

NUMERISCHE MATHEMATIK FÜR
MATHEMATIKER III¹
(Numerik gewöhnlicher Differentialgleichungen)

Prof. Dr. Hans Babovsky

Institut für Mathematik

Technische Universität Ilmenau

WS 2003/04

¹Korrekturen, Kommentare und Verbesserungsvorschläge bitte an:

babovsky@mathematik.tu-ilmenau.de

Inhaltsverzeichnis

1	Gewöhnliche Differentialgleichungen: Einführung	2
1.1	Grundbegriffe	2
1.2	Das Eulersche Polygonzugverfahren	4
1.3	Erste Fehlerbetrachtungen	9
2	Einschrittverfahren (ESV)	11
2.1	Das Konzept der Einschrittverfahren	11
2.2	Runge-Kutta-Verfahren (RKV)	14
2.3	Konvergenz von Einschrittverfahren	18
3	Lineare Mehrschrittverfahren (MSV)	22
3.1	Das Konzept der Mehrschrittverfahren	22
3.2	Die Ordnung linearer MSV	29
3.3	Homogene lineare Differenzgleichungen	32
3.4	Konsistenz und Konvergenz von MSV	34
4	Stabilität von ESV und MSV	38
4.1	Absolute Stabilität	38
4.2	Integration steifer Systeme	44
5	Numerik von Anfangswertproblemen: Ergänzungen	47
5.1	Schrittweitensteuerung	47
6	Randwertprobleme	50
6.1	Einführung	50
6.2	Das einfache Schießverfahren	52
6.3	Differenzenverfahren	63
7	Differentiell-algebraische Systeme	67
7.1	Einführung	67
7.2	Runge-Kutta-Verfahren für Index-1-Systeme	69
7.3	Systeme mit Index > 1	71

1 Gewöhnliche Differentialgleichungen: Einführung

1.1 Grundbegriffe

Die Vorlesung befasst sich mit der numerischen Lösung (explizit definierter) gewöhnlicher Differentialgleichungen. Diese sind wie folgt definiert.

[1.1] **Definition:** (a) Gegeben seien ein Gebiet $\Omega \subseteq \mathbb{R}^2$ und eine (hinreichend glatte) Funktion $f : \Omega \rightarrow \mathbb{R}$. Eine Gleichung der Form

$$x'(t) = f(t, x(t)) \quad (1.1)$$

(oder kurz: $x' = f(t, x)$) heißt **gewöhnliche Differentialgleichung 1. Ordnung** (kurz: **gDGL**).

(b) Gegeben seien ein Gebiet $\Omega \subseteq \mathbb{R}^{n+1}$ und eine Funktion $f : \Omega \rightarrow \mathbb{R}$. Eine Gleichung der Form

$$x^{(n)} = f(t, x, x', \dots, x^{(n-1)}) \quad (1.2)$$

heißt **gewöhnliche Differentialgleichung n -ter Ordnung**.

(c) Im Folgenden repräsentiere ein fett gedruckter Buchstabe einen Vektor der Länge n (n fest) (z.B. $\mathbf{x} = (x_1, \dots, x_n)^T$). Gegeben seien ein Gebiet $\Omega \subseteq \mathbb{R}^{n+1}$ und eine vektorwertige Funktion $\mathbf{f} : \Omega \rightarrow \mathbb{R}^n$. Das Gleichungssystem

$$\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t)) \quad (1.3)$$

heißt **System gewöhnlicher Differentialgleichungen** erster Ordnung (kurz: **SysgDGL**).

(d) Ist f (bzw. \mathbf{f}) von t unabhängig, so heißt die zugehörige Gleichung (bzw. das System) **autonom**.

[1.2] **Bemerkungen:** (a) Eine Funktion $x(t)$ (bzw. $\mathbf{x}(t)$) auf einem Intervall $[a, b]$ ist Lösung der gDGL (des SysgDGL), falls für jedes $t \in [a, b]$ die entsprechende Gleichung aus (1.1) \dots (1.3) erfüllt ist.

(b) gDGL's höherer Ordnung können stets als SysgDGL geschrieben werden, indem die Gleichung (1.2) umformuliert wird. Hierzu definieren wir

$$\mathbf{z}(t) = (z_1, \dots, z_n)^T := (x(t), x'(t), \dots, x^{(n-1)}(t))^T \quad .$$

Die Lösung der Gleichung (1.2) ist äquivalent zur Lösung des Systems

$$\mathbf{z}'(t) = \begin{pmatrix} z_2(t) \\ \vdots \\ z_n(t) \\ f(t, \mathbf{z}(t)) \end{pmatrix} =: \tilde{\mathbf{f}}(t, \mathbf{z}) \quad .$$

(b) Nicht autonome gDGL's (und entsprechend SysgDGL's) können stets als autonome Systeme formuliert werden. Hierzu definieren wir zu der skalaren Funktion $x(t)$ die erweiterte vektorwertige Funktion $\tilde{\mathbf{x}}(t) := (t, x(t))^T =: (\tilde{x}_1, \tilde{x}_2)^T$. Die Lösung von (1.1) ist äquivalent zur Lösung des SysgDGL

$$\tilde{\mathbf{x}}' = \tilde{\mathbf{f}}(\tilde{\mathbf{x}}) \quad \text{mit} \quad \tilde{\mathbf{f}}(\tilde{\mathbf{x}}) = (1, f(x_1, x_2))^T \quad . \quad (1.4)$$

Ist $x(t)$ eine Lösung von (1.1) und ist für ein t_0 $x(t_0) = c$, so kann aus (1.1) die Steigung $x'(t_0)$ bestimmt werden. Es ist $x'(t_0) = f(t_0, c)$. Die Tangente an die Lösung im Punkt $(t_0, x(t_0))^T$ ist damit gegeben durch den Vektor $(1, f(t, c))^T$. Dies motiviert die folgende Definition.

[1.3] Definition: Das zur gDGL $x' = f(t, x)$ definierte Vektorfeld

$$R(t, x) := \frac{1}{\sqrt{1 + f^2(t, x)}} \begin{pmatrix} 1 \\ f(t, x) \end{pmatrix}$$

heißt das **Richtungsfeld** der gDGL.

Da gDGL's Spezialfälle von SysgDGL's sind, wollen wir uns bei den folgenden Überlegungen auf letztere beschränken.

Eine Lösung $\mathbf{x}(t)$ von (1.3) ist (unter gewissen Bedingungen an \mathbf{f}) *eindeutig*, wenn sie an einem Punkt t_0 vorgegeben ist. Dies führt auf die folgende Definition.

[1.4] Definition: Unter einem **Anfangswertproblem (AWP)** verstehen wir ein SysgDGL der Form (1.3), welches ergänzt ist durch eine Bedingung der Form

$$\mathbf{x}(t_0) = \mathbf{c} \quad \text{für ein} \quad (t_0, \mathbf{c})^T \in \Omega \quad . \quad (1.5)$$

in einem Punkt t_0 .

Der folgende Satz ist ein zentrales Ergebnis der klassischen Theorie der gewöhnlichen Differentialgleichungen, welches besagt, dass Lösungen von AWP's eindeutig sind, sofern die rechten Seiten Lipschitz-stetig sind. Für einen Beweis des Satzes verweisen wir auf die Standard-Lehrbücher.

[1.5] **Satz (Existenz und Eindeutigkeit):** Die rechte Seite \mathbf{f} des SysgDGl (1.3) sei in folgendem Sinne *Lipschitz-stetig*: Es gibt eine Konstante $L > 0$ derart, dass für alle $t \in \mathbb{R}$ und alle $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ gilt: Ist $(t, \mathbf{x}_1)^T, (t, \mathbf{x}_2)^T \in \Omega$, so gilt die Abschätzung

$$\|f(t, x_1) - f(t, x_2)\| \leq L \cdot \|x_1 - x_2\| \quad . \quad (1.6)$$

(Hierbei ist $\|\cdot\|$ eine beliebige Vektornorm in \mathbb{R}^n .) Dann hat das AWP (1.3), (1.5) eine eindeutige Lösung $x(t)$. Diese lässt sich bis zum Rand von Ω fortsetzen.

Im Folgenden sei, um die Eindeutigkeit des AWP zu gewährleisten, die Lipschitz-Stetigkeit (1.6) immer vorausgesetzt. Das SysgDGl (1.3) erzeugt eine ganze Schar von Lösungen, welche **der zu (1.3) gehörige Fluss** genannt wird und welche wir mit dem Symbol Φ bezeichnen wollen. Hierbei beschreibt

$$\mathbf{x}(t) := \Phi^{t,\tau} \tilde{\mathbf{x}} \quad (1.7)$$

diejenige spezielle Lösung (*Trajektorie*), welche zum Zeitpunkt τ den Wert \tilde{x} annimmt:

$$\mathbf{x}(\tau) = \Phi^{\tau,\tau} \tilde{\mathbf{x}}.$$

Insbesondere kann die Lösung des AWP (1.3), (1.4) in Kurzform

$$\mathbf{x}(t) = \Phi^{t,t_0} \mathbf{c}$$

geschrieben werden.

1.2 Das Eulersche Polygonzugverfahren

$$\mathbf{x}(t) := \Phi^{t,\tau} \mathbf{x}_\tau$$

bezeichnet diejenige Lösung des SysgDGI, welche zu einem fest vorgegebenen Zeitpunkt τ durch einen festen Punkt \mathbf{x}_τ geht. Ist \mathbf{f} in einer Umgebung von (τ, \mathbf{x}_τ) stetig differenzierbar, so ist $\mathbf{x}(t)$ in einer Umgebung $(\tau - h, \tau + h)$ zweimal stetig differenzierbar, und es gilt

$$\mathbf{x}''(t) = \frac{d}{dt} \mathbf{f}(t, \mathbf{x}(t)) = \mathbf{f}_t(t, \mathbf{x}(t)) + \mathbf{f}_\mathbf{x}(t, \mathbf{x}(t)) \cdot \mathbf{x}'(t) = \mathbf{f}_t(t, \mathbf{x}(t)) + \mathbf{f}_\mathbf{x}(t, \mathbf{x}(t)) \cdot \mathbf{f}(t, \mathbf{x}(t)) \quad .$$

Nach der Taylorformel lässt sich $\mathbf{x}(t)$ schreiben in der Form

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(\tau) + (t - \tau) \cdot \mathbf{f}(\tau, \mathbf{x}(\tau)) + \frac{(t - \tau)^2}{2} \cdot \mathbf{x}(\tau + \theta \cdot (t - \tau)) \\ &= \mathbf{x}(\tau) + (t - \tau) \cdot \mathbf{f}(\tau, \mathbf{x}(\tau)) + \mathcal{O}(h^2) \end{aligned} \quad (1.8)$$

(für ein $\theta \in (0, 1)$). Diese Idee können wir benutzen zur Konstruktion einer ersten Näherung der Lösung des AWP (1.3), (1.4) für $t \geq t_0$. Hierzu definieren wir einen kleinen Zeitschritt Δt , sowie

$$t_i := t_0 + i \cdot \Delta t, \quad i = 1, 2, 3, \dots$$

und approximieren die Lösung durch die folgende stetige, stückweise lineare Funktion:

Initialisierung: $\eta_0 = \mathbf{x}_\Delta(t_0) := \mathbf{c};$

Iteration: Ist $\eta_i := \mathbf{x}_\Delta(t_i)$ gegeben, so definiere $\mathbf{x}_\Delta(t)$ in $[t_i, t_{i+1}]$ durch $\mathbf{x}_\Delta(t) := \eta_i + (t - t_i) \cdot \mathbf{f}(t_i, \eta_i)$; insbesondere ist $\eta_{i+1} = \eta_i + \Delta t \cdot \mathbf{f}(t_i, \eta_i)$.

Sind wir nur an Näherungen an den Knoten t_i interessiert, so ergibt sich hieraus der folgende

[1.6] Algorithmus (Eulersches Polygonzugverfahren): Wähle einen kleinen Zeitschritt Δt und definiere die zugehörigen Knoten $t_i := t_0 + i \cdot \Delta t$. Die Werte η_i an den Knoten t_i sind rekursiv definiert durch

$$\eta_0 = \mathbf{c}, \quad (1.9)$$

$$\eta_{i+1} = \eta_i + \Delta t \cdot \mathbf{f}(t_i, \eta_i). \quad (1.10)$$

[1.7] **Beispiel** (lineare homogene gDGl): Die exakte Lösung des AWP

$$x' = \lambda \cdot x, \quad x(0) = 1$$

ist $x(t) = \exp(\lambda t)$. Sie soll mit dem Eulerschen Polygonzugverfahren im Intervall $[0, T]$ approximiert werden. Hierzu wählen wir eine Zahl $N \in \mathbb{N} \setminus \{0\}$ und den Zeitschritt $\Delta t = T/N$. Mit $f(t, x) = \lambda \cdot x$ liefert der Algorithmus [1.11]

$$\begin{aligned} \eta_0 &= 1, \\ \eta_{i+1} &= \eta_i + \Delta t \cdot \lambda \cdot \eta_i = (1 + \Delta t \cdot \lambda) \eta_i. \end{aligned}$$

Durch Induktion folgt

$$\eta_i = (1 + \Delta t \lambda)^i.$$

Bemerkungen: (a) Aus der Analysis bekannt ist die Formel

$$\left(1 + \frac{1}{r}\right)^r \longrightarrow e \quad \text{für } r \rightarrow \infty.$$

Damit gilt

$$\begin{aligned} \eta_N &= \left(1 + \lambda \frac{T}{N}\right)^N = \left(\left(1 + \frac{1}{N/(\lambda T)}\right)^{N/(\lambda T)}\right)^{\lambda T} \\ &\longrightarrow e^{\lambda T} = x(T) \quad \text{für } N \rightarrow \infty. \end{aligned}$$

η_N konvergiert damit für große N gegen den korrekten Wert $x(T)$.

(b) Ist $\lambda < 0$, so gilt $x(t) = \exp(\lambda t) \rightarrow 0$. Um dieses qualitative Verhalten mit dem Eulerschen Polygonzugverfahren zu erhalten, darf der Zeitschritt Δt nicht zu groß gewählt werden. Ist nämlich $\Delta t > 2/|\lambda|$, so ist $1 + \Delta t \cdot \lambda < -1$; in diesem Fall oszilliert die Folge η_i und es gilt

$$|\eta_i| = |1 + \Delta t \cdot \lambda|^i \longrightarrow \infty \quad \text{für } i \rightarrow \infty.$$

Als notwendiges Kriterium für die Konvergenz gegen 0 für $N \rightarrow \infty$ ergibt sich damit die Bedingung

$$\Delta t < \frac{2}{|\lambda|}. \tag{1.11}$$

t	numerisch		exakt	
0.1	0.9000	0.1000	0.9094	0.0906
0.2	0.8200	0.1800	0.8352	0.1648
0.3	0.7560	0.2440	0.7744	0.2256
0.4	0.7048	0.2952	0.7247	0.2753
0.5	0.6638	0.3362	0.6839	0.3161
1.0	0.5537	0.4463	0.5677	0.4323

Tabelle 1: Lösung von Beispiel [1.13], $\alpha = \beta = 1$

[1.8] Beispiel (Lineares homogenes SysgDGL): Gelöst werden soll das AWP für $\mathbf{x}(t) = (x_1(t), x_2(t))^T$,

$$\mathbf{x}' = \begin{pmatrix} -\alpha & \beta \\ \beta & -\alpha \end{pmatrix} \cdot \mathbf{x} =: A \cdot \mathbf{x}, \quad \mathbf{x}(0) := \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Hier ist

$$\mathbf{f}(t, \mathbf{x}) = A \cdot \mathbf{x},$$

und das Eulersche Polygonzugverfahren liefert mit zur gewählten Schrittweite Δt

$$\eta_i = (I + \Delta t \cdot A)^i \cdot \eta_0 = \begin{pmatrix} 1 - \alpha \cdot \Delta t & \beta \cdot \Delta t \\ \beta \cdot \Delta t & 1 - \alpha \cdot \Delta t \end{pmatrix}^i \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Führt man eine Testrechnung durch für $\alpha = \beta = 1$ mit einer Schrittweite $\Delta t = 0.1$, so erhält man die in Tabelle 1 aufgeführten numerischen Ergebnisse, welche zumindest den qualitativen Verlauf der exakten Lösung widerspiegeln. Testrechnungen für $\alpha = \beta = 20$ sind aus Tabelle 2 abzulesen. Es zeigt sich dass die Schrittweite $\Delta t = 0.1$ hier völlig unbrauchbar ist. Die numerische Lösung ist offenbar instabil. Wie lässt sich dies erklären?

Hierzu berechnen wir die Eigenwerte $\lambda_{1,2}$ von A . Diese ergeben sich als Lösungen von

$$\det(\lambda I - A) = \begin{vmatrix} \lambda + \alpha & -\beta \\ -\beta & \lambda + \alpha \end{vmatrix} = (\lambda + \alpha)^2 - \beta^2 = 0.$$

t	num. ($\Delta t = 0.1$)		num. ($\Delta t = 0.01$)		exakt	
0.1	-1	2	0.5030	0.4970	0.5092	0.4908
0.2	5	-4	0.5000	0.5000	0.5002	0.4998
0.3	-13	14	0.5000	0.5000	0.5000	0.5000
0.4	41	-40	0.5000	0.5000	0.5000	0.5000
0.5	-121	122	0.5000	0.5000	0.5000	0.5000
1.0	29525	-29524	0.5000	0.5000	0.5000	0.5000

Tabelle 2: Lösung von Beispiel [1.13], $\alpha = \beta = 20$

Im Fall $\alpha = \beta$ sind die Eigenwerte gleich $\lambda_1 = 0$ und $\lambda_2 = -2\alpha$. Die zugehörigen Eigenvektoren sind $(1, 1)^T$ und $(1, -1)^T$. Fügen wir die Eigenvektoren zu einer Matrix T zusammen:

$$T := \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

so ist

$$A = T \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot T^{-1}.$$

Transformieren wir nun das SysgDGl $\mathbf{x}' = A \cdot \mathbf{x}$ auf das entsprechende System für $\mathbf{z} = (z_1, z_2)^T := T^{-1} \cdot \mathbf{x}$, so erhalten wir

$$\mathbf{z}' = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot \mathbf{z},$$

oder komponentenweise geschrieben:

$$\begin{aligned} z_1' &= \lambda_1 z_1, \\ z_2' &= \lambda_2 z_2. \end{aligned}$$

Dies sind zwei (skalare) lineare homogene gDGl. Aus Beispiel [1.12] wissen wir, dass zur numerischen Lösung mit Hilfe des Eulerschen Polygonzugverfahrens eine Schrittweite gewählt werden muss, für die gilt (vgl. (1.19)) $\Delta t < 2/\lambda_i$. Im Fall $\alpha = \beta = 1$ folgt

als (nach diesem Kriterium) maximal zulässige Schrittweite $\Delta t < 1/\alpha = 1$. Im Fall $\alpha = \beta = 20$ ist dies $\Delta t < 1/20 = 0.05$.

Diese beiden Beispiele sollen die folgende allgemeine Merkregel motivieren.

[1.9] Merkregel: A sei eine $n \times n$ -Matrix mit den Eigenwerten $\lambda_1, \dots, \lambda_n$. λ_{max} sei der betragsgrößte negative Eigenwert. Zur numerischen Lösung des SysgDGI

$$\mathbf{x}' = A\mathbf{x}$$

mit Hilfe des Eulerschen Polygonzugverfahrens muss ein Zeitschritt

$$\Delta t \ll 1/|\lambda_{max}| \tag{1.12}$$

gewählt werden.

Lineare SysgDGI, bei denen Eigenwerte mit sehr unterschiedlichen negativen Realteilen auftreten, heißen auch **steife Differentialgleichungssysteme**. Hier erzwingt der betragsgrößte negative Realteil eine Obergrenze für die Wahl der Schrittweite.

1.3 Erste Fehlerbetrachtungen

Gegeben sei das AWP

$$x' = f(t, x), \quad x(t_0) = c.$$

Die exakte Lösung kann mit Hilfe des Flusses beschrieben werden in der Form

$$x(t) = \Phi^{t,t_0} c.$$

Zu gegebener Schrittweite Δt und den zugehörigen diskreten Zeiten $t_i = t_0 + i \cdot \Delta t$ liefert ein numerisches Verfahren (z.B. das Eulersche Polygonzugverfahren) Näherungen η_i zu den gesuchten Werten $x(t_i)$. Der hierbei entstehende Approximationsfehler kann in zwei Anteile aufgeteilt werden.

[1.10] Fehlerarten: (a) Ausgehend von einem Näherungswert η_i für $\Phi^{t_i,t_0} c$ wird η_{i+1}

konstruiert als Näherungswert der Trajektorie durch (t_i, η_i) zum nächsten Zeitpunkt t_{i+1} , also

$$\eta_{i+1} \approx \Phi^{t_{i+1}, t_i} \eta_i.$$

Die (skalierte) Abweichung

$$\epsilon(t_i, \eta_i, \Delta t) := \frac{1}{\Delta t} \cdot \left(\Phi^{t_i + \Delta t, t_i} \eta_i - \eta_{i+1} \right) \quad (1.13)$$

nennt man **lokalen Diskretisierungsfehler**. Zu seiner Abschätzung bedient man sich üblicherweise der Taylorentwicklung des Flusses $\Phi^{t, t_i} \eta_i$ um den Punkt (t_i, η_i) .

(b) Der Startpunkt η_i zur Berechnung von η_{i+1} ist (außer für $i = 0$) ebenfalls schon mit einem Fehler behaftet, so dass η_{i+1} mit $\Phi^{t_{i+1}, t_i} \eta_i$ anstelle von $\Phi^{t_{i+1}, t_i} x(t_i)$ die "falsche" Trajektorie approximiert. Mit dem lokalen Diskretisierungsfehler als "Keimzelle" addieren sich die in jedem Zeitschritt neu entstehenden Anteile zu einem **globalen Fehler** auf, sodass unter Umständen die Teiltrajektorien von der gesuchten Lösungstrajektorie immer weiter "wegdriften". Diesen globalen Fehler versucht man mit sog. Stabilitätsuntersuchungen in den Griff zu bekommen.

