Stefan A. Funken

Numerische Mathematik IV

SKRIPT, UNIVERSITÄT ULM, SOMMERSEMESTER 2010

Vorbemerkung.

Stand. Ulm, April 2010.

Inhaltsverzeichnis

Symbolverzeichnis

Bemerkung. Mathematische Symbole sind auch (sofern möglich) alphabetisch im Stichwortverzeichnis aufgeführt.

$\mathbb{N}$	natürliche Zahlen (ohne 0)
$\mathbb{N}_0$	natürliche Zahlen vereinigt mit 0
$\mathbb{R}^d$	Euklidischer Raum von (Spalten-) Vektoren in d Komponenten

Bemerkung: Diese Seite ändert sich noch während der Vorlesung, d.h. weitere, den Studenten unbekannte Notationen, werden laufend ergänzt.

1 GEWÖHNLICHE DIFFERENTIALGLEICHUNGEN

Viele Anwendungen in Natur- und Ingenieurwissenschaften führen auf Differentialgleichungen oder Differentialgleichungssysteme, welche die zeitliche Änderung einer oder mehrerer Zustandsgrößen beschreiben. Exemplarisch seien hier die Berechnung der Flugbahn eines Raumfahrzeugs beim Wiedereintritt in die Erdatmosphäre, Räuber-Beute-Modelle oder chemische Reaktionen genannt, deren Modellierungen im Anhang hergeleitet werden.
Wir betrachten im folgenden das System von N gewöhnlichen Differentialgleichungen erster Ordnung

$$y'(t) = f(t, y(t)) \qquad (t \in [a, b]) \tag{1.1}$$

für die gesuchte Lösungsfunktion y unter der Anfangsbedingung

$$y(a) = y_0 \in \mathbb{R}^N \tag{1.2}$$

mit einem gegebenen endlichen Intervall $[a, b]$ und einer Funktion

$$f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N. \tag{1.3}$$

Das Problem (1.1), (1.2) bezeichnen wir als Anfangswertproblem.
Die Notation $y' = f(t, y)$ verwenden wir als Kurzform; dabei bedeute Differenzierbarkeit hier komponentenweise Differenzierbarkeit und es sei $y'(t) = (y_1'(t), \ldots, y_N'(t))^{\mathsf{T}}$.

1.1 EXISTENZ- UND EINDEUTIGKEITSSÄTZE

Wir wollen hier kurz auf die Existenz und Eindeutigkeit der Lösung bei Anfangswertproblemen für gewöhnliche Differentialgleichungen eingehen, auch wenn wir später stillschweigend voraussetzen, daß die betreffenden Voraussetzungen erfüllt sind und das Anfangswertproblem (1.1), (1.2) jeweils eine eindeutig bestimmte Lösung besitzt.
Bei der Eindeutigkeit einer Lösung spielt die folgende Lipschitzbedingung für Funktionen f von der Form (1.3) eine wesentliche Rolle,

$$\|f(t, u) - f(t, v)\| \le L\|u - v\| \qquad (t \in [a, b],\ u, v \in \mathbb{R}^N), \tag{1.4}$$

mit einer Konstanten $L > 0$ und einer beliebigen Vektornorm $\|\cdot\|$.
Da Stetigkeit der Funktion f die Existenz einer Lösung liefert, erhält man aus dem

Satz 1 (Existenzsatz von Peano) *Ist $f(x, y)$ in einem Gebiet D stetig, so geht durch jeden Punkt $(\xi, \eta) \in D$ mindestens eine Lösung der Differentialgleichung (1.1). Jede Lösung läßt sich bis zum Rand von D fortsetzen.*

Beispiele 2
i. Bevökerungswachstum. Sei $p(t)$ die Größe der Bevölkerung zur Zeit t, bzw. p_0 die Größe der Bevölkerung zur Zeit t_0. Die Geburtenrate sei mit $g(t, p$ bezeichnet und $s(t, p)$ sei die Sterberate. Die Funktion p löst die Anfangswertaufgabe

$$\frac{d}{dt}p(t) = (g(t, p) - s(t, p))\, p(t), \quad p(t_0) = p_0\,.$$

ii. Glatte rechte Seite f. Das Anfangswertproblem

$$\frac{d}{dt}y = 1 + y^2, \quad y(0) = 0$$

hat die Lösung $y = \tan(x)$. Als stetig differenzierbare Funktion existiert diese nur für $|x| < \pi$, obwohl $f(x, y) = 1 + y^2$ in $C^\infty(\mathbb{R}^2)$ ist. Wie verträgt sich das mit dem Satz von Piano?

iii. Mathematisches Pendel. Ein mathematisches Pendel ist die Idealisierung eines realen Pendels. Hierbei herrscht keinerlei Reibung und die gesamte Masse des Pendels (inklusive Faden) ist in einem Punkt konzentriert. Die Bewegungsgleichung für den Auslenkungswinkel (bezüglich der Ruhelage) des Pendels lautet

$$\ddot{\varphi}(t) + \frac{g}{\ell}\sin(\varphi(t)) = 0\,,$$

wobei g die Erdbeschleunigung und ℓ die Fadenlänge ist. Die gewöhnliche Differentialgleichung 2. Ordnung $\ddot{y} = f(t, y, \dot{y})$ wird durch den Ansatz

$$y = y_1, \quad \dot{y} = y_2$$

in das System einer gewöhnlichen Differentialgleichung 1. Ordnung überführt, nämlich

$$\frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ f(t, y_1, y_2) \end{pmatrix}.$$

Im Falle des mathematischen Pendels ergibt sich also das folgende System einer gewöhnlichen Differentialgleichung 1. Ordnung,

$$\frac{d}{dt}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} \varphi_2 \\ -g/\ell \cdot \sin(\varphi_1) \end{pmatrix}.$$

Der Ort des Pendels zum Zeitpunkt t ist dann $(\ell \sin\varphi(t), \ell\cos\varphi(t))$, wobei die $y - Achse$ nach unten zeigt.

iv. Doppelpendel. Hängt man an den Arm eines Pendels ein weiteres Pendel, so spricht man vom Doppelpendel. Dies ist ein beliebtes Modell zur Demonstration chaotischer Prozesse. Die Bewegungsgleichungen für die Auslenkungswinkel φ_1 und $\varphi_2 i$, welche die Fäden mit der Länge ℓ_1 und ℓ_2 mit der Vertikalen einschließen, lautet

$$\begin{aligned}(m_1 + m_2)\ell_1\ddot{\varphi}_1 + m_2\ell_2\ddot{\varphi}_2\cos(\varphi_1 - \varphi_2) + m_2\ell_2\dot{\varphi}_1^2\sin(\varphi_1 - \varphi_2) + g(m_1 + m_2)\sin(\varphi_2) &= 0\\ m_2\ell_2\ddot{\varphi}_2 + m_2\ell_1\ddot{\varphi}_1\cos(\varphi_1 - \varphi_2) - m_2\ell_1\dot{\varphi}_1^2\sin(\varphi_1 - \varphi_2) + gm_2\sin(\varphi_2) &= 0\end{aligned}$$

Die kartesischen Koordinaten (x_2, y_2) des zweiten Massepunktes kann man mittels φ_1 und φ_2 ausdrücken, nämlich

$$x_2 = \ell_1\sin\varphi_1 + \ell_2\sin\varphi_2, \quad y_2 = \ell_1\cos\varphi_1 + \ell_2\cos\varphi_2\,.$$

Hierbei liegt der Aufhängungspunkt im Ursprung und die y-Achse zeigt nach unten. Mit dem Ansatz

$$u_1 = \varphi_1\,, u_2 = \dot{\varphi}_1\,, u_3 = \varphi_2 \text{ und } u_4 = \dot{\varphi}_2$$

erhält man wie im Beispiel zuvor ein System einer gewöhnlichen Differentialgleichung 1. Ordnung. Nun jedoch von der Dimension vier.

v. Anfangswertproblem mit unendlich vielen Lösungen. Durch die rein formale Trennung der Veränderlichen und formales Integrieren erhalten wir zu der Differentialgleichung

$$y' = \sqrt{|y|}, \tag{1.5}$$

sofern wir nichtnegative Lösungen unterstellen, $2\sqrt{y} = t + c$ mit einer frei wählbaren Integrationskonstanten c. Dies motiviert die mit dem Parameter $c \in \mathbb{R}$ behaftete Funktion

$$y = \frac{(t+c)^2}{4} \tag{1.6}$$

als mögliche Lösung von (1.5) zu testen. Tatsächlich ist (1.6) differenzierbar und es gilt $y' = (t+c)/2 = \sqrt{|y|}$. Untersuchen wir nun z. B. die Anfangsbedingung

$$y(0) = 0, \tag{1.7}$$

so erhalten wir $y(t) = 0$ als offensichtliche Lösung des Anfangswertproblems (1.5) und (1.7). Es können jedoch überall weitere Funktionen der Form (1.6) abzweigen. Betrachten wir dazu folgende Funktion $y(\cdot, d)$ mit $d > 0$,

$$y(t,d) := \begin{cases} 0 & \text{für} \quad 0 \le t \le d, \\ \frac{(t-d)^2}{4} & \text{für} \quad d \le t. \end{cases} \tag{1.8}$$

Diese Funktion ist offensichtlich überall stetig und differenzierbar mit stetiger Ableitung. Sowohl für $t < d$ also auch für $t > d$ erfüllt (1.8) die Differentialgleichung (1.5) und (1.7). Da $y(d) = 0 = y'(d)$ ist, gilt (1.5) für alle t und (1.8) ist somit Lösung des Anfangswertproblems (1.5) und (1.7). Es liegen also unendlich viele Lösungen vor.

Neben der Existenz und Eindeutigkeit einer Lösung für Anfangswertprobleme (1.1), (1.2) liefert der folgende Satz auch ein wichtiges Resultat zur stetigen Abhängigkeit von den Anfangswerten.

Satz 3 (Picard-Lindelöf) *Es bezeichne $\|\cdot\|_2$ die euklidische Norm. Die Funktion $f(x,y)$ sei stetig auf dem kompakten Rechteck*

$$D := \{(x,y)|\ |x-a| \le \alpha, \|y-y_0\|_2 \le \beta\} \qquad (\alpha, \beta > 0)$$

und genüge dort einer Lipschitzbedingung (1.4) bezüglich y mit der Lipschitzkonstanten L.

a) Dann besitzt das Anfangswertproblem (1.1), (1.2) genau eine Lösung auf einem Intervall $I := [a-\gamma, a+\gamma]$ mit $\gamma = \min(\alpha, \beta/\max_{(x,y)\in D} f(x,y))$

b) Für differenzierbare Funktionen $y, \overline{y} : [a,b] \to \mathbb{R}^N$ mit

$$\begin{aligned} y'(t) &= f(t,y) \qquad (t \in [a,b]), \quad y(a) = y_0 \\ \overline{y}'(t) &= f(t,\overline{y}) \qquad (t \in [a,b]), \quad \overline{y}(a) = \overline{y}_0 \end{aligned}$$

gilt die Abschätzung

$$\|y(t) - \overline{y}(t)\|_2 \le \mathrm{e}^{L(t-a)} \|y_0 - \overline{y}_0\|_2 \qquad (t \in I).$$

1.2 Theorie der Einschrittverfahren

Das explizite Euler-Verfahren und die Methode der Taylorreihe

Beschränken wir uns zuerst auf die Situation $N = 1$. Aus geometrischen Anschauungen läßt sich recht einfach das explizite Euler-Verfahren motivieren, da die Differentialgleichung (1.1) im Punkt (t_0, y_0) mit dem Wert $y'(t_0) = f(t_0, y_0)$ die Steigung der gesuchten Funktion festlegt, d.h. es ergibt sich zu der Schrittweite h folgende Approximation

$$y(t_0 + h) \approx y(t_0) + h f(t_0, y_0). \tag{1.9}$$

Definieren wir nun Punkte

$$t_k = t_0 + kh \quad (k = 0, 1, 2, \ldots)$$

so ergibt sich durch sukzessives Anwenden von (1.9) das

explizite Euler-Verfahren

$$y_{k+1} = y_k + hf(t_k, y_k) \quad (k = 0, 1, 2, \ldots), \tag{1.11}$$

wobei y_k eine Approximation an $y(t_k)$ $(k = 0, 1, 2, \ldots)$ bezeichne. Man erhält somit eine stückweise lineare Approximation der gesuchten Funktion $y(t)$, weswegen man dieses Verfahren auch Polygonzugverfahren nennt.
Es treten nun folgende Fragen auf, mit denen wir uns im weiteren beschäftigen werden:

- Wird die Approximation „besser“, wenn man die Schrittweite h verkleinert?
- Wenn ja, in welchem Sinne und von welcher Größenordnung ist sie?
- In jedem Schritt des expliziten Euler-Verfahrens wird im allgemeinen ein Fehler $y(t_k) - y_k$ $(k = 0, 1, 2, \ldots)$ gemacht. Wie setzen sich solche Fehler in weiteren Schritten fort?

Beispiel 4 (zum expliziten Euler-Verfahren) Betrachten wir die Anfangswertaufgabe

$$y'(t) = -t \sin(\pi y(t)), \quad y(0) = \frac{1}{2}. \tag{1.12}$$

Die exakte Lösung lautet $y(t) = \frac{2}{\pi}\text{arc tan}(\mathrm{e}^{-\pi t^2/2})$. Mit dem expliziten Euler-Verfahren erhalten wir die in Tabelle 4 aufgelisteten Näherungswerte für verschieden Schrittweiten h an gleichen Stellen t_k und die zugehörigen Fehler $e_k = y(t_k) - y_k$. Man beachte, daß bei Zehntelung der Schrittweite sich auch der entsprechende Fehler zehntelt. Der Fehler nimmt proportional zur Schrittweite ab.

Tab. 1.1: Numerische Ergebnisse des expl. Euler-Verfahrens für das Anfangswertproblem (1.12) zu verschiedenen Schrittweiten h.

		Schrittweite $h = 0.1$		Schrittweite $h = 0.01$		Schrittweite $h = 0.001$	
t_k	$y(t_k)$	y_k	e_k	y_k	e_k	y_k	e_k
0.2	0.4800131	0.470009	0.010003	0.479013	0.000999	0.479913	0.000099
0.4	0.4208291	0.400848	0.019981	0.418841	0.001987	0.420630	0.000198
0.6	0.3288910	0.299518	0.029372	0.325998	0.002892	0.328602	0.000288
0.8	0.2233239	0.187647	0.035676	0.219843	0.003480	0.222976	0.000347
1.0	0.1304818	0.095713	0.034768	0.127063	0.003418	0.130140	0.000341
1.2	0.0660642	0.039488	0.026575	0.063348	0.002715	0.065792	0.000272
1.4	0.0292742	0.013118	0.016155	0.027501	0.001772	0.029095	0.000178
1.6	0.0114138	0.003451	0.007962	0.010444	0.000969	0.011315	0.000098
1.8	0.0039226	0.000698	0.003223	0.003473	0.000449	0.003876	0.000046
2.0	0.0011888	0.000104	0.001084	0.001010	0.000178	0.001170	0.000018

Die numerischen Ergebnisse wurden mit dem folgenden Matlab-Programm erstellt.

```
% explizites Euler-Verfahren mit glm. Schrittweite
 function expl_Euler(T,N,a,y0)
```

```
 y=zeros(length(y0),N+1);
 y(:,1)=y0(:)';
 h = (T-a)/N;
 for k=1:N
   y(:,k+1)=y(:,k) + h*f(a+(k-1)*h,y(:,k));
 end
 plot(linspace(a,T,N+1) ,y,'o-');
 title('Approximation mit explizitem Euler-Verfahren');

% problemabhaengige Funktion f
 function wert = f(t,y)
 wert = -t*sin(pi*y);
```

Für $h = 0,1$ ist die mit dem expliziten Eulerverfahren bestimmte Näherungslösung zum Anfangswertproblem in Abbildung (**??**) graphisch dargestellt.
Eine bedeutend bessere Approximation der gesuchten Funktion $y(t)$ in der Umgebung des Startpunktes (t_0, y_0) kann man mit der Taylorreihe mit Restglied,

$$y(t) = y(t_0) + \frac{(t-t_0)}{1!}y'(t_0) + \frac{(t-t_0)^2}{2!}y''(t_0) + \ldots + + \frac{(t-t_0)^p}{p!}y^{(p)}(t_0) + R_{p+1}$$

erhalten.

Bemerkung 5 (verschiedene Restglieder) *Besitzt y auf dem kompakten Intervall $I := [t_0, t]$ eine stetige Ableitung p-ter Ordnung, während $y^{(p+1)}$ wenigstens im Innern $\overset{o}{I}$ von I vorhanden sei. Dann gibt es nach dem Satz von Taylor mindestens eine Zahl $\xi \in \overset{o}{I}$, so daß*

$$R_{p+1} = \frac{(t-t_0)^{p+1}}{(p+1)!}y^{(p+1)}(\xi) \quad \textit{(Lagrangesches-Restglied)}$$

gelte. Es sei $q \in \mathbb{N}$. Dann gibt es ein $\vartheta \in (0,1)$ mit

$$R_{p+1} = \frac{(t-t_0)^{p+1}}{p!}\frac{(1-\vartheta)^{p+1-q}}{q}y^{(p+1)}(t_0 + \vartheta(t-t_0)) \quad \textit{(Schlömilches-Restglied)}\,.$$

Für $q = 1$ geht das Schlömilche Restglied in das Cauchysche-Restglied über. Ist $y^{(p+1)}$ auf I auch noch stetig, so gilt

$$R_{p+1} = \frac{1}{p!}\int_{t_0}^{t}(t-s)^p y^{(p+1)}(s)\,ds.$$

Durch Vernachlässigen des Restglieds R_{p+1} erhält man mit der Schrittweite $h = t - t_0$ die Rechenvorschrift für das

explizite Euler-Verfahren höherer Ordnung

$$y_{k+1} = y_k + \frac{h}{1!}y'_k + \frac{h^2}{2!}y''_k + \ldots + \frac{h^p}{p!}y_k^{(p)}\,. \tag{1.13}$$

Das Vorgehen erfordert höhere Ableitungen von y, welche sich durch sukzessives Differenzieren von y' nach t gewinnen lassen. Die Terme werden dabei jedoch so kompliziert, daß diese Methode nur in sehr einfachen Fällen angewendet wird. Ein weiterer Punkt ist die vorausgesetzte hohe Regularität, die im allgemeinen nicht gewährleistet werden kann.

Beispiel 6 (zum expliziten Euler-Verfahren höherer Ordnung) Untersuchen wir das explizite Euler-Verfahren höherer Ordnung exemplarisch anhand der Anfangswertaufgabe (1.12). Aus

$$y'(t) = -t\sin(\pi y(t))$$

ergibt sich

$$y''(t) = -\sin(\pi y(t)) + \frac{\pi t^2}{2}\sin(2\pi y(t))$$

und

$$y'''(t) = \frac{3t\pi}{2}\sin(2\pi y(t)) + \frac{t^3\pi^2}{2}(\sin(\pi y(t)) - \sin(3\pi y(t)))$$

Numerische Ergebnisse zu verschiedenen Schrittweiten sind in Tabelle 6 wiedergegeben. Eine

Tab. 1.2: Numerische Ergebnisse des expliziten Euler-Verfahrens höherer Ordnung für das Anfangswertproblem (1.12) zu verschiedenen Schrittweiten h.

		Schrittweite $h = 0.1$		Schrittweite $h = 0.01$		Schrittweite $h = 0.001$	
t_k	$y(t_k)$	y_k	e_k	y_k	e_k	y_k	e_k
0.2	0.48001314649	0.480008428	4.718e-6	0.48001313	7.805e-9	0.48001314	8.130e-12
0.4	0.42082914591	0.420781391	4.775e-5	0.42082908	5.769e-8	0.42082914	5.869e-11
0.6	0.32889107439	0.328769138	1.219e-4	0.32889094	1.260e-7	0.32889107	1.264e-10
0.8	0.22332394257	0.223223946	9.999e-5	0.22332385	8.689e-8	0.22332394	8.576e-11
1.0	0.13048188642	0.130506211	-2.432e-5	0.13048191	-2.630e-8	0.13048188	-2.634e-11
1.2	0.06606421178	0.066142854	-7.864e-5	0.06606427	-6.357e-8	0.06606421	-6.223e-11
1.4	0.02927420018	0.029323183	-4.898e-5	0.02927423	-3.550e-8	0.02927420	-3.440e-11
1.6	0.01141384191	0.011422517	-8.675e-6	0.01141384	-4.277e-9	0.01141384	-3.960e-12
1.8	0.00392269853	0.003912129	1.056e-5	0.00392268	8.922e-9	0.00392269	8.751e-12
2.0	0.00118884958	0.001176472	1.237e-5	0.00118884	9.227e-9	0.00118884	8.952e-12

Matlab-Realisierung zur Berechnung einer Approximation mit dem expliziten Euler-Verfahren höherer Ordnung sieht etwa folgendermaßen aus.

```
% Verfahren mit Taylorreihe und glm. Schrittweite
 function taylor(T,N,a,y0)
 y=zeros(length(y0),N+1);
 y(:,1)=y0(:)';
 h = (T-a)/N;
 for k=1:N
   diff = f(a+(k-1)*h,y(:,k));
   coeff = (h.^[1:size(diff,2)])./cumprod(1:size(diff,2));
   y(:,k+1)=y(:,k) + diff*coeff';
 end
 plot(linspace(a,T,N+1) ,y,'-');
 title('Approximation mit explizitem Euler-Verfahren');

% problemabhaengige Funktion f=y' und hoehere Ableitungen
 function wert = f(t,y)
 wert(:,1) = -t*sin(pi*y);
 wert(:,2) = -sin(pi*y)+pi*t.^2.*sin(2*pi*y)/2;
 wert(:,3) = 3 * pi * t .* sin(2*pi*y) / 2 ...
            - pi^2 * t.^3 .* ( sin(3*pi*y) - sin(pi*y) ) / 2 ;
```

Sowohl bei dem klassischen expliziten Euler-Verfahren als auch bei dem expliziten Euler-Verfahren höherer Ordnung haben wir eine Approximation eines „Differentialoperators“ vorgenommen, der beschreibt, wie sich eine Funktion unter den gegebenen Daten „fortsetzen“ läßt. Betrachten wir zum Vergleich nun eine Methode, in der die gesuchte Funktion approximiert wird.

