2 Einschrittmethoden

2.1 Die Eulersche Polygonzugmethode

Wir betrachten eine AWA der Form

$$u'(t) = f(t, u(t)), \quad t \in I = [t_0, t_0 + T], \quad u(t_0) = u_0. \tag{2.1.1}$$

Die Funktion $f(t,x)$ sei stetig auf $I \times \mathbb{R}^d$ und genüge einer globalen Lipschitz-Bedingung

$$\|f(t,x) - f(t,y)\| \le L_f \|x - y\|, \quad (t,x),(t,y) \in I \times \mathbb{R}^d. \tag{2.1.2}$$

Dann existiert eine eindeutig bestimmte Lösung $u(t)$ von (2.1.1) für alle $t \in I$. Letzteres folgt nach dem Fortsetzungssatz aus der linearen Beschränktheit der Funktion $f(t,x)$:

$$\|f(t,x)\| \le \|f(t,x) - f(t,0)\| + \|f(t,0)\| \le L_f \|x\| + \|f(t,0)\|.$$

Diese Lösung sei ferner hinreichend glatt. Gelegentlich werden wir auch den Fall $T \to \infty$, d. h. $I = [t_0, \infty)$, betrachten.

Zur Approximation der AWA wählt man zunächst eine Folge von diskreten Zeitpunkten $t_0 < t_1 < \ldots < t_n < \ldots < t_N = t_0 + T$ und setzt

$$I_n := [t_{n-1}, t_n], \quad h_n := t_n - t_{n-1}, \quad h := \max_{1 \le n \le N} h_n.$$

Die Eulersche Polygonzugmethode erzeugt nun ausgehend von einem Startwert $y_0^h \in \mathbb{R}^d$ eine Folge $(y_n^h)_{n \in \mathbb{N}}$ durch die rekursive Vorschrift

$$y_n^h = y_{n-1}^h + h_n f(t_{n-1}, y_{n-1}^h), \quad n = 1, \ldots, N. \tag{2.1.3}$$

Wir schreiben dies auch in Form einer Differenzengleichung

$$(L_h y^h)_n = 0, \quad n = 1, \ldots, N, \tag{2.1.4}$$

mit dem „Differenzenoperator"

$$(L_h y^h)_n := h_n^{-1}(y_n^h - y_{n-1}^h) - f(t_{n-1}, y_{n-1}^h), \tag{2.1.5}$$

für „Gitterfunktionen" $y^h = \{y_n^h\}_{n=1,\ldots,N}$.

Als Nebenprodukt unseres Beweises des Existenzsatzes von Peano (Satz 1.1) haben wir gesehen, dass wegen der Eindeutigkeit der Lösung u die Werte y_n^h für $h \to 0$ gegen die Funktionswerte $u(t_n)$ konvergieren (vorausgesetzt y_0^h konvergiert gegen u_0):

$$\max_{0 \le n \le N} \|y_n^h - u(t_n)\| \to 0 \quad (h \to 0). \tag{2.1.6}$$

Zur Abschätzung der Geschwindigkeit der Konvergenz des Diskretisierungsverfahrens (2.1.3) führen wir den sog. „Abschneidefehler" (auch „lokaler Diskretisierungsfehler" genannt) ein:

$$\tau_n^h := (L_h u^h)_n = h_n^{-1}\{u_n^h - u_{n-1}^h\} - f(t_{n-1}, u_{n-1}^h),$$

mit der Gitterfunktion $u^h = (u_n^h := u(t_n))_{0\le n\le N}$. Für den Abschneidefehler gilt offenbar wegen $u'(t) = f(t,u)$ die Beziehung

$$\tau_n^h = h_n^{-1}\int_{t_{n-1}}^{t_n} u'(t)\,dt - u'(t_{n-1}) = h_n^{-1}\int_{t_{n-1}}^{t_n}(t_n - t)u''(t)\,dt$$

und folglich

$$\|\tau_n^h\| \le \tfrac{1}{2}h_n \max_{t\in I_n}\|u''(t)\|\,. \tag{2.1.7}$$

Man spricht hier von einer Diskretisierung „erster Ordnung“.

Wenn Missverständnisse ausgeschlossen sind, schreiben wir im Folgenden einfach y_n für y_n^h und entsprechend τ_n für τ_n^h. Unter Verwendung dieser Notation genügt die exakte Lösung u der gestörten Differenzengleichung

$$u_n \;=\; u_{n-1} + h_n f(t_{n-1}, u_{n-1}) + h_n\tau_n\,. \tag{2.1.8}$$

Die Abschätzung des *globalen* Diskretisierungsfehlers $e_n = y_n - u_n$ erfolgt über einen Stabilitätssatz für das Differenzenverfahren in Analogie zum Stabilitätssatz für die AWA (Satz 1.3). Vergleich von (2.1.8) mit (2.1.3) ergibt

$$e_n \;=\; e_{n-1} + h_n\big\{f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})\big\} - h_n\tau_n$$

und unter Ausnutzung der L-Stetigkeit von $f(t,x)$,

$$\|e_n\| \le \|e_{n-1}\| + h_nL\|e_{n-1}\| + h_n\|\tau_n\|. \tag{2.1.9}$$

Durch sukzessive Anwendung dieser Beziehung erhält man die diskrete Integralungleichung („Summenungleichung”)

$$\|e_n\| \le \|e_0\| + L\sum_{\nu=0}^{n-1} h_{\nu+1}\|e_\nu\| + \sum_{\nu=1}^{n} h_\nu\|\tau_\nu\|\,. \tag{2.1.10}$$

Zur weiteren Abschätzung benötigen wir die folgende diskrete Version des Gronwallschen Lemmas (Hilfssatz 1.1).

Hilfssatz 2.1 (Diskretes Gronwallsches Lemma): *Es seien* $(w_n)_{n\ge0}$, $(a_n)_{n\ge0}$ *und* $(b_n)_{n\ge0}$ *Folgen nichtnegativer Zahlen, für die gilt* $w_0 \le b_0$ *und*

$$w_n \le \sum_{\nu=0}^{n-1} a_\nu w_\nu + b_n\,, \quad n \ge 1\,. \tag{2.1.11}$$

Ist die Folge $(b_n)_{n\ge0}$ *monoton steigend so gilt die Abschätzung*

$$w_n \le \exp\Big(\sum_{\nu=0}^{n-1} a_\nu\Big)b_n\,, \quad n \ge 1\,. \tag{2.1.12}$$

Beweis: Der Beweis könnte durch Rückführung auf das *kontinuierliche* Gronwallsche Lemma (für stückweise konstante Funktionen) bewiesen werden. Wir wollen hier aber lieber einen einfachen, direkten Beweis geben. Dazu definieren wir Zahlen $d_n \geq 0$ und S_n durch

$$S_0 := w_0 + d_0 = b_0, \qquad S_n := w_n + d_n = \sum_{\nu=0}^{n-1} a_\nu w_\nu + b_n \,.$$

Es gilt dann

$$S_n - S_{n-1} = a_{n-1} w_{n-1} + b_n - b_{n-1} \,, \quad n \geq 1.$$

Hieraus wollen wir durch Induktion nach n erschließen, dass

$$S_n \leq \exp\Big(\sum_{\nu=0}^{n-1} a_\nu \Big) b_n \,, \quad n \geq 0 \,, \tag{2.1.13}$$

wobei wie üblich im Fall $n = 0$ die Summation „leer" ist, d. h. $S_0 \leq b_0$. Zunächst ist nach Definition

$$S_0 \leq b_0 \,.$$

Sei (2.1.13) nun als richtig angenommen für $n-1$. Dann gilt wegen $b_n \geq b_{n-1}$

$$\begin{aligned} S_n &\leq S_{n-1} + a_{n-1} w_{n-1} + b_n - b_{n-1} \leq (1 + a_{n-1}) S_{n-1} + b_n - b_{n-1} \\ &\leq (1 + a_{n-1}) \exp\Big(\sum_{\nu=0}^{n-2} a_\nu \Big) b_{n-1} + b_n - b_{n-1} \\ &\leq e^{a_{n-1}} \exp\Big(\sum_{\nu=0}^{n-2} a_\nu \Big) \{ b_{n-1} + b_n - b_{n-1} \} \leq \exp\Big(\sum_{\nu=0}^{n-1} a_\nu \Big) b_n \,. \end{aligned}$$

Dies impliziert wegen $w_n \leq S_n$ die Behauptung. Q.E.D.

Im Zusammenhang mit *impliziten* Verfahren, wie z. B.. dem „impliziten Euler-Schema"

$$y_n = y_{n-1} + h_n f(t_n, y_n), \tag{2.1.14}$$

wird eine verschärfte Variante des diskreten Gronwallschen Lemmas benötigt, bei der eine *implizite* Differenzenungleichung der Form

$$w_n \;\leq\; \sum_{\nu=0}^{n} a_\nu w_\nu + b_n \,, \quad n \geq 1 \,. \tag{2.1.15}$$

angenommen wird. Unter der Annahme, dass $a_n < 1$ wird diese durch Elimination des führenden Summanden auf der rechten Seite in die folgende *explizite* Form überführt:

$$\sigma_n^{-1} w_n \;\leq\; \sum_{\nu=0}^{n-1} \sigma_\nu a_\nu \sigma_\nu^{-1} w_\nu + b_n \,, \quad n \geq 1 \,,$$

mit den Parametern $\sigma_\nu := (1-a_\nu)^{-1}$. Anwendung von (2.1.12) auf diese Situation liefert die Ungleichung

$$\sigma_n^{-1} w_n \;\le\; \exp\Big(\sum_{\nu=0}^{n-1} \sigma_\nu a_\nu\Big) b_n\,, \quad n \ge 1\,,$$

und bei Beachtung von $\sigma_n = 1 + \sigma_n a_n \le \exp(\sigma_n a_n)$ schließlich das Endresultat

$$w_n \;\le\; \exp\Big(\sum_{\nu=0}^{n} \sigma_\nu a_\nu\Big) b_n\,, \quad n \ge 1\,. \tag{2.1.16}$$

Wir fahren nun mit unserer Fehleranalyse für das Euler-Verfahren fort. Aus (2.1.10) erschließen wir mit dem diskreten Gronwallschen Lemma die *a priori* Fehlerabschätzung

$$\|e_n\| \le e^{L(t_n - t_0)} \Big\{ \|e_0\| + \sum_{\nu=1}^{n} h_\nu \|\tau_\nu\| \Big\}\,, \quad n \ge 1\,, \tag{2.1.17}$$

bzw.

$$\max_{1\le n\le N} \|e_n\| \;\le\; e^{LT} \Big\{ \|e_0\| + T \max_{1\le n\le N} \|\tau_n\| \Big\}\,. \tag{2.1.18}$$

Bei Berücksichtigung der Abschätzung (2.1.7) für τ_ν folgt also für den globalen Diskretisierungsfehler der Eulerschen Polygonzugmethode

$$\max_{1\le n\le N} \|e_n\| \;\le\; e^{LT} \left\{ \|e_0\| + \tfrac{1}{2} T \max_{1\le n\le N} \{ h_n \max_{t\in I_n} \|u''(t)\| \} \right\}\,, \tag{2.1.19}$$

d. h.: Die „(globale) Konvergenzordnung„ ist (mindestens) gleich der „(lokalen) Konsistenzordnung“. Man beachte den exponentiellen Faktor in der Abschätzung (2.1.19). Wegen ihrer geringen Genauigkeit hat die Eulersche Polygonzugmethode in der Praxis keine Bedeutung. Die Herleitung der Abschätzung (2.1.19) ist aber exemplarisch für eine große Klasse von Methoden.