Wir untersuchen den lokalen Diskretisierungsfehler für das Eulersche Polygonzugverfahren. Hierzu nehmen wir an, dass $f(t, x)$ hinreichend glatt ist (hier: mindestens zweimal stetig differenzierbar bzgl. t und x).

Ist (t_i, η_i) gegeben, so bezeichne

$$z(t) := \Phi^{t, t_i} \eta_i$$

die Trajektorie, welche durch (t_i, η_i) verläuft. Damit ist $z(\cdot)$ Lösung des AWP

$$z' = f(t, z), \quad z(t_i) = \eta_i.$$

Wir werten nun die Taylorentwicklung

$$z(t) = z(t_i) + (t - t_i) \cdot z'(t_i) + \frac{(t - t_i)^2}{2!} \cdot z''(t_i) + \dots$$

von $z(t)$ um den Punkt t_i aus. Zunächst gilt

$$z'(t_i) = f(t_i, z(t_i)) = f(t_i, \eta_i).$$

Ferner liefert die Kettenregel

$$\begin{aligned} z''(t_i) &= \left. \frac{d}{dt} z'(t) \right|_{t=t_i} = \left. \frac{d}{dt} f(t, z(t)) \right|_{t=t_i} \\ &= f_t(t_i, z(t_i)) + z'(t_i) \cdot f_x(t_i, z(t_i)) = f_t(t_i, \eta_i) + f(t_i, \eta_i) \cdot f_x(t_i, \eta_i). \end{aligned}$$

(Hierbei bezeichnen f_t und f_x die partiellen Ableitungen von f bzgl. t und x .) Aus

$$z(t_{i+1}) = \eta_i + \Delta t \cdot f(t_i, \eta_i) + \frac{\Delta t^2}{2} \cdot (f_t(t_i, \eta_i) + f(t_i, \eta_i) \cdot f_x(t_i, \eta_i)) + \mathcal{O}(\Delta t^3) \quad (1.14)$$

und

$$\eta_{i+1} = \eta_i + \Delta t \cdot f(t_i, \eta_i)$$

ergibt sich hieraus der lokale Diskretisierungsfehler

$$\begin{aligned} \epsilon(t_i, \eta_i, \Delta t) &= \frac{1}{\Delta t} (z(t_{i+1}) - \eta_{i+1}) = \frac{\Delta t}{2} \cdot (f_t(t_i, \eta_i) + f(t_i, \eta_i) \cdot f_x(t_i, \eta_i)) + \mathcal{O}(\Delta t^2) \\ &\doteq \frac{\Delta t}{2} \cdot (f_t(t_i, \eta_i) + f(t_i, \eta_i) \cdot f_x(t_i, \eta_i)) = \mathcal{O}(\Delta t). \end{aligned} \quad (1.15)$$

(Das Symbol \doteq bedeutet, dass nur der Term der führenden Ordnung in Δt berücksichtigt wurde.)

2 Einschrittverfahren (ESV)

2.1 Das Konzept der Einschrittverfahren

Der lokale Diskretisierungsfehler des Eulerschen Polygonzugverfahrens ist von der Ordnung $\mathcal{O}(\Delta t)$. Hierbei werden Näherungslösungen von gDGL's konstruiert, indem in jedem kleinen Teilintervall eine lineare Approximation konstruiert wird, wobei als Steigung die *Steigung am linken Randpunkt* gewählt wird. Das Ziel dieses Kapitels ist es, Alternativen zum Eulerschen Polygonzugverfahren zu entwickeln, welche günstiger sind bzgl. des lokalen Diskretisierungsfehlers. Die folgenden Beispiele versuchen, einen Wert für die Steigung zu finden, welcher dem Verlauf des gesamten Teilintervalls besser angepasst ist. Wie vorher sei $t_i = t_0 + i \cdot \Delta t$ mit einem vorgegebenen kleinen Zeitschritt Δt . Zur Berechnung der lokalen Diskretisierungsfehler benötigen wir die Taylorentwicklung von

$f(.,.)$ um den Punkt (t_i, η_i) . Für $\mathcal{O}(x - \eta_i) = \mathcal{O}(t - t_i) = \mathcal{O}(\Delta t)$ ist diese gegeben durch

$$\begin{aligned} f(t, x) &= f(t_i, \eta_i) + (t - t_i) \cdot f_t(t_i, \eta_i) + (x - \eta_i) \cdot f_x(t_i, \eta_i) + 0.5 \cdot (t - t_i)^2 f_{tt}(t_i, \eta_i) \\ &\quad + (t - t_i)(x - \eta_i) f_{tx}(t_i, \eta_i) + 0.5 \cdot (x - \eta_i)^2 f_{xx}(t_i, \eta_i) + \mathcal{O}(\Delta t^3) \\ &= f(t_i, \eta_i) + (t - t_i) \cdot f_t(t_i, \eta_i) + (x - \eta_i) \cdot f_x(t_i, \eta_i) + \mathcal{O}(\Delta t^2). \end{aligned} \quad (2.1)$$

[2.1] Beispiele: (a) Modifiziertes Euler-Verfahren: Hier wird als Steigung im Intervall $[t_i, t_{i+1}]$ der Funktionswert von f im Punkt $(t_i + 0.5 \cdot \Delta t, \eta_i + 0.5 \cdot \Delta t \cdot f(t_i, \eta_i))$ angenommen. (Welche geometrisch-anschauliche Bedeutung hat dieser Punkt?) Der Iterationsschritt ist damit gegeben durch die Vorschrift

$$\eta_{i+1} = \eta_i + \Delta t \cdot f(t_i + 0.5 \cdot \Delta t, \eta_i + 0.5 \cdot \Delta t \cdot f(t_i, \eta_i)). \quad (2.2)$$

Aus der Taylorentwicklung (2.1) folgt

$$\begin{aligned} \eta_{i+1} &= \eta_i + \Delta t \cdot \left(f(t_i, \eta_i) + \frac{\Delta t}{2} \cdot f_t(t_i, \eta_i) + \frac{\Delta t}{2} \cdot f(t_i, \eta_i) \cdot f_x(t_i, \eta_i) + \mathcal{O}(\Delta t^2) \right) \\ &= \eta_i + \Delta t \cdot f(t_i, \eta_i) + \frac{\Delta t^2}{2} \cdot [f_t(t_i, \eta_i) + f(t_i, \eta_i) \cdot f_x(t_i, \eta_i)] + \mathcal{O}(\Delta t^3). \end{aligned}$$

Ein Vergleich mit der Taylorentwicklung (1.18) für $z(t) = \Phi^{t, t_i} \eta_i$ ergibt als Ordnung für den lokalen Diskretisierungsfehler

$$\epsilon(t_i, \eta_i, \Delta t) = \frac{1}{\Delta t} (z(t_{i+1}) - \eta_{i+1}) = \mathcal{O}(\Delta t^2).$$

Der lokale Diskretisierungsfehler ist damit um eine Ordnung besser als im Fall des Eulerschen Polygonzugverfahrens. Allerdings sind pro Zeitschritt zwei Funktionsauswertungen von $f(.,.)$ erforderlich.

(b) Heun-Verfahren: Hier wird als Steigung der Mittelwert der Funktion $f(.,.)$ in den Punkten (t_i, η_i) und $(t_i + \Delta t, \eta_i + \Delta t \cdot f(t_i, \eta_i))$ gewählt. Die Iterationsvorschrift lautet also

$$\eta_{i+1} = \eta_i + \frac{\Delta t}{2} [f(t_i, \eta_i) + f(t_i + \Delta t, \eta_i + \Delta t \cdot f(t_i, \eta_i))]. \quad (2.3)$$

Die Taylorentwicklung (2.1) liefert

$$\begin{aligned} \eta_{i+1} &= \eta_i + \frac{\Delta t}{2} f(t_i, \eta_i) + \frac{\Delta t}{2} \cdot \left(f(t_i, \eta_i) + \Delta t \cdot f_t(t_i, \eta_i) + \Delta t \cdot f(t_i, \eta_i) \cdot f_x(t_i, \eta_i) + \mathcal{O}(\Delta t^2) \right) \\ &= \eta_i + \Delta t \cdot f(t_i, \eta_i) + \frac{\Delta t^2}{2} \cdot [f_t(t_i, \eta_i) + f(t_i, \eta_i) \cdot f_x(t_i, \eta_i)] + \mathcal{O}(\Delta t^3). \end{aligned}$$

Wie in Beispiel (a) ergibt sich damit ein lokaler Diskretisierungsfehler der Ordnung $\mathcal{O}(\Delta t^2)$.

Eine allgemeine Klasse von Approximationsverfahren für AWP's lässt sich wie folgt formulieren.

[2.2] Definition: (a) Ein **Einschrittverfahren (ESV)** für die gDGL $x' = f(t, x)$ (bzw. für das SysgDGL $\mathbf{x}' = \mathbf{f}(t, \mathbf{x})$) ist eine Formel der Form

$$\eta_{i+1} = \eta_i + \Delta t \cdot \Psi(t_i, \eta_i, \Delta t) \quad (2.4)$$

mit einer vorgegebenen (i.a. von $f(., .)$ bzw. von $\mathbf{f}(., .)$ abhängigen) Funktion Ψ .

(b) Der **lokale Diskretisierungsfehler** für das ESV ist definiert durch

$$\epsilon(t_i, \eta_i, \Delta t) = \frac{1}{\Delta t} \cdot (\Phi^{t_i+\Delta t, t_i} \eta_i - \eta_{i+1}) = \frac{1}{\Delta t} \cdot [\Phi^{t_i+\Delta t, t_i} \eta_i - \eta_i] - \Psi(t_i, \eta_i, \Delta t) \quad (2.5)$$

(vgl. Definition (1.17)).

(c) Das ESV heißt **konsistent**, wenn gilt

$$\lim_{\Delta t \rightarrow 0} \epsilon(t_i + \Delta t, \eta_i, \Delta t) = 0; \quad (2.6)$$

dies ist nach Gleichung (2.4) und der Definition von $\Phi^{t, t_i} \eta_i$ genau dann der Fall, wenn gilt

$$\lim_{\Delta t \rightarrow 0} \Psi(t_i, \eta_i, \Delta t) = f(t_i, \eta_i). \quad (2.7)$$

(d) Das ESV ist ein **Verfahren der Ordnung p** , falls

$$\epsilon(t_i, \eta_i, \Delta t) = \mathcal{O}(\Delta t^p). \quad (2.8)$$

Nach dieser Definition ist das Eulersche Polygonzugverfahren ein ESV der Ordnung 1; das modifizierte Euler-Verfahren und das Heun-Verfahren sind ESV der Ordnung 2. Weitere ESV können z.B. mit Hilfe des Ansatzes

$$\Psi(t_i, \eta_i, \Delta t) := b_1 \cdot f(t_i, \eta_i) + b_2 \cdot f(t_i + c \cdot \Delta t, \eta_i) + a \cdot \Delta t \cdot f(t_i, \eta_i) \quad (2.9)$$

konstruiert werden, wobei die Koeffizienten b_1 , b_2 , a und c frei gewählt werden können.² Geeignete Koeffizienten ergeben sich aus folgendem Satz.

[2.3] Satz: Ein ESV mit dem Ansatz (2.9) ist genau dann von (mindestens) der Ordnung 2, wenn

$$b_1 + b_2 = 1 \quad \text{und} \quad b_2 \cdot a = b_2 \cdot c = 0.5. \quad (2.10)$$

Eine höhere als die Ordnung 2 ist mit diesem Ansatz i.a. nicht zu erreichen.

Beweis: Mit dem Ansatz (2.9) hat Ψ die Taylor-Entwicklung um (t_i, η_i)

$$\begin{aligned} \Psi(t_i, \eta_i, \Delta t) &= (b_1 + b_2) \cdot f(t_i, \eta_i) \\ &\quad + b_2 \cdot \Delta t \cdot [c \cdot f_t(t_i, \eta_i) + a \cdot f(t_i, \eta_i) \cdot f_x(t_i, \eta_i)] + \mathcal{O}(\Delta t^2). \end{aligned}$$

Die Bedingungen (2.10) folgen durch Einsetzen dieser Entwicklung in das ESV (2.4) und durch Vergleich mit der Taylorentwicklung (1.18) für Φ^{t, t_i, η_i} . \square

2.2 Runge-Kutta-Verfahren (RKV)

Ein Iterationsschritt für ein ESV mit dem Ansatz (2.9) kann in die folgenden Rechenschritte zerlegt werden.

[2.4] Algorithmus (Iterationsschritt für ein ESV mit Ansatz (2.9)): Gegeben (t_i, η_i) ,

- S1 berechne $k_1 := f(t_i, \eta_i)$;
- S2 berechne $k_2 := f(t_i + c \cdot \Delta t, \eta_i + a \cdot \Delta t \cdot k_1)$;
- S3 setze $\Psi(t_i, \eta_i, \Delta t) := b_1 \cdot k_1 + b_2 \cdot k_2$.

Dies stellt den Spezialfall eines *zweistufigen Runge-Kutta-Verfahrens* dar. Dieser Ansatz soll nun verallgemeinert und analysiert werden, um eine höhere Ordnung zu erzielen.

²Alle oben aufgeführten Beispiele genügen diesem Ansatz. Für das Eulersche Polygonzugverfahren ist $b_1 = 1$ und $b_2 = 0$; für das modifizierte Euler-Verfahren ist $b_1 = 0$, $b_2 = 1$ sowie $a = c = 0.5$; für das Heun-Verfahren gilt $b_1 = b_2 = 0.5$ und $a = c = 1$.

[2.5] **Definition:** Ein **dreistufiges Runge-Kutta-Verfahren (RKV)** ist ein ESV mit dem Ansatz

$$\begin{aligned} k_1 &= f(t_i + c_1 \cdot \Delta t, \eta_i), \\ k_2 &= f(t_i + c_2 \cdot \Delta t, \eta_i + \Delta t \cdot a_{21} \cdot k_1), \\ k_3 &= f(t_i + c_3 \cdot \Delta t, \eta_i + \Delta t \cdot (a_{31} \cdot k_1 + a_{32} \cdot k_2)), \\ \Psi(t_i, \eta_i, \Delta t) &= b_1 \cdot k_1 + b_2 \cdot k_2 + b_3 \cdot k_3, \end{aligned} \quad (2.11)$$

mit geeigneten Koeffizienten a_{ij} , b_i , c_i .

[2.6] **Beispiel:** Für

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.5 \\ 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \quad \mathcal{A} = (a_{ij}) = \begin{pmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

ist

$$\begin{aligned} k_1 &= f(t_i, \eta_i), \\ k_2 &= f(t_i + 0.5 \cdot \Delta t, \eta_i + 0.5 \cdot \Delta t \cdot f(t_i, \eta_i)), \\ k_3 &= f(t_i + \Delta t, \eta_i + \Delta t \cdot [0.\bar{3} \cdot f(t_i, \eta_i) + 0.\bar{6} \cdot f(t_i + 0.5 \cdot \Delta t, \eta_i + 0.5 \cdot \Delta t \cdot f(t_i, \eta_i))]), \\ \Psi(t_i, \eta_i, \Delta t) &= 0.\bar{3} \cdot (k_1 + k_2 + k_3). \end{aligned}$$

Wir wenden dieses Verfahren an zur Lösung der **autonomen**³ gDGL

$$x' = f(x).$$

In diesem Fall lauten die Gleichungen zur Berechnung der k_i

$$\begin{aligned} k_1 &= f(\eta_i), \\ k_2 &= f(\eta_i + 0.5 \cdot \Delta t \cdot f(\eta_i)), \\ k_3 &= f(\eta_i + \Delta t \cdot [0.\bar{3} \cdot f(\eta_i) + 0.\bar{6} f(\eta_i + 0.5 \cdot \Delta t \cdot f(\eta_i))]). \end{aligned}$$

Zur Abschätzung des lokalen Diskretisierungsfehlers betrachten wir die Taylor-Entwicklungen der k_i . (Wir schreiben abkürzend f_i anstelle von $f(\eta_i)$, sowie f'_i für die Ableitung

³Eine gDGL heißt autonom, wenn die rechte Seite f nicht von t abhängt.

von $f(x)$ an der Stelle η_i .)

$$\begin{aligned}k_1 &= f_i, \\k_2 &= f_i + 0.5 \cdot \Delta t \cdot f_i f'_i + \mathcal{O}(\Delta t^2), \\k_3 &= f_i + \Delta t \cdot f_i f'_i + \mathcal{O}(\Delta t^2).\end{aligned}$$

Hieraus ergibt sich Ψ durch

$$\Psi(t_i, \eta_i, \Delta t) = 0.3 \cdot (k_1 + k_2 + k_3) = f_i + 0.5 \cdot \Delta t \cdot f_i f'_i + \mathcal{O}(\Delta t^2).$$

Vergleichen wir dies mit der Taylorentwicklung von $z(t) = \Phi^{t, t_i} \eta_i$ (vgl. (1.18)),

$$z(t_{i+1}) = \eta_i + \Delta t \cdot f_i + 0.5 \cdot \Delta t^2 \cdot f_i f'_i + \mathcal{O}(\Delta t^3),$$

so folgt für den lokalen Diskretisierungsfehler

$$\epsilon(t_i, \eta_i, \Delta t) = \mathcal{O}(\Delta t^2).$$

Damit hat das Verfahren für autonome gDGL mindestens die Ordnung 2.

[2.7] Übung: Zeigen Sie: Für autonome gDGL's hat ein dreistufiges RKV mindestens die Ordnung 2, wenn gilt

$$b_1 + b_2 + b_3 = 1 \quad \text{und} \quad b_2 a_{21} + b_3 (a_{31} + a_{32}) = \frac{1}{2}.$$

Ähnliche Abschätzungen wie in Beispiel [2.6] können auch für nicht-autonome gDGL's durchgeführt werden, auch für "höherstufige" RKV-Verfahren, welche wie folgt definiert sind.

[2.8] Definition: Ein s -stufiges **RKV** für die gDGL $x' = f(t, x)$ ist ein ESV der Form

$$\Psi(t_i, \eta_i, \Delta t) = \sum_{i=1}^s b_i k_i \tag{2.12}$$

mit

$$k_i = f \left(t_i + c_i \cdot \Delta t, \eta_i + \Delta t \cdot \sum_{j=1}^{i-1} a_{ij} k_j \right). \tag{2.13}$$

Durch Taylor-Entwicklung lassen sich leicht Bedingungen für die Ordnung s -stufiger RKV's angeben. Wir fassen zusammen.

[2.9] **Satz:** Zwischen den Koeffizienten c_i und a_{ij} gelte zusätzlich die Beziehung⁴

$$c_i = \sum_{j=1}^s a_{ij}. \quad (2.14)$$

Dann besitzt das s -stufige RKV genau dann (mindestens) die Ordnung

1, wenn

$$\sum_{i=1}^s b_i = 1; \quad (2.15)$$

2, wenn zusätzlich gilt

$$\sum_{i=1}^s b_i c_i = \frac{1}{2}; \quad (2.16)$$

3, wenn zusätzlich gilt

$$\sum_{i=1}^s b_i c_i^2 = \frac{1}{3}, \quad \sum_{i,j=1}^s b_i a_{ij} c_j = \frac{1}{6}; \quad (2.17)$$

4, wenn zusätzlich gilt

$$\sum_{i=1}^s b_i c_i^3 = \frac{1}{4}, \quad \sum_{i,j=1}^s b_i c_i a_{ij} c_j = \frac{1}{8}, \quad \sum_{i,j=1}^s b_i a_{ij} c_j^2 = \frac{1}{12}, \quad \sum_{i,j,k=1}^s b_i a_{ij} a_{jk} c_k = \frac{1}{24}. \quad (2.18)$$

[2.10] **Beispiel:** Das in der Literatur üblicherweise als das **klassische RKV** bezeichnete Schema ist vierstufig (d.h. $s = 4$) und gegeben durch

$$\mathbf{b} = \frac{1}{6} \cdot \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 0 \\ 0.5 \\ 0.5 \\ 1 \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \quad (2.19)$$

⁴Eine Begründung für diese Bedingung findet man z.B. in Abschnitt 4.2.1 von P. Deuffhard/ F. Bornemann, Numerische Mathematik II, de Gruyter, 1994.

Die explizite Formulierung lautet

$$\begin{aligned} k_1 &= f(t_i, \eta_i), \\ k_2 &= f(t_i + 0.5 \cdot \Delta t, \eta_i + 0.5 \cdot \Delta t \cdot k_1), \\ k_3 &= f(t_i + 0.5 \cdot \Delta t, \eta_i + 0.5 \cdot \Delta t \cdot k_2), \\ k_4 &= f(t_i + \Delta t, \eta_i + \Delta t \cdot k_3), \\ \Psi(t_i, \eta_i, \Delta t) &= \frac{1}{6}(k_1 + 2 \cdot k_2 + 2 \cdot k_3 + k_4). \end{aligned}$$

Man überzeugt sich leicht, dass die Gleichungen (2.15)···(2.18) erfüllt sind – es handelt sich um ein Verfahren 4. Ordnung.

[2.11] **Übung:** Zeigen Sie, dass das dreistufige Verfahren aus Beispiel [2.6] die Ordnung 2 hat.

2.3 Konvergenz von Einschrittverfahren

Bisher haben wir uns nur mit dem *lokalen Diskretisierungsfehler* befasst, dem Fehler also, der in jedem Zeitschritt Δt neu entsteht dadurch, dass die zu approximierende Lösung eines AWP durch ein Geradenstück approximiert wird. Wir werden nun feststellen, dass dieser Fehler auch maßgeblich ist für die *globale Abweichung* der Näherung von der exakten Trajektorie.

