3 Numerische Stabilität

3.1 Modellproblemanalyse

Eine Lipschitz-stetige und *(strikt) monotone* AWA

$$u'(t) = f(t, u(t)), \quad t \geq t_0, \quad u(t_0) = u_0, \tag{3.1.1}$$

hat im Falle $\sup_{t>0} \|f(t,0)\| < \infty$ eine globale, gleichmäßig beschränkte Lösung. Ist $f(t,0) \equiv 0$, so fällt diese Lösung sogar exponentiell gegen Null ab. Seien $L(t)$ die Lipschitz-Konstante und $\lambda(t)$ die Monotonie-Konstante der Funktion $f(t,\cdot)$. Wir haben gesehen (siehe Übungsaufgabe), dass das Polygonzugverfahren eine analoge Eigenschaft besitzt, wenn die strikte Schrittweitenbedingung

$$\inf_{n\geq 0} \left\{ \frac{2\lambda_{n-1}}{h_n L_{n-1}^2} \right\} > 0 \tag{3.1.2}$$

erfüllt ist. Für solche Schrittweiten ist das Verfahren also „numerisch stabil“. Anhand der skalaren Testgleichung

$$u'(t) = \lambda u(t), \quad \lambda \in \mathbb{C}, \tag{3.1.3}$$

($L = |\lambda|$) sieht man, dass die Bedingung (3.1.2) i. Allg. scharf ist. Für $\lambda \in \mathbb{R}$, $\lambda < 0$, gilt

$$y_n = (1 + h\lambda) y_{n-1} = ... = (1 + h\lambda)^n y_0,$$

d. h.: Für $h > 2|\lambda_{n-1}|/L_{n-1}^2 = 2/|\lambda|$ wächst die diskrete Lösung exponentiell, für $h = 2|\lambda_{n-1}|/L_{n-1}^2 = 2/|\lambda|$ bleibt sie beschränkt (absolutbetragsmäßig sogar konstant) und für $h < 2|\lambda_{n-1}|/L_{n-1}^2 = 2/|\lambda|$ fällt sie exponentiell. Die Testgleichung (3.1.3) wird in der weiteren Diskussion im Komplexen betrachtet, da der Parameter $\lambda \in \mathbb{C}$ für die i. Allg. komplexen Eigenwerte der Jacobi-Matrix $f'_x(t,x)$ steht.

Zur Illustration betrachten wir folgendes Beispiel

$$u'(t) = -200\, t\, u(t)^2, \quad t \geq 0, \quad u(0) = 1,$$

mit der Lösung $u(t) = (1 + 100t^2)^{-1}$. Es soll der Wert $u(3) = 1/901$ mit Hilfe des klassischen Runge-Kutta-Verfahrens approximiert werden. Nach den Ergebnissen zur Konvergenz dieses Verfahrens dürften dabei keine Probleme auftreten, insbesondere da die Lösung $u(t)$ für $t \to \infty$ sehr glatt gegen Null abfällt. Man ist daher versucht, mit relativ großen Schrittweiten zu rechnen. Bei 17-stelliger Rechnung erhält man jedoch das in Tabelle 3.1 wiedergegebene bedenkliche Resultat (N = Schrittzahl). Offensichtlich zeigt das ansonsten sehr gutartige Runge-Kutta-Verfahren bei diesem Problem eine numerische Instabilität, wenn die Schrittweite zu grob ist. Im Folgenden wollen wir uns mit der Analyse und Kontrolle solcher gefährlichen Instabilitäten beschäftigen.

Tabelle 3.1: *Beispiel numerischer Instabilität.*

N	h	$\lvert y_N - u(3)\rvert$
50	0.06	$\sim 2 \cdot 10^{-8}$
25	0.12	$\sim 2 \cdot 10^{-6}$
20	0.15	$\sim 7 \cdot 10^{-5}$
15	0.2	overflow (10^{38})
20	0.1538	$\sim 7 \cdot 10^{-5}$
19	0.1579	overflow (10^{38})

Lineare Stabilitätsanalyse

Wir nennen zunächst intuitiv ein Differenzenverfahren „numerisch stabil“ für festes h, wenn im Falle $\sup_{t>0} \|u(t)\| < \infty$ auch $\sup_{n\geq 0} \|y_n\| < \infty$. Zur Illustration sei das einfache Testproblem (3.1.3) betrachtet. Das Verhalten der Lösung $u(t) = u_0 e^{\lambda t}$ für $t \to \infty$ ist charakterisiert durch das Vorzeichen von $\operatorname{Re}\lambda$:

$$\left.\begin{array}{l} \operatorname{Re}\lambda < 0 \\ \operatorname{Re}\lambda = 0 \\ \operatorname{Re}\lambda > 0 \end{array}\right\} \Rightarrow |u(t)| = |u_0| e^{\operatorname{Re}\lambda} \left\{\begin{array}{l} \to 0 \\ \equiv |u_0| \\ \to \infty \end{array}\right. \tag{3.1.4}$$

Definition 3.1 (Absolute Stabilität): *Eine Einschrittmethode heißt „absolut stabil“ für ein $\lambda h \neq 0$, wenn sie angewendet auf das skalare Testproblem* (3.1.3) *für $\operatorname{Re}\lambda \leq 0$ beschränkte Näherungen erzeugt:* $\sup_{n\geq 0} |y_n| < \infty$.

Für die Polygonzugmethode liegt also absolute Stabilität genau dann vor, wenn für den sog. „Verstärkungsfaktor“ $\omega = \omega(\lambda h) := 1 + \lambda h$ gilt $|\omega| \leq 1$. Wir nennen allgemein

$$\mathrm{SG} = \{z = \lambda h \in \mathbb{C} : |\omega(z)| \leq 1\}$$

das „Gebiet absoluter Stabilität“ (kurz „Stabilitätsgebiet“) einer Einschrittformel. Das Stabilitätsgebiet der Polygonzugmethode ist in Abb. 3.1 dargestellt. Für ein festes λ mit $\operatorname{Re}\lambda \leq 0$ muss die Schrittweite h so bemessen sein, dass $\lambda h \in \mathrm{SG}$ ist. Andernfalls wächst die Näherungslösung y_n für $n \to \infty$ exponentiell an, obwohl die exakte Lösung beschränkt ist oder sogar exponentiell abfällt.

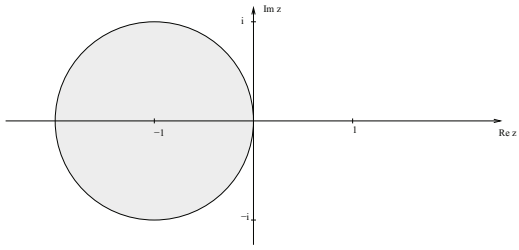

Abbildung 3.1: *Stabilitätsgebiet der Polygonzugmethode.*

Wir wollen nun die numerische Stabilität der Taylor- und der Runge-Kutta-Formeln untersuchen. Für das Testproblem (3.1.3) erhält die Taylor-Methode der Stufe R die Gestalt

$$y_n = y_{n-1} + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{r-1}(t_{n-1}, y_{n-1}) = y_{n-1} + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} \lambda^r y_{n-1}.$$

Der Verstärkungsfaktor ist also

$$\omega = \sum_{r=0}^{R} \frac{(\lambda h)^r}{r!}. \tag{3.1.5}$$

Da die Bestimmung des vollen Stabilitätsgebietes $\mathrm{SG} = \{z \in \mathbb{C} : |\omega(z)| \leq 1\}$ schwierig ist, beschränken wir uns hier auf die Betrachtung des „Stabilitätsintervalls“

$$\mathrm{SI} = \{z \in \mathbb{R} : |\omega(z)| \leq 1\}.$$

Wir finden

$$\mathrm{SI} = \begin{cases} [-2, 0] & , \quad R = 1 \\ [-2, 0] & , \quad R = 2 \\ [-2.51\ldots, 0] & , \quad R = 3 \\ [-2.78\ldots, 0] & , \quad R = 4 \end{cases}$$

Sei $F(h; t, x)$ die Verfahrensfunktion einer R-stufigen Runge-Kutta-Methode der Ordnung $m = R \leq 4$. Nach Konstruktion der Runge-Kutta-Formeln gilt dann

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

Für das Testproblem ist

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

offenbar ein Polynom in h der Ordnung $R-1$. Folglich gilt in diesem Fall

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

d. h.: Der Verstärkungsfaktor ω der Runge-Kutta-Formeln der Ordnung ($m = R \leq 4$) ist derselbe wie der der entsprechenden Taylor-Formeln. Also sind durch die obige Abbildung für $R \leq 4$ auch die Stabilitätsintervalle der R-stufigen Runge-Kutta-Formeln der Ordnung $m = R$ gegeben.

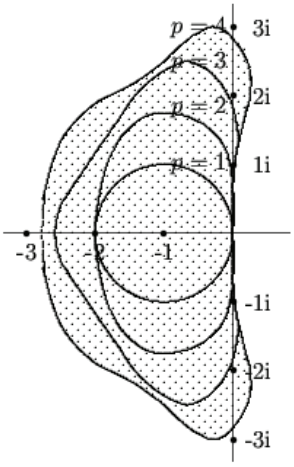

Abbildung 3.2: *Stabilitätsgebiete der (expliziten) Taylor- und Runge-Kutta-Verfahren.*

Der obigen Stabilitätsanalyse entnehmen wir, dass für $\operatorname{Re}\lambda \ll -1$ die Stabilität der durch die Runge-Kutta-Verfahren erzeugten Lösungen die Verwendung einer entsprechend kleinen Schrittweite h erfordert. In diesem Fall wäre daher die Verwendung einer Formel mit einem in der komplexen Ebene möglichst weit nach links reichendem Stabilitätsgebiet. Die in dieser Hinsicht „optimalen“ Methoden haben die folgende Eigenschaft:

Definition 3.2: *Eine Differenzenmethode heißt „A-stabil“, wenn für ihr Stabilitätsgebiet gilt*

$$\{z \in \mathbb{C} \,|\, \operatorname{Re} z \leq 0\} \subset SG. \tag{3.1.6}$$

Man kann zeigen, dass *explizite* Methoden nicht A-stabil sein können. Wir werden später sehen, dass die *implizite* Euler-Methode sowie die Trapezregel

$$\begin{aligned} y_n &= y_{n-1} + hf(t_n, y_n), \\ y_n &= y_{n-1} + \tfrac{1}{2}h\{f(t_n, y_n) + f(t_{n-1}, y_{n-1})\} \end{aligned}$$

A-stabil sind.