Bemerkung 7 *Die Idee, zum einen den Differentialoperator zu approximieren oder die Funktion anzunähern, wird uns bei den partiellen Differentialgleichungen z.B. in der Form der Finite-Differenzen-Methode (FDM) im Vergleich zur Finite-Elemente-Methode (FEM) wieder begegnen.*

Verfahren mit polynomialem Ansatz
In einem allgemeinen Näherungspunkt (t_k, y_k) lautet der gewählte Ansatz

$$y(t) = y_k + c_1(t - t_k) + c_2(t - t_k)^2 + \ldots + c_p(t - t_k)^p. \tag{1.14}$$

Dieser wird in die Differentialgleichung $y'(t) = f(t, y(t))$ eingesetzt. Ein Koeffizientenvergleich bezüglich der Potenzen von $(t - t_k) =: h$ liefert $c_1, \ldots, c_p$.

Beispiel 8 Betrachten wir wieder die Anfangswertaufgabe

$$y'(t) = -t \sin(\pi y(t)), \quad y(0) = \frac{1}{2}.$$

Die zu erfüllende Identität lautet für $p = 3$, wobei wir $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} \mp \ldots$, $\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} \mp \ldots$ berücksichtigen.

$$\begin{aligned}
&c_1 + 2c_2h + 3c_3h^2 = -(t_k + h)\sin(\pi(y_k + c_1h + c_2h^2 + c_3h^3)) \\
&= -(t_k + h)[\sin(\pi y_k)\cos(\pi(c_1h + c_2h^2 + c_3h^3) + \cos(\pi y_k)\sin(\pi(c_1h + c_2h^2 + c_3h^3))] \\
&= -(t_k + h)\sin(\pi y_k)[1 - \pi^2(c_1^2h^2 + 2c_1c_2h^3 + (c_2^2 + 2c_1c_3)h^4 + 2c_2c_3h^5 + c_3^2h^6) + \mathcal{O}(h^4)] \\
&\quad - t_k\cos(\pi y_k)[\pi(c_1h + c_2h^2 + c_3h^3) + \mathcal{O}(h^3)] \\
&= -t_k\sin(\pi y_k) - h(c_1t_k\pi\cos(\pi y_k) + \sin(\pi y_k)) + h^2(-t_k(\frac{c_1^2\pi^2}{2}\sin(\pi y_k) + \pi c_2\cos(\pi y_k)) \\
&\quad - \pi c_1\cos(\pi y_k)) + \mathcal{O}(h^3).
\end{aligned} \tag{1.15}$$

Daraus folgt:

$$\begin{aligned}
c_1 &= c_1(t_k, y_k) = -t_k\sin(\pi y_k) \\
c_2 &= c_2(t_k, y_k) = -\frac{c_1}{2}t_k\cos(\pi y_k) \\
c_3 &= c_3(t_k, y_k) = \frac{t_3}{3}[c_1^2\pi^2\sin(\pi y_k) - c_2\cos(\pi y_k).
\end{aligned}$$

Die aufwendige Rechnung der rechten Seite in (1.15) läßt sich mit Maple durch den Befehl

```
taylor(-(t+h)*sin(Pi*(y+C1*h+C2*h^2+C3*h^3)),h=0,3);
```

leicht „verifizieren“.

Die numerischen Ergebnisse zu verschiedenen Schrittweiten und der oben genannten Methode finden sich in der folgenden Tabelle:

Tab. 1.3: Numerische Ergebnisse zum Verfahren mit polynomialem Ansatz (1.14) für das Anfangswertproblem (1.12) zu verschiedenen Schrittweiten h.

		Schrittweite $h = 0.1$		Schrittweite $h = 0.01$		Schrittweite $h = 0.001$	
t	$y(t_k)$	y_k	e_k	y_k	e_k	y_k	e_k
0.2	0.4800131464	0.48000842	4.718e-6	0.480013138	7.805e-9	0.480013146	8.130e-12
0.4	0.4208291459	0.42078139	4.775e-5	0.420829088	5.769e-8	0.420829145	5.869e-11
0.6	0.3288910743	0.32876913	1.219e-4	0.328890948	1.260e-7	0.328891074	1.264e-10
0.8	0.2233239425	0.22322394	9.999e-5	0.223323855	8.689e-8	0.223323942	8.576e-11
1.0	0.1304818864	0.13050621	-2.432e-5	0.130481912	-2.630e-8	0.130481886	-2.634e-11
1.2	0.0660642117	0.06614285	-7.864e-5	0.066064275	-6.357e-8	0.066064211	-6.223e-11
1.4	0.0292742001	0.02932318	-4.898e-5	0.029274235	-3.550e-8	0.029274200	-3.440e-11
1.6	0.0114138419	0.01142251	-8.675e-6	0.011413846	-4.277e-9	0.011413841	-3.960e-12
1.8	0.0039226985	0.00391212	1.056e-5	0.003922689	8.922e-9	0.003922698	8.751e-12
2.0	0.0011888495	0.00117647	1.237e-5	0.001188840	9.227e-9	0.001188849	8.952e-12

Diskussion der Ergebnisse! Wie sieht man Konvergenz der Form $\mathcal{O}(h^3) \approx$ Fehler

Eine Matlab-Realisierung zur Berechnung einer Approximation mit polynomialem Ansatz sieht etwa folgendermaßen aus.

```
% Verfahren mit stkw. poly. Approximation der Ftkn.
% und glm. Schrittweite
 function stkw_approx(T,N,a,y0)
 y=zeros(length(y0),N+1);
 y(:,1)=y0(:)';
 h = (T-a)/N;
 for k=1:N
   cs = coeff(a+(k-1)*h,y(:,k));
   hs = (h.^[1:size(cs,2)]);
   y(:,k+1) = y(:,k) + cs*hs';
 end
 plot(linspace(a,T,N+1) ,y,'>-');
 title('Approximation durch stkw. poly. Funktion');

% problemabh"angige Koeffizientenfunktionen
 function wert = coeff(t,y)
 wert(:,1) = -t*sin(pi*y);
 wert(:,2) = -(wert(:,1).*pi.*t.*cos(pi*y)+sin(pi*y))/2;
 wert(:,3) = (-t*(-wert(:,1).^2.*pi^2.*sin(pi*y)/2+ ...
             pi*wert(:,2).*cos(pi*y))-pi*wert(:,1).*cos(pi*y))/3;
```

Vor- und Nachteile der letzten beiden Verfahren, welche auf der Entwicklung in eine Taylorreihe basieren, sind offensichtlich:

- $\oplus$ Anzahl der Glieder der Taylorreihe läßt sich im Prinzip beliebig erhöhen (Regularität der gesuchten Funktion beachten!)

⊖ großer analytischer Aufwand um höhere Ableitungen in (1.13) oder Koeffizienten $c_1, \ldots, c_p$ zu erstellen.

Um letzt genannten Nachteil zu vermeiden, werden wir später weitere Methoden motivieren und analysieren, welche direkt auf die Differentialgleichung anwendbar sind.

1.3 DISKRETISIERUNGSFEHLER, FEHLERORDNUNG

Die bisher betrachteten Verfahren zur numerischen Lösung von Anfangswertproblemen bezeichnet man als Einschrittverfahren, da vorherige Schritte des numerischen Verfahrens unberücksichtigt bleiben. Für eine Analyse solcher Einschrittverfahren stellen wir einige Aussagen in verallgemeinerter Form über Fehlerabschätzungen zusammen. Betrachten wir dazu die Rechenvorschrift

$$y_{k+1} = y_k + h\phi(t_k, y_k, h) \quad (k = 0, 1, 2, \ldots). \tag{1.16}$$

Dabei beschreibt $\phi(t_k, y_k, h)$ die zugrundeliegende Methode, z.B. im Falle des expliziten Euler-Verfahrens

$$\phi(t_k, y_k, h) = f(t_k, y_k)$$

oder in der Situation des Euler-Verfahrens höherer Ordnung

$$\phi(t_k, y_k, h) = \frac{1}{1!} f(t_k, y_k) + \frac{h}{2!}\frac{\partial}{\partial t} f(t_k, y_k) + \ldots + \frac{h^{p-1}}{p!}\frac{\partial^p}{\partial t^p} f(t_k, y_k).$$

Eine natürliche Voraussetzung an die Funktion ϕ ist sicherlich

$$\lim_{h \to 0} \frac{y_{k+1} - y_k}{h} = y'(t_k) = \phi(t_k, y_k, 0).$$

Definition 9 *Ein Einschrittverfahren* (1.16) *heißt mit einer Differentialgleichung* $y'(t) = f(t, y)$ *konsistent, falls*

$$\phi(t, y, 0) = f(t, y)$$

gilt.

Aufgabe 1 Man zeige, daß die Konsistenzbedingung sowohl für das explizite Euler-Verfahren als auch für die auf der Taylorreihe basierenden Ansätze (1.13) und (1.14) erfüllt ist.

Für die weitere Analyse setzen wir

- eine konstante Schrittweite ($t_k = a + kh, k = 0, 1, 2, \ldots$) und
- exakte Arithmetik (keine Rundungsfehler!)

voraus.
Neben der Konsistenz ist eine wichtige Größe der Diskretisierungsfehler, der lokale wie der globale (totale).

Definition 10 *Unter dem lokalen Diskretisierungsfehler* d_{k+1} *an der Stelle* t_{k+1} *versteht man den Wert*

$$d_{k+1} := y(t_{k+1}) - y(t_k) - h\phi(t_k, y(t_k), h),$$

wobei man beachte, daß auf der rechten Seite die exakten Werte $y(t_k)$ *und* $y(t_{k+1})$ *verwendet werden.*

Der lokale Diskretisierungsfehler beschreibt den Fehler, den die verwendete Rechenvorschrift in einem einzelnen Schritt macht. Da man im allgemeinen mehrere Schritte eines Verfahrens durchführt, um eine Anfangswertaufgabe zu lösen, ist für die Rechenpraxis der globale Fehler von Interesse. Sprich der Fehler, den die Näherung nach mehreren Integrationsschritten gegenüber der exakten Lösung aufweist.

Definition 11 *Als globalen Diskretisierungsfehler e_k an der Stelle t_k bezeichnet man die Differenz*

$$e_k := y(t_k) - y_k.$$

Für weitere Analysen setzen wir voraus, daß die Funktion ϕ in einem geeignet gewählten Bereich B bzgl. der Variablen y eine Lipschitz-Bedingung

$$|\phi(x,y,h) - \phi(x,\overline{y},h)| \le L|y-\overline{y}| \quad ((x,y,\overline{y},h) \in B) \tag{1.17}$$

erfüllt mit $0 < L < \infty$. Desweiteren setzen wir voraus, daß man eine Schranke D angeben kann mit

$$\max_k |d_k| \le D. \tag{1.18}$$

Nach all diesen Definitionen, Voraussetzungen und Annahmen sind wir nun in der Lage, einen Satz über den globalen Diskretisierungsfehler zu formulieren.

Satz 12 *Für den globalen Diskretisierungsfehler e_n an der festen Stelle $t_n = t_0 + nh$ gilt die Abschätzung*

$$|e_n| \le \frac{D}{hL}\mathrm{e}^{nhL} - 1 \le \frac{D}{hL}\mathrm{e}^{nhL}$$

Bemerkung 13 *Für das explizite Euler-Verfahren läßt sich*

$$|e_n| \le \frac{h}{2L} \max_{t_0 \le \xi \le t_n} |y''(\xi)| \mathrm{e}^{L(t_n - t_0)} \tag{1.19}$$

zeigen. Bei konstantem $(t_n - t_0)$ konvergiert folglich der Wert y_n an der festgehaltenen Stelle t_n proportional zu $h \to 0$.

Bevor wir obigen Satz 12 zeigen, beweisen wir folgendes Lemma:

Lemma 14 *Erfüllen die Werte e_k die Ungleichung*

$$|e_{k+1}| \le (1 + C_1)|e_k| + D \quad (k = 0, 1, 2, \ldots), \tag{1.20}$$

so gilt

$$|e_n| \le \frac{(1+C_1)^n - 1}{C_1} D + (1+C_1)^n |e_0| \tag{1.21}$$

$$\le \frac{D}{C_1}(\mathrm{e}^{nC_1} - 1) + \mathrm{e}^{nC_1}|e_0| \tag{1.22}$$

Beweis. Sukzessive Anwendung von (1.20) liefert Ungleichung (1.21)

$$\begin{aligned}
|e_n| &\le (1+C_1)|e_{n-1}| + D \\
&\le (1+C_1)(1+C_1)|e_{n-2}| + ((1+C_1)+1)D \\
&\le (1+C_1)^3|e_{n-3}| + ((1+C_1)^2 + (1+C_1) + 1)D \\
&\le (1+C_1)^n|e_0| + \frac{(1+C_1)^n - 1}{(1+C_1) - 1} D
\end{aligned}$$

Aus der Konvexität von e^t und $\mathrm{e}^t = 1 + t$ für $t = 0$ folgt

$$1 + t \leq \mathrm{e}^t$$

Mit dieser Ungleichung zusammen mit (1.21) ergibt sich (1.22), womit das Lemma bewiesen wäre. □

Beweis von Satz 12. Aus der Definition des lokalen Diskretisierungsfehlers ergibt sich nach Subtraktion der allgemeinen Rechenvorschrift (1.16)

$$e_{k+1} = e_k + h(\phi(t_k, y(t_k), h) - \phi(t_k, y_k, h)) + d_{k+1}.$$

Ausnutzen der Lipschitzbedingung (1.17) und der oberen Schranke (1.18) liefert

$$\begin{aligned} |e_{k+1}| &\leq |e_k| + h|\phi(t_k, y(t_k), h) - \phi(t_k, y_k, h)| + |d_{k+1}| \\ &\leq |e_k| + hL|y(t_k) - y_k| + |d_{k+1}| \\ &\leq (1 + hL)|e_k| + D. \end{aligned}$$

Da der globale Diskretisierungsfehler an der Stelle t_0 gleich null $(e_0 = y(t_0) - y_0 = 0)$ ist, ergibt sich unmittelbar mit Lemma 14 (setze $C = hL$)

$$\begin{aligned} |e_n| &\leq \frac{D}{hL}(\mathrm{e}^{nhL} - 1) + \mathrm{e}^{nhL}|e_0| \\ &= \frac{D}{hL}(\mathrm{e}^{nhL} - 1) \\ &\leq \frac{D}{hL}\mathrm{e}^{nhL} \end{aligned}$$

□

Definition 15 *Ein Einschrittverfahren* (1.16) *zur Lösung des Anfangswertproblems* $y' = f(t, y)$, $y(t_0) = y_0$ *besitzt die Konvergenzordnung* $p \geq 1$, *falls sich der globale Diskretisierungsfehler abschätzen läßt in der Form*

$$\max_{k \in 0,\ldots,n} ||y(t_k) - y_k||_2 \leq Ch^p$$

mit einer von der Schrittweite h *unabhängigen Konstanten* $C > 0$.

Aus Satz 12 erhält man nun leicht

Lemma 16 *Gilt für die lokalen Diskretisierungsfehler* d_k *eines Einschrittverfahren* (1.16) *die Abschätzung*

$$\max_{1 \leq k \leq n} |d_k| \leq D = Ch^{p+1} = \mathcal{O}(h^{p+1}) \quad (p \geq 1),$$

so besitzt es die Konvergenzordnung p.

Zusammenfassung: Im letzten Kapitel wurden einfache Einschrittverfahren vorgestellt. Die Unterschiede der Verfahren wurden durch die Art und Weise von Approximationen motiviert. Elementare Resultate zur Fehlerakkumulation wurden vorgestellt und numerische Beispiele haben die theoretischen Aussagen illustriert. Der Studierende sollte die Ideen hinter den Verfahren nachvollziehen können und den Inhalt der Fehleranalyse verstanden haben.

1.4 Verbesserte Polygonzugmethode, implizites Euler-Verfahren, Prädiktor-Korrektor-Verfahren

Nehmen wir an, mit der expliziten Euler-Methode seien zwei Integrationsschritte von einer gegebenen Stelle t aus gemacht worden. Zum einen mit der Schrittweite h, zum andern mit der Schrittweite $\frac{h}{2}$, dann gilt näherungsweise

$$y_h = y(t) + C_1 h + \mathcal{O}(h^2) \tag{1.23}$$

$$y_{\frac{h}{2}} = y(t) + C_1 \frac{h}{2} + \mathcal{O}(h^2) \tag{1.24}$$

Bemerkung 17 *Motivation ergibt sich aus* (1.19)

Durch Linearkombination von (1.23) und (1.24) ergibt sich ein extrapolierter Wert

$$y^* = 2y_{\frac{h}{2}} - y_h \approx y(t+h) + \mathcal{O}(h^2),$$

dessen Fehler gegenüber der exakten Lösung von zweiter Ordnung ist, was wir später auch noch genauer beweisen werden.
Wie wird diese verbesserte Polygonzugmethode nun numerisch realisiert? Ein Schritt mit der expliziten Methode von Euler mit h von t_k nach t_{k+1} liefert

$$y_{k+1}^{(1)} = y_k + hf(t_k, y_k) \tag{1.25}$$

und ein Doppelschritt mit $\frac{h}{2}$ liefert mit dem Zwischenschritt bei $t_{k+\frac{1}{2}} = t_k + \frac{h}{2}$

$$y_{k+\frac{1}{2}}^{(2)} = y_k + \frac{h}{2} f(t_k, y_k) \tag{1.26}$$

$$y_{k+1}^{(2)} = y_{k+\frac{1}{2}}^{(2)} + \frac{h}{2} f(t_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}^{(2)}) \tag{1.27}$$

und die gesuchte Approximation lautet nun

$$y_{k+1}^* = 2y_{k+1}^{(2)} - y_{k+1}^{(1)} \tag{1.28}$$

Zur Berechnung von y_{k+1}^* benötigen wir somit 2 Auswertungen von f, 4 Multiplikationen und 5 Additionen (=Subtraktionen). Setzen wir jedoch (1.25) - (1.27) geschickt in (1.28) ein, so erhalten wir

$$\begin{aligned} y_{k+1}^* = 2y_{k+1}^{(2)} - y_{k+1}^{(1)} &= 2y_{k+\frac{1}{2}}^{(2)} + hf(t_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}^{(2)}) - y_k - hf(t_k, y_k) \\ &= 2y_k + hf(t_k, y_k) + hf(t_{k+\frac{1}{2}}, y_k + \frac{h}{2} f(t_k, y_k)) - y_k - hf(t_k, y_k) \\ &= y_k + hf(t_k + \frac{h}{2}, y_k + \frac{h}{2} f(t_k, y_k)). \end{aligned}$$

Dies läßt sich algorithmisch schreiben in der Form

$$\begin{aligned} k_1 &= y_k + \frac{h}{2} f(t_k, y_k) \\ k_2 &= f(t_k + \frac{h}{2}, y_k + \frac{h}{2} k_1) \\ y_{k+1} &= y_k + hk_2. \end{aligned}$$

Nun benötigen wir 2 Funktionsauswertungen von f, 4 Multiplikationen und 4 Additionen, desweiteren ist diese Formulierung stabiler.
Dieses Verfahren nennt man verbessertes Polygonzugverfahren von Euler. Man beachte, daß auch dieses Verfahren, trotz des Zwischenschrittes, als Einschrittverfahren bezeichnet wird.
Um die lokale Konvergenzordnung bestimmen zu können, bereiten wir einige Aussagen vor, welche sich durch Differentiation der gegebenen Differentialgleichung ergeben.

$$\begin{aligned}
y' &= f(t,y)\\
y'' &= f_t + f_y y' = f_t + f f_y =: G\\
y''' &= (f_t)' + (f_y y')'\\
&= f_{tt} + f_{ty}y' + f_y(f_t + f f_y) + f(f_{ty} + f_{yy}f)\\
&= \underbrace{(f_{tt} + 2f_{ty}f + f^2 f_{yy})}_{=:H} + (f_t + f f_y) f_y\\
&= H + G f_y
\end{aligned}$$

Damit erhalten wir für den lokalen Diskretisierungsfehler des verbesserten Polygonzugverfahrens nach einer Entwicklung in Taylorreihen

$$\begin{aligned}
d_{k+1} &= y(t_{k+1}) - y(t_k) - hf(t_k + \frac{h}{2}, y(t_k) + \frac{h}{2}f(t_k, y(t_k)))\\
&= y(t_k) + hy'(t_k) + \frac{h^2}{2}y''(t_k) + \frac{h^3}{6}y'''(t_k) + \mathcal{O}(h^4)\\
&\quad - y(t_k)\\
&\quad - h\{f(t_k, y(t_k)) + \frac{h}{2}f_t(t_k, y(t_k)) + \frac{h}{2}f(t_k, y(t_k))f_y(t_k, y(t_k))\\
&\quad + \frac{1}{2}\left(\frac{h}{2}\right)^2 f_{tt} + \left(\frac{h}{2}\right)^2 f f_{ty} + \frac{1}{2}\left(\frac{h}{2}\right)^2 f^2 f_{yy} + \mathcal{O}(h^3)\}\\
&= \frac{1}{6}\{\frac{1}{4}H + G f_y\}h^3 + \mathcal{O}(h^4)\\
&= (\frac{1}{24}(H + G f_y) + \frac{3}{24}G f_y)h^3 + \mathcal{O}(h^4)
\end{aligned}$$

Somit gilt

$$|d_{k+1}| \le \frac{1}{6}(\frac{1}{4}|y'''| + \frac{3}{4}|y''||f_y|)h^3 + \mathcal{O}(h^4).$$

Sind nun y''', y'' und f_y beschränkt, so gilt

$$|d_{k+1}| = \mathcal{O}(h^3).$$

Damit ist die Fehlerordnung der verbesserten Polygonzugmethode 2.
Durch näherungsweise Integration der Differentialgleichung $y' = f(t,y)$ bezüglich der Variablen t über das Intervall $[t_k, t_{k+1}](k = 0, 1, 2, \ldots)$ erhalten wir nun weitere Einschrittmethoden. Betrachten wir dazu

$$y(t_{k+1}) - y(t_k) = \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt.$$

Verwenden wir zur Integration der rechten Seite die Trapezmethode, so erhalten wir

$$y(t_{k+1}) - y(t_k) \approx \frac{h}{2}[f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))].$$

Dies motiviert das implizite Euler-Verfahren, welches man erhält, indem man $y(t_k)$ durch y_k ersetzt. Das implizite Euler-Verfahren lautet nun

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})). \tag{1.29}$$

Da für eine allgemeine nichtlineare Differentialgleichung (1.29) nur eine implizite Gleichung für den approximierten Funktionswert y_{k+1} darstellt, spricht man hier von einer impliziten Integrationsmethode.