Im Folgenden untersuchen wir im Zeitintervall $[t_0, T]$ das AWP

$$x' = f(t, x), \quad x(t_0) = x_0. \quad (2.20)$$

Die exakte Lösung bezeichnen wir mit $x(t)$, also

$$x(t) = \Phi^{t, t_0} x_0.$$

Um einen Wert $x(t)$ ($t \in [t_0, T]$) zu approximieren, kann mit Hilfe einer Funktion Ψ ein ESV der Form (2.4) gewählt und auf eine Schrittweite $\Delta t = (t - t_0)/N$, $N \in \mathbb{N}$, angewandt werden. Untersucht werden soll, wie der **globale Diskretisierungsfehler**

$$\mathbf{e}(t, N) := \eta_N - x(t) \quad (2.21)$$

von der Schrittweite Δt (bzw. von der Zahl N) abhängt. Wir bezeichnen das ESV als **konvergent im Intervall** $[t_0, T]$, falls für alle $t \in [t_0, T]$ gilt

$$\lim_{N \rightarrow \infty} \mathbf{e}(t, N) = 0$$

Das folgende zentrale Ergebnis stellt den globalen Fehler in Relation zum lokalen Diskretisierungsfehler.

[2.12] Satz: Die Funktion Ψ sei stetig und erfülle die Lipschitz-Bedingung

$$|\Psi(t, x_1, h) - \Psi(t, x_2, h)| \leq M \cdot |x_1 - x_2| \quad (2.22)$$

mit einer Konstanten M . Dann gilt für das durch Ψ definierte ESV (2.4): Ist das ESV ein Verfahren der Ordnung p (vgl. Def. [2.2](d)), so erfüllt mit $\Delta t := (t - t_0)/N$ und mit einem geeigneten λ der globale Diskretisierungsfehler eine Abschätzung der Form

$$|\mathbf{e}(t, N)| \leq \lambda \cdot |\Delta t|^p \cdot (\exp(M(t - t_0)) - 1)/M. \quad (2.23)$$

Bevor dieser Satz bewiesen wird, soll seine Aussage kurz kommentiert werden.

[2.13] Bemerkungen: (a) Die Bedingung (2.22) lässt sich in den oben vorgestellten Verfahren leicht aus der Lipschitz-Stetigkeit (1.5) der rechten Seite $f(t, x)$ ableiten, welche wir ja immer vorausgesetzt haben. Beispielsweise folgt im skalaren Fall

$$|f(t, x_1) - f(t, x_2)| \leq L \cdot |x_1 - x_2|$$

für das Heun-Verfahren mit

$$\Psi(t, x, h) = \frac{1}{2} [f(t, x) + f(t + h, x + h \cdot f(t, x))]$$

die Abschätzung

$$\begin{aligned} |\Psi(t, x_1, h) - \Psi(t, x_2, h)| &\leq \frac{1}{2} |f(t, x_1) - f(t, x_2)| \\ &\quad + \frac{1}{2} |f(t + h, x_1 + h \cdot f(t, x_1)) - f(t + h, x_2 + h \cdot f(t, x_2))| \\ &\leq \frac{L}{2} \cdot |x_1 - x_2| + \frac{L}{2} \cdot (|x_1 - x_2| + h \cdot |f(t, x_1) - f(t, x_2)|) \\ &\leq \left(L + \frac{h \cdot L^2}{2} \right) \cdot |x_1 - x_2|. \end{aligned}$$

(b) Der globale Fehler ist wie der lokale Fehler von der Ordnung $\mathcal{O}(\Delta t^p)$. Zu beachten ist allerdings, dass die rechte Seite der Abschätzung (2.23) mit steigendem Abstand vom Anfangspunkt t_0 exponentiell wächst.

Zum Beweis des Satzes [2.12] benötigen wir das folgende Hilfsergebnis.

[2.14] Lemma: Genügt eine Zahlenfolge ξ_i , $i = 0, 1, 2, \dots$, einer Abschätzung der Form

$$|\xi_{i+1}| \leq (1 + \delta)|\xi_i| + B \quad (2.24)$$

mit den Konstanten $\delta > 0$ und $B \geq 0$, so gilt für alle $n \in \mathbb{N}$

$$|\xi_n| \leq \exp(n\delta) \cdot |\xi_0| + B \cdot \frac{\exp(n\delta) - 1}{\delta}.$$

Beweis von Lemma [2.14]: Die iterative Anwendung der Abschätzung (2.24) ergibt (mit Induktion)

$$\begin{aligned} |\xi_1| &\leq (1 + \delta)|\xi_0| + B, \\ |\xi_2| &\leq (1 + \delta)|\xi_1| + B \leq (1 + \delta)^2 \cdot |\xi_0| + B \cdot (1 + (1 + \delta)), \\ |\xi_3| &\leq (1 + \delta)|\xi_2| + B \leq (1 + \delta)^3 \cdot |\xi_0| + B \cdot (1 + (1 + \delta) + (1 + \delta)^2), \\ &\vdots \\ |\xi_n| &\leq (1 + \delta)^n \cdot |\xi_0| + B \cdot (1 + (1 + \delta) + \dots + (1 + \delta)^{n-1}) \\ &= (1 + \delta)^n \cdot |\xi_0| + B \cdot \frac{(1 + \delta)^n - 1}{\delta} \\ &\leq \exp(n\delta) \cdot |\xi_0| + B \cdot \frac{\exp(n\delta) - 1}{\delta}. \end{aligned}$$

Letztere Ungleichung folgt wegen $1 + \delta \leq \exp(\delta)$. \square

Beweis von Satz [2.12]: Mit $\Delta t = (t - t_0)/N$ definieren wir die Knoten $t_i := t_0 + i \cdot \Delta t$, die exakten Werte an den Knoten $x_i := x(t_i)$ sowie die (globalen) Diskretisierungsfehler an den Knoten $\mathbf{e}_i := \eta_i - x_i$. Aus der rekursiven Definition (2.4) für η_{i+1} und der Beziehung $x_{i+1} = \Phi^{t_{i+1}, t_i} x_i$ folgt

$$\begin{aligned} \mathbf{e}_{i+1} &= \eta_{i+1} - x_{i+1} = \eta_i + \Delta t \cdot \Psi(t_i, \eta_i, \Delta t) - \Phi^{t_{i+1}, t_i} x_i \\ &= [\eta_i - x_i] + \Delta t \cdot [\Psi(t_i, \eta_i, \Delta t) - \Psi(t_i, x_i, \Delta t)] + \Delta t \left[\Psi(t_i, x_i, \Delta t) - \frac{\Phi^{t_{i+1}, t_i} x_i - x_i}{\Delta t} \right]. \end{aligned}$$

	$\Delta t = 0.1$		$\Delta t = 0.01$		$\Delta t = 0.001$	
	Fehler	rel. Fehler	Fehler	rel. Fehler	Fehler	rel. Fehler
Euler	2.275E+05	8.479E-01	5.463E+04	2.036E-01	6.173E+03	2.300E-02
Heun	4.758E+04	1.177E-01	6.775E+02	2.525E-03	6.967E+00	2.596E-05
RKV	4.091E+02	1.525E-03	5.619E-02	2.094E-07	1.431E-06	5.331E-12

Tabelle 3: Numerische Berechnung von $x(5)$ aus Beispiel [2.15]

Nach der Lipschitz-Bedingung (2.22) ist

$$|\Psi(t_i, \eta_i, \Delta t) - \Psi(t_i, x_i, \Delta t)| \leq M \cdot |\eta_i - x_i| = M \cdot |\mathbf{e}_i|.$$

Wegen der Ordnung p des ESV ist (vgl. (2.5))

$$\left| \Psi(t_i, x_i, \Delta t) - \frac{\Phi^{t_{i+1}, t_i} x_i - x_i}{\Delta t} \right| \leq \lambda \cdot \Delta t^p$$

für ein geeignetes λ . Hieraus ergibt sich für die Folge \mathbf{e}_i die Abschätzung

$$|\mathbf{e}_{i+1}| \leq (1 + M \cdot \Delta t) \cdot |\mathbf{e}_i| + \lambda \cdot \Delta t^{p+1}.$$

Wegen Lemma [2.14] und $\mathbf{e}_0 = 0$ folgt hieraus

$$|\mathbf{e}_N| \leq \lambda \cdot \Delta t^{p+1} \cdot \frac{\exp(NM\Delta t) - 1}{M\Delta t} = \lambda \cdot \Delta t^p \cdot \frac{\exp(M(t - t_0)) - 1}{M}. \quad \square$$

[2.15] Beispiel: Wir testen drei Verfahren unterschiedlicher Ordnung, das Eulersche Polygonzugverfahren, das Heun-Verfahren und das klassische RKV, am AWP

$$x'(t) = t \cdot x(t), \quad x(0) = 1.$$

Die exakte Lösung ist gegeben durch

$$x(t) = \exp(0.5 \cdot t^2).$$

Zur Berechnung von $x(5)$ [= 2.68337...E+05] wird das Intervall $[0, 5]$ in 50, 500 und 5000 gleiche Teilintervalle unterteilt. Die globalen Diskretisierungsfehler sind in Tabelle

3 dargestellt. Daneben ist der relative Fehler $(x_{\text{exakt}}(5) - x_{\text{numerisch}}(5))/x_{\text{exakt}}(5)$ zu finden, welcher angibt, auf wie viele führende Stellen das numerische Ergebnis exakt ist. Es zeigt sich, dass das klassische RKV ohne großen Aufwand eine Genauigkeit liefert, welche nahe an die rechnerinterne Darstellungsgenauigkeit reeller Zahlen kommt.

3 Lineare Mehrschrittverfahren (MSV)

3.1 Das Konzept der Mehrschrittverfahren

Das Konzept der ESV bestand in der Berechnung einer Approximation η_{i+1} aus dem vorhergehenden Wert η_i . Um hierbei Verfahren höherer Ordnung zu erzielen (z.B. RKV), ist es in jedem Zeitschritt nötig, die Funktion $f(t, x)$ an mehreren Zwischenstellen auszuwerten. Mehrschrittverfahren dagegen stellen Berechnungsverfahren für η_{i+1} dar, für welche die bereits bekannten Werte $\eta_i, \dots, \eta_{i-k}$ verwendet werden. Die zugrunde liegenden Ideen entstammen der Interpolationstheorie.⁵ Dies illustrieren die folgenden ersten Beispiele zur Lösung des AWP

$$x' = f(t, x), \quad x(t_0) = x_0. \quad (3.1)$$

Die exakte Lösung werde wieder mit $x(t)$ bezeichnet. Wie früher seien zu einem vorgegebenen Zeitschritt Δt die diskreten Zeiten $t_i := t_0 + i \cdot \Delta t$ definiert.

[3.1] Beispiele: Durch Integration der gDGl des AWP folgt

$$\int_{t_i}^{t_{i+1}} x'(t) dt = x(t_{i+1} - t_i) = \int_{t_i}^{t_{i+1}} f(t, x(t)) dt.$$

Bezeichnen wir

$$F(t) := f(t, x(t)),$$

so erhalten wir die folgende *exakte* Formel

$$x(t_{i+1}) = x(t_i) + \int_{t_i}^{t_{i+1}} F(t) dt. \quad (3.2)$$

⁵Vgl. Vorlesung Num. Mathematik I, Abschnitte 4.1, 5.2.

Zu ihrer Integration erinnern wir uns an die Ideen zur Herleitung der Newton-Cotes-Formeln zur Integration.

(a) Wir ersetzen den Integranden $F(t)$ durch das lineare Interpolationspolynom zu den Punkten $(t_{i-1}, F(t_{i-1}))$ und $(t_i, F(t_i))$:

$$P_1(t) := F(t_i) + (t - t_i) \cdot \frac{F(t_{i-1}) - F(t_i)}{t_{i-1} - t_i} = F(t_i) + (t - t_i) \cdot \frac{F(t_i) - F(t_{i-1})}{\Delta t}.$$

Setzen wir $P_1(t)$ anstelle von $F(t)$ in die Gleichung (3.2) ein, so folgt

$$x(t_{i+1}) \approx x(t_i) + \Delta t \cdot [1.5f(t_i, x(t_i)) - 0.5f(t_{i-1}, x(t_{i-1}))].$$

Dies führt zum *expliziten* Mehrschrittverfahren

$$\eta_{i+1} = \eta_i + \Delta t \cdot [1.5f(t_i, \eta_i) - 0.5f(t_{i-1}, \eta_{i-1})].$$

(”explizit” deshalb, da η_{i+1} durch diese Formel direkt aus den Werten η_i und η_{i-1} berechnet werden kann.)

(b) Ersetzen wir $F(t)$ durch das lineare Interpolationspolynom zu den Punkten $(t_i, F(t_i))$ und $(t_{i+1}, F(t_{i+1}))$:

$$P_1(t) = F(t_i) + (t - t_i) \cdot \frac{F(t_{i+1}) - F(t_i)}{\Delta t},$$

so erhalten wir

$$x(t_{i+1}) \approx x(t_i) + \frac{\Delta t}{2} \cdot [f(t_i, x(t_i)) + f(t_{i+1}, x(t_{i+1}))]$$

mit dem entsprechenden *impliziten* Verfahren

$$\eta_{i+1} = \eta_i + 0.5 \cdot \Delta t \cdot [f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1})].$$

(”implizit” deshalb, weil die Gleichung nicht unmittelbar nach η_{i+1} aufgelöst werden kann.)

(c) Ersetzen wir schließlich $F(t)$ durch das quadratische Interpolationspolynom $P_2(t)$ zu den Punkten (t_{i-2}, F_{i-2}) , (t_{i-1}, F_{i-1}) und (t_i, F_i) :

$$P_2(t) = F_i + \frac{1}{2 \cdot \Delta t^2} \cdot \{(t - t_i)^2 \cdot [F_{i-2} - 2F_{i-1} + F_i] + \Delta t \cdot (t - t_i) \cdot [F_{i-2} - 4F_{i-1} + 3F_i]\}$$

(Rechnen Sie diese Darstellung als kleine Übung nach, z.B. mit Hilfe der Lagrange-Polynome!), so folgen die Näherung

$$x(t_{i+1}) \approx x(t_i) + \frac{\Delta t}{12} \cdot [5f(t_{i-2}, x(t_{i-2})) - 16f(t_{i-1}, x(t_{i-1})) + 23f(t_i, x(t_i))]$$

sowie das explizite Mehrschrittverfahren

$$\eta_{i+1} = \eta_i + \frac{\Delta t}{12} \cdot [5f(t_{i-2}, \eta_{i-2}) - 16f(t_{i-1}, \eta_{i-1}) + 23f(t_i, \eta_i)].$$

Die oben erhaltenen Ansätze werden wie folgt verallgemeinert.

[3.2] Definition: Ein Verfahren der Form

$$\eta_{i+r} + a_{r-1}\eta_{i+r-1} + \cdots + a_0\eta_i = \Delta t \cdot [b_r f(t_{i+r}, \eta_{i+r}) + \cdots + b_0 f(t_i, \eta_i)] \quad (3.3)$$

heißt (**lineares**) **Mehrschrittverfahren (MSV)** (genauer: **r-Schritt-Verfahren**).

Ist $b_r = 0$, so heißt das MSV **explizit**, andernfalls **implizit**.

In den **Beispielen [3.1]** beschreibt nach dieser Definition (a) ein explizites 2-Schritt-Verfahren:

$$\eta_{i+2} - \eta_{i+1} = \Delta t \cdot [1.5f(t_{i+1}, \eta_{i+1}) - 0.5f(t_i, \eta_i)].$$

Das Beispiel (b) beschreibt ein implizites 1-Schritt-Verfahren (insbesondere also ein ESV):

$$\eta_{i+1} - \eta_i = \frac{\Delta t}{2} \cdot [f(t_{i+1}, \eta_{i+1}) + f(t_i, \eta_i)].$$

Durch (c) ist ein explizites 3-Schritt-Verfahren definiert gemäß

$$\eta_{i+3} - \eta_{i+2} = \frac{\Delta t}{12} \cdot [5f(t_i, \eta_i) - 16f(t_{i+1}, \eta_{i+1}) + 23f(t_{i+2}, \eta_{i+2})].$$

Wir stellen im folgenden drei wichtige Klassen von MSV vor.

A – Adams-Bashforth-Verfahren

Dies sind explizite Verfahren, bei denen (in Verallgemeinerung zu den Beispielen [3.1](a),

$k =$	0	1	2	3	4	5
$\beta_k^{(0)}$	1					
$\beta_k^{(1)}$	$-\frac{1}{2}$	$\frac{3}{2}$				
$\beta_k^{(2)}$	$\frac{5}{12}$	$-\frac{16}{12}$	$\frac{23}{12}$			
$\beta_k^{(3)}$	$-\frac{9}{24}$	$\frac{37}{24}$	$-\frac{59}{24}$	$\frac{55}{24}$		
$\beta_k^{(4)}$	$\frac{251}{720}$	$-\frac{1274}{720}$	$\frac{2616}{720}$	$-\frac{2774}{720}$	$\frac{1901}{720}$	
$\beta_k^{(5)}$	$-\frac{475}{1440}$	$\frac{2877}{1440}$	$-\frac{7298}{1440}$	$\frac{9982}{1440}$	$-\frac{7923}{1440}$	$\frac{4277}{1440}$

Tabelle 4: Die Koeffizienten der Adams-Bashforth-Verfahren

(c) die Funktion $F(t)$ aus Gleichung (3.2) im Intervall $[t_{i+q}, t_{i+q+1}]$ durch das Interpolationspolynom zu den Knoten t_i, \dots, t_{i+q} approximiert wird. Sind $L_k, k = 0, \dots, q$, die zugehörigen Lagrangepolynome,

$$L_k(t) = \prod_{\substack{l=0 \\ l \neq k}}^q \frac{t - t_l}{t_k - t_l},$$

so ist mit den Koeffizienten $\beta_k^{(q)}$, definiert durch

$$\Delta t \cdot \beta_k^{(q)} := \int_{t_{i+q}}^{t_{i+q+1}} L_k(t) dt =_{s:=(t-t_i)/\Delta t} \Delta t \cdot \int_q^{q+1} \prod_{\substack{l=0 \\ l \neq k}}^q \frac{s - l}{k - l} ds$$

und mit $f_k := f(t_k, \eta_k)$ durch

$$\eta_{i+q+1} - \eta_{i+q} = \Delta t \cdot [\beta_q^{(q)} f_{i+q} + \beta_{q-1}^{(q)} f_{i+q-1} + \dots + \beta_0^{(q)} f_i] \tag{3.4}$$

das $(q + 1)$ -stufige **Adams-Bashforth-Verfahren** gegeben. Die Koeffizienten $\beta_k^{(q)}$ sind in Tabelle 4 dargestellt. Definieren wir den lokalen Diskretisierungsfehler wie im Fall der ESV, so hat dieses Verfahren die Ordnung $q + 1$ (vgl. Abschnitt 3.2).

Das $(q+1)$ -stufige Verfahren kann erstmalig angewendet werden, wenn die Werte η_0, \dots, η_q bekannt sind. Diese Werte müssen durch andere Verfahren, etwa ESV *der gleichen Ordnung*, berechnet werden. Das Schema des numerischen Verfahrens zur Approximation von $x(t)$ im Intervall $[t_0, T]$ sieht demnach wie folgt aus.

[3.3] Algorithmus (Adams-Bashforth-Verfahren):

$k =$	0	1	2	3	4	5
$\beta_k^{(0)}$	1					
$\beta_k^{(1)}$	$\frac{1}{2}$	$\frac{1}{2}$				
$\beta_k^{(2)}$	$-\frac{1}{12}$	$\frac{8}{12}$	$\frac{5}{12}$			
$\beta_k^{(3)}$	$\frac{1}{24}$	$-\frac{5}{24}$	$\frac{19}{24}$	$\frac{9}{24}$		
$\beta_k^{(4)}$	$-\frac{19}{720}$	$\frac{106}{720}$	$-\frac{264}{720}$	$\frac{646}{720}$	$\frac{251}{720}$	
$\beta_k^{(5)}$	$\frac{27}{1440}$	$-\frac{173}{1440}$	$\frac{482}{1440}$	$-\frac{798}{1440}$	$\frac{1427}{1440}$	$\frac{475}{1440}$

Tabelle 5: Die Koeffizienten der Adams-Moulton-Verfahren

- S0 Initialisierung: Wähle Schrittzahl $N \in \mathbb{N}$ und Ordnung $q + 1$; definiere $\Delta t := (T - t_0)/N$ und setze $\eta_0 := x_0$.
- S1 Anlaufrechnung: Wähle ESV $\Psi(t, x, h)$ mit der selben Ordnung und berechne $\eta_{i+1} := \eta_i + \Delta t \cdot \Psi(t_i, \eta_i, \Delta t)$ für $i = 0, \dots, q - 1$.
- S2 Einsatz des Adams-Bashforth-Verfahrens: Für $i := q, \dots, N - 1$ berechne

$$\eta_{i+1} := \eta_i + \Delta t \cdot \sum_{k=0}^q \beta_{q-k}^{(q)} f(t_{i-k}, \eta_{i-k}).$$

B – Adams-Moulton-Verfahren

Diese impliziten Verfahren erhält man, wenn man die Funktion $F(t)$ im Intervall $[t_{i+q-1}, t_{i+q}]$ durch das Interpolationspolynom zu den Knoten t_i, \dots, t_{i+q} approximiert. Das q -Schritt-Adams-Moulton-Verfahren ($q \geq 1$) hat die Form

$$\eta_{i+q} - \eta_{i+q-1} = \Delta t \cdot [\beta_q^{(q)} f_{i+q} + \dots + \beta_0^{(q)} f_i]. \tag{3.5}$$

Die Koeffizienten $\beta_k^{(q)}$ (einschließlich des impliziten ESV für $q = 0$) gehen aus Tabelle 5 hervor. Als **klassisches Adams-Moulton-Verfahren** wird das Verfahren für $q = 4$ bezeichnet.

