Numerische Methoden für gewöhnliche Differentialgleichungen

Lars Grüne
Lehrstuhl für Angewandte Mathematik
Mathematisches Institut
Universität Bayreuth
95440 Bayreuth
lars.gruene@uni-bayreuth.de
num.math.uni-bayreuth.de

Vorlesungsskript
6. Auflage
Sommersemester 2015

Vorwort

Dieses Skript ist im Rahmen einer gleichnamigen Vorlesung entstanden, die ich im Sommersemester 2015 an der Universität Bayreuth gehalten habe. Es ist die sechste Auflage eines Skriptes, das zuerst im Sommersemester 2003 erstellt wurde. Ich möchte mich an dieser Stelle wie stets bei all den StudentInnen bedanken, die mit zum Teil sehr ausführlichen Fehlerkorrekturen zur Verbesserung dieser dritten Auflage beigetragen haben. Neben der Verbesserungen von Fehlern wurden gegenüber der fünften Auflage die Kapitel 9–11 ergänzt, womit die Behandlung impliziter Verfahren ausgeweitet und Grundlagen geometrischer Integration in die Vorlesung aufgenommen wurden.

Die einzelnen Kapitel des Skriptes wurden auf Basis verschiedener Lehrbücher und Monographien erstellt. Dabei wurden insbesondere die Bücher von Deuflhard und Bornemann [2] und das Buch von Hairer, Lubich und Wanner [4] verwendet, allerdings wurden sowohl in Aufbau und Notation als auch bei einer Reihe von Beweisen Änderungen vorgenommen.

Eine elektronische Version dieses Skripts findet sich im WWW auf der Seite
`num.math.uni-bayreuth.de/en/team/Gruene_Lars/lecture_notes/`.

Bayreuth, Juli 2015 Lars Grüne

Inhaltsverzeichnis

Kapitel 1

Gewöhnliche Differentialgleichungen

Im Rahmen unserer numerischen Betrachtungen werden wir die benötigten theoretischen Resultate dort einführen, wo wir sie verwenden. Bevor wir mit der Numerik beginnen können, benötigen wir aber zumindest ein theoretisches Grundgerüst mit einigen Basisdefinitionen und Resultaten zu den gewöhnlichen Differentialgleichungen, das der nun folgende Abschnitt bereit stellt.

In diesem Abschnitt werden wir die grundlegenden Gleichungen definieren, mit denen wir uns im ersten Teil dieser Vorlesung beschäftigen wollen und einige ihrer Eigenschaften betrachten. Zudem werden wir zwei verschiedene grafische Darstellungsmöglichkeiten für die Lösungen kennen lernen. Für weitergehende Informationen über gewöhnliche Differentialgleichungen können z.B. die einführenden Lehrbücher [1] oder [3] empfohlen werden.

1.1 Definition

Eine gewöhnliche Differentialgleichung setzt die Ableitung einer Funktion $x : \mathbb{R} \to \mathbb{R}^n$ nach ihrem (eindimensionalen) Argument mit der Funktion selbst in Beziehung. Formal beschreibt dies die folgende Definition.

Definition 1.1 Ein *gewöhnliche Differentialgleichung* (DGL) im $\mathbb{R}^n$, $n \in \mathbb{N}$, ist gegeben durch die Gleichung

$$\frac{d}{dt}x(t) = f(t, x(t)), \tag{1.1}$$

wobei $f : D \to \mathbb{R}^n$ eine stetige Funktion ist und *Vektorfeld* genannt wird, deren Definitionsbereich D eine offene Teilmenge von $\mathbb{R} \times \mathbb{R}^n$ ist.

Eine *Lösung* von (1.1) ist eine stetig differenzierbare Funktion $x : \mathbb{R} \to \mathbb{R}^n$, die (1.1) erfüllt. □

Einige Anmerkungen zur Notation bzw. Sprechweise:

- Die unabhängige Variable t werden wir üblicherweise als Zeit interpretieren, obwohl (abhängig vom modellierten Sachverhalt) gelegentlich auch andere Interpretationen möglich sind.
- Statt $\frac{d}{dt}x(t)$ schreiben wir oft kurz $\dot{x}(t)$.
- Die Lösungsfunktion $x(t)$ nennen wir auch *Lösungskurve* oder *(Lösungs–)Trajektorie*.
- Falls das Vektorfeld f nicht von t abhängt, also $\dot{x}(t) = f(x(t))$ ist, nennen wir die Differentialgleichung *autonom*.

