(\lambda_j) \\ & & & & f(\lambda_j) \end{bmatrix} \in \mathbb{C}^{n_j \times n_j}$$

und

$$f(A) := T \operatorname{diag}\big(f(J_1), f(J_2), \dots, f(J_k)\big)\, T^{-1}.$$

Bemerkungen.

(1) Ist $f(\lambda) = \alpha_0 + \alpha_1\lambda + \cdots + \alpha_m\lambda^m$ ein Polynom, so gilt für die so definierte Matrix $f(A)$: $f(A) = \alpha_0 I + \alpha_1 A + \cdots + \alpha_m A^m$ (natürlich ist ein Polynom auf jeder Matrix definiert).

(2) Ist $f = p/q$ eine rationale Funktion mit $p \in \mathscr{P}_m$ und $q \in \mathscr{P}_k$, so ist f genau dann auf A definiert, wenn kein Eigenwert von A eine Polstelle von f ist. In diesem Fall gilt $f(A) = p(A)[q(A)]^{-1} = [q(A)]^{-1}p(A)$.

(3) Alle Aussagen von Lemma 1.6 gelten für jede beliebige Funktion f, wenn nur f auf A definiert ist.

(4) Ist f analytisch in einer Umgebung von 0 und besitzt dort die Taylor[6]-Reihe $f(\lambda) = \sum_{j=0}^{\infty} \alpha_j \lambda^j$ mit Konvergenzradius $\tau > 0$ ($\tau = \infty$ ist erlaubt), so ist f auf jeder Matrix A mit Spektralradius $\rho(A) < \tau$ definiert und es gilt

$$f(A) = \sum_{j=0}^{\infty} \alpha_j A^j = \lim_{m\to\infty} \sum_{j=0}^{m} \alpha_j A^j.$$

[6] Brook Taylor (1685–1731)

Ein bekanntes Beispiel für die letzte Bemerkung ist die Neumannsche Reihe[7]

$$(I - A)^{-1} = \sum_{j=0}^{\infty} A^j, \qquad \text{falls } \rho(A) < 1.$$

Die Exponentialfunktion einer Matrix A kann z.B. auch durch

$$\exp(A) = \sum_{j=0}^{\infty} \frac{1}{j!} A^j$$

definiert werden. (Die Reihe konvergiert für jede Matrix A, weil die zugehörige skalare Reihe einen unendlichen Konvergenzradius besitzt.)
Bei GDGen ist es oft wichtig zu wissen, wie sich $\exp(tA)$ für $t \to \infty$ verhält. Die entscheidende Größe ist die Spektralabszisse $\alpha(A)$ von A:

$$\alpha(A) := \max\{\mathrm{Re}(\lambda) \,:\, \lambda \in \Lambda(A)\}.$$

[7] Carl Neumann (1832–1925)

Satz 1.7 (Asymptotisches Verhalten von $\exp(tA)$)

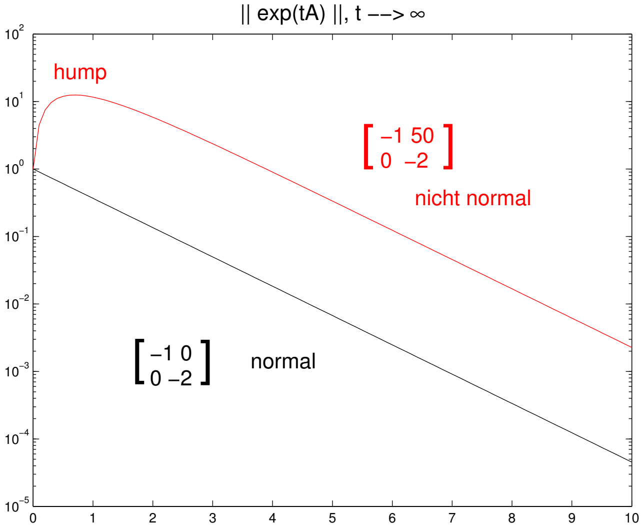
Sei $A \in \mathbb{C}^{n \times n}$.

(a) Es ist $\lim_{t\to\infty} \exp(tA) = 0$ genau dann, wenn $\alpha(A) < 0$ gilt.

(b) Wenn $\alpha(A) > 0$ ist, so ist $\exp(tA)$ für $t \to \infty$ unbeschränkt.
Ist $\alpha(A) = 0$ und jeder Eigenwert λ von A mit $\operatorname{Re}\lambda = \alpha(A)$ halbeinfach, so ist $\exp(tA)$ für $t \to \infty$ beschränkt (aber i.A. nicht konvergent).

(c) Es gilt $\|\exp(tA)\| \geqslant \exp(t\alpha(A))$ für jede Matrixnorm.
Ist A normal, dann gilt $\|\exp(tA)\|_2 = \exp(t\alpha(A))$.

Für normale Matrizen ist $\|\exp(tA)\|_2$ also eine streng monoton fallende Funktion von $t \geqslant 0$, wenn $\alpha(A) < 0$.
Ist A nicht normal, so beobachtet man die üblichen Nichtnormalitätseffekte (vollkommen analog zum Verhalten von $\|A^m\|_2$, $m \to \infty$, falls $\rho(A) < 1$).



|| exp(tA) ||, t --> ∞
hump
$\begin{bmatrix} -1 & 50 \\ 0 & -2 \end{bmatrix}$
nicht normal
$\begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$
normal
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0 1 2 3 4 5 6 7 8 9 10

Zum Abschluss soll noch ein Algorithmus zur Berechnung von $\exp(A)$ beschrieben werden.
Beachte: $\exp(A) = \lim_{m\to\infty} \sum_{j=0}^{m} A^j/j!$ ist nur geeignet, wenn $\rho(A)$ sehr klein ist; die Bestimmung von $\exp(A)$ über die Jordansche Normalform von A ist numerisch instabil oder zu aufwendig.
Der Algorithmus verwendet rationale Approximationen $(k/\ell)_{\exp}(\zeta) = p_{k,\ell}(\zeta)/q_{k,\ell}(\zeta)$ vom Typ (k,ℓ) (d.h. $p_{k,\ell} \in \mathscr{P}_k$, $q_{k,\ell} \in \mathscr{P}_\ell$) an die Exponentialfunktion, sog. Padé-Approximationen[8]. Diese sind eindeutig bestimmt durch die Vorschrift

$$\exp(\zeta) - p_{k,\ell}(\zeta)/q_{k,\ell}(\zeta) = O\left(\zeta^{k+\ell+1}\right) \quad \text{für } \zeta \to 0.$$

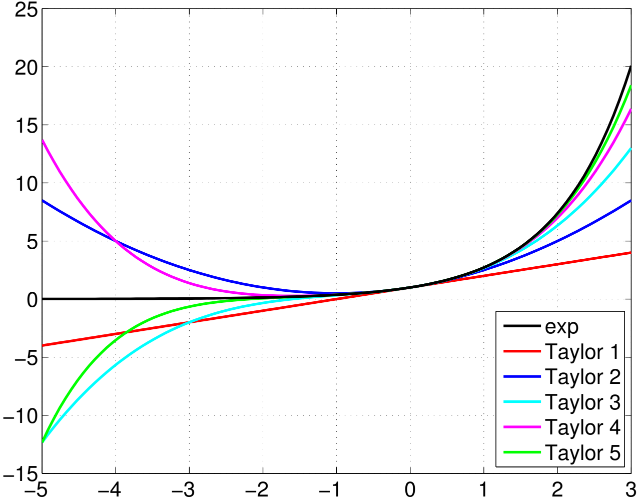
Beachte, dass hier die Taylor-Polynome für $\ell = 0$ als Spezialfall enthalten sind. Man kann die Padé-Approximationen im Fall der Exponentialfunktion explizit angeben:

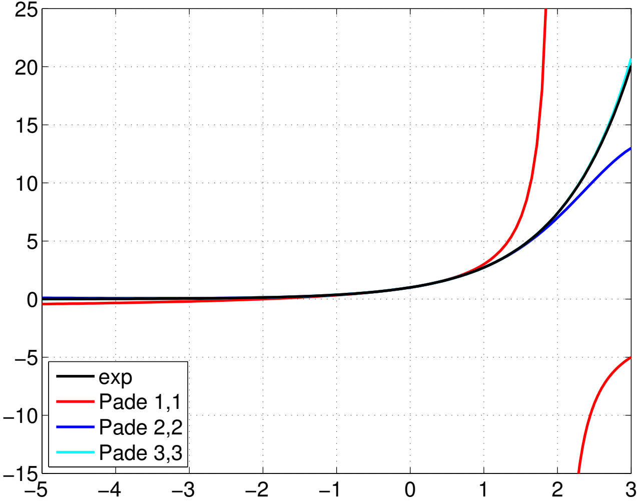
$$p_{k,\ell}(\zeta) = \sum_{j=0}^{k} \frac{(k+\ell-j)!k!}{(k+\ell)!j!(k-j)!}\,\zeta^j, \qquad q_{k,\ell}(\zeta) = \sum_{j=0}^{\ell} \frac{(k+\ell-j)!\ell!}{(k+\ell)!j!(\ell-j)!}\,\zeta^j.$$

[8] Henri Padé (1863–1953)

Matrixfunktionen

Taylorpolynome der Exponentialfunktion





25
20
15
10
5
0
−5
−10
−15
−5
−4
−3
−2
−1
0
1
2
3
exp
Pade 1,1
Pade 2,2
Pade 3,3

Satz 1.8 (Fehlerformel für Padé-Approximationen)

Für $k, \ell \in \mathbb{N}_0$ und $A \in \mathbb{C}^{n\times n}$ gilt

$$\exp(A) - (k,\ell)_{\exp}(A) = \frac{(-1)^\ell}{(k+\ell)!} A^{k+\ell+1} [q_{k,\ell}(A)]^{-1} \int_0^1 u^k (1-u)^\ell \exp((1-u)A)\,\mathrm{d}u.$$

Das bedeutet, dass $(k,\ell)_{\exp}(A)$

- sowohl für $k \to \infty$ bei festem ℓ,
- als auch für $\ell \to \infty$ bei festem k,
- als auch für $k \to \infty$ bei festem $k - \ell$

gegen $\exp(A)$ strebt.

Integrale über Matrizen $A = [a_{i,j}(u)]$ sind komponentenweise definiert:

$$\int A(u)\,\mathrm{d}u = \left[\int a_{i,j}(u)\,\mathrm{d}u\right].$$

Satz 1.8 zeigt, dass auch $(k,\ell)_{\exp}(A)$ nur dann eine akzeptable Näherung für $\exp(A)$ ist, wenn $\rho(A)$ nicht zu groß ist.

Daher verwendet man einen Trick: Kommutieren die Matrizen A, $B \in \mathbb{C}^{n\times n}$, so gilt

$$\exp(A+B) = \exp(A)\exp(B).$$

Insbesondere ist also

$$\exp(A) = \exp(A/m)^m \qquad \text{für } m = 0, 1, 2, \dots$$

Das bedeutet, dass

$$E_{k,\ell} := \left[(k,\ell)_{\exp}(A/2^j)\right]^{2^j}$$

eine Approximation an $\exp(A)$ darstellt, bei der die Padé-Approximation an der Matrix $A/2^j$ ausgewertet wird, deren Spektralradius $\rho(A)/2^j$ man durch die Wahl von j steuern kann.
Die Berechnung von $E_{k,\ell}$ erfordert $j + \max\{k,\ell\}$ Multiplikationen mit A.

Lemma 1.9

Sei $\|A\|_\infty/2^j \leqslant 1/2$. Dann ist

$$\frac{\|\exp(A) - E_{k,\ell}\|_\infty}{\|\exp(A)\|_\infty} \leqslant \varepsilon(k,\ell)\|A\|_\infty \exp(\varepsilon(k,\ell)\|A\|_\infty)$$

mit

$$\varepsilon(k,\ell) = 2^{3-(k+\ell)} \frac{k!\ell!}{(k+\ell)!(k+\ell+1)!}.$$

Bei festem $d = \max\{k,\ell\}$ (Arbeitsaufwand zur Berechnung von $E_{k,\ell}$), wird $\varepsilon(k,\ell)$ durch die Wahl $k = \ell = d$ minimiert.

Algorithmus 1 : Berechnung von $\exp(A)$.

Gegeben : A, $\delta > 0$.

$j \leftarrow \max\{0,\ 1 + \mathrm{floor}(\log_2 \|A\|_\infty)\}$.
$A \leftarrow A/2^j$.
Wähle ℓ minimal mit $\epsilon(\ell, \ell) \leqslant \delta$.
$N \leftarrow I$, $Z \leftarrow I$, $X \leftarrow I$, $c \leftarrow 1$.
for $m = 1$ **to** ℓ **do**
  $c \leftarrow c(\ell - m + 1)/((2\ell - m + 1)m)$.
  $X \leftarrow AX$; $Z \leftarrow Z + cX$; $N \leftarrow N + (-1)^m cX$.
Bestimme LU-Zerlegung von N und löse damit $NE = Z$ nach E auf.
for $m = 1$ **to** j **do**
  $E \leftarrow EE$

- Dieser Algorithmus liefert eine Approximation $E \approx \exp(A)$ derart, dass

$$E = \exp(A + \Delta A), \quad \text{wobei} \quad \|\Delta A\|_\infty \leqslant \delta \|A\|_\infty.$$

- Seine Komplexität beträgt etwa $2(\ell + j + 1/3)n^3$ flops, vgl. [Moler & Van Loan, 2003][9] und [Ward, 1977][10]
- Für große dünnbesetzte Matrizen ist unser Algorithmus jedoch ungeeignet (er verwendet die LU-Zerlegung einer Matrix der Dimension von A).
- Ähnlich wie bei linearen Gleichungssystemen, wo man selten an A^{-1} sondern vielmehr an $A^{-1}\boldsymbol{b}$, $\boldsymbol{b} \in \mathbb{C}^n$, interessiert ist, steht auch hier die Berechnung von $\exp(A)\boldsymbol{b}$ im Vordergrund. Bei großen dünn besetzten Problemen muss man auch dazu iterative Verfahren verwenden.

---

[9] C.B. Moler und C.F. Van Loan. *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev.* **45**, 3–49 (2003)

[10] R.C. Ward. *Numerical computation of the matrix exponential with accuracy estimate. SIAM J. Numer. Anal.* **14** (4) 600–610 (1977).

Inhalt

Systeme linearer Differentialgleichungen erster Ordnung

Nur wenige Systeme der Form (DG) bzw. AWPe der Form (AWP) kann man explizit lösen. Selbst lineare Systeme erster Ordnung,

$$\begin{aligned} y_1' &= a_{1,1}(t)y_1 + a_{1,2}(t)y_2 + \cdots + a_{1,n}(t)y_n + b_1(t), \\ y_2' &= a_{2,1}(t)y_1 + a_{2,2}(t)y_2 + \cdots + a_{2,n}(t)y_n + b_2(t), \\ \vdots &= \vdots \\ y_n' &= a_{n,1}(t)y_1 + a_{n,2}(t)y_2 + \cdots + a_{n,n}(t)y_n + b_n(t) \end{aligned}$$

oder kürzer

$$\boldsymbol{y}' = A(t)\boldsymbol{y} + \boldsymbol{b}(t) \quad \text{mit } A(t) = [a_{i,j}(t)] \text{ und } \boldsymbol{b}(t) = [b_j(t)], \tag{Lin}$$

gehören nur unter weiteren Einschränkungen zu diesen Ausnahmefällen.

Sind die Funktionen $a_{i,j}(t)$, $b_j(t)$ stetig über einem Intervall I und ist $\|A(t)\| \leqslant L$ für alle $t \in I$ (was wir ab jetzt stets voraussetzen), so besitzt (Lin) nach Satz 1.1 für jede Wahl der Anfangsbedingungen

$$\boldsymbol{y}(t_0) = \boldsymbol{y}_0 \qquad (t_0 \in I)$$

eine eindeutige Lösung.

Satz 1.10 (Lösungen linearer Systeme erster Ordnung)

Die Lösungen des homogenen Systems

$$\boldsymbol{y}' = A(t)\boldsymbol{y}$$

bilden einen n-dimensionalen Unterraum des $C^1(I)$. Die Differenz zweier Lösungen des inhomogenen Systems

$$\boldsymbol{y}' = A(t)\boldsymbol{y} + \boldsymbol{b}(t)$$

löst das zugehörige homogene System.

Im Spezialfall konstanter Koeffizienten

$$a_{i,j}(t) = a_{i,j} \qquad \text{für alle } t$$

lassen sich diese Lösungen angeben. Dazu betrachten wir zunächst den homogenen Fall, $\boldsymbol{b}(t) \equiv \boldsymbol{0}$: Es seien $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n$ die Einheitsvektoren im $\mathbb{R}^n$. Für $j = 1, 2, \dots, n$ löst

$$\boldsymbol{x}_j(t) := \exp(tA)\boldsymbol{u}_j$$

das AWP

$$\boldsymbol{y}' = A\boldsymbol{y}, \qquad \boldsymbol{y}(0) = \boldsymbol{u}_j.$$

Darüber hinaus sind die Funktionen $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t), \dots, \boldsymbol{x}_n(t)$ linear unabhängig und bilden deshalb eine Basis des Lösungsraums von $\boldsymbol{y}' = A\boldsymbol{y}$.
Schließlich ist die matrixwertige Funktion

$$X : \mathbb{R} \to \mathbb{R}^{n\times n}, \qquad X(t) := [\boldsymbol{x}_1(t)|\boldsymbol{x}_2(t)|\dots|\boldsymbol{x}_n(t)]$$

für alle $t \in \mathbb{R}$ invertierbar und löst das Anfangswertproblem

$$X'(t) = AX(t), \qquad X(0) = I.$$

Um das inhomogene Anfangswertproblem $\boldsymbol{y}' = A\boldsymbol{y} + \boldsymbol{b}(t)$ (der Einfachheit halber nehmen wir an, dass die Komponenten $b_j(t)$ auf ganz $\mathbb{R}$ stetig sind), $\boldsymbol{y}(0) = \boldsymbol{y}_0$, zu lösen, bedient man sich einer Technik, die unter dem Namen Variation der Konstanten bekannt ist:
Die (eindeutige) Lösung ist gegeben durch

$$\boldsymbol{y}(t) = \sum_{j=1}^{n} \left[\int_{t_0}^{t} \frac{W_j(s)}{W(s)} \,\mathrm{d}s + \boldsymbol{y}_{0,j} \right] \boldsymbol{x}_j(t),$$

mit den Wronski-Determinanten[11]

$$\begin{aligned} W(t) &= \det \begin{bmatrix} \boldsymbol{x}_1(t) & \boldsymbol{x}_2(t) & \cdots & \boldsymbol{x}_n(t) \end{bmatrix}, \\ W_j(t) &= \det \begin{bmatrix} \boldsymbol{x}_1(t) & \cdots & \boldsymbol{x}_{j-1}(t) & \boldsymbol{b}(t) & \boldsymbol{x}_{j+1}(t) & \cdots & \boldsymbol{x}_n(t) \end{bmatrix}. \end{aligned}$$

Bemerkung: Lautet die Anfangsbedingung $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, so müssen anstelle der Funktionen $\boldsymbol{x}_j$ die Funktionen $\tilde{\boldsymbol{x}}_j(t) := \exp((t - t_0)A)\boldsymbol{u}_j$ verwendet werden.

[11] Josef Wronski (1778–1853)

Eine weitere Lösungsdarstellung für inhomogene lineare Systeme mit konstanten Koeffizienten, ebenfalls unter der Bezeichnung Variation der Konstanten bekannt, lautet (mit Anfangsbedingung bei $t = t_0$)

$$\boldsymbol{y}(t) = \exp((t - t_0)A)\boldsymbol{y}_0 + \int_{t_0}^{t} \exp((t-\tau)A)\, \boldsymbol{b}(\tau)\, \mathrm{d}\tau. \tag{1.1}$$

Beachte: Diese Formel gilt auch wenn $\boldsymbol{b}$ neben t auch von $\boldsymbol{y}$ abhängt, d.h.

$$\boldsymbol{b} = \boldsymbol{b}(t, \boldsymbol{y}(t)).$$

Formel (1.1) gestattet auch die Lösung des linearisierten Problems: linearisiert man die Differentialgleichung $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ im Punkt $(t_0, \boldsymbol{y}_0)$, ergibt sich (multivariate Taylor-Entwicklung)

$$\boldsymbol{f}(t, \boldsymbol{y}) \approx \underbrace{\boldsymbol{f}(t_0, \boldsymbol{y}_0)}_{=:\boldsymbol{b}} + \underbrace{\boldsymbol{f}_t(t_0, \boldsymbol{y}_0)}_{=:\boldsymbol{a}}(t - t_0) + \underbrace{\boldsymbol{f}_{\boldsymbol{y}}(t_0, \boldsymbol{y}_0)}_{=:A}(\boldsymbol{y} - \boldsymbol{y}_0)$$

und, als Approximation in der Nähe des Linearisierungspunktes, das linearisierte AWP

$$\boldsymbol{y}'(t) = A(\boldsymbol{y} - \boldsymbol{y}_0) + (t - t_0)\boldsymbol{a} + \boldsymbol{b}, \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0. \tag{1.2}$$

Formel (1.1) liefert als Lösung von (1.2)

$$\boldsymbol{y}(t) = \boldsymbol{y}_0 + (t - t_0)A^{-1}\boldsymbol{a} + \left(e^{(t-t_0)A} - I\right)(A^{-1}\boldsymbol{b} + A^{-2}\boldsymbol{a}).$$

Inhalt

(vgl. [Braun, 1994])

- Im Mai 1945 entdeckten die Alliierten in der Kunstsammlung Hermann Görings ein bis dahin unbekanntes Gemälde von Jan Vermeer van Delft (1632–1675), nämlich „Christus und die Ehebrecherin“.
- Es dauerte nicht lange, bis der Maler Han van Meegeren als derjenige ermittelt wurde, der (über Mittelsmänner) dieses Bild an Göring verkauft hatte. Van Meegeren wurde wegen Kollaboration mit dem Feind verhaftet. Er behauptete daraufhin, dieses Bild sowie vier weitere mutmaßliche Vermeers, darunter „Christus und die Jünger in Emmaus“, selbst gemalt zu haben.
- Um diese Aussage zu bekräftigen, begann er im Gefängnis „Christus unter den Schriftgelehrten“ im Stil Vermeers zu malen. Er ging dabei sehr geschickt vor: Er kratzte von alten, wertlosen Gemälden die Farbe bis auf die Leinwand ab, vermischte die alte (und äußerst harte) Farbe mit Phenolformaldehyd, um mit ihr wieder malen zu können. Das fertige Bild wurde in einem Ofen erhitzt, wobei das Aldehyd zu Bakelit erstarrte.

- Noch vor Vollendung seiner Arbeit erfuhr van Meegeren, dass die Anklage auf Kollaboration fallen gelassen wurde und er stattdessen ein Verfahren wegen Fälschung zu erwarten hatte. Er weigerte sich daraufhin, die Vermeer-Kopie zu vollenden.
- Weil u.A. in einigen der angeblichen Vermeers-Bilder Phenolformaldehyd nachgewiesen werden konnte (eine Substanz, die bis zum Ende des 19. Jahrhunderts völlig unbekannt war), wurde van Meegeren trotzdem am 12.10.1947 zu einem Jahr Gefängnis wegen Fälschung verurteilt. Er starb kurz darauf in der Haft.
- Dessen ungeachtet waren viele Experten immer noch der Meinung, dass es sich bei „Christus und die Jünger in Emmaus“ um einen echten Vermeer handelt (aufgrund der Expertise eines bekannten Kunsthistorikers erwarb die Rembrandt-Gesellschaft dieses Werk für 174.000 US-$).
- Der Streit um die Authentizität dieses Gemäldes sollte schließlich 1967 von einer Forschergruppe an der Carnegie Mellon Universität (Pittsburgh, PA) entschieden werden.

Deren Analyse basierte auf der Tatsache, dass Künstler seit mehr als 2000 Jahren sog. Bleiweiß (Bleioxyd) verwenden, das kleine Bestandteile an radioaktivem Blei-210 und Radium-226 enthält. Um die Pittsburgher Analyse zu verstehen, sind elementare Kenntnisse über radioaktiven Zerfall erforderlich.

- Unter Radioaktivität versteht man den (ohne äußere Beeinflussung erfolgenden) Zerfall instabiler Atomkerne gewisser radioaktiver Substanzen.
- Für jede radioaktive Substanz gibt es eine charakteristische Übergangswahrscheinlichkeit λ (Zerfallskonstante), mit der im Mittel ein Atom pro Zeiteinheit zerfällt. Sind zur Zeit t also $N(t)$ radioaktive Atome vorhanden, so zerfallen im Zeitintervall $[t, t + \Delta t]$ durchschnittlich $\lambda N(t)\Delta t$ Atome.
- Für $\Delta t \to 0$ erhalten wir das Zerfallsgesetz

$$N'(t) = -\lambda N(t).$$

Die Zahl der nach einer gewissen Zeit Δt, die seit dem Zeitpunkt t_0 verstrichen ist, noch vorhanden radioaktiven Atome ist deshalb

$$N(t_0 + \Delta t) = N(t_0) \exp(-\lambda \Delta t).$$

Die Halbwertszeit, d.h. die Zeitspanne, innerhalb der die Hälfte einer gegebenen Menge radioaktiver Atome zerfällt, ergibt sich damit [setze $N(t_0 + \Delta t)/N(t_0) = 1/2$] zu

$$T_{1/2} = \Delta t = \log(2)/\lambda.$$

Da sich die Zerfallsprodukte radioaktiver Stoffe weiter umwandeln, bis ein stabiles Endglied gebildet ist, entstehen sog. Zerfallsreihen.

Für uns ist die Zerfallsreihe von Uran-238 relevant[12]:

$$
\begin{aligned}
\text{Uran-238} &\xrightarrow{4.5\cdot 10^9\text{a}} \text{Thorium-234} \xrightarrow{24\text{d}} \text{Protaktinium-234} \xrightarrow{1.2\text{min}} \text{Uran-234}\\
&\xrightarrow{2.5\cdot 10^5\text{a}} \text{Thorium-230} \xrightarrow{8\cdot 10^4\text{a}} \text{Radium-226} \xrightarrow{1.6\cdot 10^3\text{a}} \text{Radon-222}\\
&\xrightarrow{3.8\text{d}} \text{Polonium-218} \xrightarrow{3\text{min}} \text{Blei-214} \xrightarrow{27\text{min}} \text{Wismuth-214} \xrightarrow{20\text{min}} \text{Polonium-214}\\
&\xrightarrow{1.5\cdot 10^{-4}\text{s}} \text{Blei-210} \xrightarrow{22\text{a}} \text{Wismuth-210} \xrightarrow{5\text{d}} \text{Polonium-210}\\
&\xrightarrow{140\text{d}} \text{Blei-206 (nicht radioaktiv)}
\end{aligned}
$$

[12] Halbwertszeiten in Jahren [a], Tagen [d], Minuten [min] oder Sekunden [s])

Innerhalb einer Zerfallsreihe stellt sich für die Zwischensubstanzen im Laufe der Zeit ein Gleichgewichtszustand ein, das radioaktive Gleichgewicht. Voraussetzung dafür ist, dass das Ausgangselement so langsam zerfällt, dass seine Menge als konstant (oder seine Zerfallskonstante als 0) betrachtet werden kann. Um dies zu verdeutlichen, werden wir die Zerfallsreihe von Uran-238 etwas ökonomisieren,

$$\text{Uran-238} \xrightarrow{4.5\cdot 10^9\text{a}} \text{Radium-226} \xrightarrow{1.6\cdot 10^3\text{a}} \text{Blei-210} \xrightarrow{22.4\text{a}} \text{Blei-206},$$

so dass sie nur noch aus vier Elementen besteht. Wir bezeichnen mit $N_1(t)$, $N_2(t)$, $N_3(t)$ die Anzahl der Atome von Uran-238, Radium-226 bzw. Blei-210 zur Zeit t und berechnen aus den angegebenen Halbwertszeiten die zugehörigen Zerfallskonstanten,

$$\lambda_1 = 1.54 \cdot 10^{-10},\ \lambda_2 = 4.33 \cdot 10^{-4},\ \lambda_3 = 3.09 \cdot 10^{-2} \text{ (gemessen in } a^{-1}\text{).}$$

Zu lösen ist damit das System

$$
\begin{aligned}
N_1'(t) &= -\lambda_1 N_1(t),\\
N_2'(t) &= -\lambda_2 N_2(t) + \lambda_1 N_1(t),\\
N_3'(t) &= -\lambda_3 N_3(t) + \lambda_2 N_2(t),
\end{aligned}
$$

oder kürzer $\boldsymbol{N}'(t) = A\boldsymbol{N}(t)$ (mit Anfangsbedingungen $\boldsymbol{N}(t_0) = \boldsymbol{N}_0$), wobei

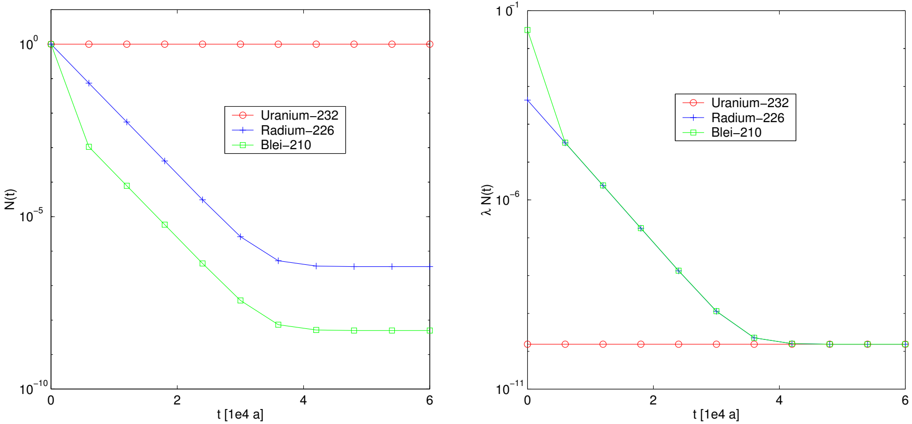
$$
A = \begin{bmatrix} -\lambda_1 & 0 & 0 \\ \lambda_1 & -\lambda_2 & 0 \\ 0 & \lambda_2 & -\lambda_3 \end{bmatrix}.
$$