Als Schwierigkeit gegenüber den Adams-Bashforth-Verfahren ergibt sich hier, dass η_{i+q} nicht direkt aus Gleichung (3.5) berechnet werden kann, da diese unbekannte Größe auch in $f_{i+q} = f(t_{i+q}, \eta_{i+q})$ auftaucht. Es bietet sich hier an, einen Löser für nichtlineare Gleichungen – etwa das Newton-Verfahren – auf die Gleichung $R(\xi) = 0$ anzuwenden mit

$$R(\xi) := \xi - \eta_{i+q-1} - \Delta t \cdot [\beta_q^{(q)} f(t_{i+q}, \xi) + \beta_{q-1}^{(q)} f_{i+q-1} + \cdots + \beta_0^{(q)} f_i]. \quad (3.6)$$

[3.4] Beispiel: Wir betrachten das Adams-Moulton-Verfahren ($q = 2$) für die gDGL $x'(t) = t + x(t)^n$. Aus dem Ansatz

$$\eta_{i+1} = \eta_i + \frac{\Delta t}{12} [5f(t_{i+1}, \eta_{i+1}) + 8f(t_i, \eta_i) - f(t_{i-1}, \eta_{i-1})]$$

folgt die Gleichung für die Unbekannte η_{i+1} ,

$$\eta_{i+1} - \frac{5 \cdot \Delta t}{12} \eta_{i+1}^n = C_i, \quad (3.7)$$

mit der Konstanten

$$C_i = \eta_i + \Delta t \cdot \left[\frac{2}{3} \eta_i^n - \frac{1}{12} \eta_{i-1}^n + t_i \right] + \frac{\Delta t^2}{3}.$$

Für $n = 1, 2$ lässt sich die Gleichung leicht nach η_{i+1} auflösen mit dem Ergebnis

$$\begin{aligned} \eta_{i+1} &= \frac{12}{12 - 5 \cdot \Delta t} \cdot C_i & (n = 1), \\ \eta_{i+1} &= \frac{6}{5 \cdot \Delta t} \left(1 - \sqrt{1 - \frac{5 \cdot \Delta t}{3} \cdot C_i} \right) & (n = 2). \end{aligned}$$

(Überlegen Sie sich, warum für $n = 2$ die andere Lösung der quadratischen Gleichung (3.7) nicht in Frage kommt! Entwickeln Sie hierzu η_{i+1} in eine Reihe bzgl. Δt .) Für andere Werte n bietet sich das Newton-Verfahren an zur Lösung der Gleichung (3.7).

Zur numerischen Lösung muss das Adams-Bashforth-Verfahren [3.3] wie folgt modifiziert werden.

[3.5] Algorithmus (Adams-Moulton-Verfahren):

$S0, S1$ (wie in [3.3])

$S2$ berechne η_{i+1} als Lösung von $R(\xi) = 0$ (z.B. mit dem Newton-Verfahren), wobei $R(\cdot)$ gegeben ist durch (3.6).

Mit den Kriterien des Abschnitts 3.2 kann festgestellt werden, dass das q -Schritt-Adams-Moulton-Verfahren von der Ordnung $q + 1$ ist.

C – Prädiktor-Korrektor-Verfahren

Die Lösung der Gleichung $R(\xi) = 0$ kann unter Umständen sehr aufwendig sein. Prädiktor-Korrektor-Verfahren⁶ verkürzen diesen Weg, indem die Nullstelle von $R(\xi)$ nicht exakt, sondern nur näherungsweise berechnet wird. Hierbei wird zunächst mit Hilfe eines expliziten Verfahrens ein Schätzwert $\bar{\eta}_{i+q}$ für η_{i+q} ermittelt (Prädiktor); anschließend wird in der Funktion $R(\xi)$ des impliziten Verfahrens der Ausdruck $f(t_{i+q}, \xi)$ ersetzt durch $f(t_{i+q}, \bar{\eta}_{i+q})$ und η_{i+q} ermittelt als (explizite) Nullstelle der so veränderten Funktion (Korrektor). Es bietet sich an, als Prädiktor ein q -Schritt-Adams-Bashforth-Verfahren und als Korrektor ein q -Schritt-Adams-Moulton-Verfahren zu verwenden. Es zeigt sich, dass die Ordnung des so veränderten Verfahrens gleich der des entsprechenden impliziten Adams-Moulton-Verfahrens ist (vgl. Abschnitt 3.2).

[3.6] Beispiel: Für $q = 3$ lautet ein Zeitschritt des Prädiktor-Korrektor-Verfahrens

$S1$ Bestimme den Prädiktorwert $\eta_{i+1}^{(P)}$ durch das Adams-Bashforth-Verfahren gemäß

$$\eta_{i+1}^{(P)} = \eta_i + \frac{\Delta t}{12} \cdot [23f(t_i, \eta_i) - 16f(t_{i-1}, \eta_{i-1}) + 5f(t_{i-2}, \eta_{i-2})];$$

$S2$ berechne η_{i+1} aus der Gleichung für das Adams-Moulton-Verfahren durch

$$\eta_{i+1} = \eta_i + \frac{\Delta t}{24} \cdot [9f(t_{i+1}, \eta_{i+1}^{(P)}) + 19f(t_i, \eta_i) - 5f(t_{i-1}, \eta_{i-1}) + f(t_{i-2}, \eta_{i-2})].$$

⁶*praedicare* (lat.) = vorhersagen, *corrigere* (lat.) = verbessern

3.2 Die Ordnung linearer MSV

Ähnlich wie in der Definition [2.2] für ESV definieren wir die Ordnung von MSV mit Hilfe des lokalen Diskretisierungsfehlers, welcher sich ergibt, wenn wir die exakte Lösung der gDGL in die Gleichung für das MSV einsetzen. Im Folgenden werde wieder für gegebenes (t_i, η_i) mit $z(t)$ die exakte Lösung der gDGL durch diesen Punkt bezeichnet:

$$z(t) := \Phi^{t, t_i} \eta_i.$$

[3.7] Definition: (a) Der **lokale Diskretisierungsfehler** für das MSV (3.3) ist definiert durch

$$\begin{aligned} \epsilon(t_i, \eta_i, \Delta t) &:= \frac{1}{\Delta t} \cdot [z(t_{i+r}) + a_{r-1}z(t_{i+r-1}) + \cdots + a_0z(t_i)] \\ &\quad - [b_r f(t_{i+r}, z(t_{i+r})) + \cdots + b_0 f(t_i, z(t_i))]. \end{aligned}$$

(b) Das MSV heißt **von der Ordnung p** , wenn

$$\epsilon(t_i, \eta_i, \Delta t) = \mathcal{O}(\Delta t^p).$$

Die Berechnung der Ordnung erfolgt wieder mit Hilfe von Taylorentwicklungen um (t_i, η_i) . Aus

$$z(t_{i+\ell}) = z(t_i) + \ell \Delta t \cdot z'(t_i) + \frac{1}{2!} (\ell \Delta t)^2 \cdot z''(t_i) + \cdots$$

und

$$f(t_{i+\ell}, z(t_{i+\ell})) = z'(t_{i+\ell}) = z'(t_i) + \ell \Delta t \cdot z''(t_i) + \frac{1}{2!} (\ell \Delta t)^2 \cdot z'''(t_i) + \cdots$$

folgt leicht die Entwicklung für den lokalen Diskretisierungsfehler

$$\epsilon(t_i, \eta_i, \Delta t) = \frac{c_0}{\Delta t} z(t_i) + c_1 z'(t_i) + c_2 \Delta t z''(t_i) + \cdots = \sum_{\ell} c_{\ell} \Delta t^{\ell-1} z^{(\ell)}(t_i)$$

mit

$$c_0 := a_0 + a_1 + \cdots + a_r, \tag{3.8}$$

$$c_1 := [a_1 + 2a_2 + \cdots + ra_r] - [b_0 + b_1 + \cdots + b_r] \tag{3.9}$$

sowie für $\ell \geq 2$

$$c_\ell := \frac{1}{\ell!} [a_1 + 2^\ell a_2 + \dots + r^\ell a_r] - \frac{1}{(\ell-1)!} [b_1 + 2^{\ell-1} b_2 + \dots + r^{\ell-1} b_r]. \quad (3.10)$$

Hierbei wurde $a_r := 1$ gesetzt. Wir fassen zusammen.

[3.8] Satz: Das MSV hat genau dann (mindestens) die Ordnung p ($p \geq 1$), wenn mit den Definitionen (3.8), (3.9) und (3.10) für c_ℓ gilt

$$c_0 = c_1 = \dots = c_p = 0.$$

Beim Entwerfen von MSV ist es nun ein Ziel, die Koeffizienten a_ℓ, b_ℓ so zu bestimmen, dass möglichst viele der c_ℓ verschwinden.

[3.9] Beispiele: (a) Ein explizites 3-Schritt-Verfahren der Form

$$\eta_{i+1} + a_1 \eta_{i-1} = \Delta t \cdot [b_0 f(t_{i-2}, \eta_{i-2}) + b_1 f(t_{i-1}, \eta_{i-1}) + b_2 f(t_i, \eta_i)]$$

mit maximaler Ordnung soll bestimmt werden. (Man beachte, dass $a_3 = 1$ und $a_0 = a_2 = 0$.) Die maximale Ordnung 3 erhält man, wenn a_1, b_0, b_1 und b_2 so bestimmt werden, dass

$$\begin{aligned} 0 &= c_0 = a_1 + 1, \\ 0 &= c_1 = a_1 + 3 - b_0 - b_1 - b_2 = 0, \\ 0 &= 2c_2 = a_1 + 9 - 2b_1 - 4b_2, \\ 0 &= 6c_3 = a_1 + 27 - 3b_1 - 12b_2. \end{aligned}$$

Die Lösung

$$a_1 = -1, \quad b_0 = 1/3, \quad b_1 = -2/3, \quad b_2 = 7/3$$

führt auf das Verfahren 3. Ordnung

$$\eta_{i+1} = \eta_{i-1} + \frac{\Delta t}{3} [7f(t_i, \eta_i) - 2f(t_{i-1}, \eta_{i-1}) + f(t_{i-2}, \eta_{i-2})].$$

(b) Zur Bestimmung einer impliziten 3-Schritt-Methode der Form

$$\eta_{i+1} + a_1\eta_{i-1} = \Delta t \cdot [b_0f(t_{i-2}, \eta_{i-2}) + b_1f(t_{i-1}, \eta_{i-1}) + b_2f(t_i, \eta_i) + b_3f(t_{i+1}, \eta_{i+1})]$$

mit maximaler Ordnung ist das Gleichungssystem

$$\begin{aligned} 0 &= c_0 = a_1 + 1, \\ 0 &= c_1 = a_1 + 3 - b_0 - b_1 - b_2 - b_3 = 0, \\ 0 &= 2c_2 = a_1 + 9 - 2b_1 - 4b_2 - 6b_3, \\ 0 &= 6c_3 = a_1 + 27 - 3b_1 - 12b_2 - 27b_3, \\ 0 &= 24c_4 = a_1 + 81 - 4b_1 - 32b_2 - 108b_3 \end{aligned}$$

zu lösen mit dem Ergebnis

$$a_1 = -1, \quad b_0 = 0, \quad b_1 = 1/3, \quad b_2 = 4/3, \quad b_3 = 1/3.$$

Das hieraus resultierende Verfahren hat die Ordnung 4 und ist – entgegen dem ursprünglichen Ansatz – wegen $b_0 = 0$ ein implizites 2-Schritt-Verfahren.

(c) Das implizite 3-Schritt-Verfahren maximaler Ordnung mit dem Ansatz

$$a_0\eta_{i-2} + a_1\eta_{i-1} + a_2\eta_i + \eta_{i+1} = \Delta t \cdot b_3 \cdot f_{i+1}$$

erhält man für

$$a_0 = -2/11, \quad a_1 = 9/11, \quad a_2 = -18/11, \quad b_3 = 6/11;$$

es ist gegeben durch

$$\frac{11}{6}\eta_{i+1} - 3\eta_i + \frac{3}{2}\eta_{i-1} - \frac{1}{3}\eta_{i-2} = \Delta t \cdot f(t_{i+1}, \eta_{i+1}).$$

Dieses Verfahren gehört in die Klasse der **BDF-** ("backward differentiation formula")

Methoden, da die linke Seite eine (Rückwärts-) Approximation der ersten Ableitung zur

Zeit t_{i+1} darstellt: Ersetzen wir η_{i+l} durch $z(t_{i+l}) := \Phi^{t_i+l, t_{i+1}}\eta_{i+1}$, so folgt

$$\begin{aligned} & \frac{11}{6}z(t_{i+1}) - 3z(t_i) + \frac{3}{2}z(t_{i-1}) - \frac{1}{3}z(t_{i-2}) \\ &= \left[\frac{11}{6} - 3 + \frac{3}{2} - \frac{1}{3} \right] \cdot z(t_{i+1}) + \left[3 - 2 \cdot \frac{3}{2} + 3 \cdot \frac{1}{3} \right] \cdot \Delta t \cdot z'(t_{i+1}) + \mathcal{O}(\Delta t^2) \\ &= \Delta t \cdot z'(t_{i+1}) + \mathcal{O}(\Delta t^2). \end{aligned}$$

[3.10] Übung: (a) Zeigen Sie: Das 3-Schritt-Adams-Bashforth-Verfahren hat die Ordnung 3, das 3-Schritt-Adams-Moulton-Verfahren die Ordnung 4.

(b) Zeigen Sie durch Taylorreihenentwicklung, dass das Prädiktor-Korrektor-Verfahren aus Beispiel [3.6] die Ordnung 4 hat. Vergleichen Sie den Rechenaufwand für dieses Verfahren mit dem für das Adams-Moulton-Verfahren der selben Ordnung.

Die Ergebnisse des Beispiels lassen sich verallgemeinern. Es gilt (ohne Beweis)

[3.11] Satz: Das q -Schritt-Adams-Bashforth-Verfahren hat die Ordnung q ; das q -Schritt-Adams-Moulton-Verfahren hat die Ordnung $q + 1$. Das zugehörige Prädiktor-Korrektor-Verfahren, zusammengesetzt aus dem q -Schritt-Adams-Bashforth-Prädiktor und dem q -Schritt-Adams-Moulton-Korrektor, hat ebenfalls die Ordnung $q + 1$.

3.3 Homogene lineare Differenzgleichungen

Lineare Differenzgleichungen, wie sie bei der Formulierung von MSV entstehen,

$$\sum_{k=0}^r a_k \eta_{i+k} = c_i$$

(vgl. Definition [3.2]), entwickeln eine "Eigendynamik". Dies ist bereits der Fall für $c_i = 0$ (entsprechend der gDGL $x' = 0$). Diese Differenzgleichungen sind der Inhalt des folgenden Abschnitts.

[3.12] Beispiel: Die Formel

$$\eta_{i+1} + 4\eta_i - 5\eta_{i-1} = 2\Delta t \cdot (2f_i + f_{i-1})$$

beschreibt ein explizites 2-Schritt-Verfahren. (Bestimmen Sie die Ordnung!) Wir wenden dieses auf das AWP $x'(t) = 0$, $x(t_0) = 1$ an. In diesem Fall ist $\eta_0 = 1$; wir nehmen an, dass η_1 durch ein ESV mit einem gewissen Rundungsfehler berechnet wurde: $\eta_1 = 1 + \epsilon \cdot \Delta t$. Für $\epsilon = 10^{-6}$ und $\Delta t = 0.05$ erhalten wir auf diese Weise $\eta_{19} = 158948$ und $\eta_{20} = -794734$, also offenbar ein oszillierendes, stark divergierendes Verhalten. Diese Instabilität wollen wir untersuchen. Zur Bestimmung der allgemeinen Lösung der

Gleichung

$$\eta_{i+1} + 4\eta_i - 5\eta_{i-1} = 0 \quad (3.11)$$

mit Startwerten $\eta_0 = c_0$ und $\eta_1 = c_1$ bestimmen wir zunächst die Nullstellen des *charakteristischen Polynoms*

$$\rho(\lambda) = \lambda^2 + 4\lambda - 5 = 0.$$

Diese sind $\lambda_1 = 1$ und $\lambda_2 = -5$. Man kann sich nun leicht überlegen, dass die allgemeine Lösung von (3.11) gegeben ist durch

$$\eta_i = \alpha\lambda_1^i + \beta\lambda_2^i = \alpha + \beta \cdot (-5)^i.$$

α und β sind eindeutig bestimmt durch die Startwerte c_0 und c_1 . Offenbar ist das instabile Verhalten des MSV dadurch verursacht, dass $\rho(\lambda)$ eine Nullstelle besitzt, deren Betrag größer als Eins ist.

Die Argumente des Beispiels [3.12] für die Lösung der homogenen Differenzgleichung lassen sich verallgemeinern. Es gilt

[3.13] Satz: Hat das Polynom r -ten Grades

$$\rho(\lambda) := a_r\lambda^r + \dots + a_1\lambda + a_0 \quad (3.12)$$

die r paarweise verschiedenen Nullstellen $\lambda_1, \dots, \lambda_r$, so lässt sich jede Lösungsfolge $(\eta_i)_{i \in \mathbb{N}}$ der **homogenen Differenzgleichung**

$$\sum_{s=0}^r a_s \eta_{i+s} = 0 \quad (3.13)$$

darstellen in der Form

$$\eta_i = \sum_{s=1}^r c_s \lambda_s^i.$$

Die Koeffizienten c_s sind eindeutig bestimmt durch die Startwerte $\eta_0, \dots, \eta_{r-1}$.

Wichtig für unsere Zwecke ist es, dass die Lösungen der homogenen Gleichung (3.13)

langsamer als linear bzgl. des Indexes anwachsen, dass also die **Wachstumsbeschränkung**

$$\lim_{i \rightarrow \infty} \frac{\eta_i}{i} = 0 \quad (3.14)$$

gilt. Hierzu gilt das folgende wichtige Resultat.

[3.14] Satz: Die Wachstumsbeschränkung (3.14) gilt genau dann für beliebige Startwerte $\eta_0, \dots, \eta_{r-1}$, wenn das folgende Stabilitätskriterium erfüllt ist.

Stabilitätskriterium: Das Polynom (3.12) besitzt nur Nullstellen λ_s mit Betrag $|\lambda_s| \leq 1$. Ist λ_s eine Nullstelle mit Betrag $|\lambda_s| = 1$, so ist λ_s *einfache* Nullstelle.

[3.15] Beispiele: (a) Das MSV des Beispiels [3.12] erfüllt nicht das Stabilitätskriterium des Satzes [3.14], da mit $\lambda_2 = -5$ eine Nullstelle gegeben ist mit Betrag größer als 1.

(b) Für das Beispiel [3.9](a) ist $\rho(\lambda)$ gegeben durch

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

Die Nullstellen sind 0, -1 und 1. Das Stabilitätskriterium ist damit erfüllt.

(c) Für das Beispiel [3.9](c) ist

$$\rho(\lambda) = \frac{11}{6}\lambda^3 - 3\lambda^2 + \frac{3}{2}\lambda - \frac{1}{3}.$$

Die Nullstellen sind $\lambda_1 = 1$ sowie $\lambda_{2/3} = 0.318182 \pm 0.283864i$. Das Stabilitätskriterium ist erfüllt.

3.4 Konsistenz und Konvergenz von MSV

Gegeben sei das lineare MSV

$$\sum_{s=0}^r a_s \eta_{i+s} = \Delta t \cdot \sum_{s=0}^r b_s f_{i+s}, \quad a_r = 1. \quad (3.15)$$

[3.16] **Definition:** Das **erste** und das **zweite charakteristische Polynom** des MSV (3.12) sind definiert durch

$$\rho(z) := \sum_{s=0}^r a_s z^s, \quad \sigma(z) := \sum_{s=0}^r b_s z^s. \quad (3.16)$$

[3.17] **Definition: (a)** Das MSV (3.15) heißt **konsistent**, falls es mindestens die Ordnung 1 hat, falls also (nach Satz [3.8]) gilt

$$\begin{aligned} c_0 &= a_0 + a_1 + \cdots + a_r = 0, \\ c_1 &= a_1 + 2a_2 + \cdots + ra_r - (b_0 + \cdots + b_r) = 0. \end{aligned}$$

Man überlegt sich leicht, dass diese Konsistenzbedingung äquivalent ist zur

Konsistenzbedingung:

$$\rho(1) = 0, \quad \rho'(1) - \sigma(1) = 0. \quad (3.17)$$

(b) Ähnlich wie bei ESV (vgl. Abschnitt 2.3) wird die Konvergenz von MSV definiert. Hierzu sei $x(\cdot)$ die exakte Lösung des AWP und $\eta_i^{(\Delta t)}$ sei die Lösung des MSV zur Schrittweite Δt . Das MSV heißt **konvergent**, wenn für beliebige $t > t_0$ und $N = 1, 2, 3, \dots$ zu den Schrittweiten $\Delta t = (t - t_0)/N$ gilt:

$$\lim_{N \rightarrow \infty} \eta_N^{(\Delta t)} = x(t).$$

Das folgende zentrale Ergebnis, nach welchem Konsistenz und Stabilität von MSV zur Konvergenz führen, geben wir zunächst ohne Beweis an.

[3.18] **Satz:** Ein MSV der Form (3.3) ist genau dann konvergent für beliebige AWP, wenn es konsistent ist und das Stabilitätskriterium des Satzes [3.14] erfüllt.

Wir kommen nun zur Fehleranalyse für gewisse explizite MSV. Wie früher bezeichne $x(t)$ die exakte Lösung des AWP

$$x' = f(t, x), \quad x(t_0) = x_0.$$

Es bezeichne

$$d_{j+1} := \frac{1}{\Delta t} \cdot \left(x(t_{j+1}) - \sum_{s=0}^{r-1} [-a_s x(t_{j+1-r+s}) + \Delta t \cdot b_s f(t_{j+1-r+s}, x(t_{j+1-r+s}))] \right)$$

den lokalen Diskretisierungsfehler des MSV beim Einsetzen der *exakten* Lösung.