1.2 Anfangswertprobleme

Eine gewöhnliche Differentialgleichung besitzt im Allgemeinen unendlich viele Lösungen. Als Beispiel betrachte die (sehr einfache) eindimensionale DGL mit $f(x,t) = x$, also

$$\dot{x}(t) = x(t)$$

mit $x(t) \in \mathbb{R}$. Betrachte die Funktion $x(t) = Ce^t$ mit beliebigem $C \in \mathbb{R}$. Dann gilt

$$\dot{x}(t) = \frac{d}{dt}Ce^t = Ce^t = x(t).$$

Für jedes feste C löst Ce^t die obige DGL, es gibt also unendlich viele Lösungen.

Um *eindeutige* Lösungen zu erhalten, müssen wir eine weitere Bedingung festlegen. Dies geschieht in der folgenden Definition.

Definition 1.2 Ein *Anfangswertproblem* für die gewöhnliche Differentialgleichung (1.1) besteht darin, zu gegebenem $t_0 \in \mathbb{R}$ und $x_0 \in \mathbb{R}^n$ eine Lösungsfunktion $x(t)$ zu finden, die (1.1) erfüllt und für die darüberhinaus die Gleichung

$$x(t_0) = x_0 \tag{1.2}$$

gilt. □

Notation und Sprechweisen:

- Für die Lösung $x(t)$, die (1.1) und (1.2) erfüllt, schreiben wir $x(t; t_0, x_0)$. Im Spezialfall $t_0 = 0$ werden wir oft kurz $x(t; x_0)$ schreiben.
- Die Zeit $t_0 \in \mathbb{R}$ bezeichnen wir als *Anfangszeit*, den Wert $x_0 \in \mathbb{R}^n$ als *Anfangswert*. Das Paar (t_0, x_0) bezeichnen wir als *Anfangsbedingung*, ebenso nennen wir die Gleichung (1.2) *Anfangsbedingung*.

Bemerkung 1.3 Eine stetig differenzierbare Funktion $x : I \to \mathbb{R}^n$ löst das Anfangswertproblem (1.1), (1.2) für ein $t_0 \in I$ und ein $x_0 \in \mathbb{R}^n$ genau dann, wenn sie für alle $t \in I$ die *Integralgleichung*

$$x(t) = x_0 + \int_{t_0}^{t} f(\tau, x(\tau))d\tau \tag{1.3}$$

erfüllt. Dies folgt sofort durch Integrieren von (1.1) bzgl. t bzw. durch Differenzieren von (1.3) nach t unter Verwendung des Hauptsatzes der Differential– und Integralrechnung. Beachte dabei, dass eine stetige Funktion x, die (1.3) erfüllt, „automatisch“ stetig differenzierbar ist, da aus der Stetigkeit von x sofort die stetige Differenzierbarkeit der rechten Seite in (1.3) und damit wegen der Gleichhet auch für x selbst folgt. □

1.3 Ein Existenz– und Eindeutigkeitssatz

Unter geeigneten Bedingungen an f können wir einen Existenz– und Eindeutigkeitssatz für Anfangswertprobleme der Form (1.1), (1.2) erhalten.

Satz 1.4 Betrachte die gewöhnliche Differentialgleichung (1.1) für ein $f : D \to \mathbb{R}^n$ mit $D \subseteq \mathbb{R} \times \mathbb{R}^n$ offen. Das Vektorfeld f sei stetig, darüberhinaus sei f Lipschitz–stetig im zweiten Argument im folgenden Sinne: Für jede kompakte Teilmenge $K \subset D$ existiere eine Konstante $L > 0$, so dass die Ungleichung

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

gilt für alle $t \in \mathbb{R}$ und $x, y \in \mathbb{R}^n$ mit (t,x), $(t,y) \in K$.

Dann gibt es für jede Anfangsbedingung $(t_0, x_0) \in D$ genau eine Lösung $x(t; t_0, x_0)$ des Anfangswertproblems (1.1), (1.2). Diese ist definiert für alle t aus einem offenen *maximalen Existenzintervall* $I_{t_0,x_0} \subseteq \mathbb{R}$ mit $t_0 \in I_{t_0,x_0}$.

Beweis: Teil 1: Wir zeigen zunächst, dass es für jede Anfangsbedingung $(t_0, x_0) \in D$ ein abgeschlossenes Intervall J um t_0 gibt, auf dem die Lösung existiert und eindeutig ist.