Nutzung der linearen Stabilitätsanalyse für allgemeine Systeme

Wir wenden uns nun der Frage nach der „numerischen Stabilität“ von Einschrittverfahren für allgemeine (nicht notwendig monotone) Systeme 1. Ordnung der Form (3.1.1) zu. Dazu müssen wir zunächst erklären, was im Folgenden unter der „Stabilität“ der Lösung einer AWA zu verstehen ist. Basierend auf der Diskussion von „Stabilität“ in Kapitel 1 führen wir folgende Begriffe ein:

Definition 3.3: *Die (globale) Lösung u einer AWA*

$$u'(t) = f(t, u(t)), \quad t \geq t_0, \quad u(t_0) = u_0, \tag{3.1.7}$$

wird „(asymptotisch) stabil“ genannt, wenn jede Lösung v der gestörten AWA

$$v'(t) = f(t, v(t)), \quad t \geq t_*, \quad v(t_*) = u(t_*) + w_*, \tag{3.1.8}$$

zu einem Zeitpunkt $t_* \geq t_0$ mit einer hinreichend kleinen Störung $\|w_*\| \leq \delta$ ebenfalls global ist und folgendes gilt:

$$\|(v-u)(t)\| \to 0 \quad (t \to \infty). \tag{3.1.9}$$

Bemerkung 3.1: In der Literatur findet man auch noch eine Reihe anderer Konzepte von „Stabilität“ für die Lösungen von AWAn. Statt der Konvergenz $\|(v-u)(t)\| \to 0$ für $(t \to \infty)$ wird manchmal nur die Beschränktheit $\sup_{t \geq t_*} \|(v-u)(t)\| \leq \varepsilon$ für beliebig kleines $\varepsilon > 0$ gefordert, wobei die Größe der Anfangsstörung an das ε gekoppelt ist: $\|w_*\| \leq \delta(\varepsilon)$. Der stärkste Stabilitätsbegriff ist der der „exponentiellen Stabilität“, bei der

$$\|(v-u)(t)\| \leq Ae^{-\alpha(t-t_*)}\|w_*\|, \tag{3.1.10}$$

d. h. exponentieller Fehlerabfall proportional zur Anfangsstörung, gefordert wird. Ein hinreichendes Kriterium für exponentielle Stabilität (und damit auch für asymptotische Stabbilität) ist, wie wir gesehen haben, die (starke) Monotonie der AWA. Da aber auch nicht-monotone AWA stabile Lösungen haben können, operieren wir im Folgenden mit dem etwas weniger einschränkenden Konzept der (asymptotischen) Stabilität.

In Anlehnung an die vorausgehende Definition führen wir nun analoge Stabilitätsbegriffe für die Diskretrisierungen von AWA durch Einschrittverfahren ein.

Definition 3.4: *Die AWA (3.1.7) sei mit einem Einschrittverfahren*

$$y_n = y_{n-1} + h_n F(h_n; t_n, y_n, y_{n-1}), \quad n \geq 0, \quad y_0 = u_0, \tag{3.1.11}$$

mit L-stetiger Verfahrensfunktion diskretisiert. Eine (globale) Lösung $(y_n)_{n \geq 0}$ heißt „(numerisch) stabil“, wenn für jede Lösung $(z_n)_{n \geq n_*}$ von

$$z_n = z_{n-1} + h_h F(h_n; t_n, z_n, z_{n-1}), \quad n \geq n_*, \quad z_{n_*} = y_{n_*} + w_*, \tag{3.1.12}$$

zu einem Zeitpunkt $t_{n_*} \geq t_0$ mit einer hinreichend kleinen Störung $\|w_*\| \leq \delta$ gilt:

$$\|z_n - y_n\| \to 0 \quad (n \to \infty). \tag{3.1.13}$$

Bemerkung 3.2: Analog zu kontinuierlichen Fall wird die Lösung $(y_n)_{n\geq 0}$ einer Differenzenapproximation als „exponentiell stabil“ bezeichnet, wenn für jede Lösung der gestörten Diffenzengleichung gilt:

$$\|z_n - y_n\| \leq Ae^{-\alpha(t_n - t_{n_*})}\|w_*\|, \quad n \geq n_*. \tag{3.1.14}$$

Die direkte Anwendung der anhand des Testproblems (3.1.3) gewonnenen Erkenntnisse zur absoluten Stabilität einer Differenzenformel für allgemeine Systeme setzt folgendes voraus:

Hypothese: *Die (globale) Lösung u der allgemeinen AWA sei asymptotisch stabil und alle Eigenwerte $\lambda(t)$ der Jacobi-Matrix $f'_x(t,u(t))$ haben die Eigenschaft $\operatorname{Re}\lambda(t) \leq 0$. Dann ist ein Differenzenverfahren mit einem Stabilitätsgebiet $\mathrm{SG} \subset \mathbb{C}$ „numerisch stabil“, wenn die Schrittweiten h_n so gewählt werden, dass gilt:*

$$h_n\,\lambda(t_n) \in \mathrm{SG}\,, \quad n \geq 0\,. \tag{3.1.15}$$

Die Berechtigung dieser Hypothese ist in allgemeinen Situationen schwer zu klären. Anhand von Beispielen zeigt sich, dass sie falsch sein kann, wenn die Jacobi-Matrix $f'_x(t,u(t))$ nicht *diagonalisierbar* ist, d. h. kein vollständiges System von Eigenvektoren besitzt. Wir wollen die wesentlichen Schritte zur Rechfertigung der Hypothese skizzieren.

(i) Zunächst wollen wir diese Frage für das kontinuierliche Problem diskutieren. Seien also u und v (globale) Lösungen der AWAn

$$\begin{aligned} u'(t) &= f(t,u(t)), \;\; t \geq t_0, \quad u(t_0) = u_0, \\ v'(t) &= f(t,v(t)), \;\; t \geq t_*, \quad v(t_*) = u(t_*) + w_*. \end{aligned}$$

mit einer „kleinen“ Störung w_*. Für die Differenz $w := v - u$ gilt dann

$$\begin{aligned} w'(t) &= f(t,v(t)) - f(t,u(t)) = \int_0^1 \frac{d}{ds} f(t,u(t)+sw(t))\,ds \\ &= \int_0^1 f'_x(t,u(t)+sw(t))\,ds\,w(t) = f'_x(t,u(t))w(t) + \mathcal{O}\big(\|w(t)\|^2\big) \end{aligned}$$

Reduktionsschritt 1 (Linearisierung): Bei Vernachlässigung des quadratischen, und damit als „klein“ anzunehmenden, Terms $\|w(t)\|^2$ genügt die Differenz w näherungsweise der linearen AWA

$$w'(t) = f'_x(t,u(t))w(t), \;\; t \geq t_*, \quad w(t_*) = w_*. \tag{3.1.16}$$

Diese beschreibt im Rahmen einer (lokalen) differentiellen Stabilitätsanalyse bei t_* das Anwachsen oder Abfallen von Störungen. Man beachte, dass $\|f'_x(t,u(t))\|$ ein Maß für die lokale Lipschitz-Stetigkeit von $f(t,x)$ ist.

Reduktionsschritt 2 (Lokalisierung): Nach „Einfrieren“ des Koeffizienten zum Zeitpunkt $t_* \geq t_0$ erhält man das autonome (lineare) System

$$w'(t) = f'_x(t_\star,u(t_\star))w(t), \;\; t \geq t_*, \quad w(t_*) = w_*. \tag{3.1.17}$$

Reduktionsschritt 3 (Separation): Ist nun die Matrix $A := f'_x(t_\star, u(t_\star))$ diagonalisierbar, so existiert eine reguläre Matrix Q, so dass

$$QAQ^{-1} = D = \mathrm{diag}(\lambda_i) \tag{3.1.18}$$

mit den Eigenwerten $\lambda_i \in \mathbb{C}\,(i = 1, \ldots, d)$ von A. Die Funktion $\bar{w}(t) := Qw(t)$ ist dann Lösung von

$$\bar{w}'(t) = QAQ^{-1}\bar{w}(t) = D\bar{w}(t)\,, \quad t \geq t_\star\,. \tag{3.1.19}$$

Dieses Diagonalsystem zerfällt in die d skalaren Gleichungen

$$\bar{w}_i'(t) = \lambda_i \bar{w}_i(t)\,, \quad t \geq t_\star\,, \quad i = 1, \ldots, d\,. \tag{3.1.20}$$

Das Verhalten der einzelnen Komponenten $\bar{w}_i$ für $t \to \infty$ ist wieder charakterisiert durch die Realteile von λ_i. Wegen der Regularität von Q folgt die Beziehung

$$\mathrm{Re}\,\lambda_i \leq 0 \;\; (i = 1, \ldots, d) \quad \Leftrightarrow \quad \|w(t)\| \leq c\|w(t_\star)\|\,, \quad t \geq t_\star\,. \tag{3.1.21}$$

Die Stabilität des diagonalisierbaren Systems (3.1.17) wird also vollständig durch die Eigenwerte λ_i von A beschrieben. Über die skizzierte Argumentationskette (Reduktionsschritte 1 – 3) wird die Stabilitätsanalyse für eine allgemeine AWA lokal auf die Untersuchung der Eigenwerte der Jacobi-Matrix $A = f'_x(t_\star, u(t_\star))$, zurückgeführt.

(ii) Die numerische Stabilitätsanalyse verläuft analog in umgekehrter Richtung. Wir diskutieren hier nur den kritischen Übergang vom skalaren Modellproblem zum allgemeinen linearen System. Die verbleibenden Schritte „Lokalisierung“ und „Linearisierung“ sind analog wie im kontinuierlichen Fall. Alle betrachteten Einschrittverfahren haben für das System (3.1.17) die Form

$$y_n = g(hA)y_{n-1}$$

mit einer rationalen Funktion $g(z)$. Z. B. ist bei den Taylor-Formeln (und bei den Runge-Kutta-Formeln mit $m = R \leq 4$)

$$g(z) = \sum_{r=0}^{R} \frac{z^r}{r!}\,.$$

Sei die Matrix A wieder als diagonalisierbar angenommen. Wir setzen $\bar{y}_n = Qy_n$ und finden

$$\bar{y}_n = Qy_n = Qg(hA)y_{n-1} = Qg(hA)Q^{-1}\bar{y}_{n-1}\,.$$

Aufgrund eines allgemeinen Satzes über analytische Matrizenfunktionen ist

$$Qg(hA)Q^{-1} \equiv g(hQAQ^{-1}) = g(hD)\,,$$

und folglich

$$\bar{y}_n = g(hD)\bar{y}_{n-1} = g(hD)^n\bar{y}_0\,,$$

bzw.

$$\bar{y}_{n,i} = g(h\lambda_i)^n\bar{y}_{0,i}\,, \quad i = 1, \ldots, d\,.$$

Wegen der eindeutigen Kopplung $\bar{y}_n \equiv Qy_n$ können wir uns bei der Stabilitätsbetrachtung also auf die *skalare* Differentialgleichung $u'(t) = \lambda u(t)$ beschränken, wobei der Parameter $\lambda \in \mathbb{C}$ die Eigenwerte der Matrix A durchläuft. Es sei betont, dass die entscheidende Voraussetzung für die Gültigkeit dieser Überlegung die angenommene Diagonalisierbarkeit der Matrix A, d. h. i. Allg. Fall der Jacobi-Matrix des Systems, ist. Andernfalls kann, wie Gegenbeispiele zeigen, die vereinfachte skalare Analyse beim Übergang zu Systemen zu Fehleinschätzungen führen. Die ebenfalls vorgenommene lokale Linearisierung sowie das „Einfrieren“ der Koeffizienten ist dagegen weniger kritisch.

