A Lösungen der Übungsaufgaben

Im Folgenden sind Lösungen für die am Ende der einzelnen Kapitel formulierten Übungsaufgaben zusammengestellt. Es handelt sich dabei nur um Lösungsvorschäge ohne Anspruch auf Vollständigkeit zur Anregung weiterer eigener Überlegungen.

A.1 Kapitel 1

Lösung A.1.1: a) Wir führen die Bezeichnungen $u_1 := v$, $u_2 := v'$, $u_3 := v''$, $u_4 := v'''$, $u_5 := u$, $u_6 := u'$, ein und erhalten das System

$$\begin{aligned} &u_1'(x) = u_2(x), \quad u_2'(x) = u_3(x), \quad u_3'(x) = u_4(x), \\ &u_4'(x) = f(x) + a(x)u_6(x), \quad u_5'(x) = u_6(x), \quad u_6'(x) = g(x) - b(x)u_1(x). \end{aligned}$$

b) Zunächst wird $u'' = g - bv$ eingesetzt und dann analog zu (a) verfahren.

Lösung A.1.2: a) Sei u Lösung der Anfangswertaufgabe

$$u'(t) = f(t, u(t)), \quad 0 \le t \le T, \quad u(0) = u_0.$$

Durch Integration über $[0, t]$ erhalten wir

$$\int_0^t u'(s)\, ds = \int_0^t f(s, u(s))\, ds,$$

und somit

$$u(t) = u(0) + \int_0^t f(s, u(s))\, ds = u_0 + \int_0^t f(s, u(s))\, ds.$$

Sei nun u Lösung der Integralgleichung

$$u(t) = u_0 + \int_0^t f(s, u(s))\, ds, \quad 0 \le t \le T.$$

Dann ist zunächst

$$u(0) = u_0 + \int_0^0 f(s, u(s))\, ds = u_0.$$

Außerdem ist u nach dem Hauptsatz der Differential- und Integralrechnung stetig differenzierbar mit

$$u'(t) = \frac{d}{dt} \int_0^t f(s, u(s))\, ds = f(t, u(t)).$$

b) Für den Integraloperator $K : C[0,T] \to C[0,T]$ gilt:

$$\begin{aligned}\|K(v)(t) - K(u)(t)\| &= \Big\|u_0 + \int_0^t f(s,v(s))\,ds - u_0 - \int_0^t f(s,u(s))\,ds\Big\| \\ &= \Big\| \int_0^t \big(f(s,v(s)) - f(s,u(s))\big)\,ds\Big\| \\ &\le \int_0^t \|f(s,v(s)) - f(s,u(s))\|\,ds \\ &\le \int_0^T \|f(s,v(s)) - f(s,u(s))\|\,ds.\end{aligned}$$

Unter Ausnutzung der L-Stetigkeit von $f(t,\cdot)$ ergibt sich

$$\begin{aligned}\|K(v)(t) - K(u)(t)\| &\le \int_0^t L\|v(s) - u(s)\|\,ds \\ &\le L\int_0^T \max_{\tau\in[0,T]} \|v(\tau) - u(\tau)\|\,ds = LT\|v-u\|_\infty.\end{aligned}$$

Da diese Abschätzung für jedes $t \in [0,T]$ gilt, folgt

$$\|K(v) - K(u)\|_\infty \le LT\|v-u\|_\infty.$$

Also ist K im Falle $q := LT < 1$ eine Kontraktion, so dass der Banachsche Fixpunktsatz anwendbar ist.

c) Es gilt die a priori Fehlerabschätzung

$$\begin{aligned}\|u^k - u\|_\infty &\le \frac{q^k}{1-q}\|u^1 - u^0\|_\infty \\ &= \frac{q^k}{1-q}\max_{t\in[0,T]}\Big\|\int_0^t f(s,u_0)\,ds\Big\| \le \frac{q^k}{1-q}\,T\|f(\cdot,u_0)\|_\infty\end{aligned}$$

und die a posteriori Fehlerabschätzung

$$\|u^k - u\|_\infty \le \frac{q}{1-q}\,\|u^k - u^{k-1}\|_\infty.$$

Lösung A.1.3: Wir schreiben die AWA in der Form $u'(t) = f(t,u(t))$. Alle Funktionen f sind stetig, es existiert also jeweils (mindestens) eine lokale Lösung.

a) Eine Lösung ist (nach Text) $u(t) = (1-t)^{-1}$. Wegen der (lokalen) L-Stetigkeit von f ist diese im Existenzintervall eindeutig. Sie wird jedoch für $t \to 1$ singulär. Gäbe es eine Lösung auch für $t \ge 1$, so müsste diese für $t < 1$ mit $(1-t)^{-1}$ übereinstimmen. Die Lösung existiert also nicht global und ist auch nicht gleichmäßig beschränkt.

b) Wegen der lokalen L-Stetigkeit der Funktion $f(t,x) = -x^2$ esistiert eine lokale, eindeutige Lösung. Dies ist gerade $u(t) = (1+t)^{-1}$. Diese Lösung ist global fortsetzbar und beschränkt.

c) Eine lokale Lösung existiert nach dem Existenzsatz von Peano. Für diese gilt wegen $u'(t) = u(t)^{1/2} \geq 0$ für alle t aus ihrem Existenzintervall notwendig $u(t) \geq 1$. Wegen der L-Stetigkeit der Funktion $f(x) = x^{1/2}$ für $x \geq 1$ ist diese Lösung also eindeutig. Diese Lösung hat die explizite Gestalt

$$u(t) = (1 + \tfrac{1}{2}t)^2$$

und ist offenbar „global“, d.h. für alle $t \geq 0$ definiert, aber nicht beschränkt.
Bemerkung: Für den Anfangswert $u(0) = 0$ gäbe es dagegen die unendlich vielen Lösungen

$$u_c(t) = \begin{cases} 0, & 0 \leq t \leq c \\ \left[\frac{1}{2}(t-c)\right]^2, & c < t. \end{cases}$$

d) Hier ist $f(t,x) = \cos(x) - 2x$. Wir können den globalen Existenzsatz aus dem Text anwenden:

$$|f(t,x)| = |\cos(x) - 2x| \leq |\cos(x)| + 2|x| \leq 2|x| + 1.$$

Außerdem ist f global L-stetig:

$$\begin{aligned} |f(t,x) - f(t,y)| &= |\cos(x) - \cos(y) + 2y - 2x| \\ &\leq |\cos(x) - \cos(y)| + 2|x-y| \\ &= |\sin(\xi)(x-y)| + 2|x-y| \qquad (\xi \in [x,y]) \\ &\leq |x-y| + 2|x-y| = 3|x-y|. \end{aligned}$$

Also ist die AWA global eindeutig lösbar. Da f außerdem der Monotoniebedingung genügt und $|f(t,0)| = 1 < \infty$ ist, folgt die Beschränktheit (sowie die exponentielle Stabilität) nach dem globalen Stabilitätssatz.

Lösung A.1.4: a) Angenommen, die AWA besitze eine eindeutige Lösung u, aber nicht die gesamte Folge $(u_h)_h$ konvergiere gegen diese. Dann existiert eine Teilfolge $(u_{h'})$ von (u_h), die nicht gegen u konvergiert und für die u auch kein Häufungspunkt ist. Auf diese Teilfolge wenden wir den Satz von Arzela-Ascoli an. Dieser liefert die Existenz einer weiteren Teilfolge, die gegen eine Lösung v konvergiert. Da nach Voraussetzung $u \neq v$ gilt, u aber die einzige Lösung der AWA sein soll, ergibt sich ein Widerspruch.

b) Der Satz von Peano und der Fortsetzungssatz behalten ihre Gültigkeit: Wir unterteilen das Zeitintervall so, dass die (endlich vielen) Unstetigkeitsstellen mit Stützstellen zusammenfallen. Danach wenden wir den Satz sukzessive für jedes Teilintervall an. Die Anfangswerte auf den Teilintervallen sind gerade die Endwerte der vorangegangenen, vorausgesetzt diese existieren überhaupt (wegen der unter Umständen nur lokalen Existenz des vorhergehenden Lösungsstückes).

Lösung A.1.5: Zum Beweis schreiben wir für jede einzelne Komponente von $f(t,x)$:

$$f_i(t,x) - f_i(t,x') = \int_0^1 \partial_s f_i(t, x' + s(x-x'))\, ds = \int_0^1 \sum_{j=1}^{d} \partial_j f_i(t, x' + s(x-x'))(x_j - x'_j)\, ds$$

und finden durch Normbildung und Ausnutzen der Verträglichkeit von euklidischer Norm und Frobenius-Norm:

$$\begin{aligned}\|f(t,x)-f(t,x')\| &\le \int_0^1 \| \sum_{j=1}^{d} \partial_j f(t,x'+s(x-x'))(x_j-x'_j) \| \, ds \\ &\le \int_0^1 \Big[\sum_{i,j=1}^{d} |\partial_j f_i(t,x'+s(x-x'))|^2 \Big]^{1/2} \|x-x'\| \, ds \\ &\le 1\, K d \, \|x-x'\| \, .\end{aligned}$$

Lösung A.1.6: Eine Differentialgleichung

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

heißt (stark) „monoton“, wenn gilt:

$$-(f(t,x)-f(t,y),x-y)_2 \ge \gamma \|x-y\|_2^2, \quad x,y \in \mathbb{R}^n.$$

Monotone AWA mit $\sup_{t\in[t_0,\infty)} \|f(t,0)\|_2 < \infty$ haben nach einem Resultat aus derm Text globale, gleichmäßig beschränkte Lösungen.

a) Für $f(t,x) := A(t)x + b(t)$ gilt, wenn $A(t)$ gleichmäßig für $t \ge t_0$ negativ definit ist:

$$-(f(t,x)-f(t,y),x-y)_2 = -(A(t)(x-y),x-y)_2 \ge \gamma \|x-y\|_2^2, \quad x,y \in \mathbb{R}^d,$$

d. h.: die AWA ist „monoton“.

b) Die Matrix $-A$ ist symmetrisch und strikt diagonal-dominant und hat folglich nur positive Eigenwerte. Also ist A negativ definit.

c) Weiter gilt im Falle $\sup_{t\ge t_0} \|b(t)\|_2 < \infty$:

$$\sup_{t\ge t_0} \|f(t,0)\|_2 = \sup_{t\ge t_0} \|b(t)\|_2 < \infty,$$

d. h.: Die Lösung der linearen AWA ist gleichmäßig beschränkt.

Lösung A.1.7 (Praktische Aufgabe): Für die verschiedenen Verfahren ergibt sich:

i) Sukzessive Approximation:

k	$u^{(k)}(1)$	$\left\lvert \frac{u^{(k)}(1)-\tan(1)}{\tan(1)} \right\rvert$
1	1.0000000	0.35790738
2	1.3333333	0.14387651
3	1.4825397	0.04807222
4	1.5369594	0.01312974
5	1.5527855	0.00296787
6	1.5565238	0.00056754
7	1.5572616	0.00009384

Bemerkung: Die sukzessive Approximation liefert eigentlich Näherungen für den ganzen Lösungsverlauf, $u(t)$, kann aber auch auf die Berechnung von $u(1)$ beschränkt werden. Der Hauptaufwand bei der Durchführung der Methode besteht in der Berechnung der Integrale $\int_0^t f(s,u^k(s))\,ds$, was in der Regel mit Quadraturformeln (am besten mit der Romberg-Methode) unter Verwendung von a posteriori Fehlerkontrolle erfolgen muß.

ii) Taylor-Methode:

R	$U_1^{(R)}(1)$	$\left\lvert\frac{U_1^{(R)}(1)-\tan(1)}{\tan(1)}\right\rvert$
1	1.0000000	0.35790738
2	1.0000000	0.35790738
3	1.3333333	0.14387651
4	1.3333333	0.14387651
5	1.4666667	0.05826416
6	1.4666667	0.05826416
7	1.5206349	0.02361155
8	1.5206349	0.02361155
9	1.5425044	0.00956931
10	1.5425044	0.00956931
11	1.5513676	0.00387829
12	1.5513676	0.00387829
13	1.5549598	0.00157181
14	1.5549598	0.00157181
15	1.5564156	0.00063703
16	1.5564156	0.00063703
17	1.5570056	0.00025818
18	1.5570056	0.00025818
19	1.5572448	0.00010464
20	1.5572448	0.00010464
21	1.5573417	0.00004241

iii) Polygonzugmethode:

h^{-1}	y_N	$\left\lvert\frac{y_N-\tan(1)}{\tan(1)}\right\rvert$
1	1.0000000	0.35790738
2	1.1250000	0.27764581
4	1.2551867	0.19405390
8	1.3669378	0.12229936
16	1.4472379	0.07073923
32	1.4974739	0.03848307
64	1.5260316	0.02014639
128	1.5413373	0.01031867
256	1.5492728	0.00522339
512	1.5533148	0.00262806
1024	1.5553548	0.00131817
2048	1.5563796	0.00066012
4096	1.5568933	0.00033032
8192	1.5571504	0.00016523
16384	1.5572790	0.00008263

Bemerkung: Die Polygonzugmethode liefert automatisch Näherungen y_n für den ganzen Lösungsverlauf und nicht nur zum Endwert $t = 1$. Der zu ihrer Durchführung notwendige Aufwand besteht im Wesentlichen in den Funktionsauswertungen $f(t_i, y_i)$ in den einzelnen Zeitschritten.

A.2 Kapitel 2

Lösung A.2.1: a) Unter *Konsistenz* (eines Einschrittverfahrens mit einer AWA) versteht man das Verschwinden des *lokalen Diskretisierungsfehlers (Abschneidefehlers)*

$$\tau_n^h \;=\; \frac{1}{h_n}(u_n - u_{n-1}) \;-\; F(h_n; t_{n-1}, u_{n-1})$$

bei kleiner werdender Schrittweite $h = \max_n(h_n)$:

$$\max_{t_n \in I} \|\tau_n^h\| \to 0 \quad \text{für} \;\; h \to 0.$$

Das Einschrittverfahren hat die *Konsistenzordnung* m, wenn gilt

$$\max_{t_n \in I} \| \tau_n^h \| = \mathcal{O}(h_n^m).$$

b1) Diese Formel entstand durch Koeffizientenvergleich aus der 2-stufigen Runge-Kutta-Formel mit der Taylormethode. Ansatz:

$$F(h_n; t_{n-1}, y_{n-1}) = c_1 f + c_2 f(t + h_n a_2, u + h_n b_{21} f).$$

Taylorentwicklung in Termen von h_n ergibt:

$$\begin{aligned} F(h_n; t_{n-1}, y_{n-1}) &= \sum_{r=0}^{\infty} \frac{h_n^r}{r!} \partial_{h_n} F \Big|_{h_n=0} \\ &= (c_1 + c_2) f + c_2 a_2 h_n f_t + c_2 b_{21} h_n f f_x + O(h_n^2). \end{aligned}$$

Koeffizientenvergleich mit der Taylormethode liefert die Konsistenzbedingung:

$$(c_1 + c_2) f + c_2 a_2 h_n f_t + c_2 b_{21} h_n f f_x + O(h_n^2) \stackrel{!}{=} f + \frac{1}{2} h_n \{f_t + f_x f\} + O(h_n^2).$$

Mit der Wahl $c_1 = 0,\ c_2 = 1$ und $a_2 = b_{21} = \frac{1}{2}$ erhält man also gerade eine Methode 2. Ordnung.

b2) Diese Formel ist eine 3-stufige Runge-Kutta-Formel

$$F(h_n; t_{n-1}, y_{n-1}) \ = \ c_1 k_1 + c_2 k_2 + c_3 k_3$$

$$\begin{aligned} k_1 &= f(t_{n-1}, y_{n-1}), \\ k_2 &= f(t_{n-1} + h_n a_2, y_{n-1} + h_n b_{21} k_1) = f(t_{n-1} + h_n a_2, y_{n-1} + h_n b_{21} f), \\ k_3 &= f(t_{n-1} + h_n a_3, y_{n-1} + h_n b_{31} k_1 + h_n b_{32} k_2) \\ &= f(t_{n-1} + h_n a_3, y_{n-1} + h_n b_{21} f + h_n b_{32} f(t_{n-1} + h a_2, y_{n-1} + h b_{21} f)). \end{aligned}$$

Taylorentwicklung der Verfahrensfunktion ergibt:

$$\begin{aligned} F(h_n; t_{n-1}, y_{n-1}) = {} & (c_1 + c_2 + c_3) f + (c_2 a_2 + c_3 a_3) h_n f_t + (c_2 b_{21} + c_3 b_{31} + c_3 b_{32}) h_n f f_x \\ & + (\tfrac{1}{2} c_2 a_2^2 + \tfrac{1}{2} c_3 a_3^2) h_n^2 f_{tt} + (c_2 a_2 b_{21} + c_3 a_3 b_{31} + c_3 a_3 b_{32}) h_n^2 f f_{tx} \\ & + (\tfrac{1}{2} c_2 b_{21}^2 + \tfrac{1}{2} c_3 b_{31}^2 + c_3 b_{31} b_{32} + \tfrac{1}{2} c_3 b_{32}^2) h_n^2 f^2 f_{xx} \\ & + c_3 a_2 b_{32} h_n^2 f_t f_x + c_3 b_{21} b_{32} h_n^2 f f_x^2 + O(h_n^3) \end{aligned}$$

Koeffizientenvergleich mit der Taylormethode

$$\begin{aligned} F(h_n; t_{n-1}, y_{n-1}) \stackrel{!}{=} {} & f + \tfrac{1}{2} h_n \{f_t + f f_x\} \\ & + \tfrac{1}{6} h_n^2 \{f_{tt} + 2 f f_{tx} + f^2 f_{xx} + f_t f_x + f f_x^2\} + O(h_n^3) \end{aligned}$$

ergibt die Konsistenzbedingungen:

$$\begin{aligned}
c_1 + c_2 + c_3 &= 1,\\
c_2 a_2 + c_3 a_3 &= \tfrac{1}{2},\\
c_2 b_{21} + c_3 b_{31} + c_3 b_{32} &= \tfrac{1}{2},\\
\tfrac{1}{2} c_2 a_2^2 + \tfrac{1}{2} c_3 a_3^2 &= \tfrac{1}{6},\\
c_2 a_2 b_{21} + c_3 a_3 b_{31} + c_3 a_3 b_{32} &= \tfrac{1}{3},\\
\tfrac{1}{2} c_2 b_{21}^2 + \tfrac{1}{2} c_3 b_{31}^2 + c_3 b_{31} b_{32} + \tfrac{1}{2} c_3 b_{32}^2 &= \tfrac{1}{6},\\
c_3 a_2 b_{32} &= \tfrac{1}{6},\\
c_3 b_{21} b_{32} &= \tfrac{1}{6}.
\end{aligned}$$

Ablesen der Koeffizienten ergibt:

$$c_1 = \tfrac{1}{10}, \quad c_2 = \tfrac{5}{10}, \quad c_3 = \tfrac{4}{10}, \quad a_2 = \tfrac{1}{3}, \quad a_3 = \tfrac{5}{6}, \quad b_{21} = \tfrac{1}{3}, \quad b_{31} = -\tfrac{5}{12}, \quad b_{32} = \tfrac{5}{4}.$$

Die so konstruierte Runge-Kutta-Formel ist demnach von 3. Ordnung.

(Alternativer Zugang: Die endgültigen Koeffizienten direkt in der Taylorentwicklung der Verfahrensfunktion einsetzen.)

Lösung A.2.2: a) Es genügt, die L-Stetigkeit der $k_r(h;t,x)$ zu zeigen. Zunächst schätzt man allgemein ab:

$$\begin{aligned}
\|k_r(h;t,x) - k_r(h;t,y)\| &= \Big\| f\Big(t + ha_r, x + h\sum_{s=1}^{R} b_{rs}k_s\Big) - f\Big(t + ha_r, y + h\sum_{s=1}^{R} b_{rs}k_s\Big)\Big\| \\
&\le L_f\Big(h\sum_{s=1}^{R} |b_{rs}|\,\|k_s(h;t,x) - k_s(h;t,y)\| + \|x - y\|\Big) \qquad (*)
\end{aligned}$$

Unter der Schrittweitenbedingung $L_f h|b_{rr}| < 1$ erhält man:

$$\begin{aligned}
&\|k_r(h;t,x) - k_r(h;t,y)\| \\
&\qquad \le \frac{1}{1 - L_f h|b_{rr}|} L_f\Big(h\sum_{s=1,s\neq r}^{R} |b_{rs}|\,\|k_s(h;t,x) - k_s(h;t,y)\| + \|x - y\|\Big) \qquad (**)
\end{aligned}$$

Diese Ungleichung erlaubt es nun, induktiv die L-Stetigkeit der k_r zu zeigen:

Induktionsanfang: Für $R = 1$ folgt sofort die Lipschitz-Stetigkeit von k_1.

Induktionsschritt: Einsetzen der Ungleichung (*) für k_R in die Ungleichungen (**) für $k_1, \cdots, k_{R-1}$ liefert $R - 1$ Ungleichungen mit modifizierten Koeffizienten:

$$\begin{aligned}
\|k_r(h;t,x) - k_r(h;t,y)\| \le L_f \Bigg(& h\sum_{s=1}^{R-1} |b_{rs}| \left(1 + \frac{|b_{rR}|\,L_f\,h}{1 - L_f h|b_{RR}|}\right) \|k_s(h;t,x) - k_s(h;t,y)\| \\
& + \left(1 + \frac{L_f\,h\,(R-1)|b_{rR}|}{1 - L_f h|b_{RR}|}\right)\|x - y\| \Bigg) \quad \text{für } r = 1, \cdots, R-1
\end{aligned}$$

Nach Induktionsvoraussetzung sind die k_r, $r = 1, \cdots, R-1$ aufgrund dieses Ungleichungssystems L-stetig. Nach Ungleichung (**) ist somit auch k_R L-stetig.

b) Durch Taylorentwicklung erhalten wir

$$\begin{aligned}
\tau_n &= \frac{u_n - u_{n-1}}{h_n} - F(h_n; t_{n-1}, u_{n-1}) \\
&= \frac{u_{n-1} + hu'_{n-1} + O(h_n^2) - u_{n-1}}{h_n} - F(h_n; t_{n-1}, u_{n-1}) \\
&= u'_{n-1} - F(h_n; t_{n-1}, u_{n-1}) + O(h_n) \\
&= f(t_{n-1}, u_{n-1}) - \sum_{r=1}^{R} c_r k_r(h_n; t_{n-1}, u_{n-1}) + O(h_n) \\
&= f(t_{n-1}, u_{n-1}) - \sum_{r=1}^{R} c_r \Big\{ f(t_{n-1}, u_{n-1}) + O(h_n) \Big\} + O(h_n) \\
&= f(t_{n-1}, u_{n-1}) \Big\{ 1 - \sum_{r=1}^{R} c_r \Big\} + O(h_n).
\end{aligned}$$

Das Verfahren ist konsistent, falls $\tau_n \to 0$ ($h_n \to 0$). Dies ist offensichtlich genau dann der Fall, wenn $\sum_{r=1}^{R} c_r = 1$ gilt.

Lösung A.2.3: Mit $\tilde{e}_n := \tilde{y}_n - u_n$ gilt

$$\tilde{e}_n = \tilde{e}_{n-1} + h_n \big(F(h_n; t_{n-1}, \tilde{y}_{n-1}) - F(h_n; t_{n-1}, u_{n-1}) \big) - h_n \tau_n + \varepsilon_n$$

und damit

$$\begin{aligned}
\|\tilde{e}_n\| &\le \|\tilde{e}_{n-1}\| + h_n \|F(h_n; t_{n-1}, \tilde{y}_{n-1}) - F(h_n; t_{n-1}, u_{n-1})\| + h_n \|\tau_n\| + \|\varepsilon_n\| \\
&\le \|\tilde{e}_{n-1}\| + h_n L \|\tilde{e}_{n-1}\| + h_n \|\tau_n\| + \|\varepsilon_n\| \\
&\le \|\tilde{e}_0\| + L \sum_{\nu=0}^{n-1} h_{\nu+1} \|\tilde{e}_\nu\| + \sum_{\nu=1}^{n} h_\nu \|\tau_\nu\| + \sum_{\nu=1}^{n} \|\varepsilon_\nu\|.
\end{aligned}$$

Mit Hilfe des diskreten Gronwallschen Lemmas erschließt man

$$\begin{aligned}
\|\tilde{e}_n\| &\le \exp\Big[L \sum_{\nu=0}^{n-1} h_{\nu+1} \Big] \Big\{ \|\tilde{e}_0\| + \sum_{\nu=1}^{n} h_\nu \|\tau_\nu\| + \sum_{\nu=1}^{n} \|\varepsilon_\nu\| \Big\} \\
&\le e^{L(t_n - t_0)} \Big\{ \|\tilde{e}_0\| + \max_{1 \le m \le n} \|\tau_m\| \sum_{\nu=1}^{n} h_\nu + \sum_{\nu=1}^{n} h_\nu h_\nu^{-1} \|\varepsilon_\nu\| \Big\} \\
&\le e^{L(t_n - t_0)} \Big\{ \|\tilde{e}_0\| + (t_n - t_0) \max_{1 \le m \le n} \|\tau_m\| + \max_{1 \le m \le n} h_m^{-1} \|\varepsilon_m\| \sum_{\nu=1}^{n} h_\nu \Big\} \\
&\le \underbrace{e^{L(t_n - t_0)}}_{=: K(t_n)} \Big\{ \|\tilde{e}_0\| + (t_n - t_0) \big(\max_{1 \le m \le n} \|\tau_m\| + \mathrm{eps} \max_{1 \le m \le n} h_m^{-1} \|y_m\| \big) \Big\},
\end{aligned}$$

was zu zeigen war.