Wir können die Lösung dieses AWPs explizit angeben:

$$
\boldsymbol{N}(t) = \exp(tA)\boldsymbol{N}_0.
$$

Die Fälschungen des Han van Meegeren

Radioaktives Gleichgewicht des vereinfachten Zerfallssystems



t [a]	$N_1(t)$	$N_2(t)$	$N_3(t)$
0	$1 \cdot 10^0$	$1 \cdot 10^0$	$1 \cdot 10^0$
10^1	$1 \cdot 10^0$	$1 \cdot 10^0$	$7 \cdot 10^{-1}$
10^3	$1 \cdot 10^0$	$6 \cdot 10^{-1}$	$9 \cdot 10^{-3}$
10^5	$1 \cdot 10^0$	$4 \cdot 10^{-7}$	$9 \cdot 10^{-9}$
10^6	$1 \cdot 10^0$	$4 \cdot 10^{-7}$	$5 \cdot 10^{-9}$
10^7	$1 \cdot 10^0$	$4 \cdot 10^{-7}$	$5 \cdot 10^{-9}$
10^8	$1 \cdot 10^0$	$4 \cdot 10^{-7}$	$5 \cdot 10^{-9}$
10^9	$9 \cdot 10^{-1}$	$3 \cdot 10^{-7}$	$4 \cdot 10^{-9}$
10^{11}	$2 \cdot 10^{-7}$	$7 \cdot 10^{-14}$	$1 \cdot 10^{-15}$
10^{12}	$1 \cdot 10^{-67}$	$5 \cdot 10^{-74}$	$6 \cdot 10^{-76}$

Natürlich gilt $\lim_{t\to\infty} N_j(t) = 0$ $(j = 1, 2, 3)$ (warum?), aber für eine sehr lange Periode (etwa $10^5 \leqslant t \leqslant 10^8$) scheint sich ein Gleichgewicht einzustellen. Die „Gleichgewichtswerte“ sind (ziemlich genau) proportional zu den Halbwertszeiten bzw. umgekehrt proportional zu den Zerfallskonstanten. Für $t = 10^7$ gilt in unserem Beispiel

$$\begin{aligned}
N_1(t)/N_2(t) &= 2.812499 \cdots 10^6 = \lambda_2/\lambda_1, \\
N_1(t)/N_3(t) &= 2.008927847142 \cdots 10^8 = \lambda_3/\lambda_1, \\
N_2(t)/N_3(t) &= 7.1428571 \cdots 10^1 = \lambda_3/\lambda_2.
\end{aligned}$$

Um diese Phänomene zu untersuchen, diagonalisieren wir A, $AT = TD$ mit $D = \operatorname{diag}(-\lambda_1, -\lambda_2, -\lambda_3)$ und

$$T = \begin{bmatrix} 1 & 0 & 0 \\ \frac{\lambda_1}{\lambda_2-\lambda_1} & 1 & 0 \\ \frac{\lambda_1\lambda_2}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)} & \frac{\lambda_2}{\lambda_3-\lambda_2} & 1 \end{bmatrix}.$$

Dann ergibt sich

$$\boldsymbol{N}(t) = \exp(tA)\boldsymbol{N}_0 = T\exp(tD)T^{-1}\boldsymbol{N}_0 = T\exp(tD)\tilde{\boldsymbol{N}}_0,$$

wobei wir $\tilde{\boldsymbol{N}}_0 = [\tilde{N}_1, \tilde{N}_2, \tilde{N}_3]^T := T^{-1}\boldsymbol{N}_0$ gesetzt haben. Entscheidend ist das Verhalten der einzigen Größe, die von t abhängt, nämlich von

$$\exp(tD) = \begin{bmatrix} \exp(-\lambda_1 t) & 0 & 0 \\ 0 & \exp(-\lambda_2 t) & 0 \\ 0 & 0 & \exp(-\lambda_3)t \end{bmatrix}.$$

Für $t \in [10^5, 10^8]$ gelten $\exp(-\lambda_1 t) \in [0.999984\ldots, 1]$ und $\exp(-\lambda_2 t), \exp(-\lambda_3 t) \in [0, 1.5\ldots 10^{-19}]$. In diesem Zeitintervall gilt also

$$\boldsymbol{N}(t) = T\exp(tD)\tilde{\boldsymbol{N}}_0 \approx T \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tilde{\boldsymbol{N}}_0 = \begin{bmatrix} \tilde{N}_1 \\ \frac{\lambda_1}{\lambda_2 - \lambda_1}\tilde{N}_1 \\ \frac{\lambda_1 \lambda_2}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)}\tilde{N}_1 \end{bmatrix}.$$

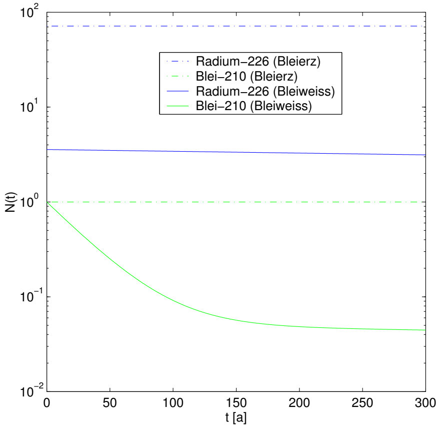
Beachtet man noch $\lambda_1 \ll \lambda_2 < \lambda_3$, so erhält man schließlich für $t \in [10^5, 10^8]$

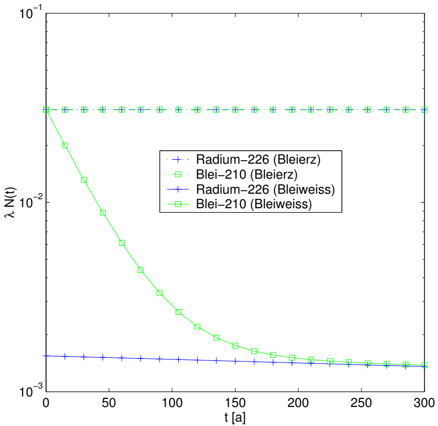
$$[N_1(t), N_2(t), N_3(t)]^T \approx \tilde{N}_1[1, \lambda_1/\lambda_2, \lambda_1/\lambda_3]^T.$$

Zurück zur Altersbestimmung von Gemälden: Wie bereits erwähnt, enthalten fast alle Gemälde Bleiweiß und damit die radioaktiven Substanzen Radium-226 und Blei-210. Bleiweiß wird aus Blei gewonnen, welches wiederum durch Schmelzen von Bleierz entsteht. Bei diesem Schmelzvorgang werden 90–95% des Radiums und seiner Tochtersubstanzen mit der Schlacke entfernt, so dass Blei-210 von seinem Nachschub abgeschnitten ist und sich mit Radium-226 nicht mehr im radioaktiven Gleichgewicht befindet. Das Blei-210 zerfällt dann sehr schnell (da es eine kurze Halbwertszeit von $T_{1/2} = 22$a besitzt), bis es mit den Resten von Radium-226 wieder im Gleichgewicht ist (nach ca. 200 Jahren).

Die Fälschungen des Han van Meegeren

Bleierz/Bleiweiß mit und ohne radioaktives Gleichgewicht





Seien jetzt t_0 der Zeitpunkt, an dem das Bleiweiß hergestellt wurde, und (wie oben) $N_2(t)$, $N_3(t)$ die Mengen von Radium-226 bzw. Blei-210 (pro g Bleiweiß). Es gilt $N_3'(t) = -\lambda_3 N_3(t) + \lambda_2 N_2(t)$. Da wir uns nur für eine Zeitspanne von 300 Jahren interessieren und Radium-226 eine Halbwertszeit von $T_{1/2} = 1600$ Jahren besitzt, können wir annehmen, dass seine Zerfallsrate (Aktivität) $\rho_2 := \lambda_2 N_2(t)$ konstant ist. Die GDG vereinfacht sich zu $N_3'(t) = -\lambda_3 N_3(t) + \rho_2$ bzw. zu $N_3'(t) + \lambda_3 N_3(t) = \rho_2$. Multiplizieren wir mit dem integrierenden Faktor $\exp(\lambda_3 t)$, so ergibt sich

$$\tfrac{d}{dt}[\exp(\lambda_3 t) N_3(t)] = \rho_2 \exp(\lambda_3 t),$$

was zu

$$N_3(t) = \frac{\rho_2}{\lambda_3}\left[1 - \exp(-\lambda_3(t - t_0))\right] + N_3(t_0)\exp(-\lambda_3(t - t_0)) \qquad (*)$$

führt. $N_3(t)$, λ_3 und ρ_2 sind bekannt (bzw. leicht zu messen). Wüssten wir die Größe von $N_3(t_0)$, könnten wir $t - t_0$ (und damit das Alter des Gemäldes) bestimmen.

Natürlich ist es unmöglich, $N_3(t_0)$ ohne Kenntnis von t_0 zu ermitteln. Wir machen von der Tatsache Gebrauch, dass $N_3(t_0)$, also die Menge an Blei-210 zum Zeitpunkt der Herstellung des Bleiweißes, ein radioaktives Gleichgewicht mit dem Radium-226 im Bleierz bildete. Lösen wir also ($*$) nach der Zerfallsrate $\lambda_3 N_3(t_0)$ von Blei-210 zur Zeit t_0 auf,

$$\lambda_3 N_3(t_0) = \lambda_3 N_3(t) \exp(\lambda_3(t - t_0)) - \rho_2 \left[\exp(\lambda_3(t - t_0)) - 1\right],$$

und nehmen $t - t_0 = 300$ a an,

$$\begin{aligned} \lambda_3 N_3(t_0) &= \lambda_3 N(t) \exp(300\lambda_3) - \rho_2 \left[\exp(300\lambda_3) - 1\right] \\ &= \lambda_3 N(t) 2^{150/11} - \rho_2 \left[2^{150/11} - 1\right] \end{aligned}$$

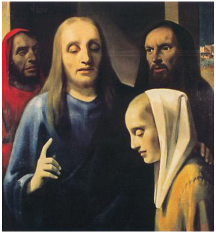
[$\exp(300\lambda_3) = \exp(300 \log(2)/T_{1,2}) = \exp(300 \log(2)/22) = 2^{150/11}$]. Um $\lambda_3 N_3(t_0)$ zu berechnen, müssen wir die gegenwärtigen Zerfallsraten $\lambda_3 N_3(t)$ von Blei-210 bzw. ρ_2 von Radium-226 bestimmen, was für einige mutmaßliche Bilder Vermeers geschehen ist.

	Zerfallsraten* von	
	Pb-210	Ra-226
„Christus und die Jünger in Emmaus“	8.5	0.80
„Fußwaschung“	12.6	0.26
„Die Notenleserin“	10.3	0.30
„Die Mandolinenspielerin“	8.2	0.17
„Die Spitzenklöpplerin“	1.5	1.40
„Der Soldat und das lachende Mädchen“	5.2	6.00

(* pro Minute und pro g Bleiweiß)

Die Fälschungen des Han van Meegeren

Hintergrund








Die Fälschungen des Han van Meegeren

Hintergrund

Legende:

(1,1) Han van Meegeren, „Christus und die Ehebrecherin“ (1941), ???

(1,2) Han van Meegeren, „Christus und die Jünger in Emmaus“ (1936/37), Museum Boymans Van Beunningen, Rotterdam

(1,3) Han van Meegeren, „Fußwaschung Christi“ (1941), Rijksmuseum, Amsterdam

(1,4) Han van Meegeren, „Die Notenleserin“ (1935/36), Rijksmuseum, Amsterdam

(2,1) Han van Meegeren, „Die Mandolinenspielerin“ (1935/36), Rijksmuseum, Amsterdam

(2,2) Jan Vermeer, „Die Spitzenklöpplerin“ (ca. 1669/70), Louvre, Paris

(2,3) Jan Vermeer, „Der Soldat und das lachende Mädchen“ (ca. 1658), Frick Collection, New York

(2,4) Jan Vermeer, „Brieflesendes Mädchen am offenen Fenster“ (ca. 1659), Gemäldegalerie „Alte Meister“, Dresden

Quellen: http://www.cacr.caltech.edu/ roy/vermeer, http://www.mystudios.com/gallery/har

Für „Christus und die Jünger in Emmaus“ ergibt sich

$$\lambda_3 N_3(t_0) = 8.5 \cdot 2^{150/11} - 0.8 \left[2^{150/11} - 1 \right] \approx 98050.$$

Es bleibt die Frage, ob dies ein akzeptabler Wert für die Zerfallsrate von Blei-210 im radioaktiven Gleichgewicht ist. Man kann nachrechnen, dass, wenn das Blei zur Zeit der Gewinnung mit einer Zerfallsrate von 100 pro Minute und g Bleiweiß zerfiel, das Erz, aus dem es stammt, einen Urananteil von 0.014 % hatte. Dies ist eine sehr hohe Urankonzentration. Andererseits gibt es (seltene) Erze, deren Urangehalt bei 2–3 % liegt. Um sicher zu gehen, nennen wir $\lambda_3 N_3(t_0)$ deshalb unakzeptabel hoch, wenn

$$\lambda_3 N_3(t_0) > 100 \cdot 3/0.014 \approx 22000$$

gilt, was bei „Christus und die Jünger in Emmaus“ offenbar der Fall ist (zum Vergleich beträgt der entsprechende Wert bei der „Spitzenklöpplerin“ $\lambda_3 N_3(t_0) \approx 1275$).

Inhalt

Wir betrachten die Bewegung eines Satelliten im Schwerefeld zweier großer Himmelskörper (z.B. Erde und Mond).

Annahmen:

- Die Bewegung aller drei Körper findet in einer Ebene statt; die beiden großen Körper rotieren in konstanter Entfernung und mit konstanter Winkelgeschwindigkeit um ihren gemeinsamen Schwerpunkt.
- Der Satellit hat somit keinen Einfluss auf die Bahnen von Erde und Mond.

Bezüglich eines mitrotierenden Koordinatensystems (in welchem Erde und Mond ruhen) mit Ursprung im gemeinsamen Schwerpunkt wird die Satellitenbahn $(x, y) = (x(t), y(t))$ beschrieben durch ein System zweier GDGen:

$$x'' = x + 2y' - \mu' \frac{x + \mu}{[(x + \mu)^2 + y^2]^{3/2}} - \mu \frac{x - \mu'}{[(x - \mu')^2 + y^2]^{3/2}},$$

$$y'' = y - 2x' - \mu' \frac{y}{[(x + \mu)^2 + y^2]^{3/2}} - \mu \frac{y}{[(x - \mu')^2 + y^2]^{3/2}}.$$

Weitere Beispiele

Satellit im Kraftfeld von Erde und Mond

- $\mu = 1/82.45$ bezeichnet den Anteil der Mondmasse an der Gesamtmasse von Erde und Mond, $\mu' = 1 - \mu$ die der Erde.
- Als Längeneinheit wählen wir die Erde-Mond-Entfernung, wobei der Mond auf der positiven und die Erde auf der negativen reellen Achse platziert werden.
- Die Zeiteinheit ist gegeben durch die Winkelgeschwindigkeit der Rotation, genauer rotiert der Mond ein Mal um die Erde in 2π Zeiteinheiten.
- Bekannt als restringiertes Dreikörperproblem, da der dritte Körper die ersten beiden nicht beeinflusst.
- Anfangsbedingung zur Zeit $t = 0$:

$$\begin{array}{ll} \text{Satellit in Position} & (x(0), y(0)) = (1.2, 0) \\ \text{mit Geschwindigkeit} & (x'(0), y'(0)) = (0, -1.05). \end{array}$$

Umschreiben in System erster Ordnung:

$$y_1 = x, \quad y_2 = y, \quad y_3 = x', \quad y_4 = y',$$

führt auf

$$\begin{aligned}
y_1' &= y_3, \\
y_2' &= y_4, \\
y_3' &= y_1 + 2y_4 - \mu' \frac{y_1 + \mu}{[(y_1 + \mu)^2 + y_2^2]^{3/2}} - \mu \frac{y_1 - \mu'}{[(y_1 - \mu')^2 + y_2^2]^{3/2}}, \\
y_4' &= y_2 - 2y_3 - \mu' \frac{y_2}{[(y_1 + \mu)^2 + y_2^2]^{3/2}} - \mu \frac{y_2}{[(y_1 - \mu')^2 + y_2^2]^{3/2}},
\end{aligned}$$

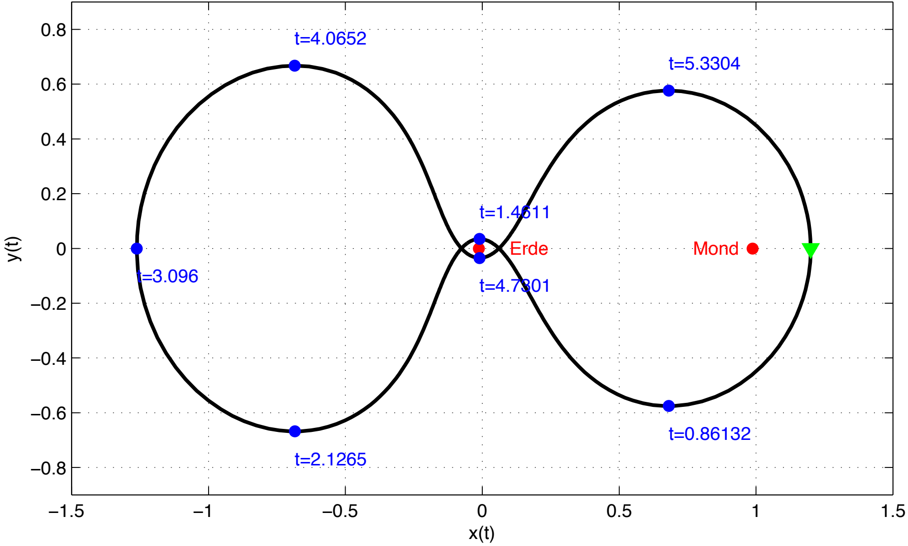
mit Anfangsbedingungen

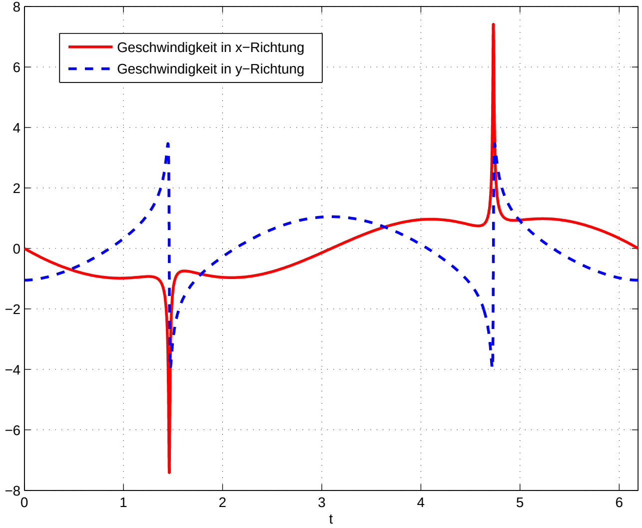
$$y_1(0) = 1.2, \quad y_2(0) = y_3(0) = 0 \quad \text{and} \quad y_4(0) = -1.05.$$

Lösung $(x(t), y(t)) = (y_1(t), y_2(t))$: geschlossene Bahn mir Periode $T \approx 6.19$.

Weitere Beispiele

Satellit im Kraftfeld von Erde und Mond





Geschwindigkeit in x−Richtung
Geschwindigkeit in y−Richtung
8
6
4
2
0
−2
−4
−6
−8
0
1
2
3
4
5
6
t

Drei Spezies S, T und U nehmen Teil an der autokatalytischen Reaktion

$$S \xrightarrow{k_1} T, \qquad T + U \xrightarrow{k_2} S + U, \qquad 2T \xrightarrow{k_3} T + U.$$

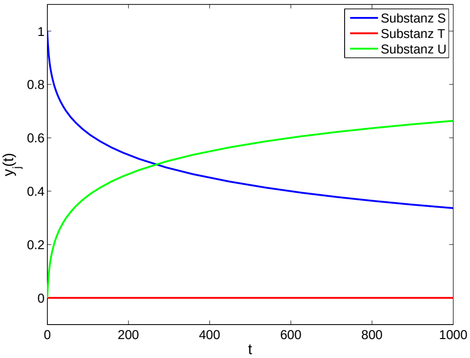
Zeitlicher Verlauf der Konzentrationen $y_1 = [S]$, $y_2 = [T]$, $y_3 = [U]$ beschrieben durch System GDGen (Massenwirkungsgesetz)

$$\begin{aligned} y_1' &= -k_1 y_1 + k_2 y_2 y_3, \\ y_2' &= k_1 y_1 - k_2 y_2 y_3 - 2k_3 y_2^2 + k_3 y_2^2 = k_1 y_1 - k_2 y_2 y_3 - k_3 y_2^2, \\ y_3' &= k_3 y_2^2. \end{aligned}$$

Reaktionsraten k_j sind ein Maß für die Geschwindigkeit mit denen die jeweilige Reaktion sich vollzieht. Sie differieren oft um mehrere Größenordnungen.

$$k_1 = 0.04, \quad k_2 = 10^4, \quad k_3 = 3 \cdot 10^7.$$

Für die Anfangebedingungen $y_1(0) = 1; y_2(0) = y_3(0) = 0$ erhalten wir:



Beachte: wegen $y_1'(t) + y_2'(t) + y_3'(t) = 0$ gilt für alle $t \geqslant t_0 = 0$, die Erhaltungsgleichung

$$y_1(t) + y_2(t) + y_3(t) = y_1(0) + y_2(0) + y_3(0) = 1.$$

Die Stabilitätsanalyse nichtlinearer dynamischer Systeme

$$\dot{\boldsymbol{u}} = \boldsymbol{f}(\boldsymbol{u}), \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0, \quad \boldsymbol{f} : \mathbb{C}^n \to \mathbb{C}^n, \tag{1.3}$$

geschieht meist durch Linearisierung um den (einen) stationären Zustand

$$\overline{\boldsymbol{u}} := \lim_{t\to\infty} \boldsymbol{u}(t).$$

Sofern ein solcher existiert ist er Lösung der Gleichung $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{0}$.

Das dynamische System (1.3) heißt lokal stabil in $\overline{\boldsymbol{u}}$, falls $\epsilon > 0$ existiert mit

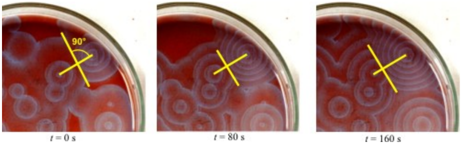
$$\lim_{t\to\infty} \boldsymbol{u}(t) = \overline{\boldsymbol{u}}, \qquad \text{sofern } \|\boldsymbol{u}(0) - \overline{\boldsymbol{u}}\| < \epsilon.$$

In vielen Fällen lässt sich die Frage nach der lokalen Stabilität von (1.3) durch Analyse der Linearisierung

$$\boldsymbol{f}(\boldsymbol{u}) \approx \boldsymbol{f}(\overline{\boldsymbol{u}}) + \boldsymbol{A}(\boldsymbol{u} - \overline{\boldsymbol{u}}), \qquad \boldsymbol{A} := \boldsymbol{f}'(\overline{\boldsymbol{u}})$$

klären, d.h. durch die Realteile der Eigenwerte von $\boldsymbol{A}$.

Die Belousov-Zhabotinsky-Reaktion ist ein Beispiel für einen sog. chemischen Oszillator, bei dem sich zeitliche Oszillationen in einem chemischen Reaktionsmechanismus zeigen.



BZ-Reaktion in einer Petri-Schale, Wellenfromt in gelb markiert.

Ein mathematisches Modell der BZ-Reaktion ist der sog. Brusselator[13], einer Evolutionsgleichung der örtlichen Variation (in einer Raumkoordinate $r \in (0,1)$) der Konzentration zweier miteinander reagierender Spezies x und y:

$$
\begin{aligned}
\partial_t x &= D_1 \partial_{rr} x + A - (B+1)x + x^2 y, & x(0,t) &= x(1,t) = A. \\
\partial_t y &= D_2 \partial_{rr} y + Bx - x^2 y, & y(0,t) &= y(1,t) = \frac{B}{A}, \\
x(r,0) &= x_0(r), & y(r,0) &= y_0(r).
\end{aligned}
$$

[13] Brussels + Oszillator, Ilya Prigogine FU Brüssel

Ein stationärer Zustand ist gegeben durch

$$\overline{x} = A, \quad \overline{y} = \frac{B}{A}.$$

Die Jacobi-Matrix an dieser Stelle ist gegeben durch

$$\boldsymbol{J} = \begin{bmatrix} D_1 \partial_{rr} + B - 1 & A^2 \\ -B & D_2 \partial_{rr} - A^2 \end{bmatrix},$$

was nach Ortsdiskretisierung eine große, dünn besetzte Matrix ergibt.

Bifurkationsproblem: Ab welchem Wert von B setzt periodisches Verhalten ein? Hierbei überqueren zwei konjugiert-komplexe Eigenwerte von $\boldsymbol{J}$ die imaginäre Achse.

(vgl. [Hairer, Norsett & Wanner, 1987, Abschnitt I.16])

Inhalt

Inhalt

Wir betrachten das AWP

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0. \tag{AWP}$$

Unter den Voraussetzungen von Satz 1.1 besitzt es eine eindeutige Lösung, sagen wir über dem Intervall I.

Wir wollen diese Lösung $\boldsymbol{y}(t)$ für $t \in [t_0, t_{\mathsf{end}}] \subseteq I$ durch das Euler-Verfahren[14], auch Euler-Cauchy-Verfahren[15] oder Polygonzug-Verfahren, den Prototyp eines numerischen Verfahrens zur Lösung von AWPen, approximieren. Wie alle numerischen Methoden, die hier besprochen werden, basiert das Euler-Verfahren auf der Idee der Diskretisierung: Statt (AWP) über $[t_0, t_{\mathsf{end}}]$ zu lösen, geben wir uns damit zufrieden, Näherungswerte für die Lösung auf einer diskreten Teilmenge

$$\{t_n \,:\, n = 0, 1, \dots, N\} \subset [t_0, t_{\mathsf{end}}]$$

zu bestimmen.

[14] Leonhard Euler (1707–1783)
[15] Augustin Louis Cauchy (1789–1857)

Wähle $N \in \mathbb{N}$, $h = (t_{\text{end}} - t_0)/N$ und definiere

$$t_n := t_0 + nh, \qquad n = 0, 1, \dots, N,$$

d.h. $t_0 < t_1 \cdots < t_{n-1} < t_N = t_{\text{end}}$. Die Zahl $h > 0$ heißt Schrittweite. (Nur der Bequemlichkeit wählen wir die Schrittweite h zunächst konstant.)

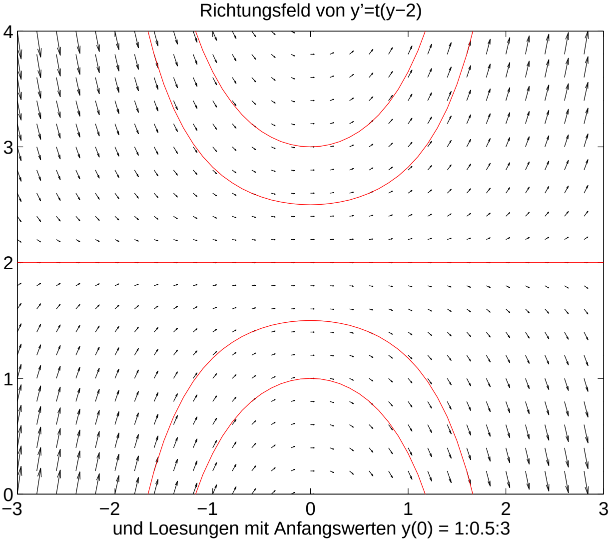
Bezeichnet nun $\boldsymbol{y}_n$ einen Näherungswert für $\boldsymbol{y}(t_n)$, $n = 0, 1, \dots, N-1$, dann ist (falls $\boldsymbol{y} \in C^2[t_0, t_{\text{end}}]$)

$$\begin{aligned}\boldsymbol{y}(t_{n+1}) &= \boldsymbol{y}(t_n + h) = \boldsymbol{y}(t_n) + h\boldsymbol{y}'(t_n) + \frac{1}{2}h^2\boldsymbol{y}''(\xi)\\ &\approx \boldsymbol{y}(t_n) + h\boldsymbol{y}'(t_n) = \boldsymbol{y}(t_n) + h\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) \approx \boldsymbol{y}_n + h\boldsymbol{f}(t_n, \boldsymbol{y}_n).\end{aligned}$$

Euler-Verfahren

$$\begin{aligned}&\boldsymbol{y}_0 \text{ gegeben,}\\ &\boldsymbol{y}_{n+1} := \boldsymbol{y}_n + h\boldsymbol{f}(t_n, \boldsymbol{y}_n), \qquad n = 0, 1, \dots, N-1.\end{aligned}$$

- Zur Veranschaulichung einer GDG 1. Ordnung $y' = f(t, y)$ wird oft das assoziierte Richtungsfeld herangezogen: In jedem Punkt (t, y) wird ein Pfeil gezeichnet, der von dort aus in Richtung der Steigung $y' = f(t, y)$ weist.
- Der Graph der Lösung des AWPs $y' = f(t, y), y(t_0) = y_0$, muss einerseits das Richtungsfeld respektieren (d.h. die Tangenten an den Graphen sind Elemente des Richtungsfelds), andererseits den Punkt (t_0, y_0) enthalten.

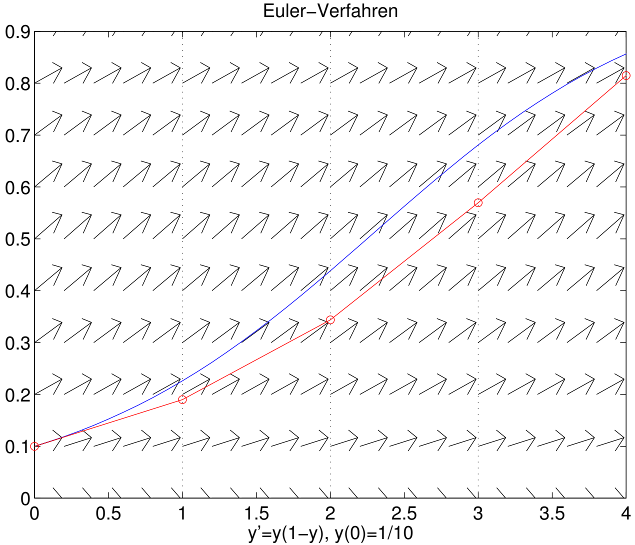


Richtungsfeld von y'=t(y−2)
4
3
2
1
0
−3
−2
−1
0
1
2
3
und Loesungen mit Anfangswerten y(0) = 1:0.5:3

Die nächste Abbildung zeigt, wie das Euler-Verfahren die logistische Gleichung

$$y' = y(1-y), \qquad y(0) = \frac{1}{10},$$

löst. Statt der exakten Trajektorie zu folgen (was natürlich unmöglich ist), produziert das Euler-Verfahren eine „stückweise lineare Lösung“ (einen Polygonzug). An der Stelle $t_0 = 0$ arbeitet das Euler-Verfahren mit der richtigen Steigung $f(t_0, y_0) = 9/100$, bereits an der Stelle $t = 1$ ist die Steigung „falsch“. In späteren Schritten entfernt man sich (potentiell) immer weiter von der exakten Lösung.



Euler−Verfahren
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3
3.5
4
y'=y(1−y), y(0)=1/10

Konvergiert das Euler-Verfahren, d.h. strebt die Näherungslösung für $h \to 0$ gegen die exakte Lösung $\boldsymbol{y}(t)$?