Beispiel 3.1: Bei dem nichtlinearen Problem vom Anfang dieses Kapitels

$$u'(t) = -200\, t\, u(t)^2\,, \quad t \in [0,3]\,, \quad u(0) = 1\,,$$

gilt entlang der Lösungstrajektorie

$$f_x(t,u(t)) = -400tu(t) = -\frac{400t}{1+100t^2}, \qquad \min_{t\in[0,3]} f_x(t,u(t)) = -20.$$

Für das klassische Runge-Kutta-Verfahren mit dem Stabilitätsintervall $SI \approx [-2.78, 0]$ impliziert dies die Schrittweitenbeschränkung $h < 2.78/20 = 0.139$. Da die L-Konstante bei diesem nichtlinearen Problem außerhalb des relativ kleinen Intervalls $[0, \frac{1}{5}]$ überall < 16 ist, wird die Instabilität im Bereich $h \sim 0.14$ nur schwach in Erscheinung treten. Tatsächlich beobachten wir in obiger Testrechnung die „Explosion“ erst bei $h \sim 0.158$.

Gegenbeispiel zur „skalaren“ Stabilitätsanalyse

Der Vollständigkeit halber geben wir ein Beispiel an, welches zeigt, dass die der auf Variablenseparation beruhenden numerischen Stabilitätsanalyse zugrundeliegende Hypothese auch *falsch* sein kann. Für Parameter $\mu < 0$, $\varepsilon > 0\,, \alpha \in \mathbb{R}$ betrachte man das System

$$u'(t) = \tilde{A}(t)u(t)\,, \quad u(0) = u^0\,, \tag{3.1.22}$$

$$\tilde{A} = \varepsilon^{-1}\, U^\star(t) A U(t)\,, \quad A = \begin{bmatrix} -1 & \mu \\ 0 & -1 \end{bmatrix}\,, \quad U(t) = \begin{bmatrix} \cos\alpha t & \sin\alpha t \\ -\sin\alpha t & \cos\alpha t \end{bmatrix}\,.$$

Die zeitabhängige Matrix $U(t)$ ist unitär, $U(t)U^\star(t) = I$ (Drehung um den Winkel $-\alpha t$ im $\mathbb{R}^2$). Der Vollständigkeit halber wollen wir die Matrix $\tilde{A}(t)$ ausrechnen:

$$\begin{aligned} \tilde{A}(t) = U^\star(t) A U(t) &= \frac{1}{\varepsilon} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \cdot \begin{bmatrix} -1 & \mu \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \\ &= \frac{1}{\varepsilon} \left(-I + \mu \begin{bmatrix} -\sin\alpha t\, \cos\alpha t & \cos^2\alpha t \\ -\sin^2\alpha t & \sin\alpha t\, \cos\alpha t \end{bmatrix} \right)\,. \end{aligned}$$

Anwendung der A-stabilen Trapezregel auf (3.1.22) ergibt

$$y_n = [I - \tfrac{1}{2}h\,\tilde{A}(t_n)]^{-1}[I + \tfrac{1}{2}h\,\tilde{A}(t_{n-1})]y_{n-1}\,. \tag{3.1.23}$$

Für transformierte Variable $v(t) = U(t)u(t)$ gilt

$$\begin{aligned} v'(t) &= U'(t)u(t) + U(t)u'(t) \\ &= [U'(t)U^\star(t) + \varepsilon^{-1}U(t)U^\star(t)A]U(t)u(t) = [U'(t)U^\star(t) + \varepsilon^{-1}A]v(t) \end{aligned}$$

und somit

$$v'(t) = B\,v(t)\,, \qquad B = \begin{bmatrix} -1/\varepsilon & \mu/\varepsilon + \alpha \\ -\alpha & -1/\varepsilon \end{bmatrix}. \tag{3.1.24}$$

Das System (3.1.24) hat die allgemeine Lösung

$$v(t) = c_1\, e^{\lambda_1 t} + c_2\, e^{\lambda_2 t} \tag{3.1.25}$$

mit den folgenden Eigenwerten und zugehörigen Eigenvektoren der Matrix B:

$$\lambda_{1,2} = -\varepsilon^{-1} \pm \sqrt{-\alpha(\alpha + \mu\varepsilon^{-1})}, \quad c_{1,2} = \begin{bmatrix} \pm\sqrt{\alpha + \mu/\varepsilon} \\ \sqrt{-\alpha} \end{bmatrix}.$$

Einschub: Setze $\alpha = -3\,, \quad \mu = 3\,, \quad \varepsilon = \frac{1}{3}$. Dann wird

$$\lambda_1 = -3 + \sqrt{3(9-3)} = 3(\sqrt{2} - 1) > 1.2\,,$$

d. h.: Die Lösung $v(t)$ von (3.1.24) wächst für $t \to \infty$ wie $e^{1.2t}$, obwohl die Eigenwerte des „aquivalenten“ Systems (3.1.22) alle negativ sind.

Wir schreiben (3.1.23) in der Form

$$y_n - y_{n-1} = \tfrac{1}{2}h\,[\tilde{A}(t_n)y_n + \tilde{A}(t_{n-1})y_{n-1}]\,. \tag{3.1.26}$$

Für die transformierte diskrete Variable $z_n = U(t_n)y_n$ gilt dann wegen

$$\begin{aligned} [I - \tfrac{1}{2}\varepsilon^{-1}hU^\star(t_{n+1})AU(t_{n+1}]y_{n+1} &= [I + \tfrac{1}{2}\varepsilon^{-1}hU^\star(t_n)AU(t_n)]y_n \\ U^\star(t_{n+1})[I - \tfrac{1}{2}\varepsilon^{-1}hA]U(t_{n+1})y_{n+1} &= U^\star(t_n)[I + \tfrac{1}{2}\varepsilon^{-1}hA]U(t_n)y_n \end{aligned}$$

auch

$$z_{n+1} = [I - \tfrac{1}{2}\varepsilon^{-1}hA]^{-1}U(t_{n+1})U^\star(t_n)[I + \tfrac{1}{2}\varepsilon^{-1}hA]z_n$$

bzw.

$$z_{n+1} = [I - \tfrac{1}{2}\varepsilon^{-1}hA]^{-1}U(h)[I + \tfrac{1}{2}\varepsilon^{-1}hA]z_n = M\,z_n. \tag{3.1.27}$$

Hierbei wurde berücksichtigt, dass (geometrisches Argument)

$$U(t_{n+1})U^\star(t_n) = \begin{bmatrix} \cos\alpha t_{n+1} & \sin\alpha t_{n+1} \\ -\sin\alpha t_{n+1} & \cos\alpha t_{n+1} \end{bmatrix} \cdot \begin{bmatrix} \cos\alpha t_n & -\sin\alpha t_n \\ \sin\alpha t_n & \cos\alpha t_n \end{bmatrix}$$

$$= \begin{bmatrix} \cos\alpha t_{n+1}\,\cos\alpha t_n + \sin\alpha t_{n+1}\,\sin\alpha t_n & -\cos\alpha t_{n+1}\,\sin\alpha t_n + \sin\alpha t_{n+1}\,\cos\alpha t_n \\ -\sin\alpha t_{n+1}\,\cos\alpha t_n + \cos\alpha t_{n+1}\,\sin\alpha t_n & \sin\alpha t_{n+1}\,\sin\alpha t_n + \cos\alpha t_{n+1}\,\cos\alpha t_n \end{bmatrix}$$

$$= \begin{bmatrix} \cos\alpha h & \sin\alpha h \\ -\sin\alpha h & \cos\alpha h \end{bmatrix} = U(h).$$

Beachte, dass M auch unabhängig von t ist. Da $U(t)$ unitär ist, $\|U(t)x\| = \|x\|$, gilt für die Lösung y_n von (3.1.23) für $t_0 = 0$:

$$\|y_n\| = \|U^\star(t_n)z_n\| = \|U^\star(t_n)M^nU(t_0)y_0\| = \|U^\star(t_n)M^nU(0)y_0\| = \|M^ny_0\|\,.$$

Wir berechnen nun die Eigenwerte von M, um entscheiden zu können, ob die diskreten Lösungen y_n anwachsen oder abnehmen. Setze $\tau = -\alpha h\,, \quad \alpha < 0\,, \quad \varepsilon = -\frac{1}{4}\alpha h^2$. Dann erhalten wir durch Taylor-Entwicklung

$$\begin{aligned}
M &= [\tfrac{1}{2}\tau I - A]^{-1} \cdot \begin{bmatrix} \cos(-\tau) & \sin(-\tau) \\ -\sin(-\tau) & \cos(-\tau) \end{bmatrix} \cdot [\tfrac{1}{2}\tau I + A] \\
&= [\tfrac{1}{2}\tau I - A]^{-1}\,[I + \tau J + \mathcal{O}(\tau^2)][\tfrac{1}{2}\tau I + A]\,, \quad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \\
&= [\tfrac{1}{2}\tau I - A]^{-1}[\tfrac{1}{2}\tau I + A] - \tau A^{-1}JA + \mathcal{O}(\tau^2) \\
&= [\tfrac{1}{2}\tau A^{-1} - I]^{-1}[\tfrac{1}{2}\tau A^{-1} + I] - \tau A^{-1}JA + \mathcal{O}(\tau^2) \\
&= -I - \tau A^{-1}[I + JA] + \mathcal{O}(\tau^2)\,.
\end{aligned}$$

Nun ist

$$A^{-1} = \begin{bmatrix} -1 & -\mu \\ 0 & -1 \end{bmatrix}\,, \qquad I + JA = \begin{bmatrix} 1 & 1 \\ -1 & 1+\mu \end{bmatrix}$$

und somit

$$M = -I - \tau \begin{bmatrix} -1+\mu & -(1+\mu+\mu^2) \\ 1 & -1-\mu \end{bmatrix} + \mathcal{O}(\tau^2)\,.$$

Die Eigenwerte von M sind näherungsweise mit den Wurzeln $\lambda_{1,2}$ des (gestörten) charakteristischen Polynoms

$$\chi(\lambda) = \lambda^2 + 2\lambda + \mu + 2$$

gegeben als

$$\mu_{1,2} = -1 - \tau\lambda_{1,2} + \mathcal{O}(\tau^2)\,,$$

wobei

$$\lambda_{1,2} = -1 \pm \sqrt{-1-\mu}\,.$$

Die Wurzel $\lambda_1 = -1 + \sqrt{-1-\mu}$ ist positiv, wenn $\mu < -2$. In diesem Fall wird (für hinreichend kleines $\tau = -\alpha h$)

$$|\mu_1| > 1\,.$$

Die A-stabile Trapezregel erzeugt dann exponentiell anwachsende Näherungen $y_n = U^\star(t_n)z_n$ zu der exponentiell abfallenden Lösung $u(t_n)$.