2.2 Allgemeine Einschrittmethoden

Der naheliegendste Weg zur Konstruktion von Differenzenformeln höherer Ordnung ist der über die Taylor-Entwicklung (skalarer Fall $d = 1$):

$$u(t) = \sum_{r=0}^{R} \frac{h^r}{r!} u^{(r)}(t-h) + \frac{h^{R+1}}{(R+1)!} u^{(R+1)}(\xi)\,, \quad \xi \in [t-h, t]\,.$$

Da u der Differentialgleichung $u' = f(t,u)$ genügt, gilt

$$u^{(r)}(t) = \Big(\frac{d}{dt}\Big)^{r-1} f(t, u(t)) \;=:\; f^{r-1}(t, u(t))\,.$$

Die „(R-stufige) Taylor-Verfahren" lauten dann

$$y_n = y_{n-1} + h_n \sum_{r=1}^{R} \frac{h_n^{r-1}}{r!} f^{(r-1)}(t_{n-1}, y_{n-1}) \,, \quad n \geq 1 \,. \tag{2.2.20}$$

Wir schreiben dies in der allgemeinen Form eines sog. „Einschrittverfahrens"

$$y_n = y_{n-1} + h_n F(h_n; t_{n-1}, y_n, y_{n-1}) \,, \tag{2.2.21}$$

bzw.

$$(L_h y^h)_n := h_n^{-1}(y_n - y_{n-1}) - F(h_n; t_{n-1}, y_n, y_{n-1}) = 0 \,, \tag{2.2.22}$$

mit der sog. „Verfahrensfunktion"

$$F(h; t, y, x) \; := \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t, x) \,.$$

Die Bezeichnung *Einschrittverfahren* erklärt sich dabei selbst. Da hier die Verfahrensfunktion nur von der unmittelbar vorausgehenden Näherung y_{n-1} abhängt, wird diese Methode *explizit* genannt. Zur Durchführung expliziter Differenzenverfahren ist in jedem Zeitschritt lediglich eine Funktionsauswertung $F(h_n; t_{n-1}, y_{n-1})$ durchzuführen, während implizite Formeln die Lösung i. Allg. nichtlinearer Gleichungssysteme erfordern. Wir werden uns daher zunächst hauptsächlich mit expliziten Verfahren beschäftigen.

Den *Abschneidefehler* der Formel (2.2.22) definiert man analog zu (2.1.7) durch

$$\tau_n := \; (L_h u^h)_n \; = \; h_n^{-1}\{u_n - u_{n-1}\} - F(h_n; t_{n-1}, u_n, u_{n-1}) \,.$$

Definition 2.1 (Konsistenz): *Die Einschrittmethode (2.2.22) heißt „konsistent (mit der AWA)" bzw. „konsistent mit Konsistenzordnung m", wenn*

$$\max_{t_n \in I} \|\tau_n\| \to 0 \quad \textit{bzw.} \quad \max_{t_n \in I} \|\tau_n\| = O(h^m) \quad (h \to 0) \,. \tag{2.2.23}$$

Offenbar hat die R-stufige Taylor-Formel für skalare AWA gerade die Konsistenzordnung $m = R$. Zur Auswertung dieser Formel müssen Ableitungen von $f(t, x)$ berechnet werden, z. B. mit den Abkürzungen $f_t := \partial f/\partial t$, $f_x := \partial f/\partial x$:

$$f^{(1)}(t, x) = [f_t + f_x f](t, x) \,, \quad f^{(2)}(t, x) = [f_{tt} + 2f_{tx} f + f_t f_x + f_{xx} f^2 + f_x^2 f](t, x) \,.$$

In der Praxis kann dies sehr aufwendig sein; man berechne z. B. $f^{(3)}(t, x)$ für $f(t, x) = (t + x^2)^{1/2} \arctan(t + x)$. Zur Vermeidung dieses Nachteils können die Ableitungen $f^{(r-1)}(t, u)$ durch Differenzenquotienten ersetzt werden, bei denen nur Auswertungen von $f(t, u)$ auftreten. Z. B. ist für die Taylor-Formel der Stufe $R = 2$

$$\begin{aligned} f^{(1)}(t, u(t)) &\approx h^{-1} \big\{ f(t + h, u(t + h)) - f(t, u(t)) \big\}, \\ &\approx h^{-1} \big\{ f(t + h, u(t) + h f(t, u(t)) - f(t, u(t)) \big\}, \end{aligned}$$

was auf die folgende Formel führt:

$$\begin{aligned} y_n &= y_{n-1} + h_n f(t_{n-1}, y_{n-1}) + \tfrac{1}{2} h_n \left\{ f(t_n,\ y_{n-1} + h_n f(t_{n-1}, y_{n-1})) - f(t_{n-1}, y_{n-1}) \right\} \\ &= y_{n-1} + h_n \left\{ \tfrac{1}{2} f(t_{n-1}, y_{n-1}) + \tfrac{1}{2} f(t_n, y_{n-1} + h_n f(t_{n-1}, y_{n-1})) \right\}. \end{aligned}$$

Wenn man bei den obigen Entwicklungsschritten die Restglieder verfolgt, findet man, dass diese Differenzenformel die Konsistenzordnung $m = 2$ besitzt, genau wie die zugehörige Taylor-Formel. Allgemein haben die so entstehenden sog. „(expliziten) Runge[1]-Kutta[2]-Verfahren" die Form

$$F(h; t, x) = \sum_{r=1}^{R} c_r k_r(h; t, x)$$

$$k_1 = f(t, x)\,, \quad k_r = f(t + h a_r, x + h \sum_{s=1}^{r-1} b_{rs} k_s)\,, \quad r = 2, \ldots, R\,,$$

mit geeignet gewählten Konstanten a_r, c_r, b_{rs}. Diese werden so bestimmt, dass mit einem möglichst großen m (im Idealfall $m = R$) gilt:

$$\sum_{r=1}^{R} c_r k_r(h; t, u(t)) = \sum_{r=1}^{m} \frac{h^{r-1}}{r!} f^{(r-1)}(t, u(t)) + O(h^m)\,.$$

Die Konsistenzordnung der entsprechenden Runge-Kutta-Formel" ist dann konstruktionsgemäß gerade m.

Beispiel 2.1: Runge-Kutta-Methoden der Stufen $R = 1, 2, 3, 4$:

- $\mathbf{R = 1}$: *Eulersche Polygonzugmethode*
- $\mathbf{R = 2}$: Durch Taylor-Entwicklung und Koeffizientenvergleich erhält man aus der Bedingung (setze $f = f(t, u)$, $f_t = f_t(t, u)$, u.s.w.)

$$\begin{aligned} c_1 f + c_2 f(t + h a_2, u + h b_{21} f) &= (c_1 + c_2) f + c_2 a_2 h f_t + c_2 b_{21} h f f_x + O(h^2) \\ &= f + \tfrac{1}{2} h \{ f_t + f_x f \} + O(h^2) \end{aligned}$$

  die Bestimmungsgleichungen $c_1 + c_2 = 1$ und $c_2 a_2 = c_2 b_{21} = \frac{1}{2}$. Als mögliche Lösungen ergeben sich z. B..:

  - $c_1 = c_2 = \frac{1}{2},\ a_2 = b_{21} = 1$ (*„Heunsches[3] Verfahren 2-ter Ordnung"*):

$$y_n = y_{n-1} + \tfrac{1}{2} h_n \{ f(t_{n-1}, y_{n-1}) + f(t_n, y_{n-1} + h_n f(t_{n-1}, y_{n-1})) \},$$

[1] Carl David Tolme Runge (1856–1927): Deutscher Mathematiker und Physiker: Promotion in Berlin bei K. Weierstrass; 1868 Prof. in Hannover; arbeitete u. ü. Spektroskopie; ab 1904 Prof. in Göttingen (Promoter F. Klein).

[2] Kutta (1867–1944): Deutscher Mathematiker; 1910–1912 Prof. an der RWTH Aachen; ab 1912 Prof. in Stuttgart; vor allem bekannt für seine Beiträge zur Aerodynamik (sog. „Kutta-Bedingung").

[3] Karl Heun (1859–1929): Deutscher Mathematiker; Promotion 1881 in Göttingen; Habilitation 1886 in München; ab 1902 Prof. für Theoretische Mechanik in Karlsruhe; Beiträge zu gew. Differentialgleichungen und speziellen Funktionen.

- $c_1 = 0\,, \quad c_2 = 1,\ a_2 = b_{21} = \frac{1}{2}$ (*„modifiziertes Euler-Verfahren"*):

$$y_n = y_{n-1} + h_n f(t_{n-1/2}, y_{n-1} + \tfrac{1}{2} h_n f(t_{n-1}, y_{n-1})).$$

– $\mathbf{R = 3}$: Für die 8 freien Parameter ergeben sich die 6 Gleichungen

$$c_1 + c_2 + c_3 = 1\,, \quad c_2 a_2 + c_3 a_3 = \tfrac{1}{2}\,, \quad c_2 a_2^2 + c_3 a_3^2 = \tfrac{1}{3}\,,$$
$$c_3 a_2 b_{32} = \tfrac{1}{6}\,, \quad b_{21} - a_2 = 0\,, \quad b_{31} - a_3 + b_{32} = 0\,.$$

Als mögliche Lösungen ergeben sich z. B..:

- $c_1 = \frac{1}{4},\ c_2 = 0,\ c_3 = \frac{3}{4},\ a_2 = \frac{1}{3},\ a_3 = \frac{2}{3},\ b_{21} = \frac{1}{3},\ b_{31} = 0,\ b_{32} = \frac{2}{3}$: (*„Heunsches Verfahren 3. Ordnung"*)

$$y_n = y_{n-1} + \tfrac{1}{4} h_n \{k_1 + 3k_3\},$$
$$k_1 = f(t_{n-1}, y_{n-1}), \quad k_2 = f(t_{n-2/3}, y_{n-1} + \tfrac{1}{3} h_n k_1),$$
$$k_3 = f(t_{n-1/3}, y_{n-1} + \tfrac{2}{3} h_n k_2).$$

- $c_1 = \frac{1}{6},\ c_2 = \frac{2}{3},\ c_3 = \frac{1}{6},\ a_2 = \frac{1}{2},\ a_3 = 1,\ b_{21} = \frac{1}{2},\ b_{31} = -1,\ b_{32} = 2$: (*„Kuttasches Verfahren 3. Ordnung"*)

$$y_n = y_{n-1} + \tfrac{1}{6} h_n \{k_1 + 4k_2 + k_3\},$$
$$k_1 = f(t_{n-1}, y_{n-1}), \quad k_2 = f(t_{n-1/2}, y_{n-1} + \tfrac{1}{2} h_n k_1),$$
$$k_3 = f(t_n, y_{n-1} - h_n k_1 + 2h_n k_2)\,.$$

– $\mathbf{R = 4}$: In diesem Fall stehen 13 freie Parameter zur Verfügung, mit denen zur Konstruktion einer Formel 4-ter Ordnung 11 Bestimmungsgleichungen zu erfüllen sind. Eine der Lösungen führt auf das „klassische Runge-Kutta-Verfahren 4-ter Ordnung":

$$y_n = y_{n-1} + \tfrac{1}{6} h_n \{k_1 + 2k_2 + 2k_3 + k_4\},$$
$$k_1 = f(t_{n-1}, y_{n-1}), \quad k_2 = f(t_{n-1/2}, y_{n-1} + \tfrac{1}{2} h_n k_1),$$
$$k_3 = f(t_{n-1/2}, y_{n-1} + 12h_n k_2), \quad k_4 = f(t_n, y_{n-1} + h_n k_3).$$

Die betrachteten Differenzenformeln sind prinzipiell auch auf Systeme anwendbar. Bei der Methode der Taylor-Entwicklung ist dabei zu berücksichtigen, dass die zeitlichen Ableitungen $f^{(i)}(t,x)$ für Systeme wesentlich komplizierter aussehen; z. B..: $f^{(1)} = f_t + f_x \cdot f$ mit der Jacobi-Matrix $f'_x(t,x) = \nabla_x f(t,x)$ der Vektorfunktion $f(t,x)$ bzgl. der variablen x. Die Runge-Kutta-Formeln sind nicht ohne weiteres auf Systeme übertragbar, da zum Abgleich der in $f^{(i)}$ auftretenden Ableitungen unter Umständen mehr Parameter notwendig sind, als zur Verfügung stehen. I. Allg. gilt, dass eine Runge-Kutta-Methode der Ordnung $m \leq 4$ für eine skalare Gleichung dieselbe Ordnung auch für Systeme hat; im Falle $m \geq 5$ ist ihre Ordnung für Systeme in der Regel reduziert.