[3.19] Satz: Für das explizite MSV gelte

$$a_s \leq 0, \quad s = 0, \dots, r-1, \quad a_r = 1.$$

Es sei

$$e_n := x(t_n) - \eta_n$$

der Fehler an der Stelle $t_n = t_0 + n \cdot \Delta t$. Definieren wir die Güte der Startwerte durch

$$G := \max_{0 \leq j \leq r-1} |x(t_j) - \eta_j|,$$

den maximalen Diskretisierungsfehler durch

$$D := \max_{j \geq r} |d_j|$$

und

$$B := \sum_{s=0}^{r-1} |b_s|,$$

so gilt die Abschätzung

$$|e_n| \leq \left(G + \frac{D}{LB} \right) \cdot \exp((t_n - t_0) \cdot LB).$$

(L ist die Lipschitz-Konstante für $f(\cdot, \cdot)$, vgl. Satz [1.8].)

Beweis: Aus den Gleichungen ($k := j + 1 - r$)

$$\eta_{j+1} = \sum_{s=0}^{r-1} [-a_s \eta_{k+s} + \Delta t \cdot f(t_{k+s}, \eta_{k+s})]$$

und

$$x(t_{j+1}) = \sum_{s=0}^{r-1} [-a_s x(t_{k+s}) + \Delta t \cdot b_s f(t_{k+s}, x(t_{k+s}))] + \Delta t \cdot d_{j+1}$$

folgt durch Subtraktion, Anwenden der Dreiecksungleichung und Ausnutzung der Lipschitz-Stetigkeit von f für $j \geq r - 1$

$$|e_{j+1}| \leq \sum_{s=0}^{r-1} (|a_s| + \Delta t \cdot |b_s| \cdot L) \cdot |e_{k+s}| + \Delta t \cdot |d_{j+1}|. \quad (3.18)$$

Für $s = 0, \dots, r - 1$ definieren wir

$$C_s := |a_s| + \Delta t \cdot L |b_s|;$$

außerdem sei

$$C := \sum_{s=0}^{r-1} C_s.$$

Aus der Konsistenzbedingung folgt $\sum_{s=0}^{r-1} a_s = -a_r = -1$ und damit

$$C = \sum_{s=0}^{r-1} |a_s| + \Delta t \cdot L \sum_{s=0}^{r-1} |b_s| = 1 + \Delta t \cdot LB \geq 1. \quad (3.19)$$

Eingesetzt in (3.18) folgt hieraus

$$\begin{aligned} |e_r| &\leq \sum_{s=0}^{r-1} C_s |e_s| + \Delta t \cdot D \leq G \cdot \sum_{s=0}^{r-1} C_s + \Delta t \cdot D = G \cdot C + \Delta t \cdot D, \\ |e_{r+1}| &\leq \sum_{s=0}^{r-1} C_s |e_{s+1}| + \Delta t \cdot D \leq G \sum_{s=0}^{r-2} C_s + C_{r-1} |e_r| + \Delta t \cdot D \\ &\leq G \left(\sum_{s=0}^{r-2} C_s + C \cdot C_{r-1} \right) + \Delta t \cdot D \cdot (C_{r-1} + 1) \leq GC^2 + \Delta t \cdot D \cdot (C + 1), \\ |e_{r+2}| &\leq \dots \leq GC^3 + \Delta t \cdot D \cdot (C^2 + C + 1) \end{aligned}$$

und allgemein (durch Induktion)

$$|e_{r+\ell}| \leq GC^{\ell+1} + \Delta t \cdot D \cdot \sum_{s=0}^{r+\ell-1} C^s.$$

Mit $n = r + \ell$ folgt wegen (3.19)

$$\begin{aligned} |e_n| &\leq G \cdot C^n + \Delta t \cdot D \sum_{s=0}^{n-1} C^s = G \cdot C^n + \Delta t \cdot D \cdot \frac{C^n - 1}{C - 1} \\ &= G \cdot (1 + \Delta t \cdot LB)^n + \frac{D}{LB} \cdot [(1 + \Delta t \cdot LB)^n - 1] \\ &\leq G \cdot \exp(n\Delta t LB) + \frac{D}{LB} [\exp(n\Delta t LB) - 1] \\ &= \left(G + \frac{D}{LB} \right) \cdot \exp(n\Delta t LB). \quad \square \end{aligned}$$

[3.20] Bemerkungen: (a) Eine ähnliche Abschätzung wie die des Satzes [3.19] kann auch für allgemeinere – insbesondere auch für implizite – MSV durchgeführt werden.

(b) Sind die ersten r Werte $\eta_0, \dots, \eta_{r-1}$ exakt gegeben, d.h. ist $G = 0$, und hat das MSV die Ordnung p , so ist $D \leq \text{const} \cdot \Delta t^p$ und es folgt (in Analogie zu Satz [2.12] für ESV)

$$|e_n| \leq \text{const} \cdot \frac{\Delta t^p}{LB} \cdot \exp((t_n - t_0) \cdot LB) = \mathcal{O}(\Delta t^p).$$

Sind die Startwerte nicht exakt bekannt, so müssen sie – um Verluste in der Ordnung zu vermeiden – mit einem (Einschritt-) Verfahren der Ordnung p ermittelt werden.

4 Stabilität von ESV und MSV

4.1 Absolute Stabilität

Während der Abschnitt 3.3 sich mit Instabilitäten befasste, welche durch die Eigen-
dynamik der linken Seiten von MSV entstanden, wollen wir jetzt Instabilitäten un-
tersuchen, welche mit den rechten Seiten der gDGl zusammenhängen. Die folgenden
Untersuchungen beziehen sich sowohl auf ESV als auch auf MSV.

[4.1] Beispiel: Wir betrachten das lineare Testproblem

$$x' = \lambda \cdot x, \quad x(0) = 1 \tag{4.1}$$

mit der exakten Lösung $x(t) = \exp(\lambda t)$ und wenden hierauf das **klassische RKV** an.
Wir erhalten

$$\begin{aligned} k_1 &= \lambda \eta_i, \\ k_2 &= \lambda \left(\eta_i + \frac{1}{2} \Delta t k_1 \right) = \left(\lambda + \frac{1}{2} \Delta t \lambda^2 \right) \eta_i, \\ k_3 &= \lambda \left(\eta_i + \frac{1}{2} \Delta t k_2 \right) = \left(\lambda + \frac{1}{2} \Delta t \lambda^2 + \frac{1}{4} \Delta t^2 \lambda^3 \right) \eta_i \\ k_4 &= \lambda (\eta_i + \Delta t k_3) = \left(\lambda + \Delta t \lambda^2 + \frac{1}{2} \Delta t^2 \lambda^3 + \frac{1}{4} \Delta t^3 \lambda^4 \right) \eta_i \end{aligned}$$

und damit

$$\eta_{i+1} = \eta_i + \frac{\Delta t}{6}(k_1 + 2k_2 + 2k_3 + k_4) = F(\lambda\Delta t) \cdot \eta_i$$

mit dem Polynom

$$F(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{3!}z^3 + \frac{1}{4!}z^4.$$

Ist $\lambda < 0$, so klingt die exakte Lösung $x(t)$ für $t \rightarrow \infty$ gegen Null ab. Dasselbe qualitative Verhalten für die numerische Lösung η_i erhalten wir nur, wenn Δt so gewählt ist, dass

$$|F(\lambda\Delta t)| < 1.$$

Wegen $|F(z)| \rightarrow \infty$ für $|z| \rightarrow \infty$ ist das aber nur möglich, wenn Δt hinreichend klein ist.

Dieses Beispiel motiviert die folgende Definition.

[4.2] Definition: (a) Führt ein ESV für das Testproblem (4.1) auf eine Rekursion der Form

$$\eta_{i+1} = F(\lambda\Delta t) \cdot \eta_i,$$

so heißt die Menge

$$B := \{z \in \mathbb{C} : |F(z)| < 1\}$$

das zum ESV gehörende **Gebiet der absoluten Stabilität**.

(b) Liegt die linke Halbebene

$$\mathbb{C}_- = \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$$

ganz in B , so heißt das Verfahren **A-stabil**.

Entsprechend dem Beispiel [4.1] sind Schrittweiten für das Testproblem so zu wählen, dass $\lambda\Delta t$ für λ mit $\operatorname{Re} \lambda < 0$ im Gebiet der absoluten Stabilität liegt. Für A-stabile Verfahren gibt es keine Schrittweitenbeschränkung.

[4.3] **Beispiele:** (a) Für das **explizite Euler-Verfahren** ist

$$\eta_{i+1} = \eta_i + \Delta t \cdot f(\eta_i) = (1 + \lambda \Delta t) \eta_i,$$

also

$$F(z) = 1 + z.$$

Das Gebiet der absoluten Stabilität ist (in der Gaußschen Zahlenebene) der Kreis um $\lambda_0 = -1$ mit Radius 1. Für $\operatorname{Re} \lambda < 0$ ist hier $\Delta t < 2/|\operatorname{Re} \lambda|$ zu wählen (vgl. Beispiel [1.12]).

(b) Für das **implizite Euler-Verfahren**

$$\eta_{i+1} = \eta_i + \Delta t \cdot f(t_{i+1}, \eta_{i+1}) = \eta_i + \lambda \Delta t \eta_{i+1}$$

ist $F(z) = 1/(1 - z)$. Für $\operatorname{Re} z < 0$ ist $|F(z)| \leq 1/(1 + |\operatorname{Re} z|) < 1$. Das Verfahren ist also A-stabil.

(c) Wir kombinieren die Verfahren (a) und (b) zu einem **Prädiktor-Korrektor-Verfahren** und definieren

$$\begin{aligned} \eta_{i+1}^{(P)} &= \eta_i + \Delta t \cdot f(t_i, \eta_i), \\ \eta_{i+1} &= \eta_i + \Delta t \cdot f(t_{i+1}, \eta_{i+1}^{(P)}). \end{aligned}$$

Einsetzen von $f(t, x) = \lambda x$ ergibt

$$\eta_{i+1} = (1 + \lambda \Delta t + (\lambda \Delta t)^2) \eta_i$$

und damit $F(z) = 1 + z + z^2$. Wegen $|F(z)| \rightarrow \infty$ für $|z| \rightarrow \infty$ ist das Verfahren nicht A-stabil.

[4.4] **Bemerkung:** Für explizite ESV ist $F(z)$ i.a. ein Polynom. Da Polynome für $|z| \rightarrow \infty$ divergieren, sind solche Verfahren nicht A-stabil. Für implizite Verfahren ist $F(z)$ i.a. eine gebrochene rationale Funktion.

Aus der Bemerkung folgt, dass z.B. das klassische RKV nicht A-stabil ist. Im nächsten Beispiel versuchen wir, die Idee der RKV auf implizite Verfahren zu verallgemeinern.

[4.5] **Beispiel:** Ein **zweistufiges implizites RKV** werde definiert durch

$$\begin{aligned} k_1 &= f\left(t_i + \frac{3 - \sqrt{3}}{6}\Delta t, \eta_i + \frac{1}{4}\Delta t \cdot k_1 + \frac{3 - 2\sqrt{3}}{12}\Delta t \cdot k_2\right), \\ k_2 &= f\left(t_i + \frac{3 + \sqrt{3}}{6}\Delta t, \eta_i + \frac{3 + 2\sqrt{3}}{12}\Delta t \cdot k_1 + \frac{1}{4}\Delta t \cdot k_2\right) \end{aligned}$$

sowie

$$\eta_{i+1} = \eta_i + \frac{\Delta t}{2} \cdot (k_1 + k_2).$$

(Warum ist dieses Verfahren implizit? Worin besteht die Verallgemeinerung gegenüber den in Abschnitt 2.2 definierten RKV?)

(i) *Berechnung von $F(z)$:* Mit $f(t, x) = \lambda x$ folgt das lineare Gleichungssystem für k_1 und k_2

$$\begin{aligned} k_1 &= \lambda \cdot \left(\eta_i + \frac{1}{4}\Delta t \cdot k_1 + \frac{3 - 2\sqrt{3}}{12}\Delta t \cdot k_2 \right), \\ k_2 &= \lambda \cdot \left(\eta_i + \frac{3 + 2\sqrt{3}}{12}\Delta t \cdot k_1 + \frac{1}{4}\Delta t \cdot k_2 \right) \end{aligned}$$

mit der Lösung

$$k_1 = \frac{1 - \frac{\sqrt{3}}{6}\lambda\Delta t}{1 - \frac{1}{2}\lambda\Delta t + \frac{1}{12}(\lambda\Delta t)^2} \cdot \lambda\eta_i, \quad k_2 = \frac{1 + \frac{\sqrt{3}}{6}\lambda\Delta t}{1 - \frac{1}{2}\lambda\Delta t + \frac{1}{12}(\lambda\Delta t)^2} \cdot \lambda\eta_i.$$

Hieraus folgt

$$F(z) = \frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{z}{2} + \frac{z^2}{12}}.$$

(ii) *Untersuchung auf A-Stabilität:* Das Gebiet der absoluten Stabilität wird begrenzt durch die Werte z mit $|F(z)| = 1$. Der Ansatz $F(z) = \exp(i\theta)$ ($\theta \in \mathbb{R}$) führt auf die quadratische Gleichung

$$\frac{z^2}{12} \cdot (1 - \exp(i\theta)) + \frac{z}{2} \cdot (1 + \exp(i\theta)) + (1 - \exp(i\theta))$$

und nach elementaren Umformungen auf

$$z^2 + 2iaz + 12 = 0 \quad \text{mit} \quad a = \frac{3 \sin(\theta)}{1 - \cos(\theta)} \in \mathbb{R}.$$

Die Lösungen

$$z = i \cdot \left(a \pm \sqrt{12 + a^2} \right)$$

liegen alle auf der imaginären Achse. Damit trennt die imaginäre Achse den Bereich der absoluten Stabilität von den Werten z mit $|F(z)| > 0$. Es folgt die A-Stabilität des Verfahrens.

In der Regel ist es sehr schwierig, den Bereich B der absoluten Stabilität explizit zu beschreiben. Einfacher ist die Untersuchung des **Stabilitätsintervalls** I_A , definiert als die Einschränkung des Bereichs auf *reellwertige negative* Werte (welche wir zur Unterscheidung mit ξ bezeichnen wollen), d.h.

$$I_A := B \cap \mathbb{R}_- = \{ \xi \in \mathbb{R} : \xi < 0 \text{ und } |F(\xi)| < 1 \}.$$

[4.6] **Beispiele:** (a) Für das **explizite** und das **implizite Euler-Verfahren** des Beispiels [4.3] folgen die Stabilitätsintervalle aus der Untersuchung des Gebiets der absoluten Stabilität als $I_A = (-2, 0)$ bzw. als $I_A = (-\infty, 0)$.

(b) Für das **Prädiktor-Korrektor-Euler-Verfahren** des Beispiels [4.3] ist

$$F(\xi) = 1 + \xi + \xi^2 = \left(\xi + \frac{1}{2} \right)^2 + \frac{3}{4}.$$

Dies ist eine nach oben geöffnete Parabel mit Minimalwert $3/4$. Damit ergeben sich die Grenzen des Stabilitätsintervalls aus der Lösung der quadratischen Gleichung $F(\xi) = 1$, und es folgt $I_A = (-1, 0)$. Damit ist das Stabilitätsintervall kleiner als das des expliziten Euler-Verfahrens!

(c) Für das **klassische RKV** ist

$$F(\xi) = 1 + \xi + \frac{1}{2}\xi^2 + \frac{1}{6}\xi^3 + \frac{1}{24}\xi^4.$$

Die Grenzen des Stabilitätsintervalls erhält man aus dem Ansatz $F(\xi) = 1$. Es folgt (aus numerischen Rechnungen, z.B. mit dem Newton-Verfahren) $I_A = (-2.7854, 0)$.

Übertragen wir nun diese Überlegungen auf MSV. Setzen wir das Testproblem $x' = \lambda x$

in das MSV (3.3) ein, so erhalten wir die homogene lineare Differenzgleichung

$$(1 - \lambda\Delta t \cdot b_r)\eta_{j+r} + \dots + (1 - \lambda\Delta t \cdot b_0)\eta_j = 0. \quad (4.2)$$

Zur Lösung bedienen wir uns – entsprechend den Überlegungen des Abschnitts 3.3 des mit (4.2) verknüpften Polynoms

$$\phi(z) = (1 - \lambda\Delta t \cdot b_r) \cdot z^r + \dots + (1 - \lambda\Delta t \cdot b_0) = \rho(z) - \lambda\Delta t \cdot \sigma(z), \quad (4.3)$$

wobei ρ und σ die in Definition [3.16] definierten charakteristischen Polynome des MSV sind. Hat ϕ r paarweise verschiedene Nullstellen z_1, \dots, z_r , so lässt sich nach Satz [3.13] die allgemeine Lösung der Differenzgleichung (4.2) darstellen in der Form

$$\eta_j = c_1 z_1^j + \dots + c_r z_r^j.$$

Offenbar ist das Abklingen der Folge η_i gegen Null genau dann gewährleistet, wenn die Beträge aller Nullstellen z_k kleiner als 1 sind. (Dasselbe gilt auch, wenn die Nullstellen nicht paarweise verschieden sind.) Dies führt zur folgenden Definition.

[4.7] Definition: Zu einem MSV heißt die Menge der Werte $\mu = \lambda\Delta t$, für welche das charakteristische Polynom

$$\phi(z) = \rho(z) - \mu \cdot \sigma(z)$$

nur Nullstellen z_i mit Beträgen $|z_i| < 1$ besitzt, das **Gebiet der absoluten Stabilität**.

Im Allgemeinen ist es schwierig, diese Gebiete analytisch exakt anzugeben. Zur numerisch-graphischen Darstellung wird man versuchen, den Rand des Stabilitätsgebiets zu bestimmen. Dieser ist dadurch gegeben, dass ϕ eine Nullstelle der Form $\exp(i\xi)$ ($\xi \in [0, 2\pi)$) hat. Der Rand ist also gegeben durch die Lösung der Gleichung

$$\rho(\exp(i\xi)) - \mu(\xi) \cdot \sigma(\exp(i\xi)) = 0.$$

Damit lässt sich das Gebiet der absoluten Stabilität eingrenzen durch die numerische Auswertung des Ausdrucks

$$\mu(\xi) := \rho(\exp(i\xi)) / \sigma(\exp(i\xi)), \quad \xi = 0 \dots 2\pi. \quad (4.4)$$

Der Wert $\mu(0)$ ist hierbei in der Regel bekannt. Nach einer der beiden Konsistenzbedingungen (3.17) gilt nämlich $\rho(1) = 0$. Ist außerdem $\rho'(1) \neq 0$, so folgt aus der zweiten Bedingung $\sigma(1) = \rho'(1) \neq 0$ und damit

$$\mu(0) = 0.$$

[4.8] Übung: Schreiben Sie ein Programm zur Berechnung der komplexwertigen Funktion $\mu(\cdot)$, definiert durch (4.4). Stellen Sie die Stabilitätsgebiete für einige der Adams-Bashforth-Verfahren graphisch dar.

4.2 Integration steifer Systeme

Die Überlegungen des vorherigen Abschnitts liefern Obergrenzen an die Schrittweite Δt , unter welchen Instabilitäten unterdrückt werden. In homogenen linearen Systemen bestimmt hierbei der Eigenwert λ_i mit dem betragsgrößten negativen Realteil die Schrittweite (vgl. hierzu die Merkregel [1.14]). Im Folgenden wollen wir an zwei Beispielen das Konzept einer Schrittweitensteuerung entwerfen, welche zu einer möglichst genauen Approximation unter möglichst geringem Rechenaufwand führt. Hierzu betrachten wir zunächst das skalare AWP

$$x' = \lambda x, \quad x(t_i) = \eta_i$$

mit der exakten Lösung

$$x(t) = \Phi^{t,t_i} \eta_i = \exp(\lambda \cdot (t - t_i)) \cdot \eta_i.$$

Die Anwendung eines ESV führt auf die Iterationsvorschrift

$$\eta_{i+1} := F(\lambda \Delta t) \cdot \eta_i$$

(vgl. Definition [4.2]). Fordern wir neben der Stabilitätsbedingung $|F(\lambda \Delta t)| < 1$, dass $F(\lambda \Delta t)$ eine gute Approximation von $\exp(\lambda \Delta t)$ ist, so führt das zu einer weiteren Einschränkung der Schrittweite.

[4.9] Beispiel: Wir betrachten das SysgDGl für $\mathbf{x} = (x_1, x_2, x_3)^T$,

$$\mathbf{x}' = A \cdot \mathbf{x}, \quad \mathbf{x}(0) = (4, 13, 1)^T,$$

mit

$$A = \begin{pmatrix} -0.5 & 32.6 & 35.7 \\ 0 & -48.0 & 9.0 \\ 0 & 9.0 & -72.0 \end{pmatrix}.$$

Mit den üblichen Methoden rechnet man (leicht) nach, dass A die drei reellen Eigenwerte $\lambda_1 = -0.5$, $\lambda_2 = -45$ und $\lambda_3 = -75$ besitzt. Die exakte Lösung setzt sich zusammen aus Anteilen, welche wie $\exp(\lambda_i \cdot t)$ gegen Null abklingen. Nach der Definition des Abschnitts 1.2 handelt es sich hier um ein steifes Differentialgleichungssystem.