Dieses Beispiel zeigt, dass zur Behandlung „nicht diagonalisierbarer“ Systeme die numerische Stabilitätstheorie der skalaren Gleichungen *nicht* ausreicht.

3.1.1 Steife Probleme

Die numerischen Stabilitätseigenschaften einer Differenzenformel sind von essentieller Bedeutung für die Integration sog. „steifer“ Probleme.

Definition 3.5 (Steifheit): *Eine AWA heißt „steif“ (entlang einer Lösung $u(t)$), wenn für die Eigenwerte $\lambda(t)$ der Jacobi-Matrix $f'_x(t,u(t))$ gilt:*

$$\kappa(t) := \frac{\max_{\mathrm{Re}\,\lambda(t)<0} |\,\mathrm{Re}\,\lambda(t)|}{\min_{\mathrm{Re}\,\lambda(t)<0} |\,\mathrm{Re}\,\lambda(t)|} \gg 1\,. \tag{3.1.28}$$

Die Größe $\kappa(t)$ wird „Steifigkeitsrate“ genannt.

Bemerkung 3.3: Die Realteile der Eigenwerte der Jacobi-Matrix $f'_x(t,u(t))$ stehen in enger Beziehung zur Lipschitz-Konstante L_f von $f(t\cdot)$:

$$\|f(t,x)-f(t,y)\| \le \max_{(t,\xi)\in D} \|f'_x(t,\xi)\|\|x-y\| \le L_f\|x-y\|,$$

$$|\,\mathrm{Re}\,\lambda_{\max}| \le |\lambda_{\max}| \le \|f'_x(t,u(t))\|.$$

Es ist zu beachten, dass bei der Bestimmung der Steifigkeitsrate nur die Eigenwerte mit *negativem* Realteil berücksichtigt werden. Diejenigen mit positivem Realteil gehören zu exponentiell wachsenden Lösungskomponenten und bedingen auf jeden Fall eine entsprechende Schrittweitenrestriktion. Steife Probleme zeichnen sich demnach durch Lösungskomponenten mit stark unterschiedlichem Abklingverhalten aus. Es ist aber nicht gerechtfertigt, eine skalare AWA als „steif“ zu bezeichnen, nur weil ihre Lipschitz-Konstante L sehr groß ist. Denn in diesem Fall müßte ja des Diskretisierungsfehlers wegen sowieso mit einer entsprechend reduzierten Schrittweite gerechnet werden.

Beispiel 3.2:

$$u'(t) = Au(t)\,, \quad u(0) = (1,0,-1)^T, \qquad A = \begin{bmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{bmatrix},$$

Die Eigenwerte von A sind $\lambda_i = -2\,, \lambda_{2,3} = -40 \pm 40i$. Die Lösung des Systems ist

$$\begin{aligned} u_1(t) &= \tfrac{1}{2}e^{-2t} + \tfrac{1}{2}e^{-40t}\,[\cos 40t + \sin 40t] \\ u_2(t) &= \tfrac{1}{2}e^{-2t} - \tfrac{1}{2}e^{-40t}\,[\cos 40t + \sin 40t] \\ u_3(t) &= -e^{-40t}\,[\cos 40t - \sin 40t]. \end{aligned}$$

Im Bereich $0 \le t \le 0.1$ variieren alle drei Lösungskomponenten schnell, so dass die Notwendigkeit einer kleineren Schrittweite $h \ll 0.1$ plausibel ist. Für $t > 0.1$ sind dagegen $u_1 \sim u_2$ nahezu identisch und variieren sehr langsam, während $u_3 \sim 0$ ist. Dies

Verhalten legt die Wahl einer gröberen Schrittweite $h \geq 0.1$ in diesem Bereich nahe. Für die explizite Euler-Methode erzwingt jedoch die Stabilitätsbedingung $|1 + 40h| < 1$ die globale Schrittweite $h < 0.025$. Tatsächlich erhalten wir bei Verwendung von $h = 0.04$ eine oszillierende Approximation von u_1 („•“ im Bild).

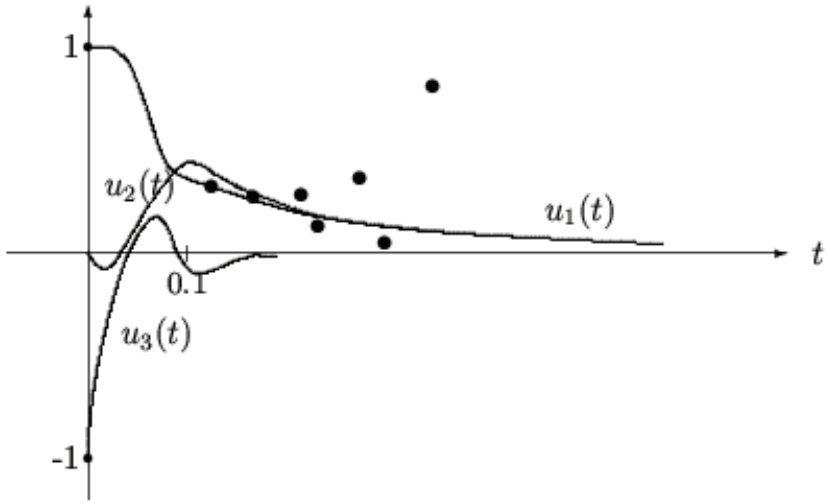

Abbildung 3.3: *Lösungskomponenten einer „steifen“ AWA und instabile numerische Approximation „•“.*

Beispiel 3.3: Bei örtlicher Diskretisierung der (1-dim.) Wärmeleitungsgleichung

$$\frac{\partial v}{\partial t}(x,t) \;=\; \frac{\partial^2 v}{\partial x^2}(x,t)\,, \qquad \begin{array}{l} v(0,t) \;=\; v(1,t) = 0 \\ v(x,0) \;=\; v^0(x) \end{array}$$

mittels des zentralen Differenzenquotionten zweiter Ordnung

$$\frac{\partial^2 v}{\partial x^2}(x,t) \sim \frac{1}{\Delta x^2}\,[\,v(x+\Delta x,t) - 2v(x,t) + v(x-\Delta x,t)\,]$$

entsteht ein System von $d = \frac{1}{\Delta x} - 1$ gewöhnlichen Differentialgleichungen in den Unbekannten $u_i(t) \sim v(x_i,t)$:

$$u_i'(t) = \frac{1}{\Delta x^2}\,[\,u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)\,]\,, \quad i = 1,\ldots,d \quad (u_0 = u_{d+1} = 0)\,.$$

Die zugehörige Koeffizientenmatrix

$$A = \frac{1}{\Delta x^2} \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & -2 & 1 \\ 0 & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{d\times d}$$

hat die Eigenwerte

$$\lambda_j = -\Big[\frac{\sin(j\pi\Delta x/2)}{\Delta x/2}\Big]^2, \quad j = 1,\dots,d, \qquad \lambda_{\max} \sim -\frac{4}{\Delta x^2}, \quad \lambda_{\min} \sim -\pi^2 .$$

Das System ist also umso steifer, je feiner die Ortsvariable diskretisiert wird. Für das explizite Euler-Verfahren erzwingt die Stabilitätsbedingung dann die Schrittweitenrelation

$$h < \tfrac{1}{2}\,\Delta x^2 .$$

3.1.2 Implizite Verfahren

Zur Integration eines steifen Systems mit nicht bekannter Steifigkeitsrate werden Differenzenformeln mit möglichst guten numerischen Stabilitätseigenschaften benötigt, d. h. möglichst A-stabile Methoden. Da *explizite* Formeln nicht A-stabil sein können, werden zur Integration steifer Systeme fast ausschließlich *implizite* Methoden verwendet. Das allgemeine implizite Einschrittverfahren hat die Gestalt

$$y_n = y_{n-1} + h_n F(h_n; t_{n-1}, y_{n-1}, y_n), \quad n \geq 1 . \tag{3.1.29}$$

Von großer praktischer Bedeutung sind die sog. „impliziten Runge-Kutta-Verfahren“

$$y_n = y_{n-1} + \sum_{r=1}^{R} c_r k_r(h_n; t_{n-1}, y_{n-1}), \quad n \geq 1, \tag{3.1.30}$$

$$k_r(h_n; t_{n-1}, y_{n-1}) = f(t_{n-1} + h_n a_r, y_{n-1} + h_n \sum_{s=1}^{R} b_{rs} k_s(h_n; t_{n-1}, y_{n-1})), \quad r = 1,\dots,R.$$

Diese Formeln sind trotz ihrer scheinbar expliziten Form natürlich implizit, da die k_r als Lösungen eines i. Allg. nichtlinearen Gleichungssystems bestimmt sind. Der einfachste Vertreter für $R = 1$ ist die „implizite Euler-Methode“

$$y_n = y_{n-1} + h_n f(t_n, y_n), \quad n \geq 1. \tag{3.1.31}$$

Für das Testproblem (3.1.3) ergibt sie

$$y_n = (1 - \lambda h)^{-n} y_0 ,$$

mit dem Verstärkungsfaktor $\omega = (1 - \lambda h)^{-1}$. Das Stabilitätsgebiet ist also das Komplement der offenen Kreisscheibe $\{z \in \mathbb{C} : |1 - z| < 1\}$, d. h.: Die implizite Euler-Methode

ist A-stabil. Für $Re\,\lambda > 0$ ist sie allerdings auch für $|1-\lambda h| \geq 1$ absolut-stabil; sie kann also Beschränktheit der Lösung $u(t)$ vorgaukeln, auch wenn $u(t)$ exponentiell wächst. In letzterem Fall (für $\mathrm{Im}\,\lambda = 0$) ist jedoch $(1-\lambda h)^{-1} < 0$, d. h.: Die Näherungswerte y_n haben oszillierende Vorzeichen für $n \to \infty$, was immer ein Zeichen für irgendwelche numerische Instabilität ist.