Einfache *implizite* Verfahren sind neben dem impliziten Euler-Verfahren die sog. „Trapezregel"

$$y_n = y_{n-1} + \tfrac{1}{2} h_n \big\{ f(t_n, y_n) + f(t_{n-1}, y_{n-1}) \big\}, \tag{2.2.24}$$

und die dazu sehr ähnliche „(Einschritt)-Mittelpunktsregel"

$$y_n = y_{n-1} + h_n f(t_n + \tfrac{1}{2}h_n, \tfrac{1}{2}(y_n + y_{n-1})). \tag{2.2.25}$$

Eine andere Variante der „Mittelpunktsformel" hat die Form (auf äquidistantem Gitter)

$$y_n = y_{n-2} + 2hf(t_{n-1}, y_{n-1})$$

Dies ist eine sog. „(explizite) Zweischrittformel". Alle drei Verfahren sind konsistent von *zweiter* Ordnung. Implizite Verfahren höherer Ordnung vom Typ der Runge-Kutta-Verfahren werden später betrachtet.

2.2.1 Lokale Konvergenz und Fehlerabschätzungen

Analog zum Polygonzugverfahren wollen wir nun die Konvergenz des allgemeinen Einschrittverfahrens (2.2.22) beweisen. Dazu dient die folgende fundamentale Bedingung:

Definition 2.2 (L-Stetigkeit): *Eine Einschrittformel heißt „Lipschitz-stetig" (kurz „L-stetig"), wenn ihre Verfahrensfunktion einer (gleichmäßigen) Lipschitz-Bedingung genügt:*

$$\|F(h;t,x,y) - F(h;t,\tilde{x},\tilde{y})\| \le L\{\|x-\tilde{x}\| + \|y-\tilde{y}\|\}\,, \tag{2.2.26}$$

für beliebige Argumente $(t,x),(t,\tilde{x}),(t,y),(t,\tilde{y}) \in I \times \mathbb{R}^d$.

Satz 2.1 (Diskreter Stabilitätssatz): *Eine Lipschitz-stetige Differenzenformel* (2.2.22) *ist „(diskret) stabil", d. h.: Für beliebige Gitterfunktionen* $y^h = \{y_n\}_{n\ge 0}$, $z^h = \{z_n\}_{n\ge 0}$ *gilt für hinreichend kleine Schrittweite* $h < \frac{1}{2}L^{-1}$ *die Abschätzung*

$$\|y_n - z_n\| \le e^{\kappa L(t_n - t_0)}\Big\{\|y_0 - z_0\| + \sum_{\nu=1}^{n} h_\nu \|(L_h y^h - L_h z^h)_\nu\|\Big\}\,, \tag{2.2.27}$$

mit der Lipschitz-Konstante L *der Verfahrensfunktion* $F(h;t,x,y)$ *und der Kontante* $\kappa = 4$ *für allgemeine implizite Methoden. Für explizite Methoden ist* $\kappa = 1$ *und die Schrittweitenbedingung kann entfallen.*

Beweis: Für zwei Gitterfunktionen $\{y_n\}_{n\ge 0}$, $\{z_n\}_{n\ge 0}$ erhalten wir durch Vergleich von

$$\begin{aligned}(L_h y^h)_n &= h_n^{-1}(y_n - y_{n-1}) - F(h_n; t_n, y_n, y_{n-1}),\\ (L_h z^h)_n &= h_n^{-1}(z_n - z_{n-1}) - F(h_n; t_n, z_n, z_{n-1})\end{aligned}$$

die Gleichung

$$y_n - z_n = y_{n-1} - z_{n-1} + h_n\big\{F(h;t_n,y_n,y_{n-1}) - F(h;t_n,z_n,z_{n-1}) + (L_h y^h - L_h z^h)_n\big\}.$$

Für das Folgende setzen wir $e_n := y_n - z_n$ und $\varepsilon_n := (L_h y^h - L_h z^h)_n$.

(i) Expliziter Fall: Unter Ausnutzung der L-Stetigkeit der Verfahrensfunktion ergibt sich

$$\|e_n\| \le \|e_{n-1}\| + h_n L\|e_{n-1}\| + h_n\|\varepsilon_n\|$$

und folglich durch rekursive Anwendung dieser Abschätzung:

$$\|e_n\| \le \|e_0\| + \sum_{\nu=0}^{n-1} Lh_{\nu+1}\|e_\nu\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|.$$

Mit Hilfe des diskreten Gronwallschen Lemmas 2.1 erhalten wir hieraus

$$\|e_n\| \le \exp\Big(L\sum_{\nu=0}^{n-1} h_{\nu+1}\Big)\Big\{\|e_0\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|\Big\} = e^{L(t_n-t_0)}\Big\{\|e_0\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|\Big\}.$$

Diese Abschätzung gilt, wie behauptet, ohne jede Bedingung an die Schrittweiten h_n.

(ii) Impliziter Fall: Für *implizite* Verfahren ergibt sich wieder unter Ausnutzung der L-Stetigkeit der Verfahrensfunktion

$$\|e_n\| \le \|e_{n-1}\| + h_n L\big\{\|e_n\| + \|e_{n-1}\|\big\} + h_n\|\varepsilon_n\|.$$

Dies impliziert mit der Setzung $h_0 := 0$ bei Beachtung der Bedingung $h < \frac{1}{2}L^{-1}$:

$$\begin{aligned}(1 - h_nL)\|e_n\| &\le (1 + h_nL)\|e_{n-1}\| + h_n\|\varepsilon_n\| \\ &= (1 - h_{n-1}L)\|e_{n-1}\| + \frac{h_n + h_{n-1}}{1 - h_{n-1}L}L(1 - h_{n-1}L)\|e_{n-1}\| + h_h\|\varepsilon_n\|,\end{aligned}$$

und weiter mit der Notation $w_n := (1 - h_nL)e_n$:

$$\|w_n\| \le \|w_{n-1}\| + \frac{h_n + h_{n-1}}{1 - h_{n-1}L}L\|w_{n-1}\| + h_h||\varepsilon_n\|$$

Rekursive Anwendung dieser Abschätzung ergibt dann

$$\|w_n\| \le \|w_0\| + \sum_{\nu=0}^{n-1} \frac{h_{\nu+1} + h_\nu}{1 - h_\nu L}L\|w_\nu\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|.$$

Mit Hilfe des diskreten Gronwallschen Lemmas erhalten wir hieraus

$$\begin{aligned}\|e_n\| &\le \frac{1}{1 - h_nL}\exp\Big(L\sum_{\nu=0}^{n-1}\frac{h_{\nu+1} + h_\nu}{1 - h_\nu L}\Big)\Big\{\|e_0\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|\Big\} \\ &\le \exp\Big(\frac{h_nL}{1 - h_nL}\Big)\exp\Big(L\sum_{\nu=0}^{n-1}\frac{h_{\nu+1} + h_\nu}{1 - h_\nu L}\Big)\Big\{\|e_0\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|\Big\} \\ &\le e^{4L(t_n-t_0)}\Big\{\|e_0\| + \sum_{\nu=1}^{n} h_\nu\|\varepsilon_\nu\|\Big\}.\end{aligned}$$

Dies vervollständigt den Beweis. Q.E.D.

Satz 2.2 (Konvergenzsatz): *Die Differenzenformel* (2.2.22) *sei L-stetig und konsistent mit der AWA. Im Falle* $\|y_0 - u_0\| \to 0$ *konvergiert dann*

$$\max_{t_n \in I} \|y_n - u(t_n)\| \to 0 \quad (h \to 0), \tag{2.2.28}$$

und für hinreichend kleine Schrittweite $h < \frac{1}{2}L^{-1}$ *gilt die a priori Fehlerabschätzung*

$$\|y_n - u(t_n)\| \leq e^{4L(t_n - t_0)} \Big\{ \|y_0 - u_0\| + \sum_{\nu=1}^{n} h_\nu \|\tau_\nu\| \Big\}, \quad 0 \leq n \leq N. \tag{2.2.29}$$

Für eine explizite Methode kann die Schrittweitenbedingung entfallen.

Beweis: Es gelten die Beziehungen

$$L_h y^h = 0\,, \quad L_h u^h = \tau^h\,,$$

so dass der diskrete Stabilitätssatz 2.1 unmittelbar die Behauptung impliziert. Q.E.D.

Satz 2.2 besagt, dass für eine L-stetige Einschrittformel die *globale* Konvergenzordnung (mindestens) gleich der *lokalen* Konsistenzordnung ist. Dies gilt also z. B.. für die Taylor-Verfahren und ebenso für die Runge-Kutta-Verfahren, die ja für L-stetiges $f(t,x)$ automatisch der Bedingung (2.2.26) genügen und auch konsistent sind unter der Bedingung $\sum_{r=1}^{R} c_r = 1$ (Übungsaufgabe). Dasselbe gilt natürlich für das impliziten Euler-Verfahren, die Trapezregel sowie die Einschritt-Mittelpunktsregel.

Bemerkung 2.1: Wir haben bisher der Einfachheit halber angenommen, dass die Funktion $f(t,x)$ und entsprechend die Verfahrensfunktion $F(t,x,y)$ einer *globalen* Lipschitz-Bedingung genügen, d. h.: Die L-Konstante kann gleichmäßig für alle $x, x' \in \mathbb{R}^d$ gewählt werden. Dies schließt z. B.. Fälle wie $f(t,x) = x^2$ und $f(t,x) = x^{1/2}$ aus. Von dieser Restriktion kann man sich durch folgende Überlegung befreien: Die AWA habe eine eindeutige Lösung auf einem Intervall $I = [t_0, t_0+T]$, und die Funktion $f(t,x)$ (und entsprechend die Verfahrensfunktion $F(t,x,y)$) genüge einer L-Bedingung auf dem „Streifen"

$$U_\rho := \Big\{ (t,x) \in I \times \mathbb{R}^d \mid \|x - u(t)\| \leq \rho \Big\}$$

um die Lösung $u(t)$. Die Funktion $f(t,x)$ wird nun so von U_ρ auf $U_\infty = I \times \mathbb{R}^d$ fortgesetzt, dass die Fortsetzung $\bar{f}(t,x)$ global L-stetig ist. Dazu sei für $(t,x) \in I \times \mathbb{R}^d$

$$x_\rho := \rho \frac{x - u(t)}{\|x - u(t)\|} + u(t)\,,$$

und

$$\bar{f}(t,x) := \begin{cases} f(t,x) & , \quad (t,x) \in U_\rho \\ f(t,x_\rho) & , \quad (t,x) \in U_\infty \setminus U_\rho \end{cases}\,.$$

Wegen der Beziehung (Übungsaufgabe)

$$\left| \frac{x}{\|x\|} - \frac{y}{\|y\|} \right| \leq \|x - y\|$$

für $x, y \in \mathbb{R}^d$, mit $\|x\| \geq 1$, $\|y\| \geq 1$ ist offenbar $\bar{f}(t,x)$ auf $I \times \mathbb{R}^d$ gleichmäßig L-stetig (bzgl. x) mit derselben L-Konstante $\hat{L}$. Sei nun $(\bar{y}_n)_n$ die durch ein Einschrittverfahren

$$\bar{y}_n = \bar{y}_{n-1} + h_n \bar{F}(h_n; t_{n-1}, \bar{y}_n, \bar{y}_{n-1})$$

gelieferte diskrete Lösung. Nach Satz 2.2 gilt dann

$$\max_{t_n \in I} \|\bar{y}_n - u(t_n)\| \to 0 \quad (h \to 0)\,.$$

Für hinreichend kleines h ist dann aber $\{(t_n, \bar{y}_n)\,,\ t_0 \leq t_n \leq t_0 + T\} \subset U_\rho$, d. h.: $\bar{y}_n \equiv y_n$.