Lösung A.2.4: a) Mit dem Ansatz $a(h) = a + ch^\alpha$ ergibt sich

$$\frac{a(h) - a}{a(h/2) - a} = \frac{ch^\alpha}{c(h/2)^\alpha} = 2^\alpha$$

und folglich

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

Wenn der Limes a nicht bekannt ist, schreiben wir

$$\begin{aligned}\frac{a(h) - a(h/2)}{a(h/2) - a(h/4)} &= \frac{a(h) - a + a - a(h/2)}{a(h/2) - a + a - a(h/4)} = \frac{ch^\alpha - c(h/2)^\alpha}{c(h/2)^\alpha - c(h/4)^\alpha} \\ &= \frac{1 - 2^{-\alpha}}{2^{-\alpha} - 4^{-\alpha}} = 2^\alpha\end{aligned}$$

und erhalten

$$\alpha = \frac{1}{\log(2)} \log\left(\Big|\frac{a(h) - a(h/2)}{a(h/2) - a(h/4)}\Big|\right).$$

b) Für die gegebenen Folgen ergibt sich:

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

und

$$\alpha \approx \frac{1}{\log(2)} \log\left(\Big|\frac{b(h) - b(h/2)}{b(h/2) - b(h/4)}\Big|\right) \approx 2.$$

Lösung A.2.5 (Praktische Aufgabe): Zu approximieren ist die Lösung der AWA

$$u'(t) = \sin(u(t)), \;\; t \in [0, 10], \;\; u(0) = 1.$$

Die exakte Lösung berechnet man mittels „TdV“:

$$\begin{aligned}\int_{u(0)}^{u(t)} \frac{1}{\sin(z)}\, dz = \int_0^t 1\, ds \quad &\Leftrightarrow \quad \big[\ln(\tan\big(\frac{z}{2}\big))\big]_{u(0)}^{u(t)} = t \\ &\Leftrightarrow \quad \ln(\tan\big(\frac{u(t)}{2}\big)) - \ln(\tan\big(\frac{u(0)}{2}\big)) = t \\ &\Leftrightarrow \quad \ln(\tan\big(\frac{u(t)}{2}\big)) = t + \ln(\tan(\frac{1}{2})) \\ &\Leftrightarrow \quad \tan\big(\frac{u(t)}{2}\big) = \exp(t + \ln(\tan(\frac{1}{2}))) \\ &\Leftrightarrow \quad u(t) = 2\arctan(e^t \tan\big(\frac{1}{2}\big)).\end{aligned}$$

Probe: Mit den Beziehungen

$$\begin{aligned}\sin(\arctan(x)) &= \frac{x}{\sqrt{1 + x^2}}, \\ \cos(\arctan(x)) &= \frac{1}{\sqrt{1 + x^2}}, \\ \sin(2x) &= 2\sin(x)\cos(x)\end{aligned}$$

gilt mit $x := e^t \tan(1/2)$:

$$u'(t) = 2 \cdot \frac{1}{1+x^2} \cdot x$$

sowie

$$\begin{aligned}\sin(u(t)) &= \sin(2\arctan(x)) \\ &= 2 \cdot \sin(\arctan(x)) \cdot \cos(\arctan(x)) \\ &= 2 \cdot \frac{x}{\sqrt{1+x^2}} \cdot \frac{1}{\sqrt{1+x^2}} \\ &= 2 \cdot \frac{1}{1+x^2} \cdot x.\end{aligned}$$

Außerdem ist der Anfangswert erfüllt: $u(0) = 2 \cdot \arctan(\tan\left(\frac{1}{2}\right)) = 2 \cdot \frac{1}{2} = 1$.

Tabelle A.1: Iterationen, absolute Fehler e und Konvergenzordnung o

iter	h	e (FE)	o (FE)	e (ME)	o (ME)	e (RK4)	o (RK4)
4	2^{-4}	$4.0 * 10^{-5}$	0.51	$9.2 * 10^{-7}$	2.22	$2.0 * 10^{-10}$	4.14
5	2^{-5}	$2.1 * 10^{-5}$	0.78	$2.2 * 10^{-7}$	2.09	$1.2 * 10^{-11}$	4.07
6	2^{-6}	$1.1 * 10^{-5}$	0.89	$5.5 * 10^{-8}$	2.04	$7.5 * 10^{-13}$	4.04
7	2^{-7}	$5.5 * 10^{-6}$	0.95	$1.4 * 10^{-8}$	2.02	$4.7 * 10^{-14}$	4.02
8	2^{-8}	$2.8 * 10^{-6}$	0.97	$3.4 * 10^{-9}$	2.01	$4 * 10^{-15}$	4.04
9	2^{-9}	$1.4 * 10^{-6}$	0.99	$8.5 * 10^{-10}$	2.00	$2 * 10^{-15}$	2.88
10	2^{-10}	$6.9 * 10^{-7}$	0.99	$2.1 * 10^{-10}$	2.00	$3 * 10^{-15}$	2.12
11	2^{-11}	$3.5 * 10^{-7}$	1.00	$5.3 * 10^{-11}$	2.00	$1 * 10^{-15}$	-0.74
12	2^{-12}	$1.7 * 10^{-7}$	1.00	$1.3 * 10^{-11}$	2.00	$6 * 10^{-15}$	-1.26
13	2^{-13}	$8.7 * 10^{-8}$	1.00	$3.3 * 10^{-12}$	2.00	$3 * 10^{-15}$	-0.74
14	2^{-14}	$4.3 * 10^{-8}$	1.00	$8.2 * 10^{-13}$	2.00	$3 * 10^{-15}$	0.62

Lösung A.2.6: Nach Definition des Abschneidefehlers ist

$$\begin{aligned}\tau_n &:= \frac{1}{h_n}\{u_n - u_{n-1}\} - f(t_n, u_n) = \frac{1}{h_n}\int_{t_{n-1}}^{t_n} u'(t)\,dt - \Big[\frac{t - t_{n-1}}{h_n}u'(t)\Big]_{t_{n-1}}^{t_n} \\ &= -\frac{1}{h_n}\int_{t_{n-1}}^{t_n} (t - t_{n-1})u''(t)\,dt\end{aligned}$$

und somit

$$\begin{aligned}\|\tau_n\| &\le \frac{1}{h_n}\int_{t_{n-1}}^{t_n}(t-t_{n-1})\|u''(t)\|\,dt\\ &\le \frac{1}{h_n}\max_{t\in I_n}\|u''(t)\|\underbrace{\int_{t_{n-1}}^{t_n}(t-t_{n-1})\,dt}_{=\frac{1}{2}h_n^2} = \frac{1}{2}h_n\max_{t\in I_n}\|u''(t)\|.\end{aligned}$$

Für u_n gilt $u_n = u_{n-1} + h_n f(t_n,u_n) + h_n\tau_n$ und damit für den Fehler $e_n := y_n - u_n$:

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

sowie unter Ausnutzung der Lipschitz-Stetigkeit von $f(t,x)$:

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

Wiederholte Anwendung liefert

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

Für $h_n < 1/L$ können wir das diskrete Gronwallsche Lemma anwenden
($\gamma := \max_{1\le i\le n}(1-Lh_i)^{-1}$):

$$\begin{aligned}\|e_n\| &\le \exp\Big[\gamma L\sum_{\nu=1}^{n}h_\nu\Big]\Big\{\|e_0\| + \sum_{\nu=1}^{n}h_\nu\|\tau_\nu\|\Big\}\\ &= e^{\gamma L(t_n-t_0)}\Big\{\|e_0\| + \sum_{\nu=1}^{n}h_\nu\|\tau_\nu\|\Big\}\\ &\le e^{\gamma L(t_n-t_0)}\Big\{\|e_0\| + \max_{1\le m\le n}\|\tau_m\|\sum_{\nu=1}^{n}h_\nu\Big\}\\ &\le e^{\gamma L(t_n-t_0)}\Big\{\|e_0\| + (t_n-t_0)\max_{1\le m\le n}\|\tau_m\|\Big\}\end{aligned}$$

Mit $\|\tau_m\| \le \frac{1}{2}h_m\max_{t\in I_m}\|u''(t)\|$ folgt

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

Lösung A.2.7: i) Die erste behauptete Ungleichung ist äquivalent zu

$$\Big\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\Big\|^2 \le \|x-y\|^2.$$

Bezeichnet $(\cdot,\cdot)$ das euklidische Skalarprodukt, so gilt (wegen $\|x\|\ge 1$, $\|y\|\ge 1$):

$$\begin{aligned}\Big\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\Big\|^2 &= \frac{1}{\|x\|^2\|y\|^2}\,(x\|y\| - y\|x\|, x\|y\| - y\|x\|)\\ &\le \frac{1}{\|x\|\|y\|}\,\big\{\|x\|^2\|y\|^2 - 2(x,y)\|x\|\|y\| + \|y\|^2\|x\|^2\big\}\\ &= 2\|x\|\|y\| - 2(x,y).\end{aligned}$$

Mit Hilfe der Young'schen Ungleichung ($ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$) erschließen wir somit

$$\Big\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\Big\|^2 \le 2\|x\|\|y\| - 2(x,y) \le \|x\|^2 - 2(x,y) + \|y\|^2 = \|x-y\|^2,$$

was zu zeigen war. Die Richtigkeit der zweiten Ungleichung sieht man wie folgt:

$$\begin{aligned}
\Big\|\frac{x}{\|x\|} - z\Big\|^2 &= \Big(\frac{x}{\|x\|} - z, \frac{x}{\|x\|} - z\Big) = \Big(\frac{x}{\|x\|} - x + x - z, \frac{x}{\|x\|} - z\Big) \\
&= \Big(\frac{x}{\|x\|} - x, \frac{x}{\|x\|} - z\Big) + \Big(x - z, \frac{x}{\|x\|} - z\Big) \\
&= 1 - \frac{(x,z)}{\|x\|} - \|x\| + (x,z) + \Big(x - z, \frac{x}{\|x\|} - z\Big) \\
&= \underbrace{\Big(1 - \frac{1}{\|x\|}\Big)}_{\ge 0,\ \text{da}\ \|x\| \ge 1}(x,z) + 1 - \|x\| + \Big(x - z, \frac{x}{\|x\|} - z\Big) \\
&\le \Big(1 - \frac{1}{\|x\|}\Big)\|x\| \underbrace{\|z\|}_{\le 1} + 1 - \|x\| + \|x - z\|\Big\|\frac{x}{\|x\|} - z\Big\| \\
&\le \|x\| - 1 + 1 - \|x\| + \|x - z\|\Big\|\frac{x}{\|x\|} - z\Big\| = \|x - z\|\Big\|\frac{x}{\|x\|} - z\Big\|
\end{aligned}$$

Kürzen liefert dann die Behauptung.

ii) Offenbar genügt es, die L-Stetigkeit von $\tilde{f}$ zu zeigen. Dabei können drei Fälle auftreten:

1. Fall $(t,x),(t,y) \in U_\rho$:
$$\|\tilde{f}(t,x) - \tilde{f}(t,y)\| = \|f(t,x) - f(t,y)\| \le L_f\|x - y\|;$$

2. Fall $(t,x),(t,y) \in (I \times \mathbb{R}^d) \setminus U_\rho$:
$$\begin{aligned}
\|\tilde{f}(t,x) - \tilde{f}(t,y)\| &= \|f(t,x_\rho) - f(t,y_\rho)\| \\
&\le L_f\|x_\rho - y_\rho\| = L_f\Big\|\rho\frac{x-u(t)}{\|x-u(t)\|} + u(t) - \rho\frac{y-u(t)}{\|y-u(t)\|} - u(t)\Big\| \\
&= L_f\rho\Big\|\frac{x-u(t)}{\|x-u(t)\|} - \frac{y-u(t)}{\|y-u(t)\|}\Big\| \\
&= L_f\rho\Big\|\frac{\rho^{-1}(x-u(t))}{\|\rho^{-1}(x-u(t))\|} - \frac{\rho^{-1}(y-u(t))}{\|\rho^{-1}(y-u(t))\|}\Big\| \\
&\le L_f\rho\|\rho^{-1}(x - u(t) - y + u(t))\| = L_f\|x - y\|;
\end{aligned}$$

3. Fall $(t,x) \in (I \times \mathbb{R}^d) \setminus U_\rho$, $(t,y) \in U_\rho$:
$$\begin{aligned}
\|\tilde{f}(t,x) - \tilde{f}(t,y)\| &= \|f(t,x_\rho) - f(t,y)\| \\
&\le L_f\|x_\rho - y\| = L_f\Big\|\rho\frac{x-u(t)}{\|x-u(t)\|} + u(t) - y\Big\| \\
&= L_f\rho\Big\|\frac{\rho^{-1}(x-u(t))}{\|\rho^{-1}(x-u(t))\|} - \rho^{-1}(y-u(t))\Big\| \\
&\le L_f\rho\|\rho^{-1}(x - u(t) - y + u(t))\| = L_f\|x - y\|.
\end{aligned}$$

Lösung A.2.8: Es gilt

$$y_n = y_{n-1} + h_n\{f(t_n, y_n) - f(t_n, 0)\} + h_n f(t_n, 0).$$

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

$$\|y_n\| \le \|y_n\|^{-1}(y_{n-1}, y_n) - \lambda_n h_n\|y_n\| + h_n\|f(t_n, 0)\|.$$

Dies ergibt

$$(1 + \lambda_n h_n)\|y_n\| \le \|y_{n-1}\| + h_n\|f(t_n, 0)\|.$$

bzw.

$$\|y_n\| \le \frac{1}{1 + \lambda_n h_n}\|y_{n-1}\| + \frac{h_n}{1 + \lambda_n h_n}\|f(t_n, 0)\|.$$

Wegen $\lambda_n \ge \lambda > 0$ (strikte Monotonie) erschließen wir hieraus per Induktion:

$$\|y_n\| \le \lambda^{-1} \max_{1\le\nu\le n} \|f(t_\nu, 0)\|.$$

Für $n = 1$ gilt trivialerweise

$$\|y_1\| \le \frac{h_1}{1 + \lambda h_1}\|\tau_1\| \le \frac{1}{\lambda}\|f(t_1, 0)|.$$

Sei die Behauptung nun richtig für $n - 1$. Dann folgt:

$$\begin{aligned}\|y_n\| &\le \frac{1}{1 + \lambda h_n}\|y_{n-1}\| + \frac{h_n}{1 + \lambda h_n}\|f(t_n, 0)\| \\ &\le \frac{1}{1 + \lambda h_n}\lambda^{-1} \max_{1\le\nu\le n-1} \|f(t_\nu, 0)\| + \frac{h_n}{1 + \lambda h_n}\|f(t_n, 0)\| \le \frac{1}{\lambda} \max_{1\le\nu\le n} \|f(t_\nu, 0)\|.\end{aligned}$$

Dies vervollständigt den Beweis.

Lösung A.2.9: a) Die L-Konstante der Fixpunktabbildung

$$g(x) := y_{n-1} + h_n F(h_n; t_n, x, y_{n-1})$$

ist $L_g = h_n L_F$ mit der L-Konstante L_F der Verfahrensfunktion $F(h; t, x, y)$ bzgl. des Arguments y. Für $h_n < L_F^{-1}$ ist die Abbildung $g(\cdot)$ eine Kontraktion und die Fixpunktiteration konvergiert nach dem Banachschen Fixpunktsatz. Man erhält durch rekursives Abschätzen mit Hilfe der L-Stetigkeit von g die a priori Fehlerabschätzung:

$$\begin{aligned}|y_n^{(k)} - y_n| &\le |y_n^{(k)} - y_n^{(k+1)}| + |y_n^{(k+1)} - y_n^{(k+2)}| + |y_n^{(k+2)} - y_n^{(k+3)}| + \cdots \\ &\le (h_n L_f)^k\, |y_n^{(1)} - y_n^{(0)}|\big\{1 + h_n L_f + (h_n L_f)^2 + \cdots\big\} \\ &= \frac{(h_n L_f)^k}{1 - h_n L_f}\, |y_n^{(1)} - y_n^{(0)}|.\end{aligned}$$

b) Wir suchen eine Nullstelle von

$$h(x) := x - y_{n-1} - h_n F(h_n; t_n, x, y_{n-1}).$$

Mit der Jacobi-Matrix

$$Jh(x) := I - hF'_x(h_n; t_n, x, y_{n-1})$$

lautet das Newton-Verfahren ausgehend von einem Startwert $y_n^{(0)}$:

$$y_n^{(k)} = y_n^{(k-1)} + \delta y_n^{(k)},$$

$$Jh\big(y_n^{(k-1)}\big)\delta y_n^{(k)} = -y_n^{(k-1)} + y_{n-1} + h_n F(h_n; t_n, y_n^{(k-1)}, y_{n-1}).$$

c) Das Newton-Verfahren ist unter der Bedingung, dass F zweimal stetig partiell differenzierbar ist und $Jh(y_n)$ regulär lokal quadratisch konvergent.

Lösung A.2.10 (Praktische Aufgabe): Nicht verfügbar.

Lösung A.2.11: a) Der Abschneidefehler

$$\tau_n \;=\; \frac{1}{h_n}(u_n - u_{n-1}) \;-\frac{1}{2}\Big(f(t_{n-1}, u_{n-1}) + f(t_n, u_n)\Big)$$

der Trapezregel erlaubt die folgende Darstellung (zweimaliges partielles Integrieren):

$$\tau_n = \frac{1}{2h_n}\int_{t_{n-1}}^{t_n} (t - t_n)(t - t_{n-1})u'''(t)\,dt.$$

Abschätzen:

$$\|\tau_n\| \le \frac{1}{12}\sup_{t\in I_n}\|u'''\|\, h_n^2.$$

b) A ist negativ definit und symmetrisch, d. h. es gibt eine Orthonormalbasis ONB bezüglich derer A als Bilinearform Diagonalgestalt besitzt mit den negativen Eigenwerten $\lambda_1, \cdots, \lambda_n$ auf der Diagonalen.

Betrachte nun die folgende Fehlerdarstellung für die Trapezregel:

$$(I - \tfrac{1}{2}h_n A)e_n = (I + \tfrac{1}{2}h_n A)e_{n-1} + h_n\tau_n.$$

Testen mit e_n und anschließender Koordinatenwechsel auf ONB liefert:

$$\Big(\big(I - \tfrac{1}{2}h_n \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}\big)\tilde{e}_n, \tilde{e}_n\Big) = \Big(\big(I + \tfrac{1}{2}h_n \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}\big)\tilde{e}_{n-1}, \tilde{e}_n\Big) + h_n\,(\tilde{\tau}_n, \tilde{e}_n)\,.$$

Es sei λ der betragsmäßig kleinste Eigenwert. Dann gilt

$$(1 + \tfrac{1}{2}h_n|\lambda|)\,(\tilde{e}_n, \tilde{e}_n) \le \Big\|I + \tfrac{1}{2}h_n \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}\Big\|\,\|\tilde{e}_{n-1}\|\,\|\tilde{e}_n\| + h_n\,(\tilde{\tau}_n, \tilde{e}_n)\,. \qquad (*)$$

Im Folgenden sei nun eine globale Schrittweitenbedingung $h_n \le h$ angenommen mit einem hinreichend kleinem h, so dass stets (die λ_i sind negativ):

$$\Big\|I + \tfrac{1}{2}h_n \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}\Big\| \le 1.$$

Division von (*) durch $\|\tilde{e}_n\|$ und anschließende Rücktransformation liefert:

$$\Big(1+\tfrac{1}{2}h_n|\lambda|\Big)\|e_n\| \le \|e_{n-1}\| + h_n\|\tau_n\|.$$

Hiermit beweist man nun induktiv die Fehlerdarstellung

$$\|e_n\| \le \frac{2}{|\lambda|}\max_{1\le\nu\le n}\|\tau_\nu\|.$$

Induktionsanfang: Mit $\|e_0\| = 0$ ergibt sich

$$\|e_1\| \le \frac{h_n}{1+\frac{1}{2}h_n|\lambda|}\|\tau_1\| \le \frac{2}{|\lambda|}\|\tau_1\|$$

Induktionsschritt:

$$\begin{aligned}\|e_n\| &\le \frac{1}{1+\frac{1}{2}h_n|\lambda|}\|e_{n-1}\| + \frac{h_n}{1+\frac{1}{2}h_n|\lambda|}\|\tau_n\| \\ &\le \left\{\frac{1}{1+\frac{1}{2}h_n|\lambda|}\,\frac{2}{|\lambda|} + \frac{h_n}{1+\frac{1}{2}h_n|\lambda|}\right\}\max_{1\le\nu\le n}\|\tau_\nu\| \\ &= \frac{2}{|\lambda|}\max_{1\le\nu\le n}\|\tau_\nu\|.\end{aligned}$$

c) Es ergibt sich also mit Hilfe von (a) und (b):

$$\|e_n\| \le \frac{1}{6}\,\frac{1}{|\lambda|}\sup_{t\in I_n}\|u'''\|\,h_n^2.$$

Dies liefert die globale Fehlerabschätzung

$$\sup_{t_n\ge 0}\|e_n\| \le \frac{1}{6}\,\frac{1}{|\lambda|}\sup_{t\ge 0}\|u'''\|\,h^2.$$

Lösung A.2.12: a) Es gilt

$$\big(1-h_n\alpha\big)\|e_n\|^2 \le \big(1-h_n\kappa\big)\|e_{n-1}\|^2 + \frac{h_n}{\alpha}\|\tau_n\|^2 \quad \forall\,\alpha>0,$$

mit $\kappa = 2\lambda - hL^2$.

Beweis: Es gilt für $e_n = u_n - y_n$:

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

Multiplizieren mit $\cdot 2\,e_n$ liefert:

$$\begin{aligned}2\|e_n\|^2 - 2\,(e_{n-1},e_n) &= 2\,h_n\big(f(t_{n-1},u_{n-1}) - f(t_{n-1},y_{n-1}),e_{n-1}\big) \\ &\quad + 2\,h_n\big(f(t_{n-1},u_{n-1}) - f(t_{n-1},y_{n-1}),e_n - e_{n-1}\big) \\ &\quad + 2h_n\,(\tau_n,e_n) \\ &\le -2h_n\lambda_n\|e_{n-1}\|^2 + 2h_nL\|e_{n-1}\|\,\|e_n - e_{n-1}\| + 2h_n\|\tau_n\|\,\|e_n\| \\ &\le -2h_n\lambda_n\|e_{n-1}\|^2 + 2h_n^2L^2\|e_{n-1}\|^2 + \|e_n - e_{n-1}\|^2 \\ &\quad + \frac{h_n}{\alpha}\|\tau_n\|^2 + h_n\,\alpha\|e_n\|^2\end{aligned}$$

Mit dem Hinweis folgt:

$$\|e_n\|^2 + \|e_{n-1} - e_n\| - \|e_{n-1}\| \le -2h_n\lambda_n\|e_{n-1}\|^2 + 2h_n^2L^2\|e_{n-1}\|^2 + \|e_n - e_{n-1}\|^2 + \frac{h_n}{\alpha}\|\tau_n\|^2 + h_n\,\alpha\|e_n\|^2$$

Umformen:

$$\big(1 - h_n\alpha\big)\|e_n\|^2 \le \big(1 - h_n\lambda + h_n^2L^2\big)\|e_{n-1}\|^2 + \frac{h_n}{\alpha}\|\tau_n\|^2.$$

□

b) Mit $\alpha = \kappa/2$ und o.B.d.A. $h \le \kappa^{-1}$ folgt induktiv:

$$\|e_n\|^2 \le \frac{8}{\kappa^2}\max_{1\le\nu\le n}\|\tau_\nu\|^2$$

Beweis:

$$\|e_1\| \le 0 + \frac{h_n}{\big(1 - \frac{1}{2}h_n\kappa\big)\frac{1}{2}\kappa}\|\tau_1\|^2 \le \frac{4}{\big(1 - \frac{1}{2}h\kappa\big)\kappa^2}\,\|\tau_1\|^2 \le \frac{8}{\kappa^2}\max_{1\le\nu\le1}\|\tau_\nu\|^2.$$

Weiterhin:

$$\begin{aligned}\|e_n\| &\le \frac{1 - h_n\kappa}{1 - \frac{1}{2}h_n\kappa}\|e_{n-1}\|^2 + \frac{h_n}{\big(1 - \frac{1}{2}h_n\kappa\big)\frac{1}{2}\kappa}\|\tau_n\|^2\\ &\le \left[\frac{1 - h_n\kappa}{1 - \frac{1}{2}h_n\kappa}\cdot\frac{8}{\kappa^2} + \frac{h_n}{\big(1 - \frac{1}{2}h_n\kappa\big)\frac{1}{2}\kappa}\right]\max_{1\le\nu\le n}\|\tau_\nu\|^2\\ &= \left[\frac{8\big(1 - h_n\kappa\big) + 4h_n\,\kappa}{\big(1 - \frac{1}{2}h_n\kappa\big)\cdot\kappa^2}\right]\max_{1\le\nu\le n}\|\tau_\nu\|^2\\ &= \frac{8}{\kappa^2}\max_{1\le\nu\le n}\|\tau_\nu\|^2\end{aligned}$$

□

c) Schließlich gilt für den Abschneidefehler des expliziten Euler-Verfahrens:

$$\|\tau_\nu\| \le \tfrac{1}{2}h_\nu \sup_{I_\nu}\|u''\|.$$

Lösung A.2.13: Durch Taylorentwicklung erhalten wir

$$\begin{aligned}u_n &= u_{n-1} + h_nu'_{n-1} + \tfrac{1}{2}h_n^2u''_{n-1} + \tfrac{1}{6}h_n^3u'''_{n-1} + O(h_n^4)\\ &= u_{n-1} + h_nf(t_{n-1},u_{n-1}) + \tfrac{1}{2}h_n^2f^{(1)}(t_{n-1},u_{n-1}) + \tfrac{1}{6}h_n^3f^{(2)}(t_{n-1},u_{n-1}) + O(h_n^4)\\ &= u_{n-1} + h_nf + \tfrac{1}{2}h_n^2\{f_t + f_xf\} + \tfrac{1}{6}h_n^3\{f_{tt} + 2f_{tx}f + f_{xx}f^2 + f_tf_x + f_x^2f\} + O(h_n^4)\end{aligned}$$

sowie

$$f(t_{n-1}+h_n, y_{n-1}+h_n f(t_{n-1},y_{n-1})) = \\ f + h_n f_t + h_n f f_x + \tfrac{1}{2}h_n^2 f_{tt} + h_n^2 f f_{tx} + \tfrac{1}{2}h_n^2 f^2 f_{xx} + O(h_n^3).$$

Für den Abschneidefehler

$$\tau_n = \frac{u_n - u_{n-1}}{h_n} - \frac{1}{2}f(t_{n-1},u_{n-1}) - \frac{1}{2}f(t_n, u_{n-1}+h_n f(t_{n-1},u_{n-1}))$$

ergibt sich damit

$$\begin{aligned}\tau_n &= f + \tfrac{1}{2}h_n\{f_t + f_x f\} + \tfrac{1}{6}h_n^2\{f_{tt} + 2f_{tx}f + f_{xx}f^2 + f_t f_x + f_x^2 f\} + O(h_n^3)\\ &\quad - \tfrac{1}{2}f - \tfrac{1}{2}f - \tfrac{1}{2}h_n f_t - \tfrac{1}{2}h_n f f_x - \tfrac{1}{4}h_n^2 f_{tt} - \tfrac{1}{2}h_n^2 f f_{tx} - \tfrac{1}{4}h_n^2 f^2 f_{xx} + O(h_n^3)\\ &= h_n^2\left\{-\tfrac{1}{12}f_{tt} - \tfrac{1}{6}f f_{tx} - \tfrac{1}{12}f^2 f_{xx} + \tfrac{1}{6}f_t f_x + \tfrac{1}{6}f_x^2 f\right\} + O(h_n^3).\end{aligned}$$

Wir erhalten somit

$$\tau_n = \tau^2(t_{n-1})h_n^2 + O(h_n^3).$$

Eine weitere Taylorentwicklung von $f(t_n - h_n, \cdot)$ liefert die behauptete Beziehung

$$\tau_n = \tau^2(t_n)h_n^2 + O(h_n^3).$$

Lösung A.2.14: (i) Die folgende Überlegung folgt der entsprechenden Passage im Text für „Schrittweitenhalbierung“. Zur Bestimmung von τ_{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 $H/4$ (vier Schritte) ergibt zum vorläufigen Zeitpunkt $t_{n+1} := t_n + H$ Näherungen y_{n+1}^H bzw. $y_{n+1}^{H/4}$. Für die Fehler gilt

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

bzw. eine analoge Identitt für $y_{n+1/4}^{H/4} - u(t_{n+1/4})$. Wir erhalten weiterhin durch rekursive Anwendung der Fehleridentität

$$\begin{aligned}y_{n+1}^{H/4} - u(t_{n+1}) &= y_{n+3/4}^{H/4} - u(t_{n+3/4}) + \tfrac{1}{4}H\Big\{F(\tfrac{1}{4}H;t_{n+3/4},y_{n+3/4}^{H/4})\\ &\quad - F(\tfrac{1}{4}H;t_{n+3/4},u(t_{n+3/4}))\Big\} - \tfrac{1}{4}H\,\tau_{n+1}^{H/4}\\ &= (1+O(H))\big\{y_{n+3/4}^{H/4} - u(t_{n+3/4})\big\} - (\tfrac{1}{4}H)^{m+1}\tau_{n+1}^m + O(H^{m+2})\\ &= (1+O(H))\big\{(1+O(H))\big\{y_{n+1/2}^{H/4} - u(t_{n+1/2})\big\} - (\tfrac{1}{4}H)^{m+1}\tau_{n+3/4}^m + O(H^{m+2})\big\}\\ &\quad - (\tfrac{1}{4}H)^{m+1}\tau_{n+1}^m + O(H^{m+2})\\ &\;\;\vdots\\ &= (1+O(H))e_n - (\tfrac{1}{4}H)^{m+1}\big\{\tau_{n+1}^m + \tau_{n+3/4}^m + \tau_{n+1/2}^m + \tau_{n+1/4}^m\big\} + O(H^{m+2}),\end{aligned}$$

und folglich

$$y_{n+1}^{H/4} - u(t_{n+1}) = (1 + O(H))e_n - 4(\tfrac{1}{4}H)^{m+1}\tau_{n+1}^m + O(H^{m+2}).$$

Dabei wurde ausgenutzt, dass sich die Hauptabschneidefunktion gemäß

$$\tau_{n+3/4}^m + \tau_{n+1/2}^m + \tau_{n+1/4}^m = 3\tau_{n+1}^m + O(H)$$

entwickeln lässt. Subtraktion dieser beiden Gleichungen ergibt

$$y_{n+1}^{H/4} - y_{n+1}^H \;=\; O(H)e_n - \tau_{n+1}^m \left\{4(\tfrac{1}{4}H)^{m+1} - H^{m+1}\right\} + O(H^{m+2})$$

bzw.

$$\tau_{n+1}^m = \frac{y_{n+1}^{H/4} - y_{n+1}^H}{H^{m+1}(1 - 4^{-m})} + O(H) + O(H^{-m})e_n\,. \tag{1.2.1}$$

Es wird nun postuliert, dass die beiden „O"-Terme rechts in (1.2.1) klein genug sind, um mit

$$\tilde{\tau}_{n+1}^m := \frac{y_{n+1}^{H/4} - y_{n+1}^H}{H^{m+1}(1 - 4^{-m})} \tag{1.2.2}$$

eine brauchbare Näherung für τ_{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 Diskussion im Text nahegelegt wird.