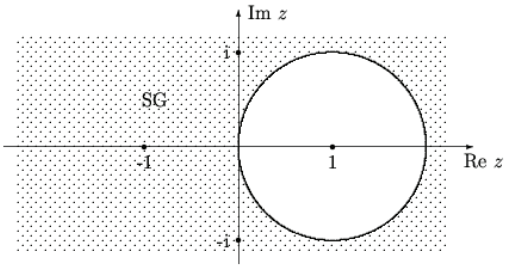

Abbildung 3.4: *Stabilitätsgebiet der impliziten Euler-Methode.*

Aufgrund der größeren Anzahl von freien Parametern der *impliziten* Runge-Kutta-Formeln lässt sich bei gegebener Stufenzahl R eine höhere Ordnung erzielen als mit *expliziten* Formeln dieser Art. Insbesondere lassen sich implizite Runge-Kutta-Formeln beliebig hoher Ordnung konstruieren, die gleichzeitig noch A-stabil sind.

Beispiel 3.4: Die 2-stufige Formel

$$\begin{aligned} y_n &= y_{n-1} + \tfrac{1}{2}h\{k_1 + k_2\}, \\ k_1 &= f\big(t_{n-1} + (\tfrac{1}{2} + \tfrac{\sqrt{3}}{6})h,\, y_{n-1} + \tfrac{1}{4}hk_1 + (\tfrac{1}{4} + \tfrac{\sqrt{3}}{6})hk_2\big)\big), \\ k_2 &= f\big(t_{n-1} + (\tfrac{1}{2} - \tfrac{\sqrt{3}}{6})h,\, y_{n-1} + (\tfrac{1}{4} - \tfrac{\sqrt{3}}{6})hk_1 + \tfrac{1}{4}hk_2\big), \end{aligned}$$

hat die Ordnung $m = 4$. Ihr Verstärkungsfaktor ist

$$\omega = \frac{1 + \frac{1}{2}\bar{h} + \frac{1}{12}\bar{h}^2}{1 - \frac{1}{2}\bar{h} + \frac{1}{12}\bar{h}^2}, \quad \bar{h} = h\lambda,$$

und ihr Stabilitätsintervall $SI = (-\infty, 0]$.

Ein Nachteil dieser „optimalen“, impliziten Runge-Kutte-Formeln ist, dass bei ihrer Anwendung in jedem Zeitschritt Gleichungssysteme der Dimension Rd gelöst werden müssen. Um dies zu vermeiden, verwendet man in der Regel sog. „diagonal-implizite“ Runge-Kutta-Verfahren (sog. „DIRK“-Formeln), welche zwar eine etwas geringere Ordnung haben, aber wegen ihrer speziellen Struktur nur die Lösung von Sytemen der Dimension d erfordern. Dies wird dadurch erreicht, dass die in der Darstellung (3.1.30) die Koeffizienten $b_{rs} = 0$, $s > r$, gewählt werden.

Beispiel 3.5: Die allgemeine 3-stufige, *diagonal-implizite* Runge-Kutta-Formel

$$\begin{aligned} y_n &= y_{n-1} + h\{c_1k_1 + c_2k_2 + c_3k_3\}, \\ k_1 &= f(t_{n-1}, y_{n-1})\,, \quad k_2 = f(t_{n-1} + a_2h,\, y_{n-1} + hb_{21}k_1 + hb_{22}k_2), \\ k_3 &= f(t_{n-1} + a_3h,\, y_{n-1} + hb_{31}k_1 + hb_{32}k_2 + hb_{33}k_3), \end{aligned}$$

hat die maximale Ordnung $m = 3$. Ihr Stabilitätsintervall ist ebenfalls $SI = (-\infty, 0]$.

3.2 Lösung monotoner Probleme: Newton-Verfahren

Das Hauptproblem bei der Anwendung impliziter Differenzenverfahren ist die Lösung der auftretenden, i. Allg. nichtlinearen Gleichungssysteme. In den einzelnen Zeitschritten hat man bei den allgemeinen R-stufigen impliziten Runge-Kutta-Methoden Gleichungssysteme der Dimension Rd zu lösen, bei den diagonal-impliziten Runge-Kutta-Formeln Systeme der Dimension d. Bei großen Systemen, $d \gg 1$, ist dies ein beträchtlicher Aufwand, der nur im Fall hochgradiger Steifheit des Problems gerechtfertigt ist.

Die Fragen nach der Existenz der diskreten Näherungen y_n und ihrer tatsächlichen Berechnung wollen wir exemplarisch anhand der impliziten Euler-Methode behandeln. Der Schritt von t_{n-1} nach t_n erfordert hier die Lösung der Fixpunktgleichung

$$y = G(y) := y_{n-1} + h_n f(t_n, y)\,. \tag{3.2.32}$$

Die Abbildung $G : \mathbb{R}^d \to \mathbb{R}^d$ ist unter der Bedingung

$$h_n L =: q < 1 \tag{3.2.33}$$

mit der Lipschitz-Konstante L von $f(t, \cdot)$ eine Kontraktion:

$$\|G(y) - G(y')\| \le h_n \|f(t_n, y) - f(t_n, y')\| \le h_n L \|y - y'\|$$

Nach dem Banachschen Fixpunktsatz existiert dann genau ein Fixpunkt $y = y_n$ von G, der mit Hilfe der „sukzessiven Approximation“

$$y^{(k+1)} = G(y^{(k)})\,, \quad k = 0, 1, 2, \ldots, \tag{3.2.34}$$

berechnet werden kann. Deren Konvergenz ist aber leider nur garantiert, wenn die Schrittweitenbedingung (3.2.33) erfüllt ist. Bei einem steifen Problem mit $L \gg 1$ ist diese Forderung aber meist zu restriktiv. In diesem Fall benötigt man für den Nachweis der Existenz der Approximationen y_n zusätzliche Struktureigenschaften der AWA. Wir diskutieren hier nur den einfachsten Fall einer „semi-monotonen“ Nichtlinearität.

Satz 3.1 (Monotone steife AWA): *Die rechte Seite $f(t, \cdot)$ der AWA sei L-stetig mit Konstante L und semi-monoton,*

$$-(f(t, x) - f(t, y)\,,\, x - y) \ge 0\,, \quad (t, x)\,,\, (t, y) \in I \times \mathbb{R}^d. \tag{3.2.35}$$

Dann existieren für beliebig gewählte Schrittweiten h_n stets die Approximationen y_n. Ferner konvergiert für jedes hinreichend kleine θ die Folge der Iterierten

$$y^{(k)} = y^{(k-1)} - \theta\big\{y^{(k-1)} - hf(t_n, y^{(k-1)}) - y_{n-1}\big\}, \quad y^{(0)} := y_{n-1}, \tag{3.2.36}$$

gegen diese Lösung y_n. Die Konvergenz ist am schnellsten für $\theta = (1 + h^2L^2)^{-1}$, wobei die a priori Fehlerabschätzung gilt:

$$\|y^{(k)} - y_n\| \le \Big(1 - \frac{1}{1 + h^2L^2}\Big)^{k/2} \|y^{(1)} - y^{(0)}\|, \quad k \ge 1. \tag{3.2.37}$$

Beweis: (i) Wir haben zu zeigen, dass für jedes feste $h > 0$ stets ein eindeutig bestimmtes $y_n \in \mathbb{R}^d$ existiert, so dass

$$y_n - hf(t_n, y_n) \;=\; y_{n-1}\,. \tag{3.2.38}$$

Die Semimonotonie der Funktion $f(t, \cdot)$ impliziert, dass die Abbildung

$$g : \mathbb{R}^d \to \mathbb{R}^d, \quad g(x) := x - hf(t_n, x)$$

strikt monoton ist mit der Monotoniekonstante $\gamma = 1$. Gemäß Korollar 1.7 existiert daher eine eindeutig bestimmte Lösung $y_n \in \mathbb{R}^d$ der Gleichung $g(y_n) = y_{n-1}$. Der Beweis dieses Resultats verwendete die Tatsache, dass die Fixpunktabbildung

$$G_\theta(x) := x - \theta\big(g(x) - y_{n-1}\big)$$

für $0 < \theta < 2(1 + h^2L^2)^{-1}$ wegen

$$\begin{aligned}
\|G_\theta(x) - G_\theta(y)\|^2 &= \|x - \theta g(x) - y_{n-1} - y + \theta g(y) + y_{n-1}\|^2 \\
&= \|(1-\theta)(x-y) + \theta h(f(t_n, x) - f(t_n, y))\|^2 \\
&= (1-\theta)^2\|x - y\|^2 + 2(1-\theta)\theta h \underbrace{(x - y, f(t_n, x) - f(t_n, y))}_{\le 0} \\
&\qquad + \theta^2h^2\|g(x) - g(y)\|^2 \\
&\le \{(1-\theta)^2 + \theta^2h^2L^2)\}\|x - y\|^2
\end{aligned}$$

eine Kontraktion ist. Deren Lipschitz-Konstante wird minimal für $\theta := (1 + h^2L^2)^{-1}$:

$$q = \big\{(1-\theta)^2 + \theta^2h^2L^2)\big\}^{1/2} = \Big(1 - \frac{1}{1+h^2L^2}\Big)^{1/2} < 1.$$

In diesem Fall konvergiert dann die Fixpunktiteration

$$y^{(k+1)} = G_\theta(y^{(k)}) = y^{(k)} - \theta\big(y^{(k)} - hf(t_n, y^{(k)}) - y_{n-1}\big)$$

gegen die Lösung von (3.2.38), wobei bekanntlich die behauptete a priori Fehlerabschätzung gilt. Q.E.D.

Das implizite Euler-Verfahrens lässt sich für „steife“ AWAn mit semi-monotoner rechter Seite also im Prinzip für beliebige Schrittweite h_n durchführen. Allerdings konvergiert in diesem Fall wegen $h_n L \gg 1$ die einfache Fixpunktiteration (3.2.36) nur sehr langsam.

Wir betrachten daher in diesem Fall als Alternative das Newton-Verfahren zur Lösung der Gleichung (3.2.32) in Form einer „Nullstellengleichung“:

$$g(y_n) := y_n - h_n f(t_n, y_n) - y_{n-1} = 0. \tag{3.2.39}$$

Dieses hat die Gestalt

$$g'(y^{(k)})y^{(k+1)} = g'(y^{(k)})y^{(k)} - g(y^{(k)})\,, \tag{3.2.40}$$

mit der Newton-Matrix

$$g'(y^{(k)}) := I - h_n f'_x(t_n, y^{(k)}).$$

In der Praxis wird das Newton-Verfahren aber in Form einer „Defektkorrekturiteration“ durchgeführt:

$$g'(y^{(k)})\delta y^{(k)} = -g(y^{(k)}), \quad y^{(k+1)} = y^{(k)} + \delta y^{(k)}. \tag{3.2.41}$$

Wenn $g'(y_n)$ regulär ist und der Startwert $y^{(0)}$ hinreichend nahe bei y_n liegt, konvergieren (unter weiteren Bedingungen an f) die Newton-Iterierten $y^{(k)} \to y_n$ $(k \to \infty)$ quadratisch:

$$\|y^{(k)} - y_n\| \le cq^{(2^k)}, \quad k \ge 1, \tag{3.2.42}$$

mit gewissen Konstanten $c > 0$ und $q \in (0,1)$. Wir wollen die Abhängigkeit dieser Konvergenz von der Lipschitz-Konstante L genauer untersuchen. Dazu benötigen wir Ergebnisse aus der Theorie des Newton-Verfahrens, welche im Folgenden in einem allgemeineren Rahmen entwickelt werden.