Formal: Zu jedem Wert von $h > 0$ gehört eine Folge von Näherungen

$$\boldsymbol{y}_n = \boldsymbol{y}_n(h), \quad n = 0, 1, \ldots, N(h) := \mathsf{floor}((t_{\mathsf{end}} - t_0)/h).$$

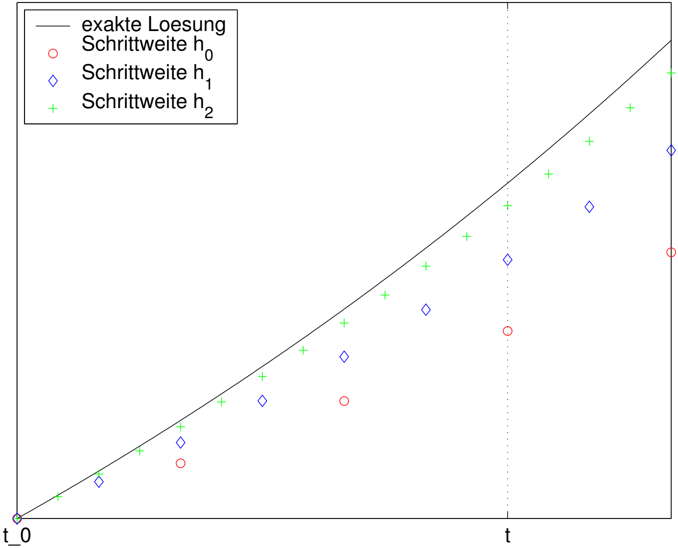
Das Verfahren heißt konvergent (in $[t_0, t_{\mathsf{end}}]$), wenn gilt

$$\lim_{h \to 0+} \max_{0 \leqslant n \leqslant N(h)} \|\boldsymbol{y}_n(h) - \boldsymbol{y}(t_n)\| = 0. \tag{Konv}$$

Satz 2.1

Unter den Voraussetzungen von Satz 1.1 konvergiert das Euler-Verfahren. Genauer:

$$\max_{0 \leqslant n \leqslant N(h)} \|\boldsymbol{y}_n(h) - \boldsymbol{y}(t_n)\| = O(h) \quad \text{für } h \to 0.$$



exakte Loesung
Schrittweite h_0
Schrittweite h_1
Schrittweite h_2
t_0
t

Die Idee dieser beiden Verfahren macht man sich leicht im Richtungsfeld klar.

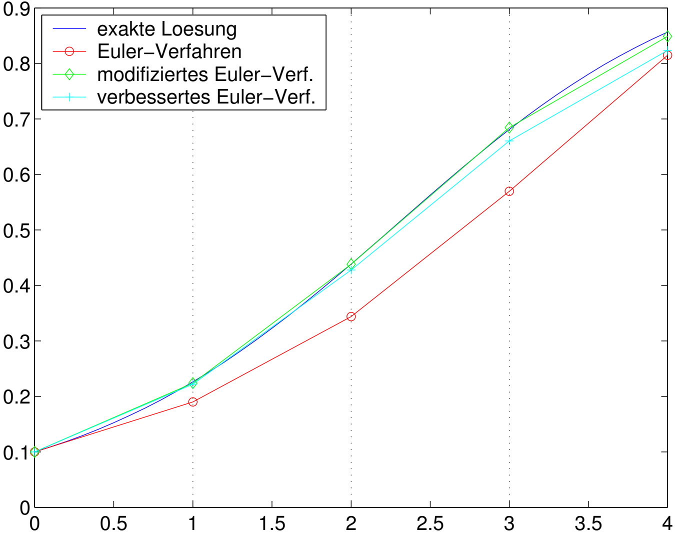
Modifiziertes Euler-Verfahren

$$\begin{aligned}
&\boldsymbol{y}_0 \text{ gegeben,}\\
&\boldsymbol{y}_{n+1} := \boldsymbol{y}_n + h\boldsymbol{f}\left(t_n + \tfrac{1}{2}h, \boldsymbol{y}_n + \tfrac{1}{2}h\boldsymbol{f}(t_n, \boldsymbol{y}_n)\right) \qquad (n = 0, 1, \dots, N-1).
\end{aligned}$$

Verbessertes Euler-Verfahren (Verfahren von Heun)

$$\begin{aligned}
&\boldsymbol{y}_0 \text{ gegeben,}\\
&\boldsymbol{y}_{n+1} := \boldsymbol{y}_n + \tfrac{1}{2}h[\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \boldsymbol{f}(t_n + h, \boldsymbol{y}_n + h\boldsymbol{f}(t_n, \boldsymbol{y}_n))]\\
&\qquad\qquad\qquad\qquad\qquad\qquad (n = 0, 1, \dots, N-1).
\end{aligned}$$

Auch das modifizierte und das verbesserte Euler-Verfahren konvergieren: Eine triviale Modifikation des Beweises von Satz 2.1 zeigt, dass für diese Verfahren sogar $\max_{0 \leqslant n \leqslant N(h)} \|\boldsymbol{y}_n(h) - \boldsymbol{y}(t_n)\| = O(h^2)$ gilt.



exakte Loesung
Euler−Verfahren
modifiziertes Euler−Verf.
verbessertes Euler−Verf.
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
0.5
1
1.5
2
2.5
3
3.5
4

Inhalt

Wir werden nun eine Reihe numerischer Verfahren zur Lösung von AWPen angeben (nicht alle davon sind für einen praktischen Einsatz geeignet), um später auf eine gewisse Anzahl von Beispielen zurückgreifen zu können.

Beispiel 1

$$\boldsymbol{y}_{n+1} - \boldsymbol{y}_n = \tfrac{1}{4} h(\boldsymbol{k}_1 + 3\boldsymbol{k}_3) \qquad \text{mit } \begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{y}_n), \\ \boldsymbol{k}_2 &= \boldsymbol{f}(t_n + \tfrac{1}{3}h, \boldsymbol{y}_n + \tfrac{1}{3}h\boldsymbol{k}_1), \\ \boldsymbol{k}_3 &= \boldsymbol{f}(t_n + \tfrac{2}{3}h, \boldsymbol{y}_n + \tfrac{2}{3}h\boldsymbol{k}_2). \end{aligned} \tag{2.1}$$

- Dieses Verfahren ist ein Einschrittverfahren (zur Berechnung von $\boldsymbol{y}_{n+1}$ ist nur die Kenntnis von $\boldsymbol{y}_n$ erforderlich).
- Es ist explizit, d.h. nach $\boldsymbol{y}_{n+1}$ aufgelöst.
- Es gehört zur Klasse der Runge-Kutta-Verfahren[16].

[16] Carl Runge (1856–1927), Martin Kutta (1867–1944)

Beispiel 2

$$\boldsymbol{y}_{n+1} - \boldsymbol{y}_n = \tfrac{1}{2}h(\boldsymbol{k}_1 + \boldsymbol{k}_2) \quad \text{mit } \begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{y}_n), \\ \boldsymbol{k}_2 &= \boldsymbol{f}(t_n + h, \boldsymbol{y}_n + \tfrac{1}{2}h\boldsymbol{k}_1 + \tfrac{1}{2}h\boldsymbol{k}_2) \end{aligned} \tag{2.2}$$

ist ebenfalls ein Einschrittverfahren der Runge-Kutta-Klasse. Im Gegensatz zu (2.1) ist es aber implizit, d.h. um $\boldsymbol{y}_{n+1}$ zu bestimmen, muss ein (im Allgemeinen nichtlineares) Gleichungssystem gelöst werden.

Beispiel 3

$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_{n+1} = \tfrac{1}{3}h\left[3\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) - 2\boldsymbol{f}(t_n, \boldsymbol{y}_n)\right] \tag{2.3}$$

ist ein explizites Zweischrittverfahren: Man benötigt $\boldsymbol{y}_n$ *und* $\boldsymbol{y}_{n+1}$, um $\boldsymbol{y}_{n+2}$ zu bestimmen. Es gehört zur Familie der linearen Mehrschrittverfahren. Neben $\boldsymbol{y}_0$ ist hier zusätzlich $\boldsymbol{y}_1$ (ein sog. Anlaufstück) erforderlich, was man sich üblicherweise mit Hilfe eines Einschrittverfahrens beschafft.

Auch

Beispiel 4

$$\boldsymbol{y}_{n+2} + \boldsymbol{y}_{n+1} - 2\boldsymbol{y}_n = \tfrac{1}{4}h\left[\boldsymbol{f}(t_{n+2}, \boldsymbol{y}_{n+2}) + 8\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) + 3f(t_n, \boldsymbol{y}_n)\right] \tag{2.4}$$

ist ein lineares Mehr- (genauer: Zwei-) Schrittverfahren. Es ist implizit.

Beispiel 5

$$\boldsymbol{y}_{n+3} + \tfrac{1}{4}\boldsymbol{y}_{n+2} - \tfrac{1}{2}\boldsymbol{y}_{n+1} - \tfrac{3}{4}\boldsymbol{y}_n = \tfrac{1}{8}h[19\boldsymbol{f}(\boldsymbol{t}_{n+2}, \boldsymbol{y}_{n+2}) + 5\boldsymbol{f}(\boldsymbol{t}_n, \boldsymbol{y}_n)] \tag{2.5}$$

stellt ein explizites lineares Dreischrittverfahren dar.

Beispiel 6

$$\begin{aligned} \boldsymbol{y}_{n+2} - \boldsymbol{y}_n &= h[\boldsymbol{f}(t_{n+2}, \boldsymbol{y}^*_{n+2}) + \boldsymbol{f}(t_n, \boldsymbol{y}_n)] \\ &\text{mit} \\ &\boldsymbol{y}^*_{n+2} - 3\boldsymbol{y}_{n+1} + 2\boldsymbol{y}_n = \tfrac{1}{2}h[\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) - 3\boldsymbol{f}(t_n, \boldsymbol{y}_n)] \end{aligned} \tag{2.6}$$

ist ein Prädiktor-Korrektor-Verfahren, bei dem ein implizites Zweischrittverfahren (der Korrektor) mit einem expliziten Zweischrittverfahren (dem Prädiktor) kombiniert wird. Das zusammengesetzte Verfahren ist explizit.

Alle Verfahrensbeispiele haben die Struktur

$$\sum_{j=0}^{k} \alpha_j \boldsymbol{y}_{n+j} = h\Phi_{\boldsymbol{f}}(\boldsymbol{y}_{n+k}, \boldsymbol{y}_{n+k-1}, \dots, \boldsymbol{y}_n, t_n; h) \tag{V}$$

(wir normieren im Folgenden durch $\alpha_k := 1$) mit

$$\Phi_{\boldsymbol{f}\equiv\boldsymbol{0}}(\boldsymbol{y}_{n+k}, \boldsymbol{y}_{n+k-1}, \dots, \boldsymbol{y}_n, t_n; h) \equiv \boldsymbol{0} \tag{V$_1$}$$

und

$$\|\Phi_{\boldsymbol{f}}(\boldsymbol{y}_{n+k}, \dots, \boldsymbol{y}_n, t_n; h) - \Phi_{\boldsymbol{f}}(\boldsymbol{y}^*_{n+k}, \dots, \boldsymbol{y}^*_n, t_n; h)\| \leqslant M \sum_{j=0}^{k} \|\boldsymbol{y}_{n+j} - \boldsymbol{y}^*_{n+j}\|. \tag{V$_2$}$$

Die Eigenschaft (V$_2$) ist eine Folge der Lipschitz-Stetigkeit von $\boldsymbol{f}$ (vgl. Satz 1.1), die immer vorausgesetzt wird.

Wir werden ausschließlich numerische Verfahren untersuchen, die die Struktur (V) besitzen und dabei den Bedingungen (V$_1$) und (V$_2$) genügen.

Inhalt

Definition 2.2 (Konvergenz)

Das Verfahren (V) heißt konvergent, wenn

$$\lim_{\substack{h\to 0\\ t=t_0+nh}} \boldsymbol{y}_n = \lim_{\substack{h\to 0\\ t=t_0+nh}} \boldsymbol{y}_n(h) = \boldsymbol{y}(t)$$

gilt, und zwar

- für alle AWPe, die den Voraussetzungen von Satz 1.1 genügen ($\boldsymbol{y}(t)$ bezeichnet die Lösung eines solchen AWPs),
- gleichmäßig für alle $t \in [t_0, t_{\text{end}}]$,
- für alle Lösungen $\{\boldsymbol{y}_n(h)\} = \{\boldsymbol{y}_n\}$ von (V) mit Anfangswerten $\boldsymbol{y}_0(h), \dots, \boldsymbol{y}_{k-1}(h)$, die $\lim_{h\to 0} \boldsymbol{y}_j(h) = \boldsymbol{y}_0$, $j = 0, \dots, k-1$, erfüllen.

Äquivalent:

$$\lim_{h\to 0} \max_{0\leqslant n\leqslant N} \|\boldsymbol{y}(t_n) - \boldsymbol{y}_n(h)\| = 0.$$

Setzt man die exakte Lösung in (V) ein, so werden linke und rechte Seite nicht übereinstimmen. Es ergibt sich ein Residuum

$$\boldsymbol{R}_{n+k} := \sum_{j=0}^{k} \alpha_j \boldsymbol{y}(t_{n+j}) - h\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_{n+k}), \boldsymbol{y}(t_{n+k-1}), \dots, \boldsymbol{y}(t_n), t_n; h).$$

$\boldsymbol{R}_{n+k}$ ist eng verknüpft mit dem lokalen Diskretisierungsfehler $\boldsymbol{T}_{n+k}$.
Unter der Lokalisierungsannahme

$$\boldsymbol{y}_{n+j} = \boldsymbol{y}(t_{n+j}) \qquad \text{für } j = 0, 1, \dots, k-1$$

liefert (V) in Schritt $n+k$:

$$\hat{\boldsymbol{y}}_{n+k} + \sum_{j=0}^{k-1} \alpha_j \boldsymbol{y}(t_{n+j}) = h\Phi_{\boldsymbol{f}}(\hat{\boldsymbol{y}}_{n+k}, \boldsymbol{y}(t_{n+k-1}), \dots, \boldsymbol{y}(t_n), t_n; h).$$

Für die exakte Lösung gilt hingegen

$$\boldsymbol{y}(t_{n+k}) + \sum_{j=0}^{k-1} \alpha_j \boldsymbol{y}(t_{n+j}) = h\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_{n+k}), \dots, \boldsymbol{y}(t_n), t_n; h) + \boldsymbol{R}_{n+k}.$$

Definiert man nun

$$\boldsymbol{T}_{n+k} := \boldsymbol{y}(t_{n+k}) - \hat{\boldsymbol{y}}_{n+k}$$

(Vorsicht: die Definition ist nicht einheitlich in der Literatur), so folgt

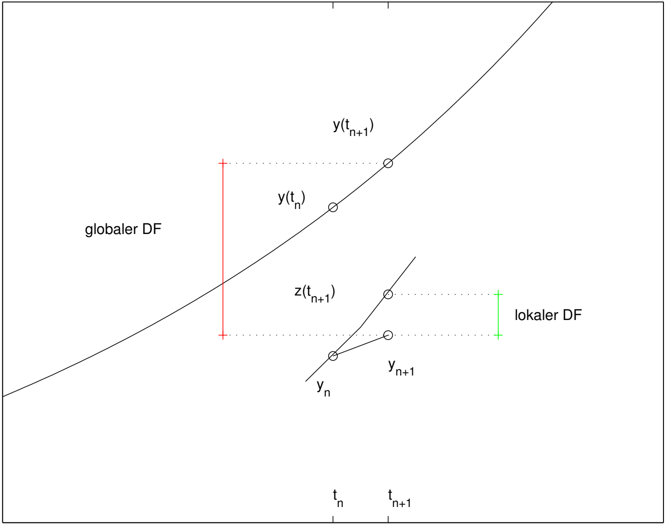
$$\begin{aligned}\boldsymbol{T}_{n+k} = h\,[&\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_{n+k}), \dots, \boldsymbol{y}(t_n), t_n; h) \\ &- \Phi_{\boldsymbol{f}}(\hat{\boldsymbol{y}}_{n+k}, \boldsymbol{y}(t_{n+k-1}), \dots, \boldsymbol{y}(t_n), t_n; h] + \boldsymbol{R}_{n+k}.\end{aligned}$$

In speziellen Fällen (lineare Mehrschrittverfahren) kann man die rechte Seite mit dem Mittelwertsatz weiter bearbeiten. Immer folgt aus (V_2)

$$\|\boldsymbol{T}_{n+k}\| \leqslant hM\|\boldsymbol{T}_{n+k}\| + \|\boldsymbol{R}_{n+k}\|$$

und damit

$$(1 - hM)\|\boldsymbol{T}_{n+k}\| \leqslant \|\boldsymbol{R}_{n+k}\|.$$


y(t_{n+1})
y(t_n)
globaler DF
z(t_{n+1})
lokaler DF
y_{n+1}
y_n
t_n
t_{n+1}

Definition 2.3 (Konsistenz)

Das Verfahren (V) heißt konsistent [mit (AWP)], wenn

$$\lim_{\substack{h\to 0\\ t=t_0+nh}} \frac{1}{h}\boldsymbol{R}_{n+k} = \boldsymbol{0}$$

für alle AWPe gilt, die den Voraussetzungen von Satz 1.1 genügen.

Satz 2.4 (Konsistenzbedingung)

Das Verfahren (V) ist genau dann konsistent, wenn gilt

$$\sum_{j=0}^{k} \alpha_j = 0 \qquad (K_1)$$

und

$$\left[\sum_{j=0}^{k} j\alpha_j\right]\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) = \Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), \dots, \boldsymbol{y}(t_n), t_n; 0). \qquad (K_2)$$

Das zum Verfahren (V) gehörende erste charakteristische Polynom ist durch

$$\rho(\zeta) = \alpha_k \zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_1 \zeta + \alpha_0$$

definiert. Hiermit läßt sich Satz 2.4 kompakter formulieren:

Satz 2.5 (Konsistenzbedingung')

Das Verfahren (V) ist genau dann konsistent, wenn gilt

$$\rho(1) = 0 \qquad (K_1')$$

$$\text{und} \qquad \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) = \frac{\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), \ldots, \boldsymbol{y}(t_n), t_n; 0)}{\rho'(1)}. \qquad (K_2')$$

Wir betrachten als Beispiel das AWP

$$y' = -y, \quad y(0) = 1 \qquad \text{mit Lösung } y(t) = \exp(-t).$$

Das implizite Zweischrittverfahren

$$y_{n+2} - (1+\alpha)y_{n+1} + \alpha y_n = \tfrac{1}{2}h[f(t_{n+2}, y_{n+2}) + (1-\alpha)f(t_{n+1}, y_{n+1}) - \alpha f(t_n, y_n)]$$

ist konsistent (für jedes $\alpha \in \mathbb{R}$):

$$\begin{aligned} \rho(1) &= 1 - (1+\alpha) + \alpha = 0, \\ \Phi_f(y(t_n), y(t_n), y(t_n), t_n; 0)/\rho'(1) &= \tfrac{1}{2}(2-2\alpha)f(t_n, y(t_n))/(2-(1+\alpha)) \\ &= f(t_n, y(t_n)). \end{aligned}$$

Lösung mit Anfangswerten $y_0 = y_1 = 1$:

$$(1 + \tfrac{1}{2}h)y_{n+2} - [1 + \alpha - \tfrac{1}{2}(1-\alpha)h]y_{n+1} + \alpha(1 - \tfrac{1}{2}h)y_n = 0 \qquad (*)$$

(Differenzengleichung zweiter Ordnung mit konstanten Koeffizienten).
Allgemeine Lösung: $y_n = c_1\xi_1^n + c_2\xi_2^n$, $\xi_1 = \alpha$ und $\xi_2 = (1 - \frac{1}{2}h)/(1 + \frac{1}{2}h)$.

Störe nun $(*)$ (Rundungsfehler, $\delta > 0$):

$$\begin{aligned}&(1+\tfrac{1}{2}h)z_{n+2} - [1+\alpha-\tfrac{1}{2}(1-\alpha)h]z_{n+1} + \alpha(1-\tfrac{1}{2}h)z_n = h\delta \\ &z_0 = 1+\delta,\ z_1 = 1+\delta.\end{aligned} \qquad (*')$$

Allgemeine Lösung:

$$z_n = c_1\xi_1^n + c_2\xi_2^n + \begin{cases}\delta/(1-\alpha), & \text{falls } \alpha \neq 1,\\ n\delta, & \text{falls } \alpha = 1.\end{cases}$$

Fall 1: $\alpha \neq 1$

$$z_n = \frac{1}{\gamma}[\mu(\delta)\xi_1^n + \nu(\delta)\xi_2^n] + \frac{\delta}{1-\alpha}$$

mit

$$\gamma = 1-\alpha-h(1+\alpha)/2, \qquad \begin{aligned}\mu(\delta) &= h[\alpha\delta/(1-\alpha)-1],\\ \nu(\delta) &= [1-\alpha(1+\delta)](1+h/2).\end{aligned}$$

Ersetze δ durch δ^*, ergibt Lösung $\{z_n^*\}$ statt $\{z_n\}$.

Fall 1a: $-1 \leqslant \alpha < 1$,

$$|z_n - z_n^*| \leqslant \left[\frac{h/(1-\alpha) + 1 + h/2}{|1 - \alpha - h(1+\alpha)/2|} + \frac{1}{1-\alpha} \right] |\delta - \delta^*|$$

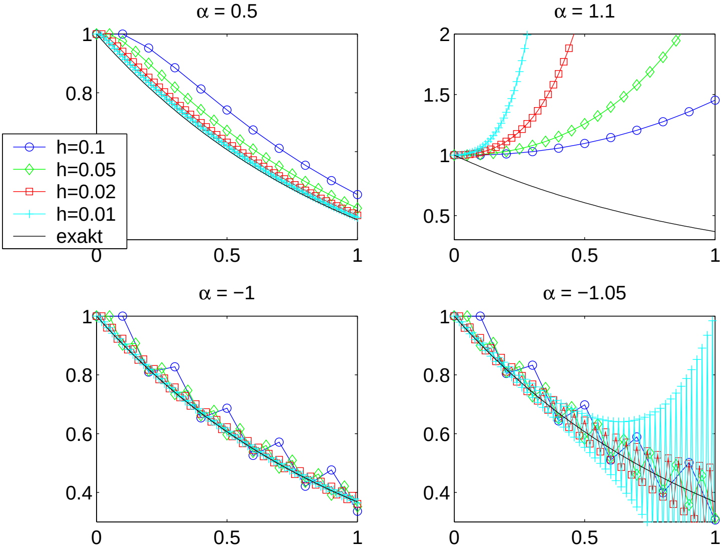
für alle $h \leqslant h_0 = 2(1-\alpha)/(1+\alpha)$.

Fall 1b: $|\alpha| > 1 \quad \{z_n - z_n^*\}$ ist unbeschränkt für $h \to 0$.

Fall 2: $\alpha = 1$

$$z_n = 1 + \frac{h-2}{2h}\delta + \frac{h+2}{2h}\delta \left(\frac{1-h/2}{1+h/2} \right)^n + n\delta$$

und $\{z_n - z_n^*\}$ ist unbeschränkt für $h \to 0$.


α = 0.5
α = 1.1
α = −1
α = −1.05
h=0.1
h=0.05
h=0.02
h=0.01
exakt
0
0.5
1
0.4
0.6
0.8
1.5
2

Definition 2.6 (Stabilität)

Seien $\{\boldsymbol{\delta}_n^{(\ell)}\}_{n=0,1,\dots,N}$, $\ell = 1,2$, zwei beliebige Störungen der Differenzengleichung (V) und seien $\{\boldsymbol{z}_n^{(\ell)}\}$, $\ell = 1,2$, die Lösungen von

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j \boldsymbol{y}_{n+j} &= h[\Phi_{\boldsymbol{f}}(\boldsymbol{y}_{n+k}, \boldsymbol{y}_{n+k-1}, \dots, \boldsymbol{y}_n, t_n; h) + \boldsymbol{\delta}_{n+k}^{(\ell)}], \\ \boldsymbol{y}_j^{(\ell)} &= \boldsymbol{y}_j(h) + \boldsymbol{\delta}_j^{(\ell)} \qquad (j = 0,1,\dots,k-1). \end{aligned} \tag{V_δ}$$

Die Methode (V) heißt stabil, wenn es positive Konstanten S und h_0 gibt mit

$$\max_{0 \leqslant n \leqslant N} \|\boldsymbol{z}_n^{(1)} - \boldsymbol{z}_n^{(2)}\| \leqslant S \max_{0 \leqslant n \leqslant N} \|\boldsymbol{\delta}_n^{(1)} - \boldsymbol{\delta}_n^{(2)}\|$$

für alle $h \in (0, h_0]$ und alle zulässigen $\boldsymbol{f}$.

Bemerkungen.

(1) Stabilität bedeutet, dass die Differenzengleichung sachgemäß gestellt ist (vgl. Satz 1.1 und 1.2, wo gezeigt wird, dass die Bedingung (Lip) garantiert, dass das (AWP) sachgemäß gestellt ist).

(2) Stabilität ist ein Muss für Rechnungen in Gleitpunktarithmetik.

Das Verfahren (V) erfüllt die Wurzelbedingung, wenn sämtliche Nullstellen ξ seines ersten charakteristischen Polynoms betragsmäßig kleiner oder gleich 1 sind, und zusätzlich aus $|\xi| = 1$ stets folgt, dass ξ eine einfache Nullstelle ist.

Satz 2.7 (Stabilität ⇔ Wurzelbedingung)

Das Verfahren (V) ist genau dann stabil, wenn es die Wurzelbedingung erfüllt.

Inhalt

Satz 2.8 (Dahlquist)

Das Verfahren (V) ist genau dann konvergent, wenn es konsistent und stabil ist. Oder kürzer:

Konsistenz & Stabilität $\Leftrightarrow$ Konvergenz.

Bemerkungen.

(1) Der Beweis zeigt, dass der globale Diskretisierungsfehler $\max_{0\leqslant n\leqslant N} \|\boldsymbol{y}(t_n) - \boldsymbol{y}_n(h)\|$ bei stabilen Verfahren von der Ordnung p ist, wenn

$$\max_{0\leqslant n\leqslant N-k} R_{n+k}/h = O(h^p) \quad \text{(Konsistenzordnung } p\text{) und}$$

$$\max_{0\leqslant n\leqslant k-1} \|\boldsymbol{y}(t_n) - \boldsymbol{y}_n(h)\| = O(h^p) \quad \text{(Anfangswerte)}$$

gelten.

(2) Ob ein Verfahren konvergent ist, kann jetzt rein algebraisch nachgeprüft werden (i.W. über die Nullstellen des ersten charakteristischen Polynoms).

Germund Dahlquist (1925–2005)

Inhalt

Ein Einschrittverfahren hat die Form

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\Phi_{\boldsymbol{f}}(\boldsymbol{y}_{n+1}, \boldsymbol{y}_n, t_n; h). \tag{ESV}$$

Es ist immer stabil, also genau dann konvergent, wenn es konsistent ist. Ist

$$\frac{1}{h}\boldsymbol{R}_{n+1} = \frac{\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n)}{h} - \Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_{n+1}), \boldsymbol{y}(t_n), t_n; h) = O(h^p) \quad (h \to 0),$$

so besitzt es (mindestens) die Konsistenzordnung p. Das Euler-Verfahren besitzt die Konsistenzordnung 1, denn:

$$\frac{\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n)}{h} - \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) = \frac{h\boldsymbol{y}'(t_n) + \frac{1}{2}h^2\boldsymbol{y}''(\xi_n)}{h} - \boldsymbol{y}'(t_n) = \tfrac{1}{2}h\boldsymbol{y}''(\xi_n).$$

Ist $\boldsymbol{f}$ p-mal stetig differenzierbar, so kann man Einschrittverfahren der Konsistenzordnung p mit der Methode des Taylor-Abgleichs konstruieren:

Man entwickelt die (unbekannte) Lösung $\boldsymbol{y}(t_n + h)$ formal nach Potenzen von h,

$$\boldsymbol{y}(t_n + h) = \boldsymbol{y}(t_n) + h\boldsymbol{y}'(t_n) + \frac{1}{2}h^2\boldsymbol{y}''(t_n) + \cdots ,$$

und nutzt aus, dass $\boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t))$ gilt, z.B.

$$\begin{aligned}
\boldsymbol{y}'(t_n) &= \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)),\\
\boldsymbol{y}''(t_n) &= \boldsymbol{f}_t(t, \boldsymbol{y}(t))|_{t=t_n} + \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{y}'(t)|_{t=t_n}\\
&= \boldsymbol{f}_t(t_n, \boldsymbol{y}(t_n)) + \boldsymbol{f}_{\boldsymbol{y}}(t_n, \boldsymbol{y}(t_n))\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)).
\end{aligned}$$

Bricht man etwa nach dem zweiten Term ab, so ergibt sich mit

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\left[\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \tfrac{1}{2}h\left(\boldsymbol{f}_t(t_n, \boldsymbol{y}_n) + \boldsymbol{f}_{\boldsymbol{y}}(t_n, \boldsymbol{y}_n)\boldsymbol{f}(t_n, \boldsymbol{y}_n)\right)\right]$$

ein Verfahren zweiter Ordnung.

Es ist klar, dass man dieses Verfahren (oder entsprechende Verfahren für $p > 2$) nur dann verwenden kann, wenn $\boldsymbol{f}$ einfach zu differenzieren ist (oder mittels automatische Differentiation).

Für die skalare GDG $y'(t) = ty(t)$ ergibt sich z.B.

$$y_{n+1} = y_n \left[1 + \tfrac{1}{2}h^2 + ht_n + \tfrac{1}{2}h^2 t_n^2\right]$$

im Fall von $p = 2$ und

$$y_{n+1} = y_n \left[1 + \tfrac{1}{2}h^2 + ht_n + \tfrac{1}{2}h^2 t_n^2 + \tfrac{1}{2}h^3 t_n + \tfrac{1}{6}h^3 t_n^3\right]$$

im Fall von $p = 3$.

Für $p = 1$ erhält man stets das Euler-Verfahren.