Das gegebene AWP soll nun mit dem klassischen RKV gelöst werden. Dieses hat nach Beispiel [4.6](c) das Stabilitätsintervall $(-2.7854, 0)$. Hierdurch und durch den betragsgrößten negativen Eigenwert λ_3 wird die Stabilitätsobergrenze für die Schrittweite Δt festgelegt durch

$$\Delta t_{max} = 2.7854/|\lambda_3| = 0.037. \quad (4.5)$$

Diese Grenze muss im Folgenden immer berücksichtigt werden. Wir fordern nun des weiteren, dass

$$|F(\lambda_i \cdot \Delta t) - \exp(\lambda_i \cdot \Delta t)| \leq \epsilon := 10^{-5}, \quad i = 1, 2, 3. \quad (4.6)$$

Dies führt (nach numerischer Auswertung der linken Seite von (4.6)) auf die Wahl $\Delta t_1 = 0.0035$. Nach 45 Zeitschritten ($t = 0.158$) sind die wie $\exp(\lambda_3 \cdot t)$ abklingenden Anteile um den Faktor $\exp(\lambda_3 \cdot 0.158) = 7.14 \cdot 10^{-6}$ reduziert und werden bei der Schrittweitenwahl nicht mehr berücksichtigt. Die nächste Zeitschrittwahl Δt_2 ergibt sich aus der Forderung (4.6) für $i = 1, 2$ und führt auf $\Delta t_2 = 0.005$. Nach weiteren 20 Zeitschritten können auch die wie $\exp(\lambda_2 t)$ abklingenden Komponenten vernachlässigt werden; die Forderung (4.5) für $i = 1$ führt auf den Zeitschritt $\Delta t_3 = 0.5$. Da dieser nicht mit der Stabilitätsforderung (4.5) im Einklang steht (die wie $\exp(\lambda_3 t)$ abklingenden Komponenten könnten zu Instabilitäten führen), wird als letzter Zeitschritt z.B. $\Delta t_3 = 0.035$ festgelegt.

Wie können die vorangegangenen Überlegungen auf allgemeinere – also insbesondere inhomogene nichtlineare – Systeme erweitert werden? Dies soll an einem Beispiel demonstriert werden.

[4.10] Beispiel: Die folgenden Gleichungen beschreiben ein System, wie es typischerweise bei der Berechnung der chemischen Reaktion dreier Substanzen auftreten kann.

$$\mathbf{x}' = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}' = \begin{pmatrix} -0.1x_1 + 100x_2x_3 \\ 0.1x_1 - 100x_2x_3 - 500x_2^2 \\ 500x_2^2 - 0.5x_3 \end{pmatrix}$$

mit den Anfangsbedingungen $\mathbf{x}(0) = (4, 2, 0.5)^T$. Um herauszufinden, mit welcher Schrittweite mit der numerischen Approximation begonnen werden soll, untersuchen wir die **Linearisierung** des Systems um den Anfangsvektor $\mathbf{x}(0)$. Hierzu setzen wir $\Delta\mathbf{x}(t) := \mathbf{x}(t) - \mathbf{x}(0)$ und vernachlässigen Terme, welche quadratisch in den Komponenten von $\Delta\mathbf{x}$ sind. Dies führt auf das linearisierte System

$$\Delta\mathbf{x}' = A \cdot \Delta\mathbf{x} + \begin{pmatrix} 99.6 \\ -2099.6 \\ 1999.75 \end{pmatrix}$$

mit

$$A = \begin{pmatrix} -0.1 & 100 \cdot x_3(0) & 100 \cdot x_2(0) \\ 0.1 & -100 \cdot x_3(0) - 1000 \cdot x_2(0) & -100 \cdot x_2(0) \\ 0 & 1000 \cdot x_2(0) & -0.5 \end{pmatrix} = \begin{pmatrix} -0.1 & 50 & 200 \\ 0.1 & -2050 & -200 \\ 0 & 2000 & -0.5 \end{pmatrix}.$$

Die Eigenwerte von A sind $\lambda_1 = -0.00025$, $\lambda_2 = -219.06$ und $\lambda_3 = -1831.54$. Für das klassische RKV ergibt sich hieraus eine Stabilitätsgrenze $\Delta t_{max} = 2.7854/1831.54 = 0.0015$. Ein Maß für die Genauigkeit für diese Schrittweite ist

$|F(\lambda_3 \Delta t_{max}) - \exp(\lambda_3 \Delta t_{max})| = 0.88$. Soll dieser Wert auf etwa 10^{-5} verringert werden, so ist eine Schrittweite $\Delta t = 0.00015$ nötig. Wegen der Nichtlinearität des Gleichungssystems muss eine Linearisierung zur Schrittweitenbestimmung jeweils nach einigen Zeitschritten wiederholt werden.

5 Numerik von Anfangswertproblemen: Ergänzungen

5.1 Schrittweitensteuerung

Bei der Wahl einer geeigneten Schrittweite h muss ein Kompromiss gefunden werden zwischen hoher Rechengenauigkeit und moderatem Rechenaufwand. Hinzu kommt, dass sich die "ideale" Schrittweite im Verlauf der Integration des Anfangswertproblems ändern kann. Das Ziel ist eine Schrittweitensteuerung, welche im Verlauf der Rechnungen die Schrittweiten korrigiert und optimiert. Hierzu ist es nötig, den lokalen Diskretisierungsfehler zu messen. Dies geschieht durch die Auswertung eines geeigneten *Kontrollverfahrens*, welches eine höhere Ordnung als das eingesetzte Verfahren hat.

Seien ein ESV und ein Kontrollverfahren gegeben durch

$$\eta_{i+t} = \eta_i + h \cdot \Psi(t, \eta_i, h), \quad (5.1)$$

$$\hat{\eta}_{i+t} = \eta_i + h \cdot \hat{\Psi}(t, \eta_i, h). \quad (5.2)$$

Sei $z(t) := \phi^{t, t_i} \eta_i$. Haben das ESV die Ordnung p und das Kontrollverfahren die Ordnung $p + 1$, so ist

$$\|\eta_{i+1} - \hat{\eta}_{i+1}\| \leq \|\eta_{i+1} - z(t_{i+1})\| + \|\hat{\eta}_{i+1} - z(t_{i+1})\| \leq C \cdot h^p + \hat{C} \cdot h^{p+1} \approx C \cdot h^p \quad (5.3)$$

und daher

$$\|\eta_{i+1} - \hat{\eta}_{i+1}\| \approx \|\eta_{i+1} - z(t_{i+1})\| = \epsilon(h, t_i, \eta_i). \quad (5.4)$$

Dies ermöglicht die Schätzung des lokalen Diskretisierungsfehlers mit Hilfe des Kontrollverfahrens.

Der Einsatz von Kontrollverfahren sollte möglichst ohne zusätzlichen Rechenaufwand erfolgen. Ist beispielsweise η_{i+1} gegeben durch ein s -stufiges Runge-Kutta-Verfahren (nach Berechnung der Steigungen k_1, \dots, k_s), so sollte das Kontrollverfahren die Werte k_1, \dots, k_s einbeziehen und ansonsten ohne weitere Werte auskommen. Da aber Runge-Kutta-Verfahren in der Regel so konstruiert sind, dass ohne weitere Funktionsauswertungen eine Ordnungssteigerung nicht möglich ist, bedient man sich z.B. des *Fehlberg-Tricks*: Man bezieht den Wert

$$\eta_{i+1} = \eta_i + h \cdot \sum_{i=1}^s b_i k_i \quad (5.5)$$

mit ein, welcher ja ohnehin (bei akzeptierter Schrittweite) berechnet werden muss.

[5.1] Beispiel: Benutzt werden soll das klassische RKV als Kontrollverfahren (also Ordnung $p = 4$) zur Steuerung eines Verfahrens dritter Ordnung. Das klassische Verfahren ist gegeben durch

$$\mathbf{b} = \frac{1}{6} \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \\ 1 \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad (5.6)$$

mit der Iterationsvorschrift

$$\eta_{i+1} = \eta_i + \frac{h}{6} \cdot (k_1 + 2k_2 + 2k_3 + k_4). \quad (5.7)$$

Versuch 1: Verfahren 3. Ordnung mit neuen Gewichten $\hat{b}_1, \dots, \hat{b}_4$. Aus den Ordnungsbedingungen ergibt sich nach Satz [2.9] das Gleichungssystem

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1/2 & 1/2 & 1 \\ 0 & 1/4 & 1/4 & 1 \\ 0 & 0 & 1/4 & 1/2 \end{pmatrix} \cdot \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{b}_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/6 \end{pmatrix}. \quad (5.8)$$

Dieses Gleichungssystem ist regulär und hat die eindeutige Lösung \mathbf{b} .

Versuch 2 (Fehlberg-Trick): Wir definieren

$$k_5 = f(t_i + h, \eta_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)) \quad (5.9)$$

und führen die Gewichte $\hat{\mathbf{b}} = (\hat{b}_1, \dots, \hat{b}_5)^T$ ein. Das Gleichungssystem für $\hat{\mathbf{b}}$, welches auf die Ordnung 3 führt, lautet nun

$$\underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1/2 & 1/2 & 1 & 1 \\ 0 & 1/4 & 1/4 & 1 & 1 \\ 0 & 0 & 1/4 & 1/2 & 1/2 \end{pmatrix}}_{\tilde{A}} \cdot \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{b}_4 \\ \hat{b}_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/6 \end{pmatrix}. \quad (5.10)$$

Eine spezielle Lösung ist

$$\hat{\mathbf{b}} = \frac{1}{6}(1, 2, 2, 1, 0)^T, \quad (5.11)$$

und es ist

$$\mathbf{ker}(\tilde{A}) = \text{span}(0, 0, 0, 1, -1)^T. \quad (5.12)$$

Ein Verfahren 3. Ordnung ist damit gegeben z.B. durch

$$\hat{\mathbf{b}} = \frac{1}{6}(1, 2, 2, 0, 1)^T. \quad (5.13)$$

6 Randwertprobleme

6.1 Einführung

Wie die bisher behandelten AWP werden die folgenden Randwertprobleme (RWP) beschrieben durch ein System von Differentialgleichungen $\mathbf{x}' = \mathbf{f}(t, \mathbf{x})$. Im Gegensatz zu jenen sind Zusatzbedingungen zur eindeutigen Bestimmung der Lösung jedoch nicht an einem einzigen Punkt ("Anfangswert") sondern an mehreren Punkten (hier: an zwei Punkten) gegeben.

[6.1] Beispiele: (a) Ein dünner wärmeleitender Stab sei zwischen zwei Wänden eingespannt. Die linke Wand ($x = 0$) habe die Temperatur T_0 , die rechte Wand ($x = L$) die Temperatur T_L . Der Temperaturverlauf $T(x)$ im Stab ist gegeben durch das RWP

$$T''(x) = 0, \quad T(0) = T_0, \quad T(L) = T_L.$$

Aus $T'' = 0$ folgt der allgemeine Ansatz $T(x) = Ax + B$ und hieraus die exakte Lösung

$$T(x) = T_0 + \frac{T_L - T_0}{L} \cdot x.$$

Durch den Ansatz $\Theta := (\theta_1, \theta_2)^T$, mit $\theta_1 = T$, $\theta_2 = T'$ kann das RWP zweiter Ordnung umformuliert werden in das System erster Ordnung

$$\Theta' = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \Theta, \quad \theta_1(0) = T_0, \quad \theta_1(L) = T_L.$$

(b) Ein dünner Balken sei an den Enden $x = 0$ und $x = L$ fest eingespannt. Zwischen $x = 0$ und $x = L$ wirke ein Kraftfeld \mathcal{F} auf den Balken (z.B. das Graviationsfeld). Die Verbiegung $y(x)$ ist gegeben durch das RWP

$$p \cdot y^{(4)} - q \cdot y'' + r \cdot y = \mathcal{F}, \quad y(0) = 0, \quad y'(0) = 0, \quad y(L) = 0, \quad y'(L) = 0.$$

Hierbei sind p , q und r Materialkonstanten. Mit Hilfe von $\Gamma := (\gamma_1, \gamma_2, \gamma_3, \gamma_4)^T$, $\gamma_1 = y$, $\gamma_2 = y' = \gamma_1'$, $\gamma_3 = y''$ und $\gamma_4 = y'''$ folgt das RWP

$$\Gamma' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -r & 0 & q & 0 \end{pmatrix} \cdot \Gamma + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mathcal{F} \end{pmatrix}, \quad \gamma_1(0) = \gamma_2(0) = \gamma_1(L) = \gamma_2(L) = 0.$$

RWP haben im allgemeinen keine eindeutige Lösung, wie das folgende Beispiel zeigt.

[6.2] Beispiel: Wir untersuchen das RWP

$$y'' = -y, \quad y(0) = y_0, \quad y(L) = y_L.$$

Die allgemeine Lösung von $y'' = -y$ lautet $y(x) = A \cdot \sin(x) + B \cdot \cos(x)$.

(a) Für die Randbedingungen $y(0) = 0$, $y(\pi/2) = 1$ existiert genau eine Lösung $y(x) = \sin(x)$.

(b) Für die Randbedingungen $y(0) = y(\pi) = 0$ existieren unendlich viele Lösungen $y(x) = c \cdot \sin(x)$, $c \in \mathbb{R}$.

(c) Für die Randbedingungen $y(0) = 0$ und $y(\pi) = 1$ existiert keine Lösung.

Randwertprobleme für vektorwertige Funktionen $\mathbf{y}(x) = (y_1(x), \dots, y_n(x))^T$, wie wir sie im Folgenden untersuchen werden, lassen sich formulieren durch ein System gewöhnlicher Differentialgleichungen,

$$\mathbf{y}'(x) = \frac{d}{dx}\mathbf{y}(x) = \mathbf{f}(x, \mathbf{y}), \quad (6.1)$$

sowie durch Randbedingungen an die Lösung \mathbf{y} in zwei Punkten $x = a$ und $x = b$. Damit erwartet werden kann, dass das RWP eindeutig lösbar ist, ist die Anzahl der Randbedingungen gleich der Dimension n des Systems. Die folgenden Typen von Randbedingungen tauchen in der Praxis häufig auf.

(a) Lineare Randbedingungen: Vorgeschrieben werden n Linearkombinationen von Komponenten von \mathbf{y} :

$$\sum_{j=1}^n \alpha_{ij} y_j(a) = \gamma_i, \quad i = 1, \dots, k, \quad \sum_{j=1}^n \beta_{ij} y_j(b) = \gamma_i, \quad i = k+1, \dots, n.$$

Mit einer geeigneten $k \times n$ -Matrix \tilde{A} und einer $(n-k) \times n$ -Matrix \tilde{B} können diese Randbedingungen geschrieben werden als $\tilde{A} \cdot \mathbf{y}(a) = (\gamma_1, \dots, \gamma_k)^T$ und $\tilde{B} \cdot \mathbf{y}(b) = (\gamma_{k+1}, \dots, \gamma_n)^T$. Werden die Matrizen \tilde{A} und \tilde{B} auf "triviale" Weise (d.h. durch Einfügen von Nullzeilen) zu $n \times n$ -Matrizen ergänzt:

$$A := \begin{pmatrix} \tilde{A} \\ 0 \end{pmatrix}, \quad B := \begin{pmatrix} 0 \\ \tilde{B} \end{pmatrix},$$

so lassen sich die Randbedingungen schreiben in der kompakten Form

$$A \cdot \mathbf{y}(a) + B \cdot \mathbf{y}(b) = \mathbf{c} \quad (6.2)$$

mit $\mathbf{c} = (\gamma_1, \dots, \gamma_n)^T$.

(b) Separierte lineare Randbedingungen: Diese sind ein Spezialfall von (a), bei denen die ersten k Komponenten von \mathbf{y} am linken Rand $x = a$ und die letzten $n - k$ Komponenten am rechten Rand $x = b$ vorgegeben sind. In diesem Fall haben die Matrizen A und B die Blockstruktur

$$A = \left(\begin{array}{c|c} \tilde{A} & 0 \\ \hline 0 & 0 \end{array} \right), \quad B = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \tilde{B} \end{array} \right),$$

wobei \tilde{A} eine $k \times k$ -Matrix und \tilde{B} eine $(n - k) \times (n - k)$ -Matrix ist.

(c) Nichtlineare Randbedingungen: Diese sind gegeben durch eine vektorwertige Funktion $r : \mathbb{R}^{2n} \rightarrow \mathbb{R}^n$ und lauten

$$r(\mathbf{y}(a), \mathbf{y}(b)) = 0. \quad (6.3)$$

6.2 Das einfache Schießverfahren

A – Ein System zweiter Ordnung

Wir betrachten das RWP auf dem Intervall $[0, 1]$ für $\mathbf{w} = (w_1, \dots, w_n)^T$,

$$\mathbf{w}'' = \mathbf{f}(x, \mathbf{w}, \mathbf{w}'), \quad \mathbf{w}(0) = \mathbf{w}_0, \quad \mathbf{w}(1) = \mathbf{w}_1.$$

Mit Hilfe von $\mathbf{y} = (y_1, \dots, y_{2n})^T$, $y_k := w_k$, $y_{n+k} := w'_k$ kann das System umgewandelt werden in das folgende System mit linearen Randbedingungen

$$\mathbf{y}' = \begin{pmatrix} \mathbf{y}_{1/2}^{(1)} \\ \mathbf{f}(x, \mathbf{y}) \end{pmatrix}, \quad \mathbf{y}_{1/2}^{(0)}(0) = \mathbf{w}_0, \quad \mathbf{y}_{1/2}^{(0)}(1) = \mathbf{w}_1, \quad (6.4)$$

wobei zur Abkürzung $\mathbf{y}_{1/2}^{(0)} = (y_1, \dots, y_n)^T = \mathbf{w}$ und $\mathbf{y}_{1/2}^{(1)} = (y_{n+1}, \dots, y_{2n})^T = \mathbf{w}'$ gesetzt wurde.

Die Idee zur Lösung des RWP ist folgendermaßen. Wenn an Stelle der Randbedingung

$\mathbf{y}_{1/2}^{(0)}(1) = \mathbf{w}_1$ an der Stelle $x = 1$ eine Randbedingung der Form $\mathbf{y}_{1/2}^{(1)}(0) = \tilde{\mathbf{w}}_0$ vorgegeben wäre, so würde es sich bei dem RWP um ein AWP handeln, welches mit den bisher behandelten Methoden numerisch gelöst werden könnte. Man könnte also versuchen, durch "Versuch und Irrtum" die fehlende Information $\mathbf{y}_{1/2}^{(1)}(0)$ zu erraten.

Lösen wir das System $\mathbf{y}' = (\mathbf{y}_{1/2}^{(1)}, \mathbf{f}(x, \mathbf{y}))^T$ mit den Anfangswerten $\mathbf{y}_{1/2}^{(0)} = \mathbf{w}_0$ und $\mathbf{y}_{1/2}^{(1)} = \mathbf{z}$ für $\mathbf{z} \in \mathbb{R}^n$, so erhalten als Lösung des AWP

$$\mathbf{y}(x) =: \mathbf{y}(x, \mathbf{z}) = \Phi^{x,0} \begin{pmatrix} \mathbf{w}_0 \\ \mathbf{z} \end{pmatrix}.$$

Der ergänzte Anfangswert \mathbf{z} ist dem RWP (5.4) angemessen gewählt, falls die Lösung die Randbedingung am rechten Rand $\mathbf{y}_{1/2}^{(0)}(1, \mathbf{z}) = \mathbf{w}_1$ erfüllt, falls also \mathbf{z} Nullstelle der Funktion

$$\mathbf{F}(\mathbf{z}) = \mathbf{y}_{1/2}^{(0)}(1, \mathbf{z}) - \mathbf{w}_1$$

ist. Um eine Nullstelle von \mathbf{F} zu finden, könnte (zumindest bei hinreichend gutem Startvektor $\mathbf{z}^{(0)}$) das Newton-Verfahren angewendet werden⁷. Eine Schwierigkeit hierbei ist es, die Jacobi-Matrix $\mathbf{DF}(\mathbf{z})$ zu bestimmen. Dies kann jedoch numerisch durchgeführt werden. Man überlegt sich leicht, dass die j -te Spalte $\mathbf{DF}(\mathbf{z})^{(j)}$ für kleine h gegeben ist durch

$$\mathbf{DF}(\mathbf{z})^{(j)} \approx \frac{1}{h} (\mathbf{F}(\mathbf{z} + h \cdot \mathbf{e}_j) - \mathbf{F}(\mathbf{z})),$$

wobei \mathbf{e}_j der j -te kanonische Einheitsvektor ist.

[6.3] Beispiel: Gelöst werden soll das skalare RWP zweiter Ordnung

$$w'' = 1 + ww', \quad w(0) = w(1) = 0.$$

Dies wird zunächst umgeformt in das äquivalente System für $\mathbf{y} = (y_1, y_2)^T = (w, w')^T$,

$$\mathbf{y}' = \begin{pmatrix} y_2 \\ 1 + y_1 y_2 \end{pmatrix}, \quad y_1(0) = y_1(1) = 0.$$

⁷vgl. Abschnitt 3.2 der Vorlesung Numerische Mathematik für Informatiker, WS 99/00

Wir installieren zunächst das klassische RKV mit Schrittweite $\Delta x = 0.01$ für das AWP

$$\mathbf{y}' = \begin{pmatrix} y_2 \\ 1 + y_1 y_2 \end{pmatrix}, \quad y_1(0) = 0, \quad y_2(0) = \xi.$$

Um die Abhängigkeit der Lösung vom Startwert zu demonstrieren, bezeichnen wir die Lösung mit $\mathbf{y}(\cdot, \xi)$. Die Aufgabe ist es, ξ so zu bestimmen, dass die zweite Randbedingung $F(\xi) := y_1(1, \xi) = 0$ erfüllt ist. Erste numerische Experimente zeigen, dass $F(0) = 0.5267$ und $F(-1) = -0.4417$. Es ist daher zu erwarten, dass $F(\cdot)$ im Intervall $(-1, 0)$ eine Nullstelle besitzt. Indem wir das AWP nacheinander für ξ und für $\xi + \Delta\xi$ ($\Delta\xi = 0.01$) lösen, erhalten wir eine Approximation der Ableitung von $F(\cdot)$:

$$F'(\xi) \approx dF(\xi) := \frac{F(\xi + \Delta\xi) - F(\xi)}{\Delta\xi}.$$

Das folgende Iterationsschema liefert nach wenigen Iterationsschritten die Nullstelle ξ_0 mit hoher Genauigkeit:

- S1 Wähle Startwert $\xi^{(0)} := -1$;
- S2 Ist $\xi^{(k)}$ gegeben, so löse das AWP nacheinander für ξ und für $\xi + \Delta\xi$ und berechne $dF(\xi)$;
- S3 Definiere $\xi^{(k+1)} := \xi^{(k)} - F(\xi)/dF(\xi)$.