Sei $g : \mathbb{R}^d \to \mathbb{R}^d$ eine differenzierbare Abbildung, für die eine Nullstelle x^* gesucht ist. Die Jacobi-Matrix $g'(\cdot)$ sei auf der Niveaumenge

$$D := D_{z^*} := \left\{x \in \mathbb{R}^d \middle|\, \|g(x)\| \le \|g(z^*)\|\right\}$$

zu einem (beliebigen) festen Punkt $z^* \in \mathbb{R}^d$ regulär mit gleichmäßig beschränkter Inverser:

$$\|g'(x)^{-1}\| \le \beta, \quad x \in D.$$

Ferner sei $g'(\cdot)$ auf D gleichmäßig L-stetig:

$$\|g'(x) - g'(y)\| \le \gamma\|x - y\|, \quad x, y \in D.$$

Mit diesen Bezeichnungen haben wir den folgenden Satz.

Satz 3.2 (Newton-Kantorovich): *Unter den vorausgehenden Voraussetzungen sei für den Startpunkt $x^{(0)} \in D$ mit $\alpha := \|g'(x^{(0)})^{-1}g(x^{(0)})\|$ die folgende Bedingung erfüllt:*

$$q := \tfrac{1}{2}\alpha\beta\gamma < 1. \tag{3.2.43}$$

Dann erzeugt die Newton-Iteration

$$x^{(k+1)} = x^{(k)} - g'(x^{(k)})^{-1} g(x^{(k)}), \quad k \geq 0,$$

eine Folge $(x^{(k)})_{k\in\mathbb{N}} \subset D$ *, welche quadratisch gegen eine Nullstelle* $x^* \in D$ *von* g *konvergiert, wobei die folgende a priori Fehlerabschätzung gilt:*

$$\|x^{(k)} - x^*\| \leq \frac{\alpha}{1 - q^{(2^k)}} q^{(2^k - 1)}, \quad k \geq 1. \tag{3.2.44}$$

Beweis: Zum Startpunkt $x^{(0)} \in D$ gehört die abgeschlossene, nicht leere Niveaumenge

$$D_0 := \left\{ x \in \mathbb{R}^d \,|\, \|g(x)\| \leq \|g(x^{(0)})\| \right\} \subset D.$$

Wir betrachten die stetige Abbildung $G : D_0 \to \mathbb{R}^d$,

$$G(x) := x - g'(x)^{-1} g(x),$$

welche gerade einen Newton-Schritt ausgehend vom Punkt x beschreibt.

(i) Wir wollen zunächst einige Hilfsresultate ableiten. Für $x \in D_0$ sei

$$x_r := x - r g'(x)^{-1} g(x), \;\; r \geq 0.$$

Für die Vektorfunktion $h(r) := g(x_r)$ gilt

$$h'(r) = -g'(x_r) g'(x)^{-1} g(x), \qquad h'(0) = -h(0).$$

Sei $R := \max\{r \in [0,1] \,|\, x_s \in D_0,\ 0 \leq s \leq r\}$. Für $0 \leq r \leq R$ ist dann

$$\begin{aligned} \|g(x_r)\| - (1-r)\|g(x)\| &\leq \|g(x_r) - (1-r)g(x)\| = \|h(r) - (1-r)h(0)\| \\ &= \Big\| \int_0^r h'(s)\, ds + r h(0) \Big\| = \Big\| \int_0^r \{h'(s) - h'(0)\}\, ds \Big\| \\ &\leq \int_0^r \|h'(s) - h'(0)\|\, ds, \end{aligned}$$

und ferner wegen $x_s - x = -s g'(x)^{-1} g(x)$:

$$\begin{aligned} \|h'(s) - h'(0)\| &= \|\{g'(x_s) - g'(x)\} g'(x)^{-1} g(x)\| \\ &\leq \gamma \|x_s - x\| \|g'(x)^{-1} g(x)\| \leq \gamma s \|g'(x)^{-1} g(x)\|^2. \end{aligned}$$

Dies ergibt

$$\|g(x_r)\| - (1-r)\|g(x)\| \leq \tfrac{1}{2} r^2 \gamma \|g'(x)^{-1} g(x)\|^2 \leq \tfrac{1}{2} r^2 \gamma \beta \|g'(x)^{-1} g(x)\| \|g(x)\|. \tag{3.2.45}$$

Mit der Größe $\alpha_x := \|g'(x)^{-1} g(x)\|$ folgt

$$\|g(x_r)\| \leq (1 - r + \tfrac{1}{2} r^2 \gamma \beta \alpha_x) \|g(x)\|. \tag{3.2.46}$$

(ii) Im Falle $\alpha_x \le \alpha$ gilt dann wegen der Voraussetzung $\frac{1}{2}\alpha\beta\gamma < 1$:

$$\|g(x_r)\| \le (1 - r + r^2)\|g(x)\|.$$

Folglich ist in diesem Fall $R = 1$ und somit $G(x) \in D_0$, d. h.: Der Newton-Schritt bringt uns nicht aus der Menge D_0 heraus. Für solche $x \in D_0$ gilt weiter

$$\begin{aligned}\|G(x) - G^2(x)\| &= \|G(x) - G(x) + g'(G(x))^{-1}g(G(x))\| \\ &\le \|g'(G(x))^{-1}\|\|g(G(x))\| \le \beta\|g(G(x))\|.\end{aligned}$$

Mit Hilfe der Abschätzung (3.2.45) für $r = 1$ folgt weiter bei Beachtung von $G(x) = x_1$:

$$\|G(x) - G^2(x)\| \le \tfrac{1}{2}\beta\gamma\|g'(x)^{-1}g(x)\|^2 = \tfrac{1}{2}\beta\gamma\|x - G(x)\|^2, \tag{3.2.47}$$

sowie

$$\|g'(G(x))^{-1}g(G(x))\| = \|G(x) - G^2(x)\| \le \tfrac{1}{2}\beta\gamma\|g'(x)^{-1}g(x)\|^2 = \tfrac{1}{2}\beta\gamma\alpha_x^2. \tag{3.2.48}$$

Für $\alpha_x \le \alpha$ überträgt sich diese Eigenschaft also auch auf $G(x)$, d. h.: $\alpha_{G(x)} \le \alpha$.

(iii) Nach diesen Vorbereitungen kommen wir nun zum Beweis des Satzes. Aus den Vorbetrachtungen ergibt sich, dass für den Startwert $x^{(0)} \in D_0$ mit $\alpha := \|g'(x^{(0)})^{-1}g(x^{(0)})\|$ alle Iterierten des Newton-Verfahrens ebenfalls in D_0 liegen und $\alpha_k := \|g'(x^{(k)})^{-1}g(x^{(k)}\| \le \alpha$ erfüllen. Hiermit erhalten wir

$$\begin{aligned}\|x^{(k+1)} - x^{(k)}\| &= \|G^2(x^{(k-1)}) - G(x^{(k-1)})\| \\ &\le \tfrac{1}{2}\beta\gamma\|G(x^{(k-1)}) - x^{(k-1)}\|^2 = \tfrac{1}{2}\beta\gamma\|x^{(k)} - x^{(k-1)}\|^2,\end{aligned}$$

und bei Iteration dieser Abschätzung:

$$\begin{aligned}\|x^{(k+1)} - x^{(k)}\| &\le \tfrac{1}{2}\beta\gamma\big(\tfrac{1}{2}\beta\gamma\|x^{(k-1)} - x^{(k-2)}\|^2\big)^2 \\ &\le (\tfrac{1}{2}\beta\gamma)^{(2^2-1)}\|x^{(k-1)} - x^{(k-2)}\|^{(2^2)} \\ &\le (\tfrac{1}{2}\beta\gamma)^{(2^2-1)}\big(\tfrac{1}{2}\beta\gamma\|x^{(k-2)} - x^{(k-3)}\|^2\big)^{(2^2)} = (\tfrac{1}{2}\beta\gamma)^{(2^3-1)}\|x^{(k-2)} - x^{(k-3)}\|^{(2^3)}.\end{aligned}$$

Fortsetzung der Iteration bis $k = 0$ ergibt mit $q = \frac{1}{2}\alpha\beta\gamma$:

$$\|x^{(k+1)} - x^{(k)}\| \le (\tfrac{1}{2}\beta\gamma)^{(2^k-1)}\|x^{(1)} - x^{(0)}\|^{(2^k)} \le (\tfrac{1}{2}\beta\gamma)^{(2^k-1)}\alpha^{(2^k)} \le \alpha q^{(2^k-1)}.$$

Für beliebiges $m \in \mathbb{N}$ folgt damit wegen $q < 1$:

$$\begin{aligned}\|x^{(k+m)} - x^{(k)}\| &\le \|x^{(k+m)} - x^{(k+m-1)}\| + \cdots + \|x^{(k+2)} - x^{(k+1)}\| + \|x^{(k+1)} - x^{(k)}\| \\ &\le \alpha q^{(2^{k+m-1}-1)} + \cdots + \alpha q^{(2^{k+1}-1)} + \alpha q^{(2^k-1)} \\ &\le \alpha q^{(2^k-1)}\big\{(q^{(2^k)})^{(2^{m-2})} + \cdots + q^{(2^k)} + 1\big\} \\ &\le \alpha q^{(2^k-1)}\sum_{j=0}^{\infty}(q^{(2^k)})^j \;\le\; \alpha q^{(2^k-1)}\frac{1}{1 - q^{(2^k)}}.\end{aligned}$$

Dies besagt, dass $(x^{(k)})_{k\in\mathbb{N}} \subset D$ eine Cauchy-Folge ist. Deren Limes $x^* \in D$ ist dann notwendig ein Fixpunkt von G bzw. Nullstelle von g:

$$x^* = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} G(x^{(k-1)}) = G(x^*).$$

Durch Grenzübergang $m \to \infty$ erhalten wir auch die Fehlerabschätzung (3.2.44). Q.E.D.