Für dieses Beispiel ergibt sich

mit Anfangsbedingung $y(0) = 1,$ d.h. mit exakter Lösung $y(t) = \exp(t^2/2)$

als (normalisierter) globaler Diskretisierungsfehler

$$h^{-p} \max_{0 \leqslant n \leqslant N} |y(t_n) - y_n|$$

für die Taylor-Verfahren der Ordnung $p \in \{1, 2, 3\}$:

N	h	$p = 1$	$p = 2$	$p = 3$
10	1.e-1	1.016e+0	4.301e-1	3.267e-1
100	1.e-2	1.090e+0	4.757e-1	3.541e-1
1000	1.e-3	1.098e+0	4.804e-1	3.569e-1
10000	1.e-4	1.099e+0	4.808e-1	3.548e-1

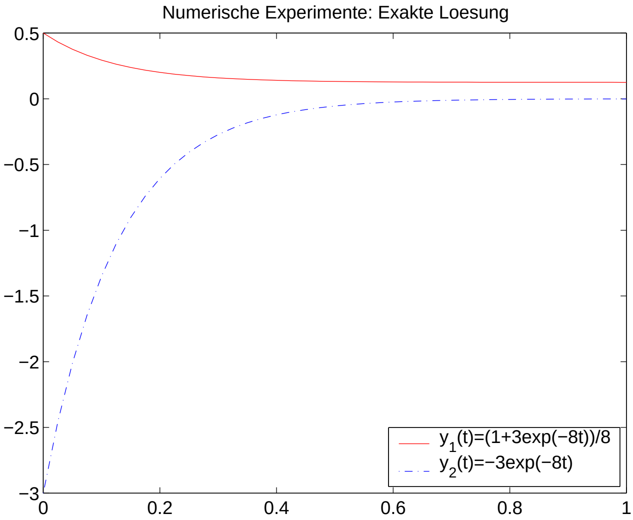
Inhalt

Wir werden jetzt die Verfahren aus Abschnitt 2 (ab Seite 104) an folgendem AWP testen:

$$y_1' = y_2 \qquad\qquad y_1(0) = \frac{1}{2},$$
$$\text{mit den Anfangsbedingungen}$$
$$y_2' = \frac{y_2(y_2 - 1)}{y_1} \qquad\qquad y_2(0) = -3.$$

Die Lösung wird für $t \in [0, 1]$ gesucht.
Die exakte Lösung lautet

$$\begin{aligned} y_1(t) &= [1 + 3\exp(-8t)]/8, \\ y_2(t) &= -3\exp(-8t). \end{aligned}$$



Numerische Experimente: Exakte Loesung
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
0
0.2
0.4
0.6
0.8
1
y1(t)=(1+3exp(−8t))/8
y2(t)=−3exp(−8t)

zu Beispiel 4:

Für $\rho(\zeta) = \zeta^2 + \zeta - 2 = (\zeta - 1)(\zeta + 2)$ gelten

$$\rho(1) = 0 \text{ und } \frac{\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), \boldsymbol{y}(t_n), \boldsymbol{y}(t_n), t_n; 0)}{\rho'(1)} = \frac{3\boldsymbol{f}(t_n, \boldsymbol{y}(t_n))}{3} = \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)).$$

Die Methode ist also konsistent, aber instabil.

t	$h = 0.1$	$h = 0.05$	$h = 0.025$	$h = 0.0125$
0.2	2.6e–02	8.2e–03	9.0e–03	1.5e–01
0.4	1.4e–01	2.1e–01	4.1e+00	1.8e+04
0.6	9.0e–01	5.9e+00	1.9e+03	2.2e+11
0.8	6.2e+00	1.7e+02	8.6e+05	Inf
1.0	4.2e+01	4.7e+03	4.0e+10	

zu Beispiel 3:

Für $\rho(\zeta) = \zeta^2 - \zeta = (\zeta - 1)\zeta$ gelten

$$\rho(1) = 0 \text{ und } \frac{\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), \boldsymbol{y}(t_n), \boldsymbol{y}(t_n), t_n; 0)}{\rho'(1)} = \frac{\boldsymbol{f}(t_n, \boldsymbol{y}(t_n))/3}{1} = \frac{\boldsymbol{f}(t_n, \boldsymbol{y}(t_n))}{3}.$$

Die Methode ist also stabil, aber inkonsistent.

t	$h = 10^{-1}$	$h = 10^{-2}$	$h = 10^{-3}$	$h = 10^{-4}$
0.2	1.3e+00	1.1e+00	1.2e+00	1.2e+00
0.4	1.1e+00	9.1e+00	9.2e+00	9.2e+00
0.6	8.0e–01	5.9e–01	5.9e–01	5.9e–01
0.8	5.5e–01	3.7e–01	3.5e–01	3.5e–01
1.0	3.8e–01	2.2e–01	2.1e–01	2.1e–01

Die Methode ist aber konsistent mit dem AWP

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})/3, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0.$$

Für den Abstand zu dessen exakter Lösung

$$\tilde{\boldsymbol{y}}(t) = [(1 + 3\exp(-8t/3))/8, \quad -3\exp(-8t/3)]^\top$$

ergibt sich:

t	$h = 10^{-1}$	$h = 10^{-2}$	$h = 10^{-3}$	$h = 10^{-4}$
0.2	3.0e–01	3.6e–02	3.8e–03	3.8e–04
0.4	4.0e–01	4.4e–02	4.4e–03	4.4e–04
0.6	3.9e–01	3.9e–02	3.9e–03	3.9e–04
0.8	3.4e–01	3.1e–02	3.1e–03	3.1e–04
1.0	2.8e–01	2.3e–02	2.3e–03	2.2e–04

.

zu Beispiel 5: Für

$$\rho(\zeta) = \zeta^3 + \zeta^2/4 - \zeta/2 - 3/4 = (\zeta - 1)(\zeta + 5/8 - \sqrt{23}i/8)(\zeta + 5/8 + \sqrt{23}i/8)$$

gelten

$$\rho(1) = 0 \text{ und } \frac{\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), \boldsymbol{y}(t_n), \boldsymbol{y}(t_n), t_n; 0)}{\rho'(1)} = \frac{3\boldsymbol{f}(t_n, \boldsymbol{y}(t_n))}{3} = \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)).$$

Die Methode ist also konsistent und stabil, damit konvergent.

t	$h = 0.1 \uparrow$	$h = 0.05 \uparrow$	$h = 0.025 \downarrow$	$h = 0.0125 \downarrow$
0.2		8.4e–03	9.3e–04	1.1e–04
0.4	2.6e–01	4.1e–02	2.4e–04	4.5e–05
0.6	1.5e+00	1.2e–01	1.6e–04	1.4e–05
0.8	8.1e+00	3.3e–01	2.1e–05	3.8e–06
1.0	4.4e+01	9.1e–01	6.8e–06	9.6e–07

zu Beispiel 6: Die Methode ist konsistent und stabil, also konvergent.

t	$h = 10^{-1}$	$h = 10^{-2}$	$h = 10^{-3}$	10^{-4}
0.2	9.0e–01	6.7e–03	2.7e–05	1.7e–07
0.4	4.0e+00	8.9e–02	1.5e–04	4.7e–06
0.6	2.3e+01	1.7e+00	3.4e–03	1.2e–04
0.8	1.4e+02	3.3e+01	8.2e–02	2.9e–03
1.0	7.9e+03	6.6e+02	2.0e+00	7.1e–02
1.2				1.7e+00
1.4				4.3e+01

zu Beispiel 1: Die Methode ist konsistent und stabil, also konvergent.

t	$h = 0.4 \uparrow$	$h = 0.2 \downarrow$	$h = 0.1 \downarrow$	$h = 0.05 \downarrow$
0.2		6.2e–01	3.9e–02	3.6e–03
0.4	7.8e+00	1.2e–01	1.5e–02	1.4e–03
0.6		2.5e–02	4.5e–03	4.4e–04
0.8	2.0e+01	5.0e–03	1.2e–03	1.2e–04
1.0		1.0e–03	2.9e–04	3.0e–05
1.2	5.0e+01			

zu Beispiel 2: Die Methode ist konsistent und stabil, also konvergent.

t	$h = 0.8 \downarrow$	$h = 0.4 \downarrow$	$h = 0.2 \downarrow$	$h = 0.1 \downarrow$
0.2			2.7e–01	5.5e–02
0.4		8.2e–01	8.6e–02	2.1e–02
0.6			2.1e–02	6.2e–03
0.8	1.6e+00	1.6e–01	4.6e–03	1.6e–03
1.0			9.6e–04	3.8e–04
1.6	8.3e–01	8.6e–03		
2.4	4.3e–01	4.6e–04		

Inhalt

Inhalt

Verfahren der Bauart

$$\sum_{j=0}^{k} \alpha_j \boldsymbol{y}_{n+j} = h \sum_{j=0}^{k} \beta_j \boldsymbol{f}_{n+j}, \qquad \text{wobei } \boldsymbol{f}_{n+j} := \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j}), \qquad \text{(LMV)}$$

heißen lineare Mehrschrittverfahren, genauer lineare k-Schritt-Verfahren.

- O.B.d.A. sei $\alpha_k = 1$ und $(\alpha_0, \beta_0) \neq (0,0)$.
- Falls $\beta_k = 0$, ist (LMV) explizit, sonst implizit.
- Bei impliziten Verfahren muss in jedem Zeitschritt ein (i.Allg. nichtlineares) Gleichungssystem der Form

  $$\boldsymbol{y}_{n+k} = h\beta_k \boldsymbol{f}(t_{n+k}, \boldsymbol{y}_{n+k}) + \sum_{j=0}^{k-1} (h\beta_j \boldsymbol{f}_{n+j} - \alpha_j \boldsymbol{y}_{n+j}) = \boldsymbol{g}(\boldsymbol{y}_{n+k}) + \boldsymbol{c}$$

  gelöst werden. Wegen

  $$\|\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{w})\| = h|\beta_k| \|\boldsymbol{f}(t_{n+k}, \boldsymbol{v}) - \boldsymbol{f}(t_{n+k}, \boldsymbol{w})\| \leqslant h\,|\beta_k|\, L\, \|\boldsymbol{v} - \boldsymbol{w}\|$$

  besitzt dies eine eindeutige Lösung, wenn $h\,|\beta_k|\,L < 1$, die mit Fixpunktiteration bestimmt werden kann.

Das Polynom

$$\sigma(\zeta) := \beta_0 + \beta_1 \zeta + \cdots + \beta_k \zeta^k \in \mathscr{P}_k$$

heißt zweites charakteristisches Polynom von (LMV) und

$$\mathscr{L}(\boldsymbol{z}(t); h) := \sum_{j=0}^{k} \left[\alpha_j \boldsymbol{z}(t + jh) - h\beta_j \boldsymbol{z}'(t + jh) \right], \quad \boldsymbol{z} \in C^1(I)$$

der mit (LMV) assoziierte Differenzenoperator.

Inhalt

Lemma 3.1

Ist z genügend oft differenzierbar, so gilt

$$\mathscr{L}(\boldsymbol{z}(t); h) = C_0 \boldsymbol{z}(t) + C_1 \boldsymbol{z}'(t) h + \cdots + C_q \boldsymbol{z}^{(q)}(t) h^q + \cdots$$

$$\text{mit} \quad C_0 = \sum_{j=0}^{k} \alpha_j = \rho(1), \quad C_1 = \sum_{j=0}^{k} [j\alpha_j - \beta_j] = \rho'(1) - \sigma(1)$$

$$\text{und} \quad C_q = \sum_{j=0}^{k} \left[\tfrac{j^q}{q!} \alpha_j - \tfrac{j^{q-1}}{(q-1)!} \beta_j \right] \quad (q = 2, 3, \ldots).$$

Für die Koeffizienten der analogen Entwicklung

$$\mathscr{L}(\boldsymbol{z}(t + \tau h); h) = \sum_{j=0}^{\infty} D_j \boldsymbol{z}^{(j)}(t + \tau h) h^j$$

gelten $C_q = \sum_{j=0}^{q} \frac{\tau^j}{j!} D_{q-j}$ $(q = 0, 1, \ldots)$.

- Der lineare Differenzenoperator $\mathscr{L}$ entspricht im Wesentlichen dem bekannten Residuum: $\boldsymbol{R}_{n+k} = \mathscr{L}(\boldsymbol{y}(t_n); h)$.
- Das lineare Mehrschrittverfahren (LMV) besitzt die genaue Konsistenzordnung p, wenn
$$C_0 = C_1 = \cdots = C_p = 0 \quad \text{und } C_{p+1} \neq 0$$
gelten. Mit den Bezeichnungen von Lemma 3.1 ist das äquivalent zu
$$D_0 = D_1 = \cdots = D_p = 0 \quad \text{und } D_{p+1} \neq 0.$$
C_{p+1} $(= D_{p+1})$ heißt dann die Fehlerkonstante des Verfahrens.
- Beachte, dass (LMV) genau dann konsistent ist (mit anderen Worten: seine Konsistenzordnung beträgt mindestens $p = 1$), wenn $\rho(1) = 0$ und $\rho'(1) = \sigma(1)$ erfüllt sind.
- (LMV) ist damit genau dann konvergent, wenn ρ die Wurzelbedingung erfüllt und $\rho(1) = 0$ sowie $\rho'(1) = \sigma(1)$ gelten (was insbesondere $\rho'(1) = \sigma(1) \neq 0$ impliziert).

Satz 3.2

Für jedes lineare k-Schritt-Verfahren sind die folgenden fünf Aussagen äquivalent:

(a) Das k-Schritt-Verfahren besitzt (mindestens) die Konsistenzordnung p.

(b) $q!\, C_q = \sum_{j=0}^{k} \left[j^q \alpha_j - q j^{q-1} \beta_j \right] = 0 \quad (q = 0, 1, \ldots, p)$.

(c) Das k-Schritt-Verfahren ist konsistent mit $y' = y$, $y(0) = 1$, von (mindestens) der Ordnung p.

(d) Die Funktion

$$\frac{\rho(\zeta)}{\log \zeta} - \sigma(\zeta)$$

hat in $\zeta = 1$ eine (mindestens) p-fache Nullstelle.

(e) $\mathscr{L}(\boldsymbol{z}(t); h) = \mathbf{0}$ für alle Polynome $\boldsymbol{z} \in \mathscr{P}_p$.

Nach der Definition der Fehlerkonstanten C_{p+1} eines LMV wissen wir lediglich dass

$$\mathscr{L}(y(t_n);h) = h^{p+1}C_{p+1}y^{(p+1)}(t_n) + O(h^{p+2}) \qquad (h \to 0).$$

Die Frage, wann auch eine Darstellung der Form $\mathscr{L}(y(t_n);h) = h^{p+1}C_{p+1}y^{(p+1)}(\xi)$ mit $t_n \leqslant \xi \leqslant t_n + h$ möglich ist, führt auf die Darstellung mittels Peano-Kern.

Lemma 3.3

Das lineare k-Schritt-Verfahren (LMV) zur Lösung von (AWP) besitze die Konsistenzordnung p. Für Funktionen $\boldsymbol{y} \in C^{(p+1)}(I)$ gilt

$$\mathscr{L}(\boldsymbol{y}(t);h) = h^{p+1}\int_0^k G(\tau)\,\boldsymbol{y}^{(p+1)}(t+\tau h)\,\mathrm{d}\tau \tag{3.1}$$

schreiben mit der Peano-Kernfunktion

$$G(\tau) = \sum_{j=0}^{k}\left[\alpha_j\frac{(j-\tau)_+^p}{p!} - \beta_j\frac{(j-\tau)_+^{p-1}}{(p-1)!}\right], \qquad u_+^k := \max\{0,u\}^k.$$

Bemerkungen:

(1) Durch früheres Abbrechen der Taylor-Reihen im Beweis von Lemma 3.3 erhält man entsprechende Darstellungen

$$\mathscr{L}(\boldsymbol{y}(t);h) = h^{q+1} \int_0^k G_q(\tau)\, \boldsymbol{y}^{(q+1)}(t+\tau h)\,\mathrm{d}\tau, \qquad 1 \leqslant q \leqslant p$$

mit

$$G_q(\tau) = \sum_{j=0}^{k} \left[\alpha_j \frac{(j-\tau)_+^q}{q!} - \beta_j \frac{(j-\tau)_+^{q-1}}{(q-1)!} \right].$$

(2) $G_q(\tau) = 0$ für $\tau \in \mathbb{R} \backslash (0,k)$.

(3) G_q ist $(q-2)$-mal stetig differenzierbar und $G_q'(\tau) = -G_{q-1}(\tau)$ (für $q = 2$ stückweise zu verstehen).

(4) $G_1(s)$ ist stückweise linear mit Sprüngen der Höhe β_j an den Stellen $\tau_j = j$, $j = 0, \dots, k$ und Steigung $-(\alpha_j + \alpha_{j+1} + \cdots + \alpha_k)$ im Intervall $(j, j+1)$.

Satz 3.4

Für den Differenzenoperator eines LMV der Konsistenzordnung p und $y \in C^{(p+1)}(I)$ gilt

$$|\mathscr{L}(y(t_n); h)| \leqslant h^{p+1} G Y$$

mit $Y := \max_{t \in I} |y^{(p+1)}(t)|$ sowie, falls der Peano-Kern $G(\tau)$ das Vorweichen in $[0, k]$ nicht wechselt,

$$G = |C_{p+1}| = \left| \int_0^k G(\tau) \, \mathrm{d}\tau \right|$$

und andernfalls

$$G = \int_0^k |G(\tau)| \, \mathrm{d}\tau.$$

LMV ohne Vorzeichenwechsel des Peano-Kerns sind beispielsweise die Familie der Adams-Bashforth bzw. Adams-Moulton Verfahren.

Satz 3.4 erlaubt Abschätzungen des lokalen Diskretisierungsfehlers:

Beispiel: Für das stabile Zweischrittverfahren

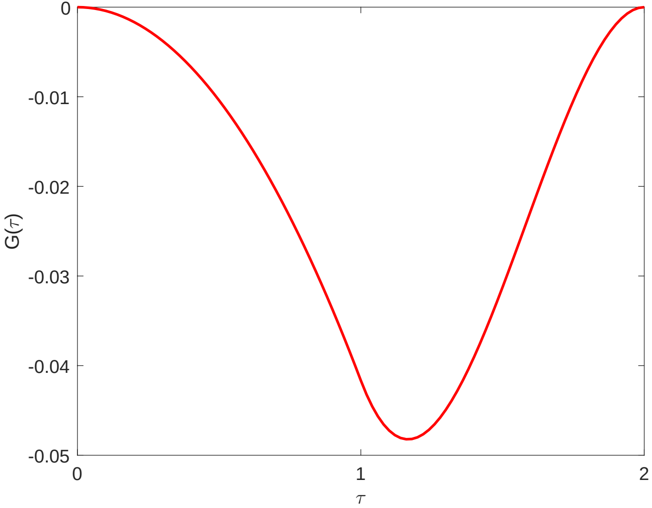
$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_{n+1} = \frac{h}{12}(5\boldsymbol{f}_{n+2} + 8\boldsymbol{f}_{n+1} - \boldsymbol{f}_n) \tag{3.2}$$

der Konsistenzordnung 3 ($C_0 = C_1 = C_2 = C_3 = 0$, $C_4 = -1/24$) erhalten wir

$$\begin{aligned} G(\tau) &= \frac{1}{24}(0-\tau)_+^2 - \frac{1}{6}(1-\tau)_+^3 - \frac{3}{4}(1-\tau)_+^2 + \frac{1}{6}(2-\tau)_+^3 - \frac{5}{24}(2-\tau)_+^2 \\ &= \begin{cases} -\frac{\tau^2}{24} & 0 \leqslant \tau \leqslant 1, \\ -\frac{1}{6}\tau^3 + \frac{19}{24}\tau^2 - \frac{7}{6}\tau + \frac{1}{2} & 1 \leqslant \tau \leqslant 2. \end{cases} \end{aligned}$$

Wie man leicht nachprüft gilt $G(\tau) \leqslant 0$ für $\tau \in [0,2]$.

Peano-Kern $G(\tau)$ des Verfahrens (3.2):



Ist $y \in C^4(I)$, so gilt nach Satz 3.4

$$|\tfrac{1}{h}R_{n+2}| = |\tfrac{1}{h}\mathscr{L}(y(t_n);h)| = h^3 \left|\int_0^2 G(\tau)\, y^{(4)}(t_n + \tau h)\,\mathrm{d}\tau\right|$$

$$\leqslant h^3|C_4| \max_{t\in I} |y^4(t)| = \frac{h^3}{24} \max_{t\in I} |y^{(4)}(t)|.$$

Alternativ folgt mit der Hölder-Ungleichung

$$|\tfrac{1}{h}R_{n+2}| \leqslant h^3 \left[\int_0^2 |G(\tau)|^\mu \mathrm{d}\tau\right]^{1/\mu} \left[\int_0^2 |y^{(4)}(t_n + \tau h)|^\nu \mathrm{d}\tau\right]^{1/\nu},$$

falls $1/\mu + 1/\nu = 1$.

Für $\mu = 1$, $\nu = \infty$ ergibt sich (entspricht Satz 3.4)

$$
\begin{aligned}
|\tfrac{1}{h} R_{n+2}| &\leqslant -h^3 \left[\int_0^2 G(\tau)\mathrm{d}\tau\right] \max_{t_n \leqslant t \leqslant t_n+2h} |y^{(4)}(t)| \\
&= \frac{h^3}{24} \max_{t_n \leqslant t \leqslant t_n+2h} |y^{(4)}(t)|
\end{aligned}
$$

sowie für $\mu = \infty$, $\nu = 1$:

$$
\begin{aligned}
|\tfrac{1}{h} R_{n+2}| &\leqslant h^3 \max_{0 \leqslant \tau \leqslant 2} |G(\tau)| \, \left[\int_0^2 |y^{(4)}(t_n + \tau h)|\mathrm{d}\tau\right] \\
&= \frac{125}{2592}\, h^3 \int_0^2 |y^{(4)}(t_n + \tau h)|\mathrm{d}\tau.
\end{aligned}
$$

Mit Hilfe von Satz 3.4 kann man auch Abschätzungen für den globalen Diskretisierungsfehler angeben. Sei

$$\boldsymbol{e}_n := \boldsymbol{y}(t_n) - \tilde{\boldsymbol{y}}_n, \qquad \tilde{\boldsymbol{y}}_n = \tilde{\boldsymbol{y}}_n(h),$$

wobei die Folge $(\tilde{\boldsymbol{y}}_n)_{n\in\mathbb{N}_0}$ die durch Rundungsfehler gestörte Differenzengleichung

$$\sum_{j=0}^{k} \alpha_j \tilde{\boldsymbol{y}}_{n+j} = h \sum_{j=0}^{k} \beta_j \boldsymbol{f}(t_{n+j}, \tilde{\boldsymbol{y}}_{n+j}) \quad + \quad \underbrace{\vartheta_n K h^{q+1}}_{\text{lokaler Rundungsfehler}} \quad , \quad \|\vartheta_n\|_\infty \leqslant 1$$

erfüllt. Wir setzen

$$A := \sum_{j=0}^{k} |\alpha_j|, \qquad B := \sum_{j=0}^{k} |\beta_j|.$$

sowie für den Fehler der Startwerte

$$E := \max_{0\leqslant j\leqslant k-1} \|\boldsymbol{e}_j\|.$$

Bezeichnet $(\gamma_j)_{j\in\mathbb{N}_0}$ die Folge der Koeffizienten der Potenzreihe

$$\frac{1}{\zeta^k\rho(\zeta^{-1})} = \frac{1}{\alpha_k + \alpha_{k-1}\zeta + \cdots + \alpha_0\zeta^k} = \sum_{j=0}^{\infty} \gamma_j \zeta^j,$$

so gilt, sofern ρ die Wurzelbedingung erfüllt ist,

$$\Gamma := \sup_{j\geqslant 0} |\gamma_j| < \infty.$$

Wir setzen ferner

$$\Gamma^* := \frac{\Gamma}{1 - h|\beta_k|L}$$

mit der Lipschitz-Konstanten L der rechten Seite $f(t, \boldsymbol{y})$.

Unter der Annahme, dass

$$h|\beta_k|L < 1 \tag{3.3}$$

lässt sich dann zeigen[17], dass

$$\|\boldsymbol{e}_n\| \leqslant \Gamma^* \left[kAE + (t_n - t_0)\left(\tfrac{1}{h}\mathscr{L}(\boldsymbol{y}(t_n); h) + h^q K\right)\right] \exp(\Gamma^* BL(t_n - t_0)).$$

Mit Hilfe der Peano-Kern-Darstellung für den lokalen Fehler $\mathscr{L}(\boldsymbol{y}(t_n); h)$ aus Satz 3.4 ergibt sich

$$\|\boldsymbol{e}_n\| \leqslant \Gamma^* \left[kAE + (t_n - t_0)(h^p GY + h^q K)\right] \exp(\Gamma^* BL(t_n - t_0)). \tag{3.4}$$

Details in [Lambert, 1991; §3.6], [Henrici, 1962; §5.3–4] .

[17] Hier wird wie schon in Kapitel 2 auf die Lösungsdarstellung für inhomogene lineare Differenzengleichungen zurückgegriffen.

(1) (3.3) ist für explizite LMV stets erfüllt, für implizite ist es eine hinreichende Bedingung für die eindeutige Lösbarkeit der Verfahrensgleichung.

(2) (3.4) zeigt den Einfluss von lokalem Fehler, Startfehler und Rundefehler; Insbesondere implizieren $E = O(h^p)$, Konsistenzordnung p sowie $q \geqslant p$ für den Gesamtfehler $\|\boldsymbol{e}_n\| = O(h^p)$.

(3) Für den Startfehler findet keine Fehlerakkumulation statt, d.h. für die Anlaufrechnung genügt ein Verfahren der Konsistenzordnung $p - 1$.

(4) Das Rundungsfehlermodell Kh^{q+1} ist unrealistisch; wir beobachten jedoch, dass man eine h-Potenz beim Übergang zum globalen Fehler verliert. Realistischer wäre eine absolute Schranke ϵ für den (lokalen) Rundungsfehler, was im globalen auf einen Anteil $\sim \epsilon/h$ führt. (Konsequenz?)

(5) Ist die Wurzelbedingung nicht erfüllt, so ist Γ nicht endlich.

(6) (3.4) fördert das Verständnis der Fehlerentwicklung, ist in der Praxis aber kaum von Nutzen. Obwohl die einzelnen Terme – zumindest a posteriori – geschätzt werden können, wird der globale Fehler durch (3.4) oft stark überschätzt.

Inhalt

- Was ist die maximale Ordnung eines konvergenten linearen k-Schritt-Verfahrens?
- $2k+2$ freie Parameter $\{\alpha_j, \beta_j\}_{j=0}^k$,
  $2k+1$ nach Normierung, $2k$ für ein explizites Verfahren.
- Konsistenz der Ordnung p führt auf $p+1$ homogene lineare Gleichungen für die Koeffizienten. Bis zu welcher Ordnung p liegt auch Stabilität vor? Erste Vermutung: $p = 2k$ [$p = 2k-1$] im impliziten [expliziten] Fall?
- 1956 beantwortet in

CONVERGENCE AND STABILITY
IN THE NUMERICAL INTEGRATION OF ORDINARY
DIFFERENTIAL EQUATIONS

GERMUND DAHLQUIST

1. Introduction and summary

1.1. Statement of the problem. Consider a class of difference equations

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \ldots + \alpha_0 y_n = h(\beta_k f_{n+k} + \ldots + \beta_0 f_n), \tag{1.1}$$

Nach Satz 3.2 ist die Konsistensordnung eines LMV von mindestens p äquivalent damit, dass die Funktion

$$\varphi(\zeta) = \frac{\rho(\zeta)}{\log \zeta} - \sigma(\zeta), \qquad (|\arg \zeta| < \pi,\ \log 1 = 0)$$

an der Stelle $\zeta = 1$ eine Nullstelle der Vielfachheit mindestens p besitzt.

Als einfache Folgerung hieraus läßt sich ein zweites charakteristisches Polynom σ bei gegebenem ersten charakteristischen Polynom ρ optimal wählen:

Satz 3.5

Sei $\rho \in \mathscr{P}_k$ mit $\rho(1) = 0$ sowie $\ell \in \{0, 1, \ldots, k\}$. Dann gibt es genau ein Polynom $\sigma \in \mathscr{P}_\ell$ sodass das zugehörige LMV die Konsistenzordnung mindestens ℓ besitzt.

Von praktischem Interesse lediglich

$\ell = k - 1$	führt auf bestmögliches explizites Verfahren,
$\ell = k$	" " implizites " .

Ausgangspunkt der Analyse der maximalen Konsistenzordnung eines stabilen LMV sind wieder dessen charakteristischen Polynome $\rho(\zeta)$ und $\sigma(\zeta)$. Folgende Variablensubstitution ist hierbei hilfreich:

$$\zeta = \frac{z+1}{z-1}, \qquad z = \frac{\zeta+1}{\zeta-1}, \tag{3.5}$$

sowie die hierdurch bestimmten Polynome

$$R(z) = \left(\frac{z-1}{2}\right)^k \rho(\zeta) = \sum_{j=0}^{k} a_j z^j, \qquad S(z) = \left(\frac{z-1}{2}\right)^k \sigma(\zeta) = \sum_{j=0}^{k} b_j z^j.$$

Lemma 3.6

Für ein stabiles LMV mit mindestens Konsistenzordnung $p = 0$ gilt

(a) $a_k = 0$ sowie $a_{k-1} = 2^{1-k}\rho'(1) \neq 0$.

(b) Alle von Null verschiedenen Koeffizienten von $R(z)$ besitzen das gleiche Vorzeichen.

Lemma 3.7

Ein LMV besitzt genau dann die Konsistenzordnung p, wenn

$$R(z)\left(\log\frac{z+1}{z-1}\right)^{-1} - S(z) = C_{p+1}\left(\frac{2}{z}\right)^{p-k} + O\left(\left(\frac{2}{z}\right)^{p-k+1}\right), \quad z \to \infty. \tag{3.6}$$

Lemma 3.8

Für die Koeffizienten der Laurent-Reihe

$$\left(\log\frac{z+1}{z-1}\right)^{-1} = \frac{z}{2} - \frac{\mu_1}{z} - \frac{\mu_3}{z^3} - \frac{\mu_5}{z^5} - \dots \tag{3.7}$$

gilt $\mu_{2j+1} > 0$ für alle $j \geqslant 0$.

Satz 3.9 (Dahlquist-Barriere)

Für die Konsistenzordnung p eines stabilen linearen k-Schritt-Verfahrens gilt

$$p \leqslant \begin{cases} k+2, & \text{falls } k \text{ gerade,} \\ k+1, & \text{falls } k \text{ ungerade,} \\ k, & \text{falls } \beta_k/\alpha_k \leqslant 0. \end{cases}$$

Optimalität bei der Ordnung LMV hat strukturelle Eigenschaften zur Folge:

Satz 3.10

Stabile LMV mit (maximaler) Konisistenzordnung $k+2$ sind symmetrisch, d.h.

$$\alpha_j = -\alpha_{k-j} \quad \text{und} \quad \beta_j = \beta_{k-j}, \qquad j = 0, \dots, k. \tag{3.8}$$

Beachte: Bei (stabilen) symmetrischen LMV gilt $\rho(\zeta) = -\zeta^k \rho(1/\zeta)$. Mit ζ ist somit auch $1/\zeta$ Nullstelle von ρ, d.h. alle Nullstellen von ρ liegen auf dem Einheitskreis und sind somit einfach.

Korollar 3.11

Ist k gerade, dann ist ein stabiles lineares k-Schritt-Verfahren mit optimaler Konsistenzordnung $k+2$ nur schwach stabil, d.h. alle Nullstellen des ersten charakteristischen Polynoms haben Betrag 1.