2.2.2 Globale Konvergenz

Die a priori Abschätzung (2.2.29) liefert eine realistische Fehlerschranke nur auf relativ kleinen Intervallen $I = [t_0, t_0 + T]$; die Größe $\exp(LT)$ wächst extrem schnell mit $T \to \infty$, und die Lipschitz-Kontstante L ist i. Allg. nur sehr grob schätzbar. Es erhebt sich also die Frage, unter welchen Bedingungen eine Abschätzung vom Typ (2.3.46) mit einer von T unabhängigen Konstante gilt.

Wir betrachten zunächst wieder als Modellfall die Eulersche Polygonzugmethode

$$y_n = y_{n-1} + h_n f(t_{n-1}, y_{n-1}) \tag{2.2.30}$$

und wenden diese an auf eine AWA, die im Folgenden Sinne L-stetig und monoton ist:

$$\|f(t,x) - f(t,x')\| \leq L(t)\,\|x - x'\|\,, \tag{2.2.31}$$

$$-(f(t,x) - f(t,x')\,,\ x - x') \geq \lambda(t)\,\|x - x'\|^2\,, \tag{2.2.32}$$

für alle $(t,x), (t,x') \in I \times \mathbb{R}^d$, mit stetigen Funktionen $L(t) \geq 0$, $\lambda(t) \geq 0$. Wir setzen $L := \max_I L(t)$ und $\lambda := \min_I \lambda(t)$. Ist die AWA „homogen", d. h.: $f(t,0) = 0$, so erhält man durch Multiplikation in der Gleichung (2.2.30) mit y_n und Beachtung von

$$2\|y_n\|^2 - 2(y_{n-1}, y_n) = \|y_n\|^2 + \|y_n - y_{n-1}\|^2 - \|y_{n-1}\|^2$$

die Identität

$$\|y_n\|^2 + \|y_n - y_{n-1}\|^2 = \|y_{n-1}\|^2 + 2h_n\big\{(f(t_{n-1}, y_{n-1}), y_{n-1}) + (f(t_{n-1}, y_{n-1}), y_n - y_{n-1})\big\}\,.$$

Unter Ausnutzung der Eigenschaften (2.2.31) und (2.2.32) folgt dann mit den Abkürzungen $L_n := L(t_n)$ und $\lambda_n := \lambda(t_n)$:

$$\begin{aligned}\|y_n\|^2 + \|y_n - y_{n-1}\|^2 &\leq \|y_{n-1}\|^2 - 2\lambda_{n-1}h_n\|y_{n-1}\|^2 + 2h_n L_{n-1}\|y_{n-1}\|\|y_n - y_{n-1}\| \\ &\leq \|y_{n-1}\|^2 - 2\lambda_{n-1}h_n\|y_{n-1}\|^2 + h_n^2 L_{n-1}^2\|y_{n-1}\|^2 + \|y_n - y_{n-1}\|^2,\end{aligned}$$

bzw.

$$\|y_n\|^2 \leq (1 + h_n^2 L_{n-1}^2 - 2\lambda_{n-1} h_n)\|y_{n-1}\|^2 \,. \tag{2.2.33}$$

Die Approximationen y_n bleibt also beschränkt bzgl. n, wenn $1+h_n^2 L_{n-1}^2 - 2\lambda_{n-1} h_n \leq 1$, d. h. wenn die Schrittweiten h_n der folgenden Bedingung genügen:

$$h_n \leq \frac{2\lambda_{n-1}}{L_{n-1}^2} \,, \quad n \geq 1 \,. \tag{2.2.34}$$

Mit Hilfe einer Verfeinerung des obigen Argumentes lässt sich die Stabilitätsaussage (2.2.33) unter der Bedingung, dass die Schrittweitenbedingung (2.2.34) gleichmäßig bzgl. n im *strikten* Sinne erfüllt ist, erweitern zu einer globalen Fehlerabschätzung der Form ($y_0 = u_0$)

$$\|y_n - u(t_n)\| \leq c \max_{1 \leq \nu \leq n} \{h_\nu \max_{I_\nu} \|u''\|\} \,. \tag{2.2.35}$$

Da der Beweis dieser Aussage für die Polygonzugmethode verhältnismäßig kompliziert ist, verzichten wir hier auf die Details und untersuchen lieber die analoge Situation für das implizite Gegenstück. Wir nehmen im Folgenden zur Vereinfachung an, dass stets $y_0 = u_0$.

Satz 2.3 (Globale Konvergenz des impliziten Euler-Verfahrens): *Die AWA sei L-stetig und monoton im Sinne von (2.2.31) und (2.2.32). Dann sind die Lösungen des implizite Euler-Verfahrens*

$$y_n = y_{n-1} + h_n f(t_n, y_n), \quad n \geq 1, \quad y_0 = u_0, \tag{2.2.36}$$

für beliebige Schrittweiten h_n wohl definiert und es gilt die globale Fehlerabschätzung

$$\|y_n - u(t_n)\| \leq \tfrac{1}{2} \min\{t_n - t_0, \lambda^{-1}\} \max_{1 \leq \nu \leq n} \{h_\nu \max_{I_\nu} \|u''\|\}, \quad t_n \geq t_0, \tag{2.2.37}$$

mit der offensichtlichen Interpretation für $\lambda = 0$.

Beweis: (i) In jedem Schritt des impliziten Euler-Verfahrens ist ein Gleichungssystem

$$y_n - h_n f(t_n, y_n) = y_{n-1} \tag{2.2.38}$$

zu lösen. Wegen der angenommenen Eigenschaften (2.2.31) und (2.2.32) ist die Abbildung $g(x) := x - h_n f(t_n, x)$ L-stetig und strikt monoton im Sinne

$$(g(x) - g(y), x - y) \geq \gamma \|x - y\|^2, \quad x, y \in \mathbb{R}^d,$$

mit einer festen Konstante $\gamma > 0$. Mit Korollar 1.7 folgt dann, dass (2.2.38) eine eindeutige Lösung besitzt.
(ii) Für den Fehler $e_n = y_n - u_n$ gilt wieder die Differenzengleichung

$$e_n = e_{n-1} + h_n\{f(t_n, y_n) - f(t_n, u_n)\} - h_n \tau_n,$$

mit dem Abschneidefehler τ_n des impliziten Euler-Verfahrens

$$\|\tau_n\| \leq \tfrac{1}{2} h_n \max_{I_n} \|u''\|.$$

Wir multiplizieren mit $\|e_n\|^{-1} e_n$ und erhalten wieder unter Ausnutzung der Monotonie

$$\|e_n\| \leq \|e_n\|^{-1}(e_{n-1}, e_n) - \lambda_n h_n \|e_n\| + h_n \|e_n\|^{-1}(\tau_n, e_n).$$

Dies ergibt

$$(1 + \lambda_n h_n)\|e_n\| \leq \|e_{n-1}\| + h_n \|\tau_n\|.$$

bzw.

$$\|e_n\| \leq \frac{1}{1 + \lambda_n h_n}\|e_{n-1}\| + \frac{h_n}{1 + \lambda_n h_n}\|\tau_n\|. \tag{2.2.39}$$

(iii) Im Falle $\lambda \geq 0$ (schwache Monotonie) summieren wir (2.2.39) über $\nu = 1, ..., n$ und erhalten unter Beachtung von $e_0 = 0$ die T-abhängige Abschätzung

$$\|e_n\| \leq \sum_{\nu=1}^{n} h_\nu \|\tau_\nu\| \leq \Big(\sum_{\nu=1}^{n} h_\nu \Big) \max_{1 \leq \nu \leq n} \|\tau_\nu\| \leq \tfrac{1}{2}(t_n - t_0) \max_{1 \leq \nu \leq n} \{h_\nu \max_{I_\nu} \|u''\|\}.$$

(iv) Im Falle $\lambda > 0$ (strikte Monotonie) erschließen wir mit Induktion aus (2.2.39) die Abschätzung

$$\|e_n\| \leq \lambda^{-1} \max_{1 \leq \nu \leq n} \|\tau_\nu\|$$

bzw.

$$\|e_n\| \leq \tfrac{1}{2} \lambda^{-1} \max_{1 \leq \nu \leq n} \{h_\nu \max_{I_\nu} \|u''\|\}.$$

Für $n = 1$ gilt trivialerweise

$$\|e_1\| \leq \frac{h_1}{1 + \lambda h_1}\|\tau_1\| \leq \frac{1}{\lambda}\|\tau_1\|.$$

Sei die Behauptung nun richtig für $n - 1$. Dann folgt mit (2.2.39):

$$\begin{aligned} \|e_n\| &\leq \frac{1}{1 + \lambda h_n}\|e_{n-1}\| + \frac{h_n}{1 + \lambda h_n}\|\tau_n\| \\ &\leq \frac{1}{1 + \lambda h_n} \lambda^{-1} \max_{1 \leq \nu \leq n-1} \|\tau_\nu\| + \frac{h_n}{1 + \lambda h_n}\|\tau_n\| \leq \frac{1}{\lambda} \max_{1 \leq \nu \leq n} \|\tau_\nu\|. \end{aligned}$$

Dies vervollständigt den Beweis. Q.E.D.

Die Argumentation im Beweis von Satz 2.3 lässt sich direkt auf solche Einschrittverfahren übertragen, deren Verfahrensfunktion Bedingungen vom Typ (2.2.31) und (2.2.32) genügen. Leider ist dies bei der Approximation monotoner AWAn mit Verfahren höherer Ordnung in der Regel nicht der Fall. Wir werden also einen anderen Zugang zur globalen Fehlerschätzung bei allgemeinen Einschrittverfahren finden müssen.

Der Ausgangspunkt dazu ist die Beobachtung, dass montone, L-stetige AWAn *exponentiell stabile* Lösungen haben. Ein „vernünftiges" Verfahren sollte nun in der Lage sein, jede exponentiell stabile Lösung global zu approximieren, unabhängig davon, ob das Problem selbst monoton ist oder nicht. Die Frage ist also: Lassen sich globale Fehlerabschätzungen der Art (2.2.35) herleiten allein aus der (angenommenen) exponentiellen Stabilität der Lösung $u(t)$ und ohne weitere Voraussetzungen an die Struktur der AWA? Dies ist tatsächlich der Fall, wie der folgende Satz zeigt.

Satz 2.4 (Globale Konvergenz): *Die L-stetige AWA habe eine (globale) exponentiell stabile Lösung $u(t)$ mit Stabilitätsparametern δ, A, α. Für jedes Einschrittverfahren, welches mit der (AWA) konsistent ist und einer Lipschitz-Bedingung genügt, gibt es dann positive Konstanten h_0 und K unabhängig von T, so dass für $h := \sup_I h_n \le h_0$ gilt:*

$$\max_{t_n \in I} \|y_n - u(t_n)\| \le K \max_{t_n \in I} \|\bar\tau_n\| . \tag{2.2.40}$$

Dabei bezeichnet $\bar\tau_n$ das Maximum des Abschneidefehlers für alle möglichen Lösungen der Differentialgleichung, die in der Umgebung $U_{\delta A} = \{(t,x) \in I \times \mathbb{R}^d , \|x - u(t)\| \le \delta A\}$ des Graphen von u verlaufen.