Aufgabe 2 Man zeige, daß im Spezialfall einer linearen Differentialgleichung erster Ordnung

$$y'(x) = a(x)y(x) + b(x)$$

mit gegebenen Funktionen $a(x)$ und $b(x)$ das implizite Euler-Verfahren auf eine lineare Gleichung für y_{k+1} führt, aus der sich eine explizite Rekursionsformel für $y_{k+1}(k = 0, 1, 2, \ldots)$ ergibt.

Erfüllt $f(x, y)$ die übliche Lipschitzbedingung bzgl. y mit der Konstanten L und gilt $\frac{1}{2}hL < 1$, so sind die Voraussetzungen für den lokalen Fixpunktsatz von Banach für die Fixpunktiteration

$$y_{k+1}^{(n+1)} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1}^{(n)})) \quad (n = 0, 1, 2, \ldots), \tag{1.30}$$

$y_{k+1}^{(0)} = y_k$ erfüllt und diese ist somit konvergent.
Untersuchen wir nun den lokalen Diskretisierungsfehler, aus dem mit Lemma 16 sofort auch eine Abschätzung für den globalen Diskretisierungsfehler folgt.

$$\begin{aligned}
|d_{k+1}| &= |y(t_{k+1}) - y_{k+1}| \\
&= |y(t_{k+1}) - y(t_k) - \frac{h}{2}(f(t_k, y(t_k)) + f(t_{k+1}, y_{k+1}))| \\
&\leq |y(t_{k+1}) - y(t_k) - \frac{h}{2}(y'(t_k) + y'(t_{k+1}))| + \frac{Lh}{2}|y(t_{k+1}) - y_{k+1}| \\
&\leq |y(t_k) + hy'(t_k) + \frac{h^2}{2}y''(t_k) + \frac{h^3}{6}y'''(t_k) + \mathcal{O}(h^4) \\
&\quad - y(t_k) - \frac{h}{2}y'(t_k) - \frac{h}{2}(y'(t_k) + hy''(t_k) + \frac{h^2}{2}y'''(t_k) + \mathcal{O}(h^3))| \\
&\quad + \frac{Lh}{2}|y(t_{k+1}) - y_{k+1}| \\
&\leq |\frac{h^3}{12}y'''(t_k) + \mathcal{O}(h^4)| + \frac{Lh}{2}|d_{k+1}|.
\end{aligned}$$

Somit gilt

$$(1 - \frac{Lh}{2})|d_{k+1}| \leq |\frac{h^3}{12}y'''(t_k) + \mathcal{O}(h^4)|.$$

Somit ist der lokale Diskretisierungsfehler von der Ordnung 3 und der globale Diskretisierungsfehler von der Ordnung 2.
Wenn das implizite Euler-Verfahren jedoch die gleiche Konvergenzordnung wie das verbesserte Polygonzugverfahren hat, welches eine explizite Methode ist, was spricht dann für das aufwendigere Verfahren, d.h. das implizite Euler-Verfahren. Der entscheidende Begriff ist hier die Stabilität, auf welche wir aber erst später eingehen werden.
Ein weiteres Einschrittverfahren erhält man, indem man die Fixpunktiteration (1.30) nach der zweiten Iteration beendet. Man erhält dann das Verfahren von Heun, welches mit modifizierter Notation wie folgt lautet.

Verfahren von Heun

$$\begin{aligned}
y_{k+1}^{(p)} &= y_k + hf(t_k, y_k) \\
y_{k+1} &= y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1}^{(p)})) \quad .
\end{aligned} \tag{1.31}$$

Man bezeichnet diese Methode als Prädiktor-Korrektor-Methode, da zuerst ein Wert in einem Prädiktor-Schritt „grob“ vorhergesagt wird und dieser dann in einem folgenden Korrektor-Schritt nachkorrigiert wird.

Aufgabe 3 Man zeige, daß die Fehlerordnung des Verfahrens von Heun gleich 2 ist.

Hinweis: Beweis ähnlich wie im Fall der verbesserten Polygonzugmethode.

Algorithmisch schreiben wir das Verfahren von Heun

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + h, y_k + hk_1) \\ y_{k+1} &= y_k + \frac{1}{2}h(k_1 + k_2) \end{aligned}$$

Bemerkung 18 *Hängt $f(t, y)$ nicht von y ab, so ist die Lösung des Anfangswertproblems $y' = f(t, y), y(t_0) = y_0$ durch das Integral $y(t) = y_0 + \int_{t_0}^{t} f(\tau)\, d\tau$ gegeben. Das Verfahren von Heun entspricht dann der Approximation von $y(t)$ mittels Trapezsummen, die verbesserte Polygonzugmethode der Mittelpunktsregel.*

Bemerkung 19 *Sowohl die verbesserte Polygonzugmethode als auch die Methode von Heun sind Beispiele von expliziten zweistufigen Runge-Kutta-Verfahren mit der Fehlerordnung 2, welche wir in verallgemeinerter Form im nächsten Kapitel vorstellen werden.*

1.5 Runge-Kutta-Verfahren

Wir wollen nun die Herleitung von Einschrittverfahren höherer Ordnung beschreiben, welche keine Ableitungen von y' benötigen (d.h. keine Methoden ähnlich zum expliziten Eulerverfahren höherer Ordnung oder zur Methode der Taylorreihe). (Da die Herleitung recht schnell kompliziert wird, beschränken wir uns hier zuerst auf dreistufige Runge-Kutta-Verfahren.)
Das Prinzip der Runge-Kutta-Verfahren ist das folgende:
Die zur Differentialgleichung äquivalente Integralgleichung lautet

$$y(t_{k+1}) - y(t_k) = \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt.$$

Analog zur Gauß-Quadratur approximieren wir das Integral durch eine allgemeine Quadraturformel

$$Q_n^{[t_k, t_{k+1}]}(f) = \sum_{i=1}^{n} w_i f(\xi_i)$$

mit Stützstellen $t_k \le \xi_1 \le \ldots \le \xi_n \le t_{k+1}$ und Gewichten $w_1, \ldots, w_n$ so, daß wir eine möglichst hohe Konvergenzordnung erhalten.
Für $n = 3$ gelangen wir zu dem allgemeinen Ansatz

$$y_{k+1} = y_k + h\{c_1 f(\xi_1, y(\xi_1)) + c_2 f(\xi_2, y(\xi_2)) + c_3 f(\xi_3, y(\xi_3))\}. \tag{1.32}$$

Desweiteren schränken wir die Wahl der Stützstellen bedingt durch die Zielsetzung eines einfachen und expliziten Verfahrens ein auf

$$\xi_1 = t_k, \xi_2 = t_k + a_2 h, \xi_3 = t_k + a_3 h, (0 < a_2, a_3 \le 1). \tag{1.33}$$

Um nun ein explizites Verfahren herzuleiten, verwenden wir die Prädiktor-Methode und ersetzen $y(\xi_1), y(\xi_2), y(\xi_3)$ durch „gute“, explizit berechenbare Näherungen.

Für die erste Stützstelle gilt

$$y(\xi_1) = y_k \tag{1.34}$$

und für die verbleibenden Stützstellen erhalten wir durch Einsetzen von Prädiktorwerten die allgemeinen Gleichungen

$$\begin{aligned} y^{(p)}(\xi_2) &= y_k + hb_{21}f(t_k, y_k), && (1.35)\\ y^{(p)}(\xi_3) &= y_k + hb_{31}f(t_k, y_k) + hb_{32}f(t_k + a_2h, y^{(p)}(\xi_2)). && (1.36) \end{aligned}$$

In algorithmischer Form schreibt sich das Verfahren (1.32) - (1.36)

$$\begin{aligned} k_1 &= f(t_k, y_k)\\ k_2 &= f(t_k + a_2h, y_k + hb_{21}k_1)\\ k_3 &= f(t_k + a_3h, y_k + h(b_{31}k_1 + b_{32}k_2))\\ y_{k+1} &= y_k + h(c_1k_1 + c_2k_2 + c_3k_3). \end{aligned} \tag{1.37}$$

Da die Funktion f pro Integrationsschritt hier dreimal ausgewertet werden muß, spricht man hier von einem dreistufigen Runge-Kutta-Verfahren.

Bemerkung 20 *Häufig findet man eine kürzere Schreibweise, bei der für mehrstufige Einschrittverfahren die Koeffizienten $a_2, a_3, b_{21}, b_{31}, b_{32}$ und c_1, c_2, c_3 in Tabellenform festgehalten werden, d.h. z.B. für Verfahren* (1.37)

0			
a_2	b_{21}		
a_3	b_{31}	b_{32}	
	c_1	c_2	c_3

Wir fordern von unserem gesuchten dreistufigen Runge-Kutta-Verfahren, daß es die spezielle Differentialgleichung $y' = 1$ exakt löst. Damit folgt für die Parameter

$$a_2 = b_{21} \quad \text{und} \quad a_3 = b_{31} + b_{32} \quad .$$

Der lokale Diskretisierungsfehler des Verfahrens (1.37) ist gegeben durch

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h(c_1\overline{k}_1 + c_2\overline{k}_2 + c_3\overline{k}_3), \tag{1.38}$$

wobei $\overline{k}_i$ sich aus k_i dadurch ergeben, daß y_k durch $y(t_k)$ ersetzt wird. Sukzessives Entwickeln der Terme $\overline{k}_1, \overline{k}_2, \overline{k}_3$ an der Stelle t_k liefert

$$\overline{k}_1 = f(t_k, y(t_k)) \tag{1.39}$$

$$\begin{aligned} \overline{k}_2 &= f(t_k + a_2h, y(t_k) + a_2hf(t_k, y(t_k))) && (1.40)\\ &= f + a_2hf_t + a_2hff_y + \frac{1}{2}a_2^2h^2f_{tt} + a_2^2h^2ff_{ty} + \frac{1}{2}a_2^2h^2f^2f_{yy} + \mathcal{O}(h^3)\\ &= f + a_2hF + \frac{1}{2}a_2^2h^2G + \mathcal{O}(h^3) \end{aligned}$$

$$\begin{aligned} \overline{k}_3 &= f(t_k + a_3h, y(t_k) + h(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)) && (1.41)\\ &= f + a_3hf_t + h(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)f_y\\ &\quad + \frac{1}{2}a_3^2h^2f_{tt} + a_3(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)h^2f_{ty} + \frac{1}{2}(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)^2h^2f_{yy} + \mathcal{O}(h^3)\\ &= f + h\{a_3f_t + (b_{31} + b_{32})ff_y\}\\ &\quad + h^2\{a_2b_{32}Ff_y + \frac{1}{2}a_3^2f_{tt} + a_3(b_{31} + b_{32})ff_{ty} + \frac{1}{2}(b_{31} + b_{32})^2f^2f_{yy}\} + \mathcal{O}(h^3)\\ &= f + a_3hF + h^2\{a_2b_{32}Ff_y + \frac{1}{2}a_3^2G\} + \mathcal{O}(h^3) \end{aligned}$$

Einsetzen von (1.39) - (1.41) in (1.38) liefert nun

$$d_{k+1} = hf(1 - c_1 - c_2 - c_3) + h^2 F\{\frac{1}{2} - a_2c_2 - a_3c_3\}$$
$$+ h^3\{Ff_y[\frac{1}{6} - a_2c_3b_{32}] + G[\frac{1}{6} - \frac{1}{2}a_2^2c_2 - \frac{1}{2}a_3^2c_3]\} + \mathcal{O}(h^4)$$

Damit das Verfahren mindestens die Konvergenzordnung 3 hat, müssen die Faktoren vor h, h^2, h^3 verschwinden. D.h. die sechs Parameter c_1, c_2, c_3, a_2, a_3 und b_{32} müssen Lösung des folgenden nichtlinearen Gleichungssystem sein:

$$\begin{aligned} c_1 + c_2 + c_3 &= 1 \\ a_2c_2 + a_3c_3 &= \frac{1}{2} \\ a_2c_3b_{32} &= \frac{1}{6} \\ a_2^2c_2 + a_3^2c_3 &= \frac{1}{3}. \end{aligned}$$

Unter der Einschränkung $a_2 \neq a_3$ und $a_2 \neq \frac{2}{3}$ ergibt sich die zweiparametrige Lösungsmenge

$$\begin{aligned} c_1 &= \frac{6_2a_3 + 2 - 3(a_2 + a_3)}{6a_2a_3} \\ c_2 &= \frac{3a_3 - 2}{6a_2(a_3 - a_2)} \\ c_3 &= \frac{2 - 3a_2}{6a_3(a_3 - a_2)} \\ b_{32} &= \frac{a_3(a_3 - a_2)}{a_2(2 - 3a_2)}. \end{aligned}$$

Kriterien für die Festlegung der Parameter können sein:

- einfache bzw. einprägsame Zahlwerte, die bei Handrechnung auch bzgl. der Rundung günstig sind,
- für spezielle Klassen von Differentialgleichungen soll der Hauptteil von d_{k+1} klein sein,
- simple Methoden der Schrittweitensteuerung sollen daraus herleitbar sein (siehe Kapitel 1.6)

Ein erstes Runge-Kutta-Verfahren dritter Ordnung erhält man nun durch die Wahl $a_2 = \frac{1}{3}$ und $a_3 = \frac{2}{3}$. Das resultierende Verfahren von Heun dritter Ordnung in algorithmischer bzw. tabellarischer Form lautet:

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + \frac{1}{3}h, y_k + \frac{1}{3}hk_1) \\ k_3 &= f(t_k + \frac{2}{3}h, y_k + \frac{2}{3}hk_2) \\ y_{k+1} &= y_k + \frac{h}{4}(k_1 + 3k_3) \end{aligned}$$

0			
$\frac{1}{3}$	$\frac{1}{3}$		
$\frac{2}{3}$	0	$\frac{2}{3}$	
	$\frac{1}{4}$	0	$\frac{3}{4}$

Eine weitere Methode erhält man nun durch die ebenso ausgezeichnete Wahl der Konstanten zu $a_2 = \frac{1}{2}$ und $a_3 = 1$. Dies liefert die Methode von Kutta dritter Ordnung

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + \frac{1}{2}h, y_k + \frac{1}{2}hk_1) \\ k_3 &= f(t_k + h, y_k - hK_1 + 2hk_2) \\ y_{k+1} &= y_k + \frac{h}{6}(k_1 + 4k_2 + k_3) \end{aligned} \tag{1.42}$$

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

Übungsaufgabe: In welchem Zusammenhang steht die Simpsonregel zur Methode (1.42)?

1.6 SCHRITTWEITENSTEUERUNG

Da zur Zahldarstellung in einem Rechner nur endlich viele Ziffern verwendet werden können, ist es offensichtlich, daß h wegen der auftretenden Rundungsfehler nicht beliebig klein gemacht werden kann. Andererseits ist aber auch eine beliebige Erhöhung der Konvergenzordnung p, da dann der Rechenaufwand stark ansteigt. In vielen Fällen ist es aber auch gar nicht notwendig, für alle Schritte ein kleines h bzw. ein großes p zu verwenden, da die Lösung $y(t)$ für verschiedene t unterschiedliches Verhalten zeigt. Das Ziel der adaptiven Schrittweitensteuerung ist es, die Schrittweite der Eigenschaft der Lösung anzupassen. Das Prinzip sieht dabei wie folgt aus. Es wird versucht, den lokalen Diskretisierungsfehler zu schätzen und dementsprechend die Schrittweite für den nächsten Schritt so zu wählen, daß er in einem kleinen Toleranzbereich bleibt. Somit ergibt sich bei geschätzten großen lokalen Diskretisierungsfehlern eine kleine Schrittweite und vice versa. Wie erhält man aber nun die wichtige Information über den lokalen Diskretisierungsfehler?
Man verwendet dazu einfach zwei Verfahren unterschiedlicher Ordnung. Um den Mehraufwand zu minimieren, ist man daran interessiert, Paare von Einschrittverfahren zu finden, bei denen keine oder nur wenige zusätzliche Auswertungen der Funktion f notwendig sind, die „grobe“ Information mit geringem Aufwand als Nebenprodukt des gewählten Verfahrens anfällt.
Die verbesserte Polygonzugmethode und das Verfahren von Kutta erfüllen diese Forderung. Die Näherung des lokalen Diskretisierungsfehler ergibt sich mit

$$d_{k+1}^{(K)} = y(t_{k+1}) - y_{k+1}^{(K)} = \mathcal{O}(h^4) \quad \Rightarrow \quad y(t_{k+1}) = y_{k+1}^{(K)} + \mathcal{O}(h^4)$$

zu

$$d_{k+1}^{(VP)} = y_{k+1}^{(K)} - y_{k+1}^{(VP)} + \mathcal{O}(h^4)$$

und somit

$$d_{k+1}^{(VP)} \approx \frac{h}{6}(k_1 + 4k_2 + k_3) - hk_2 = \frac{h}{6}(k_1 - 2k_2 + k_3).$$

Mit einer zusätzlichen Funktionsauswertung für k_3 liefert der Term $\frac{h}{6}(k_1 - 2k_2 + k_3)$ eine Näherung an den lokalen Diskretisierungsfehler, welche um eine Ordnung größer ist als der lokale Diskretisierungsfehler selbst.
Ein Algorithmus mit Schrittweitensteuerung basierend auf der Approximation des lokalen Diskretisierungsfehlers könnte wie folgt aussehen.

```
function adaptiv_np(tn,t0,y0,h,tol)
%
% Adaptiv_VP(3.6,0,[1.2;0;0;-1.049357509830350],0.1,0.002)
%
 y(:,1)=y0(:);
 t=t0;
 T=t;
 H=[];
 while t<tn
   flag = 1;
   while flag
     k1 = f(t,y(:,end));
     k2 = f(t+h,y(:,end)+h*k1);
     k3 = f(t+h/2,y(:,end)+h*(k1+k2)/4);
     dk = norm(h*(k1-2*k2+k3)/6);
     if dk > 1.2 * tol
       h = 0.9 * h;
     elseif dk < 0.8 *tol
       h = 1.1 * h;
     else
       flag = 0;
     end
   end
   y = [y,y(:,end)+h*k2];
   t=t+h;
   T=[T,t];
   H=[H,h];

   subplot(1,2,1); plot(y(1,:),y(3,:),'r-'); title('Phasendiagramm');
   subplot(1,2,2); plot([T(1:end-1);T(2:end)],[H;H],'g-'); title('Schrittweite');
   drawnow

 end
 subplot(1,2,1)
% plot(T,y,'r-');
 plot(y(1,:),y(3,:),'r-');
 subplot(1,2,2)
 plot([T(1:end-1);T(2:end)],[H;H],'g-');

% problemabhaengige Funktion f
 function wert = f(t,y)
 mu = 1/82.45;
 z1 = ((y(1)+mu)^2+y(3)^2)^(3/2);
 z2 = ((y(1)-1+mu)^2+y(3)^2)^(3/2);
 wert = [y(2); y(1)+2*y(4)-(1-mu)*(y(1)+mu)/z1-mu*(y(1)-1+mu)/z2;
         y(4); y(3)-2*y(2)-(1-mu)*y(3)/z1-mu*y(3)/z2];
```

Beispiel 21 (R. D. Grigorieff) An einem Beispiel aus der Ingenieur-Wissenschaft wollen wir den obigen Algorithmus anwenden. Die folgende Differentialgleichung beschreibt eine periodische

Satellitenbahn im Erde-Mond-System, wobei $\mu = (82.45)^{-1}$ die relative Mondmasse ist:

$$\ddot{u} = u + 2\dot{v} - (1-\mu)\frac{u+v}{[(u+\mu)^2+v^2]^{\frac{3}{2}}} - \mu\frac{u-(1-\mu)}{[(u-1+\mu)^2+v^2]^{\frac{3}{2}}}$$
$$\ddot{v} = v - 2\dot{u} - (1-\mu)\frac{v}{[(u+\mu)^2+v^2]^{\frac{3}{2}}} - \mu\frac{v}{[(u-1+\mu)^2+v^2]^{\frac{3}{2}}}$$

$$\begin{array}{ll} u(0) = 1.2 & u'(0) = 0 \\ v(0) = 0 & v'(0) = -1.049357509830350 \quad . \end{array}$$

Ersetzt man im Algorithmus den ersten „plot“-Befehl durch

```
plot(y(1,:),y(2,:),-);
```

so erhält man für eine verwendete Toleranz $tol = 0.01$ (Anfangsschrittweite $h = 0.1$) die folgende graphische Ausgabe:
GRAPHIK

Die historisch älteste, und deshalb oft als klassische Runge-Kutta-Methode vierter Ordnung bezeichnet, lautet

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

Das Verfahren hat simple Parameterwerte und hat die Eigenschaft, daß für die sukzessive Bestimmung der Werte k_i $(i \geq 2)$ nur die unmittelbar vorangehenden Werte k_{i-1} benötigt werden, und der Algorithmus somit äußerst Speicherplatz sparend ist. Neben diesem Runge-Kutta-Verfahren vierter Ordnung existieren zahlreiche Varianten mit verschiedenen Zielsetzungen. Eine zu oben analoge Schrittweitensteuerung durch Kombination von drei- und vierstufigen Runge-Kutta-Methoden ist nicht möglich, weil die betreffenden Gleichungssysteme keine Lösungen zulassen, so daß k_1, k_2 und k_3 in beiden Verfahren identisch sind.