(ii) Analog zum Vorgehen im Text ergibt sich mit einer Schrittweite H:

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

und somit unter Beachtung von $(1 - O(H))^{-1} = 1 + O(H)$

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

Für die Schrittweite $\frac{H}{2}$ erhalten wir

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

und somit

$$y_{n+1}^{\frac{H}{2}} - u_{n+1} = (1+O(H))\Big\{y_{n+\frac{1}{2}}^{\frac{H}{2}} - u_{n+\frac{1}{2}}\Big\} - \Big(\frac{H}{2}\Big)^{m+1}\tau_{n+1}^m + O(H^{m+2}).$$

Durch Einsetzen ergibt sich

$$\begin{aligned} y_{n+1}^{\frac{H}{2}} - u_{n+1} &= (1+O(H))\Big\{(1+O(H))e_n - \Big(\frac{H}{2}\Big)^{m+1}\tau_{n+\frac{1}{2}}^m + O(H^{m+2})\Big\} \\ &\quad - \Big(\frac{H}{2}\Big)^{m+1}\tau_{n+1}^m + O(H^{m+2}) \\ &= (1+O(H))e_n - 2\Big(\frac{H}{2}\Big)^{m+1}\tau_{n+1}^m + O(H^{m+2}), \end{aligned}$$

wobei wir ausgenutzt haben, dass $\tau_{n+\frac{1}{2}}^m = \tau_{n+1}^m + O(H)$. Durch Subtraktion beider Gleichungen erhält man

$$y_{n+1}^{\frac{H}{2}} - y_{n+1}^H = O(H)e_n - \tau_{n+1}^m\Big\{2\Big(\frac{H}{2}\Big)^{m+1} - H^{m+1}\Big\} + O(H^{m+2})$$

bzw.

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

Analog zum Text wird nun wieder postuliert, dass die beiden „$\mathcal{O}$“-Terme klein sind.

Lösung A.2.15: Subtraktion der beiden Gleichungen liefert:

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

Unter der Annahme $\tau_n \approx \tau_n(y_n)$, sowie

$$\sup_{t_0\le t_n\le T}\|f(t_n, .)\|_\infty =: \kappa < \infty$$

liefert dies:

$$\|e_n - E_n\| \le \|e_{n-1} - E_{n-1}\| + h_n\kappa\,\|e_n - E_n\| + h_n c\|e_n\|^2 + h_n\,\mathcal{O}(\|e_n\|^3)$$

Umformen:

$$\|e_n - E_n\| \le \|e_{n-1} - E_{n-1}\| + \frac{h_n\kappa}{1-h_n\kappa}\,\|e_{n-1} - E_{n-1}\| + \frac{h_n}{1-h_n\kappa}\big(c\|e_n\|^2 + \mathcal{O}(\|e_n\|^3)\big).$$

Rekursiv Einsetzen:

$$\|e_n - E_n\| \le \sum_{\nu=0}^{n-1}\frac{h_{\nu+1}\kappa}{1-h_{\nu+1}\kappa}\,\|e_\nu - E_\nu\| + \sum_{\nu=1}^{n}\frac{h_\nu}{1-h_\nu\kappa}\big(c\|e_\nu\|^2 + \mathcal{O}(\|e_\nu\|^3)\big).$$

Diskrete Gronwallsche Ungleichung unter einer Schrittweitenbedingung $h\kappa \leq 2^{-1}$ anwenden:

$$\begin{aligned}\|e_n - E_n\| &\leq \exp\big(2\,\kappa(t_n - t_0)\big)\,2\Big\{\sum_{\nu=1}^{n-1} h_\nu\, c\|e_\nu\|^2 + \sum_{\nu=1}^{n-1} h_\nu \mathcal{O}(\|e_\nu\|^3)\Big\} \\ &\leq \exp\big(2\,\kappa(t_n - t_0)\big)\,2\Big\{(t_n - t_0)\, c \max_{1\leq\nu\leq n}\|e_\nu\|^2 + (t_n - t_0)\max_{1\leq\nu\leq n}\mathcal{O}(\|e_\nu\|^3)\Big\},\end{aligned}$$

also

$$\|e_n - E_n\| = \mathcal{O}\big(\max_{1\leq\nu\leq n}\|e_\nu\|^2\big).$$

Lösung A.2.16 (Praktische Aufgabe): Wir haben die folgenden Resultate:

Tabelle A.2: Lokale vs. globale Verfeinerung auf $I = [-3, -1]$ für versch. Toleranzen (Heun)

	TOL	$\#Int$	$\#Eval$	h_{min}	h_{max}	$\max\|e_n\|$
local	10^{-5}	75	872	$7.5\cdot10^{-3}$	$2.0\cdot10^{-1}$	$3.4\cdot10^{-6}$
global	10^{-5}	160	600	$1.3\cdot10^{-2}$	$1.3\cdot10^{-2}$	$6.6\cdot10^{-6}$
local	10^{-9}	7536	90404	$7.2\cdot10^{-5}$	$2.1\cdot10^{-3}$	$1.4\cdot10^{-12}$
global	10^{-9}	20480	81880	$9.8\cdot10^{-5}$	$9.8\cdot10^{-5}$	$4.1\cdot10^{-10}$

Tabelle A.3: Lokale vs. globale Verfeinerung auf $I = [-3, -1]$ für versch. Toleranzen (RK)

	TOL	$\#Int$	$\#Eval$	h_{min}	h_{max}	$\max\|e_n\|$
local	10^{-5}	5	40	$2.0\cdot10^{-1}$	$8.0\cdot10^{-1}$	$1.5\cdot10^{-4}$
global	10^{-5}	20	80	$1.0\cdot10^{-1}$	$1.0\cdot10^{-1}$	$6.6\cdot10^{-7}$
local	10^{-9}	53	1228	$1.5\cdot10^{-2}$	$9.4\cdot10^{-2}$	$8.0\cdot10^{-9}$
global	10^{-9}	160	1200	$1.3\cdot10^{-2}$	$1.3\cdot10^{-2}$	$1.9\cdot10^{-10}$

Tabelle A.4: Lokale Verfeinerung auf $I = [-3, 1]$ für versch. Fehlerkonstanten (Heun)

	TOL	K	$\#Int$	$\#Eval$	h_{min}	h_{max}	$\max\|e_n\|$
local	10^{-5}	10	5766	62288	$6.0\cdot10^{-5}$	$7.4\cdot10^{-2}$	$3.5\cdot10^{-3}$
local	10^{-5}	2500	83733	854306	$3.8\cdot10^{-6}$	$9.3\cdot10^{-3}$	$1.0\cdot10^{-6}$
global	10^{-5}	-	81920	327600	$4.9\cdot10^{-5}$	$4.9\cdot10^{-5}$	$2.8\cdot10^{-6}$

Tabelle A.5: Lokale Verfeinerung auf $I = [-3, 1]$ für versch. Fehlerkonstanten (RK)

	TOL	K	$\#Int$	$\#Eval$	h_{min}	h_{max}	$\max \|e_n\|$
local	10^{-5}	10	118	2212	$2.3 \cdot 10^{-3}$	$8.0 \cdot 10^{-1}$	$1.6 \cdot 10^{0}$
local	10^{-5}	2500	357	7660	$1.5 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$	$1.5 \cdot 10^{-3}$
global	10^{-5}	-	640	4960	$6.3 \cdot 10^{-3}$	$6.3 \cdot 10^{-3}$	$4.5 \cdot 10^{-6}$
local	10^{-9}	10	897	19372	$5.9 \cdot 10^{-4}$	$7.9 \cdot 10^{-2}$	$4.2 \cdot 10^{-5}$
local	10^{-9}	2500	3469	73720	$1.5 \cdot 10^{-4}$	$4.0 \cdot 10^{-2}$	$3.4 \cdot 10^{-7}$
global	10^{-9}	-	10240	81760	$3.9 \cdot 10^{-4}$	$3.9 \cdot 10^{-4}$	$2.5 \cdot 10^{-10}$

A.3 Kapitel 3

Lösung A.3.1: Angewendet auf das skalare Modellproblem $u'(t) = \lambda u(t)$ liefern die Verfahren folgende Verstärkungsfaktoren:

1.) $y_{n+1} = y_n + \frac{1}{2}h_n \{\lambda y_{n+1} + \lambda y_n\}$ führt auf

$$y_{n+1} = \frac{1 + \frac{1}{2}h_n\lambda}{1 - \frac{1}{2}h_n\lambda} y_n.$$

Der Verstärkungsfaktor lautet also $\omega(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$. Somit ist

$$|\omega(z)| \leq 1 \quad \Leftrightarrow \quad \left|1 + \tfrac{1}{2}z\right| \leq \left|1 - \tfrac{1}{2}z\right|.$$

- 1. Fall: $z < 2$:
$$-1 + \tfrac{1}{2}z \leq 1 + \tfrac{1}{2}z \leq 1 - \tfrac{1}{2}z$$
liefert die Bedingungen $0 \leq 2$ und $z \leq 0$.

- 2. Fall: $z \geq 2$:
$$1 - \tfrac{1}{2}z \leq 1 + \tfrac{1}{2}z \leq -1 + \tfrac{1}{2}z$$
liefert die Bedingungen $z \geq 0$ und $2 \leq 0$.

Das Stabilitätsintervall ist also SI $= (-\infty, 0]$.

2.) $y_{n+1} = y_n + h_n\lambda \left(y_n + \frac{1}{2}h\lambda\right)$ führt auf

$$y_{n+1} = \left(\tfrac{1}{2}(h_n\lambda)^2 + h_n\lambda + 1\right) y_n.$$

Der Verstärkungsfaktor lautet also $\omega(z) = \frac{1}{2}z^2 + z + 1$. Somit ist

$$|\omega(z)| \leq 1 \quad \Leftrightarrow \quad \left|\tfrac{1}{2}z^2 + z + 1\right| \leq 1.$$

Wegen $\frac{1}{2}z^2+z+1=\frac{1}{2}\left(z^2+2z+2\right)=\frac{1}{2}(z+1)^2+\frac{1}{2}$ ist dies äquivalent zu

$$\tfrac{1}{2}(z+1)^2+\tfrac{1}{2}\le 1 \quad\Leftrightarrow\quad |z+1|\le 1.$$

Das Stabilitätsintervall ist also SI $=[-2,0]$.

3.) Zunächst bestimmen wir

$$f^{(1)}(t,x)=f_t(t,x)+f(t,x)\cdot f_x(t,x)=0+\lambda x\cdot\lambda=\lambda^2 x.$$

Damit erhalten wir $y_{n+1}=y_n+\frac{1}{6}h_n\left\{2\lambda y_{n+1}+4\lambda y_n+h_n\lambda^2 y_n\right\}$ und somit

$$\begin{aligned} y_{n+1} &= \frac{\frac{1}{6}(h\lambda)^2+\frac{2}{3}h_n\lambda+1}{1-\frac{1}{3}h_n\lambda}y_n=\frac{\frac{1}{6}\left((h_n\lambda)^2+4h_n\lambda+6\right)}{1-\frac{1}{3}h_n\lambda}y_n \\ &= \frac{\frac{1}{6}\left((h_n\lambda+2)^2+2\right)}{1-\frac{1}{3}h_n\lambda}y_n=\frac{\frac{1}{6}\left((h_n\lambda+2)^2+2\right)}{\frac{1}{6}(6-2h_n\lambda)}y_n=\frac{(h_n\lambda+2)^2+2}{6-2h_n\lambda}y_n. \end{aligned}$$

Der Verstärkungsfaktor lautet also $\omega(z)=\frac{(z+2)^2+2}{6-2z}$. Somit ist

$$|\omega(z)|\le 1 \quad\Leftrightarrow\quad |(z+2)^2+2|\le|6-2z|.$$

– 1. Fall: $z<3$:

$$-6+2z\le(z+2)^2+2\le 6-2z \quad\Leftrightarrow\quad -6+2z\le z^2+4z+6\le 6-2z$$

liefert die Bedingungen $z^2+2z+12\ge 0 \;\Leftrightarrow\; (z+1)^2+11\ge 0 \;\Leftrightarrow\; (z+1)^2\ge -11$ und $z^2+6z\le 0 \;\Leftrightarrow\; (z+3)^2-9\le 0 \;\Leftrightarrow\; |z+3|\le 3$.

– 2. Fall: $z\ge 3$:

$$6-2z\le(z+2)^2+2\le -6+2z \quad\Leftrightarrow\quad 6-2z\le z^2+4z+6\le -6+2z$$

liefert die Bedingungen $z^2+6z\ge 0 \;\Leftrightarrow\; (z+3)^2-9\ge 0 \;\Leftrightarrow\; |z+3|\ge 3$ und $z^2+2z+12\le 0 \;\Leftrightarrow\; (z+1)^2+11\le 0 \;\Leftrightarrow\; (z+1)^2\le -11$.

Das Stabilitätsintervall ist also SI $=[-6,0]$.

Lösung A.3.2: i) Das entstehende System 1. Ordnung lautet

$$\begin{bmatrix} u_1'(t)\\ u_2'(t)\end{bmatrix}=\begin{bmatrix} u_2(t)\\ f(t,u_1(t),u_2(t))\end{bmatrix}.$$

Die zugehörige Jacobi-Matrix ergibt sich zu

$$J:=\begin{bmatrix} \frac{\partial u_2(t)}{\partial u_1} & \frac{\partial u_2(t)}{\partial u_2}\\ \frac{\partial f(t,u_1(t),u_2(t))}{\partial u_1} & \frac{\partial f(t,u_1(t),u_2(t))}{\partial u_2}\end{bmatrix}=\begin{bmatrix} 0 & 1\\ \frac{\partial f(t,u_1(t),u_2(t))}{\partial u_1} & \frac{\partial f(t,u_1(t),u_2(t))}{\partial u_2}\end{bmatrix}.$$

Nach Voraussetzung ist $\frac{\partial f(t,u_1,u_2)}{\partial u_1} =: c \geq 0$. Setzen wir weiter $d := \frac{\partial f(t,u_1,u_2)}{\partial u_2}$, so gilt:

$$\det(\lambda I - J) = \det \begin{bmatrix} \lambda & -1 \\ -c & \lambda - d \end{bmatrix} = \lambda(\lambda - d) - c = \lambda^2 - d\lambda - c \stackrel{!}{=} 0.$$

Umformen liefert

$$\lambda_{1/2} = \frac{d}{2} \pm \sqrt{\frac{d^2}{4} + c}.$$

Unter den gegebenen Voraussetzungen ist $\frac{d^2}{4} + c \geq 0$ und J hat somit nur reelle Eigenwerte.

ii) Im Falle reeller Eigenwerte muss nur das Stabilitätsintervall betrachtet werden. Da ohne Einschränkung $\lambda_2 \leq 0$, muss die Schrittweite h also so bemessen sein, dass $\lambda_2 h \in \mathrm{SI}$ ist.

Lösung A.3.3: a) Die Matrix-Exponentialfunktion e^A und der Matrix-Sinus $\sin(A)$ sind definiert als die Grenzwerte im $\mathbb{R}^{d\times d}$ der matrixwertigen Reihen:

$$e^A := \sum_{k=0}^{\infty} \frac{A^k}{k!}, \qquad \sin(A) := \sum_{k=0}^{\infty} (-1)^k \frac{A^{2k+1}}{(2k+1)!}$$

Diese Reihen haben die gemeinsame Majorante $\sum_{k=0}^{\infty} \|A\|^k/k!$, welche für beliebige Matrix A (absolut) konvergiert. Also sind beide Matrix-Reihen für beliebige Matrix A konvergent (im Sinne der Matrizen-Konvergenz = Norm-Konvergenz = elementweise Konvergenz) und stellen die definierten Funktionen dar. Die dritte Funktion hat die Reihendarstellung („Neumannsche Reihe“)

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Diese Reihe konvergiert für jede Matrix mit Norm $\|A\| < 1$, wobei $\|\cdot\|$ eine beliebige Matrizennorm (submultiplikative Matrix-Norm) ist. Folglich konvergiert die Reihe für jede Matrix mit $\mathrm{spr}(A) := \max\{|\lambda| : \lambda \text{ Eigenwert von } A\} < 1$. Für eine solche Matrix gilt dann $\lim_{k\to\infty} \|A^k\| \leq \lim_{k\to\infty} \|A\|^k = 0$ und folglich

$$(I - A)\sum_{k=0}^{m} A^k = \sum_{k=0}^{m} A^k - \sum_{k=1}^{m+1} A^k = I - A^{m+1} \to I \quad (m \to \infty).$$

Also ist der Limes $\lim_{m\to\infty} \sum_{k=0}^{m} A^m$ gerade die Inverse von $I - A$.

b) Zunächst stellen wir fest, dass

$$(QAQ^{-1})^i = QA\underbrace{Q^{-1}Q}_{=I}AQ^{-1}\ldots QAQ^{-1} = QA^iQ^{-1}$$

gilt. Damit gilt für alle n:

$$Q\Big[\sum_{i=0}^{n} a_i A^i\Big]Q^{-1} = \sum_{i=0}^{n} a_i Q A^i Q^{-1}.$$

Konvergiert nun $g(A) = \lim_{n\to\infty} \sum_{i=0}^{n} a_i A^i$, so ist das Cauchy-Kriterium erfüllt. Damit folgt dann

$$\Big\| \sum_{i=m+1}^{n} a_i \big(QAQ^{-1}\big)^i \Big\| = \Big\| Q \Big[\sum_{i=m+1}^{n} a_i A^i \Big] Q^{-1} \Big\| \leq \|Q\| \|Q^{-1}\| \Big\| \sum_{i=m+1}^{n} a_i A^i \Big\|.$$

Also ist dann auch das Cauchy-Kriterium für $g(QAQ^{-1})$ erfüllt und die Reihe konvergiert.

Lösung A.3.4 (Praktische Aufgabe): Nicht verfügbar.

Lösung A.3.5: 1.) Die symmetrische Matrix A besitzt ein Orthonormalsystem aus Eigenvektoren, d. h.: Es gibt eine orthogonale Matrix Q mit $A = QDQ^T$ und $D = \operatorname{diag}(\lambda_i)$ mit den Eigenwerten λ_i von A. Dann ist

$$p(A) = p\left(QDQ^T\right) = Qp(D)Q^T \quad \text{bzw.} \quad q(A) = q\left(QDQ^T\right) = Qq(D)Q^T.$$

Damit folgt:

$$\begin{aligned} g(hA) &= q(hA)^{-1} \cdot p(hA) = q\left(hQDQ^T\right)^{-1} \cdot p\left(hQDQ^T\right) \\ &= [Qq(hD)QT]^{-1} Qp(hD)Q^T = Qq(hD)^{-1}Q^T Qp(hD)Q^T \\ &= Qq(hD)^{-1}p(hD)Q^T = Qg(hD)Q^T. \end{aligned}$$

2.) Mit $y_n = g(hA)y_{n-1}$ folgt

$$\begin{aligned} Q^T y_n &= Q^T g(hA) Q Q^T y_{n-1} \\ &= g(hD) Q^T y_{n-1}. \end{aligned}$$

Dies erlaubt die Abschätzung:

$$\begin{aligned} & \|Q^T y_n\|_2 \leq \|g(hD)\|_2 \|Q^T y_{n-1}\|_2 \\ \iff \quad & \|y_n\|_2 \leq \|g(hD)\|_2 \|y_{n-1}\|_2 \quad (Q \text{ orthogonal}). \end{aligned}$$

Weiterhin gilt für die von der euklidischen Norm induzierten natürlichen Matrizennorm:

$$\|g(hD)\|_2 = \|\operatorname{diag}\big(g(h\lambda_i)\big)\|_2 \leq \max_{1\leq i\leq d} |g(h\lambda_i)|.$$

Damit ergibt sich:

$$\|y_n\|_2 \leq \max_{1\leq i\leq d} |g(h\lambda_i)|^n \|y_0\|.$$

3.) Für das System

$$\begin{bmatrix} u(t) \\ v(t) \end{bmatrix}' = \underbrace{\begin{bmatrix} -10 & 9 \\ 9 & -10 \end{bmatrix}}_{=:A} \begin{bmatrix} u(t) \\ v(t) \end{bmatrix}$$

sind die Eigenwerte der Matrix A zu bestimmen:

$$\det \begin{bmatrix} \lambda+10 & -9 \\ -9 & \lambda+10 \end{bmatrix} = (\lambda+10)^2 - 81 = \lambda^2 + 20\lambda + 100 - 81 = \lambda^2 + 20\lambda + 19.$$

Die Eigenwerte sind also gegeben durch

$$\lambda_{1/2} = -10 \pm \sqrt{100-19} = -10 \pm 9, \quad \text{d. h.} \quad \lambda_1 = -1, \quad \lambda_2 = -19.$$

Jetzt schreiben wir das klassische 4-stufige Runge-Kutta-Verfahren für $f(t,x) = Ax$ als $y_n = g(h_nA)y_{n-1}$:

$$\begin{aligned}
k_1 &= Ay_{n-1}, \\
k_2 &= A\left(y_{n-1} + \tfrac{1}{2}h_nAy_{n-1}\right) = Ay_{n-1} + \tfrac{1}{2}h_nA^2y_{n-1}, \\
k_3 &= A\left(y_{n-1} + \tfrac{1}{2}h_n\left(Ay_{n-1} + \tfrac{1}{2}h_nA^2y_{n-1}\right)\right) \\
&= Ay_{n-1} + \tfrac{1}{2}h_nA^2y_{n-1} + \tfrac{1}{4}h_n^2A^3y_{n-1}, \\
k_4 &= A\left(y_{n-1} + h_n\left(Ay_{n-1} + \tfrac{1}{2}h_nA^2y_{n-1} + \tfrac{1}{4}h_n^2A^3y_{n-1}\right)\right) \\
&= Ay_{n-1} + h_nA^2y_{n-1} + \tfrac{1}{2}h_n^2A^3y_{n-1} + \tfrac{1}{4}h_n^3A^4y_{n-1}.
\end{aligned}$$

Damit ergibt sich insgesamt:

$$\begin{aligned}
y_n &= y_{n-1} + \tfrac{1}{6}h_n\{k_1 + 2k_2 + 2k_3 + k_4\} \\
&= \left(I + h_nA + \tfrac{1}{2}h_n^2A^2 + \tfrac{1}{6}h_n^3A^3 + \tfrac{1}{24}h_n^4A^4\right)y_{n-1}.
\end{aligned}$$

Es gilt also $g(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$. Damit das System noch numerisch stabil integriert wird, muss nun

$$\max\{|g(-h)|, |g(-19h)|\} = \max|g(-19h)| < 1$$

sein. Da $g(z) > 0$ für alle $z \in \mathbb{R}$, ist dies äquivalent zu

$$1 - 19h + \tfrac{361}{2}h^2 - \tfrac{6859}{6}h^3 + \tfrac{130321}{24}h^4 < 1.$$

bzw.

$$130321h^4 - 27436h^3 + 4332h^2 - 456h < 0.$$

Für $h \neq 0$ ist dies gleichbedeutend mit

$$130321h^3 - 27436h^2 + 4332h - 456 < 0.$$

Mit dem Newton-Verfahren ermitteln wir nun eine Nullstelle von $f(h) := 130321h^3 - 27436h^2 + 4332h - 456$:

$$h_{(i+1)} = h_{(i)} - \frac{f(h_{(i)})}{f'(h_{(i)})} = h_{(i)} - \frac{130321h_{(i)}^3 - 27436h_{(i)}^2 + 4332h_{(i)} - 456}{390963h_{(i)}^2 - 54872h_{(i)} + 4332}$$

führt mit $h_{(0)} = 0$ zu

$$h_{(1)} = 0{,}1053, \quad h_{(2)} = 0{,}1579, \quad h_{(3)} = 0{,}1474, \quad h_{(4)} = 0{,}1466, \quad h_{(5)} = 0{,}1466.$$

Da $f(h)$ keine weiteren reellen Nullstellen hat, wird das System für $h < 0{,}1466$ numerisch stabil integriert.