Beweis: Wir führen den Beweis durch Induktion nach t. Sei o.B.d.A. $h \le 1$ angenommen. Zunächst wird ein $\Delta \ge 1$ so gewählt, dass

$$A\, e^{-\alpha(\Delta - h)} \le \tfrac{1}{2}, \quad h \in (0,1]. \tag{2.2.41}$$

Die *lokale* Konvergenzaussage (2.2.29) liefert für alle $t_n \in [t_0, t_0 + \Delta]$ die Abschätzung

$$\|y_n - u_n\| \le \Delta e^{\kappa L \Delta} \max_{t_1 \le t_\nu \le t_n} \|\bar\tau_\nu\| . \tag{2.2.42}$$

Wir setzen nun

$$K := 2\Delta e^{\kappa L \Delta} \tag{2.2.43}$$

und wählen dann h_0 hinreichend klein, so dass für $t_n \ge t_1$ aus

$$\|y_n - u_n\| \le K \max_{t_1 \le t_\nu \le t_n} \|\bar\tau_\nu\| ,$$

für $h \le h_0$, notwendig $\|y_n - u_n\| < \delta$ folgt.

Nach diesen Vorbereitungen sei nun angenommen, dass die Behauptung (2.2.40) richtig ist für alle $t_\nu \in (t_0, t_n]$ mit irgendeinem $t_n \ge t_0 + \Delta$. Dann betrachten wir $w_\star = y_n - u_n$ als Störung von $u(t)$ zum Zeitpunkt $t_\star = t_n$. Für $h \le h_0$ ist automatisch $\|w_\star\| < \delta$, und die Stabilitätseigenschaft von $u(t)$ liefert für die gestörte Lösung $v(t)$ die Abschätzung

$$\|v(t) - u(t)\| \le A\, e^{-\alpha(t - t_n)} \|y_n - u_n\| , \quad t \ge t_n . \tag{2.2.44}$$

Wir fassen nun y_ν für $t_\nu \le t_n$ als Näherung von $v(t_\nu)$ auf und können wegen

$$\|v(t) - u(t)\| \le A\delta , \quad t \ge t_n$$

den lokalen Konvergenzsatz wie folgt anwenden:

$$\max_{t_n \le t_\nu \le t_{n+m}} \|y_\nu - v_\nu\| \le \Delta e^{L\Delta} \max_{t_n \le t_\nu \le t_{n+m}} \|\bar{\tau}_\nu\| , \tag{2.2.45}$$

wobei $m \in \mathbb{N}$, so dass $t_{n+m} \le t_n + \Delta \le t_{n+m+1}$. Dann folgt für $h \le h_0$:

$$\begin{aligned} \|y_{n+m} - u_{n+m}\| &\le \|y_{n+m} - v_{n+m}\| + \|v_{n+m} - u_{n+m}\| \\ &\le \Delta e^{L\Delta} \max_{t_n \le t_\nu \le t_{n+m}} \|\bar{\tau}_\nu\| + A e^{-\alpha(\Delta - h)} \|y_n - u_n\| , \end{aligned}$$

bzw. wegen (2.2.41) und der Induktionsannahme,

$$\|y_{n+m} - u_{n+m}\| \le \left(\Delta e^{L\Delta} + \tfrac{1}{2}K\right) \max_{t_1 \le t_\nu \le t_{n+m}} \|\bar{\tau}_\nu\| .$$

Mit unserer Definition von K ergibt dies

$$\|y_{n+m} - u_{n+m}\| \le K \max_{t_1 \le t_\nu \le t_{n+m}} \|\bar{\tau}_\nu\| ,$$

was den Schluß von t_n nach $t_n + \Delta$ vervollständigt. Q.E.D.

Bemerkung 2.2: In Satz 1.8 wurde gezeigt, dass exponentiell stabile Lösungen L-stetiger, autonomer AWAn stationäre Limiten haben, d. h.: $\lim_{t\to\infty} u(t) = u_\infty$ mit $f(u_\infty) = 0$ und $\|u(t) - u_\infty\| = O(e^{-\alpha t})$ $(t \to \infty)$. Durch Kombination der Argumente im Beweis von Satz 1.8 und Satz 2.4 sollte sich zeigen lassen, dass in diesem Fall auch die durch L-Stetige, konsistente Einschrittverfahren erzeugten Näherungslösungen $(y_n)_{n\in\mathbb{N}}$ (in einem diskreten Sinne) exponentiell stabil sind und gegen stationäre Limiten konvergieren, $\lim_{n\to\infty} y_n = y_\infty$, welche wiederum Approximationen des kontinuierlichen stationären Limes u_∞ sind. Der Beweis für diese naheliegende Aussage ist aber noch nicht geführt worden; s. hierzu auch die Diskussion im nächten Kapitel über „numerische Stabilität".

2.3 Schrittweitenkontrolle

Das Hauptproblem bei der Durchführung von Differenzenverfahren zur Lösung einer AWA ist die Bestimmung geeigneter Schrittweiten h_n zur Gewährleistung einer vorgeschriebenen Approximationsgüte. Die im Konvergenzsatz 2.2 angegebene Fehlerabschätzung erlaubt es, aus Schranken für den „lokalen" Abschneidefehler τ_n auf das Verhalten des „globalen" Diskretisierungsfehlers $e_n = y_n - u(t_n)$ zu schließen. Unter Annahme exakter (d. h. rundungsfehlerfreier) Rechnung gilt bei Verwendung fehlerfreier Startwerte auf dem Intervall $I = [t_0, t_0 + T]$ die *a priori* Fehlerabschätzung

$$\max_{t_n \in I} \|e_n\| \le K \sum_{t_n \in I} h_n \|\tau_n\| \le KT \max_{t_n \in I} \|\tau_n\| , \tag{2.3.46}$$

mit einer Konstante $K = K(T) \approx \exp(LT)$. Obwohl im Extremfall K exponentiell mit der Intervallänge T wächst, nehmen wir im Folgenden an, dass K von moderater Größe ist. Wir haben gesehen, dass diese Annahme z. B.. für monotone AWAn berechtigt ist.

Zur praktischen Auswertung der a priori Fehlerabschätzung (2.3.46) benötigt man möglichst scharfe Schranken für $\|\tau_n\|$. Bei den Taylor-Verfahren m-ter Ordnung gilt z. B..

$$\|\tau_n\| \leq \frac{1}{(m+1)!}\, h^m \max_{t\in I_n} \|u^{m+1}(t)\|\,,$$

und es wären somit höhere Ableitungen der unbekannten exakten Lösung $u(t)$ abzuschätzen. Dies ist aber selbst bei Ausnutzung der Beziehung $u^{(m+1)}(t) = f^{(m)}(t, u(t))$ und Kenntnis von Schranken für $u(t)$ kaum mit vertretbarem Aufwand möglich. Daher werden in der Praxis meist *a posteriori Schätzungen* für die Abschneidefehler verwendet, die man aus den berechneten Näherungswerten für $u(t)$ erhält. Die zugehörige Theorie ist naturgemäß stark heuristisch geprägt. Allgemein für Differenzenverfahren anwendbar ist die sog. „Methode der Schrittweitenhalbierung", die im Folgenden beschrieben wird. Sie entspricht dem üblichen Vorgehen zur Fehlerschätzung bei der numerischen Quadratur. Wir beschränken die Diskussion der Einfachheit halber auf *explizite* Methoden.

Ausgangspunkt ist eine Darstellung des Abschneidefehlers $\tau_n = \tau(t_n)$ auf dem Intervall $[t_{n-1}, t_n]$ zur Schrittweite h_n in der Form

$$\tau_n = \tau^{(m)}(t_n)\, h_n^m + O(h_n^{m+1}) \tag{2.3.47}$$

mit einer von h_n unabhängigen Funktion $\tau^{(m)}(t)$, der sog. „Hauptabschneidefehlerfunktion", und einem Restglied höherer Ordnung. Bei den Taylor-Verfahren ist z. B..

$$\tau^{(m)}(t_n) = \frac{1}{(m+1)!}\, u^{(m+1)}(t_{n-1})\,.$$

Ähnliche Darstellungen gelten auch für die Runge-Kutta-Verfahren (Übungsaufgabe).

Wir wollen nun Strategien angeben, mit deren Hilfe während der Rechnung die Schrittweiten h_n so gewählt werden, dass zu einer vorgegebenen Fehlertoleranz $TOL > 0$ auf dem Intervall I die Schranke

$$\max_{t_n\in I} \|e_n\| \leq TOL\,, \tag{2.3.48}$$

realisiert wird. Die Toleranz TOL sollte dabei deutlich größer als die Maschinengenauigkeit eps gewählt sein, genauer (siehe Übungsaufgabe):

$$TOL > \max_{t_n\in I} \left\{ h_n^{-1}\, \|y_{n-1}\|\, eps \right\}\,.$$

Ausgangspunkt ist die *a priori* Fehlerabschätzung (2.3.46). Wir setzen $K = 1$ und nehmen an, dass Schätzungen für die lokalen Abschneidefehler τ_n bzw. für die Hauptabschneidefehlerfunktion $\tau_n^{(m)} = \tau^{(m)}(t_n)$ bekannt sind. Wie solche zu berechnen sind, wird anschließend diskutiert. Es bieten sich nun zwei Strategien zur Schrittweitensteuerung an:

Strategie I: Die Schrittweiten h_n werden gemäß

$$K h_n^m \|\tau_n^{(m)}\| \approx \frac{TOL}{T} \quad bzw. \quad h_n \approx \Big(\frac{TOL}{KT\|\tau_n^{(m)}\|}\Big)^{1/m}$$

gewählt, so dass wie gewünscht folgt:

$$\max_{t_n\in I}\|e_n\| \approx K\sum_{t_n\in I} h_n\left\{h_n^m\|\tau_n^{(m)}\|\right\} \approx \frac{TOL}{T}\sum_{t_n\in I} h_n = TOL.$$

Die Anzahl der durchzuführenden Zeitschritte ergibt sich dann zu

$$N = \sum_{t_n\in I} h_n h_n^{-1} \approx \sum_{t_n\in I} h_n\Big(\frac{KT\|\tau_n^{(m)}\|}{TOL}\Big)^{1/m} = \Big(\frac{KT}{TOL}\Big)^{1/m}\sum_{t_n\in I} h_n\|\tau_n^{(m)}\|^{1/m}\,.$$

Unter Berücksichtigung der Beziehung $\tau_n^{(m)} \approx u^{(m+1)}(t_{n-1})$ folgt also in etwa, dass

$$N \approx \Big(\frac{KT}{TOL}\Big)^{1/m}\int_I \|u^{(m+1)}\|^{1/m}\,dt\,.$$

Strategie II: Die Schrittweiten h_n werden gemäß

$$Kh_n^{m+1}\|\tau_n^{(m)}\| \approx \frac{TOL}{N} \quad bzw. \quad h_n \approx \Big(\frac{TOL}{KN\|\tau_n^{(m)}\|}\Big)^{1/(m+1)}$$

gewählt mit der (noch unbekannten) Gesamtzahl N der durchzuführenden Zeitschritte, so dass ebenfalls folgt:

$$\max_{t_n\in I}\|e_n\| \approx K\sum_{t_n\in I}\left\{h_n^{m+1}\|\tau_n^{(m)}\|\right\} \approx \frac{TOL}{N}\sum_{t_n\in I} 1 = TOL.$$

Die Anzahl N ergibt sich dann analog wie oben zu

$$N \approx \sum_{t_n\in I} h_n\Big(\frac{KN\|\tau_n^{(m)}\|}{TOL}\Big)^{1/(m+1)} = \Big(\frac{KN}{TOL}\Big)^{1/(m+1)}\sum_{t_n\in I} h_n\|\tau_n^{(m)}\|^{1/(m+1)}\,.$$

Unter Berücksichtigung der Beziehung $\tau_n^{(m)} \approx u^{(m+1)}(t_{n-1})$ ergibt sich diesmal

$$N^{m/(m+1)} = N^{1-1/(m+1)} \approx \Big(\frac{K}{TOL}\Big)^{1/(m+1)}\int_I \|u^{m+1}\|^{1/(m+1)}\,dt\,,$$

und folglich

$$N \approx \Big(\frac{K}{TOL}\Big)^{1/m}\left(\int_I \|u^{(m+1)}\|^{1/(m+1)}dt\right)^{(m+1)/m}.$$

Da N a priori nicht bekannt ist, muss es zunächst geschätzt und dann im Verlaufe von mehreren Durchläufen angepasst werden. Diese Schrittweitenstrategie erscheint also aufwendiger als die erste.