Die erreichbare Fehlerordnung p einer expliziten m-stufigen Runge-Kutta-Methode ist für $p \leq 9$ gemäß der folgenden Tabelle gegeben.

Tab. 1.4: Maximale Fehlerordnung von expliziten Runge-Kutta-Verfahren

$m =$	1	2	3	4	5	6	7	8	9
$p =$	1	2	3	4	4	5	6	6	7

Von Fehlberg stammt eine geschickte Kombination von Runge-Kutta-Verfahren 4. und 5. Ordnung, in welcher die Runge-Kutta-Methode 4. Ordnung fünf der ohnehin sechs der erforderlichen Auswertungen verwendet.

0					
$\frac{1}{4}$	$\frac{1}{4}$				
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$			
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$		
1	$\frac{8341}{4104}$	$-\frac{32832}{4104}$	$\frac{29440}{4104}$	$-\frac{845}{4104}$	
$\frac{1}{2}$	$-\frac{6080}{20520}$	$\frac{41040}{20520}$	$-\frac{28352}{20520}$	$\frac{9295}{20520}$	$-\frac{5643}{20520}$
y_{k+1}	$\frac{2375}{20520}$	$\frac{11264}{20520}$	$\frac{10985}{20520}$	$-\frac{4104}{20520}$	0
$d_{k+1} \approx$	$\frac{1045}{376200}$	$-\frac{11264}{376200}$	$-\frac{10985}{376200}$	$\frac{7524}{376200}$	$\frac{13680}{376200}$

1.7 Implizite Runge-Kutta-Verfahren

Zur numerischen Lösung von steifen Differentialgleichungen (siehe späteres Kapitel) sind spezielle Methoden erforderlich. Zu diesen gehören die impliziten Runge-Kutta-Verfahren, welche dadurch charakterisiert sind, daß die Steigungen $k_1, k_2, \ldots$ durch ein implizites Gleichungssystem definiert werden. Da die Herleitung von mehrstufigen impliziten Runge-Kutta-Methoden sehr aufwendig ist, wollen wir solche Verfahren hier definieren.
Ein zweistufiges, implizites Runge-Kutta-Verfahren vierter Ordnung ist in algorithmischer Form

$$\begin{aligned} k_1 &= f(t_k + \frac{3-\sqrt{3}}{6}h, y_k + \frac{1}{4}hk_1 + \frac{3-2\sqrt{3}}{12}hk_2) \\ k_2 &= f(t_k + \frac{3+\sqrt{3}}{6}h, y_k + \frac{3+s\sqrt{3}}{12}hk_1 + \frac{1}{4}hk_2) \\ y_{k+1} &= y_k + \frac{h}{2}(k_1 + k_2) \end{aligned} \tag{1.43}$$

oder in tabellarischer Form

$\frac{3-\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{3-s\sqrt{3}}{12}$
$\frac{3+\sqrt{3}}{6}$	$\frac{3+s\sqrt{3}}{12}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

Allgemein gilt, daß ein m-stufiges implizites Runge-Kutta-Verfahren durch geeignete Wahl der $m(m+1)$ freien Parameter die maximal erreichbare Fehlerordnung $2m$ besitzt.
Die Fixpunktiteration zum Lösen von (1.43) ist konvergent für alle Schrittweiten h, die der Bedingung

$$hBL < 1 \quad \text{mit } B := \max_i \sum_j |b_{ij}|$$

genügen und L die Lipschitz-Konstante der Funktion $f(x, y)$ darstellt.

Bemerkung 22 *Die impliziten Runge-Kutta-Verfahren besitzen eine Stabilitätseigenschaft, die bei der Integration von steifen Differentialgleichungssystemen entscheidend ist. Für allgemeine Differentialgleichungssysteme sind die impliziten Runge-Kutta-Methoden trotz ihrer hohen Fehlerordnung wenig attraktiv, weil in jedem Integrationsschritt ein im allgemeinen nichtlineares Gleichungssystem für die m Unbekannten k_i $(1 \leq i \leq m)$ zu lösen ist.*

1.8 Mehrschrittverfahren

Ein Nachteil der Einschrittverfahren ist die hohe Anzahl der Funktionsauswertungen, die für eine hohe Konsistenzordnung und somit für eine hohe Konvergenzordnung nötig sind. Dieser Nachteil tritt bei den folgenden Mehrschrittverfahren nicht auf. Es bleibt im allgemeinen jedoch nur ein deutlicher Vorteil, solange die Schrittweite fest gewählt wird. Für Einschrittverfahren haben wir schon in Kapitel 1.6 gesehen, daß adaptive Schrittweitensteuerung dort recht einfach durchzuführen ist. Dies läßt sich für Mehrschrittverfahren nicht so einfach realisieren.
Es sei $t_j = t_0 + jh$ $(j = 0, 1, 2, \ldots)$.

Ein k-Schrittverfahren zur Lösung des AWP $y'(t) = f(t, y(t))$, $y(t_0) = y_0$ ist ein Verfahren zur Berechnung von Näherungswerten y_{j+k} zu $y(t_{j+k})$ aus $y_j, \ldots, y_{j+k-1}$ durch Lösen der Gleichung

$$\sum_{\ell=0}^{k} \alpha_\ell y_{j+\ell} = h\phi(t_j, y_j, \ldots, y_{j+k}, h, f) \tag{1.44}$$

für $j = 0, 1, 2, \ldots$ mit geeigneten Startwerten $y_0, \ldots, y_{k-1}$. Dabei ist $\frac{1}{h}\sum_{\ell=0}^{k} \alpha_\ell y_{j+\ell}$ als eine Näherung an die Ableitung y' (bisher $\frac{1}{h}(y_{j+1} - y_j)$) und $\phi(t_j, y_j, \ldots, y_{j+k}, h, f)$ als Näherung an $f(t, y(t))$ aufzufassen.
Ein lineares k-Schrittverfahren ist von der Form

$$\sum_{\ell=0}^{k} \alpha_\ell y_{j+\ell} = h\sum_{\ell=0}^{k} \beta_\ell f(t_{j+\ell}, y_{j+\ell}) \quad (j = 0, 1, 2, \ldots) \tag{1.45}$$

mit geeigneten Startwerten $y_0, \ldots, y_{k-1}$. Gilt $\beta_k = 0$ so spricht man von einem expliziten Verfahren, ansonsten von einem impliziten.
Die im allgemeinen nichtlineare Gleichung in y_{j+k} im impliziten linearen k-Schrittverfahren kann aus (1.45) iterativ berechnet werden. Es sei f Lipschitz-stetig in y mit einer Lipschitz-Konstanten L. Dann gilt für die Iteration nach dem Banachschen Fixpunktsatz

$$\begin{aligned} y_{j+k}^{(r+1)} &= h\frac{\beta_k}{\alpha_k} f(t_{j+k}, y_{j+k}^{(r)}) - \frac{1}{\alpha_k}(h\sum_{\ell=0}^{k-1} \beta_\ell f(t_{j+\ell}, y_{j+\ell}) - \sum_{\ell=0}^{k-1} \alpha_\ell y_{j+\ell}) \\ &=: \psi(y_{j+k}^{(r)}), \end{aligned}$$

daß diese konvergent ist, falls $|\frac{\beta_k}{\alpha_k}|hL < 1$ gilt.
Untersuchen wir nun zuerst, ob die Diskretisierung der Differentialgleichung lokal vernünftig ist.

Definition 23 *Der lokale Diskretisierungsfehler des Mehrschrittverfahrens* (1.44) *ist*

$$d_{j+k}(t_j, h, f) = \sum_{\ell=0}^{k} \alpha_\ell y(t_{j+\ell}) - h\phi(t_j, y(t_j), \ldots, y(t_{j+k-1}), y_{j+k}, h, f).$$

Das Mehrschrittverfahren heißt mit einer Differentialgleichung $y' = f(t, y)$ konsistent, falls

$$\lim_{h\to 0} \frac{d_{j+k}(t_j, h, f)}{h} = 0$$

gilt. Es heißt konsistent von der Ordnung p, falls

$$d_{j+k}(t_j, h, f) = \mathcal{O}(h^{p+1})$$

gilt.

Beispiel 24 (eines Mehrschrittverfahrens) Man betrachte die Taylor-Entwicklungen

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

Subtraktion beider Gleichungen liefert

$$y(t+h) - y(t-h) = 2hy'(t) + \mathcal{O}(h^3) = 2hf(t, y(t)) + \mathcal{O}(h^3).$$

Dies motiviert das lineare 2-Schrittverfahren

$$y_{k+1} = 2hf(t_k, y_k) + y_{k-1}, \quad k = 1, 2, 3, \ldots,$$

welches als Quadratur der Mittelpunktsregel entspricht. Die Konsistenzordnung ist 2, was sich aus der Herleitung sofort ergibt.

Im Gegensatz zu Einschrittverfahren impliziert Konsistenz nicht immer Konvergenz, welches wir später untersuchen werden.
Betrachten wir nun zuerst die Entwicklung der wichtigsten Mehrschrittverfahren, die man über numerische Integration herleiten kann, ähnlich wie Runge-Kutta-Verfahren.
Es gilt

$$y(t_{j+k}) - y(t_{j+r-\ell}) = \int_{t_{j+r-\ell}}^{t_{j+k}} f(t, y(t))\, dt. \tag{1.46}$$

Dabei sei $y(t_{j+k})$ der Wert, welcher approximiert werden soll. Es ist nun eine naheliegende Idee, den Integranden f durch ein Interpolationspolynom $P_{r,j}(s) = P_f(s|t_j, \ldots, t_{j+r})$ vom Grade r zu ersetzen, welches an den Stellen $t_j, \ldots, t_{j+r}$ interpoliert.
Dabei betrachtet man das Integral über $(t_{j+r-\ell}, t_{j+k})$, während nur auf $[t_j, t_{j+r}]$ interpoliert wird. Dieses Vorgehen ist graphisch in Abbildung 1.1 dargestellt.

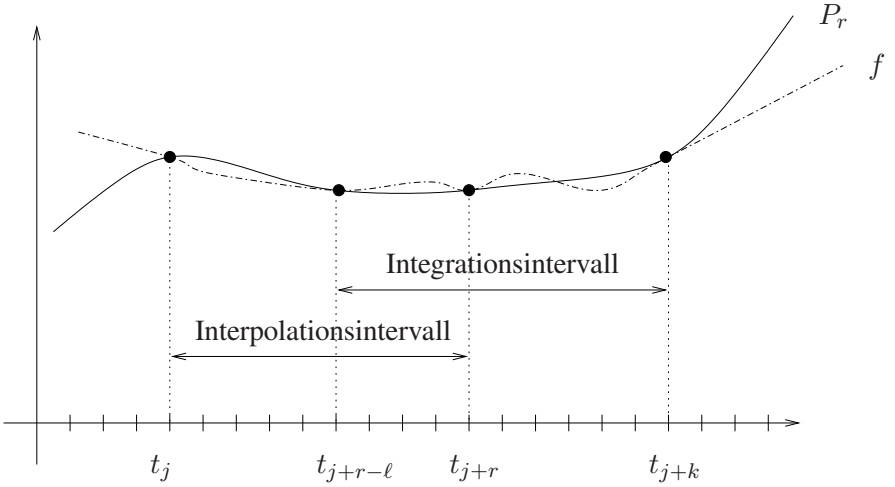

Abb. 1.1: Exemplarische Darstellung von Interpolations- und Integrationsintervall

Die Lagrangesche Darstellungsform für das Interpolationspolynom $P_{r,j}(s)$ sieht folgendermaßen aus:

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

Also erhalten wir, falls $P_{r,j}(s)$ in (1.46) für f eingesetzt wird:

$$y_{j+k} - y_{j+r-\ell} \approx h \sum_{i=0}^{r} f(t_{j+i}, y_{j+i}) \beta_i^{(r,\ell)}, \tag{1.47}$$

wobei

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

ist. Dabei erhält man den letzten Schritt durch Substitution $s = t_j + (r + \tilde{s})h$.
Die Rechenvorschrift lautet somit

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

Übliche Wahlen für r und ℓ:

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

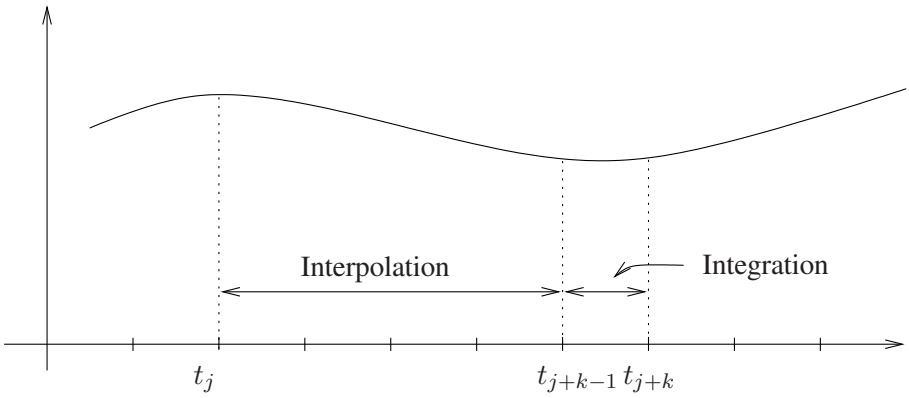

Abb. 1.2: Allgemeine Darstellung des Adams-Bashforth

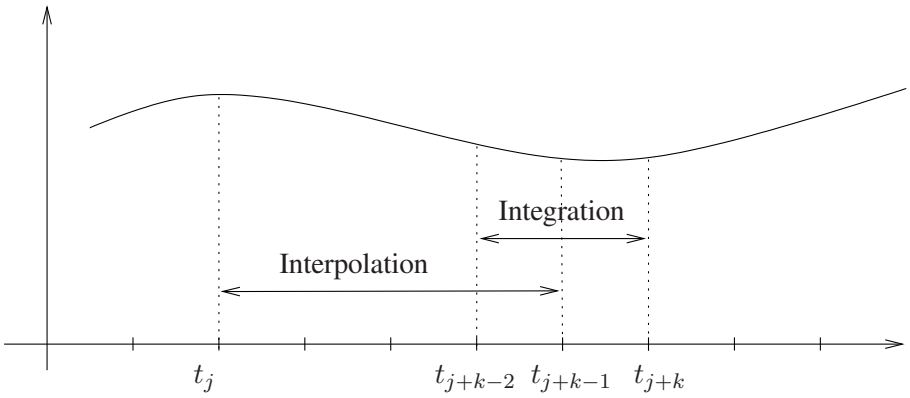

Abb. 1.3: Allgemeine Darstellung des Nyström

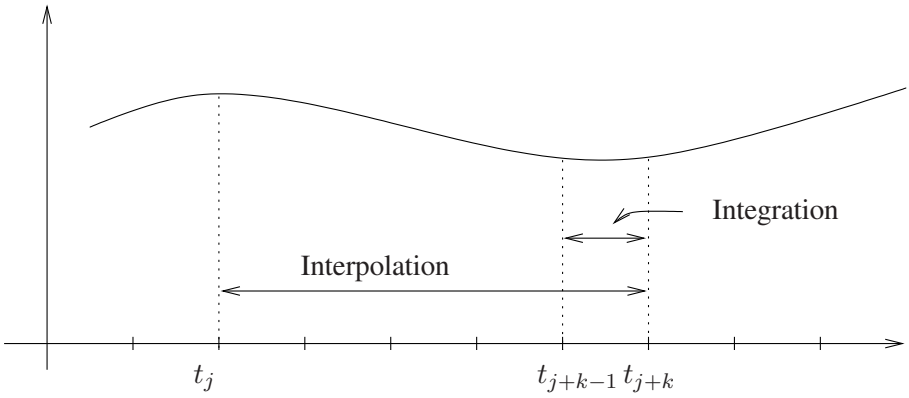

Abb. 1.4: Allgemeine Darstellung des Adams-Moulton

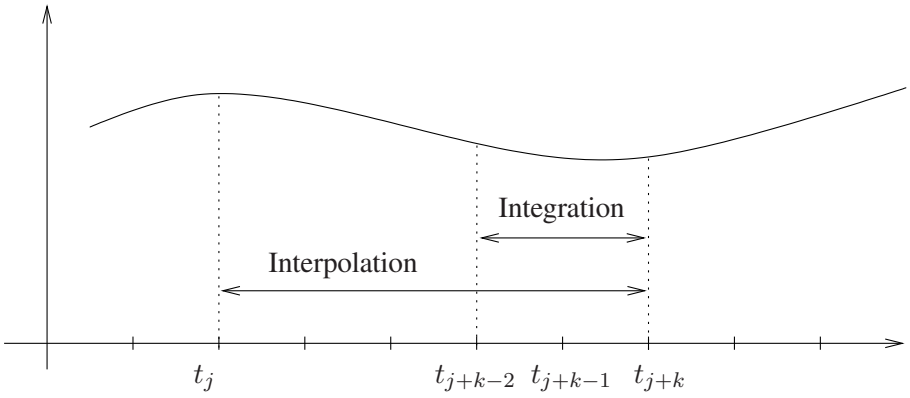

Abb. 1.5: Allgemeine Darstellung des Milne-Simpson

Beispiel 25 Adams-Bashforth:

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

$r = k-1 = 0$: Hiermit ergibt sich: $y_{j+1} = y_j + hf_j$ (explizites Euler-Verfahren)

$r = k-1 = 1$: Hiermit ergibt sich: $y_{j+2} = y_{j+1} + h(\frac{3}{2} f_{j+1} - \frac{1}{2} f_j)$

Nyström:

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

$r = k-1 = 0$: Hiermit ergibt sich: $y_{j+1} = y_{j-1} + 2hf_j$ (Mittelpunktsregel)

Adams-Moulton:

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

$r = k = 0$: Hiermit ergibt sich: $y_j = y_{j-1} + hf_j$ (implizites Euler-Verfahren)

$r = k = 1$: Hiermit ergibt sich: $y_{j+1} = y_j + \frac{h}{2}(f_{j+1} - f_j)$

$r = k = 2$: Hiermit ergibt sich: $y_{j+2} = y_{j+1} + \frac{h}{12}(5f_{j+2} + 8f_{j+1} - f_j)$

Milne-Simpson:

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

Zusätzlich zu der Langrangschen Darstellungsformel für das Interpolationspolynom $P_{r,j}(s)$ gibt es die Newtonsche Interpolationsformel, welche für die paarweise verschiedenen $(r+1)$ Stützstellen $t_j, t_{j+1}, \ldots, t_{j+r}$ lautet

$$P_{r,j}(s) = c_0 + c_1(s-t_j) + c_2(s-t_j)(s-t_{j+1}) + \ldots + c_n(s-t_j)\ldots(s-t_{j+r}). \quad (1.48)$$

Die unbekannten Koeffizienten $c_0, \ldots, c_n$ lassen sich zu den $(r+1)$ Stützpunkten $(t_{j+\ell}, f_{j+\ell})$, $(\ell = 0, \ldots, r)$ prinzipiell aus den Interpolationsbedingungen

$$\begin{array}{lcll} P_{r,j}(t_j) & = & c_0 & = f_j \\ P_{r,j}(t_{j+1}) & = & c_0 + c_1(t_{j+1} - t_j) & = f_{j+1} \\ P_{r,j}(t_{j+2}) & = & c_0 + c_1(t_{j+2} - t_j) + c_2(t_{j+2} - t_j)(t_{j+2} - t_{j+1}) & = f_{j+2} \\ \text{etc.} & & & \end{array} \quad (1.49)$$

sukzessive berechnen, da das lineare Gleichungssystem (1.49) Linksdreiecksgestalt besitzt.