Mit zehnstelliger Genauigkeit erhalten wir nach 5 Schritten den Wert $\xi_0 = -0.5041659279$ sowie die zugehörige Lösung des RWP.

Wir wenden uns nun der prinzipiellen Lösbarkeit des *skalaren* RWP

$$w'' = f(x, w, w') \quad , \quad w(0) = w(1) = 0 \tag{6.5}$$

mit Hilfe des Schieß- und des Newton-Verfahrens zu. Die Lösung des AWP

$$v'' = f(x, v, v') \quad , \quad v(0) = 0 \quad , \quad v'(0) = \xi \tag{6.6}$$

bezeichnen wir mit v_ξ . Zudem definieren wir in Analogie zu oben $F(\xi) := v_\xi(1)$. Definitionsbereich $D(F)$ von F ist die Menge aller ξ , für die v_ξ für das gesamte Intervall $[0, 1]$ existiert. v_ξ ist Lösung des RWP, falls $F(\xi) = 0$.

[6.4] **Satz:** Die Funktion f sei stetig differenzierbar bezüglich w und w' mit den partiellen Ableitungen f_w und $f_{w'}$. Dann ist F differenzierbar mit $F'(\xi) = u_\xi(1)$, wobei u_ξ die Lösung des AWP

$$u'' = f_w(x, v_\xi, v'_\xi)u + f_{w'}(x, v_\xi, v'_\xi)u' \quad , \quad u(0) = 0 \quad , \quad u'(0) = 1 \quad (6.7)$$

ist.⁸

Beweis: Für $\xi \neq \alpha$ seien v_ξ und v_α die Lösungen des AWP (5.6) mit $v'_\xi(0) = \xi$ und $v'_\alpha(0) = \alpha$. Wir definieren

$$u_\alpha(x) := \frac{v_\alpha(x) - v_\xi(x)}{\alpha - \xi} \quad . \quad (6.8)$$

Dann ist offenbar

$$\frac{F(\alpha) - F(\xi)}{\alpha - \xi} = \frac{v_\alpha(1) - v_\xi(1)}{\alpha - \xi} = u_\alpha(1) \quad , \quad (6.9)$$

und durch Grenzübergang $\alpha \rightarrow \xi$ folgt

$$F'(\xi) = \lim_{\alpha \rightarrow \xi} u_\alpha(1) \quad . \quad (6.10)$$

u_α erfüllt offenbar die Randbedingungen

$$u_\alpha(0) = 0 \quad , \quad u'_\alpha(0) = 1 \quad . \quad (6.11)$$

Außerdem gilt für festes $x \in (0, 1)$

$$(\alpha - \xi)u''_\alpha = v''_\alpha - v''_\xi = f(x, v_\alpha, v'_\alpha) - f(x, v_\xi, v'_\xi) = \phi(1) - \phi(0) \quad (6.12)$$

mit

$$\phi(t) = f\left(x, v_\xi + t(v_\alpha - v_\xi), v'_\xi + t(v'_\alpha - v'_\xi)\right) \quad . \quad (6.13)$$

Mit $v := v_\xi + t(v_\alpha - v_\xi)$ folgt

$$\begin{aligned} \phi'(t) &= f_w(x, v, v')(v_\alpha - v_\xi) + f_{w'}(x, v, v')(v'_\alpha - v'_\xi) \\ &= (\alpha - \xi)f_w(x, v, v')u_\alpha + f_{w'}(x, v, v')u'_\alpha \end{aligned} \quad (6.14)$$

⁸Man beachte, dass dies ein lineares AWP und damit eindeutig lösbar ist.

sowie oben eingesetzt

$$\begin{aligned} u''_\alpha &= \frac{\phi(1) - \phi(0)}{\alpha - \xi} = \frac{1}{\alpha - \xi} \int_0^1 \phi'(t) dt \\ &= \left(\int_0^1 f_w(x, v, v') dt \right) u_\alpha + \left(\int_0^1 f_{w'}(x, v, v') dt \right) u'_\alpha . \end{aligned} \quad (6.15)$$

Für $\alpha \rightarrow \xi$ streben v und v' gleichmäßig gegen v_ξ und v'_ξ , und für die Integrale gilt

$$\int_0^1 f_w(x, v, v') dt \rightarrow f_w(x, v_\xi, v'_\xi) \quad , \quad \int_0^1 f_{w'}(x, v, v') dt \rightarrow f_{w'}(x, v_\xi, v'_\xi) \quad . \quad (6.16)$$

Aus der Theorie gewöhnlicher Differentialgleichungen folgt, dass u_α gegen die Lösung des AWP (5.7) konvergiert. \circ

Wir setzen unseres frühere Beispiel fort.

[5.3] Beispiel (Fortsetzung): Es ist $f(w, w') = 1 + ww'$ und damit $f_w = w'$ und $f_{w'} = w$. Laut Satz ist das AWP

$$v'' = 1 + vv' \quad , \quad v(0) = 0 \quad , \quad v'(0) = \xi \quad (6.17)$$

zu ergänzen durch das AWP

$$u'' = v'u + vu' \quad , \quad u(0) = 0 \quad , \quad u'(0) = 1 \quad . \quad (6.18)$$

Dann ist $F(\xi) = v(1)$ und $F'(\xi) = u(1)$; das Newton-Verfahren liefert damit als Verbesserung den neuen Wert

$$\tilde{\xi} = \xi - v(1)/u(1) \quad . \quad (6.19)$$

Die Differentialgleichungen für v und u sind gekoppelt, da die rechte Seite für u'' auch von v abhängt. Sie sollten daher als ein einziges System behandelt werden. Hierzu definieren wir

$$\mathbf{y} = (y_1, y_2, y_3, y_4)^T := (v, v', u, u')^T \quad . \quad (6.20)$$

Zu lösen ist das AWP

$$\mathbf{y}' = (y_2, 1 + y_1y_2, y_4, y_2y_3 + y_1y_4)^T \quad , \quad \mathbf{y}(0) = (0, \xi, 0, a)^T \quad . \quad (6.21)$$

k	ξ	$v(1)$	$u(1)$
0	0	0.5266687	1.115790
1	-0.472014	0.03101425	1.0083094
2	-0.502772664	0.00133740045	1.004480724
3	-0.504104098	0.00005933825	1.00432274670

Tabelle 6: Newton-Verfahren für Beispiel [5.3]

Wir demonstrieren die Ergebnisse eines Iterationsverfahrens. Startwert ist $\xi = 0$. Formel (5.19) gibt an, wie ξ nach dem Newton-Verfahren verbessert werden. Tabelle 6 gibt die relevanten Werte der Iterationen an.

Eine Schwierigkeit bei der Durchführung des Schießverfahrens ist es, ein geeignetes "Fenster" für die ergänzenden Anfangswerte zu finden, für die eine Lösung des AWP existiert und innerhalb dessen die Lösung des RWP gefunden werden kann. Beispielsweise kann leicht gezeigt werden, dass das AWP (5.17) für keinen Wert ξ eine globale Lösung (d.h. eine Lösung im gesamten Intervall \mathbb{R}_+) besitzt. Eine obere Schranke für diejenigen Werte, bei denen das AWP eine Lösung im Intervall $[0,1]$ hat, ist etwa bei $\xi_{\text{sup}} = 4.8$ gegeben.

B – Ein lineares RWP zweiter Ordnung

Die Überlegungen des vorherigen Abschnitts vereinfachen sich im Fall linearer SysgDGL. Wir demonstrieren dies am folgenden skalaren linearen RWP zweiter Ordnung

$$w''(x) = p(x) + q(x)w(x) + r(x)w'(x), \quad w(0) = \alpha, \quad w(1) = \beta, \quad (6.22)$$

welches wie vorher umgewandelt werden kann in das System

$$\mathbf{y}' = \begin{pmatrix} 0 & 1 \\ q & r \end{pmatrix} \cdot \mathbf{y} + \begin{pmatrix} 0 \\ p \end{pmatrix} \quad (6.23)$$

für $\mathbf{y} = (y_1, y_2)^T = (w, w')^T$ mit den Randbedingungen $y_1(0) = \alpha$ und $y_1(1) = \beta$.

Wegen der Linearität der Gleichung (5.23) sind konvexe Linearkombinationen von Lö-

sungen des SysgDGI wieder Lösungen des Systems: Seien w_1, w_2 Lösungen von (5.22) und $\lambda_1, \lambda_2 \in \mathbb{R}$ mit $\lambda_1 + \lambda_2 = 1$. Man rechnet leicht nach, dass dann auch $w = \lambda_1 w_1 + \lambda_2 w_2$ Lösung von (5.23) ist. Erfüllen w_1 und w_2 die Anfangsbedingung $w_1(0) = w_2(0) = \alpha$, so gilt auch $w(0) = \alpha$.

Wir können zwei Lösungen dazu verwenden, eine Lösung des Randwertproblems zu konstruieren. Zunächst konstruieren wir die Lösung $\overset{\circ}{\mathbf{y}} = (\overset{\circ}{y}_1, \overset{\circ}{y}_2)^T$ des Systems (5.23) zu den Anfangswerten $\overset{\circ}{\mathbf{y}}(0) = (\alpha, 0)^T$. Schließlich berechnen wir noch eine weitere Lösung $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2)^T$ von (5.23) zum Anfangswert $\bar{\mathbf{y}}(0) = (\alpha, 1)^T$. Die konvexe Linearkombination $\mathbf{y} = \lambda \cdot \overset{\circ}{\mathbf{y}} + (1 - \lambda) \cdot \bar{\mathbf{y}}$ ist Lösung von (5.23) und erfüllt die Randbedingungen $y_1(0) = \alpha$ und $y_1(1) = \lambda \overset{\circ}{y}_1(1) + (1 - \lambda)\bar{y}_1(1)$. Ist $\bar{y}_1(1) \neq \overset{\circ}{y}_1(1)$, so ist mit $\lambda := (\beta - \bar{y}_1(1)) / (\overset{\circ}{y}_1(1) - \bar{y}_1(1))$ die Lösung des RWP (5.22) gefunden. Wir fassen die Schritte im folgenden Algorithmus zusammen.

[6.5] Algorithmus zur Lösung des linearen RWP:

- S1 Initialisierung: Wähle ein ESV (z.B. klassisches RKV) und (mit einem hinreichend großen N) eine Schrittweite $\Delta t = 1/N$.
- S2 Löse numerisch das AWP (5.6), $\overset{\circ}{\mathbf{y}}(0) = (\alpha, 0)^T$; ist $\overset{\circ}{y}_1(1) = \beta$, so ist $\overset{\circ}{\mathbf{y}}$ eine Lösung des RWP; andernfalls:
- S3 Löse numerisch das AWP (5.6), $\bar{\mathbf{y}}(0) = (\alpha, 1)^T$.
- S4 Ist $\bar{y}_1(1) \neq \overset{\circ}{y}_1(1)$, so berechne die Lösung des RWP:

$$\mathbf{y} := \frac{\beta - \bar{y}_1(1)}{\overset{\circ}{y}_1(1) - \bar{y}_1(1)} \cdot \overset{\circ}{\mathbf{y}} - \frac{\beta - \overset{\circ}{y}_1(1)}{\overset{\circ}{y}_1(1) - \bar{y}_1(1)} \cdot \bar{\mathbf{y}};$$

andernfalls Ausgabe: das RWP hat keine Lösung.

Ein Nachteil des einfachen Schießverfahrens ergibt sich aus dem folgenden Beispiel.

[6.6] Beispiel: Gegeben sei das lineare RWP

$$-u'' + c^2 u = c^2 x, \quad u(0) = 0, \quad u(1) = 0 \quad (6.24)$$

mit einer Konstanten $c \in \mathbb{R}$.

Transformation $w = u - x$ führt auf das das homogene Problem

$$-w'' + c^2 w = 0, \quad w(0) = 0, \quad w(1) = -1. \quad (6.25)$$

mit der Lösung

$$w(x) = \xi_1 \exp(cx) + \xi_2 \exp(-cx) \quad (6.26)$$

mit geeigneten Koeffizienten ξ_1, ξ_2 . Es folgt

$$u(x) = x + \xi_1 \exp(cx) + \xi_2 \exp(-cx). \quad (6.27)$$

Die Anwendung des einfachen Schießverfahrens führt auf AWP's der Form

$$-v_\alpha'' + c^2 v_\alpha = c^2 x, \quad v_\alpha(0) = 0, \quad v_\alpha'(0) = \alpha \quad (6.28)$$

mit Lösungen der Form

$$v_\alpha(x) = x + \eta_1 \exp(cx) + \eta_2 \exp(-cx). \quad (6.29)$$

Einsetzen der Anfangsbedingungen führt auf die Lösung

$$v_\alpha(x) = x + \frac{\alpha - 1}{2c} (\exp(cx) - \exp(-cx)). \quad (6.30)$$

Zur Lösung des ursprünglichen RWP's ist eine Nullstelle des Funktionals

$$F(\alpha) = v_\alpha(1) = \frac{\alpha - 1}{c} \sinh(c) \quad (6.31)$$

zu finden. Für $c \gg 1$ ist $\sinh(c) \approx 0.5 \cdot \exp(c)$. Die numerische Berechnung kann leicht auf Floating point-Fehlermeldungen führen.

C – Mehrzielverfahren

Der bedeutendste Nachteil des einfachen Schießverfahrens besteht darin, dass – teilweise mit großem Aufwand – ein "Fenster" gesucht werden muss, innerhalb dessen Lösungen des AWP im gesamten Intervall $[0,1]$ existieren. Einen Ausweg bieten hier die Mehrzielverfahren, bei denen das Gesamtintervall $[0,1]$ in mehrere Teilintervalle zerlegt wird. In jedem linken Randpunkt eines Teilintervalls wird das AWP neu gestartet. Zu

lösen ist dabei ein Nullstellenproblem, welches nicht nur dafür sorgt, dass die Randbedingung für $x = 1$ korrekt erfüllt ist, sondern auch, dass die Funktion in den inneren Teilintervall-Randpunkten einmal stetig differenzierbar ist.

Wir betrachten wieder das RWP

$$w'' = f(x, w, w') \quad , \quad w(0) = w(1) = 0$$

und führen auf dem Intervall $[0,1]$ das folgende Gitter ein:

$$\Delta = \{x_0, x_1, \dots, x_n\} \quad \text{mit} \quad 0 = x_0 < x_1 < \dots < x_n = 1 \quad .$$

Auf den Intervallen $[x_{i-1}, x_i]$ ($i = 1, \dots, n$) werden folgende AWP's gelöst.

$$v_i'' = f(x, v_i, v_i') \quad , \quad v_i(x_{i-1}) = \eta_{i-1} \quad , \quad v_i'(x_{i-1}) = \alpha_{i-1} \quad .$$

Durch die Randbedingungen für w vorgeschrieben sind die Werte $\eta_0 = w(0)$ sowie $v_n(x_n) = w(1) = 0$. Wir definieren den Vektor

$$\mathbf{z} = (\alpha_0, \eta_1, \alpha_1, \dots, \eta_{n-1}, \alpha_{n-1})^T \in \mathbb{R}^{2n-1}$$

sowie die Abbildung $F : \mathbb{R}^{2n-1} \rightarrow \mathbb{R}^{2n-1}$,

$$F(\mathbf{z}) = \begin{pmatrix} v_1(x_1) - \eta_1 \\ v_1'(x_1) - \alpha_1 \\ \vdots \\ v_{n-1}(x_{n-1}) - \eta_{n-1} \\ v_{n-1}'(x_{n-1}) - \alpha_{n-1} \\ v_n(1) \end{pmatrix} .$$

AWP

$$w'' = 1 + ww', \quad w(x_0) = \eta, \quad w'(x_0) = \alpha. \quad (6.32)$$

Um die Abhängigkeit von den Anfangswerten zu dokumentieren, schreiben wir $w(x) = w(x, \eta, \alpha)$. Außerdem definieren wir

$$U(x) := \partial_\eta w(x), \quad u(x) := \partial_\alpha w(x). \quad (6.33)$$

Für diese Funktionen können wir leicht die folgenden AWP's herleiten.

$$U'' = U \cdot w' + w \cdot U', \quad U(x_0) = 1, \quad U'(x_0) = 0, \quad (6.34)$$

$$u'' = u \cdot w' + w \cdot u', \quad u(x_0) = 0, \quad u'(x_0) = 1. \quad (6.35)$$

Wir wollen nun auf das RWP

$$w'' = 1 + ww', \quad w(0) = w(1) = 0 \quad (6.36)$$

das Mehrzielverfahren anwenden, wobei das Intervall $[0, 1]$ in zwei gleich große Teilintervalle zerlegt wird: $x_0 = 0, x_1 = 0.5, x_2 = 1$. Definieren wir $x_0 := 0$ und $x_1 := 0.5$, so sind folgende AWP's zu lösen.

$$w_i'' = 1 + w_i w_i', \quad i = 0, 1 \quad (6.37)$$

mit den Zusatzbedingungen

$$w_0(0) = 0, \quad w_1(1) = 0, \quad w_0(0.5) = w_1(0.5) =: \eta_1, \quad w_0'(0.5) = w_1'(0.5) =: \alpha_1. \quad (6.38)$$

Unbekannt sind zunächst die Größen $\alpha_0 := w'(0)$ sowie η_1 und α_1 . Zur Lösung des Zweiziel-Verfahrens definieren wir die beiden AWP's

$$\frac{\partial}{\partial x} \begin{pmatrix} w_0 \\ w_0' \\ u_0 \\ u_0' \end{pmatrix} = \begin{pmatrix} w_0' \\ 1 + w_0 w_0' \\ u_0' \\ u_0 w_0' + w_0 u_0 \end{pmatrix}, \quad \frac{\partial}{\partial x} \begin{pmatrix} w_1 \\ w_1' \\ U_1 \\ u_1 \\ U_1' \\ u_1' \end{pmatrix} = \begin{pmatrix} w_1' \\ 1 + w_1 w_1' \\ U_1' \\ u_1' \\ U_1 w_1' + w_1 U_1' \\ u_1 w_1' + w_1 u_1' \end{pmatrix} \quad (6.39)$$

mit

$$(w_0(0), w'_0(0), u_0(0), u'_0(0)) = (0, \alpha_0, 0, 1), \quad (6.40)$$

$$(w_1(0.5), w'_1(0.5), U_1(0.5), u_1(0.5), U'_1(0.5), u'_1(0.5)) = (\eta_1, \alpha_1, 1, 0, 0, 1). \quad (6.41)$$

Zur Anwendung des Newton-Verfahrens definieren wir den Vektor \mathbf{z} und die Abbildung $F(\mathbf{z})$ durch

$$\mathbf{z} := \begin{pmatrix} \alpha_0 \\ \eta_1 \\ \alpha_1 \end{pmatrix}, \quad F(\mathbf{z}) = \begin{pmatrix} w_0(0.5) - \eta_1 \\ w'_0(0.5) - \alpha_1 \\ w_1(1) \end{pmatrix}. \quad (6.42)$$

Die Jacobi-Matrix von F ist gegeben durch

$$DF(\mathbf{z}) = \begin{pmatrix} u_0(0.5) & -1 & 0 \\ u'_0(0.5) & 0 & -1 \\ 0 & U_1(1) & u_1(1) \end{pmatrix} \quad (6.43)$$

6.3 Differenzenverfahren

Im Folgenden untersuchen wir das RWP auf dem Intervall $[0, 1]$,

$$-w''(x) + b(x)w'(x) + c(x)w(x) = f(x) \quad , w(0) = w(1) = 0 \quad , \quad (6.44)$$

wobei b , c und f stetige Funktionen sind, und

$$\forall x \in [0, 1] : c(x) \geq 1 \quad .$$

Unter diesen Voraussetzungen besitzt das RWP (5.24) eine eindeutige Lösung⁹.

Wir wollen nun die Differentialgleichung nicht wie früher in ein System erster Ordnung umwandeln, sondern mit Hilfe geeigneter Differenzenquotienten diskretisieren. Hierzu zerlegen wir zunächst das Intervall $[0, 1]$ in n gleiche Teilintervalle und definieren den zugehörigen Diskretisierungsparameter $h := 1/n$. Zur Abkürzung schreiben wir $x_j := j \cdot h$, $w_j := w(x_j)$, $w''_j := w''(x_j)$ ($j = 0, \dots, n$) etc. Diskrete Approximationen der ersten

⁹vgl. H. Heuser: Gewöhnliche Differentialgleichungen, Kapitel VI, Teubner 1992, Stuttgart.

und zweiten Ableitungen von w erhalten wir leicht aus der Taylorentwicklung. Für die erste Ableitung definieren wir die Approximationen

$$\begin{aligned} D_h^+[w](x) &= \frac{w(x+h) - w(x)}{h} \quad , \\ D_h^-[w](x) &= \frac{w(x) - w(x-h)}{h} \quad , \\ D_h[w](x) &= \frac{w(x+h) - w(x-h)}{2h} \quad , \end{aligned}$$

welche wir auch als *rechtsseitigen*, *linksseitigen* bzw. *zentralen Differenzenquotienten* bezeichnen. Eine Approximation für die zweite Ableitung erhalten wir durch den (zentralen) Differenzenquotienten

$$D_h^2[w](x) = \frac{D_h^+[w](x) - D_h^-[w](x)}{h} = \frac{w(x+h) - 2w(x) + w(x-h)}{h^2} \quad .$$

Setzen wir die Approximationen

$$\begin{aligned} w'_i &\approx \frac{w_{i+1} - w_{i-1}}{2h} \quad , \\ w''_i &\approx \frac{w_{i+1} - 2w_i + w_{i-1}}{h^2} \end{aligned}$$

in (5.24) ein, so führt dies mit den Randbedingungen auf ein lineares Gleichungssystem für $\mathbf{w} = (w_1, \dots, w_n)^T$ mit der Inhomogenität $\mathbf{f} = (f_1, \dots, f_n)^T$.

[6.8] Beispiel: Im einfachsten Fall $b = c \equiv 0$ lautet dies

$$h^{-2} \underbrace{\begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}}_{=:L_h} \mathbf{w} = \mathbf{f} \quad . \quad (6.45)$$

Aus der Taylorformel folgen die folgenden Abschätzungen.