Beide beschriebenen Strategien zur Schrittweitenwahl sind asymptotisch gleich effizient, d. h.: Die globale Fehlertoleranz TOL wird mit $N \approx TOL^{-1/m}$ Zeitschritten

erreicht. Allerdings ergeben sich leichte Unterschiede bei den Konstanten. Wir wollen deren Bedeutung für $m = 1$ (Eulersche Polygonzugmethode) diskutieren. Für Strategie I gilt dann

$$N \approx \frac{KT}{TOL} \int_I \|u''\| \, dt \,,$$

und für Strategie II entsprechend

$$N \approx \frac{K}{TOL} \left(\int_I \|u''\|^{1/2} dt \right)^2 \leq \frac{KT}{TOL} \int_I \|u''\| \, dt \,.$$

Der Unterschied besteht also im wesentlichen darin, wie die Regularität der exakten Lösung in die Schrittzahl eingeht. Strategie II ist hinsichtlich der Anzahl der erzeugten Zeitschritte offenbar dann ökonomischer als Strategie I, wenn

$$\left(\int_I \|u''\|^{1/2} \, dt \right)^2 << T \int_I \|u''\| \, dt \,.$$

Dies ist etwa der Fall für *singuläre* Lösungen, deren zweite Ableitungen nicht integrabel sind; z. B..: $u(t) = (1-t)^{1/2}$.

2.3.1 Schätzung des Abschneidefehlers

Wichtigster Bestandteil der obigen Schrittweitenstrategien sind gute Schätzungen für die Hauptabschneidefehlerfunktion $\tau_n^{(m)}$. Diese kann man etwa mit Hilfe des im Folgenden beschriebenen Prozesses gewinnen. Sei zum Zeitpunkt t_n die Näherung y_n berechnet, so dass

$$\max_{t_\nu \in [t_0, t_n]} \|y_\nu - u(t_\nu)\| \leq TOL \,.$$

Zur Bestimmung von $\tau_{n+1}^{(m)}$ und damit der neuen Schrittweite h_{n+1} wählen wir zunächst eine Schätzschrittweite H (etwa $H = 2h_n$) . Anwendung des Einschrittverfahrens zum Startwert y_n mit den Schrittweiten H (ein Schritt) und $H/2$ (zwei Schritte) ergibt zum vorläufigen Zeitpunkt $t_{n+1} := t_n + H$ Näherungen y_{n+1}^H bzw. $y_{n+1}^{H/2}$. Für die Fehler gilt

$$\begin{aligned} y_{n+1}^H - u(t_{n+1}) &= e_n + H \left\{ F(H; t_n, y_n) - F(H; t_n, u_n) \right\} - H \tau_{n+1}^H \\ &= (1 + O(H)) \, e_n - H^{m+1} \tau_{n+1}^{(m)} + O(H^{m+2}) \end{aligned}$$

sowie analog für $y_{n+1/2}^{H/2} - u(t_{n+1/2})$. Wir erhalten weiter

$$\begin{aligned} y_{n+1}^{H/2} - u(t_{n+1}) &= y_{n+1/2}^{H/2} - u(t_{n+1/2}) + \tfrac{1}{2} H \Big\{ F(\tfrac{1}{2}H; t_{n+1/2}, y_{n+1/2}^{H/2}) \\ &\qquad - F(\tfrac{1}{2}H; t_{n+1/2}, u(t_{n+1/2})) \Big\} - \tfrac{1}{2} H \, \tau_{n+1}^{H/2} \\ &= (1 + O(H)) \left\{ y_{n+1/2}^{H/2} - u(t_{n+1/2}) \right\} - (\tfrac{1}{2}H)^{m+1} \tau_{n+1}^{(m)} + O(H^{m+2}) \\ &= (1 + O(H)) \left\{ (1 + O(H)) e_n - (\tfrac{1}{2}H)^{m+1} \tau_{n+1/2}^{(m)} + O(H^{m+2}) \right\} \\ &\qquad - (\tfrac{1}{2}H)^{m+1} \tau_{n+1}^{(m)} + O(H^{m+2}) \end{aligned}$$

und folglich

$$y_{n+1}^{H/2} - u(t_{n+1}) = (1 + O(H))e_n - 2(\tfrac{1}{2}H)^{m+1}\tau_{n+1}^{(m)} + O(H^{m+2}).$$

Dabei wurde ausgenutzt, dass sich die Hauptabschneidefehlerfunktion gemäß

$$\tau_{n+1/2}^{(m)} = \tau_{n+1}^{(m)} + O(H)$$

entwickeln lässt. Subtraktion dieser beiden Gleichungen ergibt

$$y_{n+1}^{H/2} - y_{n+1}^{H} \;=\; O(H)e_n - \tau_{n+1}^{(m)} \left\{2(\tfrac{1}{2}H)^{m+1} - H^{m+1}\right\} + O(H^{m+2})$$

bzw.

$$\tau_{n+1}^{(m)} = \frac{y_{n+1}^{H/2} - y_{n+1}^{H}}{H^{m+1}(1 - 2^{-m})} + O(H) + O(H^{-m})e_n\,. \tag{2.3.49}$$

Bis hierin war die Analyse noch mathematisch korrekt. Nun wird postuliert, dass die beiden „O"-Terme rechts in (1.2.1) klein genug sind, um mit

$$\tilde{\tau}_{n+1}^{(m)} := \frac{y_{n+1}^{H/2} - y_{n+1}^{H}}{H^{m+1}(1 - 2^{-m})} \tag{2.3.50}$$

eine brauchbare Näherung für $\tau_{n+1}^{(m)}$ zu erhalten. Dazu wird oft $e_n = 0$ angenommen, d. h.: Man betrachtet den Abschneidefehler entlang der diskreten Approximation $(y_n)_n$ anstatt entlang der „richtigen" Lösung $u(t)$. Alternativ kann man sich auch auf die Annahme einer höheren Approximationsordnung $e_n = O(H^{m+1})$ abstützen, was durch die folgende Diskussion nahegelegt wird.

2.3.2 Adaptive Schrittweitensteuerung

Mit der obigen Schätzung für $\tau_{n+1}^{(m)}$ wird nun gemäß einer der oben angegebenen Strategien eine neue Schrittweite h_{n+1} bestimmt, also etwa als (Strategie I):

$$h_{n+1} = \left(\frac{TOL}{KT\|\tilde{\tau}_{n+1}^{(m)}\|}\right)^{1/m}. \tag{2.3.51}$$

Zur Kontrolle wird noch überprüft, dass nicht $h_{n+1} \ll H$, was die Brauchbarkeit der Schätzung $\tilde{\tau}_{n+1}^{(m)}$ in Frage stellen würde. Insgesamt ergibt sich also der folgende Algorithmus zur adaptiven Schrittweitenwahl und Fehlerkontrolle:

(i) Sei die Näherung $y_n \sim u(t_n)$ berechnet, mit der letzten Schrittweite h_n. Wähle $H = 2h_n$ und setze probeweise $t_{n+1} := t_n + H$.

(ii) Berechne y_{n+1}^{H} und $y_{n+1}^{H/2}$, und bestimme die Schätzung des Abschneidefehlers und die daraus resultierende Schrittweite h_{n+1} etwa aus (2.3.51).

(iii) Überprüfe, ob $h_{n+1} \ll \frac{1}{2}H = h_n$ (z. B..: $h_{n+1} \leq \frac{1}{4}H$).

a) Wenn ja: Die Schätzung für $\tau_{n+1}^{(m)}$ ist zu grob. Wiederhole Schritt (i) mit $H = 2h_{n+1}$. (Beende die Rechnung, falls $H < h_{min}$!).

b) Wenn nein: Setze $h_{n+1} = H, t_{n+1} = t_n + H$ und akzeptiere die beste verfügbare Näherung $y_{n+1} := y_{n+1}^{H/2}$ zu $u(t_{n+1})$.

Eine noch bessere Näherung zu $u(t_{n+1})$ erhält man durch eine Linearkombination der beiden Werte $y_{n+1}^{H/2}$ und y_{n+1}^{H} („Prinzip der Extrapolation zum Limes $H = 0$“):

$$u(t_{n+1}) = \frac{2^m y_{n+1}^{H/2} - y_{n+1}^{H}}{2^m - 1} + O(H^{m+1}) .$$

Heuristische Grundlage dieses Schritts ist die postulierte „asymptotische” Entwicklung

$$y_{n+1}^{H} = u(t_{n+1}) + a^m(t_{n+1})H^m + O(H^{m+1}) \tag{2.3.52}$$

mit einer H-unabhängigen Funktion $a^m(t)$. Wir werden uns später noch eingehender mit der Extrapolation bei der Lösung von AWAn befassen.

Bemerkung: Die Schrittweitenkontrolle durch *Schrittweitenhalbierung* ist prinzipiell für jede Einschrittmethode anwendbar. Sie ist orientiert am *lokalen Abschneidefehler*,

$$\tau_n := h_n^{-1}\Big\{u(t_n) - u(t_{n-1})\Big\} - F(h_n; t_{n-1}, u(t_{n-1})) ,$$

den man durch Einsetzen der exakten Lösung u in die Differenzengleichung erhält, und basiert auf der *diskreten* Stabilität des L-stetigen Differenzenoperators. Dieser Ansatz führt zunächst auf *a priori* Fehlerabschätzungen, die erst danach durch Schätzung des Abschneidefehlers τ_n in verwendbare *a posteriori* Fehlerabschätzungen umgewandelt werden. Die Methode zur Schrittweitenwahl durch lokale *Extrapolation* ist im Prinzip auch für *implizite* Einschrittformeln anwendbar (Übungsaufgabe).

Ein alternativer Zugang bedient sich des *Residuums* der diskreten Lösung $(y_n)_n$, welches man durch Einsetzen einer geeigneten Interpolierenden y^h (etwa stückweise linear) von $(y_n)_n$ in die Differentialgleichung erhält:

$$R(y^h) := y^{h\prime} - f(t, y^h) , \quad t \in I .$$

Damit genügt y^h der gestörten Gleichung

$$y^{h\prime} = f(t, y^h) + R(y^h) , \quad t \in I ,$$

und man erhält über die Stabilität des Differentialoperators (Satz 1.4) direkt eine *a posteriori* Abschätzung für den Fehler $e := y^h - u$ durch das bekannte Residuum $R(y^h)$:

$$\max_{t\in I} \|y^h(t) - u(t)\| \leq e^{L_f T}\left\{\|y_0^h - u_0\| + T \max_{t\in I} \|R(y^h)\|\right\} . \tag{2.3.53}$$

Hierbei besteht aber das Problem, dass unter Umständen, insbesondere bei Verfahren höherer Ordnung, das heuristisch gebildete Residuum nicht mit der richtigen Ordnung gegen Null geht und der Fehler somit grob überschätzt wird. Diesen Zugang zur Fehlerschätzung werden wir später im Zusammenhang mit den sog. *Galerkin-Verfahren* zur Lösung von AWAn weiter verfolgen.