Definition 26 *Die eindeutig durch $(r+1)$ Stützpunkte $(t_{j+\ell}, f_{j+\ell})$ $(\ell = 0, \ldots, r)$ festgelegten Werte c_l in* (1.48) *bezeichnen wir durch*

$$c_\ell := f[t_j, t_{j+1}, \ldots, t_{j+\ell}], \quad (\ell = 0, 1, \ldots, r).$$

Unter Verwendung der Startwerte $f_k = f[t_k]$ $(k = j, j+1, \ldots, j+r)$ lautet das Bildungsgesetz für die Koeffizienten c_ℓ im Prinzip

$$f[t_{i_0}, t_{i_1}, \ldots, t_{i_k}] = \frac{f[t_{i_1}, \ldots, t_{i_k}] - f[t_{i_0}, t_{i_1}, \ldots, t_{i_{k-1}}]}{t_{i_k} - t_{i_0}}. \quad (1.50)$$

Die $f[t_{i_0},\dots,t_{i_k}]$ bezeichnet man als k-te dividierte Differenzen, zugehörig zu den Stellen $t_{i_0},\dots,t_{i_k}$. Die Auswertung der Rekursionsformel (1.50) *erfolgt am zweckmäßigsten im Schema der dividierten Differenzen unter Verwendung der Startwerte $f[t_k]=f_k$, $(k=j,\dots,j+r)$.*

$$\begin{array}{c|c|c|c|c|c} t_j & f[t_j] & & & & \\ & & f[t_j,t_{j+1}] & & & \\ t_{j+1} & f[t_{j+1}] & & & & \\ & & & f[t_j,\dots,t_{j+r-1}] & & \\ \vdots & \vdots & & & f[t_j,\dots,t_{j+r}] & \\ & & \dots & & & \\ t_{j+r-1} & f[t_{j+r-1}] & & f[t_{j+1},\dots,t_{j+r}] & & \\ & & f[t_{j+r-1},t_{j+r}] & & & \\ t_{j+r} & f[t_{j+r}] & & & & \end{array} \tag{1.51}$$

Für äquidistante Stützstellen $t_k=t_0+kh$ $(k=0,1,2,\dots)$ erhält die Newtonsche Darstellungsformel (1.48) eine besonders einprägsame Darstellung, da die Nenner für jede Kolonne in dem Schema (1.51) konstant sind. Es gilt mit $f[t_k]=f_k$

$$\begin{aligned}\nabla^1 f_k &:= f_k - f_{k-1} = hf[t_{k+1},t_k]\\ \nabla^2 f_k &:= \nabla^1 f_k - \nabla^1 f_{k-1} = 2h^2 f[t_{k-2},t_{k-1},t_k]\\ \nabla^\ell f_k &:= \nabla^{\ell-1} f_k - \nabla^{\ell-1} f_{k-1} = \ell! h^\ell f[t_{k-\ell},\dots,t_k]\end{aligned}$$

Die Rechenvorschrift zur Berechnung der c_k lautet dann

```
for k=0:r
  c(k)=f(j+k)
end
for k=1:r
  for j=r:-1:k
    c(j)=c(j)-c(j-1)
  end
end
fak=1
hpower=h
for k=1:r
  fak=fak*h;
  c(k)=c(k)/fak/hpower
  hpower=hpower*h;
end
```

Man erhält damit analog zu (1.47) das Verfahren beschrieben durch

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

wobei

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

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

Was sind nun die Vor- und Nachteile beider Darstellungsformeln?

Vorteile der Lagrange-Form:

- weniger Operationen pro Schritt
- Änderung der Schrittweite einfacher

Vorteile der Newton Form:

- Die numerische Schätzung des Diskretisierungsfehlers kann man wegen der hier einfachen Restglieddarstellung der Polynominterpolation recht leicht erhalten.
- Die Variation der Ordnung ist leicht möglich durch „Anhängen“ von höheren Differenzen.
- Numerisch ist diese Darstellung vorteilhafter, da die Differenzen ∇^i immer kleiner werden und so der Rundungsfehlereinfluß gemindert wird, während bei der Lagrange-Darstellung die verschiedenen Vorzeichen der Koeffizienten problematisch werden können.

Bevor wir uns nun der Konsistenzordnung von Mehrschrittverfahren widmen, betrachten wir vorher noch eine Aussage über den Interpolationsfehler.

Satz 27 *Es sei $f \in C^n[a,b]$, $f^{(n+1)}(x)$ existiere für alle $x \in (a,b)$. Es sei $a \le x_0 < x_1 < \ldots < x_n \le b$. Dann gilt*

$$f(x) - P_n(x,f) = \frac{(x-x_0)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

wobei $\min(x,x_0,\ldots,x_n) < \xi < \max(x,x_0,\ldots,x_n)$ ist.

Beweis. Nach Konstruktion von $P_n(x,f)$ gilt $P_n(x_k,f) = f(x_k)$ $(k = 0,1,\ldots,n)$. Es sei x fest und ungleich $x_0, x_1, \ldots, x_n$. Es sei

$$K(x) = \frac{f(x) - P_n(x,f)}{(x-x_0)\cdots(x-x_n)} \tag{1.52}$$

und wir betrachten die Funktion

$$W(t) = f(t) - P_n(t,f) - (t-x_0)\cdots(t-x_n)K(x). \tag{1.53}$$

Die Funktion $W(t)$ verschwindet an den Stellen $t = x_0, \ldots, t = x_n$ und durch (1.52) auch an der Stelle $t = x$. Nach dem verallgemeinerten Satz von Rolle verschwindet die Funktion $W^{(n+1)}(t)$ an einer Stelle ξ mit $\min(x,x_0,\ldots,x_n) < \xi < \max(x,x_0,\ldots,x_n)$. Jedoch gilt nach (1.53)

$$W^{(n+1)}(t) = f^{(n+1)}(t) - (n+1)!K(x),$$

so daß

$$0 = W^{(n+1)}(\xi) = f^{(n+1)}(\xi) - (n+1)!K(x)$$

und wir somit

$$K(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) \tag{1.54}$$

erhalten. Nach Einsetzen von (1.54) in (1.52) erhalten wir die Behauptung. □

Bemerkung 28 *Eine gute Näherung an $f^{(n+1)}(\xi)$ ist die $(n+1)$ Differenz, d.h.*

$$h^{-(n+1)}\nabla^{n+1} f(x_0) \approx f^{(n+1)}(\xi).$$

Betrachten wir wieder die Anfangswertaufgabe $y'(t) = f(t,y(t))$, $y(t_0) = y_0$. In Analogie zum Einschrittverfahren definieren wir die Begriffe Konsistenz und Konsistenzordnung.

Definition 29 *Wir definieren*

$$\tau(\hat{t},h,f)=\frac{1}{h}\sum_{\ell=0}^{k}\alpha_\ell f(\hat{t}+\ell h)-\phi(\hat{t},y(\hat{t}),\dots,y(\hat{t}+kh),h,f).$$

Das Mehrschrittverfahren (1.44) *heißt konvergent, falls für jedes f, dessen partielle Ableitungen auf $[t_0,t_0+a]\times\mathbb{R}^d$ existieren, stetig und beschränkt sind, gilt*

$$\lim_{h\to 0}\tau(\hat{t},h,f)=0\qquad(\hat{t}\in[t_0,t_0+a]).$$

Es heißt konsistent von der Ordnung p, falls $\tau(\hat{t},h,f)=\mathcal{O}(h^p)$ gilt.

Bemerkung 30 *Für explizite lineare k-Mehrschrittverfahren entspricht $\frac{h}{\alpha_k}\tau(\hat{t},h,f)$ dem lokalen Diskretisierungsfehler in einer kanonischen Verallgemeinerung der Definition 23. Man beachte den Unterschied in ϕ bzgl. „y_k“ und „$y(t_k)$“.*

Die aus der Polynominterpolation hergeleiteten Verfahren lassen sich durch die Parameter (r,ℓ) identifizieren. Mit Hilfe der Restgliedabschätzung der Polynominterpolation aus Satz 27 läßt sich die Konsistenzordnung leicht bestimmen.

Satz 31 *Es sei U eine Umgebung der Anfangswertaufgabe $y'=f(t,y(t)$, $y(t_0)=y_0$. Dann sind k-Schrittverfahren vom Typ (r,ℓ) konsistent für $f\in C^{r+1}(U)$ mit der Ordnung $p=r+1$.*

Bevor wir Satz 31 beweisen noch einige Bemerkungen.

Bemerkungen 32

i) Aus Satz 31 folgt, daß das Adams-Bashforth-Verfahren und das Nyström-Verfahren konsistent sind von der Ordnung $p=k$ und das Adams-Moulton- bzw. das Milne-Simpson-Verfahren konsistent sind von der Ordnung $p=k+1$.

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

Beweis von Satz 31. Anstelle eines beliebigen $\hat{t}$ schreiben wir $\hat{t}=t_j$, $\tau_j=\tau(t_j,h,f)$ und setzen $d=1$. Dann gilt mit Substitution $s=t_{j+r}+hx$:

$$\begin{aligned}
\tau_j&=\frac{1}{h}(y_{j+k}-y_{j+r-\ell})-\sum_{i=0}^{r}\gamma_i^{(r,\ell)}\nabla^i f_{j+r}\\
&=\frac{1}{h}\int_{t_{j+r-\ell}}^{t_{j+k}}f(s,y(s))\,ds-\frac{1}{h}\int_{t_{j+r-\ell}}^{t_{j+k}}P_{r,j}(s)\,ds\\
&=\frac{1}{h}\int_{t_{j+r-\ell}}^{t_{j+k}}(s-t_j)\cdot\ldots\cdot(s-t_{j+r})\frac{D^{r+1}f(\cdot,y(\cdot))(\xi(s))}{(r+1)!}\,ds\\
&=\frac{1}{(r+1)!}\int_{-\ell}^{k-r}h^{r+1}x(x+1)\cdot\ldots\cdot(x+r)D^{r+1}f(\cdot,y(\cdot))(\xi(t_{j+r}+hx))\,dx\\
&=h^{r+1}(-1)^{r+1}\int_{-\ell}^{k-r}\binom{-x}{r+1}D^{r+1}f(\cdot,y(\cdot))(\xi(t_{j+r}+hx))\,dx.
\end{aligned}$$

Somit erhalten wir

$$|\tau_j|\leq h^{r+1}|\gamma_{r+1}^{(r,\ell)}|\sup_{x\in(-\ell,k-r)}|D^{r+1}(f(\cdot,y(\cdot)))(\xi(t_j+r+hx))|.$$

□

Bemerkung 33 *Für genügend kleines h läßt sich der Term*

$$|D^{r+1}(f(\cdot, y(\cdot)))(\xi(t_j + r + hx))|$$

gut durch $\nabla^{r+1} f$ approximieren.

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

$$\frac{1}{h}\sum_{i=1}^{k} \alpha_i y_{j+i} \quad = \quad \phi(t_j, y_j, \ldots, y(t_j + k), h, f) \tag{1.55}$$

$$\overset{\text{speziell}}{=} \quad \sum_{i=0}^{k} \beta_i f_{j+i}. \tag{1.56}$$

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

$$\begin{aligned} \frac{1}{h}\sum_{i=0}^{k} \alpha_i y(\hat{t} + ih) &= \phi(\hat{t}, y(\hat{t}), \ldots, y(\hat{t} + kh), h, f) + \tau(\hat{t}, y(\hat{t}), h, f) \\ \text{speziell} \quad &= \sum_{i=1}^{k} \beta_i y'(\hat{t} + ih) + \tau(\hat{t}, y(\hat{t}), h, f) \end{aligned} \tag{1.57}$$

gerade

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

Damit hier $\tau(\hat{t}, y(\hat{t}), h, f) = \mathcal{O}(h)$ gilt, muß offenbar mindestens gelten

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

Im linearen Fall ist die Konsistenz gesichert, falls die zusätzliche Bedingung gilt

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

Man erhält $\tau(\hat{t}, h, f) = \mathcal{O}(h^2)$, wenn weiter gilt

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

usw. Definiert man

$$\rho(z) := \sum_{i=0}^{k} \alpha_i z^i \quad \text{und} \quad \sigma(z) := \sum_{i=0}^{k} \beta_i z^i$$

so gilt

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

Man bezeichnet $\rho(z)$ als erstes charakteristisches Polynom des k-Schrittverfahrens (**??**) und $\sigma(z)$ als zweites charakteristisches Polynom des linearen k-Schrittverfahrens (**??**).
Mit obigen Überlegungen erhält man folgende Ergebnisse:

Lemma 34 *Sei $y \in C^1([t_0, t_0 + a])$, $\rho(1) = 0$. Es gelte*

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

Dann ist das Verfahren (**??**) *konsistent.*

Beweis folgt direkt aus (1.57). □

Das obige Lemma bedeutet, daß ϕ eine „vernünftige“ Inkrementfunktion ist, z. B. bei dem Einschrittverfahren $y_{j+1} = y_j + h\phi$ ist $\rho(z) = z - 1$, $\rho'(1) = 1$, und wir erhalten wieder $\lim_{h\to\infty} \phi(t, y(t), h, f) = f(t, y(t))$.

Lemma 35 *Das lineare k-Schrittverfahren* (**??**) *ist konsistent genau dann, wenn $\rho(1) = 0$ und $\rho'(1) = \sigma(1)$.*

Beweis folgt aus (1.57) und unseren dortigen Überlegungen. □

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

Satz 36 *Folgende Aussagen sind für das lineare k-Schrittverfahren äquivalent:*

i) Es gilt $\sum_{i=0}^{k}(i^\ell\alpha_i - \ell i^{\ell-1}\beta_i) = 0 \quad (\ell = 0, \ldots, p)$.

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

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

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

Beweis.
„i) $\iff$ ii)“: Nach (1.57) ist

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

Hieraus folgt sofort: i) $\iff$ ii).
„iv) $\iff$ iii)“: Die Lösung zu $y'(t) = y(t)$, $y(0) = 1$ ist $y(t) = e^t$. Für diesen speziellen Fall gilt

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

für $z = e^h$. Dies ist von der Ordnung p genau dann, wenn $\psi(z)$ in 1 eine p-fache Nullstelle hat, d. h. $\psi(z) = h^p\psi^{(p)}(1) + \ldots$
„iii) $\Longleftrightarrow$ ii)“: Betrachte die Entwicklung in „i) $\Longleftrightarrow$ ii)“ speziell für $y = e^t$. Gilt iii), dann muß für alle h gelten

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

Also gilt i).
„i) $\Longleftrightarrow$ iii)“: Wir wissen „i) $\Longleftrightarrow$ ii)“, „ii) $\Longleftrightarrow$ iii)“ trivial. □

Definition 37 *Die Größe*

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

*heißt Fehlerkonstante des linearen k-Schrittverfahrens (**??**) der Ordnung p.*
Ist p minimal mit $c_p \neq 0$, so ist p die genaue Konsistenzordnung.
$\widehat{c_p} = c_p/\sigma(1)$ heißt normierte Fehlerkonstante.

Lemma 38

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

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

Beweis. Man betrachte für α_i, β_i das folgende Gleichungssystem

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 2 & \cdots & k & -1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & 2^2 & \cdots & k^2 & 0 & -2\cdot 1 & -2\cdot 2^1 & \cdots & -2r \\ 0 & 1 & 2^3 & \cdots & k^3 & 0 & -3\cdot 1 & -3\cdot 2^2 & \cdots & -3r^2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2^p & \cdots & k^p & 0 & -p\cdot 1 & -p\cdot 2^{p-1} & \cdots & -p\cdot r^{p-1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{pmatrix} = 0 \qquad (1.58)$$

Nach Satz 36 ist zu zeigen, daß dieses System für $p = k + r$ eindeutig lösbar ist, wenn $\alpha_k = 1$ festgelegt wird. In diesem Fall kann man das System schreiben als

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

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

Betrachten wir das homogene Gleichungssystem

$$A^{\mathsf{T}} \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_p \end{pmatrix} = 0. \tag{1.59}$$

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

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

d. h. das Polynom q vom Grad p hat $k + r + 1 = p + 1$ Nullstellen. Das ist aber nur für $q = 0$ möglich. Also hat (1.59) nur die triviale Lösung. A^{T} bzw. A ist damit nichtsingulär und das Gleichungssystem (1.58) ist daher mit $\alpha_k = 1$ eindeutig lösbar. Damit wäre i) bewiesen.
Beweisen wir nun ii). Sei ρ mit $\rho(1) = 0$ gegeben. Dann sind in (1.58) die Werte $\alpha_0, \dots, \alpha_k$ bekannt, und da $\sum_{i=0}^{k} \alpha_i = 0$ gilt, kann man aus dem System die erste Gleichung streichen und erhält durch Umschreiben das folgende Gleichungssystem für die Koeffizienten $\beta_0, \dots, \beta_r$:

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

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

1.9 Konvergenz von Mehrschrittverfahren

Wir haben gesehen, daß es sehr einfach ist, lineare Mehrschrittverfahren von hoher Konsistenzordnung zu konstruieren. Anders als im Fall der Einschrittverfahren ist es hier aber so, daß die Konsistenz allein noch keine Konvergenz impliziert. Betrachten wir dazu ein Beispiel.

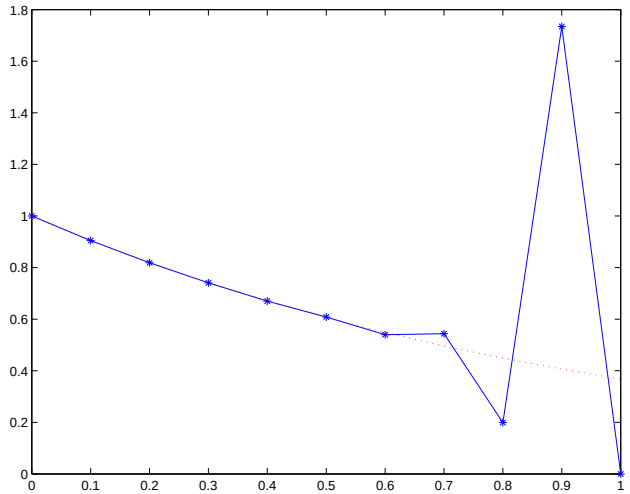

Abb. 1.6: Exakte Lösung und Näherung zu der Lösung von (1.60) mit (1.61) und $h = 0.1$.

Beispiel 39 Zur Lösung der Anfangswertaufgabe

$$y'(t) = -y(t), \quad y(0) = 1 \tag{1.60}$$

betrachten wir das lineare 2-Schrittverfahren

$$y_{j+2} + 4y_{j+1} - 5y_j = h(4f_{j+1} + 2f_j) \tag{1.61}$$

mit exakten Anfangswerten $y_0 = 1$, $y_1 = e^{-h}$. Das Verfahren ist nach Satz 36 von der Ordnung 3. In Abbildung 1.6 ist der exakte Verlauf der Lösungskurve und Approximation, welche mit obigem Verfahren und Schrittweite $h = 0.1$ berechnet wurde, dargestellt. Für $t > 0.8$ ist die Lösung überhaupt nicht mehr akzeptabel. Ein Verkleinern der Schrittweite auf $h = 0.01$ (Rundungsfehler sind hier noch nicht von Belang) liefert die in Abbildung 1.7 dargestellte Lösung. Man beachte dabei die Skalierung der Ordinate.

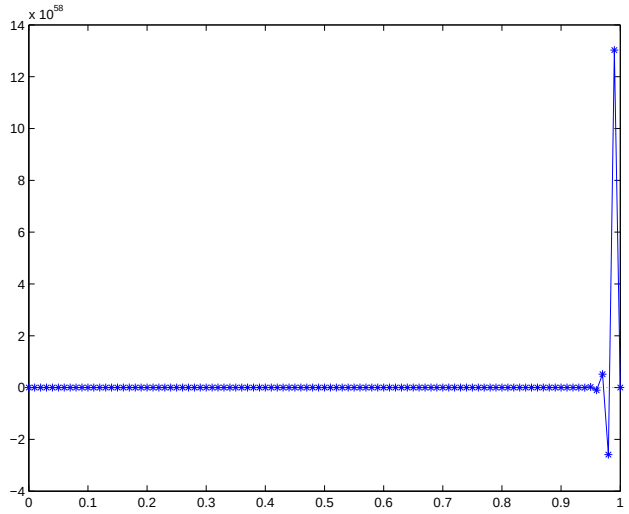

Abb. 1.7: Näherung zu der Lösung von (1.60) mit (1.61) und $h = 0.01$.

Merke: Das Verfahren ist zwar konsistent, liefert aber keine Konvergenz!

Um dieses Verhalten von Mehrschrittverfahren allgemeiner zu studieren, machen wir einen kleinen Exkurs in den Bereich Differenzengleichungen. Eine Gleichung der Form

$$\alpha_k y_{j+k} + \alpha_{k-1} y_{j+k-1} + \ldots + \alpha_0 y_j = 0, \qquad \alpha_k \neq 0 \tag{1.62}$$

für alle $j \in \mathbb{N}$, mit $\alpha_\nu, y_\nu \in \mathbb{C}$ oder $\mathbb{R}$, nennt man lineare homogene Differenzengleichung k-ter Ordnung mit konstanten Koeffizienten. Eine Folge $(y_\ell)_{\ell \in \mathbb{N}}$ heißt Lösung, wenn die y_ν der Gleichung (1.62) genügen. Die Lösungen bilden einen Vektorraum. Man sieht sofort:

- Es gibt zu jedem Satz von Startwerten $y_0, \ldots, y_{k-1}$ genau eine Lösung.
- Es gibt k linear unabhängige Vektoren $y^{(1)}, y^{(2)}, \ldots, y^{(k)}$, $y^{(i)} = (y_0^{(i)}, \ldots, y_k^{(i)})^\mathsf{T} \in \mathbb{C}^{k+1}$, die senkrecht auf $(\alpha_0, \ldots, \alpha_k)^\mathsf{T}$ stehen.
- Zu jedem $y^{(i)}$ gibt es eine Lösungsfolge, die gerade mit $y^{(i)}$ beginnt. Der Lösungsraum hat die Dimension k.

Wir wollen in diesem Abschnitt die allgemeine Lösung von (1.62) darstellen. Zur Motivation betrachten wir einmal für den Fall $\alpha_k = 1$, den wir o.B.d.A. annehmen können, folgende aus (1.62)

resultierende Beziehung

$$Y_{j+1} := \begin{pmatrix} y_{j+1} \\ \vdots \\ y_{j+k} \end{pmatrix} = \begin{pmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ \alpha_0 & -\alpha_1 & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{pmatrix} \begin{pmatrix} y_j \\ \vdots \\ y_{j+k-1} \end{pmatrix} =: AY_j.$$

Damit erhält man

$$Y_n = AY_{n-1} = AAY_{n-2} = \ldots = A^n Y_0.$$

Man betrachte nun das zur Differenzengleichung (1.62) gehörige Polynom

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

Es seien $z_1, \ldots, z_\ell$ die paarweise verschiedenen Nullstellen mit den Vielfachheiten $\sigma_1, \ldots, \sigma_\ell$ mit $\sum_{i=0}^{\ell} \sigma_i = k$, d. h.

$$\rho(z) = \alpha_k (z - z_1)^{\sigma_1} \cdot \ldots \cdot (z - z_\ell)^{\sigma_\ell}.$$

Um im weiteren die Schreibweise zu vereinfachen, führen wir folgende Notationen ein:
Wir definieren

$$y_j^{(p,q)} := j(j-1) \cdot \ldots \cdot (j-q+1) z_p^{j-q} \qquad (j \in \mathbb{N})$$

mit $j(j-1) \cdot \ldots \cdot (j-q+1) = 1$, falls $q = 0$, und die Folge

$$y^{(p,q)} = (y_j^{(p,q)})_{j \in \mathbb{N}}.$$

Es gilt also

$$y^{(p,0)} = (1, z_p, z_p^2, z_p^3, \ldots)$$

und für $q \neq 0$

$$y^{(p,q)} = (0, 0, \ldots, 0, q!, (q+1)q \cdot \ldots \cdot 2z_p, (q+2) \cdot \ldots \cdot 3z_p^2, \ldots) \tag{1.63}$$

mit q führenden Nullen.
Mit dieser Notation läßt sich die allgemeine Lösungs von (1.62) wie folgt schreiben.