Alternativ kann man sich auch die Kenntnis des Stabilitätsintervalls der Runge-Kutta-Formel ($\mathrm{SI} = [-2.78\ldots, 0]$) zunutze machen und hieraus die maximal zulässige Schrittweite bestimmen.

Lösung A.3.6: Wendet man die Trapezregel auf das Modellproblem $u'(t) = \lambda u(t)$ an, so erhält man

$$y_n = y_{n-1} + \tfrac{1}{2}h_n\lambda y_n + \tfrac{1}{2}h_n\lambda y_{n-1}$$

bzw. umgeformt

$$y_n = \frac{1 + \frac{1}{2}h_n\lambda}{1 - \frac{1}{2}h_n\lambda} y_{n-1}.$$

Der Verstärkungsfaktor ist also $\omega(z) := \dfrac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$. Es gilt dann

$$|\omega(z)| \le 1 \quad \Leftrightarrow \quad \left|1 + \tfrac{1}{2}z\right| \le \left|1 - \tfrac{1}{2}z\right| \quad \Leftrightarrow \quad \left|1 + \tfrac{1}{2}z\right|^2 \le \left|1 - \tfrac{1}{2}z\right|^2.$$

Dies wiederum ist äquivalent zu

$$\begin{aligned} & \left(1 + \tfrac{1}{2}\operatorname{Re} z\right)^2 + \left(\tfrac{1}{2}\operatorname{Im} z\right)^2 \le \left(1 - \tfrac{1}{2}\operatorname{Re} z\right)^2 + \left(\tfrac{1}{2}\operatorname{Im} z\right)^2 \\ \Longleftrightarrow \quad & 1 + \operatorname{Re} z + \tfrac{1}{4}(\operatorname{Re} z)^2 \le 1 - \operatorname{Re} z + \tfrac{1}{4}(\operatorname{Re} z)^2 \\ \Longleftrightarrow \quad & 2 \operatorname{Re} z \le 0. \end{aligned}$$

Also ist $SG_{\mathrm{TR}} = \{z \in \mathbb{C} |\ \operatorname{Re} z \le 0\}$.

Lösung A.3.7: a) Für das Modellproblem $u'(t) = \lambda u(t)$ ist $k_1 = \lambda y_{n-1}$ und

$$k_2 = \lambda\left(y_{n-1} + \tfrac{1}{2}h\lambda y_{n-1} + \tfrac{1}{2}hk_2\right) = \lambda y_{n-1} + \tfrac{1}{2}h\lambda^2 y_{n-1} + \tfrac{1}{2}h\lambda k_2$$

bzw.

$$k_2 = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\lambda y_{n-1}.$$

Einsetzen ergibt

$$\begin{aligned} y_n &= y_{n-1} + \tfrac{1}{2}h\lambda y_{n-1} + \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\tfrac{1}{2}h\lambda y_{n-1} = \left(1 + \tfrac{1}{2}h\lambda + \frac{1 + \frac{1}{2}h\lambda}{2 - h\lambda}h\lambda\right) y_{n-1} \\ &= \left(1 + \tfrac{1}{2}h\lambda\right)\left(1 + \frac{h\lambda}{2 - h\lambda}\right) y_{n-1} = \left(1 + \tfrac{1}{2}h\lambda\right)\frac{2 - h\lambda + h\lambda}{2 - h\lambda} y_{n-1} \\ &= \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda} y_{n-1}. \end{aligned}$$

Mit dem Verstärkungsfaktor $\omega(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$ gilt dann nach einer der vorausgegangenen Aufgaben:

$$|\omega(z)| \leq 1 \quad \Leftrightarrow \quad \operatorname{Re} z \leq 0,$$

d. h.: Das Stabilitätsgebiet der semi-implizite Runge-Kutta-Formel zweiter Ordnung ist $SG = SG_{\mathrm{TR}} = \{z \in \mathbb{C} |\ \operatorname{Re} z \leq 0\}$.

b) Betrachten wir zunächst weiter diese Runge-Kutta-Formel: Hier muss nur k_2 mit dem Newton-Verfahren bestimmt werden:

$$G(k) := f\left(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) + \tfrac{1}{2}hk\right) - k.$$

Für das Newton-Verfahren benötigen wir die Jacobi-Matrix von G:

$$G'(k) = \tfrac{1}{2}hf_x\left(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) + \tfrac{1}{2}hk\right) - I.$$

Die Newton-Iteration zur Bestimmung von k_2 lautet dann:

$$G'(k^{(i)})k^{(i+1)} = G'(k^{(i)})k^{(i)} - G(k^{(i)}), \quad i = 1, 2, \ldots,$$

also:

$$\begin{aligned}\left(\tfrac{1}{2}hf_x(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) + \tfrac{1}{2}hk^{(i)}) - I\right) k^{(i+1)} \\ = \tfrac{1}{2}hf_x\left(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) + \tfrac{1}{2}hk^{(i)}\right) k^{(i)} \\ - f\left(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) + \tfrac{1}{2}hk^{(i)}\right).\end{aligned}$$

Bei der Trapezregel wird y_n direkt mit dem Newton-Verfahren berechnet:

$$\begin{aligned}G(y) &:= y_{n-1} + \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) + \tfrac{1}{2}hf(t_{n-1} + h, y) - y, \\ G'(y) &= \tfrac{1}{2}hf_x(t_{n-1} + h, y) - I.\end{aligned}$$

Es lautet somit:

$$\begin{aligned}\left(\tfrac{1}{2}hf_x(t_{n-1} + h, y^{(i)}) - I\right) y^{(i+1)} = \tfrac{1}{2}hf_x(t_{n-1} + h, y^{(i)})y^{(i)} \\ - y_{n-1} - \tfrac{1}{2}hf(t_{n-1}, y_{n-1}) - \tfrac{1}{2}hf(t_{n-1} + h, y^{(i)}).\end{aligned}$$

Um den Aufwand pro Zeitschritt zu vergleichen, untersuchen wir zunächst die Anzahl der benötigten Funktionsauswertungen und arithmetischen Operationen:

		Fkt.sausw.	Mult.	Add.
Start:	RK2 :	1	2	2
	Trapez:	1	2	2
G:	RK2:	1	1	2
	Trapez:	1	1	2
G':	RK2:	1	2	2
	Trapez:	1	1	1
Ende:	RK2:	0	1	1
	Trapez:	0	0	0

Zu Beginn können wir nämlich jeweils $t_{n-1}+h$, $\frac{h}{2}$ sowie $y_{n-1}+\frac{h}{2}f(t_{n-1},y_{n-1})$ berechnen.

Beide Verfahren haben also pro Iteration etwa den gleichen Aufwand. Um die Konvergenzgeschwindigkeit des Newton-Verfahrens zu beurteilen, betrachten wir wegen

$$\|x_t - z\| \leq \frac{M}{2m}\|x_t - x_{t-1}\|^2, \quad M := \max\|G''\|, \quad m := \min\|G'\|$$

die zweiten Ableitungen der Verfahrensfunktionen:

$$\begin{aligned} G''(k) &= \tfrac{1}{4}h^2 f_{xx}(\dots) \quad \text{(RK2)},\\ G''(y) &= \tfrac{1}{2}h f_{xx}(\dots) \quad \text{(Trapez)}. \end{aligned}$$

Da für beide Verfahren $m = 1 + O(h)$ gilt, konvergiert das Newton-Verfahren für die Runge-Kutta-Formel schneller.

Lösung A.3.8 (Praktische Aufgabe): Nicht verfügbar.

Lösung A.3.9: Anwenden des mehrdimensionalen Mittelwertsatzes auf g liefert:

$$g(x^*) = g(x^k) + g'(\xi^k)\cdot(x^* - x^k), \quad \xi^k \in \overline{x^k x^*}.$$

Umformen und $g(x^*) = 0$ ausnutzen:

$$0 = g(x^k) + g'(x^k)\cdot(x^* - x^k) + \big(g'(\xi^k) - g'(x^k)\big)\cdot(x^* - x^k).$$

Umstellen:

$$g'(x^k)^{-1}g(x^k) + x^* - x^k = -g'(x^k)^{-1}\big(g'(\xi^k) - g'(x^k)\big)\cdot(x^* - x^k).$$

Damit erhält man:

$$\begin{aligned} x^* - x^{k+1} &= x^* - x^k + g'(x^k)^{-1}g(x^k)\\ &= -g'(x^k)^{-1}\big(g'(\xi^k) - g'(x^k)\big)\cdot(x^* - x^k). \end{aligned}$$

Also

$$\frac{\|x^* - x^{k+1}\|}{\|x^* - x^k\|} \leq \|g'(x^k)^{-1}\|\|g'(\xi^k) - g'(x^k)\|. \tag{I}$$

Hieraus folgert man, dass es ein δ gibt, so dass mit einer Wahl $x^0 \in K_\delta(x^*)$ das Newtonverfahren stets superlinear konvergiert:

$\|g'(x^k)^{-1}\|$ ist auf $K_\delta(x^*)$ (für δ klein genug) nach Annahme beschränkt. Weiterhin folgt aufgrund der Stetigkeit von g', dass durch hinreichend kleines δ stets gilt:

$$\Big[\max_{\mu\in K_\delta(x^*)}\|g'(\mu)^{-1}\|\Big]\|g'(x) - g'(u)\| \leq \kappa < 1 \quad \forall x, y \in K_\delta(x^*).$$

Mit $x_0 \in K_\delta(x^*)$ liegt also induktiv durch (I) auch jedes x^k, und damit auch jedes $\xi^k \in \overline{x^k x^*}$ in $K_\delta(x^*)$. (I) liefert schließlich

$$\frac{\|x^* - x^{k+1}\|}{\|x^* - x^k\|} \leq \kappa < 1.$$

Das Verfahren ist also kontraktiv. Die Superlinearität folgt dann sofort aus

$$\|g'(\xi^k) - g'(x^K)\| \to 0 \quad \text{für } x^k \to x^*.$$

Lösung A.3.10: a) Beim Anwenden des semi-impliziten Runge-Kutta-Verfahrens ist in jedem Schritt das implizite Gleichungssystem

$$k_2 = f\big(t_n, y_{n-1} + \tfrac{1}{2}h\,k_1 + \tfrac{1}{2}h\,k_2\big)$$

zu lösen, d. h.: Gesucht ist eine Nullstelle von $g(\cdot)$ mit

$$g(k) := k - \underbrace{f\big(t_n, y_{n-1} + \tfrac{1}{2}h\,k_1 + \tfrac{1}{2}h\,k\big)}_{:=\hat{f}(k)}.$$

$\hat{f}(k)$ ist nun semi-monoton:

$$\begin{aligned} -\Big(\hat{f}(k) - \hat{f}(\tilde{k}), k - \tilde{k}\Big) &= -\frac{2}{h}\Big(f\big(t_n, y_{n-1} + \tfrac{1}{2}h\,k_1 + \tfrac{1}{2}h\,k\big) - f\big(t_n, \cdots + \tfrac{1}{2}h\,\tilde{k}\big) \\ &\qquad \big(\cdots + \tfrac{1}{2}h\,k\big) - \big(\cdots + \tfrac{1}{2}h\,\tilde{k}\big)\Big) \\ &\geq 0, \end{aligned}$$

aufgrund selbiger Eigenschaft für f. Analog zum Text folgt hieraus nun die Anwendbarkeit des Satzes von Newton-Kantorovich.

b) Im Fall diagonal-impliziter Runge-Kutta-Verfahren sind $R-1$ nichtlineare Gleichungssysteme der Dimension d zu lösen:

$$k_r = f\big(t_{n-1} + h\,a_r, y_{n-1} + h\sum_{s=1}^{r-1} b_{rs}\,k_s + hb_{rr}\,k_r\big).$$

Die Argumentation verläuft dann völlig analog.

Lösung A.3.11: Wir überprüfen die Voraussetzungen des Satzes über das gedämpfte Newton-Verfahren, wobei wir voraussetzen, dass die rechte Seite f unserer AWA semimonoton ist, und ihre erste Ableitung f_x L-stetig ist mit Konstante L'. Die Verfahrensfunktion des Newton-Verfahrens ist gerade (vgl. eine vorausgegangene Aufgabe):

$$G(k) = f\left(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}f\left(t_{n-1}, y_{n-1}\right) + \tfrac{1}{2}hk\right) - k$$

mit der Ableitung:

$$G'(k) = \tfrac{1}{2}hf_x\left(t_{n-1} + h, y_{n-1} + \tfrac{1}{2}f\left(t_{n-1}, y_{n-1}\right) + \tfrac{1}{2}hk\right) - I$$

Die Funktion G ist streng-monoton, denn aufgrund der Semi-Monotonie von f gilt:

$$\begin{aligned}-\left(G\left(k\right)-G\left(\bar{k}\right),k-\bar{k}\right) &= -\left(f\left(t_n,\ldots+\tfrac{1}{2}hk\right)-f\left(t_n,\ldots+\tfrac{1}{2}h\bar{k}\right),k-\bar{k}\right)\\ &\quad -\left(\bar{k}-k,k-\bar{k}\right)\\ &\geq \|k-\bar{k}\|^2.\end{aligned}$$

Hieraus folgt (s. Text), dass

$$-\left(G'\left(x\right)y,y\right)\geq\|y\|^2,\quad y\in\mathbb{R}^d,$$

und damit die Regularität von $G'\left(x\right)$ (Null kann nicht Eigenwert sein.). Dies sieht man wie folgt: Mit dem Fundamentalsatz der Differential- und Integralrechnung gilt für beliebigen Richtungsvektor $e\in\mathbb{R}^d$ und kleines $h\in\mathbb{R}$:

$$G(x+he)-G(x)=\int_0^1\frac{d}{ds}G(x+she)\,ds=\int_0^1 G'(x+she))\,ds\,he.$$

Damit gilt

$$h^2\|e\|^2\leq-(G(x+he)-G(x),he)=-h^2\Big(\int_0^1 G'(x+she))\,ds\,e,e\Big),$$

und nach Kürzen durch h^2 folgt für $h\to 0$:

$$\|e\|^2\leq-(G'(x)e,e).$$

Analog zum Text erschließen wir weiter:

$$\begin{aligned}\|G'\left(x\right)^{-1}\|^2 &= \sup_{y\neq0}\frac{\|G'\left(x\right)^{-1}y\|^2}{\|y\|^2}\\ &\leq \sup_{y\neq0}\frac{\left(G'\left(x\right)G'\left(x\right)^{-1}y,G'\left(x\right)^{-1}y\right)}{\|y\|^2}\\ &\leq \sup_{y\neq0}\frac{\|G'\left(x\right)^{-1}y\|}{\|y\|}=\|G'\left(x\right)^{-1}\|\end{aligned}$$

Und folglich $\|G'\left(x\right)^{-1}\|\leq 1$. Wir können also $\beta=1$ verwenden. Ferner ist G' L-stetig, denn:

$$\begin{aligned}\|G'\left(k\right)-G'\left(\bar{k}\right)\| &= \tfrac{1}{2}h\left\|f_x\left(t_n,\ldots+\tfrac{1}{2}hk\right)-f_x\left(t_n,\ldots+\tfrac{1}{2}h\bar{k}\right)\right\|\\ &\leq \tfrac{1}{2}hL'\left\|\tfrac{1}{2}hk-\tfrac{1}{2}h\bar{k}\right\|\\ &= \tfrac{1}{4}h^2L'\|k-\bar{k}\|\end{aligned}$$

Und somit $\gamma=\frac{1}{4}h^2L'$. Wir erhalten damit aus dem Satz über das gedämpfte Newton-Verfahren die Schrittweiten:

$$\lambda_k=\min\left(1,\frac{4}{\alpha_k h^2L'}\right).$$

Lösung A.3.12 (Praktische Aufgabe): Nicht verfügbar.

A.4 Kapitel 4

Lösung A.4.1: i) Die angegebene LMM hat die Koeffizienten ($R = 3$) $\alpha_3 = 1$, $\alpha_2 = \alpha$, $\alpha_1 = -\alpha$, $\alpha_0 = -1$, $\beta_3 = 0$, $\beta_2 = \frac{3+\alpha}{2}$, $\beta_1 = \frac{3+\alpha}{2}$, $\beta_0 = 0$. Ihr erstes charakteristisches Polynom lautet damit

$$\rho(\lambda) = \lambda^3 + \alpha\lambda^2 - \alpha\lambda - 1.$$

Es hat wegen $\lambda^3 + \alpha\lambda^2 - \alpha\lambda - 1 = (\lambda - 1)\left(\lambda^2 + (\alpha + 1)\lambda + 1\right)$ die Nullstellen

$$\lambda_1 = 1, \quad \lambda_{2/3} = -\tfrac{\alpha+1}{2} \pm \tfrac{1}{2}\sqrt{\alpha^2 + 2\alpha - 3}.$$

Für $-3 < \alpha < 1$ folgt

$$\lambda_{2/3} = -\tfrac{\alpha+1}{2} \pm \tfrac{1}{2}\sqrt{\alpha^2 + 2\alpha - 3} = -\tfrac{\alpha+1}{2} \pm \tfrac{1}{2}i\sqrt{3 - 2\alpha - \alpha^2}$$

und somit

$$|\lambda_{2/3}|^2 = \left(\tfrac{\alpha+1}{2}\right)^2 + \tfrac{3}{4} - \tfrac{1}{2}\alpha - \tfrac{1}{4}\alpha^2 = 1.$$

Wegen $3 - 2\alpha + \alpha^2 > 0$ für $-3 < \alpha < 1$ sind alle Nullstellen paarweise verschieden und damit die LMM nullstabil im Bereich $-3 < \alpha < 1$.

Umgekehrt ist für $\alpha > 3$ bzw. $\alpha < -1$ der Radikant in obiger Formel für $\lambda_{2,3}$ positiv und die $\lambda_{2,3}$ somit reell. Im Fall $\alpha < -3$ ist $-\alpha > 3$ und wegen

$$\lambda_2 = -\tfrac{\alpha+1}{2} + \tfrac{1}{2}\sqrt{\alpha^2 + 2\alpha - 3} \geq \tfrac{-\alpha-1}{2} > \tfrac{3-1}{2} = 1$$

ist dann $|\lambda_2| > 1$ und die LMM kann nicht nullstabil sein. Ebenso ist für $\alpha > 1$:

$$\lambda_3 = -\tfrac{\alpha+1}{2} - \tfrac{1}{2}\sqrt{\alpha^2 + 2\alpha - 3} \leq -\tfrac{\alpha+1}{2} < -\tfrac{1+1}{2} = -1$$

$|\lambda_3| > 1$ und die LMM kann ebenso nicht nullstabil sein. Im Grenzfall $\alpha = -3$ ist $\lambda_1 = \lambda_2 = \lambda_3 = 1$; ebenso folgt aus $\alpha = 1$ sofort $\lambda_2 = \lambda_3 = -1$.

ii) Um die Konsiztenzordnung der Formel zu bestimmen, betrachten wir

$$C_0 := \sum_{r=0}^{3} \alpha_r, \quad C_i := \frac{1}{i!}\sum_{r=0}^{3} r^i\alpha_r - \frac{1}{(i-1)!}\sum_{r=0}^{3} r^{i-1}\beta_r.$$

Zunächst ist

$$\begin{aligned}
C_0 &= -1 - \alpha + \alpha + 1 = 0,\\
C_1 &= (-\alpha + 2\alpha + 3) - \left(0 + \tfrac{3+\alpha}{2} + \tfrac{3+\alpha}{2} + 0\right) = 0.\\
C_2 &= \tfrac{1}{2}\left(-\alpha + 4\alpha + 9\right) - \left(\tfrac{3+\alpha}{2} + 3 + \alpha\right) = \tfrac{3}{2}\alpha + \tfrac{9}{2} - \tfrac{3}{2}\alpha - \tfrac{9}{2} = 0.
\end{aligned}$$

Weiter erhält man

$$\begin{aligned}
C_3 &= \tfrac{1}{6}\left(-\alpha + 8\alpha + 27\right) - \tfrac{1}{2}\left(\tfrac{3+\alpha}{2} + 6 + 2\alpha\right) = \tfrac{7}{6}\alpha + \tfrac{9}{2} - \tfrac{5}{4}\alpha - \tfrac{15}{4} = -\tfrac{1}{12}\alpha + \tfrac{3}{4}.\\
C_4 &= \tfrac{1}{24}\left(-\alpha + 16\alpha + 81\right) - \tfrac{1}{6}\left(\tfrac{3+\alpha}{2} + 12 + 4\alpha\right) = \tfrac{5}{8}\alpha + \tfrac{27}{8} - \tfrac{3}{4}\alpha - \tfrac{27}{12} = -\tfrac{1}{8}\alpha + \tfrac{9}{8},
\end{aligned}$$

Die LMM ist also für $\alpha \in (-3, 1)$ von 2. Ordnung und hätte theoretisch eine Konsistenzordnung von 4 für die Wahl $\alpha = 9$. Hier geht allerdings die Nullstabilität verloren.

Lösung A.4.2: Zunächst formen wir die Differentialgleichung durch Einführen der Hilfsfunktionen $u_0(t) := u(t),\ u_1(t) := u'(t)$ um in ein System erster Ordnung:

$$\begin{bmatrix} u_0'(t) \\ u_1'(t) \end{bmatrix} = \begin{bmatrix} u_1(t) \\ -20u_1(t) - 19u_0(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -19 & -20 \end{bmatrix}}_{=:A} \begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix}, \quad t \geq 0, \quad \begin{bmatrix} u_0(0) \\ u_1(0) \end{bmatrix} = \begin{bmatrix} 1 \\ -10 \end{bmatrix}.$$

Um die Kontraktionseigenschaft der Fixpunktiteration zu sichern, benötigt man die L-Konstante der Iterationsabbildung

$$g(x) = y_{n-1} + \tfrac{1}{12}h\left(5f(t_n, x) + 8f_{n-1} - f_{n-2}\right).$$

Mit der L-Konstante L_f von f ist diese gerade $L_g = |\beta_2|hL_f = \frac{5}{12}hL_f$. Die Schrittweite h muss also so bestimmt werden, dass

$$h < (|\beta_2|L_f)^{-1} = \tfrac{12}{5}L_f^{-1}.$$

Zur Bestimmung von L_f sei festgestellt:

$$\|f(t,x) - f(t,y)\| \leq \|A\|\|x - y\|,$$

mit einer beliebigen, verträglichen Matrizennorm $\|.\|$. Um hiermit eine möglichst kleine Lipschitzkonstante $L_f = \|A\|$ zu erhalten, wählen wir die maximale Spaltensummennorm: $\|A\|_1 = 21$ ($\|A\|_\infty = 39$). Es ergibt sich damit

$$h < \frac{1}{|\beta_2|L_f} = \frac{12}{21 \cdot 5} = \frac{4}{35} \approx 0{,}114.$$

Lösung A.4.3: Zu bestimmen ist zunächst das Stabilitätspolynom

$$\pi(\lambda; z) := \sum_{r=0}^{R} \left[\alpha_r - z\beta_r\right] \lambda^r$$

und dann dessen Nullstellen.

Es ist $y_n - y_{n-2} = 2hf_{n-1}$ und somit $R = 2,\ \alpha_2 = 1,\ \alpha_1 = 0,\ \alpha_0 = -1,\ \beta_2 = 0,\ \beta_1 = 2,\ \beta_0 = 0$. Damit ist

$$\pi(\lambda; z) = \lambda^2 - 2z\lambda - 1.$$

Die Nullstellen ergeben sich zu

$$\lambda_{1/2} = z \pm \sqrt{z^2 + 1}.$$

Betrachten wir zunächst das Stabilitätsintervall: Für $z > 0$ ist $\lambda_1 > 1$, für $z < 0$ ist $\lambda_2 < -1$. Für $z = 0$ folgt $\lambda_{1/2} = \pm 1$. Da dies zwei einfache Nullstellen sind, ist das Stabilitätsintervall $\mathrm{SI} = \{0\}$.

Um das Stabilitätsgebiet zu bestimmen, bedienen wir uns der im Text beschriebenen Vereinfachung. Seien $\lambda_1(0)$ und $\lambda_2(0)$ die Nullstellen des Polynoms $\pi(\lambda;0) = \lambda^2 - 1$. Es ist also $\lambda_1(0) = 1,\ \lambda_2(0) = -1$. Wir machen den Ansatz

$$\lambda_1(z) := 1 + \gamma_1 z + O(z^2),$$
$$\lambda_2(z) := -1 + \gamma_2 z + O(z^2).$$

Einsetzen in $\pi(\lambda; z) = 0$ liefert

$$1 + 2\gamma_1 z - 2z - 1 + O(z^2) = 0 \quad \text{sowie} \quad 1 - 2\gamma_2 z + 2z - 1 + O(z^2) = 0$$

bzw. $\gamma_1 = 1$ sowie $\gamma_2 = 1$. Damit ergibt sich $\lambda_1(z) \approx 1 + z$ und $\lambda_2(z) \approx -1 + z$. Nun muss gelten ($z := x + iy$)

$$|\lambda_1(z)| \le 1 \quad \Leftrightarrow \quad (x+1)^2 + y^2 \le 1 \qquad \text{sowie} \qquad |\lambda_2(z)| \le 1 \quad \Leftrightarrow \quad (x-1)^2 + y^2 \le 1.$$

Dies ist nur erfüllt für $(x, y) = (0, 0)$.

Zusatz: Eine genauere Analyse zeigt jedoch: Für $z := iy$ ist $\lambda_{1/2} = iy \pm \sqrt{1 - y^2}$. Im Fall $|y| < 1$ ist $\lambda_1 \neq \lambda_2$ und $|\lambda_{1/2}| = y^2 + 1 - y^2 = 1$. Für $y = \pm 1$ liegt eine doppelte Nullstelle $\lambda = \pm i$ mit $|\lambda| = 1$ vor. Falls $|y| > 1$, so ist

$$\lambda_{1/2} = iy \pm \sqrt{1 - y^2} = i\big(y \pm \sqrt{y^2 - 1}\big),$$

und mindestens eine der Nullstellen hat einen Betrag größer als 1. Ebenso hat mindestens eine der Nullstellen einen Betrag größer als 1, falls $x \neq 0$. Somit folgt $\mathrm{SG} = \{z \in \mathbb{C} \,|\, z = (0, y),\ y \in [-1, 1]\}$.

Lösung A.4.4 (Praktische Aufgabe): Nicht verfügbar.

Lösung A.4.5: Die Rückwärtsdifferenzenformeln der Stufen $R = 1, 2, 3$ lauten:

- $R = 1$: $y_n - y_{n-1} = hf_n$, Koeffizienten $\alpha_1 = 1,\ \alpha_0 = -1,\ \beta_1 = 1$.
- $R = 2$: $y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = \frac{2}{3}hf_n$, Koeffizienten $\alpha_2 = 1,\ \alpha_1 = -\frac{4}{3},\ \alpha_0 = \frac{1}{3},\ \beta_2 = \frac{2}{3}$.
- $R = 3$: $y_n - \frac{18}{11}y_{n-1} + \frac{9}{11}y_{n-2} - \frac{2}{11}y_{n-3} = \frac{6}{11}hf_n$, Koeffizienten $\alpha_3 = 1,\ \alpha_2 = -\frac{18}{11},\ \alpha_1 = \frac{9}{11},\ \alpha_0 = -\frac{2}{11},\ \beta_3 = \frac{6}{11}$.