(Jedes) stabile k-Schritt-Verfahren ($k = 2\ell$) mit Konsistenzordnung $k+2$ kann man wie folgt konstruieren:

(1) Setze

$$\rho(\zeta) := (\zeta - 1)(\zeta + 1) \prod_{j=1}^{(k-2)/2} (\zeta - \zeta_j)(\zeta - \overline{\zeta}_j)$$

mit paarweise verschiedenen ζ_j, $|\zeta_j| = 1$, $\operatorname{Im} \zeta_j > 0$.

(2) Bestimme die ersten Koeffizienten der Taylor-Entwicklung

$$\frac{\left(\frac{z-1}{2}\right)^k \rho\left(\frac{z+1}{z-1}\right)}{\log \frac{z+1}{z-1}} = \sum_{j=0}^{\infty} b_j z^j \quad \text{und setze} \quad S(z) := \sum_{j=0}^{k} b_j z^j.$$

(3) Setze

$$\sigma(\zeta) := (\zeta - 1)^k S\left(\frac{\zeta + 1}{\zeta - 1}\right).$$

- Das einzige stabile Zweischrittverfahren der Ordnung 4 ist die Simpson-Regel[18]

$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_n = \tfrac{h}{3}(\boldsymbol{f}_{n+2} + 4\boldsymbol{f}_{n+1} + \boldsymbol{f}_n).$$

- Für $k = 4$ ist z.B.

$$\boldsymbol{y}_{n+4} - \boldsymbol{y}_n = \tfrac{h}{90}(56\boldsymbol{f}_{n+4} - 31\boldsymbol{f}_{n+3} + 96\boldsymbol{f}_{n+2} - 31\boldsymbol{f}_{n+1} + 56\boldsymbol{f}_n)$$

  ein stabiles Verfahren der Ordnung 6.

- In der Praxis spielen diese Beispiele (wie alle linearen 2ℓ-Schritt-Verfahren der Ordnung $2\ell + 2$) keine Rolle (vgl. dazu Abschnitt 6).

[18] Thomas Simpson (1710–1761)

Inhalt

Die Idee der Adams-Bashforth-Verfahren[19]:

$$\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n) = \int_{t_n}^{t_{n+1}} \boldsymbol{y}'(t)\,\mathrm{d}t = \int_{t_n}^{t_{n+1}} \boldsymbol{f}(t, \boldsymbol{y}(t))\,\mathrm{d}t.$$

Ersetze $\boldsymbol{f}(t, \boldsymbol{y}(t))$ durch ein Polynom $\boldsymbol{q}_{k-1}(t) \in \mathscr{P}_{k-1}$, das die k Datenpaare

$$(t_n, \boldsymbol{f}_n),\ (t_{n-1}, \boldsymbol{f}_{n-1}),\ \dots,\ (t_{n-k+1}, \boldsymbol{f}_{n-k+1})$$

interpoliert. In der Lagrange[20]-Darstellung ist dieses durch

$$\boldsymbol{q}_{k-1}(t) = \sum_{j=0}^{k-1} \boldsymbol{f}_{n-j} \prod_{\substack{\ell=0\\ \ell\neq j}}^{k-1} \frac{t - t_{n-\ell}}{t_{n-j} - t_{n-\ell}}$$

gegeben.

[19] John Couch Adams (1819–1892), Francis Bashforth, 1819–1912)
[20] Joseph-Louis Lagrange (1736–1813)

k-te Adams-Bashforth-Formel

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \sum_{j=0}^{k-1} \beta_{k,j} \boldsymbol{f}_{n-j} \tag{A-B}$$

$$\text{mit } \beta_{k,j} = \frac{1}{h} \int_{t_n}^{t_{n+1}} \prod_{\substack{\ell=0 \\ \ell \neq j}}^{k-1} \frac{t - t_{n-\ell}}{t_{n-j} - t_{n-\ell}} \,\mathrm{d}t = \int_0^1 \prod_{\substack{\ell=0 \\ \ell \neq j}}^{k-1} \frac{s+\ell}{\ell - j} \,\mathrm{d}s.$$

Die Adams-Moulton-Verfahren[21] konstruiert man fast genauso. Der einzige Unterschied besteht darin, dass ein Interpolationspolynom $\boldsymbol{q}_k$ vom Grad k zu den $(k+1)$ Daten

$$(t_{n+1}, \boldsymbol{f}_{n+1}),\ (t_n, \boldsymbol{f}_n),\ (t_{n-1}, \boldsymbol{f}_{n-1}),\ \dots,\ (t_{n-k+1}, \boldsymbol{f}_{n-k+1})$$

als Approximation an $\boldsymbol{f}(t, \boldsymbol{y}(t))$ verwendet wird.

[21] Forest Ray Moulton (1872–1952),

k-te Adams-Moulton-Formel:

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \sum_{j=0}^{k} \beta^*_{k,j} \boldsymbol{f}_{n+1-j} \quad \text{mit } \beta^*_{k,j} = \int_{-1}^{0} \prod_{\substack{\ell=0 \\ \ell \neq j}}^{k} \frac{s+\ell}{\ell - j} \, \mathrm{d}s. \tag{A-M}$$

Satz 3.12

Das Adams-Bashforth-Verfahren (A-B) ist ein explizites lineares k-Schritt-Verfahren. Es ist stabil und besitzt die Konsistenzordnung k. Das Adams-Moulton-Verfahren (A-M) ist ein implizites lineares k-Schritt-Verfahren. Es ist stabil und besitzt die Konsistenzordnung $k+1$.

Koeffizienten für Adams-Bashforth-Verfahren:

k						
1	1					
2	$\frac{3}{2}$	$-\frac{1}{2}$				
3	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$			
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$		
5	$\frac{1901}{720}$	$-\frac{2774}{720}$	$\frac{2616}{720}$	$-\frac{1274}{720}$	$\frac{251}{720}$	
6	$\frac{4277}{1440}$	$-\frac{7923}{1440}$	$\frac{9982}{1440}$	$-\frac{7298}{1440}$	$\frac{2877}{1440}$	$-\frac{475}{1440}$

Beispielsweise ist

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \tfrac{h}{2}(3\boldsymbol{f}_n - \boldsymbol{f}_{n-1})$$

das Adams-Bashforth-Verfahren für $k = 2$.

Koeffizienten für Adams-Moulton-Verfahren:

k							
1	$\frac{1}{2}$	$\frac{1}{2}$					
2	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$				
3	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$			
4	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$		
5	$\frac{475}{1440}$	$\frac{1427}{1440}$	$-\frac{798}{1440}$	$\frac{482}{1440}$	$-\frac{173}{1440}$	$\frac{27}{1440}$	
6	$\frac{19087}{60480}$	$\frac{65112}{60480}$	$-\frac{46461}{60480}$	$\frac{37504}{60480}$	$-\frac{20211}{60480}$	$\frac{6312}{60480}$	$-\frac{863}{60480}$

Beispielsweise ist

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{h}{2}(\boldsymbol{f}_{n+1} + \boldsymbol{f}_n) \qquad \text{(Trapezregel)}$$

das Adams-Moulton-Verfahren für $k = 1$ und

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{h}{12}(5\boldsymbol{f}_{n+1} + 8\boldsymbol{f}_n - \boldsymbol{f}_{n-1})$$

das für $k = 2$.

Natürlich kann man auch in

$$\boldsymbol{y}(t_{n+k}) - \boldsymbol{y}(t_{n+k-\ell}) = \int_{t_{n+k-\ell}}^{t_{n+k}} \boldsymbol{y}'(t)\mathrm{d}t = \int_{t_{n+k-\ell}}^{t_{n+k}} \boldsymbol{f}(t, \boldsymbol{y}(t))\,\mathrm{d}t$$

($\ell = 1, 2, \ldots$) den Integrand durch ein Interpolationspolynom ersetzen, um lineare Mehrschrittverfahren zu konstruieren. (Für $\ell = 1$ ergeben sich die Adams-Formeln.) Für $\ell = 2$ erhält man so die expliziten Nyström-Verfahren, z.B. die Mittelpunktsregel

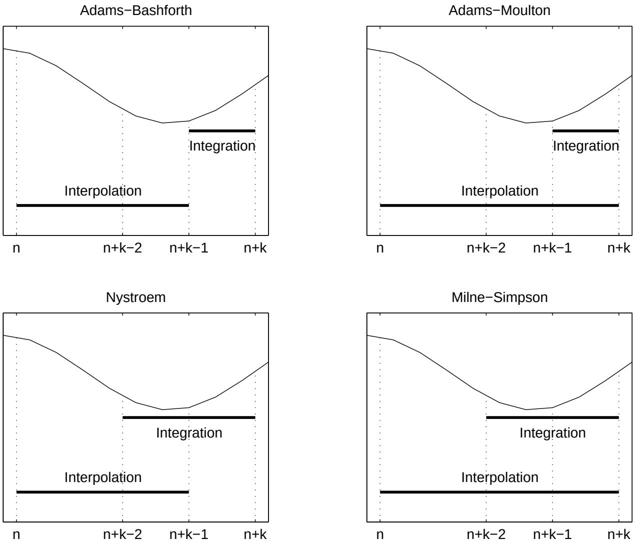
$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_n = 2h\boldsymbol{f}_{n+1},$$

bzw. die impliziten Milne-Simpson-Verfahren, wie etwa die Simpson-Regel

$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_n = \tfrac{h}{3}\left(\boldsymbol{f}_{n+2} + 4\boldsymbol{f}_{n+1} + \boldsymbol{f}_n\right).$$

Lineare Mehrschrittverfahren

Interpolation und Integration bei den Adams-artigen LMV



Inhalt

Löst man ein AWP durch ein implizites Mehrschrittverfahren, so muss in jedem Zeitschritt ein Gleichungssystem der Form

$$\boldsymbol{y}_{n+k} = h\beta_k \boldsymbol{f}(t_{n+k}, \boldsymbol{y}_{n+k}) + \boldsymbol{c}$$

(vgl. Abschnitt 1) gelöst werden. Ist $h|\beta_k|L < 1$, so konvergiert die Fixpunktiteration

$$y_{n+k}^{(\nu)} = h\beta_k \boldsymbol{f}(t_{n+k}, \boldsymbol{y}_{n+k}^{(\nu-1)}) + \boldsymbol{c} \qquad \nu = 1, 2 \ldots$$

gegen die (eindeutige) Lösung. Einen Startwert $\boldsymbol{y}_{n+k}^{(0)}$ kann man mit einem expliziten Verfahren berechnen. Man verwendet dazu ein Verfahren gleicher Ordnung.

Kombiniert man etwa die $(k+1)$-te Adams-Bashforth-Formel (Prädiktor) mit der k-ten Adams-Moulton-Formel (Korrektor), so liest sich ein Zeitschritt des resultierenden Prädiktor-Korrektor-Verfahrens wie folgt:

(P): $\quad \boldsymbol{y}_{n+1}^{(0)} = \boldsymbol{y}_n + h \sum_{j=0}^{k} \beta_{k+1,j} \boldsymbol{f}_{n-j}$,

For $\nu = 0, 1, 2, \dots$:

$$\boldsymbol{f}_{n+1}^{(\nu)} = \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}^{(\nu)}),$$

(K): $\quad \boldsymbol{y}_{n+1}^{(\nu+1)} = \boldsymbol{y}_n + h\beta_{k,0}^* \boldsymbol{f}_{n+1}^{(\nu)} + h \sum_{j=1}^{k} \beta_{k,j}^* \boldsymbol{f}_{n+1-j}$.

Man bricht die Iteration (K) ab, wenn $\|\boldsymbol{y}_{n+1}^{(\nu+1)} - \boldsymbol{y}_{n+1}^{(\nu)}\|$ „genügend klein" ist. Dann setzt man $\boldsymbol{y}_{n+1} = \boldsymbol{y}_{n+1}^{(\nu+1)}$ und (für weitere Zeitschritte) $\boldsymbol{f}_{n+1} = \boldsymbol{f}_{n+1}^{(\nu)}$ (oder alternativ: $\boldsymbol{f}_{n+1} = \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}^{(\nu+1)})$).

- In der Praxis kann man so nur selten vorgehen, da völlig unklar ist, wie viele Schritte von (K) durchgeführt werden müssen, um das Abbruchkriterium zu erfüllen. Stattdessen wird man nur eine feste (kleine) Zahl μ von diesen Iterationsschritten durchführen.
- Bezeichnen p bzw. p^* die Konsistenzordnungen von Prädiktor und Korrektor (in unserem Beispiel $p = p^* = k$), dann ist die Konsistenzordnung des zusammengesetzten Verfahrens gleich der des Korrektors, wenn $p \geqslant p^*$ oder $\mu > p^* - p$ gilt.

Inhalt

Der Begriff absolute Stabilität befasst sich anstelle der Stabilität eines Verfahrens im Grenzwert $h \to 0$, mit dessen Verhalten für lange Integrationsintervalle bei fester Schrittweite $h > 0$.

Qualitativ: wie klein muss man bei gegebener AWA und gegebenem Verfahren die Schrittweite wählen, damit die numerische Approximation der Lösung sich zumindest qualitativ richtig verhält.

Beispiel 1: Für die AWA

$$y'(t) = -\sin t, \qquad y(0) = 1 \tag{3.9}$$

mit Lösung $y(t) = \cos t$ (und $L = 0$) beträgt das Residuum R_{n+1} des expliziten Euler-Verfahrens an der Stelle t_n

$$R_{n+1} = \mathscr{L}(y(t_n); h) = -\frac{h^2}{2}\, y''(t_n) + O(h^3) = \frac{h^2}{2} \cos t_n + O(h^3),$$

sodass der globale Fehler für $t \in [t_0, t_{\mathsf{end}}]$ beschränkt ist durch (vgl. Beweis Satz 2.1)

$$\max_{t_0 \leqslant nh \leqslant t_{\mathsf{end}}} |e_n| \leqslant h \max_{t_0 \leqslant t \leqslant t_{\mathsf{end}}} |\cos t| = h.$$

Für einen Fehler $|e| \leqslant 10^{-3}$ bei der Integration bis $t_{\text{end}} = 2$ müßte also eine Schrittweite von $h = 10^{-3}$ ausreichen. Man erhält für $N = 2000$

$$y_N = -4.166014 \cdot 10^{-1}, \quad |y(2) - y_N| = 4.547667 \cdot 10^{-4}.$$

Beispiel 2: Wir modifizieren obige AWA zu

$$y'(t) = \lambda(y - \cos t) - \sin t, \qquad y(0) = 1 \tag{3.10}$$

mit derselben Lösung $y(t) = \cos t$. Man rufe sich in Erinnerung, dass beim expliziten Euler-Verfahren der globale Fehler der Rekursion

$$|e_{n+1}| \leqslant (1 + hL)|e_n| + O(h^2)$$

genügt, mit L der Lipschitz-Konstanten der rechten Seite. Für $\lambda = -10$ erhalten wir nun

$$y_N = -4.170721 \cdot 10^{-1}, \quad |y(2) - y_N| = 1.611611 \cdot 10^{-5}.$$

Beispiel 3: Wir betrachten obiges Beispiel für $\lambda = -2100$. In diesem Fall ergibt sich

$$y_N = 1.597768 \cdot 10^{76}, \quad |y(2) - y_N| = 1.452516 \cdot 10^{76}.$$

Verschiedene Schrittweiten für dieses AWP liefern

h	Fehler bei $t_{\mathsf{end}} = 2$
0.001	$1.7e + 76$
0.000976	$3.1e + 36$
0.00095	$8.6e - 04$
0.0008	$7.3e - 04$
0.0004	$3.6e - 04$

- Bekanntlich streben die Lösungen $\boldsymbol{y}(t)$ von $\boldsymbol{y}' = A\boldsymbol{y}$, $A \in \mathbb{R}^{m\times m}$ (konstant) gegen $\mathbf{0}$ für $t \to \infty$, wenn $\operatorname{Re}\lambda < 0$ für alle Eigenwerte λ von A gilt.
- Wir suchen Bedingungen an ein numerisches Verfahren (zunächst LMV), so dass die Näherungslösungen dasselbe asymptotische Verhalten besitzen.
- Dazu eine Bezeichnung: Seien ρ und σ die charakteristischen Polynome eines LMV; dann heißt

$$\pi(\zeta;\hat{h}) := \rho(\zeta) - \hat{h}\sigma(\zeta), \qquad \hat{h} = h\lambda$$

Stabilitätspolynom des Verfahrens.

Lemma 3.13

Es seien $\{y_n\}$ die Näherungen eines linearen k-Schritt-Verfahrens

$$\sum_{j=0}^{k} \alpha_j \boldsymbol{y}_{n+j} = h \sum_{j=0}^{k} \beta_j \boldsymbol{f}_{n+j} \quad (n = 0, 1, 2, \ldots)$$

zur Lösung von $\boldsymbol{y}' = A\boldsymbol{y}$ (inkl. Anfangsbedingungen). Bei festem h gilt

$$\lim_{n\to\infty} \|\boldsymbol{y}_n\| = 0$$

genau dann, wenn alle Nullstellen von $\pi(\zeta; \hat{h}) = \pi(\zeta; h\lambda)$ (als Polynom in ζ betrachtet) betragsmäßig echt kleiner als 1 sind und zwar für jedes $\lambda \in \Lambda(A)$.

- Das Verfahren heißt absolut stabil für $\hat{h}$, wenn alle Nullstellen ζ von $\pi(\cdot;\hat{h})$ die Beziehung $|\zeta| < 1$ erfüllen.
- Die Menge

$$\mathscr{R}_A := \{\hat{h} \in \mathbb{C} : \pi(\cdot;\hat{h}) \text{ hat nur Nullstellen in } |\zeta| < 1\}$$

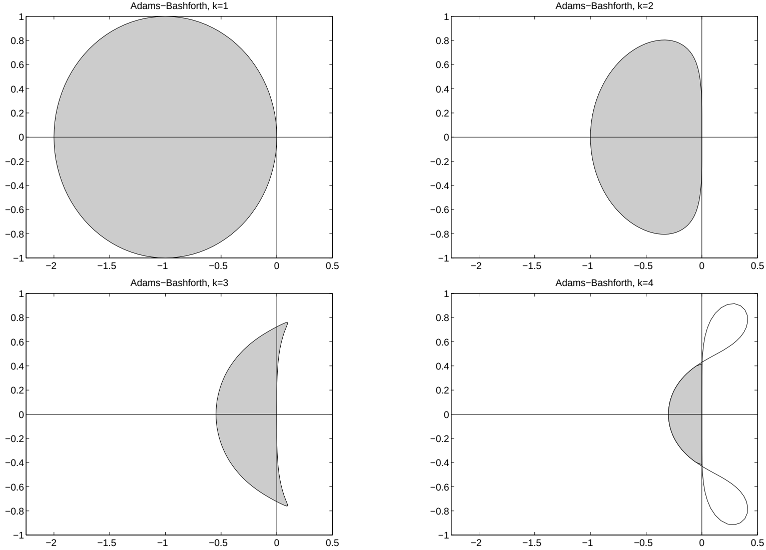
  heißt Stabilitätsgebiet des Verfahrens.
- Das Verfahren heißt A-stabil (absolut stabil), wenn $\mathscr{R}_A$ die linke Halbebene $\{\operatorname{Re}\zeta < 0\}$ enthält.

Bemerkungen.

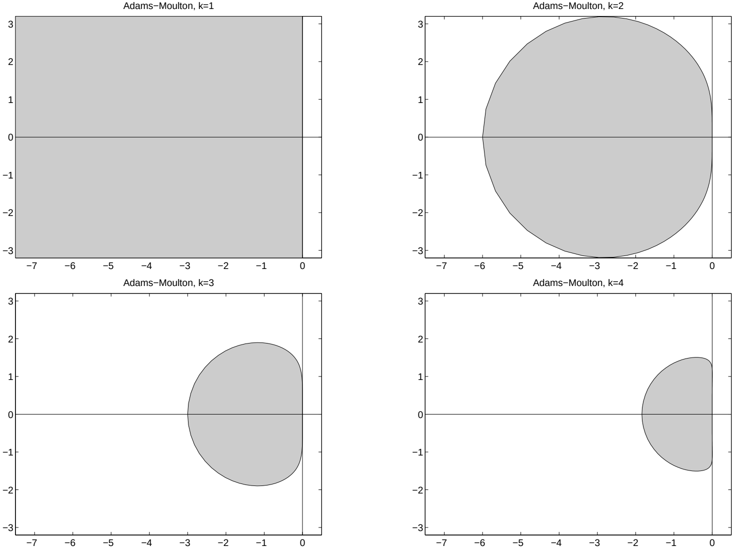
(1) Für jedes konvergente LMV gibt es ein $\hat{h}_0 > 0$, so dass $\mathscr{R}_A \cap [0,\hat{h}_0] = \varnothing$.

(2)

$$\partial\mathscr{R}_A \subseteq \left\{\hat{h} \in \mathbb{C} : \hat{h} = \frac{\rho(\mathrm{e}^{i\phi})}{\sigma(\mathrm{e}^{i\phi})},\ 0 \leqslant \phi \leqslant 2\pi\right\}.$$



Adams−Bashforth, k=1
Adams−Bashforth, k=2
Adams−Bashforth, k=3
Adams−Bashforth, k=4
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−2
−1.5
−1
−0.5
0
0.5



Adams−Moulton, k=1
Adams−Moulton, k=2
Adams−Moulton, k=3
Adams−Moulton, k=4
3
2
1
0
−1
−2
−3
−7
−6
−5
−4
−3
−2
−1
0

Ein LMV mit charakteristischen Polynomen $\rho(\zeta)$ und $\sigma(\zeta)$ heißt irreduzibel, falls $\rho(\zeta)$ und $\sigma(\zeta)$ keine gemeinsamen Nullstellen besitzen.

Lemma 3.14

Ist ein lineares Mehrschrittverfahren A-stabil, so gilt

$$\operatorname{Re} \frac{\rho(\zeta)}{\sigma(\zeta)} \geqslant 0 \quad \text{für } |\zeta| \geqslant 1. \tag{3.11}$$

Für irreduzible LMV gilt auch die Umkehrung, d.h. (3.11) impliziert A-Stabilität.

Satz 3.15 (Dahlquist, 1963)

Für die Konsistenzordnung p eines A-stabilen LMV gilt $p \leqslant 2$. Gilt $p = 2$, so gilt für die Fehlerkonstante des Verfahrens $C \leqslant -\frac{1}{12}$. Die Trapezregel ist das einzige A-stabile LMV mit Fehlerkonstante $C = -\frac{1}{12}$.

Inhalt

Idee der BDF-Verfahren (Backward Differentiation Formulas), oder auch Gear-Verfahren[22]:
Um eine Näherung für $\boldsymbol{y}(t_{n+1})$ zu gewinnen, approximieren wir $t \mapsto \boldsymbol{f}(t, \boldsymbol{y}(t))$ durch das Interpolationspolynom $P_k \in \mathscr{P}_k$ mit den $k+1$ Wertepaaren

$$(t_j, \boldsymbol{f}(t_j, \boldsymbol{y}_j)) = (t_j, \boldsymbol{f}_j), \qquad j = n-k+1, \dots, n+1.$$

Es ergibt sich

$$\boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)) \approx P_k(t) = \sum_{j=0}^{k} \ell_j \left(\frac{t_{n+1} - t}{h} \right) \boldsymbol{y}_{n+1-j} \quad \text{mit} \quad \ell_j(t) = \prod_{\substack{s=0 \\ s \neq j}}^{k} \frac{t-s}{j-s}.$$

Jetzt approximieren wir $\boldsymbol{y}'(t_{n+1}) \approx P_k'(t_{n+1}) = \sum_{j=0}^{k} \left(-\frac{1}{h}\right) \ell_j'(0) \boldsymbol{y}_{n+1-j}$ und setzen diese Näherung ein in

$$P_k'(t_{n+1}) \approx \boldsymbol{y}'(t_{n+1}) = \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \approx \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) = \boldsymbol{f}_{n+1}$$

[22] Charles William Gear, (1935–)

und erhalten die Verfahrensgleichung eines linearen k-Schritt-Verfahrens

$$\sum_{j=0}^{k} (-\ell_j'(0))\boldsymbol{y}_{n+1-j} = h\boldsymbol{f}_{n+1},$$

das BDF(k)-Verfahren. In Standardform:

$$\boldsymbol{y}_{n+1} + \sum_{j=1}^{k} \frac{\ell_j'(0)}{\ell_0'(0)}\boldsymbol{y}_{n+j-1} = -\frac{h}{\ell_0'(0))}\boldsymbol{f}_{n+1}.$$

Nach Konstruktion besitzt es die Konsistenzordnung k.

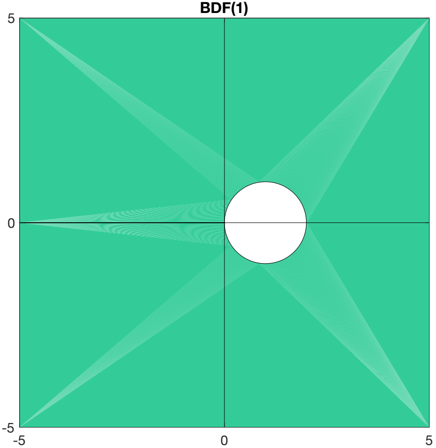
Satz 3.16 (Stabilität von BDF-Verfahren)

Das BDF(k)-Verfahren ist genau dann stabil (und damit konvergent), wenn $k \leqslant 6$.

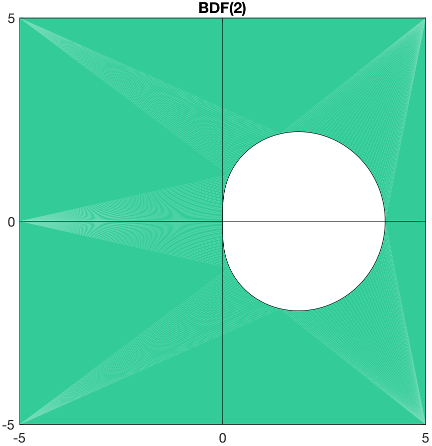
Koeffizienten und Fehlerkonstanten der BDF-Verfahren:

k	α_6	α_5	α_4	α_3	α_2	α_1	α_0	β_k	C_{k+1}
1						1	-1	1	$-\frac{1}{2}$
2					1	$-\frac{4}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$-\frac{2}{9}$
3				1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$	$\frac{6}{11}$	$-\frac{3}{22}$
4			1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	$\frac{12}{25}$	$-\frac{12}{125}$
5		1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	$\frac{60}{137}$	$-\frac{10}{137}$
6	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$	$\frac{60}{147}$	$-\frac{20}{343}$

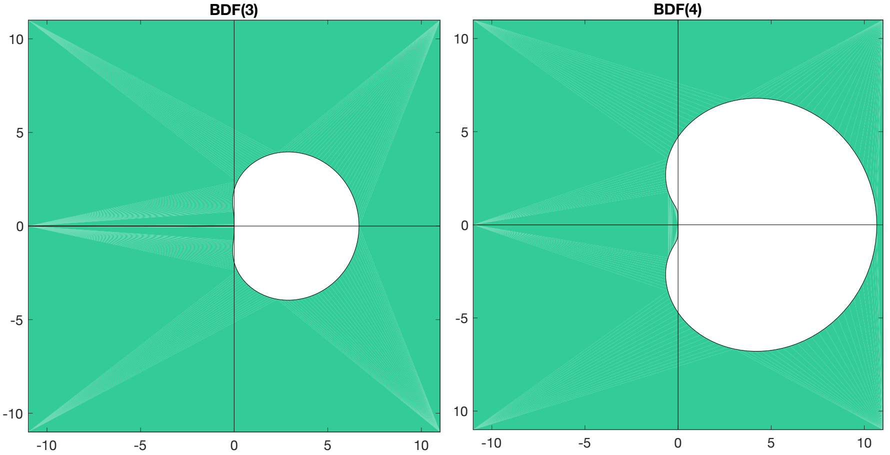
Für $k = 1$ erhält man das implizite Euler-Verfahren $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})$.
BDF-Verfahren zeichnen sich durch „große“ Stabilitätsbereiche aus.



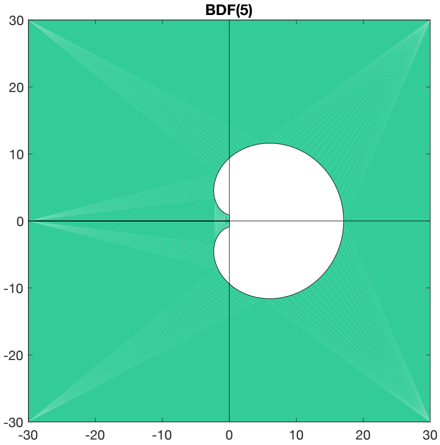
BDF(1)
5
0
-5
-5
0
5



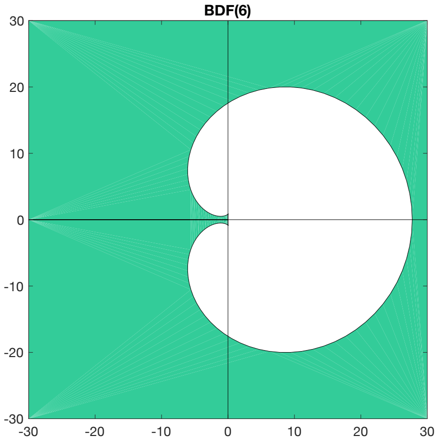
BDF(2)
5
0
-5
-5
0
5



BDF(3)
BDF(4)
10
5
0
-5
-10
-10
-5
0
5
10



BDF(5)
30
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30



BDF(6)
30
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30

Inhalt

Inhalt

Ausgangspunkt wie immer (Substitution: $s = t + \tau h$, $0 \leqslant \tau \leqslant 1$)

$$\begin{aligned} \boldsymbol{y}(t+h) = & \ \boldsymbol{y}(t) + [\boldsymbol{y}(t+h) - \boldsymbol{y}(t)] \ = \ \boldsymbol{y}(t) + \int_t^{t+h} \boldsymbol{y}'(s) \, \mathrm{d}s \\ = & \ \boldsymbol{y}(t) + h \int_0^1 \boldsymbol{y}'(t + \tau h) \, \mathrm{d}\tau. \end{aligned}$$

Approximiere durch Quadraturformel

$$\int_0^1 g(\tau) \, \mathrm{d}\tau \approx \sum_{j=1}^m \beta_j g(\gamma_j). \qquad (*)$$

Damit zumindest $g \equiv 1$ exakt integriert wird, fordern wir $\sum_{j=1}^m \beta_j = 1$.