Lemma 40 *Jede Lösung von* (1.62) *hat die Form*

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

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

Beweis. Im ersten Schritt zeigen wir, daß $y^{(p,q)}$ eine Lösung von (1.62) ist. Da z_p eine σ_p-fache Nullstelle von $\rho(z)$ ist, gilt für $0 \leq q \leq \sigma_p - 1$

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

Im zweiten Schritt zeigen wir, daß die k Folgen $y^{(p,q)}$ für $p = 1, \ldots, \ell$ und $0 \le q \le \sigma_p - 1$ linear unabhängig sind. Für die ersten k Elemente von $y^{(p,q)}$ aus (1.63) finden wir

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

Betrachten wir das Hermite-Interpolationspolynom mit

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

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

1.10 Konvergenz- und Stabilitätsbedingungen für Mehrschrittverfahren

In diesem Abschnitt wollen wir die Konvergenz der Mehrschrittverfahren analysieren. Dabei erlaubt uns die Theorie, den Einfluß von Rundungsfehlern gleich mitzuberücksichtigen. Statt y_j berechnen wir

$$\tilde{y}_j = y(t_j) + \varepsilon_j, \quad j = 0, 1, \ldots, k-1.$$

Für die weiteren $\tilde{y}_j$ erhalten wir dann aus dem k-Schrittverfahren

$$\sum_{r=0}^{k} \alpha_r \tilde{y}_{j+r} = h\phi(t_j, \tilde{y}_j, \ldots, \tilde{y}_{j+k}, h, f) + h\varepsilon_{j+k} \quad . \tag{1.64}$$

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

$$\varepsilon(\hat{t}, h_j) = \varepsilon_j, \quad \tilde{y}_{h_j}(\hat{t}) = \tilde{y}_j(\hat{t}, \varepsilon). \tag{1.65}$$

Definition 41 *Das durch $\rho(z) = \sum_{r=0}^{k} \alpha_r z^r$ und die Funktion ϕ gegebene Mehrschrittverfahren heißt konvergent, falls für alle $t \in [t_0, t_0 + a]$, für alle $f \in C^1(U)$ (U Umgebung der Lösung) und für alle $\varepsilon(t, h)$, für die es ein $\psi(h) \ge 0$ gibt mit $\psi(h) \to 0$ für $h \to 0$, so daß*

$$|\varepsilon(t, h)| \le \psi(h)$$

gilt, so daß

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

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

Satz 42 (Nullstellenbedingung) *Falls $\phi(t, y_j, \ldots, y_{j+k}, h, 0) = 0$ für alle $(t, y_j, \ldots, y_{j+k})$ (gilt z.B. für lineare Mehrschrittverfahren) und falls zu $\rho(z) = \sum_{r=0}^{k} \alpha_r z^r$ und ϕ gehörende Mehrschrittverfahren konvergent ist, so gilt die Stabilitätsbedingung (auch Nullstellenbedingung):*
Ist z eine Nullstelle von ρ, so ist $|z| < 1$ oder $|z| = 1$ und in diesem Fall ist die Nullstelle einfach.

Bemerkung 43 *In der Literatur findet man für den Begriff der Nullstellenbedingung auch die Bezeichnung Wurzelbedingung.*

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

$$\tilde{y}_i = \varepsilon_i, \quad i = 0, \ldots, k-1$$
$$\sum_{\nu=0}^{k} \alpha_\nu \tilde{y}_{j+\nu} = h_j \varepsilon_{j+k} = 0$$

bestimmt. Nach Lemma 40 löst $(1, z, z^2, z^3, \ldots)$ obige Differenzengleichung. Wegen $\tilde{y}_j = \varepsilon_i = h_j z^i$ für $i = 0, \ldots, k-1$ folgt daher für $\tilde{y}_\nu = h_j z^\nu$. Die Konvergenz des Mehrschrittverfahrens ergibt nun

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

Das gilt aber nur, wenn $|z|^j$ beschränkt ist, d.h. falls $|z| \leq 1$ gilt. Angenommen nun gilt, daß z mit $|z| = 1$ eine mehrfache Nullstelle wäre. Wähle dann $\varepsilon_0 = 0, \varepsilon_\nu = \nu h_j z^{\nu-1}$ für $\nu = 1, \ldots, k-1$ so folgt, daß

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

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

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

Dies geht nur für $|z| < 1$. Also haben wir einen Widerspruch und z mit $|z| = 1$ kann nur eine einfache Nullstelle sein. □

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

Lemma 44 *Sei $\rho(M) = \max_{\lambda EW von M} |\lambda|$ der Spektralradius von M für $M \in \mathbb{C}^{n\times n}$.*

(i) Zu jedem $M \in \mathbb{C}^{n\times n}$ und jedem $\varepsilon > 0$ gibt es eine Vektornorm $||\cdot||$, so daß für die zugehörige Matrixnorm $||M|| = \sup_{||x||=1} ||Mx||$ gilt

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

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

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

Beweis. Sei $TMT^{-1} = J \in \mathbb{C}^{n\times n}$ die Jordansche Normalform von M mit

$$J = diag(C_k(\lambda_i)), \quad C_k(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{k\times k}.$$

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

$$D_\varepsilon^{-1} J D_\varepsilon = diag(C_k^{(\varepsilon)}(\lambda_i))$$

mit

$$C_k^{(\varepsilon)}(\lambda_i) = \begin{pmatrix} \lambda_i & \varepsilon & & \\ & \lambda_i & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{k\times k}.$$

Dann gilt für $||\cdot||_\infty$

$$||D_\varepsilon^{-1} T M T^{-1} D_\varepsilon||_\infty = ||D_\varepsilon^{-1} J D_\varepsilon||_\infty \leq \rho(M) + \varepsilon.$$

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

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

Beweisen wir nun (ii). Dazu ordnen wir die Eigenwerte von M absteigend an, d.h. $|\lambda_1| = |\lambda_2| = \ldots = |\lambda_s| > |\lambda_{s+1}| \geq \ldots \geq |\lambda_n|$ und setzen $\varepsilon = \rho(M) - |\lambda_{s+1}|$. Dann ist

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

und mit D_ε und T wie eben ist dann

$$||D_\varepsilon^{-1} T M T^{-1} D_\varepsilon||_\infty = ||D_\varepsilon^{-1} J D_\varepsilon||_\infty = \rho(M).$$

Die gesuchte Vektornorm ist daher dieselbe wie im Beweis von (i). □

Satz 45 *Das durch $\rho(z)$ und ϕ gegebene Mehrschrittverfahren sei konsistent und erfülle die Lipschitzbedingung:*
Für alle $f \in C^1(Ix\mathbb{R})$ gelte, daß es $h_0 > 0$ und ein $L < \infty$ gebe mit

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

für alle $h < h_0$ und alle $\eta_{j+\nu}, \bar{\eta}_{j+\nu} \in \mathbb{R}$.
Dann gilt, daß das Verfahren genau dann konvergent ist, wenn es die Stabilitätsbedingung erfüllt.

Beweis. Daß aus der Konvergenz die Stabilitätsbedingung folgt, haben wir schon in Satz 42 bewiesen. Sei nun die Stabilitätsbedingung erfüllt. Es seien $\tilde{y}_j, \varepsilon_j$ wie in (1.64) und (1.65) definiert. Dann gilt für den Fehler $e_j = \tilde{y}_j - y(t_j)$

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

und

$$\begin{aligned} \sum_{\nu=0}^{k} \alpha_\nu e_{j+\nu} &= h[\phi(t_j, \tilde{y}_j, \ldots, \tilde{y}_{j+k}, h, f) - \phi(t_j, y(t_j), \ldots, y(t_{j+k}), h, f)] \\ &\quad + h(\varepsilon_{j+k} - \tau_{j+k}) \\ &=: c_{j+k} \end{aligned} \tag{1.66}$$

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

$$|c_{j+k}| \leq hL \sum_{\nu=0}^{k} |e_{j+\nu}| + h(|\varepsilon_{j+k} + |\tau_{j+k}|). \tag{1.67}$$

Wir definieren

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

Dabei sei o.B.d.A. $\alpha_k = 1$. Dann ist (1.66) äquivalent zu

$$E_{j+1} = ME_j + C_{j+k}. \tag{1.68}$$

Nun ist $(-1)^k \rho(z)$ das charakteristische Polynom von M. Aus der Konsistenz des Mehrschrittverfahrens folgt $\rho(1) = 0$. Daher ist 1 Eigenwert von M. Wegen der Stabilitätsbedingung sind alle Eigenwerte vom Betrag ≤ 1 und die vom Betrag 1 sind einfach. Insbesondere gilt also $\rho(M) = 1$. Nach Lemma 44 gibt es eine Vektornorm $||\cdot||_\sim$ mit $||M||_\sim = \sup_{||x||=1} ||Mx||_\sim = \rho(M) = 1$, also

$$||Mx||_\sim \leq ||x||_\sim \quad \text{für alle } x \in \mathbb{C}^k. \tag{1.69}$$

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

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

für alle $x = (x_1, \ldots, x_k)^T \in \mathbb{C}^k$. Damit folgt

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

und

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

Aus (1.67) folgt

$$|c_{j+k}| \leq hLR(||E_j||_\sim + ||E_{j+1}||_\sim) + h(|\varepsilon_{j+k}| + |\tau_{j+k}|), \tag{1.70}$$

da $\sum_{\nu=0}^{k} |e_{j+\nu}| \leq \sum_{\nu=0}^{k-1} |e_{j+\nu}| + \sum_{\nu=1}^{k} |e_{j+\nu}|$.
Anwendung von (1.68) und (1.69) liefert mit (1.70)

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

Subtraktion liefert dann

$$(1 - hLR^2)||E_{j+1}||_\sim \leq (1 - hLR^2)||E_j||_\sim + hR(|\varepsilon_{j+k}| + |\tau_{j+k}|),$$

welches äquivalent ist zu

$$||E_{j+1}||_\sim \leq \frac{1 + hLR^2}{1 - hLR^2}||E_j||_\sim + \frac{hR}{1 - hLR^2}(|\varepsilon_{j+k}| + |\tau_{j+k}|). \tag{1.71}$$

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

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

Es ergibt sich aus (1.71) somit für $h \leq \frac{1}{2LR^2}$

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

Anwendung von Lemma 44 ergibt

$$\begin{aligned} ||E_{j+1}||_\sim &= (||E_0||_\sim + \sum_{\nu=0}^{j} 2hR(|\varepsilon_{j+k}| + |\tau_{j+k}|))\mathrm{e}^{\sum_{\nu=0}^{j} 4hLR^2} \\ &\leq (||E_0||_\sim + 2hR(j+1)(\varepsilon + \tau))\mathrm{e}^{(j+1)4hLR^2} \end{aligned}$$

mit $\varepsilon = \max_j |\varepsilon_{j+k}|$, $\tau = \max_j |\tau_{j+k}|$. D.h. man hat für $t \neq t_0$, $h = \frac{t-t_0}{j+k}$, $h < \frac{1}{2LR^2}$

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

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

Korollar 46 *Ist das Mehrschrittverfahren ein konsistentes Verfahren p-ter Ordnung und sind die Anfangsfehler ε_i ebenfalls von der Ordnung p, dann gilt auch für den globalen Diskretisierungsfehler*

$$|y_{h_n}(t,\varepsilon) - y(t)| = \mathcal{O}(h^p)$$

für alle $h_n = \frac{t-t_0}{n}$ mit n groß genug.

Definition 47 *Genügt $\rho(z)$ der Stabilitätsbedingung (Satz 42), so heißt das Verfahren null-stabil. Diese Stabilität hängt nur von ρ, nicht von ϕ ab.*

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

Satz 48 (Dahlquist) *Ein lineares k-Schrittverfahren erfülle die Stabilitätsbedingung und besitze für $f \in C^p(U)$ die Ordnung p. Dann ist $p \leq k+2$.*

Beweis siehe Grigorieff, Band 2. □

1.11 Prädiktor-Korrektor-Verfahren

Bisher haben wir explizite und implizite Mehrschrittverfahren unterschieden. Dabei ist bei glatten Problemen der Vorteil der expliziten Mehrschrittverfahren bzgl. ihrer Komplexität oder des geringeren Rechenaufwandes offensichtlich. Zur Lösung der impliziten Verfahren wurden Fixpunktiterationen verwendet. Im allgemeinen iteriert man jedoch nicht bis zur Konvergenz (genau betrachtet ist es durch die Zahlendarstellung im allgemeinen auch nicht möglich), sondern nur eine feste Anzahl von Schritten. Bei Prädiktor-Korrektor-Verfahren meist in der Größenordnung der Konsistenz oder geringer. Um einen guten Startwert, den Prädiktor, für die Fixpunktiteration zu erhalten, wählt man ein explizites Verfahren.
Im weiteren präzisieren wir nun, welche Konsistenzordnung Prädiktor-Korrektor-Verfahren haben.
Gegeben sei ein explizites Verfahren (Prädiktor)

$$\sum_{i=0}^{k} \alpha_i^* y_{j+i} = h \sum_{i=0}^{k-1} \beta_i^* f(t_{j+i}, y_{j+i}), \quad \alpha_k^* = 1$$

und ein implizites Verfahren (Korrektor)

$$\sum_{i=0}^{k} \alpha_i y_{j+i} = h \sum_{i=0}^{k} \beta_i f(t_{j+i}, y_{j+i}), \quad \alpha_k = 1.$$

Hiermit definieren wir das folgende

Prädiktor-Korrektor-Verfahren

1. Berechnung der Näherung $y_{j+k}^{(P)}$.
   Gegeben seien $y_j, \ldots, y_{j+k-1}$, so berechne

$$y_{j+k}^{(P)} = -\sum_{i=0}^{k-1} \alpha_i^* y_{j+i} + h \sum_{i=0}^{k-1} \beta_i^* f(t_{j+i}, y_{j+i}) \tag{1.73}$$

   (Anwendung des Prädiktors)

2. Für $m = 1, 2, \ldots, m_0$ iteriere gemäß

$$y_{j+k}^{(m)} = -\sum_{i=0}^{k-1} \alpha_i y_{j+i} + h \sum_{i=0}^{k-1} \beta_i f(t_{j+i}, y_{j+i}) + h\beta_k f(t_{j+k}, y_{j+k}^{(m-1)}) \tag{1.74}$$

   mit $y_{j+k}^{(0)} = y_{j+k}^{(P)}$. (Anwendung des Korrektors)

Sei m_0 die Anzahl der Korrektorschritte.
Man unterscheidet zwei Varianten von Prädiktor-Korrektor-Verfahren:

i) Nach Anwendung des Korrektors wird $f(t_{j+k}, y_{j+k}^{(m_0)})$ noch ausgewertet. Dieser f-Wert wird dann bei der Anwendung des Prädiktors für die nächste Stelle t_{j+k+1} verwendet. Dieses Verfahren wird mit $P(EC)^{m_0}E$ bezeichnet. Dabei steht P für Prädiktor, E für Auswerten (Evaluate) von f und C für Korrektor.

ii) Nach Anwendung des Korrektors wird mit $f(t_{j+k}, y_{j+k}^{(m_0-1)})$ weitergerechnet. Damit wird eine Auswertung von f gespart. Dieses Vorgehen wird mit $P(EC)^{m_0}$ bezeichnet.

Satz 49 (Konsistenzordnung von Prädiktor-Korrektor-Verfahren) *Ist f hinreichend oft differenzierbar, genügt es insbesondere einer Lipschitzbedingung bzgl. y, dann ist die Ordnung des $P(EC)^{m_0}E$ bzw. $P(EC)^{m_0}$ Verfahrens $\min(p, m_0 + p^*)$, wobei p^* und p die Konsistenzordnungen des Prädiktor- bzw. Korrektorverfahrens sind.*

Bemerkung 50 *Falls $m_0 + p^* > p$, so ist die Fehlerkonstante die des Korrektors. Es kann also ein Prädiktor mit kleinerer Ordnung gewählt werden und mittels genügender Iteration die Ordnung des Korrektors mit dessen Fehlerkonstante erreicht werden.*

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

Adams-Bashforth-Moulton-Verfahren (3,3) (1.75)

$$\begin{aligned} y_{k+1}^{(P)} &= y_k + \tfrac{h}{12}(23f_k - 16f_{k-1} + 5f_{k-2}) && \text{Adams-Bashforth } \mathcal{O}(h^3) \\ y_{k+1} &= y_k + \tfrac{h}{12}(9f(t_{k+1}, y_{k+1}^{(P)}) + 19f_k - 5f_{k-1} + f_{k-2}) && \text{Adams-Moulton } \mathcal{O}(h^4) \end{aligned}$$

Satz 49 liefert, daß diese Prädiktor-Korrektor-Methode die Konsistenzordnung 4 hat. Die Größe des lokalen Diskretisierungsfehlers hängt jedoch von den Hauptanteilen der beiden verwendeten Methoden ab. Dabei kann im speziellen der Hauptanteil der Adams-Bashforth der dominierende Anteil sein.
Diese Situation kann verbessert werden, falls als Prädiktor eine Adams-Bashforth-Methode mit gleicher Ordnung wie die Korrektorformel verwendet wird.

Wir kombinieren die explizite k-Schrittmethode von Adams-Bashforth als Prädiktor mit der impliziten 3-Schrittmethode von Adams-Moulton.

Adams-Bashforth-Moulton-Verfahren (4,3) (1.76)

$$y_{k+1}^{(P)} = y_k + \frac{h}{24}(55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}) \qquad \text{Adams-Bashforth } \mathcal{O}(h^4)$$

$$y_{k+1} = y_k + \frac{h}{24}(9f(t_{k+1}, y_{k+1}^{(P)}) + 19f_k - 5f_{k-1} + f_{k-2}) \qquad \text{Adams-Moulton } \mathcal{O}(h^4)$$

Bei solchen Kombinationen von Prädiktor-Korrektor-Verfahren ist stets der Koeffizient des Hauptanteils der Korrektorformel maßgebend. Der lokale Diskretisierungsfehler der Methode (1.76) ist daher kleiner als derjenige in (1.75).

Beispiel 51 Wir behandeln die Anfangswertaufgabe $y' = te^{t-y}$, $y(1) = 0$ mit den beiden A-B-M-Methoden (1.75) und (1.76) und wählen die Schrittweiten $h = 0.1, 0.01, 0.001$. Um die beiden Verfahren unter gleichen Bedingungen vergleichen zu können, wurden in beiden Fällen die drei Startwerte y_1, y_2, y_3 mit der klassischen Runge-Kutta-Methode berechnet. In Abbildung 1.8 sind die Fehler $e = \max_{0\le j\le n} |y_j - y(t_j)|$ in Relation zur Schrittweite für beide Verfahren dargestellt. Beide Graphen haben näherungsweise in der doppeltlogarithmischen Skalierung die Steigung -4, was einer Konvergenzordnung 4 entspricht. Es wird auch deutlich, daß der globale Fehler für Methode (1.76) kleiner ist als derjenige für (1.75), was unsere Motivation bestätigt.