Für die Konvergenz ist Konsistenz und Nullstabilität zu überprüfen.

$R = 1$: Konsistenz folgt aus

$$\sum_{r=0}^{1} \alpha_r = -1 + 1 = 0, \qquad \sum_{r=0}^{1} r\alpha_r = -1 = \sum_{r=0}^{1} \beta_r.$$

Das erste charakteristische Polynom ist

$$\rho(\lambda) = \sum_{r=0}^{1} \alpha_r \lambda^r = \lambda - 1,$$

und hat die Nullstelle $\lambda = 1$. Somit ist die BDF nullstabil und damit konvergent.

$R = 2$:

$$\sum_{r=0}^{2} \alpha_r = \tfrac{1}{3} - \tfrac{4}{3} + 1 = 0, \qquad \sum_{r=0}^{2} r\alpha_r = -\tfrac{4}{3} + 2 = \tfrac{2}{3} = \sum_{r=0}^{2} \beta_r.$$

Das erste charakteristische Polynom ist

$$\rho(\lambda) = \sum_{r=0}^{2} \alpha_r \lambda^r = \lambda^2 - \tfrac{4}{3}\lambda + \tfrac{1}{3},$$

mit den Nullstellen $\lambda_1 = 1$ und $\lambda_2 = \frac{1}{3}$. Somit ist die BDF nullstabil und damit konvergent.

$R = 3$:

$$\sum_{r=0}^{3} \alpha_r = -\tfrac{2}{11} + \tfrac{9}{11} - \tfrac{18}{11} + 1 = 0, \qquad \sum_{r=0}^{3} r\alpha_r = \tfrac{9}{11} - \tfrac{36}{11} + 3 = \tfrac{6}{11} = \sum_{r=0}^{3} \beta_r.$$

Das erste charakteristische Polynom ist

$$\rho(\lambda) = \sum_{r=0}^{3} \alpha_r \lambda^r = \lambda^3 - \tfrac{18}{11}\lambda^2 + \tfrac{9}{11}\lambda - \tfrac{2}{11},$$

und hat die Nullstellen $\lambda_1 = 1, \quad \lambda_{2/3} = \frac{7}{22} \pm \frac{1}{22} i\sqrt{39}$. Wegen

$$|\lambda_{2/3}|^2 = \tfrac{7^2}{22^2} + \tfrac{39}{22^2} = \tfrac{49+39}{484} = \tfrac{2}{11}$$

ist die BDF nullstabil und damit konvergent.

Für die Berechnung der Startwerte wird ein Verfahren der Ordnung $p^* \geq p - 1$ benötigt (p Ordnung der BDF). Bestimme also die Ordnungen der Rückwärtsdifferenzenformeln für $R = 2, 3$: ($C_0 = C_1 = 0$ bereits für Konsistenz nötig)

$R = 2$:

$$C_2 = \tfrac{1}{2} \sum_{r=0}^{2} r^2 \alpha_r - \sum_{r=0}^{2} r\beta_r = -\tfrac{2}{3} + 2 - \tfrac{4}{3} = 0,$$
$$C_3 = \tfrac{1}{6} \sum_{r=0}^{2} r^3 \alpha_r - \tfrac{1}{2} \sum_{r=0}^{2} r^2 \beta_r = -\tfrac{2}{9} + \tfrac{4}{3} - \tfrac{4}{3} = -\tfrac{2}{9}$$

Die Ordnung ist also $p = 2$. Es genügt also z. B. das implizite Eulerverfahren als Startprozedur.

$R = 3$:

$$C_2 = \tfrac{1}{2}\sum_{r=0}^{3} r^2\alpha_r - \sum_{r=0}^{3} r\beta_r = \tfrac{9}{22} - \tfrac{36}{11} + \tfrac{9}{2} - \tfrac{18}{11} = 0,$$
$$C_3 = \tfrac{1}{6}\sum_{r=0}^{3} r^3\alpha_r - \tfrac{1}{2}\sum_{r=0}^{3} r^2\beta_r = \tfrac{3}{22} - \tfrac{24}{11} + \tfrac{9}{2} - \tfrac{27}{11} = 0,$$
$$C_4 = \tfrac{1}{24}\sum_{r=0}^{3} r^4\alpha_r - \tfrac{1}{6}\sum_{r=0}^{3} r^3\beta_r = \tfrac{3}{88} - \tfrac{12}{11} + \tfrac{27}{8} - \tfrac{27}{11} = -\tfrac{3}{22}$$

Die Ordnung ist also $p = 3$. Es genügt also z. B. die Trapezregel als Startprozedur.

Lösung A.4.6: Wir zitieren zunächst einige Resultate aus dem Text: Für die Iterierten $y_n^{(k)}$ im Korrektor-Verfahren gilt die Abschätzung:

$$\|y_n^{(k)} - y_n\| \le q^k \|y_n^{(0)} - y_n\|$$

wobei y_n die exakte Lösung des Korrektors und $q = h\beta_R^{(K)}$ ist. Ferner gilt

$$y_n^{(0)} - u_n = h\tau_{(P)}^h \left(1 - O\left(h\right)\right),$$
$$y_n - u_n = h\tau_{(C)}^h \left(1 - O\left(h\right)\right)$$

mit den Abschneidefehlern $\tau_{(P)}^h$, $\tau_{(C)}^h$ des Prädiktor-, Korrektorverfahrens. Für die Abschneidefehler gilt (bis auf das Vorzeichen):

$$\tau_{(P)}^h = C_{m^{(P)}+1}^{(P)} h^{m^{(P)}} u_n^{\left(m^{(P)}+1\right)} + O\left(h^{m^{(P)}+1}\right)$$

$$\tau_{(C)}^h = C_{m^{(C)}+1}^{(C)} h^{m^{(C)}} u_n^{\left(m^{(C)}+1\right)} + O\left(h^{m^{(C)}+1}\right)$$

Damit erhalten wir nun:

$$\begin{aligned}
\|y_n^{(k)} - u_n\| &\le \|y_n^{(k)} - y_n\| + \|y_n - u_n\| \\
&\le q^k \|y_n^{(0)} - y_n\| + \|y_n - u_n\| \\
&\le q^k \|y_n^{(0)} - u_n\| + \left(q^k + 1\right) \|y_n - u_n\| \\
&\le q^k C_{m^{(P)}+1}^{(P)} h^{m^{(P)}+1} \left\|u_n^{\left(m^{(P)}+1\right)}\right\| + O\left(h^{m^{(P)}+2+k}\right) \\
&+ \left(q^k + 1\right) C_{m^{(C)}+1}^{(C)} h^{m^{(C)}+1} \left\|u_n^{\left(m^{(C)}+1\right)}\right\| + \left(q^k + 1\right) O\left(h^{m^{(C)}+2}\right)
\end{aligned}$$

und somit, falls $m^{(P)} + k \le m^{(C)}$:

$$\|y_n^{(k)} - u_n\| \le c h^{m^{(P)}+1+k} + O\left(h^{m^{(P)}+2+k}\right)$$

und im Fall $m^{(P)} + k > m^{(C)}$:

$$\|y_n^{(k)} - u_n\| \le C_{m^{(C)}+1} h^{m^{(C)}+1} \left\|u_n^{\left(m^{(C)}+1\right)}\right\| + O\left(h^{m^{(C)}+2}\right)$$

und somit die erste Behauptung. Die zweite Behauptung ergibt sich durch Ablesen aus der letzten Ungleichung.

Lösung A.4.7: Das betrachtete System lautet

$$\begin{bmatrix} u'(t) \\ v'(t) \\ w'(t) \end{bmatrix} = \underbrace{\begin{bmatrix} -10 & -100 & 0 \\ 100 & -10 & 0 \\ 1 & 1 & -t \end{bmatrix}}_{=: A(t)} \begin{bmatrix} u(t) \\ v(t) \\ w(t) \end{bmatrix}.$$

Zu bestimmen sind nun die Eigenwerte von $A(t)$:

$$\begin{aligned} \det(\lambda I - A(t)) &= \det \begin{bmatrix} \lambda+10 & 100 & 0 \\ -100 & \lambda+10 & 0 \\ -1 & -1 & \lambda+t \end{bmatrix} = (\lambda+t) \det \begin{bmatrix} \lambda+10 & 100 \\ -100 & \lambda+10 \end{bmatrix} \\ &= (\lambda+t)\left((\lambda+10)^2+10000\right) = (\lambda+t)(\lambda^2+20\lambda+10100). \end{aligned}$$

Die Eigenwerte sind also

$$\lambda_1 = -t \quad \text{sowie} \quad \lambda_{2/3} = -10 \pm \sqrt{100-10100} = -10 \pm 100i.$$

Um eine möglichst hohe Ordnung zu bekommen, fordern wir von der LMM nur A(α)-Stabilität. Bestimme also den Winkel α: Wegen $\lambda_1 \in \mathbb{R}$ und $\lambda_3 = \overline{\lambda_2}$ erhalten wir

$$\alpha = \arctan \tfrac{100}{10} = \arctan 10 \approx 84{,}3°.$$

Es kommen also nur Rückwärtsdifferenzenformeln der Stufe $R \leq 3$ in Frage. Nach dem Text erhält man bei Verwendung der 3-stufigen BDF nach Konstruktion die Ordnung $m = 3$.

Lösung A.4.8 (Praktische Aughabe): Nicht verfügbar.

Lösung A.4.9: a) *Exponentielle Stabilität* ist eine Eigenschaft der Lösung u einer AWA. Diese besagt, dass hinreichend kleine Störungen exponentiell abklingen, genauer: Es gibt $\delta,\ A,\ \lambda > 0$, so dass für alle Störungen $\|w^*\| \leq \delta$ zu einem Zeitpunkt $t^* \geq t_0$ für die Lösung der gestörten AWA $v(t) = f(t, v(t)),\ t \geq t^*,\ v(t) = u(t) + w^*$, stets gilt

$$\|u(t) - v(t)\| \leq Ae^{-\lambda(t-t^*)}.$$

b) *Diskrete Stabilität* ist eine Eigenschaft eines Ein- oder Mehrschrittverfahrens, welche es einem erlaubt die Differenz zweier Gitterfunktionen $\{y_n\}_{n\geq 0},\ \{z_n\}_{n\geq 0}$ mit Hilfe der Verfahrensvorschrift L_h darzustellen:

$$\|y_n - z_n\| \leq Ke^{\Gamma(t_n - t_0)}\Big\{ \max_{0\leq\nu\leq R-1} + \sum_{\nu=R}^{n} h_\nu \|L_h y_h - L_h z_h\| \Big\}.$$

Aus diskreter Stabilität folgt mit Konsistenz die (lokale) Konvergenz einer Methode.

c) Numerische Stabilität ist eine Eigenschaft einer (globalen) Lösung $y_{nn\geq 0}$ eines Einschrittverfahrens, $y_n = y_{n-1} + h_n F(h_n; t_n, y_n, y_{n-1})$, $n \geq 0, y_0 = u_0$ (oder Mehrschrittverfahrens): $\{y_n\}$ heißt numerisch stabil, falls für jede Lösung $\{z_n\}_{n\geq n^*}$ von

$$z_n = z_{n-1} + h_n F(h_n; t_n, z_n, z_{n-1}), n \geq n^*,\ z_{n^*} = u_{n^*} + w^*$$

mit einer hinreichend kleinen Störung $\|w^*\| \leq \delta$ stets gilt:

$$\|z_n - y_n\| \to 0,\ \text{für } n \to \infty\,.$$

d) Eine lineare Mehrschrittmethode heißt *Null-stabil*, falls für die Nullstellen λ_i des ersten charakteristischen Polynoms $\rho(\lambda) = \sum_{r=0}^{R} \alpha_r \lambda^r v$ stets gilt: $|\lambda_i| \leq 1$, und falls λ_i eine mehrfache Nullstelle ist, sogar $|\lambda_i| < 1$. Null-Stabilität ist eine notwendige (und zusammen mit L-Stetigkeit der Verfahrensfunktion hinreichende) Bedingung für die diskrete Stabilität von linearen Mehrschrittmethoden.

e), f) A-Stabilität und A(0)-Stabilität sind Begriffe aus der numerischen Stabilitätsanalyse. Ein Ein- oder Mehrschrittverfahren heißt A-stabil, falls die gesamte negative, komplexe Halbebene $\{z : \operatorname{Re} z \leq 0\}$ im Stabilitätsgebiet liegt, und A(0)-stabil, falls dies die für die negative, reelle Achse zutrifft.

A.5 Kapitel 5

Lösung A.5.1: a) Wir betrachten das Intervall $[0, 1]$ und wollen $u(1)$ berechnen. Wir verwenden die Grundschrittweite H, d. h. wir unterteilen das Intervall $[0, 1]$ in $N = \frac{1}{H}$ Teile. Auf jeden Teilintervall verwenden wir jetzt den folgenden Algorithmus:

1. Es sei $y(t_k)$ mit $t_k = kH$ und $0 \leq k \leq N - 1$ berechnet (Ziel ist es, per Extrapolation der Zwischenwerte $a(h_0), \ldots, a(h_5)$ mit $h_0 = H/2, \ldots, h_5 = H/16$ den nächsten Wert $y(t_{k+1})$ zu berechnen)
2. Setze $n_0 = 2, n_1 = 4, n_2 = 6, n_3 = 8, n_4 = 12, n_5 = 16$.
3. Setze $i = 0$.
4. Berechne
$$\eta(t_k + \nu h_i; h_i), \quad h_i = \frac{H}{n_i}, \quad \nu = 1, \ldots, n_i + 1,$$
mit Hilfe der Mittelpunktsregel gestartet durch die Polygonzugmethode,
$$\begin{aligned} \eta(t_k + h_i; h_i) &= y_{t_k} + h_i f(t_k, y(t_k)) \\ \eta(t_k + (\nu+1)h_i; h_i) &= \eta(t_k + (\nu-1)h_i; h_i) + 2h_i f(t_k + \nu h_i, \eta(t_k + \nu h_i; h_i)), \end{aligned}$$
für $\nu = 1, \ldots n_i$, und setze
$$a(h_i) = \tilde{\eta}(t_{k+1}; h_i) = \tfrac{1}{4}\{\eta(t_{k+1} - h_i; h_i) + 2\eta(t_{k+1}; h_i) + \eta(t_{k+1} + h_i; h_i)\}.$$

5. Falls $i \leq 4$ setze $i \leftarrow i+1$ und gehe zu (4).

6. Berechne mit $a(h_i) = T_{i0}$ die Werte T_{ii} des Extrapolationstableaus mit Hilfe der Rekursionsformel
$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{(h_{i-k}/h_i)^\gamma - 1}.$$
und erhalte $T_{5,5}$ als Näherung für $u(t_{k+1})$.

Ordnung des Verfahren: Mithilfe des Satzes von Gragg erhalten wir mit $m = 5$ die Ordnung 12.

b) Anzahl der Funktionsauswertungen in dem eben beschriebenen Algorithmus für den Schritt von t_k auf t_{k+1}: Für jedes i:

- Polygonzugmethode: 1 Funktionsauswertung (nur beim ersten Schritt, d. h. für i=0 nötig)
- n_i-mal Anwendung der Mittelpunktsregel: n_i Funktionsauswertungen.
- Mittelung: keine weiteren Funktionsauswertungen nötig.

Damit ergeben sich bei der hier betrachteten Folge der n_i
$$(2+1) + 4 + 6 + 8 + 12 + 16 = 49$$
Funktionsauswertungen. Da wir insgesamt auf $N = \frac{1}{H}$ Intervallen extrapolieren, benötigen wir also $49/H$ Funktionsauswertungen, um eine Näherung für $u(1)$ zu erhalten.

Lösung A.5.2: Analog zum klassischen Graggschen Extrapolationsverfahren berechnet man y_{k+1} mit einer Basisschrittweite H ausgehend vom bereits berechneten y_k durch Extrapolation:

1. Setze $n_0 = 2,\ n_1 = 4,\ n_2 = 6,\ n_3 = 8,\ n_4 = 12,\ n_5 = 16.$
2. Setze $i = 0$.
3. Es ist $\eta\big(t_k; h_i\big) = y_k$. Berechne
$$\eta\big(t_k + \nu h_i; h_i\big),\ h_i = \frac{H}{n_i},\ \nu = 1, \cdots, n_i$$
mit Hilfe des impliziten Eulerverfahrens:
$$\eta\big(t_k + \nu h_i; h_i\big) = \eta\big(t_k + (\nu-1)h_i; h_i\big) + f\big(t_k + \nu h_i, \eta(t_k + \nu h_i; h_i)\big).$$

4. Setze $a(h_i) = \eta(t_k + n_i h_i; h_i)$.
5. Falls $i < 5$ setze $i \leftarrow i+1$ und gehe zu (3).
6. Bestimme $y_{k+1} = a(0)$ durch Extrapolation von $a(h)$:
$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{(h_{i-k}/h_i)^1 - 1},\ T_{i0} = a(h_i).$$
Setze $y_{k+1} = T_{55}$.

a) Die Ordnung des zugrundeliegenden Verfahrens ist $\gamma = 1$, so dass sich bei $m = 5$ die Ordnung 6 ergibt.

b) Für einen Extrapolationsschritt, d. h.: $m = 1$, hat man mit $n_0 = 2$, $n_1 = 4$ zu berechnen:

$$n_0: \quad \begin{cases} \eta\big(t_k + \frac{H}{2}; \frac{H}{2}\big) = y_k + \frac{H}{2} f\big(t_k + \frac{H}{2}, \eta\big(t_k + \frac{H}{2}; \frac{H}{2}\big)\big), \\ \eta\big(t_k + H; \frac{H}{2}\big) = \eta\big(t_k + \frac{H}{2}; \frac{H}{2}\big) + \frac{H}{2} f\big(t_k + H, \eta\big(t_k + H; \frac{H}{2}\big)\big). \end{cases}$$

$$n_1: \quad \begin{cases} \eta\big(t_k + \frac{H}{4}; \frac{H}{4}\big) = y_k + \frac{H}{4} f\big(t_k + \frac{H}{4}, \eta\big(t_k + \frac{H}{4}; \frac{H}{4}\big)\big), \\ \qquad\qquad \vdots \\ \eta\big(t_k + H; \frac{H}{4}\big) = \eta\big(t_k + \frac{3H}{4}; \frac{H}{4}\big) + \frac{H}{4} f\big(t_k + H, \eta\big(t_k + H; \frac{H}{4}\big)\big). \end{cases}$$

Angewendet auf das skalare Testproblem

$$u'(t) = \lambda u(t)$$

erhält man

$$\eta\big(t_k + H; \tfrac{H}{2}\big) = \big(1 - \tfrac{1}{2} H\lambda\big)^{-2} y_k,$$
$$\eta\big(t_k + H; \tfrac{H}{4}\big) = \big(1 - \tfrac{1}{4} H\lambda\big)^{-4} y_k.$$

Extrapolation:

$$y_{k+1} = \frac{2^1 \eta\big(t_k + H; \frac{H}{4}\big) - \eta\big(t_k + H; \frac{H}{2}\big)}{2^1 - 1} \quad = \Big[2\big(1 - \tfrac{1}{4} H\lambda\big)^{-4} - \big(1 - \tfrac{1}{2} H\lambda\big)^{-2}\Big] y_k.$$

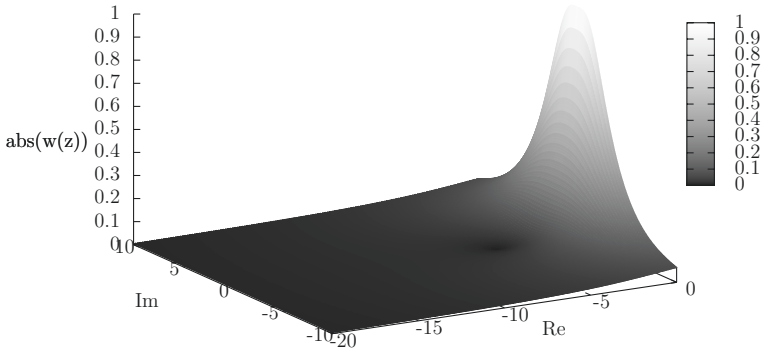
Der Verstärkungsfaktor ist also:

$$w(z) = 2\big(1 - \tfrac{1}{4} z\big)^{-4} - \big(1 - \tfrac{1}{2} z\big)^{-2}.$$

Diskussion für die reelle Achse:

Es ist $w(0) = 1$ und $\lim_{x \to -\infty} w(x) = 0$. Die Funktion $w(x)$ besitzt im Negativen genau ein lokales Minimum bei $x_0 \approx -8.25$ mit $w(x_0) \approx -0.015$ und genau einen Wendepunkt bei $x_1 \approx -11.7$. Somit ist stets $|w(x)| \leq 1$ für alle $x \leq 0$.

Diskussion für die komplexe Halbebene: Man betrachte den folgenden Plot von $|w(z)|$:



Lösung A.5.3: Sei $y_n = \eta(t_n; \frac{h}{2})$ berechnet. Mit einem expliziten Euler-Schritt bestimmen wir

$$\eta(t_n + \tfrac{h}{2}; \tfrac{h}{2}) = \eta(t_n; \tfrac{h}{2}) + \tfrac{h}{2} f\left(t_n, \eta(t_n; \tfrac{h}{2})\right) = y_n + \tfrac{h}{2} f(t_n, y_n)$$

und anschließend mit der Mittelpunktsregel

$$\begin{aligned}\eta(t_n + h; \tfrac{h}{2}) &= \eta(t_n; \tfrac{h}{2}) + 2\tfrac{h}{2} f\left(t_n + \tfrac{h}{2}, \eta(t_n + \tfrac{h}{2}; \tfrac{h}{2})\right) \\ &= y_n + h f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\right)\end{aligned}$$

sowie

$$\begin{aligned}\eta(t_n + \tfrac{3}{2}h; \tfrac{h}{2}) &= \eta(t_n + \tfrac{h}{2}; \tfrac{h}{2}) + 2\tfrac{h}{2} f\left(t_n + h, \eta(t_n + h; \tfrac{h}{2})\right) \\ &= y_n + \tfrac{h}{2} f(t_n, y_n) + h f\left(t_n + h, y_n + h f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\right)\right).\end{aligned}$$

Nun wird gesetzt

$$\begin{aligned}y_{n+1} &:= \tfrac{1}{4}\left\{\eta(t_n + \tfrac{h}{2}; \tfrac{h}{2}) + 2\eta(t_n + h; \tfrac{h}{2}) + \eta(t_n + \tfrac{3}{2}h; \tfrac{h}{2})\right\} \\ &= \tfrac{1}{4}\Big\{ y_n + \tfrac{h}{2} f(t_n, y_n) + 2y_n + 2h f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\right) \\ &\quad + y_n + \tfrac{h}{2} f(t_n, y_n) + h f\left(t_n + h, y_n + h f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\right)\right) \Big\} \\ &= y_n + h\Big\{ \tfrac{1}{4} f(t_n, y_n) + \tfrac{1}{2} f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\right) \\ &\quad + \tfrac{1}{4} f\left(t_n + h, y_n + h f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\right)\right) \Big\} \\ &= y_n + h\left\{\tfrac{1}{4} k_1 + \tfrac{1}{2} k_2 + \tfrac{1}{4} k_3\right\}\end{aligned}$$

mit

$$k_1 := f(t_n, y_n), \quad k_2 := f(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} k_1), \quad k_3 := f(t_n + h, y_n + h k_2).$$

Entsprechend dem Extrapolationssatz hat diese Formel die Ordnung 2 ($m = 0$, da keine Extrapolationsschritte durchgeführt werden). Alternativ stellt man fest, dass die Taylorreihe der Runge-Kutta-Formel mit der von f bis zur Ordnung 2 übereinstimmt.

Angewendet auf das Modellproblem $u'(t) = f(t, u(t))$ mit $f(t, x) = \lambda x$ erhält man

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{4}h\lambda y_n + \tfrac{1}{2}h\lambda\left(y_n + \tfrac{1}{2}h\lambda y_n\right) + \tfrac{1}{4}h\lambda\left(y_n + h\lambda\left(y_n + \tfrac{1}{2}h\lambda y_n\right)\right) \\ &= \left(1 + h\lambda + \tfrac{1}{2}(h\lambda)^2 + \tfrac{1}{8}(h\lambda)^3\right) y_n. \end{aligned}$$

Es ist also

$$\omega(z) = 1 + z + \tfrac{1}{2}z^2 + \tfrac{1}{8}z^3 = \tfrac{1}{8}\left(z^3 + 4z^2 + 8z + 8\right) = \tfrac{1}{8}(z+2)\left((z+1)^2 + 3\right).$$

Wegen $(z+1)^2 + 3 > 0$, unterscheiden wir die folgenden Fälle:

$z \geq -2$: Dann ist $|\omega(z)| = \frac{1}{8}(z+2)\left((z+1)^2+3\right)$ und es gilt $|\omega(z)| \leq 1$, wenn $\frac{1}{8}(z+2)\left((z+1)^2+3\right) \leq 1$ bzw. $z^3+4z^2+8z \leq 0$. Wegen $z^3+4z^2+8z = z\left(z^2+4z+8\right) = z\left((z+2)^2+4\right)$ ist dies nur für $z \in [-2, 0]$ erfüllt.

$z < -2$: Dann ist $|\omega(z)| = -\frac{1}{8}(z+2)\left((z+1)^2+3\right)$ und es gilt $|\omega(z)| \leq 1$, wenn $-\frac{1}{8}(z+2)\left((z+1)^2+3\right) \leq 1$ bzw. $z^3+4z^2+8z+16 \geq 0$. Wir suchen also eine Nullstelle von $g(z) := z^3 + 4z^2 + 8z + 16$. Mit $g'(z) = 3z^2 + 8z + 8$ und dem Startwert $z_0 := -2$ liefert das Newtonverfahren

$$z_{n+1} := z_n - \frac{g(z_n)}{g'(z_n)} = \frac{2z_n^3 + 4z_n^2 - 16}{3z_n^2 + 8z_n + 8}$$

die folgenden Approximationen: $z_1 = -4$, $z_2 = -3.333$, $z_3 = -3.111$, $z_4 = -3.088$, $z_5 = -3.087$, $z_6 = -3.087$. Wegen $|\omega(-3)| = \frac{7}{8} < 1$ ist die Bedingung erfüllt für $z \in [-3.1, 0]$.

Zusammengenommen erhalten wir für das Stabilitätsintervall $\mathrm{SI} = [-3.1, 0]$.

Lösung A.5.4 (Praktische Aufgabe): Nicht verfügbar.