Daraus folgt

$$\begin{aligned} \boldsymbol{y}(t+h) \approx\ & \boldsymbol{y}(t) + h \sum_{j=1}^{m} \beta_j \boldsymbol{y}'(t+\gamma_j h) \\ =\ & \boldsymbol{y}(t) + h \sum_{j=1}^{m} \beta_j \boldsymbol{f}(t+\gamma_j h, \boldsymbol{y}(t+\gamma_j h)). \end{aligned} \tag{RK-1}$$

Problem: $\boldsymbol{y}(t+\gamma_j h) = \boldsymbol{y}(t) + h \int_0^{\gamma_j} \boldsymbol{y}'(t+\tau h)\, \mathrm{d}\tau$ sind unbekannt. Näherungen wieder durch Quadraturformeln, aber mit den alten Knoten γ_j $(j = 1, \dots, m)$ aus ($*$) (sonst würden sich neue „Unbekannte“ $\boldsymbol{y}(t + \text{Knoten} \cdot h)$ ergeben).

$$\int_0^{\gamma_j} g(\tau)\, d\tau \approx \sum_{\ell=1}^{m} \alpha_{j,\ell}\, g(\gamma_\ell) \qquad (j = 1, \dots, m). \tag{$**$}$$

Damit zumindest $g \equiv 1$ exakt integriert wird, fordern wir

$$\sum_{\ell=1}^{m} \alpha_{j,\ell} = \gamma_j \qquad (j = 1, \dots, m).$$

Damit ergibt sich

$$\begin{aligned}\boldsymbol{y}(t+\gamma_j h) &\approx \boldsymbol{y}(t) + h\textstyle\sum_{\ell=1}^m \alpha_{j,\ell}\boldsymbol{y}'(t+\gamma_\ell h) \\ &= \boldsymbol{y}(t) + h\textstyle\sum_{\ell=1}^m \alpha_{j,\ell}\boldsymbol{f}(t+\gamma_\ell h, \boldsymbol{y}(t+\gamma_\ell h)).\end{aligned} \tag{RK-2}$$

Abkürzung: $\tilde{\boldsymbol{k}}_j := \boldsymbol{f}(t+\gamma_j h, \boldsymbol{y}(t+\gamma_j h))$ $(j = 1, \ldots, m)$.

(RK-2): $\tilde{\boldsymbol{k}}_j \approx \boldsymbol{f}\left(t+\gamma_j h, \boldsymbol{y}(t) + h\sum_{\ell=1}^m \alpha_{j,\ell}\tilde{\boldsymbol{k}}_\ell\right)$ $(j = 1, \ldots, m)$.

(RK-1): $\boldsymbol{y}(t+h) \approx \boldsymbol{y}(t) + h\sum_{j=1}^m \beta_j \tilde{\boldsymbol{k}}_j$.

m-stufiges Runge-Kutta-Verfahren (RKV)

$$\begin{aligned}\boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h\sum_{j=1}^m \beta_j \boldsymbol{k}_j \qquad \text{mit} \\ \boldsymbol{k}_j &= \boldsymbol{f}\left(t_n+\gamma_j h, \boldsymbol{y}_n + h\sum_{\ell=1}^m \alpha_{j,\ell}\boldsymbol{k}_\ell\right) \qquad (j = 1, \ldots, m).\end{aligned} \tag{RKV}$$

Butcher-Tableau
(John Charles Butcher, $*$1933)

$$\begin{array}{c|ccc} \gamma_1 & \alpha_{1,1} & \cdots & \alpha_{1,m} \\ \vdots & \vdots & & \vdots \\ \gamma_m & \alpha_{m,1} & \cdots & \alpha_{m,m} \\ \hline & \beta_1 & \cdots & \beta_m \end{array}$$

Beispiele.

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

symbolisiert ein zweistufiges explizites RKV, nämlich das verbesserte Euler-Verfahren.
(Ein RKV ist explizit, wenn $\alpha_{j,\ell} = 0 \ \forall j \leqslant \ell$ gilt.)

$$\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & 5/12 \\ \hline & 1/4 & 3/4 \end{array}$$

symbolisiert ein zweistufiges implizites RKV:

$$\begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}\left(t_n, \boldsymbol{y}_n + \tfrac{1}{4}h\boldsymbol{k}_1 - \tfrac{1}{4}h\boldsymbol{k}_2\right), \\ \boldsymbol{k}_2 &= \boldsymbol{f}\left(t_n + \tfrac{2}{3}h, \boldsymbol{y}_n + \tfrac{1}{4}h\boldsymbol{k}_1 + \tfrac{5}{12}h\boldsymbol{k}_2\right), \end{aligned}$$

(„zwei“ i.A. nichtlineare Gleichungen für $\boldsymbol{k}_1$ und $\boldsymbol{k}_2$)

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \tfrac{1}{4}h(\boldsymbol{k}_1 + 3\boldsymbol{k}_2).$$

(Das Verfahren (2.2) aus Abschnitt 2 ist ein weiteres implizites zweistufiges RKV.)

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

symbolisiert ein dreistufiges explizites RKV.

Verfahren dritter Ordnung von Kutta

$$
\begin{aligned}
\boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{y}_n), \\
\boldsymbol{k}_2 &= \boldsymbol{f}\left(t_n + \tfrac{1}{2}h, \boldsymbol{y}_n + \tfrac{1}{2}h\boldsymbol{k}_1\right), \\
\boldsymbol{k}_3 &= \boldsymbol{f}(t_n + h, \boldsymbol{y}_n - h\boldsymbol{k}_1 + 2h\boldsymbol{k}_2), \\
\boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \tfrac{1}{6}h(\boldsymbol{k}_1 + 4\boldsymbol{k}_2 + \boldsymbol{k}_3).
\end{aligned}
$$

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

symbolisiert ein dreistufiges explizites RKV.

Verfahren dritter Ordnung von Heun

$$\begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{y}_n), \\ \boldsymbol{k}_2 &= \boldsymbol{f}\left(t_n + \tfrac{1}{3}h, \boldsymbol{y}_n + \tfrac{1}{3}h\boldsymbol{k}_1\right), \\ \boldsymbol{k}_3 &= \boldsymbol{f}\left(t_n + \tfrac{2}{3}h, \boldsymbol{y}_n + \tfrac{2}{3}h\boldsymbol{k}_2\right), \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \tfrac{1}{4}h(\boldsymbol{k}_1 + 3\boldsymbol{k}_3). \end{aligned}$$

(Vgl. (2.1) aus Abschnitt 2.)

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

symbolisiert ein vierstufiges explizites RKV.

Klassisches Runge-Kutta-Verfahren

$$
\begin{aligned}
\boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{y}_n), \\
\boldsymbol{k}_2 &= \boldsymbol{f}(t_n + \tfrac{1}{2}h, \boldsymbol{y}_n + \tfrac{1}{2}h\boldsymbol{k}_1), \\
\boldsymbol{k}_3 &= \boldsymbol{f}(t_n + \tfrac{1}{2}h, \boldsymbol{y}_n + \tfrac{1}{2}h\boldsymbol{k}_2), \\
\boldsymbol{k}_4 &= \boldsymbol{f}(t_n + h, \boldsymbol{y}_n + h\boldsymbol{k}_3), \\
\boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \tfrac{1}{6}h(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4).
\end{aligned}
$$

Alternative Form von RKV

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \sum_{j=1}^{m} \beta_j \boldsymbol{f}(t_n + \gamma_j h, \tilde{\boldsymbol{y}}_j)$$

$$\text{mit} \quad \tilde{\boldsymbol{y}}_j = \boldsymbol{y}_n + h \sum_{\ell=1}^{m} \alpha_{j,\ell} \boldsymbol{f}(t_n + \gamma_\ell h, \tilde{\boldsymbol{y}}_\ell) \quad (j = 1, \dots, m). \tag{RKV*}$$

Setze $\boldsymbol{k}_j = \boldsymbol{f}(t_n + \gamma_j h, \tilde{\boldsymbol{y}}_j)$.

- $\boldsymbol{k}_j \approx \boldsymbol{y}'(t_n + \gamma_j h)$
- $\tilde{\boldsymbol{y}}_j \approx \boldsymbol{y}(t_n + \gamma_j h)$.

Inhalt

Jedes RKV hat die Form

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\Phi_{\boldsymbol{f}}(\boldsymbol{y}_n, t_n; h) \quad \text{mit} \quad \Phi_{\boldsymbol{f}}(\boldsymbol{y}_n, t_n; h) = \sum_{j=1}^{m} \beta_j \boldsymbol{k}_j.$$

Es ist ein Einschrittverfahren ($\rho(\zeta) = \zeta - 1$), also stabil und (vgl. Abschnitt 3) genau dann konsistent, wenn

$$\Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), t_n; 0) = \boldsymbol{f}(t_n, \boldsymbol{y}(t_n))\rho'(1)$$

erfüllt ist, was hier zu $\sum_{j=1}^{m} \beta_j = 1$ äquivalent ist.

Ein RKV ist deshalb genau dann konvergent, wenn

$$\sum_{j=1}^{m} \beta_j = 1$$

gilt.

Um die Konsistenzordnung eines RKVs zu bestimmen (oder um m-stufige RKV mit möglichst hoher Konsistenzordnung zu konstruieren), sind wie im Fall der Taylor-Verfahren (siehe Abschnitt 5) komplizierte Rechnungen erforderlich. Wir untersuchen als Beispiel explizite dreistufige RKV,

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma_2 & \gamma_2 & 0 & 0 \\ \gamma_3 & \gamma_3 - \alpha_{3,2} & \alpha_{3,2} & 0 \\ \hline & \beta_1 & \beta_2 & \beta_3 \end{array},$$

und entwickeln

$$\frac{1}{h}\boldsymbol{R}_{n+1} = \frac{\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n)}{h} - \Phi_{\boldsymbol{f}}(\boldsymbol{y}(t_n), t_n; h) = \frac{\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n)}{h} - \sum_{j=1}^{3} \beta_j \boldsymbol{k}_j$$

nach Potenzen von h (unter der Voraussetzung, dass $\boldsymbol{y}$ bzw. $\boldsymbol{f}$ genügend oft differenzierbar sind).

Für skalare AWPe ergibt sich mit den Abkürzungen

$$F := f_t + f_y f \quad \text{und} \quad G := f_{tt} + 2f_{ty} f + f_{yy} f^2$$

(alle Ableitungen von f werden an der Stelle $(t_n, y(t_n))$ ausgewertet) die Beziehung

$$\frac{y(t_{n+1}) - y(t_n)}{h} = f + \frac{1}{2}Fh + \frac{1}{6}(G + f_y F)h^2 + O(h^3).$$

Anderseits ist

$$\begin{aligned}
k_1 &= f(t_n, y(t_n)) = f, \\
k_2 &= f(t_n + h\gamma_2, y(t_n) + h\gamma_2 k_1) = f + h\gamma_2 F + \tfrac{1}{2}h^2\gamma_2^2 G + O(h^3), \\
k_3 &= f(t_n + h\gamma_3, y(t_n) + h(\gamma_3 - \alpha_{3,2})k_1 + h\alpha_{3,2}k_2) \\
&= f + h\gamma_3 F + h^2(\gamma_2\alpha_{3,2}F f_y + \tfrac{1}{2}\gamma_3^2 G) + O(h^3).
\end{aligned}$$

Das bedeutet:

$$\begin{aligned}\tfrac{1}{h}\boldsymbol{R}_{n+1} &= \Big[1 - \textstyle\sum_{j=1}^{3}\beta_j\Big]\,f + \left[\tfrac{1}{2} - \beta_2\gamma_2 - \beta_3\gamma_3\right]Fh \\ &\quad + \left[(\tfrac{1}{3} - \beta_2\gamma_2^2 - \beta_3\gamma_3^2)\tfrac{1}{2}G + (\tfrac{1}{6} - \beta_3\gamma_2\alpha_{3,2})Ff_y\right]h^2 + O(h^3).\end{aligned}$$

Folgerungen:

1. Das Euler-Verfahren ist das einzige einstufige explizite RKV der Ordnung 1 ($\beta_1 = 1$). Es gibt kein einstufiges explizites RKV höherer Ordnung.
2. Die zweistufigen expliziten RKV der Ordnung 2 sind durch

$$\beta_1 + \beta_2 = 1 \quad \text{und} \quad \beta_2\gamma_2 = \tfrac{1}{2}$$

charakterisiert. Beispiele sind das modifizierte ($\beta_1 = 0$, $\beta_2 = 1$, $\gamma_2 = \frac{1}{2}$) und das verbesserte Euler-Verfahren ($\beta_1 = \beta_2 = \frac{1}{2}$, $\gamma_2 = 1$). Kein explizites zweistufiges RKV besitzt die Ordnung 3.

3. Explizite dreistufige RKV der Ordnung 3 sind durch die vier Gleichungen

$$\beta_1 + \beta_2 + \beta_3 = 1, \qquad \beta_2\gamma_2^2 + \beta_3\gamma_3^2 = \tfrac{1}{3},$$
$$\beta_2\gamma_2 + \beta_3\gamma_3 = \tfrac{1}{2}, \qquad \beta_3\gamma_2\alpha_{3,2} = \tfrac{1}{6}$$

   charakterisiert. (Man kann zeigen, dass keine dieser Methoden die Ordnung 4 besitzt.) Beispiele sind das Verfahren von Heun ($\beta_1 = \frac{1}{4}$, $\beta_2 = 0$, $\beta_3 = \frac{3}{4}$, $\gamma_2 = \frac{1}{3}$, $\gamma_3 = \alpha_{3,2} = \frac{2}{3}$) und das Verfahren von Kutta ($\beta_1 = \frac{1}{6}$, $\beta_2 = \frac{2}{3}$, $\beta_3 = \frac{1}{6}$, $\gamma_2 = \frac{1}{2}$, $\gamma_3 = 1$, $\alpha_{3,2} = 2$).

4. Ähnliche (kompliziertere) Rechnungen zeigen, dass es eine zweiparametrige Familie expliziter vierstufiger RKV der Ordnung 4 gibt, von denen keines die Ordnung 5 besitzt. Ein Beispiel ist das klassische Runge-Kutta-Verfahren.

Weitere Beispiele sind

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 0 \\
2/3 & -1/3 & 1 & 0 & 0 \\
1 & 1 & -1 & 1 & 0 \\
\hline
 & 1/8 & 3/8 & 3/8 & 1/8
\end{array}
\qquad \text{(3/8-Regel)}
$$

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
2/5 & 2/5 & 0 & 0 & 0 \\
3/5 & -3/20 & 3/4 & 0 & 0 \\
1 & 19/44 & -15/44 & 40/44 & 0 \\
\hline
 & 55/360 & 125/360 & 125/360 & 55/360
\end{array}
\qquad \text{(Formel von Kuntzmann).}
$$

- Die oben beschriebene Methode, die Ordnung eines RKVs zu bestimmen, wird für Verfahren höherer Ordnung schnell unübersichtlich: Die Koeffizienten eines expliziten Verfahrens der Ordnung 3 müssen 4 Gleichungen erfüllen (s.o.), während bei einem Verfahren der Ordnung 8 bereits 200 nicht-lineare Gleichungen überprüft werden müssen.
- Die sog. Butcher-Theorie[23] erleichtert mit Hilfe graphentheoretischer Bäume die Buchhaltung bei den partiellen Ableitungen von f und erlaubt eine elegante Berechnung der Ordnung eines gegebenen RKVs (sie liefert aber keine Methode, ein Verfahren mit gewünschter Ordnung zu konstruieren).
- Wir beschränken uns hier darauf, *notwendige Ordnungsbedingungen* abzuleiten, die sich aus den speziellen AWPen

$$y' = y + t^{\ell-1}, \quad y(0) = 0, \qquad \ell \in \mathbb{N},$$

ergeben.

[23] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations. Runge-Kutta and General Linear Methods.* John Wiley & Sons, Chichester 1987

Satz 4.1 (Notwendige Ordnungsbedingungen für RKV)

Das durch das Butcher-Tableau

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

definierte RKV besitze die Ordnung p. Dann gelten

$$\boldsymbol{b}^\top A^k C^{\ell-1} \boldsymbol{e} = \frac{(\ell-1)!}{(\ell+k)!} = \frac{1}{\ell(\ell+1)\dots(\ell+k)}$$

für $\ell = 1, 2, \dots, p$ und $k = 0, 1, \dots, p-\ell$.

Dabei sind $\boldsymbol{b} := [\beta_1, \beta_2, \dots, \beta_m]^\top$, $A := [\alpha_{j,\nu}]_{1 \leqslant j,\nu \leqslant m}$, $C := \operatorname{diag}(\gamma_1, \gamma_2, \dots, \gamma_m)$, und $\boldsymbol{e} := [1, 1, \dots, 1]^\top \in \mathbb{R}^m$.

Spezialfälle der notwendigen Bedingungen aus Satz 4.1 sind (für $k = 0$)

$$\boldsymbol{b}^\top C^{\ell-1}\boldsymbol{e} = \sum_{j=1}^{m} \beta_j \gamma_j^{\ell-1} = \frac{1}{\ell} \qquad \text{für } \ell = 1, 2, \ldots, p$$

sowie (für $\ell = 1$ mit $k \leftarrow k+1$)

$$\boldsymbol{b}^\top A^{k-1}\boldsymbol{e} = \frac{1}{k!} \qquad \text{für } k = 1, 2, \ldots, p.$$

Bemerkung. Ein explizites m-stufiges RKV besitzt höchstens die Konsistenzordnung m, denn hier ist $A^m = O$ (A ist echte untere Dreiecksmatrix). Für die optimale Ordnung $p(m)$ eines expliziten m-stufigen RKVs gilt sogar $p(m) \leqslant m-1$ falls $m \geqslant 5$, genauer:

m	1	2	3	4	5	6	7	8	9	10	11	12
$p(m)$	1	2	3	4	4	5	6	6	7	7	8	9

.

Inhalt

Wir wenden ein m-stufiges RKV auf die Testgleichung $y' = \lambda y$ an und erhalten

$$y_{n+1} = \left[1 + \hat{h}\boldsymbol{b}^\top (I_m - \hat{h}A)^{-1}\boldsymbol{e}\right] y_n =: R(\hat{h}) y_n, \qquad (\hat{h} = \lambda h)$$

so dass (bei festem h)

$$\lim_{n\to\infty} y_n = 0 \text{ (für alle } y_0\text{)} \quad \Leftrightarrow \quad |R(\hat{h})| < 1.$$

In Analogie zu Abschnitt 6 definieren wir den Stabilitätsbereich eines RKVs durch

$$\mathscr{R}_A := \{\hat{h} \in \mathbb{C} : |R(\hat{h})| < 1\}.$$

Für ein beliebiges m-stufiges RKV gilt

$$R(\hat{h}) = 1 + \hat{h}\boldsymbol{b}^\top (I_m - \hat{h}A)^{-1}\boldsymbol{e} = 1 + \sum_{j=1}^{\infty} \hat{h}^j \boldsymbol{b}^T A^{j-1}\boldsymbol{e}.$$

- Besitzt das Verfahren die Ordnung p, so folgt

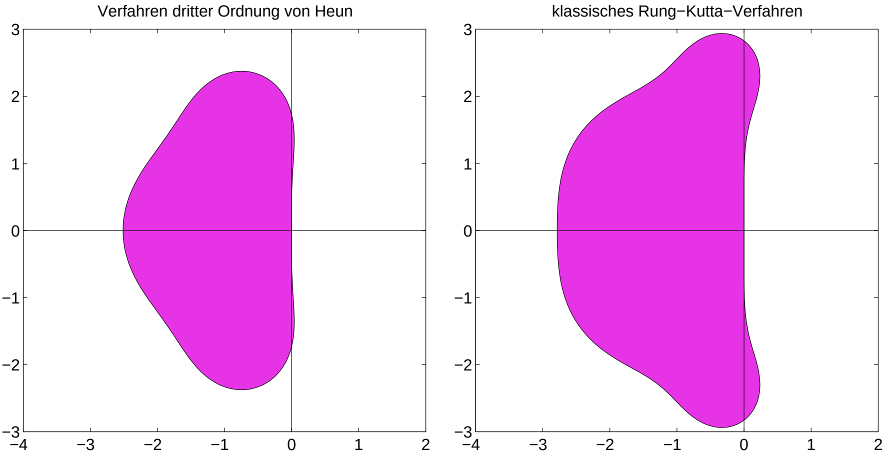
$$R(\hat{h}) = \sum_{j=0}^{p} \frac{1}{j!}\hat{h}^j + \sum_{j=p+1}^{\infty} \hat{h}^j \boldsymbol{b}^\top A^{j-1}\boldsymbol{e}.$$

- Ist das RKV explizit, so folgt

$$R(\hat{h}) = 1 + \sum_{j=1}^{m} \hat{h}^j \boldsymbol{b}^T A^{j-1}\boldsymbol{e}.$$

- Insbesondere hängt der Stabilitätsbereich eines m-stufigen expliziten RKVs der Ordnung m $(1 \leqslant m \leqslant 4)$ wegen $R(\hat{h}) = \sum_{j=0}^{m} \frac{1}{j!}\hat{h}^j$ nicht von den Koeffizienten des Verfahrens ab.
- Außerdem besitzt kein explizites RKV einen unbeschränkten Stabilitätsbereich (denn R ist in diesem Fall ein Polynom).

Beispiele für Stabilitätsgebiete zweier RKV:



Inhalt

- Kein Verfahren zur Lösung von AWPen arbeitet in der Praxis mit einer *konstanten* Schrittweite.
- Man wird vielmehr versuchen, die Schrittweite an das Verhalten der Lösung $\boldsymbol{y}$ anzupassen (ändert sich $\boldsymbol{y}$ in einem Bereich schnell, so ist dort eine kleine Schrittweite angebracht; in Bereichen, in denen $\boldsymbol{y}$ kaum variiert, ist eine größere Schrittweite ausreichend).
- Wir werden hier eine Schrittweitensteuerung vorstellen, die zum Ziel hat, den Konsistenzfehler $\boldsymbol{K}_{n+1} := \frac{1}{h}\boldsymbol{R}_{n+1}$ (wird in der Literatur oft lokaler Diskretisierungsfehler genannt, vgl. Abschnitt 3) zu kontrollieren:

$$\|\boldsymbol{K}_n\| \sim \texttt{tol}, \quad n = 1, 2, \dots,$$

  mit einer vorgebenen Toleranz `tol`.
- Bei Systemen von DGen (insbesondere dann, wenn die Lösungskomponenten von unterschiedlicher Größenordnung sind) wird man für jede Komponente eine eigene absolute Fehlertoleranz und global eine relative Fehlertoleranz festsetzen.

Das folgende Lemma besagt, dass mit dem Konsistenzfehler auch der (eigentlich interessierende) globale Diskretisierungsfehler kontrolliert wird.

Lemma 4.2

Für den globalen Diskretisierungsfehler $\boldsymbol{e}_n = \boldsymbol{y}(t_n) - \boldsymbol{y}_n$ eines Einschrittverfahrens gilt

$$\|\boldsymbol{e}_n\| \leqslant (t_n - t_0)\,\kappa_n\,\exp(M(t_n - t_0)).$$

Dabei ist $\kappa_n := \max\{\|\boldsymbol{K}_j\| \,:\, j = 0, 1, \ldots, n\}$ und M die Lipschitzkonstante der Verfahrensfunktion (vgl. (V_2) aus Abschnitt 2).

Um den Konsistenzfehler zu schätzen, verwendet man zwei Methoden unterschiedlicher Konsistenzordnungen (sagen wir p und q mit $p < q$), um $\boldsymbol{y}_n$ aus $\boldsymbol{y}_{n-1}$ zu berechnen:

$$\boldsymbol{y}_n = \boldsymbol{y}_{n-1} + h\Phi_f(\boldsymbol{y}_{n-1}, t_{n-1}; h) \qquad \text{bzw.}$$
$$\hat{\boldsymbol{y}}_n = \boldsymbol{y}_{n-1} + h\hat{\Phi}_f(\boldsymbol{y}_{n-1}, t_{n-1}; h).$$

Für die zugehörigen Konsistenzfehler gelten (Lokalisierungsannahme):

$$\boldsymbol{K}_n = \frac{\boldsymbol{y}(t_n) - \boldsymbol{y}(t_{n-1})}{h} - \Phi_f(\boldsymbol{y}(t_{n-1}), t_{n-1}; h) = O(h^p),$$
$$\hat{\boldsymbol{K}}_n = \frac{\boldsymbol{y}(t_n) - \boldsymbol{y}(t_{n-1})}{h} - \hat{\Phi}_f(\boldsymbol{y}(t_{n-1}), t_{n-1}; h) = O(h^q).$$

Daraus folgt

$$\boldsymbol{K}_n - \hat{\boldsymbol{K}}_n = \hat{\Phi}_f(\boldsymbol{y}(t_{n-1}), t_{n-1}; h) - \Phi_f(\boldsymbol{y}(t_{n-1}), t_{n-1}; h) = \tfrac{1}{h}(\hat{\boldsymbol{y}}_n - \boldsymbol{y}_n) + O(h^p).$$

Wegen $\boldsymbol{K}_n - \hat{\boldsymbol{K}}_n = \boldsymbol{K}_n(1 + O(h^{q-p})) \sim \boldsymbol{K}_n$ erhalten wir aus

$$\tfrac{1}{h}\|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n\| \sim \|\boldsymbol{K}_n\|$$

eine (grobe) Schätzung für $\|\boldsymbol{K}_n\|$.

- Ist $\frac{1}{h}\|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n\| > \texttt{tol}$, so wird die Schrittweite h verworfen und mit

$$\left(\frac{\tilde{h}}{h}\right)^p = \alpha \frac{h\,\texttt{tol}}{\|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n\|} \tag{$*$}$$

  eine neue Schrittweite $\tilde{h}$ bestimmt (α ist hier ein Sicherheitsfaktor, etwa $\alpha = 0.9$).
- Ausgehend von $\boldsymbol{y}_{n-1}$ werden jetzt neue Näherungen $\boldsymbol{y}_n$ und $\hat{\boldsymbol{y}}_n$ (an der Stelle $t_{n-1} + \tilde{h}$) berechnet. Diesen Prozess wiederholt man so lange, bis $\frac{1}{h}\|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n\| \leqslant \texttt{tol}$ erfüllt ist. Dann wird ($*$) verwendet, um eine neue (größere) Schrittweite für den nächsten Schritt ($n \to n+1$) vorzuschlagen.
- Die Wahl von $\tilde{h}$ nach ($*$) motiviert sich folgendermaßen:

$$\begin{aligned} &\text{benutzte Schrittweite } h: && \tfrac{1}{h}\|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n\| \sim \|\boldsymbol{K}_n\| = ch^p + O(h^{p+1}) \sim ch^p, \\ &\text{erwünschte Schrittweite } \tilde{h}: && \texttt{tol} = \|\boldsymbol{K}_n\| = c\tilde{h}^p + O(\tilde{h}^{p+1}) \sim c\tilde{h}^p. \end{aligned}$$

- Um den Aufwand in Grenzen zu halten, verwendet man zur Berechnung von $\boldsymbol{y}_n$ und $\hat{\boldsymbol{y}}_n$ zwei RKV (verschiedener Ordnungen), deren Butcher-Tableaus sich nur im Vektor $\boldsymbol{b}$ unterscheiden (d.h. A und $\boldsymbol{c}$ sind für beide Verfahren gleich, so dass die Größen $\boldsymbol{k}_j$ nur einmal berechnet werden müssen).
- Man spricht von eingebetteten RKV und schreibt

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^\top \\ \hline & \hat{\boldsymbol{b}}^\top \end{array}, \quad \text{z.B.} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}.$$

Im Beispiel wird ein RKV der Ordnung 1 (das Euler-Verfahren) in ein RKV der Ordnung 2 (das verbesserte Euler-Verfahren) eingebettet.

Ein populäres Beispiel ist die Fehlberg 4(5)-Formel:

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

Hier werden zwei sechsstufige RKV der Ordnungen 4 bzw. 5 kombiniert.

Ein weiteres Beispiel ist das Verfahren von Dormand-Prince 4(5):

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

Hier wird ein sechstufiges RKV der Ordnung 4 in ein siebenstufiges RKV der Ordnung 5 eingebettet.

Inhalt

- Ist die Matrix A eines m-stufigen RKVs keine echte untere $\triangle$-Matrix (ist das RKV also implizit), so muss in jedem Zeitschritt ein nicht-lineares Gleichungssystem der Form

$$\begin{array}{ccc} \boldsymbol{k}_1 & = & \boldsymbol{f}(t_n + \gamma_1 h, h\sum_{\ell=1}^{m} \alpha_{1,\ell} \boldsymbol{k}_\ell) \\ \vdots & \vdots & \vdots \\ \boldsymbol{k}_m & = & \boldsymbol{f}(t_n + \gamma_m h, h\sum_{\ell=1}^{m} \alpha_{m,\ell} \boldsymbol{k}_\ell) \end{array} \qquad (\square)$$

  gelöst werden.
- Dieses System hat also mn Unbekannte, wenn $\boldsymbol{y}' = f(t, \boldsymbol{y})$ aus n Gleichungen besteht.
- Mit Hilfe des Banachschen Fixpunktsatzes erkennt man, dass ($\square$) für genügend kleine h eindeutig lösbar ist.
- In der Praxis wird man dieses System aber nicht mit der Fixpunktiteration, sondern mit einem Newton- bzw. Quasi-Newton-Verfahren lösen.

- Ist A eine untere (aber keine echte untere) $\triangle$-Matrix, so nennt man das zugehörige RKV halb-implizit. Das System ($\square$) zerfällt dann in m Systeme mit jeweils n Unbekannten.
- Implizite RKV werden oft mit Hilfe von Gauß-Quadraturformeln konstruiert. Dies sind Formeln der Bauart

$$\int_a^b g(\tau)\,\mathrm{d}\tau = \sum_{j=1}^{m} \beta_j g(\gamma_j) + R_m(g).$$

  Hier werden die Gewichte β_j und Knoten γ_j so gewählt, dass $R_m(p) = 0$ für Polynome p möglichst hohen Grades d erfüllt ist. Man kann zeigen (vgl. Numerik I), dass eine optimale Wahl auf $d = 2m - 1$ führt (man sagt die Quadraturformel hat Exaktheitsgrad $2m - 1$).
- Die zugehörigen RKV (auch sie werden Gauß-Formeln genannt) haben die Ordnung $2m$. Beachte, dass kein m-stufiges RKV eine höhere Ordnung besitzen kann. Warum?

Für $m = 1$ ergibt sich die implizite Mittelpunktsregel $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$,

welche die Ordnung 2 besitzt.

Für $m = 2$ und $m = 3$ ergeben sich die Gauß-Formeln

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \text{bzw.} \qquad \begin{array}{c|ccc} \frac{5-\sqrt{15}}{10} & \frac{5}{36} & \frac{10-3\sqrt{15}}{45} & \frac{25-6\sqrt{15}}{180} \\ \frac{1}{2} & \frac{10+3\sqrt{15}}{72} & \frac{2}{9} & \frac{10-3\sqrt{15}}{72} \\ \frac{5+\sqrt{15}}{10} & \frac{25+6\sqrt{15}}{180} & \frac{10+3\sqrt{15}}{45} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

mit den Konsistenzordnungen 4 bzw. 6.