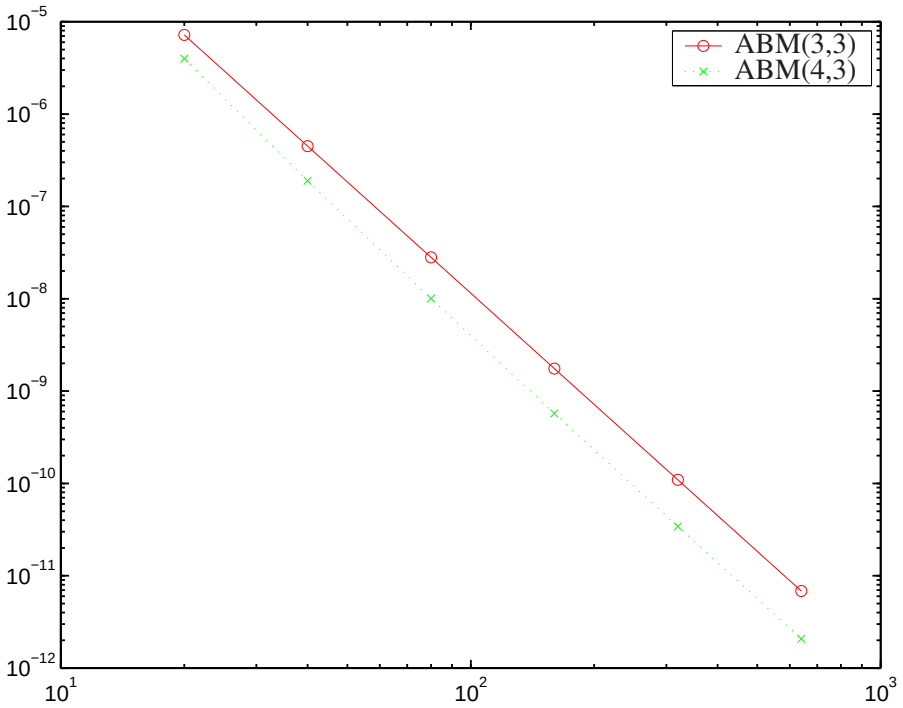

Abb. 1.8: Vergleich von ABM(3,3) und ABM(4,3) für verschiedene Schrittweiten

```
function yN = abm2(t0,y0,tN,N)
% ABM - Praediktor-Korrektor-Verfahren
% fuer AWA y' = f(t,y), y(t0) = y0 mit N Schritten
y = zeros(length(y0),N+1);
y(:,1)=y0(:);
fj = zeros(length(y0),N+1);
fj(:,1) = f(t0,y0);
h = (tN-t0)/N;
% Startrechnung: klassischer Runge-Kutta 4. Ordnung
```

```
for j = 1:3
  k1 = f(t0+(j-1)*h,y(:,j));
  k2 = f(t0+(j-1)*h+h/2,y(:,j) + h/2*k1);
  k3 = f(t0+(j-1)*h+h/2,y(:,j) + h/2*k2);
  k4 = f(t0+j*h,y(:,j) + h*k3);
  y(:,j+1) = y(:,j) + h*(k1 + 2*k2 + 2*k3 + k4)/6;
  fj(:,j+1) = f(t0+j*h,y(:,j+1));
end
% Praediktor-Korrektor-Schritte
for j = 4:N
  yp=y(:,j)+h*(-9*fj(:,j-3)+37*fj(:,j-2)-59*fj(:,j-1)+55*fj(:,j))/24;
  yc=y(:,j)+h*(fj(:,j-2)- 5*fj(:,j-1)+19*fj(:,j)+9*f(t0+j*h,yp))/24;
  y(:,j+1) = yc;
  fj(:,j+1) = f(t0+j*h,y(:,j+1));
end
yN = y(:,N+1);
```

1.12 Stabilitätsbegriffe, Stabilitätsbereiche

Bisher wurden Analysen gemacht, welche die Situation von gegen Null gehender Schrittweiten wiedergeben. Jede Approximation wird jedoch mit Schrittweiten $h \neq 0$ bestimmt, von der man im allgemeinen nicht weiß, ob sie genügend klein ist, daß die asymptotische Theorie das Verhalten der Lösungen qualitativ richtig beschreibt. Statt an Aussagen für genügend kleines h ist man in diesem Abschnitt primär an der Beschreibung des Verhaltens der Näherungslösung für die gerade gewählte Schrittweite interessiert.

Wir werden nur einige wichtige Stabilitätskonzepte in diesem Abschnitt diskutieren. Zum einen betrachten wir stabile Verfahren, bei denen die absoluten Fehler $y_j - y(t_j)$ mit wachsendem j kleiner werden. Dies ist die sogenannte absolute Stabilität. Zum anderen bezeichnen wir ein Verfahren als stabil, wenn die Fehler nicht stärker wachsen als die Lösung. Dies ist die relative Stabilität.

Zuvor betrachten wir noch eine spezielle Frage zur Stabilität. Wie hängt die Lösung $y(t)$ der Anfangswertaufgabe von Störungen in den Anfangsdaten ab? Satz 3 liefert uns die Eigenschaft

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

d.h. daß die Lösung stetig von den Anfangsdaten abhängt.

Wie ein folgendes Beispiel zeigen wird, ist dies jedoch nur eine lokale Eigenschaft. Wir betrachten die Differentialgleichung

$$y' = 10(y - \frac{t^2}{1+t^2}) + \frac{2t}{(1+t^2)^2}. \tag{1.77}$$

Für den Anfangswert $y(0) = 0$ ergibt sich die exakte Lösung $y(t) = \frac{t^2}{1+t^2}$ und für $\tilde{y}(0) = -\varepsilon$ ergibt sich $\tilde{y}(t) = \frac{t^2}{1+t^2} - \varepsilon \mathrm{e}^{10t}$ als Lösung. Es ist offensichtlich, daß für kein $\varepsilon \neq 0$ y und $\tilde{y}$ asymptotisch das gleiche Verhalten haben. Dies ist aber ein Problem der Anfangswertaufgabe und unabhängig vom Approximationsverfahren. Dieses Phänomen bezeichnet man häufig als inhärente Instabilität.

Die inhärente Instabilität kann man nur so in den Griff bekommen, daß man mit Methoden hoher Fehlerordnung und mit hoher Rechengenauigkeit arbeitet, um sowohl die Diskretisierungs- als auch die Rundungsfehler genügend klein zu halten.

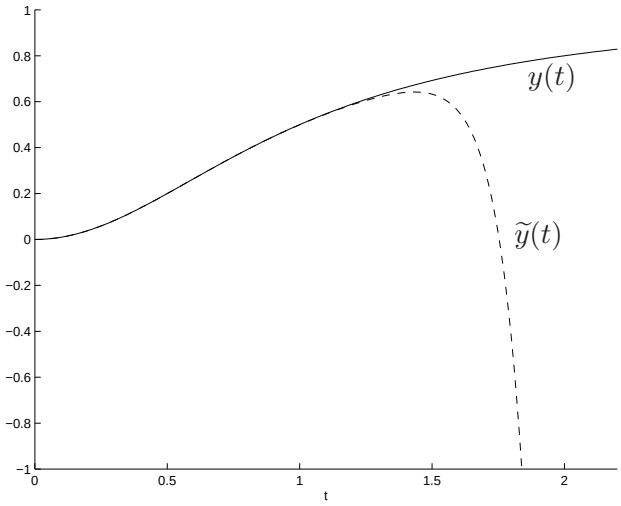

Abb. 1.9: Exakte Lösung und mittels des klassischen Runge-Kutta-Verfahrens zu der Lösung von (1.77) und $h = 0.01$ bestimmte Näherung.

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

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

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

$$\sum_{r=0}^{k} \alpha_r \tilde{y}_{j+r} = h \sum_{i=0}^{k} \beta_r f_{j+r} + h\varepsilon_{j+k}. \tag{1.78}$$

Es sei wieder $e_j = y(t_j) - y_j$ für $j = 0, 1, 2, \ldots$. Man beachte, daß für $j = 0, \ldots, k-1$ dies gerade die Startfehler sind, die aus dem Einschrittverfahren kommen. Da gilt

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

(mit $\tau_{j+k} = \tau(t_{j+k}, y, f)$), folgt mit (1.78)

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

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

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

Damit erhält man die Differenzengleichung

$$\sum_{r=0}^{k}(\alpha_r - h\beta_r\lambda)e_{j+r} = \phi. \tag{1.79}$$

Eine spezielle Lösung von (1.79) ist

$$\hat{e}_j = -\frac{1}{h\lambda\sum_{r=0}^{k}\beta_r}\phi$$

für alle j, denn $\sum_{r=0}^{k}\alpha_r = 0$.
Dann ist die allgemeine Lösung von (1.79)

$$e_j = \eta_j - \frac{1}{h\lambda\sum_{r=0}^{k}\beta_r}\phi, \tag{1.80}$$

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

$$\sum_{r=0}^{k}(\alpha_r - h\lambda\beta_r)\eta_{j+r} = 0 \tag{1.81}$$

sind. Die η_j bestimmen das Wachstumsverhalten der Fehler e_j. Daher untersuchen wir jetzt die Lösungen von (1.81).

Bemerkungen 52

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

ii) Der Fall $\lambda > 0$ ist im allgemeinen nicht sehr interessant, da die in den Natur- und Ingenieurwissenschaften durch Differentialgleichungen beschriebenen Vorgänge in der Regel exponentiell abklingende Komponenten aufweisen. Dies ist für $\lambda < 0$ der Fall.

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

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

und man hat n unabhängige Differentialgleichungen vom obigen Typ. Untersuchungen für den eindimensionalen Fall geben also auch Auskunft über das Verhalten von Systemen. Dabei entsprechen oszillierende, exponentiell abklingende Komponenten von Systemen von Differentialgleichungen komplexen Werten von λ. Dabei ist im allgemeinen wieder der Fall $Re(\lambda) < 0$ von größerem Interesse.

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

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

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

Beispiel 53 Betrachte für $y' = \lambda y$, $y(0) = 1$ das Verfahren $y_{j+2} = y_j + 2h\lambda y_{j+1}$ (Mittelpunktsregel). Das ist ein null-stabiles Verfahren der Ordnung 2, denn $\rho(z) = z^2 - 1 = (z+1)(z-1)$ und mit $\alpha_0 = -1$, $\alpha_1 = 0$, $\alpha_2 = 1$, $\beta_0 = 0$, $\beta_1 = 2$, $\beta_2 = 0$ folgt

$$\begin{aligned} \ell :=0 \qquad & \sum \alpha_i = \rho(1) = 0, \\ \ell :=1 \quad & \alpha_1 + 2\alpha_2 - \beta_0 - \beta_1 - \beta_2 = 0, \\ \ell :=2 \qquad & \alpha_1 + 4\alpha_2 - 2\beta_1 - 4\beta_2 = 0. \end{aligned}$$

Die y_j sind Lösungen der Differenzengleichung

$$y_{j+2} - 2h\lambda y_{j+1} - y_j = 0.$$

Mit den exakten Anfangswerten $y_0 = 1$, $y_1 = \mathrm{e}^{\lambda h}$ erhalten wir mit Lemma 40

$$y_j = c_1 z_1^j + c_2 z_2^j,$$

wobei $z_{1,2} = h\lambda \pm \sqrt{1 + h^2\lambda^2}$ die Nullstellen von $\pi(z) = z^2 - 2h\lambda z - 1$ sind und c_1, c_2 aus $y_0 = 1$, $y_1 = \mathrm{e}^{\lambda h}$ bestimmt werden:

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

Falls man wieder z_1, z_2, c_1 und c_2 nach h entwickelt, so erhält man für $y_h(t)$, der Näherung an der Stelle t mit $h = t/j$

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

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

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

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

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

Nach (1.80), (1.81) und Lemma 40 ist

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

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

Dieses Beispiel motiviert die folgende Definition.

Definition 54 *Ein lineares k-Schrittverfahren heißt stark stabil, falls 1 einfache Nullstelle von ρ ist und alle $(k-1)$ weiteren Nullstellen von ρ vom Betrag kleiner 1 sind. Es heißt schwach stabil, falls die Stabilitätsbedingung gilt und mehrere Nullstellen den Betrag 1 haben.*

Beispiel 55 Die Nyström- und die Milne-Simpson-Verfahren ($\rho(z) = z^k - z^{k-2}$) sind also schwach stabil.

Beachte: Die Begriffe stark und schwach stabil sind problemunabhängig, es kann jedoch sein, daß sie erst für sehr kleine h greifen. Es steht die Idee dahinter, daß die Wurzeln $z_j(h\lambda)$ von $\pi(z, h\lambda)$ für $h \to 0$ gegen die Wurzeln von $\rho(z)$ streben. Insbesondere heißt dann die Wurzel $z_1(h\lambda)$ von $\pi(z, h\lambda)$, für die $\lim_{h\to 0} z_1(h\lambda) = 1$ gilt, Hauptwurzel von π.
Die Konvergenz läßt sich für stark stabile Verfahren genauer beschreiben. Falls das Verfahren die Ordnung p hat, so gilt für die Hauptwurzel $z_1(h\lambda)$ von π

$$z_1(h\lambda) = \mathrm{e}^{h\lambda} + \mathcal{O}(h^{p+1}). \tag{1.82}$$

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

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

Also folgt

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

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

$$\pi(\mathrm{e}^{\lambda h}, h\lambda) = (\alpha_k - h\lambda\beta_k)(\mathrm{e}^{h\lambda} - z_1(h\lambda)) \cdot \ldots \cdot (\mathrm{e}^{h\lambda} - z_k(h\lambda)). \tag{1.83}$$

Wegen $\mathrm{e}^{h\lambda} \overset{h\to 0}{\to} 1$ und starker Stabilität gilt

$$\mathrm{e}^{h\lambda} - z_j(h\lambda) \overset{h\to 0}{\not\to} 1 \quad \text{für } j > 1.$$

D.h.

$$|(\underbrace{\alpha_k}_{\neq 0} - h\lambda\beta_k)(\mathrm{e}^{h\lambda} - z_2(h\lambda)) \cdot \ldots \cdot (\mathrm{e}^{h\lambda} - z_k(h\lambda))| > M$$

für kleine h. Daher folgt aus (1.83)

$$\mathrm{e}^{h\lambda} - z_1(h\lambda) = \mathcal{O}(h^{p+1})$$

und damit ergibt sich (1.82). Für genügend kleines h ist wegen der starken Stabilität

$$|z_1(h\lambda)| > |z_\ell(h\lambda)| \quad (\ell = 2, \ldots, k).$$

Das Wachstum der Fehler richtet sich also nach dem Wachstum von $|z_1(h\lambda)|^j$. Nach (1.82) gilt nun für $\lambda < 0$ und kleine $h > 0$ die Abschätzung $|z_1(h\lambda)| < 1$. In diesem Fall haben wir für den Fehler e_j sogar eine Fehlerdämpfung, da

$$|z_i(h\lambda)|^j \overset{j\to\infty}{\to} 0 \quad (i \in \{1, \ldots, k\}).$$

Beachte: Dies sind die absoluten Fehler und nicht die relativen Fehler. Die Dämpfung der absoluten Fehler folgt damit nur aus dem Abklingen der Lösung $\mathrm{e}^{h\lambda}$.

Definition 56 *Ein Mehrschrittverfahren heißt absolut stabil für $h\lambda$, falls $|z_j(h\lambda)| < 1$ für alle $j = 1, \ldots, k$, andernfalls heißt es absolut instabil. Der Bereich*

$$D_\lambda = \{h\lambda : |z_i(h\lambda)| < 1 (i = 1, \ldots, k)\}$$

heißt Bereich absoluter Stabilität.

Bemerkung 57 *Wegen* (1.82) *ist jedes konsistente, null-stabile Verfahren für alle hinreichend kleinen $\lambda h > 0$ absolut instabil. In diesem Fall wächst der Fehler, aber auch die Lösung. Der Begriff der absoluten Stabilität ist ein Konzept für $h\lambda$ negativ, d.h. für abklingende Lösungen $\mathrm{e}^{\lambda t}$ mit $\lambda < 0$. Solche Anteile sind in vielen Differentialgleichungen zumindest lokal (in der Linearisierung $y' = Ay + f$) enthalten.*

Es stellt sich nun die Frage, ob es immer einen Bereich absoluter Stabilität gibt. Dazu schauen wir uns ein Beispiel an.

Beispiel 58 Betrachten wir wieder (wie in Beispiel 53) die Mittelpunktsregel

$$y_{j+2} = y_j + 2h\lambda y_{j+1}$$

mit $\pi(z) = z^2 - 2h\lambda z - 1$. Die Nullstellen von π sind gegeben durch

$$z_1 = h\lambda + \sqrt{1 + h^2\lambda^2}$$
$$z_2 = h\lambda - \sqrt{1 + h^2\lambda^2}.$$

Für $h\lambda < 0$ folgt $|z_2| > 1$, für $h\lambda > 0$ folgt $|z_1| > 1$. Also kann nie $|z_1| < 1$ und $|z_2| < 1$ gelten. Es gibt daher keinen Bereich der absoluten Stabilität. Dasselbe folgt für das Milne-Simpson-Verfahren

$$\begin{aligned} y_{j+2} &= y_j + \frac{h}{3}(f_{j+2} + 4f_{j+1} + f_j) \\ &= y_j + h\lambda(\frac{1}{3}y_{j+2} + \frac{4}{3}y_{j+1} + \frac{1}{3}y_j) \end{aligned}$$

mit $\pi(z) = (1 - \frac{h\lambda}{3})z^2 - \frac{4}{3}h\lambda z - (1 + \frac{1}{3}h\lambda)$ und den Nullstellen

$$z_{1,2} = \frac{2h\lambda \pm \sqrt{3(3 + h^2\lambda^2)}}{3 - h\lambda}.$$

Bemerkung 59

i) Man kann zeigen, daß „optimale“ Verfahren, das sind solche mit $p = k + 2$, keine Bereiche absoluter Stabilität besitzen.

ii) Alle stark stabilen Verfahren haben Bereiche absoluter Stabilität, da alle parasitären Nullstellen von $\rho(z)$ betraglich < 1 sind und $\lim_{h\to 0} \pi(z, h\lambda) = \rho(z)$ für alle z gilt.

iii) Falls alle parasitären Nullstellen im Ursprung liegen, erwartet man einen großen Bereich absoluter Stabilität, z.B. beim Adams-Moulton-Verfahren $y_{j+k} - y_{j+k-1} = h\sum \ldots$, denn $\rho(z) = z^k - z^{k-1} = z^{k-1}(z - 1)$. Bei impliziten Verfahren sind diese Bereiche zum Teil erheblich (bis zu 10 mal) größer als bei expliziten Verfahren (siehe Grigorieff).

Beispiel 60 Bereiche absoluter Stabilität beim impliziten und expliziten Euler-Verfahren

i) implizites Euler-Verfahren

$$\begin{aligned} y_{j+1} &= y_j + hf(t_{j+1}, y_{j+1}) \\ &= y_j + h\lambda y_{j+1} \quad (\text{für } f(t, y) = \lambda y) \\ \pi(z) &= (1 - h\lambda)z - 1 \\ z(h\lambda) &= \frac{1}{1 - h\lambda} \end{aligned}$$

Der Bereich absoluter Stabilität ist daher gegeben durch

$$D_\lambda = \{h\lambda : |1 - h\lambda| > 1\}.$$

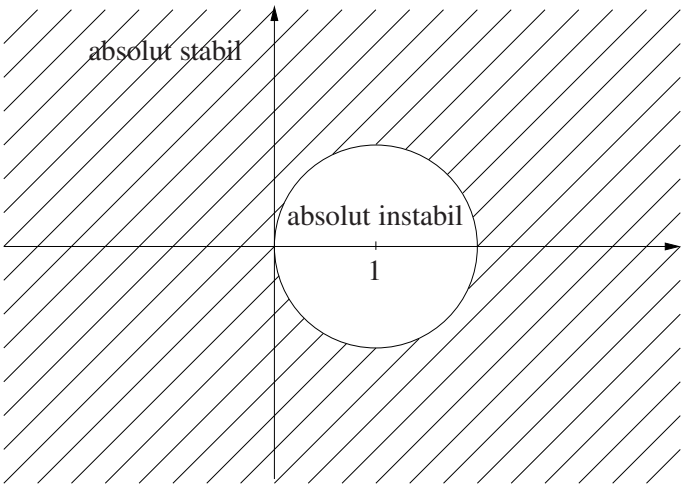

Abb. 1.10: Stabilitätsbereiche beim impliziten Euler-Verfahren

ii) explizites Euler-Verfahren

$$\begin{aligned} y_{j+1} &= y_j + h\lambda y_j \\ \pi(z) &= z - (1 + h\lambda) \\ z(h\lambda) &= 1 + h\lambda \end{aligned}$$

Der Bereich absoluter Stabilität ist daher gegeben durch

$$D_\lambda = \{h\lambda : |1 + h\lambda| < 1\}.$$

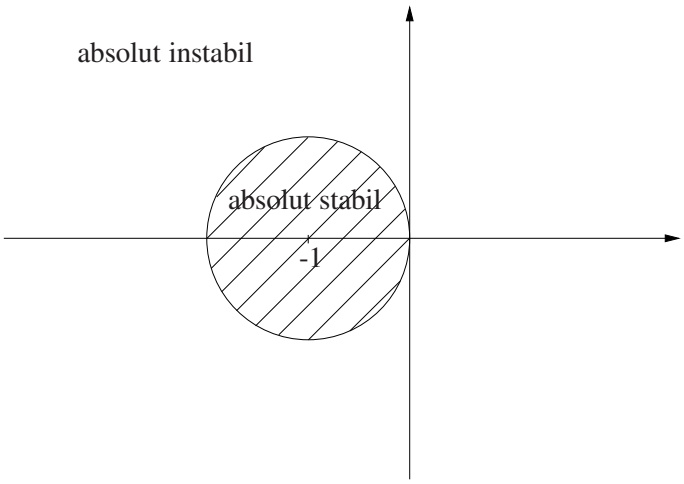

Abb. 1.11: Stabilitätsbereiche beim expliziten Euler-Verfahren

Daß der Bereich der absoluten Stabilität für den impliziten Fall positive Werte $h\lambda$ enthält, ist kein Widerspruch zu Bemerkung 57. Dort wurde ausgesagt: Jedes konsistente null-stabile Verfahren ist für alle hinreichend kleinen $h\lambda$ absolut instabil.

Bisher war unsere Blickrichtung auf $\mathrm{e}^{\lambda t}$ mit $Re(\lambda)$ negativ ausgerichtet. Der Parameter λ stand dabei für die Linearisierung $\partial f/\partial y$. Im allgemeinen ist $\partial f/\partial y$ nicht konstant und man erhält λ als geeignete Schranke für $\partial f/\partial y$ oder als typischen Wert, der über einem Teil des Integrationsbereichs stimmt. Falls λ verändert werden muß, so muß sich das maximal erlaubte h ändern, so daß $h\lambda$ noch im Bereich absoluter Stabilität liegt.

Betrachten wir nun $\lambda > 0$. In diesem Fall sagen (**??**) und Bemerkung **??**, daß die Fehler e_j wachsen. Da dies die Lösung der linearisierten Gleichung $y' = \lambda y + g(t)$ auch tut, ist das Anwachsen des Fehlers akzeptabel, ja sogar unvermeidbar, da Rundungsfehler stets in der Größenordnung (Maschinengenauigkeit) · (Größe der Werte) liegen. Der absolute Fehler sollte jedoch nicht wesentlich schneller als die Lösung wachsen.
Daher die folgende Definition.

Definition 61 *Ein lineares k-Schrittverfahren heißt für $h\lambda$ relativ stabil, falls $|z_j(h\lambda)| < |z_1(h\lambda)|$ für alle $j = 2, \ldots, k$ (z_1 Hauptwurzel), sonst heißt es relativ instabil. Den Bereich relativer Stabilität definiert man als*

$$R_\lambda = \{h\lambda : |z_j(h\lambda)| < |z_1(h\lambda) \text{ für alle } j > 1\}$$

Bemerkung 62

i) Für relativ stabile Verfahren wachsen die rarasitären Zweige langsamer als der Hauptzweig, die Hauptwurzel darf größer als 1 sein.

ii) Der Bereich R_λ kann aus disjunkten Mengen bestehen.