Bemerkung: Die kritische Schwäche der allgemeinen heuristischen Schrittweitenkontrolle für das implizite Euler-Verfahren basierend auf der a priori Fehlerabschätzung (2.3.46) ist die möglicherweise starke Unterschätzung der Fehlerkonstante K, wenn sie einfach willkürlich gesetzt wird. Auf der anderen Seite orientieren sich analytische a priori Abschätzungen von K zwangsläufig am schlimmsten Fall und führen zu grober Überschätzung des tatsächlichen Fehlers und damit zu ineffizienter Schrittweitenkontrolle. Ein Ansatz zur möglichen Überwindung dieses Problems basiert auf der Beziehung

$$e_n = e_{n-1} + h_n f'(t_n, y_n)e_n + h_n\tau_n(u) + h_n O(e_n^2), \tag{2.3.54}$$

für den Fehler $e_n = u_n - y_n$, mit Anfangswert $e_0 = 0$. Mit einer Schätzung des Abschneidefehlers $\tau_n(y_n) \approx \tau_n(u)$ (erhalten etwa mit Hilfe lokaler Extrapolation) kann die Lösung E_n der linearisierten Fehlergleichung

$$E_n = E_{n-1} + h_n f'(t_n, y_n)E_n + h_n\tau_n(y_n), \quad 0 \le n \le N, \tag{2.3.55}$$

verwendet werden, um eine Schätzung für den Fehler $E_n \approx e_n$ zu gewinnen.

2.3.3 Numerischer Test

Für die AWA

$$u'(t) = -200\,t\,u(t)^2\,, \quad t \ge -3\,, \quad u(-3) = 1/901\,,$$

mit der Lösung $u(t) = (1 + 100\,t^2)^{-1}$ wurde der Wert $u(0) = 1$ approximiert mit

- dem Runge-Kutta-Verfahren 2. Ordnung

$$y_n = y_{n-1} + \tfrac{1}{2}h_n\{k_1 + k_2\}\,, \quad k_1 = f(t_{n-1}, y_{n-1})\,, \quad k_2 = f(t_n\,, y_{n-1} + h_n k_1)\,.$$

- mit dem „klassischen" Runge-Kutta-Verfahren 4. Ordnung

$$\begin{aligned} y_n &= y_{n-1} + \tfrac{1}{6}h_n\{k_1 + 2k_2 + 2k_3 + k_4\},\\ k_1 &= f(t_{n-1}, y_{n-1}), \quad k_2 = f(t_{n-1/2}, y_{n-1} + \tfrac{1}{2}h_n k_1),\\ k_3 &= f(t_{n-1/2}, y_{n-1} + \tfrac{1}{2}h_n k_2), \quad k_4 = f(t_n, y_{n-1} + h_n k_3). \end{aligned}$$

Die Schrittweitensteuerung erfolgte dabei gemäß der obigen Strategie

$$h_{n+1}^m \frac{y_{n+1}^{H/2} - y_{n+1}^H}{H^{m+1}(1 - 2^{-m})} \sim TOL = \mathrm{eps}\,\frac{|y_n|}{h_{n+1}}\,.$$

Bei 17-stelliger Rechnung ergaben sich folgende Resultate:

Rechnung mit variabler Schrittweite

Ordnung	eps	$h_{\min}$	$h_{\max}$	Fehler	# Auswertungen
$m=2$	10^{-9}	$2,5\cdot 10^{-4}$	$3,8\cdot 10^{-3}$	$1,3\cdot 10^{-6}$	~ 16.000
	10^{-13}	$7,3\cdot 10^{-6}$	$1,2\cdot 10^{-4}$	$2,7\cdot 10^{-6}$	~ 384.000
$m=4$	10^{-9}	$6,6\cdot 10^{-4}$	$1,0\cdot 10^{-1}$	$2,9\cdot 10^{-6}$	~ 1.200
	10^{-17}	$1,9\cdot 10^{-4}$	$2,9\cdot 10^{-3}$	$1,7\cdot 10^{-10}$	~ 2.000

Rechnung mit fester Schrittweite

Ordnung	h	Fehler	# Auswertungen
$m=2$	$5\cdot 10^{-5}$	$3\cdot 10^{-6}$	~ 120.000
$m=4$	$5\cdot 10^{-3}$	$3\cdot 10^{-6}$	~ 2.000

2.4 Übungsaufgaben

Aufgabe 2.1: a) Man rekapituliere den Begriff der „Konsistenz“ und den der „Konsistenzordnung“ einer (expliziten) Einschrittformel $y_n = y_{n-1} + hF(h; t_{n-1}, y_{n-1})$ zur Approximation der Differentialgleichung $u'(t) = f(t, u(t))$.

b) Man gebe die Konsistenzordnungen der folgenden Differenzenformeln an:

(i) Modifizierte Euler-Formel:

$$y_n = y_{n-1} + hf\big(t_{n-1} + \tfrac{1}{2}h, y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1})\big);$$

(ii) 3-stufige Runge-Kutta-Formel:

$$y_n = y_{n-1} + \tfrac{1}{10}h\{k_1 + 5k_2 + 4k_3\}, \quad k_1 = f(t_{n-1}, y_{n-1}),$$
$$k_2 = f\big(t_{n-1} + \tfrac{1}{3}h, y_{n-1} + \tfrac{1}{3}hk_1\big), \quad k_3 = f\big(t_{n-1} + \tfrac{5}{6}h, y_{n-1} - \tfrac{5}{12}hk_1 + \tfrac{5}{4}hk_2\big).$$

Aufgabe 2.2: Das allgemeine (explizite oder implizite) Runge-Kutta-Verfahren hat die Form

$$y_n = y_{n-1} + h_n F(h_n; t_{n-1}, y_{n-1})$$

mit der Verfahrensfunktion

$$F(h; t, x) = \sum_{r=1}^{R} c_r k_r(h; t, x), \quad k_r(h; t, x) = f\Big(t + ha_r, x + h\sum_{s=1}^{R} b_{rs}k_s\Big),$$

mit Konstanten c_r, a_r, b_{rs}. Im Fall $b_{rs} = 0$ für $s \geq r$ ist das Schema explizit. Man zeige:

a) Dieses Verfahren genügt der Lipschitz-Bedingung (L_h)

$$\|F(h;t,x) - F(h;t,\tilde{x})\| \le L\|x - \tilde{x}\|,$$

wenn die Funktion $f(t,x)$ (bzgl. x) Lipschitz-stetig ist.

b) Das Verfahren ist genau dann konsistent, wenn $\sum_{r=1}^{R} c_r = 1$ ist.

Aufgabe 2.3: Bei der Durchführung einer expliziten (L-stetigen) Einschrittmethode wird wegen des unvermeidbaren Rundungsfehlers eine gestörte Rekursion

$$\tilde{y}_n = \tilde{y}_{n-1} + h_n F(h_n; t_{n-1}, \tilde{y}_{n-1}) + \varepsilon_n, \quad n \ge 1,$$

gelöst. Die „lokalen" Fehler verhalten sich dabei wie $\|\varepsilon_n\| \sim \text{eps}\,\|y_n\|$, wobei eps die sog. „Maschinengenauigkeit" (maximaler relativer Rundungsfehler) bezeichnet. Man beweise die Abschätzung (Stabilitätssatz)

$$\|\tilde{y}_n - u(t_n)\| \le K(t_n)\Big\{\|\tilde{y}_0 - u_0\| + (t_n - t_0)\max_{1\le m\le n}\|\tau_m\| + \text{eps}\max_{1\le m\le n} h_m^{-1}\|y_m\|\Big\},$$

wobei τ_m den Abschneidefehler der Differenzenformel bezeichnet.

Bemerkung: Dies zeigt, dass bei einer Verkleinerung der Schrittweiten h_n über eine gewisse Grenze hinaus der Gesamtfehler wieder anwachsen wird. Ferner wird die Wahl der Fehlertoleranz $\varepsilon \sim \text{eps}\,\|y_n\|/h_n$ bei der automatischen a posteriori Schrittweitenkontrolle nahegelegt.

Aufgabe 2.4: In vielen Fällen kann die Konvergenzordnung eines Grenzprozesses

$$a(h) \to a \ \ (h \to 0), \quad a(h) - a = O(h^\alpha),$$

nur experimentell bestimmt werden. Dazu werden bei bekanntem Limes a für zwei Werte h und $h/2$ die Fehler $a(h)-a$ und $a(h/2)-a$ berechnet und dann die Ordnung α über den formalen Ansatz $a(h) - a = ch^\alpha$ aus der folgenden Formel ermittelt:

$$\alpha = \frac{1}{\log(2)} \log\left(\left|\frac{a(h)-a}{a(h/2)-a}\right|\right).$$

a) Man rekapituliere die Rechtfertigung dieser Formel und überlege, wie man vorgehen könnte, wenn kein exakter Limes a bekannt ist.

b) Man bestimme die inhärenten Konvergenzordnungen für die folgenden Werte:

h	$a(h)$	$b(h)$
2^{-1}	7.188270827204928	8.89271737217539
2^{-2}	7.095485351135761	8.971800326329658
2^{-3}	7.047858597600531	8.992881146463981
2^{-4}	7.023726226390662	8.998220339291473
2^{-5}	7.011579000356371	8.999559782988968
2^{-5}	7.005485409034109	8.999895247704067
Limes	$a(0) = 7.0$	$b(0) =?$

Aufgabe 2.5 (Praktische Aufgabe):
Man berechne Näherungslösungen für die AWA

$$u'(t) = \sin(u(t)), \quad t \geq 0, \quad u(0) = 1,$$

mit Hilfe

- der Polygonzugmethode,
- der modifizierten Euler-Formel,
- des „klassischen“ Runge-Kutta-Verfahrens 4-ter Ordnung,

jeweils für die (konstanten) Schrittweiten $h_i = 2^{-i}, i = 1, ..., 8$.

Man bestimme „experimentell“ die Konvergenzordnungen p der Verfahren für die Approximation des Lösungswertes $u(10)$ aus den berechneten Näherungen $y_N^{(i)} \sim u(10)$ zu Schrittweiten h_i nach der Formel

$$p = -\frac{1}{\log(2)} \log\Big(\frac{y_N^i - y_N^{i-1}}{y_N^{i-1} - y_N^{i-2}}\Big).$$

Aufgabe 2.6: Man betrachte das implizite Euler-Schema

$$y_n = y_{n-1} + h_n f(t_n, y_n), \quad t_n \geq t_0, \quad y_0 \approx u_0,$$

zur Diskretisierung der üblichen L-stetigen AWA $u'(t) = f(t, u(t)), t \geq t_0, u(t_0) = u_0$. Man beweise mit den Mitteln aus dem Text unter der Annahme einer geeigneten Schrittweitenbedingung die a priori Fehlerabschätzung (mit einem geeigneten $\gamma > 0$)

$$\|y_n - u(t_n)\| \leq e^{\gamma L(t_n - t_0)} \Big\{ \|y_0 - u_0\| + \tfrac{1}{2} T \max_{1 \leq m \leq n} \big\{ h_m \max_{t \in [t_{m-1}, t_m]} \|u''(t)\| \big\} \Big\}.$$

Aufgabe 2.7: (i) Man beweise zunächst die beiden Abschätzungen

$$\Big\| \frac{x}{\|x\|} - \frac{y}{\|y\|} \Big\| \le \|x-y\|, \qquad \Big\| \frac{x}{\|x\|} - z \Big\| \le \|x-z\|,$$

für beliebige Vektoren $x, y, z \in \mathbb{R}^d$ mit $\|x\| \ge 1$, $\|y\| \ge 1$, $\|z\| \le 1$. (Hinweis: Zum Nachweis dieser Abschätzungen darf „geometrisch“ argumentiert werden; analytische Beweise sind aber auch willkommen.)

(ii) Die Funktion $f(t,x)$ genüge für ein $\rho > 0$ auf dem „Schlauch“

$$U_\rho := \big\{ (t,x) \in I \times \mathbb{R}^d \,\big|\, \|x - u(t)\| \le \rho \big\}, \quad I = [t_0, t_0 + T],$$

um die Lösung $u(t)$ der zugehörigen AWA $u'(t) = f(t,u(t)),\ t \in I,\ u(t_0) = u_0$, der üblichen Lipschitz-Bedingung mit Konstante L_f. Für $(t,x) \in (I \times \mathbb{R}^d) \setminus U_\rho$ sei gesetzt

$$x_\rho := \rho \, \frac{x - u(t)}{\|x - u(t)\|} + u(t),$$

so dass stets $(t, x_\rho) \in U_\rho$ ist. Man zeige, dass dann die modifizierte Funktion

$$\tilde{f}(t,x) := \begin{cases} f(t,x), & (t,x) \in U_\rho, \\ f(t,x_\rho), & (t,x) \in (I \times \mathbb{R}^d) \setminus U_\rho, \end{cases}$$

auf ganz $I \times \mathbb{R}^d$ stetig und sogar global Lipschitz-stetig ist mit derselben Konstante L_f.