Bei Gauß-Radau-Integrationsformeln wählt man einen Knoten als (entweder linken oder rechten) Endpunkt des Integrationsintervalls.

- Die übrigen Knoten und alle Gewichte werden so bestimmt, dass sich ein möglichst hoher Exaktheitsgrad ergibt.
- Man kann zeigen, dass eine Gauß-Radau-Formel mit m Knoten den Exaktheitsgrad $2m-2$ besitzt. Daher haben die zugehörigen impliziten RKV die Konsistenzordnung $2m-1$.
- Zu dieser Klasse gehören die Verfahren

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \text{und} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

- Schließlich kann man noch beide Enden des Integrationsintervalls als Knoten wählen und die übrigen Daten so bestimmen, dass die zugehörige Integrationsformel den Exaktheitsgrad $2m-3$ (bzw. das zugehörige implizite RKV die Konsistenzordnung $2m-2$) besitzt. Man spricht von Gauß-Lobatto-Formeln. Ein Beispiel ist die Trapezregel (für $m=2$).

- Ein Vorteil von impliziten gegenüber expliziten RKV ist ihr wesentlich größerer Stabilitätsbereich (wird ausführlicher im nächsten Kapitel diskutiert).
- Wir betrachten die Trapezregel

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \text{d.h.} \qquad \begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{y}_n), \\ \boldsymbol{k}_2 &= \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_n + h(\boldsymbol{k}_1 + \boldsymbol{k}_2)/2), \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h(\boldsymbol{k}_1 + \boldsymbol{k}_2)/2, \end{aligned}$$

oder kürzer:

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h(\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}))/2.$$

Die zugehörige Stabilitätsfunktion ist $R(\hat{h}) = (1 + \hat{h}/2)/(1 - \hat{h}/2)$ und es gilt:

$$|R(\hat{h})| < 1 \Leftrightarrow |1 + \hat{h}/2| < |1 - \hat{h}/2| \Leftrightarrow \operatorname{Re} \hat{h} < 0.$$

Die Trapezregel ist daher A-stabil.

Inhalt

Kollokationsmethoden sind spezielle implizite RKV, die – auf Grund ihrer Konstruktion – sehr viel leichter zu analysieren sind als allgemeine RKV.

Mit gegebenen

$$0 \leqslant \gamma_1 < \gamma_2 < \cdots < \gamma_m \leqslant 1 \text{ setzen wir } t^{(j)} := t_n + \gamma_j h \quad (j = 1, 2, \ldots, m)$$

und suchen ein „Polynom“ $\boldsymbol{p}$ vom Grad $\leqslant m$ (bei Systemen von k DGen ist das ein „Vektor“ $\boldsymbol{p} = [p_1, p_2, \ldots, p_k]^\top$ aus k Polynomen vom Grad höchstens m), das die Interpolationsbedingungen

$$\boldsymbol{p}(t_n) = \boldsymbol{y}_n, \qquad \boldsymbol{p}'(t^{(j)}) = \boldsymbol{f}(t^{(j)}, \boldsymbol{p}(t^{(j)})) \quad (j = 1, 2, \ldots, m)$$

erfüllt. Die Näherung $\boldsymbol{y}_{n+1}$ an der Stelle t_{n+1} wird dann definiert durch

$$\boldsymbol{y}_{n+1} := \boldsymbol{p}(t_{n+1}).$$

$\boldsymbol{p}'$ ist ein Polynom vom Grad $m-1$, das durch die letzten m dieser Interpolationsbedingungen eindeutig bestimmt ist. In Lagrange-Form besitzt es die Darstellung

$$\boldsymbol{p}'(t_n+\tau h) = \textstyle\sum_{j=1}^m \ell_j(t_n+\tau h)\boldsymbol{k}_j$$

$$\text{mit } \ell_j(t_n+\tau h) = \textstyle\prod_{j\neq i=1}^m \dfrac{\tau-\gamma_i}{\gamma_j-\gamma_i} \quad \text{und} \quad \boldsymbol{k}_j := \boldsymbol{p}'(t^{(j)}).$$

Jetzt folgt für jedes $i \in \{1,2,\dots,m\}$

$$\begin{aligned}\boldsymbol{p}(t^{(i)}) - \boldsymbol{p}(t_n) &= \int_{t_n}^{t^{(i)}} \boldsymbol{p}'(t_n+sh)\,\mathrm{d}s = \int_{t_n}^{t^{(i)}} \textstyle\sum_{j=1}^m \ell_j(t_n+sh)\boldsymbol{k}_j\,\mathrm{d}s \\ &= h\textstyle\sum_{j=1}^m \left(\displaystyle\int_0^{\gamma_i} \ell_j(r)\,\mathrm{d}r\right)\boldsymbol{k}_j =: h\textstyle\sum_{j=1}^m \alpha_{i,j}\boldsymbol{k}_j.\end{aligned}$$

Analog:

$$\boldsymbol{p}(t_{n+1}) - \boldsymbol{p}(t_n) = h\textstyle\sum_{j=1}^m \left(\displaystyle\int_0^1 \ell_j(r)\,\mathrm{d}r\right)\boldsymbol{k}_j =: h\textstyle\sum_{j=1}^m \beta_j\boldsymbol{k}_j.$$

Daraus folgt

$$\boldsymbol{y}_{n+1} = \boldsymbol{p}(t_{n+1}) = \boldsymbol{p}(t_n) + h \textstyle\sum_{j=1}^{m} \beta_j \boldsymbol{k}_j = \boldsymbol{y}_n + h \sum_{j=1}^{m} \beta_j \boldsymbol{k}_j$$
$$\text{mit } \boldsymbol{k}_j = \boldsymbol{p}'(t^{(j)}) = \boldsymbol{f}(t^{(j)}, \boldsymbol{p}(t^{(j)})) = \boldsymbol{f}(t_n + \gamma_j h, \boldsymbol{y}_n + \textstyle\sum_{\ell=1}^{m} \alpha_{j,\ell} \boldsymbol{k}_\ell).$$

Mit anderen Worten: Jedes Kollokationsverfahren ist ein (implizites) RKV. Implementiert wird es in der Form (RKV) bzw. (RKV*), d.h. zu gegebenen γ_j ($j = 1, 2, \dots, m$) bestimmt man zunächst

$$\alpha_{i,j} = \int_0^{\gamma_i} \ell_j(r)\,\mathrm{d}r \quad \text{und} \quad \beta_j = \int_0^1 \ell_j(r)\,\mathrm{d}r \qquad (i, j = 1, 2 \dots, m).$$

Nicht jedes implizite RKV ist ein Kollokationsverfahren.

Beispiel. $\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & -5/12 \\ \hline & 1/4 & 3/4 \end{array}$ repräsentiert kein Kollokationsverfahren.

Satz 4.3 (Konsistenzordnung bei Kollokationsverfahren)

Für ein m-stufiges Kollokationsverfahren mit dem Butcher-Tableau

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

sind die folgenden drei Aussagen äquivalent:

(a) Das Verfahren besitzt die Konsistenzordnung $m + p$.

(b) $\int_0^1 \tau^j \prod_{j=1}^m (\tau - \gamma_j) \, \mathrm{d}\tau = 0$ für $j = 0, 1, \ldots, p-1$.

(c) $\boldsymbol{b}^\top C^{\ell-1} \boldsymbol{e} = 1/\ell$ für $\ell = 1, 2, \ldots, m+p$.
Dabei sind $C := \operatorname{diag}(\gamma_1, \gamma_2, \ldots, \gamma_m)$ und $\boldsymbol{e} := [1, 1, \ldots, 1]^\top \in \mathbb{R}^m$.

Inhalt

Inhalt

Es gibt keine zufriedenstellende Definition der Bauart

„eine DG heißt steif, wenn ...“.

Wir beschreiben verschiedene Aspekte des Phänomens „Steifheit einer DG“ an Beispielen.

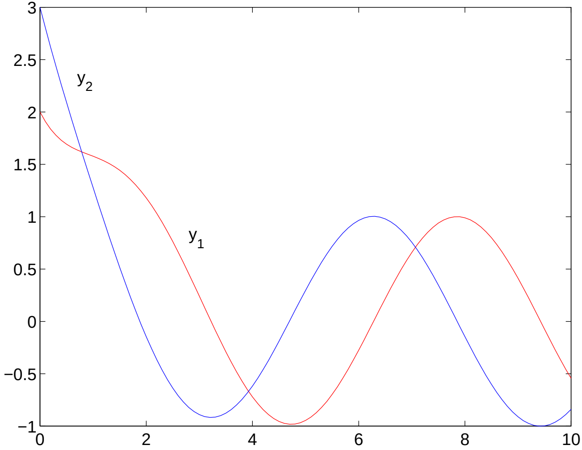
Beispiel 1. Die beiden AWPe

$$\boldsymbol{y}' = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} \boldsymbol{y} \quad + \begin{bmatrix} 2\sin t \\ 2(\cos t - \sin t) \end{bmatrix}, \qquad \boldsymbol{y}(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \tag{AWP$_1$}$$

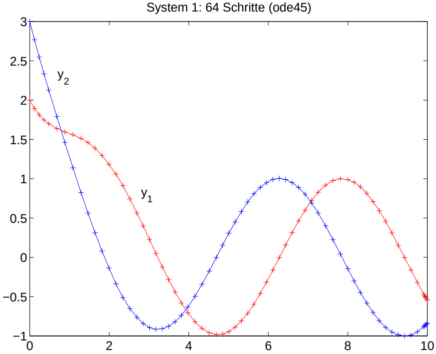
$$\boldsymbol{y}' = \begin{bmatrix} -2 & 1 \\ 998 & -999 \end{bmatrix} \boldsymbol{y} + \begin{bmatrix} 2\sin t \\ 999(\cos t - \sin t) \end{bmatrix}, \quad \boldsymbol{y}(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \tag{AWP$_2$}$$

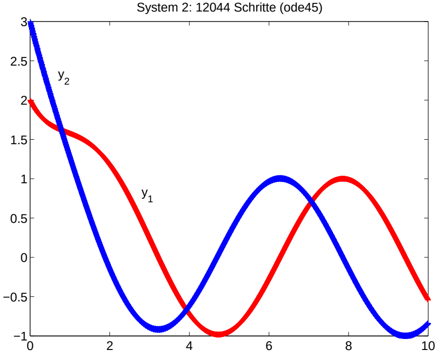
besitzen dieselbe Lösung.

$$\boldsymbol{y}(t) = \exp(-t)\begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}\sin t\\ \cos t\end{bmatrix}.$$



Wir lösen beide Probleme für $t \in [0, 10]$ mit der Matlab-Routine `ode45` (basierend auf dem Verfahren von Dormand-Prince, einem eingebettetes RKV, bei dem zwei RKV der Ordnungen 4 bzw. 5 kombiniert werden), wobei wir eine relative Toleranz von 0.01 vorgeben.





Obwohl die exakte Lösung in beiden Fällen dieselbe ist, erfordert die Lösung von (AWP_2) etwa 200-mal mehr Aufwand als die von (AWP_1):

AWP	$h_{\min}$	$h_{\max}$	$h_{\varnothing}$	Schritte
(AWP_1)	2.27e-2	2.03e-1	1.56e-1	64
(AWP_2)	5.98e-4	2.88e-3	8.30e-3	12044

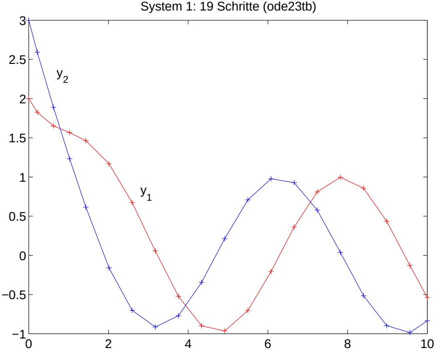
.

Lösen wir stattdessen beid Probleme mit der Matlab-Routine `ode23tb` (die die Trapezregel mit der BDF-Formel der Ordnung 2 kombiniert), so ergibt sich (relative Toleranz wie oben = 0.01):

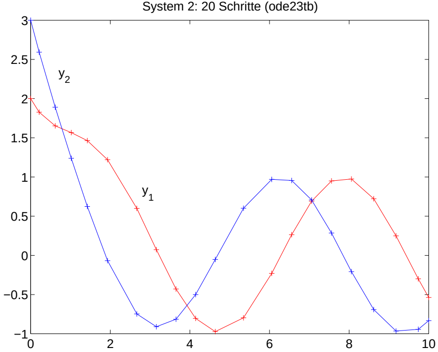
AWP	$h_{\min}$	$h_{\max}$	$h_{\varnothing}$	Schritte
(AWP_1)	2.13e-1	5.81e-1	5.26e-1	19
(AWP_2)	2.13e-1	7.36e-1	5.00e-1	20

.

Beide Probleme werden ohne Schwierigkeiten gelöst.



System 1: 19 Schritte (ode23tb)
y2
y1
3
2.5
2
1.5
1
0.5
0
−0.5
−1
0
2
4
6
8
10



System 2: 20 Schritte (ode23tb)
y2
y1
3
2.5
2
1.5
1
0.5
0
−0.5
−1
0
2
4
6
8
10

Die Unterschiede zwischen den beiden AWPen werden sichtbar, wenn wir alle Lösungen der zugehörigen Systeme betrachten. Im ersten Fall ergibt sich

$$\boldsymbol{y} = \kappa_1 \exp(-t) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \kappa_2 \exp(-3t) \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} \sin t \\ \cos t \end{bmatrix},$$

während im zweiten Fall

$$\boldsymbol{y} = \kappa_1 \exp(-t) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \kappa_2 \exp(-1000t) \begin{bmatrix} 1 \\ -998 \end{bmatrix} + \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}$$

die allgemeine Lösung ist. (Beide Systeme sind inhomogen und linear mit konstanten Koeffizienten, wobei die Eigenwerte der Koeffizientenmatrix

- im ersten Fall $\lambda_1 = -1$, $\lambda_2 = -3$
- und im zweiten Fall $\lambda_1 = -1$, $\lambda_2 = -1000$

sind.) Die Lösungen der konkreten AWPe sind jeweils durch $\kappa_1 = 2$ und $\kappa_2 = 0$ (!) gegeben.

- In beiden Fällen ist die Lösung aus einem Gleichgewichtsanteil, nämlich $[\sin t, \, \cos t]^\top$, und einem transienten Anteil der Form $\exp(\lambda_1 t)\boldsymbol{a}_1 + \exp(\lambda_2 t)\boldsymbol{a}_2$ zusammengesetzt. Die Unterschiede liegen in der Geschwindigkeit, mit der der transiente Teil (für $t \to \infty$) verschwindet.
- Obwohl in der exakten Lösung der AWPe gar nicht erkennbar ist, wie schnell der abklingende Anteil verschwindet ($\kappa_2 = 0$), bestimmt er die erforderliche Schrittweite.
- Würde man in (AWP$_2$) die AB in $\boldsymbol{y}(0) = [0, \, 1]^\top$ ändern, so würde $\kappa_1 = \kappa_2 = 0$ folgen, d.h. der transiente Teil ist in der Lösung $\boldsymbol{y}(t) = [\sin t, \, \cos t]^\top$ vollständig unsichtbar. Selbst dann wäre `ode45` nur mit extrem kleiner Schrittweite in der Lage, die Aufgabe zu lösen.

- Wir betrachten die Stabilitätsbereiche $\mathscr{R}_A$: Für `ode45` kann man zeigen, dass $\mathscr{R}_A \cap \mathbb{R} \sim [-3, 0]$ (Stabilitätsintervall). Will man für (AWP$_1$) also $\lambda h \in \mathscr{R}_A$ (für alle Eigenwerte λ) garantieren, genügt es, $h < 1$ zu fordern (durchschn. Schrittweite $h_\varnothing \sim 0.156$ resultiert aus der geforderten Genauigkeit!).
- Dagegen muss für (AWP$_2$) $h < 0.003$ gefordert werden, damit $\lambda h \in \mathscr{R}_A$ für alle λ folgt und es sind Stabilitätsforderungen, die zu der kleinen durchschnittlichen Schrittweite von $h_\varnothing \sim 0.008$ führen.
- Das Verfahren `ode23tb` ist absolut stabil, so dass hier λh für alle $h > 0$ (in jedem der beiden Fälle (AWP$_1$) und (AWP$_2$)) im Stabilitätsbereich liegt.

Für $\boldsymbol{y}' = A\boldsymbol{y} + \boldsymbol{b}(t)$ heißt

$$\frac{\max_{\lambda \in \Lambda(A)} |\operatorname{Re} \lambda|}{\min_{\lambda \in \Lambda(A)} |\operatorname{Re} \lambda|}$$

Steifigkeitsquotient des linearen DG-Systems.

Ein lineares DG-System mit konstanten Koeffizienten heißt steif, wenn seine Eigenwerte alle negativen Realteil besitzen und sein Steifigkeitsquotient groß ist.

Um die Problematik dieser gebräuchlichen „Definition“ zu erläutern, betrachten wir ein weiteres AWP (mit derselben Lösung):

$$\boldsymbol{y}' = \begin{bmatrix} -2 & 1 \\ -1.999 & 0.999 \end{bmatrix} \boldsymbol{y} + \begin{bmatrix} 2\sin t \\ -0.999(\cos t - \sin t) \end{bmatrix}, \quad \boldsymbol{y}(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}. \qquad (\mathsf{AWP}_3)$$

Die Matrix dieses Systems besitzt die Eigenwerte

$$\lambda_1 = -1, \quad \lambda_2 = -0.001$$

und sein Steifigkeitsquotient beträgt 1000 (wie bei (AWP_2)).

Trotzdem hat `ode45` keine ernsten Probleme:

AWP	$h_{\min}$	$h_{\max}$	$h_{\varnothing}$	Schritte
(AWP_3)	1.19e-1	2.50e-1	2.27e-1	44 (`ode45`)
(AWP_3)	1.25e-1	4.92e-1	4.55e-1	22 (`ode23tb`)

Ob ein System steif ist oder nicht kann also nicht immer aus dem Steifigkeitsquotienten abgelesen werden.

Auch „Definitionen“ wie

> „*ein System ist steif, wenn* $\max|\operatorname{Re}\lambda|$ *groß ist (etwa* $\max|\operatorname{Re}\lambda| \gg 1$*)*“

sind natürlich wenig hilfreich (die Variablentransformation $t \mapsto 0.001t$ macht aus (AWP_2) ein Problem, das dieselbe Steifigkeit besitzt, bei dem aber $\max|\operatorname{Re}\lambda| = 1$ gilt).

Pragmatische „Definitionen“ für Steifigkeit sind:

> *Ein System ist steif, wenn Stabilitätsanforderungen — und nicht Genauigkeitsanforderungen — die Größe der Schrittweite bestimmen.*

> *Ein System heißt steif, wenn gewisse Komponenten der Lösung sehr viel schneller abklingen als andere.*

> *Ein System heißt steif, wenn explizite Verfahren nicht funktionieren.*

Beispiel 2. Die Differentialgleichung

$$y'(t) = \lambda(y(t) - g(t)) + g'(t) \tag{$*$}$$

besitzt die allgemeine Lösung $y(t) = \gamma \exp(\lambda t) + g(t)$ ($\gamma \in \mathbb{R}$).

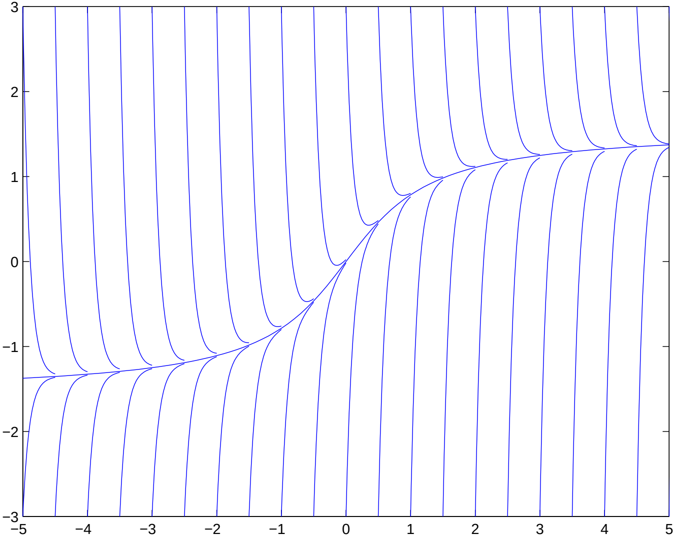
Wir wählen g als glatte Funktion, z.B. $g(t) = \arctan t$ und $\lambda = -10$. Auch hier setzt sich die Lösung aus einem glatten ($g(t)$ = Gleichgewichtsanteil) und einem schnell abklingenden Teil ($\gamma \exp(\lambda t)$ = transienter Anteil) zusammen.

Wir wählen die AB $y(0) = 0$ und approximieren die exakte Lösung

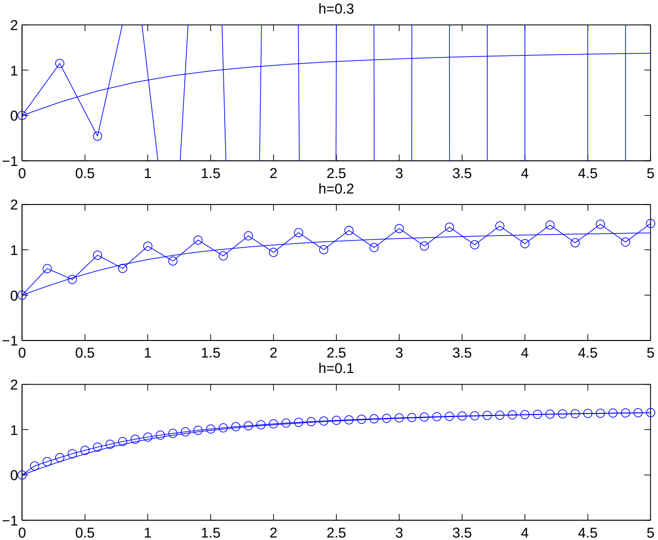
$$y(t) = \arctan t \quad \text{für} \quad t \in [0, 5]$$

mit dem

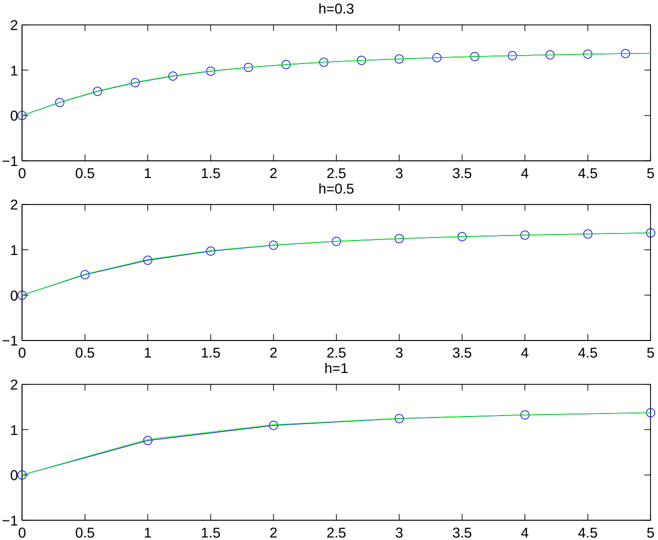
- expliziten Euler-Verfahren ($\mathscr{R}_A = \{\hat{h} : |\hat{h} + 1| < 1\}$) und dem
- impliziten Euler-Verfahren ($\mathscr{R}_A = \mathbb{C} \backslash \{\hat{h} : |\hat{h} - 1| \leqslant 1\}$).


3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4
5

Explizit:



Implizit:



Inhalt

Erinnerung. Ein numerisches Verfahren zur Lösung von

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0,$$

heißt A-stabil, wenn folgendes gilt:

Wendet man das Verfahren auf ein lineares Problem $\boldsymbol{y}' = A\boldsymbol{y}$ $(t \in [0, \infty))$ an und liegen die Eigenwerte von A alle in der linken Halbebene $\{\zeta : \mathrm{Re}\,\zeta < 0\}$, dann soll für die Näherungslösung stets $\lim_{n\to\infty} \|\boldsymbol{y}_n\| = 0$ gelten (die exakte Lösung erfüllt $\lim_{t\to\infty} \|\boldsymbol{y}(t)\| = 0$).

In konkreten Fällen wird die äquivalente Bedingung

$$\mathscr{R}_A \supseteq \{\hat{h} : \mathrm{Re}\,\hat{h} < 0\}$$

überprüft. Zur Definition des Stabilitätsbereichs $\mathscr{R}_A$ vergleiche Abschnitt 6 für lineare Mehrschrittverfahren und Abschnitt 3 für RKV (oder allgemeine Einschrittverfahren).

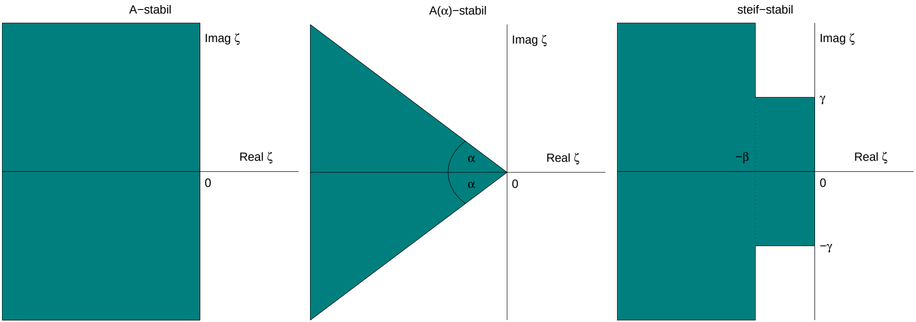
Abgeschwächte Stabilitätsbegriffe sind

- $A(\alpha)$-stabil für $\alpha \in (0, \pi/2) \quad \Leftrightarrow \quad \mathscr{R}_A \supseteq \{\hat{h} : -\alpha < \pi - \arg \hat{h} < \alpha\}$,
- A_0-stabil $\quad \Leftrightarrow \quad \mathscr{R}_A \supseteq (-\infty, 0)$,
- steif-stabil $\quad \Leftrightarrow \quad \mathscr{R}_A \supseteq \mathscr{R}_1(\beta) \cup \mathscr{R}_2(\beta, \gamma)$ für positive β und γ, wobei

$$\mathscr{R}_1(\beta) := \{\hat{h} : \operatorname{Re} \hat{h} < -\beta\} \text{ und}$$
$$\mathscr{R}_2(\beta, \gamma) := \{\hat{h} : -\beta \leqslant \operatorname{Re} \hat{h} < 0, |\operatorname{Im} \hat{h}| \leqslant \gamma\}.$$

Außerdem heißt ein Einschrittverfahren

- L-stabil, wenn es A-stabil ist und zusätzlich gilt: Bei Anwendung auf die Testgleichung
$$y' = \lambda y, \quad \operatorname{Re} \lambda < 0,$$
resultiert eine Näherung
$$y_{n+1} = R(h\lambda) y_n \quad \text{mit} \quad \lim_{h\lambda \to -\infty} |R(h\lambda)| = 0.$$

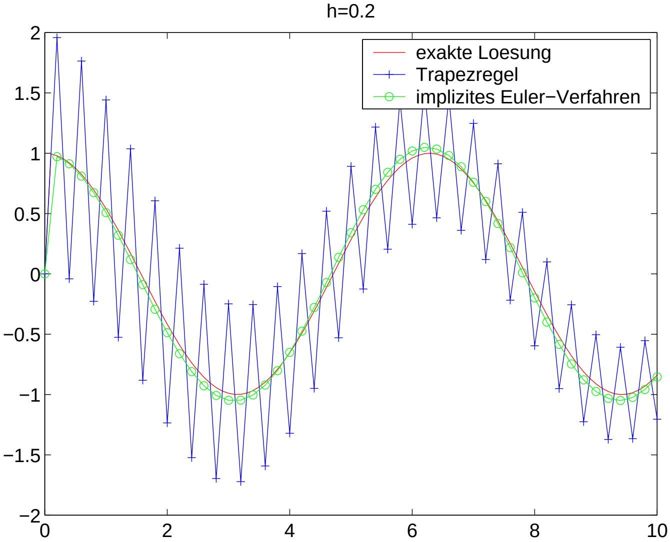


Offensichtlich gilt $(0 < \alpha \leqslant \arg(\beta + i\gamma))$

$$L\text{-stabil} \Rightarrow A\text{-stabil} \Rightarrow \text{steif-stabil} \Rightarrow A(\alpha)\text{-stabil} \Rightarrow A_0\text{-stabil}.$$

- Die Trapezregel
$$y_{n+1} = y_n + \frac{1}{2}h(f_{n+1} + f_n), \qquad \text{d.h.} \qquad R(\hat{h}) = \frac{1+\hat{h}/2}{1-\hat{h}/2},$$
ist A-stabil, aber nicht L-stabil.
- Das implizite Euler-Verfahren
$$y_{n+1} = y_n + hf_{n+1}, \qquad \text{d.h.} \qquad R(\hat{h}) = \frac{1}{1-\hat{h}},$$
ist dagegen L-stabil.

Wir illustrieren den Unterschied, indem wir mit beiden Verfahren die Lösung von (AWP$_2$) aus Abschnitt 1 approximieren. Um den transienten Anteil der Lösung besser sichtbar zu machen, werden die ABen $\boldsymbol{y}(0) = [0,0]^\top$ vorgeschrieben. Die exakte Lösung ist dann
$$\boldsymbol{y}(t) = \tfrac{-1}{999}\exp(-t)\begin{bmatrix}1\\1\end{bmatrix} + \tfrac{1}{999}\exp(-1000t)\begin{bmatrix}1\\-998\end{bmatrix} + \begin{bmatrix}\sin t\\ \cos t\end{bmatrix}.$$
Die folgende Abbildung zeigt die zweite Komponente der (Näherungs)lösung für $h = 0.2$.


h=0.2
exakte Loesung
Trapezregel
implizites Euler−Verfahren
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
2
4
6
8
10

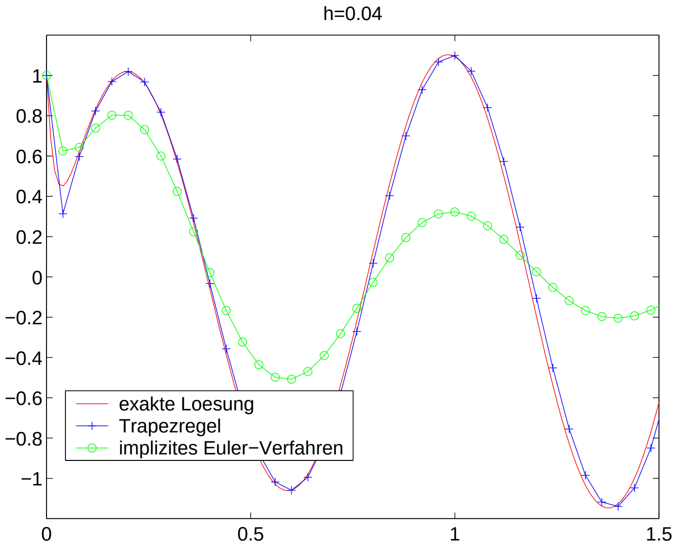
Das Beispiel

$$\boldsymbol{y}' = \begin{bmatrix} 42.2 & 50.1 & -42.1 \\ -66.1 & -58.0 & 58.1 \\ 26.1 & 42.1 & -34.0 \end{bmatrix} \boldsymbol{y}, \qquad \boldsymbol{y}(0) = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix},$$

mit der exakten Lösung

$$\boldsymbol{y}(t) = \begin{bmatrix} \exp(0.8t)\,\sin(8t) + \exp(-50t) \\ \exp(0.1t)\,\cos(8t) - \exp(-50t) \\ \exp(0.1t)\,(\sin(8t) + \cos(8t)) + \exp(-50t) \end{bmatrix}$$

zeigt, dass L-stabile Verfahren nicht immer besser sind als A-stabile (die folgende Abbildung zeigt die erste Komponente der (Näherungs)lösung für $h = 0.04$ und $t \in [0, 1.5]$). Beim impliziten Euler-Verfahren müsste man $h < 0.0031\ldots$ wählen, um eine akzeptable Näherung zu erhalten (warum?).



h=0.04
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
0.5
1
1.5
exakte Loesung
Trapezregel
implizites Euler−Verfahren

Inhalt

Wendet man ein Einschrittverfahren (ESV) auf die Testgleichung $y' = \lambda y$ an, so gilt $y_{n+1} = R(\lambda h) y_n$. Dabei ist R ein Polynom (bei expliziten Verfahren) oder eine (gebrochen) rationale Funktion in $\hat{h} = \lambda h$.