Beispiel 63 Wiederholt betrachten wir die Mittelpunktsregel

$$y_{j+2} = y_j + 2h\lambda y_{j+1}$$

mit $\pi(z) = z^2 - 2h\lambda z - 1$ und den Nullstellen $z_1 = h\lambda + \sqrt{1 + h^2\lambda^2}$ und $z_2 = h\lambda - \sqrt{1 + h^2\lambda^2}$. Für $h\lambda > 0$ ist $|z_2| < |z_1|$. Der Bereich relativer Stabilität ist daher die gesamte rechte Halbebene.

Beispiel 64 Man betrachte das konsistente und null-stabile Verfahren

$$y_{j+2} - y_j = \frac{h}{2}(f_{j+1} + 3f_j)$$

wieder für $f(t, y) = \lambda y$. Hier ist $\pi(z, h\lambda) = z^2 - \frac{1}{2}h\lambda z - (1 + \frac{3}{2}h\lambda)$ mit den Wurzeln

$$z_1 = \frac{1}{4}h\lambda + \sqrt{\frac{1}{16}h^2\lambda^2 + \frac{3}{2}h\lambda + 1}$$
$$z_2 = \frac{1}{4}h\lambda - \sqrt{\frac{1}{16}h^2\lambda^2 + \frac{3}{2}h\lambda + 1}$$

Entwickeln wir wie üblich nach h ($\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{4}\frac{x^2}{2} + \frac{3}{8}\frac{x^3}{6} \mp \ldots$)

$$z_1 = 1 + h\lambda + \mathcal{O}(h^2)$$
$$z_2 = -1 - \frac{1}{2}h\lambda + \mathcal{O}(h^2)$$

Intervall absoluter Stabilität ist daher $(\alpha, 0)$ mit geeignetem $\alpha < 0$, Intervall relativer Stabilität ist $(0, \beta)$ mit geeignetem $\beta > 0$. Nach dem absoluten Stabilitätskriterium würden wir dieses Verfahren also nur für $\partial f/\partial y < 0$ verwenden. Dann nimmt der absolute Fehler ab, aber nicht so schnell wie die Lösung, denn der Fehler nimmt ab wie $|z_2^j| = |1 + \frac{1}{2}h\lambda - \mathcal{O}(h^2)|^j$, die Lösung $\mathrm{e}^{\lambda t_j} = \mathrm{e}^{\lambda(t_0+jh)} = \mathrm{e}^{\lambda t_0}(\mathrm{e}^{\lambda h})^j = \mathrm{e}^{\lambda t_0}(1+h\lambda+\mathcal{O}(h^2))^j$ nimmt ab wie $|z_1|^j = |1+h\lambda+\mathcal{O}(h^2)|^j$. Nach dem relativen Stabilitätsbegriff würden wir dieses Verfahren nur für $\partial f/\partial y < 0$ verwenden und einen Fehler akzeptieren, der wächst, aber nicht schneller als die Lösung. Hier scheint also das Konzept der relativen Stabilität günstiger zu sein.
Beide Konzepte haben Vor- und Nachteile. Die Aussagen bei absoluter Stabilität sind tendenziell zu optimistisch, die der relativen Stabilität oft schwer herzuleiten und gelegentlich pessimistisch. Die Menge der „guten“ $h\lambda$ ist in Wahrheit größer als der Stabilitätsbereich angibt.

Zusammenfassend läßt sich sagen:
Für kleine $h\lambda > 0$ sind konsistente, null-stabile Verfahren absolut instabil ($z_1(h\lambda) = \mathrm{e}^{\lambda h} + \mathcal{O}(h^{p+1})$). Ein relatives Konzept ist daher angebracht. Für $h\lambda < 0$ sind relativ stabile Mehrschrittverfahren für kleine $|h\lambda|$ auch absolut stabil (wegen $z_1(h\lambda) = \mathrm{e}^{\lambda h} + \mathcal{O}(h^{p+1})$). Das relative Konzept ist also im allgemeinen vorzuziehen. Das absolute Konzept ist jedoch für steife Differentialgleichungen wichtig, wie wir es im folgenden Abschnitt sehen werden.

1.13 Steife Differentialgleichungen

Bisher haben wir immer auf steife Differentialgleichungen verwiesen, um die Notwendigkeit von impliziten Ein- und Mehrschrittverfahren. Daraus läßt sich eine laxe Beschreibung steifer Differentialgleichungen ableiten, nämlich „Differentialgleichungen bezeichnet man als steif, wenn explizite Verfahren versagen." Im folgenden wollen wir nun auf eine präzisere Definition und eine Motivation des Begriffs „Steifheit einer Differentialgleichung" eingehen.
Probleme der Kontrolltheorie und des Ablaufs chemischer Reaktionen zeigen häufig dieses Phänomen. Betrachten wir daher im folgenden Beispiel den Zerfall von Ozon in höheren Luftschichten, welcher durch eine gewöhnliche Differentialgleichung beschrieben wird.

Beispiel 65 Ozon O_3 zerfällt in höheren Luftschichten unter der Einwirkung der Sonnenstrahlung in der durch

$$\begin{aligned} O_3 + O_2 &\underset{k_2}{\overset{k_1}{\rightleftharpoons}} O + 2O_2 \\ O_3 + O &\overset{k_3}{\rightarrow} \qquad O_2 \end{aligned}$$

beschriebenen Form. Dabei werden die Reaktionsgeschwindigkeiten k_i $(i = 1, 2, 3)$ als bekannt vorausgesetzt.
Bezeichnen wir nun mit $y_1 = [O_3]$, $y_2 = [O]$, $y_3 = [O_2]$ die Konzentrationen der miteinander reagierenden Gase, so läßt sich die Reaktion durch die Lösungen des Differentialgleichungssystems

$$\begin{aligned} \dot{y}_1 &= -k_1 y_1 y_3 + k_2 y_2 y_3^2 - k_3 y_1 y_2 \\ \dot{y}_2 &= \quad k_1 y_1 y_2 - k_2 y_2 y_3^2 - k_3 y_1 y_2 \\ \dot{y}_3 &= -k_1 y_1 y_3 + k_2 y_2 y_3^2 + k_3 y_1 y_2 \end{aligned}$$

beschreiben. Für eine Modellierung unter vereinfachten Annahmen verweisen wir auf den Anhang. Vereinfachend nehmen wir an, daß sich die Konzentration von Sauerstoff $[O_2]$ nicht ändert, d.h. $\dot{y}_3 = 0$. Skalierung der kinetischen Parameter k_j liefert dann

$$\begin{aligned} \dot{y}_1 &= -y_1 - y_1 y_2^2 + 294 y_2, \qquad & y_1(0) &= 1 \\ \dot{y}_2 &= \frac{(y_1 - y_1 y_2)}{98} - 3y_2, \qquad & y_2(0) &= 1 \\ 0 &\leq t \end{aligned}$$

Mit dem expliziten, impliziten Euler-Verfahren und dem klassischen Runge-Kutta-Verfahren berechnete Näherungen sind in Abbildung **??** dargestellt.

Was ist nun das typische an Problemen der chemischen Reaktionskinetik oder der Kontrolltheorie. Betrachten wir dazu zwei weitere Beispiele.

Beispiel 66 Sei

$$y' = Ay, \quad y(0) = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \text{ mit } A = \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{pmatrix} \tag{1.84}$$

Die Eigenwerte von A lauten -2, $-40+40i$, $-40-40i$. Die exakte Lösung des Differentialgleichungssystems ist gegeben durch $y = (u, v, w)^T$ mit

$$\begin{aligned} u(t) &= \frac{1}{2}\mathrm{e}^{-2t} + \frac{1}{2}\mathrm{e}^{-40t}(\cos 40t + \sin 40t) \\ v(t) &= \frac{1}{2}\mathrm{e}^{-2t} - \frac{1}{2}\mathrm{e}^{-40t}(\cos 40t + \sin 40t) \\ w(t) &= -\mathrm{e}^{-40t}(\cos 40t - \sin 40t) \end{aligned}$$

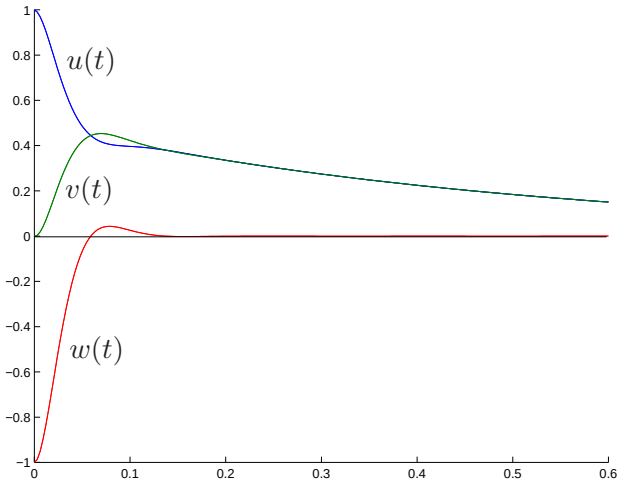

Abb. 1.12: Exakte Lösung von (1.84).

Ab 0.2 gibt es keine großen Schwankungen mehr im Lösungsverlauf. Man möchte daher eine nicht zu kleine Schrittweite bei der numerischen Lösung benutzen. Betrachten wir das Euler-Verfahren mit der Schrittweite $h = 0.05$, so erhalten wir für $u(t)$ auf $[0.2, 0.6]$ die Näherungen (**??**) in der folgenden Graphik.

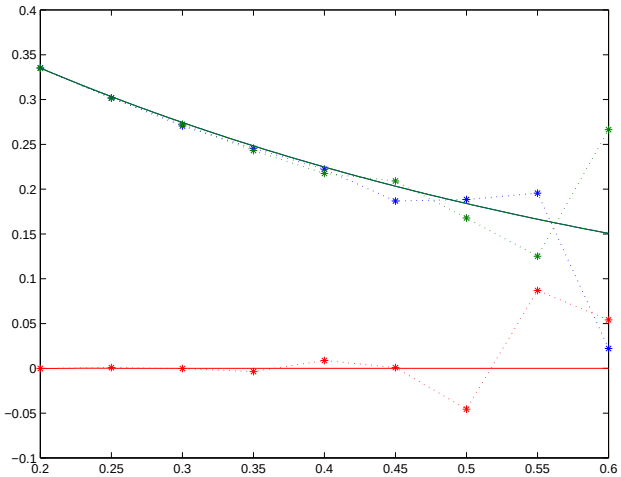

Abb. 1.13: Exakte und Näherungslösung von (1.84).

Hier ist der letzte Näherungswert nicht mehr akzeptabel.

Beispiel 67 Wenn man dagegen das Problem mit

$$A = \begin{pmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad y(0.2) = \begin{pmatrix} \frac{1}{2}\mathrm{e}^{-0.4} \\ \frac{1}{2}\mathrm{e}^{-0.4} \\ 0 \end{pmatrix}$$

betrachtet, so ist die Lösung gegeben durch $y = \frac{1}{2}(\mathrm{e}^{-2t}, \mathrm{e}^{-2t}, 0)^T$. Dies ist für $t > 0.2$ nicht von der Lösung aus dem Beispiel 66 zu unterscheiden $(\max_{t>0.2} |y_{(\times)} - y_{(\times\times)}|)$.
Wendet man hierauf das Euler-Verfahren mit Schrittweite $h = 0.05$ an, so erhält man eine befriedigende Näherungslösung.
Der Grund des so unterschiedlichen Verhaltens liegt in den verschiedenen Bereichen der absoluten Stabilität. Damit für das erste Beispiel $h\lambda$ für alle Eigenwerte von A in diesem Bereich liegen, muß $h < 0.018$ sein $(h < 1/|-39 \pm 40i|)$. Verantwortlich für diese Einschränkung sind die Eigenwerte $-40 \pm 40i$, obwohl ihre Anteile an der Lösung ab 0.2 fast verschwinden. Für das zweite Beispiel sind die Eigenwerte $-2, -2, 0$ und $h < 1$ sichert absolute Stabilität.

Wir betrachten nun allgemein das $n \times n$ Differentialgleichungssystem

$$y' = Ay + \phi(t)$$

(z.B. aus der Linearisierung eines Differentialgleichungssystems). Die Matrix A sei diagonalisierbar und $Re(\lambda_i) < 0$ für alle Eigenwerte λ_i von A. Dann hat die Lösung die Form

$$y(t) = \sum_{i=1}^{n} c_i \mathrm{e}^{\lambda_i t} v^{(i)} + \psi(t),$$

wobei $v^{(i)}$ der Eigenvektor von A zum Eigenwert λ_i ist. Wegen $Re(\lambda_i) < 0$ geht der erste Teil für $t \to \infty$ gegen 0. Für $t \to \infty$ strebt $y(t)$ also gegen $\psi(t)$. Dabei bezeichnet man $\psi(t)$ als stationäre Lösung.
Bei Aufgaben dieser Art ist es das Ziel, diese stationäre Lösung zu bestimmen. Wir müsssen dazu die numerische Lösung so lange verfolgen, bis der am langsamsten abklingende Term $c_i \mathrm{e}^{\lambda_i t} v^{(i)}$ vernachlässigbar klein ist. Falls $|Re(\lambda_i)|$ sehr klein ist, ist der Integrationsweg sehr lang. Falls außerdem ein $|Re(\lambda_j)|$ sehr groß ist, so muß für diesen langen Weg eine sehr kleine Schrittweite h verwendet werden, damit $h\lambda_j$ im Bereich absoluter Stabilität liegt. Falls also

$$|Re(\lambda_1)| \geq \ldots \geq |Re(\lambda_n)| \quad \text{und} \quad |Re(\lambda_1)| \gg |Re(\lambda_n)|,$$

so muß sehr lange mit sehr kleiner Schrittweite gerechnet werden. Das hier beschriebene Problem bezeichnet man als das Problem der Steifheit (stiffness, auch Steifigkeit bei manchen Autoren).

Definition 68 *Ein System $y' = Ay + \phi(t)$ heißt steif, falls*

i) $Re(\lambda_i) < 0$ *für alle Eigenwerte λ_i von A*

ii) $\max_{i=1,\ldots,n} |Re(\lambda_i)| \gg \min_{i=1,\ldots,n} |Re(\lambda_i)|$

Der Quotient $\max |Re(\lambda_i)| / \min |Re(\lambda_i)|$ *heißt Steifigkeitsquotient.*

Bemerkung 69 *Allgemeiner redet man für*

$$y' = f(t, y) = f(t, y_0) + \frac{\partial f}{\partial y}(t, y_0)(y - y_0) + \ldots$$

von der Steifheit (auf dem Intervall I), falls die Eigenwerte von $\partial f / \partial y$ $i)$ und $ii)$ auf I erfüllen.

Im ersten Beispiel dieses Abschnitts ist der Steifigkeitsquotient 20, in der Praxis ist 10^6 nicht unüblich. Das Problem ist dann, λh in den Bereich absoluter Stabilität zu bekommen.

Definition 70 *Ein Verfahren heißt A-stabil, falls der Bereich absoluter Stabilität die gesamte linke Halbebene umfaßt.*

Bei einem A-stabilen Verfahren entfallen also die Schwierigkeiten mit den steifen Systemen. Aber

Satz 71 (Dahlquist)

i) Ein explizites lineares Mehrschrittverfahren ist niemals A-stabil.

ii) Die Ordnung eines A-stabilen impliziten linearen Mehrschrittverfahrens ist ≤ 2.

iii) Das A-stabile Verfahren 2. Ordnung mit der kleinsten Fehlerkonstante ist die Trapezregel $u_{j+1} - u_j = h/2(f_{j+1} - f_j)$.

Beweis siehe z.B. Hairer, Wanner. □

Dieser Satz motiviert die folgenden Abschwächungen.

Definition 72 *Ein Verfahren heißt $A(\alpha)$-stabil, $\alpha \in]0, \pi/2[$, falls der Bereich absoluter Stabilität $W_\alpha = \{h\lambda | -\alpha < \pi - \arg(h\lambda) < \alpha\}$ umfaßt. Es heißt $A(0)$-stabil, falls es $A(\alpha)$-stabil ist für hinreichend kleines $\alpha \in]0, \pi/2[$.*

ZEICHNUNG
Beachte: Für λ mit $Re(\lambda) < 0$ liegt $h\lambda$ entweder für alle h in W_α oder für alle h außerhalb. Zum Beispiel ist $h\lambda$ für $\lambda = iz$, $z \in \mathbb{R}$, stets außerhalb von W_α, für $\lambda = z$, $z \in \mathbb{R}$ ist $h\lambda$ stets innerhalb von W_α. Falls wir wissen, daß alle Eigenwerte eines steifen Systems in W_α liegen, so kann ein $A(\alpha)$-stabiles Verfahren ohne Stabilitätseinschränkungen (an die Schrittweite) verwendet werden. Falls alle Eigenwerte reell sind, z.B. bei symmetrischen Jacobimatrizen $\partial f/\partial y$, so kann jedes $A(0)$-stabile Verfahren benutzt werden. Man kann zeigen:

Satz 73 (Widlund)

i) Ein explizites lineares Mehrschrittverfahren ist niemals $A(0)$-stabil.

ii) Die Trapezregel ist das einzige $A(0)$-stabile k-Schrittverfahren mit Ordnung $\geq k + 1$.

iii) Für alle $\alpha \in [0, \pi/2[$ gibt es ein $A(\alpha)$-stabiles lineares k-Schrittverfahren der Ordnung p mit $k = p = 3$ und $k = p = 4$.

Bemerkung 74 *Explizite Runge-Kutta-Verfahren haben i.a. einen sehr kleinen Bereich absoluter Stabilität. Implizite Runge-Kutta-Methoden sind A-stabil.*

Beispiel 75 Implizites Runge-Kutta-Verfahren mit $m = 1$ für $y' = \lambda y$.
Wir betrachten das Verfahren

$$y_{j+1} = y_j + k_1$$

mit

$$k_1 = hf(t_j + \frac{1}{2}h, y_j + \frac{1}{2}k_1).$$

In diesem Fall also

$$k_1 = \lambda h(y_j + \frac{1}{2}k_1),$$

so daß gilt

$$(1 - \frac{1}{2}h\lambda)k_1 = h\lambda y_j.$$

Damit ergibt sich

$$y_{j+1} = y_j + \frac{h\lambda}{1 - \frac{1}{2}h\lambda} y_j = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda} y_j$$

und

$$\pi(z, h\lambda) = z - \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}.$$

Falls $\lambda < 0$, so ist die Nullstelle stets betraglich < 1.
Analog für $m = 2$:

$$u_{j+2} = \frac{1 + \frac{1}{2}h\lambda + \frac{1}{12}h^2\lambda^2}{1 - \frac{1}{2}h\lambda + \frac{1}{12}h^2\lambda^2} u_j.$$

Allgemein gilt: Bei expliziten Runge-Kutta-Verfahren sind die Koeffizienten in $\pi(z, h\lambda)$ Polynome in $h\lambda$ (Taylorpolynom von $\mathrm{e}^{h\lambda}$). Bei impliziten Runge-Kutta-Verfahren sind sie rationale Funktionen in $h\lambda$ (Padé-Approximation von $\mathrm{e}^{h\lambda}$). Diese sind vom Betrag < 1 für $Re(\lambda) < 0$, also sind die Verfahren A-stabil.

Beispiel 76 Explizites Runge-Kutta-Verfahren mit $m = 2$ für $y' = \lambda y$.
Wir betrachten das modifizierte Euler-Verfahren

$$\begin{aligned} y_{j+1} - y_j &= hf(t_j + \frac{1}{2}h, y_j + \frac{1}{2}hf(t_j, y_j)) \\ &= h\lambda(1 + \frac{1}{2}h\lambda)y_j \\ &= (h\lambda + \frac{h^2\lambda^2}{2})y_j. \end{aligned}$$

Damit ergibt sich

$$\pi(z, h\lambda) = z - (1 + h\lambda + \frac{h^2\lambda^2}{2}).$$

Das Intervall absoluter Stabilität ist daher $(-2, 0)$.

Mehr über steife Differentialgleichungen findet man in Hairer, Wanner.

Literaturverzeichnis

[A] R.A. ADAMS, "Sobolev Spaces", Pure Appl. Math. 65, Academic Press, New York, 1975.

[G] G. H. GOLUB, C. F. VAN LOAN, Matrix Computations, 3. ed., Hopkins Univ. Press, 1996.

[H] W. HACKBUSCH, Iterative Lösung großer schwachbesetzter Gleichungssysteme, 2. Auflage, Teubner-Verlag, Stuttgart, 1993.

[HH] G. HÄMMERLIN, K.-H. HOFFMANN, Numerische Mathematik, 4. Auflage, Springer-Verlag, Berlin u.a., 1994.

[P] R. PLATO, Numerische Mathematik kompakt, Vieweg-Verlag.

[Sc] H. R. SCHWARZ, Numerische Mathematik, 4. Auflage, Teubner-Verlag, Stuttgart, 1997.

[St] J. STÖR, R. BULIRSCH, Numerische Mathematik 1 und 2, Springer-Verlag, Berlin u.a., 1994.

[SW] K. STREHMEL, R. WEINER, Numerik gewöhnlicher Differentialgleichungen, Teubner-Verlag, Stuttgart, 1995.

[TS] W. TÖRNIG, P. SPELLUCCI, Numerische Mathematik für Ingenieure und Physiker, Band 1 & 2, Springer-Verlag, Berlin u.a.

[W] W. WALTER, Gewöhnliche Differentialgleichungen, 6. Auflage, Springer-Verlag, Berlin u.a., 1996.

Index