[6.9] Lemma: (a) Für $u \in \mathcal{C}^2[0, 1]$ gilt

$$|D_h^\pm[u](x) - u'(x)| \leq \frac{1}{2} \|u''\|_\infty h \quad (6.46)$$

(b) Für $u \in \mathcal{C}^3[0, 1]$ gilt

$$|D_h[u](x) - u'(x)| \leq \frac{1}{6} \|u'''\|_\infty h^2 \quad (6.47)$$

(c) Für $u \in \mathcal{C}^4[0, 1]$ gilt

$$|D_h^2[u](x) - u''(x)| \leq \frac{1}{12} \|u^{(4)}\|_\infty h^2 \quad (6.48)$$

Die Diskretisierung des Randwertproblems (6.44) führt auf das Gleichungssystem $L_h \mathbf{w} = \mathbf{f}$ mit einer Matrix der Form

$$L_h = \begin{pmatrix} d_1 & s_1 & & 0 \\ r_2 & d_2 & \ddots & \\ & \ddots & \ddots & s_{n-2} \\ 0 & & r_{n-1} & d_{n-1} \end{pmatrix} \quad (6.49)$$

Die Koeffizienten d_i , r_i und s_i hängen von der Wahl der Approximation der ersten Ableitung ab. Beispielsweise gilt bei der Wahl des zentralen Differenzenquotienten

$$d_i = 2 + h^2 c(x_i), \quad r_i = -1 - hb(x_i)/2, \quad s_i = -1 + hb(x_i)/2. \quad (6.50)$$

Ein Gütekriterium für Differenzenverfahren lässt sich wie früher mit Hilfe der Konsistenzordnung formulieren.

[6.10] Definition: Das Differenzenverfahren hat die *Konsistenzordnung* q , wenn für jede hinreichend glatte Lösung \mathbf{u} des Randwertproblems gilt

$$\|L_h \mathbf{u} - \mathbf{f}\|_\infty \leq C \cdot h^q. \quad (6.51)$$

Durch Vergleich mit der Taylor-Entwicklung ergibt sich unmittelbar

[6.11] Satz: Die Lösung u des Randwertproblems sei viermal stetig differenzierbar. Das Differenzenverfahren hat die Konvergenzordnung $q = 2$ bei Verwendung des zentralen, und $q = 1$ bei Verwendung des links- oder rechtsseitigen Differenzenquotienten.

Die oben hergeleiteten Matrizen L_h haben (mit einer geeigneten positiven Diagonalmatrix D) die Form

$$L_h = -D(I - B), \quad (6.52)$$

wobei die Störung B nur nichtnegative Koeffizienten (für h hinreichend klein) und eine Zeilensummennorm ≤ 1 hat. Man kann zeigen, dass L_h invertierbar ist und die Inverse gegeben ist durch die *Neumann-Reihe*

$$L_h^{-1} = -(I + B + B^2 + \dots) \cdot D^{-1} \quad .$$

Im allgemeineren Rahmen fügt sich L_h in die wohlbekanntene Theorie der M -Matrizen.

Die Matrix L_h hat für hinreichend kleinen Diskretisierungsparameter h die folgenden Kennzeichen:

- L_h ist Tridiagonalmatrix;
- die Diagonalelemente sind strikt negativ, die Nebendiagonalelemente strikt positiv;
- $L_h = (\ell_{ij})$ ist schwach diagonaldominant, d.h. für alle Zeilen gilt $|\ell_{ii}| \geq \sum_{j \neq i} |\ell_{ij}|$, und es gibt mindestens eine Zeile i_0 , für welche gilt $|\ell_{i_0 i_0}| > \sum_{j \neq i_0} |\ell_{i_0 j}|$.

Damit ist L_h ein Spezialfall von Matrizen, welche in der Literatur als M -Matrizen bezeichnet werden. Die Inversen solcher Matrizen sind durch Neumann-Reihen darstellbar. Das Gleichungssystem (5.25) ist also eindeutig lösbar, z.B. mit dem Jacobi-Iterationsverfahren, aber auch mittels Gauß-Elimination (ohne Pivotisierung); für sehr kleine h wird allerdings die Kondition von L_h sehr schlecht, weshalb in der Praxis geeignetere Varianten verwendet werden.

Auf M -Matrizen stößt man auch bei der Diskretisierung sog. *partieller Differentialgleichungen* vom elliptischen Typ.

7 Differentiell-algebraische Systeme

7.1 Einführung

Wir betrachten AWP's, welche kombiniert sind mit algebraischen Gleichungen.

[7.1] Definition: (a) Ein *differentiell-algebraisches System* (kurz: *DA-System*) für die Funktionen $\mathbf{y}(t) \in \mathbb{R}^d$ und $\mathbf{z}(t) \in \mathbb{R}^p$ ist ein Gleichungssystem der Form

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{y}(0) = \mathbf{y}_0, \quad (7.1)$$

$$0 = \mathbf{g}(\mathbf{y}, \mathbf{z}), \quad \mathbf{z}(0) = \mathbf{z}_0 \quad (7.2)$$

mit $\mathbf{f} : \mathbb{R}^{d+p} \rightarrow \mathbb{R}^d$ und $\mathbf{g} : \mathbb{R}^{d+p} \rightarrow \mathbb{R}^p$. Die Anfangswerte heißen *konsistent*, falls gilt

$$\mathbf{g}(\mathbf{y}_0, \mathbf{z}_0) = 0. \quad (7.3)$$

(Dies sei in Folgenden immer vorausgesetzt!)

(b) Das System hat *Index Eins*, falls die Ableitungsmatrix

$$\frac{\partial}{\partial \mathbf{z}} \mathbf{g}(\mathbf{y}_0, \mathbf{z}_0) \in \mathbb{R}^{p \times p} \quad (7.4)$$

regulär ist, andernfalls einen *Index größer als Eins*.

[7.2] Beispiele: (a) Die Modellierung eines bestimmten Verstärkerschaltkreises¹⁰ mit Hilfe der Kirchhoffschen Gesetze führt auf ein Gleichungssystem der Form

$$0 = -i_B + (u_0 - u_B)/r_1 + (u_S - u_B)'c_1 - u_B/r_2 \quad (7.5)$$

$$0 = (u_L - u_C)'c_2 + (u_0 - u_C)/r_C - i_C \quad (7.6)$$

$$0 = -u_E'c_E + i_B + i_C - u_E/r_E \quad (7.7)$$

$$0 = -u_L/r_L + (u_C - u_L)'c_2. \quad (7.8)$$

Durch Zusammenfassung der unbekanntenen Größen zu einem Vektor:

$$\mathbf{x} = (u_B, u_C, u_E, u_L)^T \quad (7.9)$$

¹⁰vgl. M. Hanke-Boureois: Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens, Abschnitt 64.1

erhalten wir ein System der Form

$$M\mathbf{x}' = \phi(\mathbf{x}) \quad (7.10)$$

mit der singulären Matrix

$$M = \begin{pmatrix} c_1 & & & \\ & c_2 & -c_2 & \\ & & c_E & \\ & -c_2 & & c_2 \end{pmatrix}. \quad (7.11)$$

Multiplikation mit der Matrix

$$P = \begin{pmatrix} 1 & & & \\ & 1 & -1 & \\ & & 1 & \\ & -1 & & 1 \end{pmatrix} \quad (7.12)$$

führt auf das DA-System

$$c_1 u'_B = \phi_1 \quad (7.13)$$

$$2c_2 \cdot (u_C - u_L)' = \phi_2 - \phi_4 \quad (7.14)$$

$$c_E u'_E = \phi_3 \quad (7.15)$$

$$0 = \phi_2 + \phi_4 = -u_C/r_C - u_L/r_L + \xi \quad (7.16)$$

für die vektorwertige Funktion $\mathbf{y} = (u_B, u_C - u_L, u_E)^T$ und die skalare Funktion $z = u_C/r_C + u_L/r_L$. Wegen $\partial_z g(z) = -1$ hat das System den Index 1.

(b) Ein Schlitten bewege sich unter der Kraft $F(t, \mathbf{x}, \mathbf{x}')$ (Gravitations-, Reibungskräfte etc.) einen Berg hinab. Das Profil des Berges sei implizit als Niveaufläche einer Funktion $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ gegeben:

$$g(\mathbf{x}) = g(x_1, x_2, x_3) = 0. \quad (7.17)$$

Die Bahn des Schlittens wird beschrieben durch die Newton-Gleichung

$$m\mathbf{x}'' = \text{Gesamtkraft} = F + Z. \quad (7.18)$$

Hierbei ist Z eine Zwangskraft, welche den Schlitten auf der Erdoberfläche hält. Diese Zwangskraft wirkt immer senkrecht auf die Oberfläche:

$$Z = \lambda \cdot \nabla_{\mathbf{x}} g, \quad (7.19)$$

wobei $\lambda = \lambda(t)$ unbekannt ist. Damit ergeben sich als DA-System für $(\mathbf{x}, \lambda)^T$ die sog. *Euler-Lagrange-Gleichungen*

$$m\mathbf{x}'' = F(t, \mathbf{x}, \mathbf{x}') + \lambda \cdot \nabla_{\mathbf{x}} g(\mathbf{x}) \quad (7.20)$$

$$0 = g(\mathbf{x}). \quad (7.21)$$

Da $g(\cdot)$ nicht explizit von λ abhängt, hat das System einen Index > 1 .

[7.3] Bemerkung: Index-1-Probleme sind (bei konsistenten Anfangswerten) zumindest lokal (d.h. für kleine Zeitintervalle) lösbar. Wegen der Regularität von $\partial_{\mathbf{z}} \mathbf{g}$ kann nämlich nach dem Satz über implizite Funktionen die Bedingung $\mathbf{g}(\mathbf{y}, \mathbf{z}) = 0$ lokal umformuliert werden in $\mathbf{z} = \phi(\mathbf{y})$ mit einer geeigneten Funktion ϕ . Für ϕ gilt in der Nähe von \mathbf{y}_0 die Approximation

$$\phi(\mathbf{y}) \approx \mathbf{z}_0 - (\nabla_{\mathbf{z}} \mathbf{g}(\mathbf{y}_0, \mathbf{z}_0))^{-1} \cdot (\nabla_{\mathbf{y}} \mathbf{g}(\mathbf{y}_0, \mathbf{z}_0)) \cdot (\mathbf{y} - \mathbf{y}_0). \quad (7.22)$$

(Wie lautet ϕ in Beispiel [7.2](a)?) Das DA-System ist damit lokal äquivalent zu

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \phi(\mathbf{y})), \quad \mathbf{y}(0) = \mathbf{y}_0. \quad (7.23)$$

7.2 Runge-Kutta-Verfahren für Index-1-Systeme

Wir betrachten zunächst eine Modifikation des DA-Systems, welche durch Einführung einer kleinen Störung ϵ auf das folgende Anfangswertproblem führt.

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{y}(0) = \mathbf{y}_0, \quad (7.24)$$

$$\epsilon \mathbf{z}' = \mathbf{g}(\mathbf{y}, \mathbf{z}), \quad \mathbf{z}(0) = \mathbf{z}_0 \quad (7.25)$$

Die Lösung dieses Systems für die vektorwertige Funktion $(\mathbf{y}, \mathbf{z})(t)$ mit Hilfe von (impliziten) Runge-Kutta-Verfahren soll nun untersucht werden. Beschrieben seien dies Wie früher durch die Matrix $\mathcal{A} = (a_{ij})$ sowie den Vektor \mathbf{b} . Der Vektor \mathbf{c} spielt keine

Rolle, da es sich um ein autonomes System handelt. Bezeichnen wir mit $(\eta_i, \zeta_i)^T$ die numerische Näherung von $(\mathbf{y}, \mathbf{z})(t_i)$ und mit $\mathbf{k}_\ell = (\lambda_\ell, \mu_\ell)^T$ die Steigungen, welche beim Runge-Kutta-Verfahren berechnet werden müssen, so folgt

$$\lambda_\ell = \mathbf{f} \left(\eta_i + h \sum_{j=1}^s a_{\ell j} \lambda_j, \zeta_i + h \sum_{j=1}^s a_{\ell j} \mu_j \right) \quad (7.26)$$

$$\epsilon \mu_\ell = \mathbf{g} \left(\eta_i + h \sum_{j=1}^s a_{\ell j} \lambda_j, \zeta_i + h \sum_{j=1}^s a_{\ell j} \mu_j \right) \quad (7.27)$$

sowie für den nächsten Zeitschritt

$$\eta_{i+1} = \eta_i + h \sum_{\ell=1}^s b_\ell \lambda_\ell \quad (7.28)$$

$$\epsilon \zeta_{i+1} = \epsilon \zeta_i + h \sum_{\ell=1}^s b_\ell \mu_\ell \quad (7.29)$$

Wir wollen nun versuchen, in Gleichung (7.27) den Grenzübergang $\epsilon \rightarrow 0$ zu vollziehen. Hierzu stellen wir zunächst fest, dass für glatte \mathbf{g} und kleine h

$$\mathbf{g}(\tilde{\eta}, \tilde{\zeta} + h\tilde{\mu}) \approx \mathbf{g}(\tilde{\eta}, \tilde{\zeta}) + h \cdot D_\zeta \mathbf{g}(\tilde{\eta}, \tilde{\zeta}) \cdot \tilde{\mu}. \quad (7.30)$$

Setzen wir

$$\mu := (\mu_1 \dots \mu_s) \in \mathbb{R}^{p \times s}, \quad (7.31)$$

so führt die Linearisierung von (7.29) für $\epsilon \rightarrow 0$ auf das lineare Gleichungssystem

$$h \cdot \mathcal{A} \cdot (D\mathbf{g}(\tilde{\eta}_i, \zeta_i) \cdot \mu)^T = - \underbrace{(\mathbf{g}(\tilde{\eta}_i, \zeta_i), \dots, \mathbf{g}(\tilde{\eta}_i, \zeta_i))^T}_{s\text{-mal}} \quad (7.32)$$

mit

$$\tilde{\eta}_i = \eta_i + h \sum_{j=1}^s a_{\ell j} \lambda_j. \quad (7.33)$$

Damit dieses System nach μ auflösbar ist, müssen wir fordern, dass \mathcal{A} invertierbar ist. Dies schließt insbesondere explizite Runge-Kutta-Verfahren aus. Andererseits folgt für invertierbares \mathcal{A} aus der Index-1-Eigenschaft und dem Satz über implizite Funktionen, dass (lokal) das System (7.26), (7.27) auflösbar ist. Genauere Analysen ergeben¹¹

¹¹vgl. Satz 82.1 in M. Hanke-Bourgeois

[7.4] **Satz:** Voraussetzungen: (i) \mathbf{f} und \mathbf{g} seien hinreichend glatt; $D_\zeta \mathbf{g}$ sei überall invertierbar und die Inverse $(D_\zeta \mathbf{g})^{-1}$ sei gleichmäßig beschränkt.

(ii) Das Runge-Kutta-Verfahren $(\mathcal{A}, \mathbf{b}, \mathbf{c})$ habe die Ordnung q ; \mathcal{A} sei invertierbar. Weiterhin sei $a_{sj} = b_j$ für $j = 1, \dots, s$. (RKV's mit diesen Eigenschaften heißen *steifgenau*.) Dann gilt für $T > 0$: Für hinreichend kleine h ist das Verfahren wohldefiniert; es ist

$$\|\mathbf{y}(t_i) - \eta_i\| \leq \mathcal{O}(h^q), \quad \|\mathbf{z}(t_i) - \zeta_i\| \leq \mathcal{O}(h^q) \quad \text{für alle } t_i = ih \in [0, T]. \quad (7.34)$$

[7.5] **Bemerkung:** Für nicht steifgenaue Runge-Kutta-Verfahren bleibt die Ordnung q erhalten, wenn die Gleichung $\mathbf{g}(\mathbf{y}_i, \mathbf{z}_i) = 0$ ersetzt wird durch $\mathbf{g}(\mathbf{y}_{i+1}, \mathbf{z}_{i+1}) = 0$.

7.3 Systeme mit Index > 1

Wir beschränken uns hier auf die Untersuchung des Systems (vgl. Beispiel (7.2)(b)) für $\mathbf{z} = (\mathbf{x}, \mathbf{v}, \lambda) = (\mathbf{x}, \mathbf{x}', \lambda)$

$$m\mathbf{v}' = F + \ell \cdot \nabla_{\mathbf{x}} \mathbf{g} \quad (7.35)$$

$$\mathbf{x}' = \mathbf{v} \quad (7.36)$$

$$0 = \mathbf{g}(\mathbf{x}), \quad (7.37)$$

wobei wir zur Vereinfachung anstelle des dreidimensionalen Raums nur zwei Dimensionen annehmen, also

$$\mathbf{x} =: (x, y)^T, \quad \mathbf{v} =: (v, w)^T, \quad (7.38)$$

und F konstant sowie $m = 1$ gewählt wird. Damit wird die rechte Seite

$$\mathbf{f}(\mathbf{x}, \mathbf{v}) = \left(v \quad w \quad F + \ell \cdot \partial_x g(x, y), F + \ell \cdot \partial_y g(x, y) \right)^T. \quad (7.39)$$

Zur Anwendung kommt das einfachste Runge-Kutta-Verfahren, welches die Voraussetzungen des Satzes erfüllt – das implizite Euler-Verfahren

$$\mathbf{k}_0 = \mathbf{f}(\eta_i + h \cdot \mathbf{k}_0), \quad (7.40)$$

$$0 = \mathbf{g}(\eta_i + h \cdot \mathbf{k}_0), \quad (7.41)$$

$$\eta_{i+1} = \eta_i + h \cdot \mathbf{k}_0. \quad (7.42)$$

Wir können dies umformulieren zu

$$\eta_{i+1} = \eta_i + h \cdot \eta_{i+1} \quad (7.43)$$

$$0 = g(\eta_{i+1}). \quad (7.44)$$

Ausführlich lauten die Gleichungen

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h \cdot \mathbf{v}_{i+1} \quad (7.45)$$

$$\mathbf{v}_{i+1} = \mathbf{v}_i + h \ell_{i+1} \cdot \nabla_{\mathbf{x}} g(\mathbf{x}_{i+1}) \quad (7.46)$$

$$0 = g(\mathbf{x}_{i+1}). \quad (7.47)$$

Führen wir die folgenden Linearisierungen ein:

$$g(\mathbf{x}_{i+1}) \approx g(\mathbf{x}_i) + h \cdot \nabla_{\mathbf{x}}^T g(\mathbf{x}_i) \cdot (\mathbf{x}_{i+1} - \mathbf{x}_i) \quad (7.48)$$

$$\nabla_{\mathbf{x}} g(\mathbf{x}_{i+1}) \approx \nabla_{\mathbf{x}} g(\mathbf{x}_i) + h \cdot g''(\mathbf{x}_i) \cdot (\mathbf{x}_{i+1} - \mathbf{x}_i) \quad (7.49)$$

sowie die Approximation

$$\ell_{i+1} - \ell_i = \mathcal{O}(h), \quad (7.50)$$

so erhalten wir für $\mathbf{z}_{i+1} = (\mathbf{x}_{i+1}, \mathbf{v}_{i+1}, \ell_{i+1})^T$ ein lineares Gleichungssystem der Form

$$J \cdot \mathbf{z}_{i+1} = \mathbf{r}_i \quad (7.51)$$

mit der Funktionalmatrix

$$J = \begin{pmatrix} I & -hI & 0 \\ -h^2 \ell_i g''(\mathbf{x}_i) & I & -h \nabla_{\mathbf{x}} g(\mathbf{x}_i) \\ h \nabla_{\mathbf{x}}^T g(\mathbf{x}_i) & 0 & 0 \end{pmatrix} \quad (7.52)$$

Wir erkennen, dass J für kleine h fast singulär ist, was eine numerische Lösung problematisch macht. (Warum?) Daneben lässt sich auch zeigen, dass das Verfahren nicht mehr die Konsistenzordnung 1 hat. Daher verwerfen wir diesen Ansatz und suchen eine Alternative.

Ein *halbexplizites* Euler-Verfahren kann wie folgt konstruiert werden.

Schritt 1: Zunächst wird durch ein explizites Verfahren ein Prädiktor für \mathbf{x}_{i+1} bestimmt gemäß

$$\mathbf{x}_{i+1}^P = \mathbf{x}_i + h \cdot \mathbf{v}_i. \quad (7.53)$$

Schritt 2: Mit dem Ansatz

$$\mathbf{x}_{i+1} := \mathbf{x}_{i+1}^P + \alpha \cdot \nabla g(\mathbf{x}_{i+1}^P) \quad (7.54)$$

wird das (eindimensionale) Nullstellenproblem

$$g(\mathbf{x}_{i+1}) = 0 \quad (7.55)$$

für α gelöst.

Schritt 3: Eine Gleichung für \mathbf{v}_{i+1} erhalten wir aus dem impliziten Ansatz

$$\mathbf{v}_{i+1} = \mathbf{v}_i + h \cdot (F + \lambda_{i+1} \nabla g(\mathbf{x}_{i+1})). \quad (7.56)$$

Eine Gleichung für λ_{i+1} erhalten wir aus der Forderung, dass \mathbf{v}_{i+1} im Tangentialraum, also senkrecht dem Normalenvektor $\nabla g(\mathbf{x}_{i+1})$ steht.

$$\nabla g^T(\mathbf{x}_{i+1}) \cdot \mathbf{v}_{i+1} = 0. \quad (7.57)$$

Hieraus folgt das Gleichungssystem

$$\begin{pmatrix} I & -\nabla g(\mathbf{x}_{i+1}) \\ \nabla g^T(\mathbf{x}_{i+1}) & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v}_{i+1} \\ h\lambda_{i+1} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_i + hF \\ 0 \end{pmatrix}. \quad (7.58)$$

Für dieses Verfahren kann für hinreichend glattes g die Konvergenzordnung 1 bewiesen werden.¹²

¹²vgl. Proposition 82.2 in M. Hanke-Bourgeois