A.6 Kapitel 6

Lösung A.6.1: Wir formen das System um in

$$\begin{aligned} Mu'(t) &= b(t) - Au(t) - N(u(t))u(t) - Bp(t), \\ 0 &= B^T u(t). \end{aligned}$$

Da M regulär ist, können wir mit M^{-1} multiplizieren und erhalten

$$\begin{aligned} u'(t) &= M^{-1}b(t) - M^{-1}Au(t) - M^{-1}N(u(t))u(t) - M^{-1}Bp(t), \\ 0 &= B^T u(t). \end{aligned}$$

Mit den Bezeichnungen $\tilde{b}(t) := M^{-1}b(t)$, $\tilde{A} := M^{-1}A$, $\tilde{N}(u(t)) := M^{-1}N(u(t))$ und $\tilde{B} := M^{-1}B$ erhalten wir nach Differenzieren der zweiten Gleichung

$$\begin{aligned} u'(t) &= \tilde{b}(t) - \tilde{A}u(t) - \tilde{N}(u(t))u(t) - \tilde{B}p(t), \\ 0 &= B^T u'(t) = B^T \Big(\tilde{b}(t) - \tilde{A}u(t) - \tilde{N}(u(t))u(t) - \tilde{B}p(t) \Big) . \end{aligned}$$

Bezeichnen wir die Spalten von $\tilde{N}(u(t))$ mit $\tilde{N}_i(u(t))$ und die Zeilen von $u(t)$ mit $u_i(t)$, so können wir die dies umformen zu

$$\begin{aligned} u'(t) &= \tilde{b}(t) - \tilde{A}u(t) - \tilde{N}(u(t))u(t) - \tilde{B}p(t), \\ 0 &= B^T \Big(\tilde{b}(t) - \tilde{A}u(t) - \sum_{i=1}^{n} \tilde{N}_i(u(t)) \cdot u_i(t) - \tilde{B}p(t) \Big). \end{aligned}$$

Durch nochmaliges Differenzieren der zweiten Gleichung erhalten wir

$$\begin{aligned} u'(t) &= \tilde{b}(t) - \tilde{A}u(t) - \tilde{N}(u(t))u(t) - \tilde{B}p(t), \\ 0 &= B^T\tilde{b}'(t) - B^T\tilde{A}u'(t) - B^T\tilde{B}p'(t) \\ &\quad - B^T \sum_{i=1}^{n} \Big\{ \Big(\sum_{j=1}^{n} \frac{\partial \tilde{N}_i(u(t))}{\partial u_j} \cdot u_j'(t) \Big) \cdot u_i(t) + \tilde{N}_i(u(t)) \cdot u_i'(t) \Big\}. \end{aligned}$$

Nach erneutem Ersetzen von $u'(t)$ in der zweiten Gleichung durch die erste erhält man, falls $B^T\tilde{B} = B^TM^{-1}B$ regulär ist, eine Gleichung zur Bestimmung von $p'(t)$ in Abhängigkeit von $u(t)$ und $p(t)$. Die DAE hat also den Index 2 und ist lösbar, wenn $B^TM^{-1}B$ regulär ist, was bedeutet, dass B Rang m hat.

Lösung A.6.2 (Praktische Aufgabe): Nicht verfügbar.

Lösung A.6.3:

a) (Lokale) L-Stetigkeit von $f(t,x)$ bzgl. des Arguments x;

b) d-dimensionaler Vektorraum;

c) Eine stetige Funktion $w(t) \geq 0$, die durch ihr Zeitintegral beschränkt ist, hat höchsten exponentielles Wachstum;

d) Ja, nach Fortsetzungssatz, da $f(t,x)$ bzgl. x gleichmäßig beschränkt ist;

e) $y_n = y_{n-1} + h_nF(h_n; t_n, y_n, y_{n-1})$, $n \geq 1$, $y_0 = u^0$;
$\tau_n = h_n^{-1}(u_n - u_{n-1}) - F(h_n; t_n, u_n, u_{n-1})$, $u_n := u(t_n)$;

f) $y_n = y_{n-1} + \frac{1}{2}h_n\{f(t_n, y_n) + f(t_{n-1}, y_{n-1})\}$, Ordnung $m = 2$;

g) $\sum_{r=0}^{R} \alpha_{R-r}y_{n-r} = h\sum_{r=0}^{R} \beta R - rf_{n-r}$, $f_m := f(t_m, y_m)$, $\pi(z;\bar{h}) = \sum_{r=0}^{R}\{\alpha_r - \bar{h}\beta_r\}z^r$;

h) Wenn der Quotient von kleinstem und größtem (negativen) Realteil der Eigenwerte der Jacobi-Matrix $f_x'(t,u(t))$ entlang der Lösungstrajektorie sehr groß ist;

i) „Null-stabil“: Alle Nullstellen des 1. charakteristischen Polynoms $\rho(z) = \sum_{r=0}^{R} \alpha_r z^r$ erfüllen $|\lambda| \leq 1$, sowie $|\lambda| < 1$, wenn sie mehrfach sind.
„A-stabil“: das Stabilitätsgebiet enthält die „negative“ komplexe Halbebene.
„A(0)-stabil“: Das Stabilitätsgebiet enthält die „negative“ reelle Halbachse..

j) $SG_{\text{expl.Euler}} = \{z \in \mathbb{C} \,|\, |z+1| \leq 1\}$, $SG_{\text{impl.Euler}} = \{z \in \mathbb{C} \,|\, |z-1| \geq 1\}$, $SG_{\text{Trapez}} = \{z \in \mathbb{C} \,|\, \operatorname{Re} z \leq 0\}$;

k) Minimalzahl der zeitlichen Ableitungen die zur Überführung der DAE in eine „normal“ AWA erforderlich sind.

A.7 Kapitel 7

Lösung A.7.1: i) Beim dG(1)-Verfahren besitzt U auf dem Intervall $I_n = (t_{n-1}, t_n]$ die Darstellung

$$U(t) = U_{n-1}^{+} + \frac{t - t_{n-1}}{h_n} \left(U_n^{-} - U_{n-1}^{+}\right).$$

Einsetzen in die 2. Gleichung liefert mit $f(t,x) = Ax + b$ und $y_n = U_n^{-}$:

$$\begin{aligned}
& U_n^{-} - U_{n-1}^{+} = \frac{2}{h_n} \int_{t_{n-1}}^{t_n} AU(t) \cdot (t - t_{n-1}) + b \cdot (t - t_{n-1}) \,\mathrm{d}t \\
\iff \qquad & y_n - U_{n-1}^{+} = \frac{2}{h_n} \left(AU_{n-1}^{+} \cdot \frac{1}{2} h_n^2 + \frac{1}{3} \frac{1}{h_n} h_n^3 A (U_n^{-} - U_{n-1}^{+}) \right) + h_n b \\
\iff \qquad & \left(I - \tfrac{2}{3} h_n A\right) y_n = \left(I + \tfrac{1}{3} h_n A\right) U_{n-1}^{+} + h_n b.
\end{aligned}$$

Und für die 1. Gleichung:

$$\begin{aligned}
& U_n^{-} = U_{n-1}^{-} + \int_{t_{n-1}}^{t_n} AU(t) + b \,\mathrm{d}t \\
\iff \qquad & y_n = y_{n-1} + h_n A U_{n-1}^{+} + \tfrac{1}{2} h_n A (y_n - U_{n-1}^{+}) + h_n b \\
\iff \qquad & \left(I - \tfrac{1}{2} h_n A\right) y_n = y_{n-1} + \tfrac{1}{2} h_n A U_{n-1}^{+} + h_n b.
\end{aligned}$$

Einsetzen der 2. in die 1. Gleichung liefert schließlich:

$$\left(I - \tfrac{2}{3} h_n A + \tfrac{1}{6} h_n^2 A^2\right) y_n = \left(I + \tfrac{1}{3} h_n A\right) y_{n-1} + \left(I - \tfrac{1}{6} h_n A\right) h_n b.$$

ii) Der Abschneidefehler ist definiert durch

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

Man entwickelt in diesem Fall:

$$\begin{aligned}
u(t_n) &= u(t_{n-1}) + h_n u'(t_{n-1}) + \tfrac{1}{2} h_n^2 u''(t_{n-1}) + \mathcal{O}(h_n^3) \\
&= u(t_{n-1}) + h_n f\big|_{t_{n-1}} + \tfrac{1}{2} h_n^2 \left(f_t + f_x f\right)\big|_{t_{n-1}} + \mathcal{O}(h_n^3) \\
&= u(t_{n-1}) + h_n \big(A u(t_{n-1}) + b\big) + \tfrac{1}{2} h_n^2 A \big(A u(t_{n-1}) + b\big) + \mathcal{O}(h_n^3).
\end{aligned}$$

Weiterhin:

$$u(t_{n-1} + h_n F(h_n; t_{n-1}, u(t_{n-1})) \\ = \left(I - \tfrac{2}{3}h_n A + \tfrac{1}{6}h_n^2 A^2\right)^{-1} \left[\left(I + \tfrac{1}{3}h_n A\right)u(t_{n-1}) + \left(I - \tfrac{1}{6}h_n A\right)h_n b\right].$$

Mit

$$\left(I - \tfrac{2}{3}h_n A + \tfrac{1}{6}h_n^2 A^2\right)^{-1} = I + \tfrac{2}{3}h_n A + \tfrac{5}{18}h_n^2 A^2 + \mathcal{O}(h_n^3)$$

ergibt sich:

$$\begin{aligned} &u(t_{n-1} + h_n F(h_n; t_{n-1}, u(t_{n-1})) \\ &\quad = \left(I + \tfrac{2}{3}h_n A + \tfrac{5}{18}h_n^2 A^2 + \mathcal{O}(h_n^3)\right)\left[\left(I + \tfrac{1}{3}h_n A\right)u(t_{n-1}) + \left(I - \tfrac{1}{6}h_n A\right)h_n b\right] \\ &\quad = \left(I + h_n A + \tfrac{1}{2}h_n^2 A^2 + \mathcal{O}(h_n^3)\right)u(t_{n-1} + \left(I + \tfrac{1}{2}h_n A\right)h_n b + \mathcal{O}(h_n^3). \end{aligned}$$

Man sieht durch Subraktion:

$$h_n \tau_n = O(h_n^3),$$

d. h. das Verfahren ist von 2. Ordnung.

Bemerkung:

$$\left(I - \tfrac{2}{3}h_n A + \tfrac{1}{6}h_n^2 A^2\right)y_n = \left(I + \tfrac{1}{3}h_n A\right)y_{n-1}$$

ist ein sogenanntes „subdiagonales Padé-Schema 2. Ordnung“: Unter einer entsprechenden Begriffsbildung kann man zeigen, dass ein *Lösungsoperator* $e^{(t-t_0)A}$ existiert, der angewendet auf den Startwert u_0,

$$u(t) = e^{(t-t_0)A}u_0,$$

eine Lösung der AWA

$$u'(t) = Au(t), \quad t \geq t_u, \quad u(t_0) = u_0$$

liefert (Der Einfachheit halber sei $b = 0$ gesetzt). Das Padé-Schema entsteht nun durch eine stückweise rationale Approximation des Lösungsoperators:

$$e^{(t_n - t_0)A} = e^{h_n A} \cdot \ldots \cdot e^{h_n A} \approx \frac{P(h_n A)}{Q(h_n A)} \cdot \ldots \cdot \frac{P(h_n A)}{Q(h_n A)},$$

mit gewissen Polynomen P und Q.

iii) Angewendet auf das skalare Testproblem ergibt sich:

$$\left(1 - \tfrac{2}{3}h_n\lambda + \tfrac{1}{6}h_n^2\lambda^2\right)y_n = \left(1 + \tfrac{1}{3}h_n\lambda\right)y_{n-1}.$$

Der Verstärkungsfaktor lautet:

$$w(z) = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}.$$

Bestimmung des Stabilitätsintervalls: Es soll sein:

$$\left|1+\tfrac{1}{3}z\right| \le \left|1-\tfrac{2}{3}z+\tfrac{1}{6}z^2\right| = \left|\tfrac{1}{6}(z-2)^2+\tfrac{1}{3}\right|,$$

also

$$\iff \qquad \left|1+\tfrac{1}{3}z\right| \le 1-\tfrac{2}{3}z+\tfrac{1}{6}z^2.$$

Fallunterscheidung:

1. Fall: $z \ge -3$:

$$\begin{aligned} & 1+\tfrac{1}{3}z \le 1-\tfrac{2}{3}z+\tfrac{1}{6}z^2 \\ \iff \quad & 0 \le z\cdot(z-6) \\ \iff \quad & z \ge 6 \quad \text{oder} \quad z \le 0 \\ \iff \quad & z \in [-3,0]\cup[6,\infty)\,. \end{aligned}$$

2. Fall: $z \le -3$:

$$\begin{aligned} & -1-\tfrac{1}{3}z \le 1-\tfrac{2}{3}z+\tfrac{1}{6}z^2 \\ \iff \quad & 0 \le 11+(z-1)^2 \\ \iff \quad & z \in (-\infty,3) \end{aligned}$$

Damit ergibt sich $\mathrm{SI} = \mathbb{R}\setminus(0,6)$.

Lösung A.7.2: i) Es gibt eine Konstante $\kappa \in \mathbb{R}^+$, so dass die Ungleichung

$$\sup_{t\in I}|v(t)| \le \kappa\left(\int_I |v'(t)|\,dt + \Big|\int_I v(t)\,dt\Big|\right) \quad \forall v \in P_r(I)$$

auf dem Einheitsintervall $I=(0,1)]$ gilt.
Hierzu zeigen wir, dass die rechte Seite eine Norm auf dem endlichdimensionalen Polynomraum $P_r(I)$ darstellt. Die Existenz von κ folgt dann aus der Normäquivalenz. Homogenität und Dreiecksungleichung sind klar aufgrund selbiger Eigenschaft des Absolutbetrags und die Definitheit folgt aus:

$$\begin{aligned} & \int_I |v'(t)|\,dt + \Big|\int_I v(t)\,dt\Big| = 0 \\ \Longrightarrow \quad & v = \text{const.} \qquad \Big|\int_I v(t)\,dt\Big| = 0 \\ \Longrightarrow \quad & v \equiv 0. \end{aligned}$$

Sei nun $I_n = (t_{n-1}, t_n]$ ein beliebiges Intervall. Definiere

$$\chi:\ (0,1] \to (t_{n-1},t_n], \quad t = \chi(\hat{t}) = t_{n-1} + h_n\,\hat{t}.$$

Dies induziert einen „Pullback“: $P_r(I_n) \to P_r(I)$ durch

$$\hat{\varphi}(\hat{t}) = \varphi\big(\chi(\hat{t})\big) \quad \text{für } \varphi \in P_r(I_n).$$

Weiterhin ist

$$\hat{\varphi}'(\hat{t}) = \varphi'(t)\chi'(\hat{t}) = \varphi'(t)h_n.$$

Hiermit können die einzelnen Terme der Ungleichung entsprechend transformiert werden:

$$\begin{aligned}
\sup_{t\in I_n} |\varphi(t)| &= \sup_{\hat{t}\in I} |\varphi(\chi(\hat{t}))| = \sup_{\hat{t}\in I} |\hat{\varphi}(\hat{t})|, \\
\int_{I_n} \varphi(t)\,dt &= \int_I \varphi(\chi(\hat{t}) \cdot |\det J\chi^{-1}(\hat{t})|\,d\hat{t} = \int_I \hat{\varphi}(\hat{t})\,d\hat{t} \cdot \frac{1}{h_n} \\
\int_{I_n} |\varphi'(t)|\,dt &= \int_I |\varphi'(\chi(\hat{t})| \cdot |\det J\chi^{-1}(\hat{t})|\,d\hat{t} = \int_I |\hat{\varphi}(\hat{t})\,h_n|\,d\hat{t} \cdot \frac{1}{h_n}
\end{aligned}$$

Einsetzen der Identitäten in die Ungleichung auf I ergibt also zusammenfassend die Behauptung:

$$\sup_{t\in I_n} |v(t)| \le \kappa \left(\int_{I_n} |v'(t)|\,dt + \Big| \int_{I_n} v(t)\,dt \Big| \right) \quad v \in P_r(I_n).$$

ii) Diese Ungleichung gilt auch gleichmäßig für Funktionen $v \in C^1\big(\bar{I}_n\big)$. Wir argumentieren, dass sich das obige Skalierungsargument direkt überträgt. Man zeigt zunächst wieder die Gültigkeit der Ungleichung

$$\sup_{t\in I} |v(t)| \le \kappa \left(\int_I |v'(t)|\,dt + \Big| \int_I v(t)\,dt \Big| \right) \quad v \in C^1\big(\bar{I}\big).$$

Dies muss allerdings aufgrund der unendlichen Dimesion von $C^1\big(\bar{I}\big)$ über ein etwas anderes Argument erfolgen: Sei $v \in C^1\big(\bar{I}\big)$ beliebig. Dann gibt es nach dem Mittelwertsatz ein $\xi \in I$, so dass

$$v(\xi) = \int_I v(t)\,dt.$$

Es gilt dann:

$$\begin{aligned}
v(t) &= v(\xi) + \int_\xi^t v'(\tilde{t})\,d\tilde{t} \\
\Longrightarrow \quad |v(t)| &= |v(\xi)| + \int_I |v'(\tilde{t})|\,d\tilde{t} \\
\Longrightarrow \quad \sup_{t\in I} |v(t)| &= \Big| \int_I v(t)\,dt \Big| + \int_I |v'(t)|\,dt
\end{aligned}$$

Das Skalierungsargument überträgt sich ohne Modifikation: Definiere analog zu oben den „Pullback“: $C^1\big(\hat{I}_n\big) \to C^1\big(\hat{I}\big)$ durch

$$\hat{\varphi}(\hat{t}) = \varphi\big(\chi(\hat{t})\big) \quad \text{für } \varphi \in C^1\big(\hat{I}_n\big).$$

Dieser ist offensichtlich wohldefiniert und es gilt wieder

$$\hat{\varphi}'(\hat{t}) = \varphi'(t)\chi'(\hat{t}) = \varphi'(t)h_n,$$

so dass die restliche Argumentation aus (i) ohne weitere Modifikation anwendbar ist.

Lösung A.7.3: Beim $\mathrm{dG}_{\exp}(\mathrm{r})$-Verfahren werden nach *rechts* halboffene Teilintervalle $I_n = [t_{n-1}, t_n)$ verwendet anstelle der nach links halboffenen Intervalle des dG(r)-Verfahrens. Dies führt zu einem modifizierten Ansatz der Form:

$$\sum_{n=1}^{N} \left\{ \int_{I_n} (U' - f(t,U), \varphi)\, dt + \big([U]_n, \varphi_n^-\big) \right\} = 0, \qquad (*)$$

bzw. bei Wahl einer Testfunktion φ mit $\varphi \equiv 0$ auf $I_m \neq I_n$:

a) Die Wahl stückweiser konstanter Ansatz- und Testfunktionen im Fall $r = 0$,

$$y_n := U_{n+1}(t) = U_n^+ = U_{n+1}^-,$$
$$\varphi \equiv 1 \text{ auf } I_n,$$

mit dem Startwert $y_0 := u_0$ reduziert diese Gleichung auf:

$$y_n = \int_{I_n} f(t, y_{n-1})\, dt + y_{n-1}.$$

Durch Approximation des Integrals mit der Boxregel,

$$\int_{I_n} f(t, y_{n-1})\, dt \approx h_n f(t_{n-1}, y_{n-1})$$

erhält man die explizite Polygonzugmethode:

$$y_{n+1} = y_n + h_n f(t_n, y_n).$$

b) Es gilt

$$U_n^+ = U_{n-1}^+ + \int_{I_n} f(t, U(t))\, dt,$$

$$u(t_n) = u(t_{n-1}) + \int_{I_n} f(t, u(t))\, dt.$$

Subtraktion liefert:

$$U_n^+ - u(t_n) = U_{n-1}^+ - u(t_{n-1}) + \int_{I_n} \big(f(t, U_{n-1}^+) - f(t, u(t))\big)\, dt$$

Unter Ausnutzung der L-Stetigkeit von f:

$$\left|U_n^+ - u(t_n)\right| \le \left|U_{n-1}^+ - u(t_{n-1})\right| + \int_{I_n} L_f \left|U_{n-1}^+ - u(t)\right| dt\,. \qquad (**)$$

Der Integrand des Integrals auf der rechten Seite erlaubt die Darstellung

$$\begin{aligned} U_{n-1}^+ - u(t) &= U_{n-1}^+ - u(t_{n-1}) - \int_{t_{n-1}}^{t} u'(\tilde{t})\, d\tilde{t} \\ \Longrightarrow \quad \left|U_{n-1}^+ - u(t)\right| &\le \left|U_{n-1}^+ - u(t_{n-1})\right| + \int_{t_{n-1}}^{t_n} \left|u'(\tilde{t})\right| d\tilde{t} \\ &\le \left|U_{n-1}^+ - u(t_{n-1})\right| + h_n \sup_{I_n} |u'|. \end{aligned}$$

Einsetzen in (**):

$$\begin{aligned} \left|U_n^+ - u(t_n)\right| &\le \left|U_{n-1}^+ - u(t_{n-1})\right| + \int_{I_n} L_f \left(\left|U_{n-1}^+ - u(t_{n-1})\right| + h_n \sup_{I_n} |u'| \right) dt \\ &\le \left|U_{n-1}^+ - u(t_{n-1})\right| + h_n L_f \left|U_{n-1}^+ - u(t_{n-1})\right| + h_n^2 L_f \sup_{I_n} |u'|. \end{aligned}$$

Rekursives Einsetzen liefert (man beachte $|U_0^+ - u(t_0)| = 0$):

$$\left|U_n^+ - u(t_n)\right| \le \sum_{\nu=0}^{n-1} h_{\nu+1} L_f \left|U_\nu^+ - u(t_\nu)\right| + \sum_{\nu=1}^{n} h_\nu^2 L_f \sup_{I_\nu} |u'|.$$

Nach dem Gronwallschen Lemman gilt folglich:

$$\left|U_n^+ - u(t_n)\right| \le \exp\left(L_f(t_n - t_0)\right) \sum_{\nu=1}^{n} h_\nu^2 L_f \sup_{I_\nu} |u'|,$$

und damit

$$\left|U_n^+ - u(t_n)\right| \le \exp\left(L_f(t_n - t_0)\right) L_f(t_n - t_0) \max_{1\le\nu\le n} \left\{ h_\nu \sup_{I_\nu} |u'| \right\},$$

Sei nun $t \in I_{n+1}$:

$$\begin{aligned} \left|U_n^+ - u(t)\right| &\le \left|u(t) - u(t_n)\right| + \left|U_n^+ - u(t_n)\right| \\ &\le \left| \int_{t_n}^{t} u'(\tilde{t})\, d\tilde{t} \right| + \left|U_n^+ - u(t_n)\right| \\ &\le h_{n+1} \sup_{I_{n+1}} |u'| + \left|U_n^+ - u(t_n)\right| \\ &\le \left\{ 1 + L_f(t_n - t_0) \exp\left(L_f(t_n - t_0)\right) \right\} \max_{1\le\nu\le n+1} \left\{ h_\nu \sup_{I_\nu} |u'| \right\}, \end{aligned}$$

Also letztendlich:

$$\sup_{t\in I} \left|U(t) - u(t)\right| \le \left\{ 1 + L_f T e^{L_f T} \right\} \max_{1\le\nu\le N} \left\{ h_\nu \sup_{I_\nu} |u'| \right\},$$

c) Im Fall stückweiser linearer Ansatz- und Testfunktionen ($r = 1$) hat $U(t)$ die Gestalt:

$$U(t) = h_n^{-1}(t - t_{n-1})U_n^- - h_n^{-1}(t - t_n)U_{n-1}^+$$

Testen von (*) mit $\varphi \equiv 1$ und $\varphi \equiv h_n^{-1}(t_n - t)$ liefert:

$$U_n^+ - U_{n-1}^+ = \int_{I_n} f(t, U)\, dt,$$

$$U_n^- - U_{n-1}^+ = \frac{2}{h_n} \int_{I_n} f(t, U)(t_n - t)\, dt.$$

Approximiere die Integrale mit der Trapezregel:

$$\int_{I_n} f(t, U)\, dt \approx \tfrac{1}{2} h_n \left(f(t_{n-1}, U_{n-1}^+) + f(t_n, U_n^-) \right),$$

$$\frac{2}{h_n} \int_{I_n} f(t, U)(t_n - t)\, dt \approx h_n f(t_{n-1}, U_{n-1}^+).$$

Einsetzen liefert:

$$U_n^+ = U_{n-1}^+ + \frac{1}{2} h_n \left\{ k_1 + k_2 \right\}, \quad k_1 = f(t_{n-1}, U_{n-1}^+), \quad k_2 = f(t_n, U_{n-1}^+ + h_n k_1).$$

Dies ist gerade das *Heunsche Verfahren 2. Ordnung.*

Lösung A.7.4: i) Die Spezialfälle $r = 0, 1$ folgen direkt aus dem allgemeinen Beweis in (ii). Alternativ kann man die Behauptungen auch durch direktes Nachrechnen verifizieren. So z. B. erlaubt $r = 0$ die folgende Diskussion:

Nach dem Mittelwertsatz gibt es ein $\xi \in I$ mit $u(\xi) = \frac{1}{I_n} \int_{I_n} u(t)\, dt$. Weiterhin gilt

$$u(t) = \int_\xi^t u'(s)\, ds + u(\xi).$$

Damit ergibt sich:

$$|u(t) - u(\xi)| \le \int_{t_{n-1}}^{t_n} |u'(s)|\, ds \le h_n \sup_{t \in I_n} |u'(t)|.$$

Durch Testen der Bestimmungsgleichung von $\pi_r u$ mit $\varphi \equiv 1$ sieht man $\pi_r u = u(\xi)$, so dass sich unmittelbar ergibt:

$$\sup_{t \in I_n} |u(t) - \pi_r u| \le h_n \sup_{t \in I_n} |u'(t)|.$$

ii) Sei $f \in C(\bar{I}_n)$ mit der Eigenschaft

$$\int_{I_n} f(t) \cdot \varphi(t)\, dt = 0 \quad \forall \varphi \in P_r(I_n),$$

Angenommen f hat höchstens r Nullstellen. Sei nun $\{\tau_i,\ i = 1, \cdots, m\}$ die (evtl. leere) Menge aller Nullstellen von f an der die Funktion f das Vorzeichen wechselt. (Aufgrund der höchstens r Nullstellen von f ist stets $f(t) \neq 0$ für alle $t \neq t_i$ in einer Umgebung einer Nullstelle t_i, und der Vorzeichenwechsel damit wohldefiniert.) Definiere

$$\psi(t) := \prod_{i=1}^{m} (t - \tau_i) \quad \in P_r(I_n)$$

($\psi(t) :\equiv 1$, falls keine τ_i existieren). Die Funktion $f(t)\psi(t)$ besitzt damit nur noch Nullstellen an der sie das Vorzeichen nicht wechselt und es gilt:

$$f(t)\psi(t) \geq 0, \text{ oder } f(t)\psi(t) \leq 0$$

gänzlich auf I_n. Weiterhin ist $\psi \not\equiv 0$ und $f \not\equiv 0$ (f hat höchstens r reelle Nullstellen). Somit gilt:

$$\int_{I_n} f(t)\psi(t)\mathrm{d}t > 0, \text{ oder } \int_{I_n} f(t)\psi(t)\mathrm{d}t < 0.$$

Damit muss für den Polynomgrad m von ψ aufgrund der Orthogonalitätseigenschaft aber $m \geq r + 1$ gelten: *Widerspruch!*

Seien nun $u \in C^r(\bar{I}_n)$ und $\pi_r u$ die L^2-Interpolierende von u. Dann hat $u - \pi_r u$ nach obiger Darstellung mindestens $r + 1$ Nullstellen. Seien $\tau_0, \cdots, \tau_r$ $r + 1$ willkürlich ausgewählte, paarweise verschiedene Nullstellen von $u - \pi_r u$. Fasse nun das Nullpolynom $p \equiv 0$ als Lagrangeinterpolierende in diesen Punkten von $u - \pi_r u$ auf. Nach der *Fehlerdarstellung der Lagrange-Interpolation*,

$$\exists \xi_t : \quad (u - \pi_r u)(t) - p(t) = \frac{(u - \pi_r u)^{(r+1)}(\xi_t)}{(r+1)!} \prod_{j=0}^{r} (t - \tau_j),$$

folgt damit sofort die Abschätzung:

$$\sup_{t \in I_n} |(u - \pi_r u)(t)| \leq \frac{1}{(r+1)!} h_n^{(r+1)} \sup_{t \in I_n} |u^{(r+1)}|.$$

Lösung A.7.5 (Praktische Aufgabe): Nicht verfügbar.