Besitzt das ESV die Konsistenzordnung p, so ist

$$R(\hat{h}) - \exp(\hat{h}) = O(\hat{h}^{p+1}) \qquad (\hat{h} \to 0).$$

Natürlicher Ansatz: Bestimme eine rationale Funktion R vom Typ (k, ℓ) (k = Zählergrad, ℓ = Nennergrad), so dass der Exponent μ in

$$R(\hat{h}) - \exp(\hat{h}) = O(\hat{h}^{\mu})$$

maximal wird. Es zeigt sich, dass das Maximum $\mu = k+\ell+1$ durch die in Abschnitt 4 eingeführten Padé-Approximationen $(k, \ell)_{\exp}$ realisiert wird.

$(k,\ell)_{\exp}(\zeta)$ **für** $0 \leqslant k,\ell \leqslant 3$**:**

k,ℓ	0	1	2	3
0	1	$\frac{1}{1-\zeta}$	$\frac{1}{1-\zeta+\frac{1}{2}\zeta^2}$	$\frac{1}{1-\zeta+\frac{1}{2}\zeta^2-\frac{1}{6}\zeta^3}$
1	$1+\zeta$	$\frac{1+\frac{1}{2}\zeta}{1-\frac{1}{2}\zeta}$	$\frac{1+\frac{1}{3}\zeta}{1-\frac{2}{3}\zeta+\frac{1}{6}\zeta^2}$	$\frac{1+\frac{1}{4}\zeta}{1-\frac{3}{4}\zeta+\frac{1}{4}\zeta^2-\frac{1}{24}\zeta^3}$
2	$1+\zeta+\frac{1}{2}\zeta^2$	$\frac{1+\frac{2}{3}\zeta+\frac{1}{6}\zeta^2}{1-\frac{1}{3}\zeta}$	$\frac{1+\frac{1}{2}\zeta+\frac{1}{12}\zeta^2}{1-\frac{1}{2}\zeta+\frac{1}{12}\zeta^2}$	$\frac{1+\frac{2}{5}\zeta+\frac{1}{20}\zeta^2}{1-\frac{3}{5}\zeta+\frac{3}{20}\zeta^2-\frac{1}{60}\zeta^3}$
3	$1+\zeta+\frac{1}{2}\zeta^2+\frac{1}{6}\zeta^3$	$\frac{1+\frac{3}{4}\zeta+\frac{1}{4}\zeta^2+\frac{1}{24}\zeta^3}{1-\frac{1}{4}\zeta}$	$\frac{1+\frac{3}{5}\zeta+\frac{3}{20}\zeta^2+\frac{1}{60}\zeta^3}{1-\frac{2}{5}\zeta+\frac{1}{20}\zeta^2}$	$\frac{1+\frac{1}{2}\zeta+\frac{1}{10}\zeta^2+\frac{1}{120}\zeta^3}{1-\frac{1}{2}\zeta+\frac{1}{10}\zeta^2-\frac{1}{120}\zeta^3}$

Eine rationale Approximation $R(\zeta)$ an $\exp(\zeta)$ heißt

- A-akzeptabel, wenn $|R(\zeta)| < 1$ für alle $\zeta \in \mathbb{C}$ mit $\operatorname{Re} \zeta < 0$,
- A_0-akzeptabel, wenn $|R(\zeta)| < 1$ für alle $\zeta \in \mathbb{R}$ mit $\zeta < 0$,
- L-akzeptabel, wenn R A-akzeptabel ist und $R(\zeta) \to 0$ für $\zeta \to -\infty$ gilt.

Satz 5.1

Es gilt

1. $(k,k)_{\exp}$ ist A-akzeptabel (G. Birkhoff, R.S. Varga 1962).
2. $(k,\ell)_{\exp}$ ist A_0-akzeptabel für $k \leqslant \ell$ (R.S. Varga 1961).
3. $(k,\ell)_{\exp}$ ist L-akzeptabel für $k \in \{\ell-1, \ell-2\}$ (B.L. Ehle, 1969).
4. $(k,\ell)_{\exp}$ ist genau dann A-akzeptabel, wenn $k \in \{\ell, \ell-1, \ell-2\}$ (Ehle-Vermutung; G. Wanner, E. Hairer, S.P. Nørsett 1978).

Der Beweis der vierten Aussage basiert auf der Theorie der Ordnungssterne. Ist $R(\zeta)$ eine rationale Funktion, so heißt

$$\mathscr{S}_R := \{\zeta \in \mathbb{C} : |R(\zeta)| > |\exp(\zeta)|\}$$

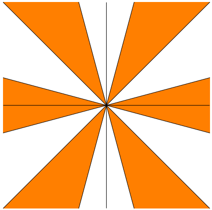
der zugehörige Ordnungsstern. Sei $\mathscr{T}_R := \mathbb{C} \backslash \mathscr{S}_R$ sein Komplement.

Vier Eigenschaften.

(1) R ist genau dann eine Approximation der Ordnung p an die Exponentialfunktion, d.h.

$$R(\zeta) - \exp(\zeta) = O(\zeta^{p+1}) \quad (\zeta \to 0),$$

wenn Folgendes gilt: Für $\zeta \to 0$ besteht sowohl $\mathscr{S}_R$ als auch $\mathscr{T}_R$ aus genau $p+1$ Sektoren mit Winkel jeweils $\pi/(p+1)$. Die Sektoren von $\mathscr{S}_R$ und $\mathscr{T}_R$ trennen sich.

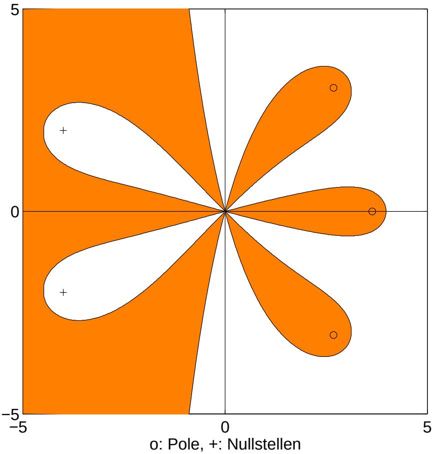


$(p = 5)$

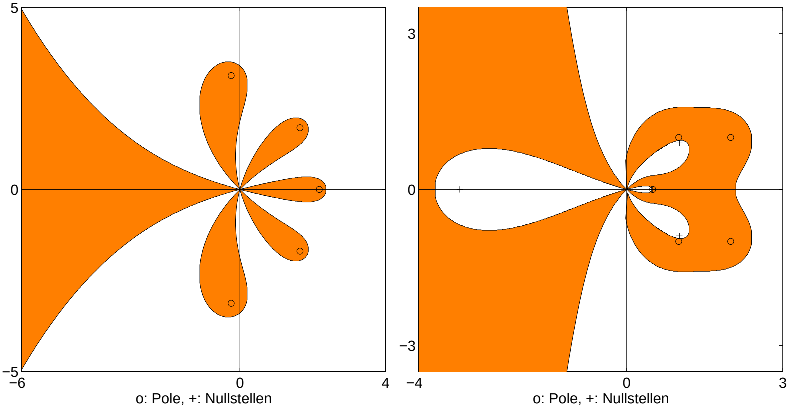
(2) Der Rand von $\mathscr{S}_R$ enthält genau zwei unbeschränkte Zweige.

(3) Komponenten von $\mathscr{S}_R$, die k Sektoren enthalten, nennt man Finger der Ordnung k. Komponenten von $\mathscr{T}_R$, die k Sektoren enthalten, heißen duale Finger der Ordnung k.
Es gilt: Jeder beschränkte Finger der Ordnung k enthält mindestens k Pole von R (Vielfachheiten mitzählen). Jeder beschränkte duale Finger der Ordnung k enthält mindestens k Nullstellen von R (Vielfachheiten mitzählen).

(4) R ist genau dann A-akzeptabel, wenn $\mathscr{S}_R$ keine Punkte der imaginären Achse enthält und R keine Pole in der linken Halbebene $\operatorname{Re} \zeta < 0$ besitzt.



Ordnungssterne der Padé-Approximationen $(2,3)_{\exp}$ (links) und $(1,4)_{\exp}$ (rechts).



Ordnungssterne der Padé-Approximation $(0,5)_{\exp}$ (links)
und einer (nicht A-akzeptablen) $(4,5)$-Approximation R der Ordnung 5 (rechts)
$(R(\zeta) = [\frac{41}{24}\zeta^4 + \frac{7}{6}\zeta^3 - 9\zeta^2 + 14\zeta - 5]/[\zeta^5 + \frac{13}{2}\zeta^4 + 18\zeta^3 - \frac{51}{2}\zeta^2 + 19\zeta - 5])$.

Inhalt

Satz 5.2 (Zweite Dahlquist-Barriere)

Es gilt:

- Kein explizites lineares MSV ist A-stabil.
- Ein A-stabiles lineares MSV besitzt höchstens die Ordnung 2.
- Das A-stabile lineare MSV mit Konsistenzordnung 2 und kleinster Fehlerkonstante ist die Trapezregel.

Satz 5.3 (C.W. Cryer, 1973)

Es gilt:

- Kein explizites lineares MSV ist A_0-stabil.
- Es gibt A_0-stabile lineare MSV beliebig hoher Konsistenzordnung.

Beispielsweise sind die BDF-Verfahren aus Abschnitt 7 alle A_0-stabil.

Inhalt

Entscheidend (vgl. Abschnitt 3)

$$R(\hat{h}) = 1 + \hat{h}\boldsymbol{b}^\top(I - \hat{h}A)^{-1}\boldsymbol{e} = \frac{\det(I - \hat{h}(A - \boldsymbol{e}\boldsymbol{b}^T))}{\det(I - \hat{h}A)}.$$

	Ordnung	Typ von R	Stabilität
Gauß	$2m$	(m, m)	A-stabil
Gauß-Radau	$2m-1$	$(m-1, m)$	L-stabil
Gauß-Lobatto	$2m-2$	$(m-1, m-1)$	A-stabil

Inhalt

Alle bisher getroffenen Stabilitätsaussagen basieren auf der Untersuchung linearer Systeme mit konstanten Koeffizienten (lineare Stabilitätstheorie). Oft wird versucht, diese Aussagen mit folgender Argumentation zu verallgemeinern:

Die Lösung $\boldsymbol{y}$ des AWPs $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$, $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$ kann durch die Lösung der linearen Variationsgleichung

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}^*(t)) + \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}^*(t))[\boldsymbol{y} - \boldsymbol{y}^*(t)] \qquad (*)$$

approximiert werden. Lokal ist $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}^*)$ nahezu zeitunabhängig, so dass $(*)$ näherungsweise die Form $\boldsymbol{y}' = A\boldsymbol{y} + \boldsymbol{g}(t)$ besitzt.

Zwei Beispiele sollen zeigen, dass Skepsis gegenüber diesem "frozen Jacobian argument" angebracht ist.

Beispiel 1.

$$\boldsymbol{y}' = \begin{bmatrix} -1 - 9\cos^2(6t) + 6\sin(12t) & 12\cos^2(6t) + 4.5\sin(12t) \\ -12\sin^2(6t) + 4.5\sin(12t) & -1 - 9\sin^2(6t) - 6\sin(12t) \end{bmatrix} \boldsymbol{y}.$$

Die Systemmatrix $A(t)$ besitzt die (t-unabhängigen!) Eigenwerte

$$\lambda_1 = -1, \quad \lambda_2 = -10.$$

Obwohl die Eigenwerte negativ sind, ist die allgemeine Lösung

$$\boldsymbol{y}(t) = \kappa_1 \exp(2t) \begin{bmatrix} \cos(6t) + 2\sin(6t) \\ 2\cos(6t) - \sin(6t) \end{bmatrix} + \kappa_2 \exp(-13t) \begin{bmatrix} \sin(6t) - 2\cos(6t) \\ 2\sin(6t) + \cos(6t) \end{bmatrix}$$

sicher nicht monoton fallend für $t \geqslant 0$.

Beispiel 2.

$$\boldsymbol{y}' = \begin{bmatrix} \frac{-1}{2t} & \frac{2}{t^3} \\ \frac{-t}{2} & \frac{-1}{2t} \end{bmatrix} \boldsymbol{y}, \qquad t \geqslant 1. \tag{$\Diamond$}$$

Die Systemmatrix $A(t)$ besitzt die Eigenwerte

$$\lambda_{1,2} = \frac{-1 \pm 2i}{2t},$$

die für $t \geqslant 1$ negativen Realteil besitzen. Die allgemeine Lösung ist

$$\boldsymbol{y}(t) = \kappa_1 \begin{bmatrix} t^{-3/2} \\ -\frac{1}{2} t^{1/2} \end{bmatrix} + \kappa_2 \begin{bmatrix} 2t^{-3/2} \log t \\ t^{1/2}(1 - \log t) \end{bmatrix}.$$

Für $\kappa_1 = 1$, $\kappa_2 = 0$ ist

$$\|\boldsymbol{y}\|_2^2 = t^{-3} + \tfrac{1}{4} t$$

streng monoton wachsend (für $t \geqslant 12^{1/4} = 1.86\dots$).

Definition 5.4

Das System $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ heißt dissipativ in $[t_0, t_{\text{end}}]$, wenn

$$\langle \boldsymbol{f}(t, \boldsymbol{y}) - \boldsymbol{f}(t, \tilde{\boldsymbol{y}}), \boldsymbol{y} - \tilde{\boldsymbol{y}} \rangle \leqslant 0$$

für alle $t \in [t_0, t_{\text{end}}]$ und alle $\boldsymbol{y}$, $\tilde{\boldsymbol{y}}$ aus dem Definitionsbereich von $\boldsymbol{f}$ gilt. $\langle \cdot, \cdot \rangle$ bezeichnet ein (z.B. das Euklidische) Innenprodukt im $\mathbb{R}^n$.

Satz 5.5

Ist das System $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ dissipativ in $[t_0, t_{\mathrm{end}}]$, dann ist es auch kontraktiv in $[t_0, t_{\mathrm{end}}]$. Dies bedeutet: Für je zwei Lösungen $\boldsymbol{y}$, $\tilde{\boldsymbol{y}}$ von $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ mit Anfangsbedingungen $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, $\tilde{\boldsymbol{y}}(t_0) = \tilde{\boldsymbol{y}}_0$, $\boldsymbol{y}_0 \neq \tilde{\boldsymbol{y}}_0$, gilt

$$\|\boldsymbol{y}(t_2) - \tilde{\boldsymbol{y}}(t_2)\| \leqslant \|\boldsymbol{y}(t_1) - \tilde{\boldsymbol{y}}(t_1)\|$$

für alle t_1, t_2 mit $t_0 \leqslant t_1 \leqslant t_2 \leqslant t_{\mathrm{end}}$. ($\| \cdot \| := \langle \cdot, \cdot \rangle^{1/2}$)

Ob ein System dissipativ (kontraktiv) ist, kann mit Hilfe logarithmischer Normen entschieden werden: Sei $\|\cdot\|$ eine (von einer Vektornorm induzierte) Matrixnorm im $\mathbb{R}^{n\times n}$. Dann heißt

$$\mu(A) = \mu_{\|\cdot\|}(A) := \lim_{h\to 0+} \frac{\|I_n + hA\| - 1}{h} \qquad (A \in \mathbb{R}^{n\times n})$$

die zu $\|\cdot\|$ gehörige logarithmische Norm. Ist $\|\cdot\| = \|\cdot\|_p$, so schreibt man $\mu_p(\cdot)$ für die zugehörige logarithmische Norm.

Eigenschaften.

(1) Die logarithmische Norm ist – trotz ihres Namens – keine Norm.

(2) Wird $\|\cdot\|$ von einem Innenprodukt $\langle\cdot,\cdot\rangle$ induziert, so gilt

$$\mu(A) = \max_{\|\boldsymbol{x}\|=1} \langle A\boldsymbol{x}, \boldsymbol{x}\rangle.$$

(3) $\mu_2(A) = \rho(\frac{1}{2}(A + A^\top))$ (ρ = Spektralradius).

(4) Ist $\boldsymbol{f}(t,\boldsymbol{y}) = A(t)\boldsymbol{y}$, so ist

$$\langle \boldsymbol{f}(t,\boldsymbol{y}) - \boldsymbol{f}(t,\tilde{\boldsymbol{y}}), \boldsymbol{y} - \tilde{\boldsymbol{y}}\rangle = \langle A(t)(\boldsymbol{y} - \tilde{\boldsymbol{y}}), \boldsymbol{y} - \tilde{\boldsymbol{y}}\rangle \leqslant \mu(A(t))\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2.$$

Satz 5.6

Seien $\|\cdot\|$ eine Norm im $\mathbb{R}^n$ und $\nu(t) : \mathbb{R} \to \mathbb{R}$ eine stückweise stetige Funktion mit

$$\mu(\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y})) \leqslant \nu(t)$$

für alle $t \in [t_0, t_{\text{end}}]$ und alle $\boldsymbol{y}$ aus dem Definitionsbereich von $\boldsymbol{f}$. Dann gilt für je zwei Lösungen $\boldsymbol{y}$, $\tilde{\boldsymbol{y}}$ von $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ mit Anfangsbedingungen $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, $\tilde{\boldsymbol{y}}(t_0) = \tilde{\boldsymbol{y}}_0$, $\boldsymbol{y}_0 \neq \tilde{\boldsymbol{y}}_0$,

$$\|\boldsymbol{y}(t_2) - \tilde{\boldsymbol{y}}(t_2)\| \leqslant \exp\left(\int_{t_1}^{t_2} \nu(s)\,\mathrm{d}s\right) \|\boldsymbol{y}(t_1) - \tilde{\boldsymbol{y}}(t_1)\|$$

für alle t_1, t_2 mit $t_0 \leqslant t_1 \leqslant t_2 \leqslant t_{\text{end}}$.

Beispiel.

Für das System ($\Diamond$) von Seite 279 gilt:

$$\boldsymbol{f_y}(t, \boldsymbol{y}) = A(t) = \begin{bmatrix} \frac{-1}{2t} & \frac{2}{t^3} \\ \frac{-t}{2} & \frac{-1}{2t} \end{bmatrix}.$$

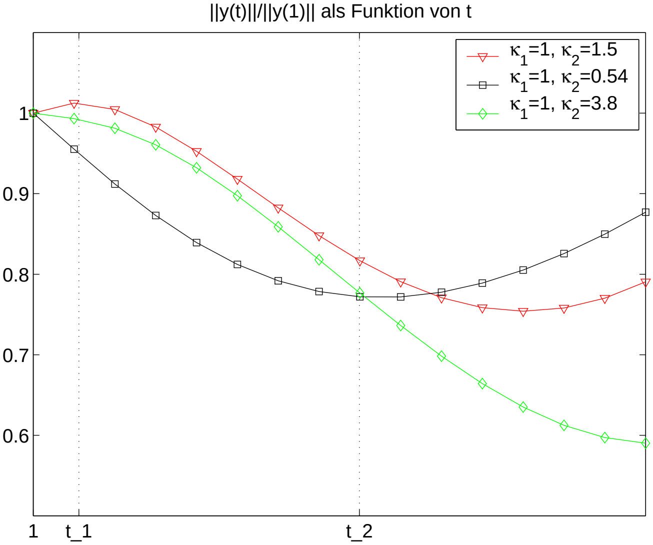
Es folgt

$$\mu_2(A(t)) = \rho(\tfrac{1}{2}(A(t) + A(t)^\top)) = \max\left\{\left|\tfrac{-1}{2t} \pm \left(\tfrac{1}{t^3} - \tfrac{t}{4}\right)\right|\right\}$$

und damit $\mu_2(A(t)) \leqslant 0$ genau dann, wenn

$$t \in [t_1, t_2] \approx [1.1, 1.8] \quad \text{mit} \quad t_{1,2} = (\sqrt{5} \mp 1)^{1/2}.$$

Das System ist daher in $[t_1, t_2]$ dissipativ (und kontraktiv). Folglich ist $\|\boldsymbol{y}(t)\|_2$ für jede Lösung $\boldsymbol{y}$ in $[t_1, t_2]$ monoton fallend.



||y(t)||/||y(1)|| als Funktion von t
κ1=1, κ2=1.5
κ1=1, κ2=0.54
κ1=1, κ2=3.8
1
0.9
0.8
0.7
0.6
1
t_1
t_2

Definition 5.7

Ein Runge-Kutta-Verfahren gegeben durch das Butcher-Tableau $\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^\top \end{array}$ heißt algebraisch stabil, wenn die symmetrischen Matrizen

$$B := \operatorname{diag}(\beta_1, \beta_2, \dots, \beta_m) \quad \text{und} \quad M := BA + A^\top B - \boldsymbol{b}\boldsymbol{b}^\top$$

beide positiv semidefinit sind.

Satz 5.8

Seien $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ ein dissipatives System und $\{\boldsymbol{y}_n\}$, $\{\tilde{\boldsymbol{y}}_n\}$, $\boldsymbol{y}_0 \neq \tilde{\boldsymbol{y}}_0$, Näherungslösungen, die aus der Anwendung eines Runge-Kutta-Verfahrens resultieren. Dann gilt: Ist das Verfahren algebraisch stabil, so sind die Näherungen kontraktiv, d.h.

$$\|\boldsymbol{y}_{n+1} - \tilde{\boldsymbol{y}}_{n+1}\| \leqslant \|\boldsymbol{y}_n - \tilde{\boldsymbol{y}}_n\| \qquad (n = 0, 1, \dots).$$

Beispiele.

- Die m-stufigen Gauß-Formeln ($m \in \{1,2,3\}$) aus Abschnitt 5 sind alle algebraisch stabil (hier ist M die Nullmatrix!).
- Die Gauß-Radau-Formel
$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$
ist algebraisch stabil, denn:
$$M = \frac{1}{16}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \text{ d.h. } \Lambda(M) = \{0, 1/8\}.$$
- Die Trapezregel (aufgefasst als implizites zweistufiges RKV) ist nicht algebraisch stabil ($\Lambda(M) = \{\pm 1/2\}$).

Inhalt

Diskretisierungen von Reaktions-Diffusions-Transport Problemen (z.B. reaktive Strömung) führen oft auf steife nichtlineare Anfangswertaufgaben der Form

$$\boldsymbol{u}'(t) = \boldsymbol{F}(\boldsymbol{u}(t)). \tag{$*$}$$

Nach Linearisierung um $\boldsymbol{u}(t_n)$

$$\boldsymbol{F}(\boldsymbol{u}(t)) = \boldsymbol{A}_n[\boldsymbol{u}(t) - \boldsymbol{u}(t_n)] + \boldsymbol{F}(\boldsymbol{u}(t_n)) + \boldsymbol{R}(\boldsymbol{u}(t_n)) \qquad \boldsymbol{A}_n := \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{u}}(\boldsymbol{u}(t_n))$$

lautet $(*)$

$$\boldsymbol{u}'(t) = \boldsymbol{A}_n[\boldsymbol{u}(t) - \boldsymbol{u}(t_n)] + \boldsymbol{F}(\boldsymbol{u}(t_n)) + \boldsymbol{R}(\boldsymbol{u}(t)) \tag{$**$}$$

- Zeitschrittverfahren behandeln typischerweise den linearen Anteil implizit (z.B. SIMPLE-Verfahren)
- Erfordert Lösung eines LGS in jedem Zeitschritt.

Alternative: Lösung von $(**)$ lautet

$$\boldsymbol{u}(t_n + \Delta t) = \boldsymbol{u}(t_n) + \underbrace{(e^{\Delta t \boldsymbol{A}_n} - \boldsymbol{I})\boldsymbol{A}_n^{-1}\boldsymbol{F}(\boldsymbol{u}(t_n))}_{=:\phi(\boldsymbol{A})\boldsymbol{b}}$$
$$+ \int_{t_n}^{t_n+\Delta t} e^{(t_n+\Delta t - s)\boldsymbol{A}_n}\boldsymbol{R}(\boldsymbol{u}(s))\,ds$$

mit

$$\phi(z) = \frac{e^{\Delta t z} - 1}{z}, \qquad \boldsymbol{b} = \boldsymbol{F}(\boldsymbol{u}(t_n)).$$

Krylov-Approximation von $\phi(\boldsymbol{A}_n)\boldsymbol{b}$ in Kombination mit verschiedenen Approximationen des Integraltermes führen auf sog. exponentielle Integratoren zur expliziten Zeitintegration steifer Systeme.

Hamiltonsche Systeme

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}(p,q), \quad \dot{q}_i = \frac{\partial H}{\partial p_i}(p,q),$$

besitzen zwei entscheidende Eigenschaften:

(a) Die Lösungen erhalten den Wert der Hamiltonschen Funktion $H(p,q)$.

(b) Der zugehörige Fluß ist symplektisch, d.h. er erhält die 2-Form

$$\omega^2 = \sum_{i=1}^{n} \mathrm{d}p_i \wedge \mathrm{d}q_i$$

Wir betrachten den harmonischen Oszillator

$$H(p,q) = \frac{1}{2}(p^2 + k^2 q^2), \text{ d.h. } \dot{p} = -k^2 q, \ \dot{q} = p.$$

Anwendung von Euler, implizitem Euler, verbessertem Euler und (impliziter) Mittelpunktsregel:

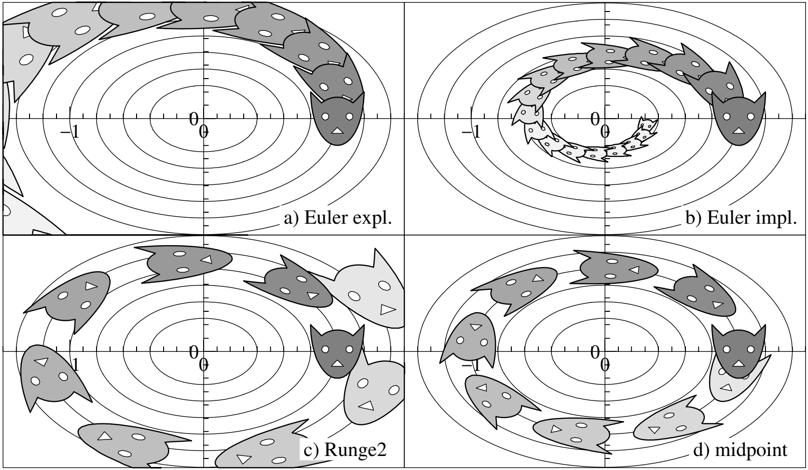


Fig. 16.1. Destruction of symplecticity of a Hamiltonian flow, $k = (\sqrt{5}+1)/2$

aus Hairer, Norsett & Wanner, 2008

Simulation verschieder mathematischer Modelle (CME, SDE, ODE) der autokatalytischen chemischen Reaktion $S \to \varnothing$. ausgehend von 100 Molekülen.

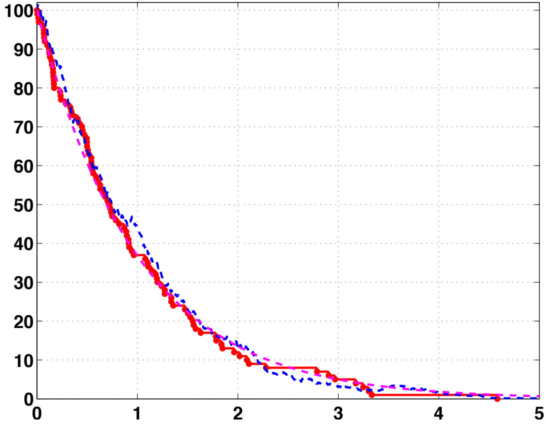


Bild: D. Higham

CME $$p_i'(t) = \lambda(i+1)p_{i+1}(t) - \lambda i p_i(t)$$

ergibt $\mathbf{E}[\boldsymbol{X}(T)] = Ne^{-\lambda t}$, $\mathbf{Var}[\boldsymbol{X})t)] = Ne^{-\lambda t}(1-e^{-\lambda t})$

SDE $$\mathrm{d}\boldsymbol{Y}(t) = -\lambda \boldsymbol{Y}(t)\mathrm{d}t - \sqrt{\lambda \boldsymbol{Y}(t)}\,\mathrm{d}W(t)$$

ergibt $\mathbf{E}[\boldsymbol{Y}(T)] = Ne^{-\lambda t}$, $\mathbf{Var}[\boldsymbol{Y})t)] = Ne^{-\lambda t}(1-e^{-\lambda t})$

ODE $$z'(t) = -\lambda z(t),\ z(0) = N$$

ergibt $z(t) = Ne^{-\lambda t}$.

Dieselben Simulationen ausgehend von 10 Molekülen.

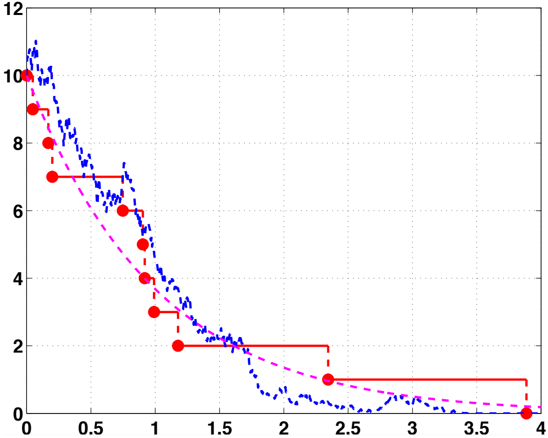


Bild: D. Higham