In der obigen Situation des impliziten Euler-Schrittes ist $g(x) := x - hf(t_n, x) - y_{n-1}$ und damit

$$g'(x) = I - hf_x(t_n, x).$$

Aufgrund der angenommenen Semi-Monotonie von $f(t,\cdot)$ gilt

$$\begin{aligned}(g'(x)y, y) &= \|y\|^2 - h(f_x(t_n, x)y, y)\\ &= \|y\|^2 - h\lim_{\varepsilon\to 0} \varepsilon^{-1}(f(t_n, x+\varepsilon y) - f(t_n, x), y) \geq \|y\|^2.\end{aligned} \tag{3.2.49}$$

Daher ist $g'(x)$ regulär, und es folgt unter Verwendung von (3.2.49):

$$\begin{aligned}\|g'(x)^{-1}\|^2 &= \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{\|g'(x)^{-1}y\|^2}{\|y\|^2} \leq \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{(g'(x)g'(x)^{-1}y, g'(x)^{-1}y)}{\|y\|^2}\\ &\leq \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{\|g'(x)^{-1}y\|}{\|y\|} = \|g'(x)^{-1}\|,\end{aligned}$$

bzw. $\|g'(x)^{-1}\| \leq 1$. In diesem Fall haben wir also stets

$$\beta := \sup_{x\in\mathbb{R}^d} \|g'(x)^{-1}\| \leq 1, \quad \alpha \leq \|g(x^{(0)})\| =: \alpha'.$$

Allerdings ist nach wie vor

$$\|g'(x) - g'(y)\| = h\|f_x(t_n, x) - f_x(t_n, y)\| \leq hL'\|x - y\|$$

mit der Lipschitz-Konstante L' von $f_x(t_n,\cdot)$. Dies führt zu folgendem Resultat.

Korollar 3.1 (Newton-Verfahren): *Unter den vorausgehenden Voraussetzungen sei für den Startpunkt $y^{(0)} \in \mathbb{R}^d$ mit $\alpha' := \|y^{(0)} - hf(t_n, y^{(0)}) - y_{n-1}\|$ die folgende Bedingung erfüllt:*

$$q := \tfrac{1}{2}\alpha' hL' < 1. \tag{3.2.50}$$

Dann erzeugt die Newton-Iteration (3.2.40) eine Folge $(y^{(k)})_{k\in\mathbb{N}}$, welche quadratisch gegen y_n konvergiert und es gilt die Fehlerabschätzung

$$\|y^{(k)} - y_n\| \leq \frac{\alpha'}{1 - q^{(2^k)}} q^{(2^k-1)}, \quad k \geq 1. \tag{3.2.51}$$

Wir sehen, dass das Newton-Verfahren im Fall einer steifen, semi-momotonen AWA zwar für alle Schrittweiten h_n quadratisch konvergiert, aber der Einzugsbereich der Konvergenz proportional zu $(hL')^{-1}$ schrumpft. Da für eine steife AWA im allg. $L \gg 1$ sowie $L' \gg 1$ ist, wird damit das Konvergenzproblem praktisch nur verschoben. Im nächsten Schritt wollen wir versuchen, das Newton-Verfahren zu „globalisieren“, d. h. den Einzugsbereich der Konvergenz auf ganz $\mathbb{R}^d$ zu erweitern. Dazu gehen wir wieder in den oben definierten abstrakten Rahmen zurück und betrachten zur Lösung der Gleichung $g(x) = 0$ mit der Abbildung $g : \mathbb{R}^d \to \mathbb{R}^d$ das sog. „gedämpfte“ Newton-Verfahren

$$x^{(k+1)} = x^{(k)} - \lambda_k g'(x^{(k)})^{-1} g(x^{(k)}), \quad k \geq 1, \tag{3.2.52}$$

mit Parametern $\lambda_{k-1} \in (0, 1]$. Dafür haben wir folgendes Resultat.

Satz 3.3 (gedämpftes Newton-Verfahren): *Unter den vorausgehenden Voraussetzungen erzeugt für jeden Startpunkt $x^{(0)} \in D$ die gedämpfte Newton-Iteration (3.2.52) mit*

$$\lambda_k := \min\Big\{1, \frac{1}{\alpha_k \beta \gamma}\Big\}, \qquad \alpha_k := \|g'(x^{(k)})^{-1} g(x^{(k)})\|,$$

eine Folge $(x^{(k)})_{k\in\mathbb{N}}$, für welche nach k_* Schritten $q_* := \frac{1}{2}\alpha_{k_*}\beta\gamma < 1$ erfüllt ist, so dass ab dann $x^{(k)}$ quadratisch konvergiert.

Beweis: Wir verwenden wieder die Bezeichnungen aus dem Beweis von Satz 3.2. Für eine Newton-Iterierte $x^{(k)} \in D_0$ gilt mit $\alpha_k := \|g'(x^{(k)})^{-1} g(x^{(k)})\| \leq \alpha$ die Abschätzung

$$\|g(x_r^{(k)})\| \leq (1 - r + \tfrac{1}{2} r^2 \alpha_k \beta\gamma) \|g(x^{(k)})\|, \quad 0 \leq r \leq 1.$$

Man beachte, dass $x_1^{(k)} = x^{(k+1)}$. Für $\frac{1}{2}\alpha_k\beta\gamma < 1$ ist die Hauptvoraussetzung von Satz 3.2 erfüllt, d. h.: Die Folge $(x^{(l)})_{l\geq k}$ konvergiert quadratisch gegen eine Nullstelle von g. Sei nun angenommen, dass $\frac{1}{2}\alpha_k\beta\gamma \geq 1$. Dann wird der Vorfaktor in obiger Abschätzung minimal für

$$r_* = \frac{1}{\alpha_k\beta\gamma} > 0 : \quad 1 - r_* + \tfrac{1}{2} r_*^2 \alpha_k\beta\gamma \;\leq\; 1 - \frac{1}{2\alpha_k\beta\gamma} < 1.$$

Bei Wahl von $r_k := (\alpha_k\beta\gamma)^{-1}$ ist also $(x^{(k)})_{k\in\mathbb{N}} \subset D_0$, und die Norm $\|g(x^{(k)})\|$ fällt streng monoton:

$$\|g(x^{(k+1)})\| \leq \Big(1 - \frac{1}{2\alpha_k\beta\gamma}\Big) \|g(x^{(k)})\|.$$

Nach endlich vielen, $k_* \geq 1$, Iterationsschritten ist dann $\frac{1}{2}\alpha_{k_*}\beta\gamma < 1$, und die quadratische Konvergenz der weiteren Folge $(x^{(k)})_{k\geq k_*}$ folgt wieder aus Satz 3.2. Q.E.D.

Aus Satz 3.3 erhalten wir das folgende Korollar für die vorliegende, spezielle Situation.

Korollar 3.2 (Gedämpftes Newton-Verfahren): *Unter den vorausgehenden Voraussetzungen erzeugt für jeden Startpunkt* $y^{(0)} \in \mathbb{R}^d$ *die gedämpfte Newton-Iteration*

$$y^{(k+1)} = y^{(k)} - \lambda_k g'(y^{(k)})^{-1} g(y^{(k)}), \quad k \geq 1, \tag{3.2.53}$$

mit

$$\lambda_k := \min\Big\{1, \frac{1}{\alpha'_k hL'}\Big\}, \qquad \alpha'_k := \|y^{(k)} - hf(t_n, y^{(k)}) - y_{n-1}\|$$

eine Folge $(y^{(k)})_{k\in\mathbb{N}}$*, für welche nach* $k_* \approx |\log(hL')| hL'$ *Schritten die Bedingung* $q_* := \frac{1}{2}\alpha'_{k_*} hL' < 1$ *erfüllt ist, so dass ab dann* $y^{(k)}$ *quadratisch gegen* y_n *konvergiert.*

Beweis: Aus dem Beweis von Satz 3.3 entnehmen wir die Abschätzung

$$\|g(y^{(k+1)})\| \leq \Big(1 - \frac{1}{2\alpha_k \beta\gamma}\Big)\|g(y^{(k)})\|.$$

Im vorliegenden Fall gilt wegen $\beta = \|g'(x)^{-1}\| \leq 1$:

$$\alpha_k = \|g'(x^{(k)})^{-1} g(x^{(k)})\| \leq \|g(x^{(k)})\| =: \alpha'_k \leq \alpha'_0.$$

Mit $\gamma = hL'$ erhalten wir also

$$1 - \frac{1}{2\alpha_k\beta\gamma} \;\leq\; 1 - \frac{1}{2\alpha'_0 hL'} < 1.$$

Die Bedingung $q := \frac{1}{2}\alpha_k\beta\gamma \leq \frac{1}{2}\alpha'_0\beta\gamma < 1$ ist dann erfüllt für

$$\frac{1}{2}\Big(1 - \frac{1}{2\alpha'_0 hL'}\Big)^k \alpha'_0 hL' < 1 \quad \Leftrightarrow \quad \Big(1 - \frac{1}{2\alpha'_0 hL'}\Big)^k < \frac{2}{\alpha'_0 hL'},$$

bzw.

$$k > \frac{|\log(4\sigma)|}{\log(1-\sigma)} \approx |\log(hL')|\, hL', \qquad \sigma := \frac{1}{2\alpha'_0 hL'},$$

wobei $hL' \gg 1$ bzw. $\sigma \ll 1$ angenommen wird. Q.E.D.

Die bisher für das implizite Euler-Verfahren abgeleiteten Resultate basieren im wesentlichen auf der Semi-Monotonie der Funktion $f(t,\cdot)$. Sie lassen sich direkt auf andere, implizite Verfahren übertragen, wenn entsprechend die jeweilige Verfahrensfunktion $F(h;t,z,\cdot)$ semi-monoton ist. Dies ist automatisch der Fall z.B. für die Trapezregel und für die später betrachteten „Rückwärtsdifferenzenformeln“ (implizite, lineare Mehrschrittmethoden). I. Allg., nicht-monotonen Fall kommt man um restriktive Bedingungen an die Qualität der Startpunkte $y^{(0)}$ bzw. an die Schrittweiten h_n nicht herum.

3.3 Übungsaufgaben

Aufgabe 3.1: Man gebe die Stabilitätsintervalle der folgenden Einschrittformeln an:

$$\begin{aligned} &a) \qquad y_{n+1} = y_n + \tfrac{1}{2}h\big\{f(t_{n+1}, y_{n+1}) + f(t_n, y_n)\big\}, \\ &b) \qquad y_{n+1} = y_n + hf\big(t_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hf(t_n, y_n)\big), \\ &c) \qquad y_{n+1} = y_n + \tfrac{1}{6}h\big\{2f(t_{n+1}, y_{n+1}) + 4f(t_n, y_n) + hf^{(1)}(t_n, y_n)\big\}, \end{aligned}$$

wobei $f^{(1)}(t,x) = f'_t(t,x) + f(t,x)f'_x(t,x)$.