Lösung A.7.6: Die a priori-Fehlerabschätzung für das dG(0)-Verfahren lautet

$$\sup_I \|e\| \leq K_{\mathrm{dG}} \max_{1 \leq n \leq N} \Big\{ h_n \sup_{I_n} \|u'\| \Big\},$$

die analoge Fehlerabschätzung für das implizite Eulerverfahren:

$$\sup_{t_n \in I} \|e_n\| \leq K_{\mathrm{iE}} \max_{1 \leq n \leq N} \Big\{ h_n \sup_{I_n} \|u''\| \Big\},$$

mit Konstanten K_{dG}, K_{iE}, die exponentiell von $L_f(t)$ abhängen.

Die L-Konstante von $f(x) = x^2$ verhält sich aufgrund

$$\|x^2 - y^2\| \leq 2 \max\{\|x\|, \|y\|\} \|x - y\|$$

entlang des Lösungsverlaufs $u(t)$ wie

$$L(t) = 2 \sup_{s\in[t_0,t]} |u(s)| = \frac{2}{1-t}$$

Dies ergibt für die Konstanten K_{dG}, bzw. K_{iE}:

$$K_{\mathrm{dG}} = C_{\mathrm{dG}} \exp\left(\int_{t_0}^{t_N} \frac{2}{1-s}\, dts\right) = C_{\mathrm{dG}} \exp\left(-2\log(1-t_N)\right) = C_{\mathrm{dG}} \frac{1}{(1-t_N)^2}.$$

i) Es ist

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

und damit

$$\sup_{t\in I_n} |u'(t)| = \frac{1}{(1-t_n)^2}, \quad \sup_{t\in I_n} |u''(t)| = \frac{2}{(1-t_n)^3}.$$

Es ergibt sich also:

$$\text{dG(0):} \qquad \sup_I \|e\| \leq C_{\mathrm{dG}} \frac{1}{(1-t_N)^2} \max_{1\leq n\leq N} \left\{\frac{h_n}{(1-t_n)^2}\right\},$$

$$\text{Impl. Euler:} \qquad \sup_{t_n\in I} \|e_n\| \leq C_{\mathrm{iE}} \frac{1}{(1-t_N)^2} \max_{1\leq n\leq N} \left\{\frac{2\,h_n}{(1-t_n)^3}\right\}.$$

ii) Wir betrachten die Schrittweitenbedingung

$$\sup_I \|e\| \leq K_{\mathrm{dG}} \max_{1\leq n\leq N} \left\{\frac{h_n}{(1-t_n)^2}\right\} \overset{!}{\leq} \mathrm{TOL}.$$

Hinreichend hierfür ist die stärkere Bedingung für $1 \leq n \leq N$:

$$K_{\mathrm{dG}} \frac{h_n}{(1-t_n)^2} \overset{!}{=} \mathrm{TOL}.$$

Approximiert man diese Beziehung nun kontinuierlich, so erhält man:

$$\tilde{h}(t) \overset{!}{=} \frac{\mathrm{TOL}}{K_{\mathrm{dG}}} (1-t)^2.$$

Die kontinuierliche Approximation der Schrittanzahl N verläuft nach folgender Motivation:

$$N = \sum_{1\leq n\leq N} 1 = \sum_{1\leq n\leq N} h_n^{-1} h_n = \sum_{1\leq n\leq N} h(t_n)^{-1} h(t_n) \approx \int_0^{t_N} \tilde{h}(t)^{-1}\, dt.$$

Dies ergibt

$$\tilde{N}_{\mathrm{dG}} = \frac{C_{\mathrm{dG}}}{\mathrm{TOL}}(1-t_N)^{-3},$$

und analog

$$\tilde{N}_{\mathrm{iE}} = \frac{C_{\mathrm{iE}}}{\mathrm{TOL}}(1-t_N)^{-4}.$$

Während der Aufwand des dG(0)-Verfahrens also kubisch in $(1-t_N)^{-1}$ ist, ergibt sich für die Abschätzung des impliziten Eulerverfahrens eine Abhängigkeit in vierter Potenz.

Lösung A.7.7: Nicht verfügbar.

Lösung A.7.8: Beim cG(1)-Verfahren macht man den Ansatz

$$U \in C(I) \cap S_h(1) = \left\{ v \in C(I) : v\big|_{I_n} \in P_1(I_n) \right\}$$

mit Testfunktionen

$$\varphi \in S_h^{(0)}(I) := \left\{ v : I \to \mathbb{R} \ : \ v\big|_{I_n} \in P_0(I_n) \right\}.$$

Wir suchen also $U \in C(I) \cap S_h(1)$ mit

$$(u_0, \varphi_0^+) = A(U,\varphi) := \sum_{n=1}^{N} \int_{I_n} (U' - AU(t) - b, \varphi)\ dt + (U_0, \varphi_0^+) \quad \forall \varphi \in S_h^{(0)}(I).$$

Bemerkung: Hieraus folgt intervallweise:

$$U_n - U_{n-1} = \int_{I_n} AU(t) + b\ dt,$$

bzw. mit $U = h_n^{-1}(t-t_{n-1})U_n + h_n^{-1}(t_n - t)U_{n-1}$:

$$U_n - U_{n-1} = \tfrac{1}{2} h_n (AU_n + AU_{n-1} + 2b).$$

Das primale Problem lautet in obiger Notation: Gesucht ist $u \in C^1(I)$ mit

$$A(u,\varphi) = (u_0, \varphi) \qquad \forall \varphi \in \left\{ v : I \to \mathbb{R} \ : \ v\big|_{I_n} \in C_c(I_n) \right\}.$$

Man überlegt sich analog zum Text, dass das duale Problem die Gestalt hat: Gesucht ist $z \in C^1(I)$ mit

$$L^*(z,\varphi) = (e_N/\|e_N\|, \varphi_N^-) \quad \forall \varphi \in \left\{ v : I \to \mathbb{R} \ : \ v\big|_{I_n} \in C_c(I_n) \right\},$$

mit der adjungierten Linearisierung:

$$L^*(z,\varphi) = \sum_{n=1}^{N} \int_{I_n} (-z' - A^* z, \varphi)\, dt + (z_N, \varphi_N^-)$$

und dem Fehler $e(t) = u(t) - U(t)$.
Man definiert weiterhin das Residuum $\rho(U, V)$ als

$$\rho(U, V) = (u_0, V_0^+) - A(U, V).$$

Bemerkung: $\rho(U, V) = 0$ für alle $V \in S_h^{(0)}(I)$ nach Definition der diskreten Lösung U.
Der Schlüssel zur a posteriori-Fehlerschätzung ist die Feststellung, dass mit $\varphi = e$ gilt:

$$L^*(z, e) = (e_N/\|e_N\|, e_N^-) = \|e_N\|,$$

sowie

$$\begin{aligned} L^*(z, e) &= \sum_{n=1}^{N} \int_{I_n} (-z' - A^* z, e)\, dt + (z_N, e_N^-) \\ &= \sum_{n=1}^{N} \int_{I_n} (z, e' - Ae - b + b)\, dt - (z_0, e_0^+) \\ &= \sum_{n=1}^{N} \int_{I_n} (z, u' - Au - b)\, dt - (z_0, u_0^+) - \sum_{n=1}^{N} \int_{I_n} (z, U' - AU - b)\, dt + (z_0, U_0^+) \\ &= 0 + \rho(U, z). \end{aligned}$$

Also

$$\rho(U, z) = \|e_N\|.$$

Analog zum Vorgehen im Text setzt man wieder $Z := \tilde{P}_0 z \in S_h^{(0)}$ mit:

$$(\tilde{P}_0 z)_0^+ = z(t_0).$$

Damit ist dann $\rho(U, z) = \rho(U, z - Z)$. Ausgeschrieben:

$$\begin{aligned} \|e_N\| &= \rho(U, z) \\ &= \rho(U, z - Z) \\ &= -\sum_{n=1}^{N} \int_{I_n} (U' - AU - b, z - Z)\, dt. \end{aligned}$$

Und letztendlich:

$$\begin{aligned} & \|e_N\| = \rho(U, z) = -\sum_{n=1}^{N} \int_{I_n} (U' - AU - b, z - \tilde{P}_0 z)\, dt \\ \Longrightarrow \quad & \|e_N\| \le \sum_{n=1}^{N} \sup_{I_n} \|U' - AU - b\| \int_{I_n} \|z - \tilde{P}_0 z)\|\, dt. \end{aligned}$$

Lösung A.7.9: a) In der vorliegenden Situation bestimmen wir

$$K(T) \approx (1-T)^{-2}, \qquad |\tau_n^0(U)| \approx \sup_{I_n} |u''| \approx (1-t_n)^{-3},$$

und folglich $h_n \approx (1-T)(1-t_n)^{3/2} N^{-1/2}\,\mathrm{TOL}^{1/2}$. Dies liefert

$$\begin{aligned} N \;=\; \sum_{n=1}^{N} h_n h_n^{-1} \;&\approx\; N^{1/2}(1-T)^{-1}\mathrm{TOL}^{-1/2} \sum_{n=1}^{N} h_n (1-t_n)^{-3/2} \\ &\approx\; N^{1/2}(1-T)^{-3/2}\mathrm{TOL}^{-1/2}, \end{aligned}$$

und schließlich $N \approx (1-T)^{-3}\mathrm{TOL}^{-1}$.

b) Wir gehen aus von der a posteriori Fehlerabschätzung

$$|e_N^-| \le \sum_{n=1}^{N} h_n^2 \rho_n(U)\omega_n(z), \quad \rho_n(U) := h_n^{-1}|[U]_n|, \quad \omega_n(z) := h_n^{-1}\int_{I_n} |z'|\,dt,$$

mit der zugehörigen dualen Lösung z, und die daraus resultierende (implizite) Schrittweitenformel

$$h_n := \left(\frac{\mathrm{TOL}}{N\rho_n(U)\omega_n(z)}\right)^{1/2}.$$

Weiter gilt (s. Text)

$$\rho_n(U) := h_n^{-1}|[U]_n| \approx h_n^{-1}|u_n - u_{n-1}| \le h_n^{-1}\int_{I_n} |u'|\,dt \le (1-t_n)^{-2},$$

sowie bei Beachtung von $z(t) = (1-t)^2(1-t_n)^{-2}$:

$$\omega_n(z) := h_n^{-1}\int_{I_n} |z'|\,dt = 2h_n^{-1}(1-t_n)^{-2}\int_{I_n} (1-t)\,dt \le 2(1-t_n)^{-1}.$$

Damit ergibt sich

$$h_n \approx \frac{\mathrm{TOL}^{1/2}}{N^{1/2}}(1-t_n)^{3/2}.$$

Für N folgt dann die Beziehung

$$\begin{aligned} N = \sum_{n=1}^{N} h_n h_n^{-1} &\approx \sum_{n=1}^{N} h_n \frac{N^{1/2}}{\mathrm{TOL}^{1/2}}(1-t_n)^{-3/2} \\ &\approx \frac{N^{1/2}}{\mathrm{TOL}^{1/2}}\int_0^T (1-t)^{-3/2}\,dt \approx \frac{N^{1/2}}{\mathrm{TOL}^{1/2}}(1-T)^{-1/2} \end{aligned}$$

bzw.

$$N \approx \frac{1}{\mathrm{TOL}}\frac{1}{1-T}.$$

Lösung A.7.10 (Praktische Aufgabe): s. [10], Chapter 2.

A.8 Kapitel 8

Lösung A.8.1: a) Zunächst erhalten wir durch Einsetzen in die allgemeine Lösung $u(t) = A\sin(t) + B\cos(t) + 1$

$$u(0) = B + 1, \quad u(\tfrac{\pi}{2}) = A + 1, \quad u(\pi) = 1 - B.$$

a) $u(0) = u(\frac{\pi}{2}) = 0$: Dies liefert die Bedingungen

$$B + 1 = 0 \quad \Leftrightarrow \quad B = -1 \qquad \text{sowie} \qquad A + 1 = 0 \quad \Leftrightarrow \quad A = -1.$$

Es existiert also genau eine Lösung zu diesen Randwerten.

b) $u(0) = u(\pi) = 0$: Dies liefert die Bedingungen

$$B + 1 = 0 \quad \Leftrightarrow \quad B = -1 \qquad \text{sowie} \qquad 1 - B = 0 \quad \Leftrightarrow \quad B = 1.$$

Es existiert also keine Lösung zu diesen Randwerten.

c) $u(0) = u(\pi) = 1$: Dies liefert die Bedingungen

$$B + 1 = 1 \quad \Leftrightarrow \quad B = 0 \qquad \text{sowie} \qquad 1 - B = 1 \quad \Leftrightarrow \quad B = 0.$$

Es existert also für jedes A eine Lösung, also unendlich viele.

b) Mit den Funktionen u und $v := u'$ lautet das umgeformte System erster Ordnung

$$u'(t) - v(t) = 0, \quad v'(t) + u(t) = 1,$$

bzw. in Matrixschreibweise

$$\begin{bmatrix} u'(t) \\ v'(t) \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} u(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Die Fundamentalmatrix dieses Systems ist

$$Y(t) = \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix}$$

Die drei gegebenen Randbedingungen lauten

$$a) \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u(\frac{\pi}{2}) \\ v(\frac{\pi}{2}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$b) \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u(\pi) \\ v(\pi) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$c) \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u(\pi) \\ v(\pi) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Damit ergibt sich in allen drei Fällen

$$B_a + B_b Y(b) = \begin{bmatrix} 1 & 0 \\ \cos(b) & \sin(b) \end{bmatrix}, \qquad \det(B_a + B_b Y(b)) = \sin(b).$$

Somit ist die Matrix $B_a + B_b Y(b)$ im Falle a) regulär und in den Fällen b) und c) nicht regulär.

Lösung A.8.2: Analog zum Vorgehen im Text bei der Betrachtung des Sturm-Liouville-Problems mit Dirichlet-Randbedingungen lässt sich auch das Neumann-Problem in der Standardform schreiben als

$$\begin{bmatrix} u_0'(t) \\ u_1'(t) \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ r(t) & q(t) \end{bmatrix} \begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix} = \begin{bmatrix} 0 \\ f(t) \end{bmatrix},$$
$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_0(a) \\ u_1(a) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_0(b) \\ u_1(b) \end{bmatrix} = \begin{bmatrix} g_a \\ g_b \end{bmatrix}.$$

Es reicht deshalb wieder zu zeigen, dass das homogene Problem

$$-u''(t) + q(t)u'(t) + r(t)u(t) = 0, \quad t \in I = [a,b], \quad u'(a) = u'(b) = 0$$

nur die triviale Lösung besitzt. Dazu multiplizieren wir die Gleichung mit u und integrieren über I:

$$\int_I -u''(t)u(t)\,dt + \tfrac{1}{2}\int_I q(t)\left(u(t)^2\right)'\,dt + \int_I r(t)u(t)^2\,dt = 0.$$

Durch partielle Integration erhalten wir daraus unter Beachtung der Randbedingungen

$$\int_I \left(u'(t)\right)^2\,dt + \int_I \left(r(t) - \tfrac{1}{2}q'(t)\right)u(t)^2\,dt + \tfrac{1}{2}q(t)u(t)^2\Big|_a^b = 0.$$

Unter den Voraussetzungen

$$\min_{t\in I}\left(r(t) - \tfrac{1}{2}q'(t)\right) \geq 0, \quad q(a) \leq 0, \quad q(b) \geq 0,$$

wobei eine der Ungleichungen strikt gelten muss, folgt damit $u \equiv 0$.

Lösung A.8.3: (a) Nach dem Hauptsatz der Differential- und Integralrechnung gilt für $t \in [a,b]$

$$v(t) = \int_a^t v'(s)\,ds + v(a).$$

Damit folgt

$$|v(t)| \leq \int_a^t |v'(s)|\,ds + |v(a)| \leq \int_a^b |v'(s)|\,ds + |v(a)|.$$

Da die rechte Seite der Ungleichung unabhängig von t ist, ergibt sich durch Maximumsbildung über t direkt die Behauptung.

(b) Da v insbesondere stetig ist, folgt aus dem Mittelwertsatz der Integralrechnung die Existenz einer Stelle $t_0 \in [a,b]$, so dass

$$(b-a)v\,(t_0) = \int_a^b v\,(s)\;ds.$$

Nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$v\,(t) = \int_{t_0}^t v'\,(s)\;ds + v\,(t_0)\,.$$

Insgesamt ergibt sich also

$$|v\,(t)\,| \le \int_a^b |v'\,(s)\,|\,ds + |v\,(t_0)\,| = \int_a^b |v'\,(s)\,|\,ds + \frac{1}{b-a}\left|\int_a^b v\,(s)\;ds\right|.$$

A.9 Kapitel 9

Lösung A.9.1: i) Wir betrachten zunächst den Spezialfall $A = I$. Für alle $x \in \mathbb{K}^n$ gilt

$$\|(I+B)x\| \ge \|x\| - \|Bx\| \ge (1 - \|B\|)\|x\|.$$

Wegen $1 - \|B\| > 0$ ist also $I+B$ injektiv und folglich regulär. Mit der Abschätzung

$$\begin{aligned} 1 &= \|I\| = \|(I+B)(I+B)^{-1}\| = \|(I+B)^{-1} + B(I+B)^{-1}\| \\ &\ge \|(I+B)^{-1}\| - \|B\|\,\|(I+B)^{-1}\| = \|(I+B)^{-1}\|\,(1-\|B\|) > 0 \end{aligned}$$

erhält man die behauptete Ungleichung für den Fall $A = I$.

ii) Nun zum Fall einer allgemeinen Matrix A: Wegen $\|B\| \le \|A^{-1}\|^{-1}$ ist $\|A^{-1}B\| \le \|A^{-1}\|\|B\| < 1$. Nach dem o. a. Resultat ist also $I + A^{-1}B$ regulär, und es gilt

$$\|(I+A^{-1}B)^{-1}\| \le \frac{1}{1-\|A^{-1}B\|} \le \frac{1}{1-\|A^{-1}\|\|B\|}.$$

Dann ist auch $A + B = A(I + A^{-1}B)$ regulär, und es gilt

$$\|(A+B)^{-1}\| \le \|A^{-1}(I+A^{-1}B)^{-1}\|\|\| \le \|\|A^{-1}\|\|(I+A^{-1}B)^{-1}\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|B\|}.$$

Lösung A.9.2: Wir betrachten die zweidimensionale lineare RWA

$$u'(t) - Au(t) = f(t), \quad t \in [a,b], \quad u_1(a) = \alpha, \quad u_2(b) = \beta.$$

Diese lautet in Standardform

$$\begin{bmatrix} u_1'(t) \\ u_2'(t) \end{bmatrix} - \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} = \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix},$$
$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(a) \\ u_2(a) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1(b) \\ u_2(b) \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Beim Gegenschießen machen wir für die diskrete Lösung $y(t)$ den folgenden Ansatz:

$$y(t) = y_i(t) + Y_i(t)s_i, \quad t \in I_i = [t_{i-1}, t_i],$$

wobei $y_i(t)$ bzw. $Y_i(t)$ als Lösungen der folgenden AWA gegeben sind:

$$\begin{aligned} y_1'(t) - Ay_1(t) &= f(t), & t \in [t_0, t_1], \quad y_1(t_0) &= 0, \\ Y_1'(t) - AY_1(t) &= 0, & t \in [t_0, t_1], \quad Y_1(t_0) &= I, \\ y_2'(t) - Ay_2(t) &= f(t), & t \in [t_1, t_2], \quad y_2(t_{\mathbf{2}}) &= 0, \\ Y_2'(t) - AY_2(t) &= 0, & t \in [t_1, t_2], \quad Y_2(t_{\mathbf{2}}) &= I. \end{aligned}$$

Ziel ist es nun, die Parametervektoren s_1 und s_2 so bestimmen, dass $y(t)$ die Randbedingungen erfüllt und auf dem gesamten Intervall $[a, b]$ stetig ist, d. h. dass gilt

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} (y_1(a) + Y_1(a)s_1) + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} (y_2(b) + Y_2(b)s_2) = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

sowie

$$y_1(t_1) + Y_1(t_1)s_1 = y_2(t_1) + Y_2(t_1)s_2.$$

Da wir in diesem Fall separierte Randbedingungen haben, können wir s_{11} und s_{22} direkt aus der Gleichung für die Randbedingungen ablesen. Unter Beachtung von $y_1(a) = y_2(b) = 0$ und $Y_1(a) = Y_1(b) = I$ erhalten wir nämlich

$$s_{11} = \alpha \quad \text{und} \quad s_{22} = \beta.$$

Somit sind nur noch s_{12} und s_{21} mithilfe der Schließbedingungen zu bestimmen. Aus

$$Y_1(t_1)s_1 - Y_2(t_1)s_2 = y_2(t_1) - y_1(t_1)$$

erhalten wir mit $s_{11} = \alpha$ und $s_{22} = \beta$

$$\begin{bmatrix} Y_1(t_1) & -Y_2(t_1) \end{bmatrix} \begin{bmatrix} \alpha \\ s_{12} \\ s_{21} \\ \beta \end{bmatrix} = \begin{bmatrix} y_2(t_1) - y_1(t_1) \end{bmatrix}.$$

Diese Bedingung kann man in ein 2×2 Gleichungssystem für die gesuchten Werte s_{12} und s_{21} umschreiben. Ist dieses eindeutig lösbar, so erhalten wir

$$s_1 = \begin{bmatrix} \alpha \\ \tilde{\alpha} \end{bmatrix}, \quad s_2 = \begin{bmatrix} \tilde{\beta} \\ \beta \end{bmatrix}.$$

Bemerkung: Bei Verwendung der üblichen Mehrzielmethode (d. h. kein Gegenschießen) hätten wir ein 3×3-Gleichungssystem erhalten, da wir zusätzlich zu s_{12} und s_{21} noch s_{22} mithilfe der Randbedingungen bestimmen müssten.

Lösung A.9.3: Wir formen das Anfangswertproblem um in ein System 1. Ordnung:

$$\begin{bmatrix} u_0'(t) \\ u_1'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 100 & 0 \end{bmatrix} \begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix}, \quad \begin{bmatrix} u_0(0) \\ u_1(0) \end{bmatrix} = \begin{bmatrix} 1 \\ s \end{bmatrix}.$$

Wir bestimmen die Eigenwerte von A:

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda & -1 \\ -100 & \lambda \end{bmatrix} = \lambda^2 - 100 = (\lambda + 10)(\lambda - 10).$$

A hat also die Eigenwerte 10 und -10. Die allgemeine Lösung des Anfangswertproblems lautet daher

$$\begin{bmatrix} u_0(t;s) \\ u_1(t;s) \end{bmatrix} = c_1 e^{10t} \begin{bmatrix} 1 \\ 10 \end{bmatrix} + c_2 e^{-10t} \begin{bmatrix} 1 \\ -10 \end{bmatrix}.$$

Damit erhält man für die Koeffizienten c_1 und c_2 aufgrund der Anfangswerte die Bedingungen

$$\begin{bmatrix} 1 \\ s \end{bmatrix} = \begin{bmatrix} u_0(0;s) \\ u_1(0;s) \end{bmatrix} = \begin{bmatrix} c_1 + c_2 \\ 10c_1 - 10c_2 \end{bmatrix}$$

und somit

$$c_1 = \frac{10+s}{20}, \quad c_2 = \frac{10-s}{20}.$$

Die Lösung der Anfangswertaufgabe lautet also

$$u(t;s) = \frac{10+s}{20} e^{10t} + \frac{10-s}{20} e^{-10t}.$$

Damit wird $e^{-30} = u(3;s^*) = \frac{10+s^*}{20} e^{30} + \frac{10-s^*}{20} e^{-30}$ und durch Auflösen nach s^* erhalten wir $s^* = -10$ und damit als Lösung der Randwertaufgabe

$$u(t) = u(t;s^*) = e^{-10t}.$$

Hat man statt $s^* = -10$ ein $\tilde{s}$ mit $|\frac{\tilde{s} - s^*}{s^*}| \leq \epsilon$, d. h. $|\tilde{s} + 10| \leq 10\epsilon$, so folgt

$$\begin{aligned} |u(3;\tilde{s}) - u(3;s^*)| &= \left| \frac{10+\tilde{s}}{20} e^{30} + \frac{10-\tilde{s}}{20} e^{-30} - \frac{10+s^*}{20} e^{30} - \frac{10-s^*}{20} e^{-30} \right| \\ &= \left| \frac{\tilde{s} - s^*}{20} \right| \left(e^{30} - e^{-30} \right) = \left| \frac{\tilde{s} + 10}{20} \right| \left(e^{30} - e^{-30} \right) \leq \frac{e^{30} - e^{-30}}{2} \epsilon. \end{aligned}$$

Mit $u(3;s^*) = e^{-30}$ ergibt sich für den relativen Fehler in $u(3;\tilde{s})$:

$$\left|\frac{u(3;\tilde{s}) - u(3;s^*)}{u(3;s^*)}\right| \leq \frac{e^{60}-1}{2}\epsilon.$$

Lösung A.9.4 (Praktische Aufgabe): Nicht verfügbar.