Aufgabe 2.8: Die (nicht-autonome) AWA $u'(t) = f(t,u(t)),\ t \ge t_0,\quad u(t_0) = u_0$, genüge der üblichen (globalen) Lipschitz- und Monotonie-Bedingung. Man zeige mit den Argumenten aus dem Text, dass dann die implizite Euler-Methode

$$y_n = y_{n-1} + h_n f(t_n, y_n),\ \ n \ge 1, \quad y_0 = u_0,$$

ohne Schrittweitenrestriktion Näherungen y_n liefert, welche im Fall $\sup_{t \ge t_0} \|f(t,0)\| \le M$ beschränkt bleiben:

$$\sup_{n \ge 1} \|y_n\| \le M'.$$

(Hinweis: Man passe die Argumentation im Text zum Nachweis der globalen a priori Fehlerabschätzung für das implizite Euler-Verfahren an die vorliegende Fragestellung an, ohne letzteres Resultat explizit zu verwenden.)

Aufgabe 2.9: Die Durchführung eines impliziten, L-stetigen Einschrittverfahrens

$$y_n = y_{n-1} + h_n F(h_n; t_n, y_n, y_{n-1}),\ \ n \ge 1, \quad y_0 = u_0,$$

zur Lösung der L-stetigen AWA $u'(t) = f(t,u(t)),\ t \ge 0,\ u(0) = u_0$, der Dimension d erfordert in jedem Zeitschritt die Lösung eines i. Allg. nichtlinearen Gleichungssystems.

a) Man formuliere (mit Begründung) eine Bedingung an die Zeitschrittweiten h_n, unter der die Konvergenz der einfachen Fixpunktiteration

$$y_n^{(k)} = y_{n-1} + h_n F(h_n; t_n, y_n^{(k-1)}, y_{n-1}), \quad k \geq 1, \quad y_n^{(0)} = y_{n-1},$$

gesichert ist, und gebe eine Fehlerabschätzung für diese Fixpunktiteration an.

b) Unter welcher Zusatzbedingung an die Verfahrensfunktion $F(h_n; t_n, y_n, y_{n-1})$ ist die Lösbarkeit des nichtlinearen Gleichungssystems unabhängig von der Wahl der Schrittweite h_n stets gesichert?

c) Wie lautet das Newton-Verfahren zur Lösung dieses Gleichungssystems und unter welchen Bedingungen konvergiert es quadratisch (s. Literatur)?

Aufgabe 2.10 (Praktische Aufgabe):
a) Man berechne Näherungslösungen für die AWA

$$u'(t) = -200\, t\, u(t)^2, \quad t_0 := -3 \leq t \leq 3, \quad u(-3) = \frac{1}{901},$$

mit Hilfe der expliziten „Polygonzugmethode“

$$y_n = y_{n-1} + h f(t_{n-1}, y_{n-1}), \quad n = 1, \ldots, N := 4/h,$$

für die (konstanten) Schrittweiten $h = 2^{-i}, i = 5, \ldots, 10$. Man vergleiche die berechneten Werte zum Zeitpunkt $t = 1$ mit dem Wert $u(1)$ der exakten Lösung $u(t) = (1+100t^2)^{-1}$ in einem logarithmischen Plot (Logarithmus des absoluten Fehlers als Funktion von h bzw. $i = 0, 1, 2, \ldots$).

b) Man wiederhole die Rechnung mit der sog. „(impliziten) Trapezregel“

$$y_n = y_{n-1} + \frac{1}{2} h \big\{ f(t_n, y_n) + f(t_{n-1}, y_{n-1}) \big\}, \quad n = 1, \ldots, N := 6/h,$$

und vergleiche die beobachteten „Konvergenzordnungen“:

$$|u(3) - y_N| = \mathcal{O}(h^p).$$

Wie verhält sich das dieses Verfahren für die gröbere Schrittweite $h = 2^{-4}$? Man versuche, den beobachteten Effekt zu erklären.

c) Man untersuche die Konvergenz der folgenden, aus den mit der Trapezregel gewonnenen Werte $y_N^{(i)}$ zur Schrittweite h_i gebildeten Approximationen

$$\tilde{y}_N^{(i)} := \frac{1}{3} \big\{ 4 y_N^{(i)} - y_N^{(i-1)} \big\}, \quad i = 2, ..., 8.$$

Die beobachteten Phänomene werden im Verlauf des Textes erklärt werden.

Aufgabe 2.11: Betrachtet werde die lineare AWA

$$u'(t) = Au(t) + b(t), \quad t \geq 0, \quad u(0) = u_0,$$

mit der Matrix $A = (a_{ij})_{i,j=1}^{d}$ mit den Elementen

$$a_{ii} = -2, \;\; a_{i,i\pm1} = 1, \;\; a_{ij} = 0 \;\text{ sonst},$$

und der Vektorfunktion $b(t) = (b_j(t))_{j=1}^{d}$ mit den Komponenten $b_j(t) = \sin(j\pi\, t)$. Lässt sich die (globale) Lösung $u(t)$ dieser AWA mit Hilfe der Trapezregel

$$y_n = y_{n-1} + \tfrac{1}{2}h_n\big\{Ay_{n-1} + b_{n-1} + Ay_n + b_n\big\},$$

gleichmäßig in der Zeit approximieren, d. h.: Gilt unter einer globalen Schrittweitenbedingung $h_n \leq h$ bei hinreichend kleinem h eine globale Fehlerabschätzung der Form

$$\sup_{t_n \geq 0} \|y_n - u_n\| \leq Kh^2,$$

und wie sieht dabei die Fehlerkonstante K aus?

Aufgabe 2.12: Man zeige für (global) L-stetige und (strikt) monotone AWAn im Sinne des Textes unter der Schrittweitenbedingung

$$h = \sup_{n\geq 1} h_n < \frac{2\lambda}{L^2},$$

dass das explizite Euler-Verfahren, $y_n = y_{n-1} + h_n f(t_{n-1}, y_{n-1})$, $n \geq 1$, $y_0 = u_0$, global konvergiert, d. h. es gilt eine globale Fehlerabschätzung der Form

$$\|y_n - u(t_n)\| \leq c \max_{1\leq\nu\leq n} \{h_\nu \max_{I_\nu} \|u''\|\}, \quad t_n \geq t_0.$$

Hinweis: Man versuche (angelehnt an den Beweis der globalen Konvergenz des impliziten Euler-Verfahrens aus dem Text) die Abschätzung ($\kappa = 2\lambda - hL^2$)

$$\|e_n\|^2 \leq \frac{8}{\kappa^2} \max_{1\leq\nu\leq n} \|\tau_\nu\|^2 \quad (\text{für } \kappa h < 1).$$

induktiv zu beweisen. Hierbei könnten sich die Beziehung $2\|e_n\|^2 - 2(e_{n-1}, e_n) = \|e_n\|^2 + \|e_n - e_{n-1}\|^2 - \|e_{n-1}\|^2$ und die Youngsche Ungleichung, $2ab \leq \varepsilon^{-1}a^2 + \varepsilon b^2$, als nützlich erweisen.

Bemerkung: Diese Aussage folgt auch aus dem „globalen“ Konvergenzsatz des Textes, da unter den gestellten Bedingungen die Lösung der AWA exponentiell stabil ist. Dies zeigt die Leistungsfähigkeit dieses allgemeinen Satzes.

Aufgabe 2.13: Man zeige exemplarisch für das Heunsche Verfahren 2-ter Ordnung

$$y_n = y_{n-1} + \tfrac{1}{2}h_n\big\{f(t_{n-1}, y_{n-1}) + f(t_n, y_{n-1} + h_n f(t_{n-1}, y_{n-1}))\big\},$$

dass der Abschneidefehler einer expliziten Runge-Kutta-Formel m-ter Ordnung eine Darstellung der Form

$$\tau_n = \tau^{(m)}(t_n)h_n^m + \mathcal{O}(h_n^{m+1})$$

erlaubt, wobei die sog. „führende Abschneidefunktion“ $\tau^{(m)}(t)$ nicht von h_n abhängt. (Hinweis: Man treibe einfach die zur der Ermittlung der Konsistenzordnung der Differenzenformel angesetzten Taylor-Entwicklungen um eine Stufe weiter.)

Aufgabe 2.14: Man rekapituliere die im Text angegebene „Methode der Schrittweitenhalbierung“ zur Schätzung des Abschneidefehlers *expliziter* Einschrittverfahren und beantworte dabei die folgenden Fragen:

(i) Wie lauten die Formeln, wenn statt mit „Schrittweitenhalbierung“ mit „Schrittweitenviertelung“ gearbeitet wird?

(ii) Ist diese Methode auch für *implizite* Einschrittverfahren

$$y_n = y_{n-1} + h_n F(h_n; t_{n-1}, y_{n-1}, y_n)$$

mit L-stetiger Verfahrensfunktion $F(h; t, \cdot, \cdot)$ anwendbar?

Aufgabe 2.15: Die kritische Schwäche der allgemeinen heuristischen Schrittweitenkontrolle für das implizite Euler-Verfahren basierend auf der allgemeinen lokalen a priori Fehlerabschätzung aus dem Text ist die möglicherweise starke Unterschätzung der Fehlerkonstante K, wenn sie einfach willkürlich gesetzt wird. Auf der anderen Seite orientieren sich analytische a priori Abschätzungen von K zwangsläufig am schlimmsten Fall und führen zu grober Überschätzung des tatsächlichen Fehlers und damit zu ineffizienter Schrittweitenkontrolle.

Ein Ansatz zur möglichen Überwindung dieses Problems basiert auf der Beziehung

$$e_n = e_{n-1} + h_n f'(t_n, y_n)e_n + h_n\tau_n(u) + h_n\mathcal{O}(e_n^2),$$

für den Fehler $e_n = u_n - y_n$, mit Anfangswert $e_0 = 0$. Mit einer Schätzung des Abschneidefehlers $\tau_n(y_n) \approx \tau_n(u)$ (erhalten etwa mit Hilfe lokaler Extrapolation) kann die Lösung E_n der linearisierten Fehlergleichung

$$E_n = E_{n-1} + h_n f'(t_n, y_n)E_n + h_n\tau_n(y_n), \quad 0 \le n \le N,$$

verwendet werden, um eine Schätzung für den Fehler $E_n \approx e_n$ zu gewinnen. Man zeige, dass für diese Schätzung gilt:

$$\max_{0\le n\le N} \|e_n - E_n\| = \mathcal{O}\big(\max_{0\le n\le N} \|e_n\|^2\big).$$

Aufgabe 2.16 (Praktische Aufgabe):
Man berechne eine Näherungslösung für die AWA

$$u'(t) = -200\,t\,u(t)^2, \quad t \ge -3, \quad u(-3) = \frac{1}{901},$$

auf dem Intervall $I = [-3, 3]$ mit Hilfe der Heunschen Formel 2. Ordnung unter Verwendung der Strategie zur Schrittweitensteuerung aus dem Text. Als angestrebte Fehlertoleranz wähle man $\varepsilon = 10^{-5}$, als Fehlerkonstante $K = 10$ und als Startschrittweite $h_0 = 10^{-2}$. Man beurteile die Güte der Schrittweitensteuerung durch Vergleich mit der exakten Lösung

$$u(t) = \frac{1}{1 + 100t^2}.$$

Für welche konstante Schrittweite würde man dieselbe Genauigkeit erzielen, und wieviele Funktionsauswertungen $f(t, x)$ sind jeweils erforderlich?