Aufgabe 3.2: Aus einer skalaren Differentialgleichung 2-ter Ordnung

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

mit einer differenzierbaren Funktion $f(t,x,y)$ gewinnt man durch Einführung der Hilfsfunktionen $u_1 := u, u_2 := u'$ ein System von Gleichungen 1-ter Ordnung. Man zeige, dass die Jacobi-Matrix dessen rechter Seite im Falle $\partial_x f \geq 0$ nur reelle Eigenwerte hat. Welche Konsequenzen hat dies für die Approximierbarkeit dieses Problems mit Differenzenformeln?

Aufgabe 3.3: Sei $f(\cdot) : D \subset \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ eine analytische, d. h. durch eine konvergente Potenzreihe darstellbare, Matrixfunktion:

$$f(A) = \sum_{i=0}^{\infty} a_i A^i.$$

a) Für die Exponentialfunktion $f(A) = e^{-A}$, die Sinusfunktion $f(A) = \sin(A)$ und die Inversenfunktion $f(A) = (I - A)^{-1}$ gebe man die jeweiligen Potenzreihen und deren Konvergenzradien an.

b) Man zeige, dass mit jeder regulären Matrix $Q \in \mathbb{R}^{d\times d}$ gilt:

$$Qf(A)Q^{-1} = f(QAQ^{-1}).$$

Wenn die Argumentation für eine allgemeine, analytische Funktion $f(\cdot)$ zu schwierig erscheint, beschränke man sich auf den Fall einer rationalen Funktion.

Aufgabe 3.4 (Praktische Aufgabe): Man löse die 3-dimensionale, steife AWA

$$u'(t) = Au(t), \quad t \geq 0, \quad u(0) = (1, 0, -1)^T,$$

mit der Systemmatrix

$$A = \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{pmatrix}$$

und der Lösung

$$u_1(t) = \frac{1}{2}e^{-2t} + \frac{1}{2}e^{-40t}\{\cos(40t) + \sin(40t)\},$$
$$u_2(t) = \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-40t}\{\cos(40t) + \sin(40t)\},$$
$$u_3(t) = -e^{-40t}\{\cos(40t) - \sin(40t)\}$$

mit Hilfe

a) des klassischen (expliziten) Runge-Kutta-Verfahrens 4. Ordnung,

b) der (impliziten) Trapezregel 2. Ordnung (mit „direkter“ Gleichungssystemlösung durch Gauß-Elimination)

Zu berechnen ist der Vektor $u(2) \in \mathbb{R}^3$ auf 10 Dezimalstellen. Man versuche, in beiden Fällen möglichst sparsam zu arbeiten. Mit welchem Verfahren lässt sich diese Aufgabe (mit äquidistanter Schrittweite) am effizientesten, d. h. in geringster Zeit, lösen?

Aufgabe 3.5: Jede der im Text betrachteten Einschrittmethoden nimmt angewendet auf ein lineares (autonomes) System $u'(t) = Au(t)$ die Form $y_n = g(hA)y_{n-1}$ an, mit einer rationalen Funktion $g(\cdot)$, d. h. es gibt Polynome q, p, so dass

$$g(hA) = q(hA)^{-1}p(hA).$$

a) Man zeige, dass im Fall einer symmetrischen Matrix A die Darstellung

$$g(hA) = Qg(hD)Q^T$$

gilt, mit einer *orthogonalen* Matrix Q und $D = \mathrm{diag}(\lambda_i)$, $\lambda_1, \cdots, \lambda_n$ Eigenwerte von A. (Hinweis: Symmetrische Matrizen besitzen ein Orthonormalsystem von Eigenvektoren.)

b) Für den Fall, dass die Matrix A symmetrisch ist, zeige man weiterhin bzgl. der euklidischen Norm die Abschätzung

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

mit den Eigenwerten λ_i von A.

c) Man bestimme mit Hilfe von (b) die maximale Schrittweite h, für die das „klassische“ 4-stufige Runge-Kutta-Verfahren das System

$$u'(t) = -10u(t) + 9v(t), \quad v'(t) = 9u(t) - 10v(t)$$

noch numerisch stabil integriert.

Aufgabe 3.6: Man beweise, dass die Trapezregel

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

A-stabil ist, d. h.: Ihr Stabilitätsgebiet enthält die negative komplexe Halbebene. Genauer gilt sogar

$$SG = \{z \in \mathbb{C} \,|\, \mathrm{Re} z \leq 0\}.$$

(Hinweis: Die Beziehung $SI = \{z \in \mathbb{R} \,|\, z \leq 0\}$ ist evident. Die stärkere Aussage für das ganze Stabilitätsgebiet SG kann man durch direkte Rechnung ableiten.)

Aufgabe 3.7: Man zeige, dass die semi-implizite Runge-Kutta-Formel 2-ter Ordnung

$$y_n = y_{n-1} + \tfrac{1}{2}h\big\{k_1 + k_2\big\}, \quad k_1 = f(t_{n-1}, y_{n-1}), \;\; k_2 = f(t_n, y_{n-1} + \tfrac{1}{2}hk_1 + \tfrac{1}{2}hk_2),$$

A-stabil ist. Man vergleiche den Rechenaufwand (pro Zeitschritt) für diese Methode mit dem für die gleichfalls A-stabile Trapezregel, wenn zur Auflösung der impliziten Gleichungen das Newton-Verfahren verwendet wird.

Aufgabe 3.8 (Praktische Aufgabe): a) Man berechne eine Näherungslösung für die AWA

$$\begin{aligned} u'(t) &= -50u(t) + 49v(t), \quad u(0) = 1, \\ v'(t) &= 49u(t) - 50v(t), \quad v(0) = 1, \end{aligned}$$

mit Hilfe der Trapezregel

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

sowie der modifizierten Euler-Formel

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

für die (konstanten) Schrittweiten $h = 2^{-i}, i = 1, \ldots, 8$. Man vergleiche die berechneten Werte zum Zeitpunkt $t = 3$ mit dem Wert $u(3)$ der exakten Lösung. Dazu berechne man entweder die exakte Lösung analytisch oder erzeuge einen sehr genauen Referenzwert durch Rechnung mit der feinen Gitterweite $h = 2^{-10}$.

b) Man berechne die Lösung mit einem relativen Fehler kleiner als 10^{-3} mit Hilfe einer geeignet erscheinenden Methode aus dem Text (mit äquidistanter Schrittweite). Dabei soll der numerische Aufwand (Zahl der Auswertungen der Funktion f) möglichst gering sein.

Aufgabe 3.9: Für zweimal stetig differenzierbare Abbildungen $g : D \subset \mathbb{R}^d \to \mathbb{R}^d$ mit invertierbarer Jacobi-Matrix $g'(\cdot)$ konvergiert das Newton-Verfahren lokal quadratisch gegen eine Nullstelle x^*. Man zeige, dass es für (nur) stetig differenzierbare Abbildungen immer noch „super-linear“ konvergiert,

$$\frac{\|x^k - x^*\|}{\|x^{k-1} - x^*\|} \to 0 \quad (k \to \infty);$$

es ist also i. Allg. asymptotisch schneller als die einfache Fixpunktiteration. Zur Vereinfachung nehme man an, dass g auf ganz $\mathbb{R}^d$ definiert ist und dort die geforderten Eigenschaften besitzt. Ferner darf die Existenz einer Nullstelle x^* von $g(\cdot)$ angenommen werden.

Aufgabe 3.10: Die im Text entwickelte Theorie des Newton-Verfahrens basiert auf der Annahme der Semi-Monotonie der rechten Seite $f(t,\cdot)$ in der Differentialgleichung bzw. der Verfahrensfunktion $F(h;t,x,\cdot)$ bzgl. des „impliziten“ Arguments:

$$-\big(f(t,y_1)-f(t,y_2),y_1-y_2\big)\geq 0,\quad y_1,y_2\in\mathbb{R}^d.$$

a) Man untersuche die Anwendbarkeit dieser Resultate zur Lösung der impliziten Gleichungssysteme bei Verwendung des semi-impliziten Runge-Kutta-Verfahrens

$$y_n=y_{n-1}+\tfrac{1}{2}h\{k_1+k_2\},\quad k_1=f(t_{n-1},y_{n-1}),\ \ k_2=f(t_n,y_{n-1}+\tfrac{1}{2}hk_1+\tfrac{1}{2}hk_2)$$

bei angenommener Semi-Monotonie und zweimaligen Differenzierbarkeit von $f(t,\cdot)$.

b) Wie sieht es im Fall eines diagonal-impliziten Runge-Kutta-Verfahrens höherer Ordnung aus?

c) Ist zu erwarten, dass solch ein Ergebnis auch für voll implizite Runge-Kutta-Verfahren gültig bleibt, d. h. Semi-Monotonie des zu lösenden *Rd*-dimensionalen, nichtlinearen Gleichungssystems?

Aufgabe 3.11: Für die Newton-Iteration zur Lösung der nichtlinearen Gleichungen bei der Durchführung des impliziten Euler-Verfahrens ist im Text die Schrittweitenstrategie

$$\lambda_k=\min\Big(1,\frac{1}{\alpha_k hL}\Big),\quad \alpha_k:=\|g'(y^{(k)})^{-1}g(y^{(k)})\|,$$

entwickelt worden. Man entwickle unter analogen Voraussetzungen wie im Text eine entsprechende Strategie für die Newton-Iteration bei dem semi-impliziten Runge-Kutta-Verfahren aus Aufgabe 7.2:

$$y_n=y_{n-1}+\frac{1}{2}h\big\{k_1+k_2\big\},\quad k_1=f(t_{n-1},y_{n-1}),\ \ k_2=f(t_n,y_{n-1}+\tfrac{1}{2}hk_1+\tfrac{1}{2}hk_2).$$

Aufgabe 3.12 (Praktische Aufgabe): Man approximiere die (globale) Lösung der 2-dimensionalen AWA

$$\begin{aligned}u_1'(t)&=\sin(u_1(t))\sin(u_2(t)),\ \ t\geq 0,\quad u_1(0)=3,\\ u_2'(t)&=\sin(u_1(t))\sin(u_2(t)),\ \ t\geq 0,\quad u_2(0)=4,\end{aligned}$$

mit Hilfe der Trapezregel

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

mit äquidistanten Schrittweiten $h=2^{-i}$, $i=4,\ldots,10$. Die in jedem Zeitschritt auftretenden nichtlinearen Gleichungssysteme werden mit dem Newton-Verfahren (ohne Dämpfung) gelöst. Die Lösung konvergiert für $t\to\infty$ gegen einen konstanten Vektor; dessen Wert soll bestimmt werden.