A.10 Kapitel 10

Lösung A.10.1: Wir betrachten die RWA

$$\begin{aligned} u'(t) - A(t)u(t) &= f(t), \quad t \in [a,b], \\ B_a u(a) + B_b u(b) &= g. \end{aligned}$$

1. Für die Polygonzugmethode gilt

   $$L_h y_n := h^{-1}(y_n - y_{n-1}) - A(t_{n-1})y_{n-1} = f(t_{n-1})$$

   bzw.

   $$(-I - hA(t_{n-1}))y_{n-1} + y_n = hf(t_{n-1}).$$

   Unter Hinzunahme der Randbedingung

   $$B_a y_0 + B_b y_N = g$$

   erhalten wir damit das System ($A_n := A(t_n)$, $f_n := f(t_n)$)

   $$\begin{bmatrix} B_a & 0 & \dots & 0 & B_b \\ -I - hA_0 & I & & & \\ & -I - hA_1 & I & & \\ & & \ddots & \ddots & \\ & & & -I - hA_{N-1} & I \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} = \begin{bmatrix} g \\ hf_0 \\ \vdots \\ \vdots \\ hf_{N-1} \end{bmatrix}.$$

   Um dieses System aufzustellen, benötigt man jeweils N Auswertungen von $A(t)$ und $f(t)$. Das entstehende LGS hat eine Blockstrucktur, so dass wir das LGS mit Aufwand $O(d^3 N)$ lösen können.

2. Im Falle des Schießverfahrens bestimmen wir zunächst die diskreten Fundamentalmatrizen Y_n^h aus den Lösungen der AWA

   $$Y'(t) - A(t)Y(t) = 0, \quad Y(a) = I$$

   sowie die Lösung y_0 der AWA

   $$y_0'(t) - A(t)y_0(t) = f(t), \quad y_0(a) = 0.$$

Hierzu werden ebenfalls nur N Auswertungen von A und f benötigt, da die Gitterweite für alle AWA gleich ist. Die Anzahl der Matrix-Vektor Multiplikationen ist aber $(d+1)\,N$ anstatt N.

Anschließend bestimmen wir mit $Q^h = B_a + B_b Y_N^h$:

$$Q^h s^h = g - B_b y_{0,N}^h$$

Die Lösung dieses LGS erfordert einen Aufwand $O\,(d^3)$. Anschließend erhalten wir unsere Näherungslösung u_n^h mittels:

$$u_n^h = y_{0,n}^h + Y_n^h s^h,$$

also durch weitere N Matrix-Vektor Multiplikationen ($O\,(Nd^2)$). Ferner benötigen wir für dieses vorgehen die Lösungen der $d+1$ AWA. Also einen Speicherplatz $O\,(Nd^2)$ anstatt $O\,(Nd)$ wie im Fall (a).

Lösung A.10.2: (i) Zunächst schreiben wir die RWA als System erster Ordnung (in Standardform):

$$\begin{bmatrix} u_0'(t) \\ u_1'(t) \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_0(0) \\ u_1(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_0(1) \\ u_1(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Für die Trapezregel gilt

$$L_h y_n := \frac{1}{h}(y_n - y_{n-1}) - \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (y_{n-1} + y_n) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

bzw.

$$\begin{bmatrix} -1 & -\frac{h}{2} \\ -\frac{h}{2} & -1 \end{bmatrix} y_{n-1} + \begin{bmatrix} 1 & -\frac{h}{2} \\ -\frac{h}{2} & 1 \end{bmatrix} y_n = \begin{bmatrix} 0 \\ -h \end{bmatrix}.$$

Zusammen mit den Randbedingungen ergibt sich also folgendes Gleichungssystem:

$$\begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} & \cdots & \cdots & \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} -1 & -\frac{h}{2} \\ -\frac{h}{2} & -1 \end{bmatrix} & \begin{bmatrix} 1 & -\frac{h}{2} \\ -\frac{h}{2} & 1 \end{bmatrix} & & \\ & \ddots & \ddots & \\ & & \begin{bmatrix} -1 & -\frac{h}{2} \\ -\frac{h}{2} & -1 \end{bmatrix} & \begin{bmatrix} 1 & -\frac{h}{2} \\ -\frac{h}{2} & 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} y_0 \\ \vdots \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 \\ -h \end{bmatrix} \\ \vdots \\ \begin{bmatrix} 0 \\ -h \end{bmatrix} \end{bmatrix}$$

Die Randwertaufgabe ist als Spezialfall eines Sturm-Liouville-Problems ($p = 1,\ q = 0,\ r = 1$) wegen

$$1 + (1-0)^2 \min_{t\in[0,1]} \{1-0\} = 2 > 0$$

eindeutig lösbar.

(ii) Zunächst schreiben wir die RWA als System erster Ordnung (in Standardform):

$$\begin{bmatrix} u_0'(t) \\ u_1'(t) \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_0(0) \\ u_1(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_0(1) \\ u_1(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Für die Trapezregel gilt

$$L_h y_n := \frac{1}{h}(y_n - y_{n-1}) - \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} (y_{n-1} + y_n) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

bzw.

$$\begin{bmatrix} -1 & -\frac{h}{2} \\ 0 & -1-\frac{h}{2} \end{bmatrix} y_{n-1} + \begin{bmatrix} 1 & -\frac{h}{2} \\ 0 & 1-\frac{h}{2} \end{bmatrix} y_n = \begin{bmatrix} 0 \\ -h \end{bmatrix}.$$

Zusammen mit den Randbedingungen ergibt sich also folgendes Gleichungssystem:

$$\begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} & \cdots & \cdots & \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} -1 & -\frac{h}{2} \\ 0 & -1-\frac{h}{2} \end{bmatrix} & \begin{bmatrix} 1 & -\frac{h}{2} \\ 0 & 1-\frac{h}{2} \end{bmatrix} & & \\ & \ddots & \ddots & \\ & & \begin{bmatrix} -1 & -\frac{h}{2} \\ 0 & -1-\frac{h}{2} \end{bmatrix} & \begin{bmatrix} 1 & -\frac{h}{2} \\ 0 & 1-\frac{h}{2} \end{bmatrix} \end{bmatrix} \begin{bmatrix} y_0 \\ \vdots \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 \\ -h \end{bmatrix} \\ \vdots \\ \begin{bmatrix} 0 \\ -h \end{bmatrix} \end{bmatrix}$$

Die Randwertaufgabe ist als Spezialfall eines Sturm-Liouville-Problems ($p = 1,\ q = 1,\ r = 0$) wegen

$$1 + (1-0)^2 \min_{t\in\ [0,1]} \{0-0\} = 1 > 0$$

eindeutig lösbar. Da die Trapezregel für AWAn von zweiter Ordnung ist, ist sie dies nach dem Äquivalenzsatz aus der Vorlesung ebenfalls für RWA.

Lösung A.10.3: a) Die RWA besitzt eine eindeutige Lösung, denn das entsprechende homogene Problem besitzt nur die triviale Lösung $u \equiv 0$, wie man wie folgt sieht: Sei u eine Lösung von

$$-u''(t) + 100u'(t) = 0, \quad t \in [0,1], \quad u(0) = u(1) = 0.$$

Multiplikation mit u und Integration über $I = [0,1]$ liefert

$$-\int_0^1 u''(t)u(t)\,dt + 100\int_0^1 u'(t)u(t)\,dt = 0$$

bzw. nach partieller Integration

$$-u'(t)u(t)\Big|_0^1 + \int_0^1 \big(u'(t)\big)^2\,dt + 50\int_0^1 \big(u(t)^2\big)'\,dt = 0.$$

Einsetzen der Randwerte liefert

$$\int_0^1 \big(u'(t)\big)^2\,dt = 0$$

und damit $u' \equiv 0$. Wegen $u(0) = u(1) = 0$ folgt also $u \equiv 0$.

b) Für die Gitterfunktion $(y_n^h)_n$ ergibt sich durch Anwendung der zentralen Differenzenquotienten

$$-h^{-2}\{y_{i+1}^h - 2y_i^h + y_{i-1}^h\} + 50h^{-1}\{y_{i+1} - y_{i-1}\} = 1, \quad i = 1,\dots,N.$$

Die Randbedingungen liefern $y_0^h = y_{N+1}^h = 0$. Durch Elimination der Randwerte y_0^h und y_N^h erhält man so das folgende Gleichungssystem:

$$h^{-2}\begin{bmatrix} 2 & 50h-1 & & & & \\ -50h-1 & 2 & 50h-1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -50h-1 & 2 & 50h-1 & \\ & & & -50h-1 & 2 & \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

Die Matrix ist offenbar diagonal-dominant, falls gilt

$$2 \geq |-50h-1| + |50h-1| = 50h + 1 + |50h-1|.$$

Fall 1: $50h - 1 \geq 0 \Leftrightarrow h \geq \frac{1}{50}$:

$$2 \geq 50h + 1 + 50h - 1 = 100h \Leftrightarrow h \leq \frac{1}{50}.$$

Fall 2: $50h - 1 < 0 \Leftrightarrow h < \frac{1}{50}$:

$$2 \geq 50h + 1 + 1 - 50h = 2.$$

Die Matrix ist also diagonal-dominant, falls $h \leq \frac{1}{50}$. Strikte Diagonaldominanz liegt nie vor.

c) Für die Gitterfunktion $(y_n^h)_n$ ergibt sich in diesem Fall

$$-h^{-2}\{y_{i+1}^h - 2y_i^h + y_{i-1}^h\} + 100h^{-1}\{y_i - y_{i-1}\} = 1, \quad i = 1,\dots,N.$$

Elimination der Randwerte y_0^h und y_{N+1}^h führt dann auf das Gleichungssystem

$$h^{-2}\begin{bmatrix} 2+100h & -1 & & & & \\ -100h-1 & 2+100h & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -100h-1 & 2+100h & -1 & \\ & & & -100h-1 & 2+100h & \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}.$$

Diese Matrix ist also wegen

$$2+100h \geq |-100h-1| + |-1| = 100h+2$$

für jede Schrittweite h diagonal-dominant. Strikte Diagonaldominanz liegt auch hier nie vor.

Lösung A.10.4: Für die Lösung der Differenzengleichungen (2. Ordnung)

$$-(\hat{\epsilon}+\tfrac{1}{2})y_{n-1} + 2\hat{\epsilon} - (\hat{\epsilon}-\tfrac{1}{2}h)y_{n+1} = 0, \quad y_0 = 1, \quad y_{N+1} = 0$$

machen wir einen Ansatz der Form $y_n = \lambda^n$. Mögliche Werte für λ sind dann gerade die Lösungen $\lambda_\pm$ der Gleichung

$$-\epsilon\lambda^2 + (2\epsilon+h)\lambda - (\epsilon+h) = 0,$$

wobei wir $\hat{\epsilon} = \epsilon + \frac{1}{2}h$ verwendet haben. Berücksichtigung der Randbedingungen $y_0 = 1$ und $y_{N+1} = 0$ in dem Ansatz

$$y_n = c_+\lambda_+^n + c_-\lambda_-^n$$

ergibt die Bedingungen $c_+ + c_- = 1$ sowie $c_+\lambda_+^{N+1} + c_-\lambda_-^{N+1} = 0$ und damit

$$c_- = \frac{\lambda_+^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}}, \quad c_+ = -\frac{\lambda_-^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}}.$$

Die Lösung hat also die Gestalt

$$y_n = \frac{\lambda_+^{N+1}\lambda_-^n - \lambda_-^{N+1}\lambda_+^n}{\lambda_+^{N+1} - \lambda_-^{N+1}}.$$

Im konkreten Fall sind die Lösungen $\lambda_\pm$ gegeben als

$$\lambda_\pm = \frac{-2\epsilon - h \pm \sqrt{(2\epsilon+h)^2 - 4\epsilon(\epsilon+h)}}{-2\epsilon} = \frac{-2\epsilon - h \pm h}{-2\epsilon}, \text{ d. h. } \lambda_+ = 1, \quad \lambda_- = \frac{\epsilon+h}{\epsilon}.$$

Die Lösung besitzt also die konkrete Darstellung

$$y_n = \frac{\lambda_-^{N+1} - \lambda_-^n}{\lambda_-^{N+1} - 1} \quad \text{mit } \lambda_- = \frac{\epsilon+h}{\epsilon} > 1 > 0.$$

Insbesondere ist damit $y_n \geq 0$ (oszillationsfrei) monoton fallend.

A.11 Kapitel 11

Lösung A.11.1:

1. Der Satz von Picard-Lindelöf fordert die L-Stetigkeit der AWA und ergibt neben der lokalen Existenz einer Lösung auch deren Eindeutigkeit.

2. $u(t) = (1-t)^{-1}$.

3. Die lokale Lösung einer (stetigen) AWA ist in der Zeit fortsetzbar, bis ihr Graph an den Rand des Definitionsbereich der rechten Seite $f(t,x)$ stößt.

4. Ja.

5. Der Abschneidefehler entsteht durch Einsetzen der exakten Lösung in die Differenzenformel:
$$h_n t_n^h := u_n - u_{n-1} - h_n F(h_n; t_n, u_n, u_{n-1}).$$

6. Wenn für die Eigenwerte $\lambda(t)$ der Jacobi-Matrix $f'_x(t,u(t))$ entlang der Lösungstrajektorie gilt:
$$\frac{\max_{\mathrm{Re}\lambda(t)<0} |\mathrm{Re}\lambda(t)|}{\min_{\mathrm{Re}\lambda(t)<0} |\mathrm{Re}\lambda(t)|} \gg 1.$$

7. Wenn sie konsistent und null-stabil ist.

8. Die LMM
$$\sum_{r=0}^{R} \alpha_{R-r} y_{n-r} = h \sum_{r=0}^{R} \beta_{R-r} f_{n-r}$$
hat das Stabilitätspolynom
$$\pi(\lambda, qh) = \sum_{r=0}^{R} \{\alpha_r - qh\lambda\beta_r\}\lambda^r.$$

9. Die Ordnung ist $m = 2$.

10. Die Simpson- und die Mittelpunkts-Methode haben ein triviales Stabilitätsgebiet.

11. Die Mittelpunkts-Methode ist explizit und erlaubt eine asymptotische Fehlerentwicklung nach geraden Potenzen von h.

12. Eine Funktion $a(h)$ wird für eine Folge von Werten $h_0 > h_1 > \cdots > h_m > 0$ ausgewertet. Der Grenzwert $a(0)$ wird durch den Wert $p_m(0)$ des Interpolationspolynoms m-ten Grades zu den Stützwerten $\{h_0, a(h_0)\}, \ldots, \{h_m, a(h_m)\}$ approximiert.

13. Für die Schrittweiten $h_i = h/i$ gilt $\limsup_{i\to\infty} h_{i+1}/h_1 = 1$, so dass in der Fehlerabschätzung für das Extrapolationsverfahren die Konstante im Restglied nicht beschränkt bleibt.

14. A-Stabilität bedeutet, dass das Stabilitätsgebiet der Diffenzenformal die ganze „negative“ komplexe Halbenene enthält. $A(\alpha)$-Stabilität verlangt dies nur für einen Sektor mit Halbwinkel $\alpha > 0$.

15. Bessere Stabilität beim Lösen der AWAn wegen kürzerer Integrationsintervallen.

16. Die algebraische Nebenbedingung enthält explizit die „algebraische“ Variable und ihre Jacobi-Matrix bzgl. dieser ist regulär.

17. Die Fundamentalmatrix $Y(t)$ ist die Lösung der linearen Matrix-AWA $Y'(t) = A(t)Y(t),\ t \in [a,b],\ Y(a) = I$.

18. Ein „reguläres Sturm-Liouville-Problem“ ist eine lineare RWA 2-ter Ordnung

$$-[pu']'(t) + q(t)u'(t) + r(t)u(t) = f(t),\ \ t \in [a,b], \quad \text{Randbedingungen.}$$

19. Es wird $u(a) = g_a$ und $u(b) = g_b$ gefordert.

20. Die Trapezregel angewendet auf eine lineare RWA hat die Gestalt:

$$\begin{aligned} y_n - y_{n-1} &= \tfrac{1}{2}h\big\{A(t_n)y_n + A(t_{n-1})y_{n-1} + f(t_n) + f(t_{n-1})\big\}, \quad n = 1,\dots,N, \\ B_a y_0 + B_b y_N &= g. \end{aligned}$$

Lösung A.11.2: Das Graggsche Extrapolationsverfahren basiert auf der asymptotischen Entwicklung

$$y_n = u(t_n) + \sum_{k=1}^{m} h^{2k}\left\{a_k(t_n) + (-1)^n b_k(t_n)\right\} + \mathcal{O}(h^{2m+2})$$

der Mittelpunktsregel gestartet mit der Polygonzugmethode. Diese erlaubt somit bei Wahl gerader Koeffizienten $n_i = 2,4,6,8,12,16,\cdots$ in $h_i = \frac{H}{n_i}$ eine Extrapolation zum Limes in 2. Ordnung.

Man berechnet also konkret:

1. Zu einer Grundschrittweite H mit ganzen Zahlen $n_i = 2,4,6,8,\cdots$ werden

$$\eta(t_n + \nu h_i; h_i),\ h_i = H/n_i,\ \nu = 1,\cdots,n_i+1$$

berechnet mit

$$\begin{aligned} \eta(t_n + h_i; h_i) &= y_n + h_i f(t_n, y_n), \\ \eta(t_n + (\nu+1)h_i; h_i) &= \eta(t_n + (\nu-1)h_i; h_i) + 2h_i f(t_n + \nu h_i, \eta(t_n + \nu h_i; h_i) \end{aligned}$$

.

2. Mittelungsprozedur:

$$a(h_i) = \tfrac{1}{4}\left(\eta(t_{n+1} = h_i; h_i) + \eta(t_{n+1}; h_i) + \eta(t_{n+1} + h_i; h_i)\right).$$

3. Extrapolation:

$$T_{i0} = a(h_i),$$
$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{(h_{i-k}/h_i)^2 - 1},$$

sowie dem finalen Setzen von

$$y_{n+1} = T_{mm}.$$

Lösung A.11.3: Die adaptive Schrittweitenwahl beruht auf einer Entwicklung des Abschneidefehlers,

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

und einer Schätzung des *Hauptabschneidefehlers* τ^m durch

$$\tau^m(t_n) \approx \frac{y_n^{H/2} - y_n^H}{H^{m+1}(1 - 2^{-m})} \qquad (*)$$

mit Hilfe der Näherungen y_n^H (mit Schrittweite H) und $y_n^{H/2}$ (mit halber Schrittweite $H/2$) von $u(t_{n-1} + H)$.

Aus der a priori-Fehlerabschätzung

$$\max_{t_n \in I} \|e_n\| \leq KT \max_{t_n \in I} \|t_n\|$$

gewinnt man die Schrittweitenbedingung

$$h_n \approx \left(\frac{\mathrm{TOL}}{KT \|t^m(t_n)\|} \right)^{1/m}, \qquad (**)$$

so dass $\max_{t_n \in I} \|e_n\| \leq \mathrm{TOL}$.

Dies motiviert z. B. den folgenden Algorithmus:

1. Sei y_n berechnet, Setze $H := 2h_n$.
2. Berechne y_{n+1}^H, $y_{n+1}^{H/2}$ und schätze mit (*) und (**) die notwendige Schrittweite $\tilde{h}_n$ ab.
3. $h_n \ll \frac{1}{2}H$? Falls ja und $\tilde{h}_n > h_{\min}$, setze $H = 2h_{n+1}$ und mache bei (2.) weiter.
4. Ansonsten: Setze $h_{n+1} = H$, $t_{n+1} = t_n + H$,
$$y_{n+1} = \frac{2^m y_{n+1}^{H/2} - y_{n+1}^H}{2^m - 1}.$$
5. Falls $t_{n+1} < T$ Weiter bei (1.)

Lösung A.11.4: a) Die AWA heißt steif, wenn mit den Eigenwerten $\lambda(t)$ der Jacobi-Matrix $f'_x(t, u(t))$ entlang der Lösungstrajektorie gilt:

$$\kappa(t) = \frac{\max_{\mathrm{Re}\lambda(t)<0} |\mathrm{Re}\lambda(t)|}{\min_{\mathrm{Re}\lambda(t)<0} |\mathrm{Re}\lambda(t)|} \gg 1.$$

b) Eine LMM ist „A-stabil“, falls die linke komplexe Halbebene ganz im Stabilitätsgebiet liegt. Eine LMM ist „A(0)-stabil“, falls die negative reelle Achse ganz im Stabilitätsgebiet liegt. Sind die Eigenwerte $\lambda(t)$ der Jacobi-Matrix $f'_x(t, u(t))$ stets reellwertig, so ist keine Schrittweitenrestriktion nötig um numerisch stabil zu integrieren.

c) Die Eigenwerte von A sind $\lambda_1 = -1$ und $\lambda_2 = -399$, so dass

$$\kappa(t) = 399 \gg 1,$$

d. h. die AWA ist steif. Das Stabilitätsintervall der Polygonzugmethode ist $\mathrm{SI} = [-2, 0]$. Um mit der Polygonzugmethode numerisch stabil zu integrieren fordern wir:

$$\lambda_i \cdot h \in [-2, 0] \quad \Longrightarrow \quad h \leq \frac{2}{399}.$$

Die Trapezregel ist A-stabil, d. h. nach (b) ist keine Schrittweitenbedingung notwendig.

Lösung A.11.5: Zur Lösung der nichtlinearen RWA

$$\begin{aligned} &u'(t) = f(t, u(t)), \quad t \in I, \\ &r(u(a), u(b)) = 0 \end{aligned}$$

definiert beim beim Schießverfahren $y(t; s)$ als (eindeutige) Lösung der AWA

$$\begin{aligned} &y'(t) = f(t, y(t)), \quad t \in I, \\ &y(a) = s. \end{aligned}$$

Das Lösen der RWA ist dann äquivalent zum Finden einer Nullstelle von

$$F(s) := r(s, y(b; s)) = 0.$$

Die Nullstelle sucht man dabei mit Hilfe des Newtonverfahrens. Hierbei ist in der Jacobi-Matrix

$$F'(s) = r'_x(s, y(b; s)) + r'_y(s, y(b; s)) \cdot y'_s(b; s)$$

der unbekannte Term $G(t; s) := y'_s(t; s)$ nach dem Satz von der differentiellen Stabilität bestimmt durch die Matrix-AWA

$$\begin{aligned} &G(t; s)'(t) = f'_x(t, y(t; s))G(t; s)(t), \quad t \in I, \\ &G(a; s) = I. \end{aligned}$$

Dies motiviert das folgende Vorgehen:

Sei $s^{(0)}$ ein geeigneter Startwert. Berechne $s^{(i+1)}$ aus der vorangegangenen Schätzung $s^{(i)}$ durch

1. Löse AWA

$$y'(t) = f(t, y(t)), \quad t \in I,$$
$$y(a) = s^{(i)}.$$

Setze damit

$$F(s^{(i)}) := r(s^{(i)}, y(b; s^{(i)})).$$

2. Löse die lineare Matrix-AWA

$$G(t; s^{(i)})'(t) = f'_x(t, y(t; s^{(i)}))G(t; s^{(i)})(t), \quad t \in I,$$
$$G(a; s^{(i)}) = I$$

und setze

$$F'(s^{(i)}) := r'_x(ss^{(i)}, y(b; ss^{(i)})) + r'_y(ss^{(i)}, y(b; ss^{(i)})) \cdot G(b; s^{(i)}).$$

3. Löse abschließend

$$F'(s^{(i)})s^{(i+1)} = F'(s^{(i)})s^{(i)} - F(s^{(i)}).$$

Lösung A.11.6 (Praktische Aufgabe): Zur Approximation des (nicht-steifen) Lorenz-Systems verwenden wir das dG(1)-Verfahren mit adaptiver Schrittweitenwahl gemäß der im Text beschriebenen Strategie (für Kontrolle des „Endzeitfehlers“ $e(T)$ mit Fehlertoleranz $\mathrm{TOL} = 0.2$),

$$\|e_N^-\| \leq C_i^{(2)} C_s^{(2)}(t_N) \max_{1 \leq n \leq N} \left\{ h_n \|[U]_{n-1}\| + h_n^3 \|\ddot{f}(t, U)\|_n \right\}. \tag{1.11.3}$$

Die dabei auftretende Stabilitätskonstante wächst exponentiell mit der Zeit wie

$$C_S^{(1)}(T) \approx \exp(\tfrac{2}{3}T).$$

Für $T = 25$ ist dies $C_s^{(2)}(25) \approx 10^8$ und für $T = 35$ bereits $C_s^{(2)}(35) \approx 10^{13}$. Dies bedeutet, dass verlässliche Rechnungen mit dem dG(1)-Verfahren mit einer akzeptable Schrittweite $h_n \geq 10^{-7}$ bis zum Zeitpunkt $T = 35$ möglich sind; s. Abb. A.11.6. Weit darüberhinaus kommt man mit dem (impliziten) dG(1)-Verfahren wegen seiner niedrigen Ordnung und der weiter wachsenden Stabilitätskonstante $C_s^{(2)}(40) \approx 10^{16}$ nicht mehr, da dann wegen des enorm kleinen lokalen Zeitschritts der Rundungsfehler dominiert (bei Rechnung in REAL*8-Arithmetik). Abb. A.11.6 zeigt, dass die Leistungsgrenze scheint bei etwa $T_{crit} \approx 37$ liegt. Zum Vergleich wurde eine genaue Referenzlösung auf dem Intervall $I = [0, 40]$ erzeugt mit Hilfe eines „Hochleistungs-ODE-Lösers“ (entwickelt von Reinhold v.Schwerin, HS Ulm) auf Basis von Adams-Formeln variabler Ordnung (maximale Ordnung 12) mit adaptiver Schrittweitenkontrolle (Fehlertoleranz 10^{-29}) und Rechnung mit REAL*16-Arithmetik.

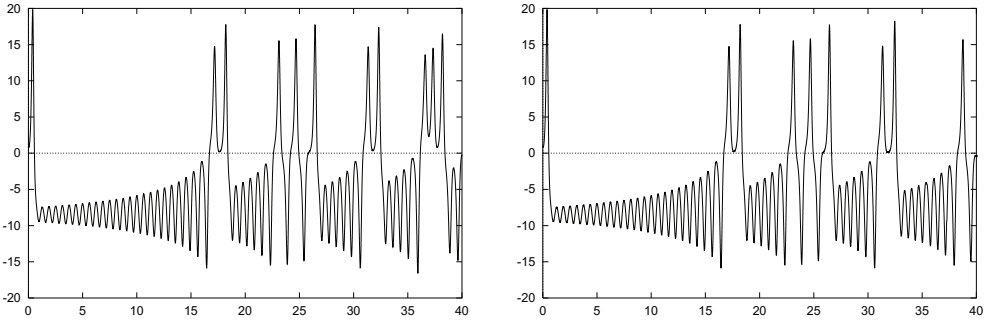


Abbildung A.1: Referenzlösung (x-Komponente) (links) und „falsche“ Lösung (x-Komponente) (rechts) auf dem Intervall $I = [0, 40]$ berechnet mit dem adaptiven dG(1)-Verfahren mit TOL $= 10^{-4}$.