Dazu wählen wir ein beschränktes abgeschlossenes Intervall I um t_0 und ein $\varepsilon > 0$, so dass die kompakte Umgebung $U = I \times \overline{B}_\varepsilon(x_0)$ von (t_0, x_0) in D liegt (dies ist möglich, da D eine offene Menge ist). Da f stetig ist und U kompakt ist, existiert eine Konstante M, so dass $\|f(t,x)\| \leq M$ für alle $(t,x) \in U$ gilt. Wir wählen nun $J = [t_0 - \delta, t_0 + \delta]$ wobei $\delta > 0$ so gewählt ist, dass $J \subseteq I$ gilt und $L\delta < 1$ sowie $M\delta < \varepsilon$ erfüllt ist, wobei L die Lipschitz-Konstante von f für $K = U$ ist. Alle somit konstruierten Mengen sind in Abbildung 1.1 dargestellt.

Nun verwenden wir zum Beweis der Existenz und Eindeutigkeit der Lösung auf J den Banachschen Fixpunktsatz auf dem Banachraum $C(J, \mathbb{R}^n)$ mit der Norm

$$\|x\|_\infty := \sup_{t \in J} \|x(t)\|.$$

Auf $C(J, \mathbb{R}^d)$ definieren wir die Abbildung

$$T : C(J, \mathbb{R}^d) \to C(J, \mathbb{R}^d), \quad T(x)(t) := x_0 + \int_{t_0}^{t} f(\tau, x(\tau)) d\tau.$$

Beachte, dass für jedes $t \in J$ und jedes $x \in B := C(J, \overline{B}_\varepsilon(x_0))$ die Ungleichung

$$\begin{aligned} \|T(x)(t) - x_0\| &= \left\| \int_{t_0}^{t} f(\tau, x(\tau)) d\tau \right\| \leq \left| \int_{t_0}^{t} \underbrace{\|f(\tau, x(\tau))\|}_{\leq M, \text{ weil } (\tau, x(\tau)) \in \overline{U}} d\tau \right| \\ &\leq \delta M \leq \varepsilon \end{aligned}$$

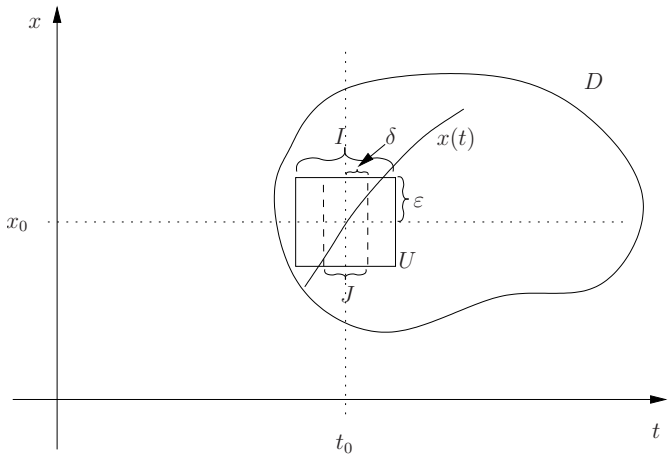

Abbildung 1.1: Mengen im Beweis von Teil 1

gilt, weswegen T die Menge B in sich selbst abbildet.

Um den Banachschen Fixpunktsatz auf dieser Menge anzuwenden, müssen wir zeigen, dass $T : B \to B$ eine Kontraktion ist, also dass

$$\|T(x) - T(y)\|_\infty \leq k\|x - y\|_\infty$$

gilt für alle $x, y \in B$ und ein $k < 1$. Diese Eigenschaft folgt für $k = L\delta < 1$ aus

$$\begin{aligned}
\|T(x) - T(y)\|_\infty &= \sup_{t\in J} \left\| \int_{t_0}^{t} f(\tau, x(\tau))d\tau - \int_{t_0}^{t} f(\tau, y(\tau))d\tau \right\| \\
&\leq \sup_{t\in J} \left| \int_{t_0}^{t} \underbrace{\|f(\tau, x(\tau)) - f(\tau, y(\tau))\|}_{\leq L\|x(\tau)-y(\tau)\| \leq L\|x-y\|_\infty} d\tau \right| \\
&\leq \sup_{t\in J} |t - t_0| L\|x - y\|_\infty = \delta L\|x - y\|_\infty.
\end{aligned}$$

Also sind die Voraussetzungen des Banachschen Fixpunktsatzes erfüllt, weswegen T einen eindeutigen Fixpunkt $x \in B$, also eine „Fixpunktfunktion“, besitzt. Da diese Fixpunktfunktion x nach Konstruktion von T die Integralgleichung (1.3) erfüllt, ist sie nach Bemerkung 1.3 eine stetig differenzierbare Lösung des Anfangswertproblems.

Es bleibt zu zeigen, dass diese eindeutig ist, dass also kein weiterer Fixpunkt $y \in C(J, \mathbb{R}^d)$ existiert. Aus dem Banachschen Fixpunktsatz folgt bereits, dass in $B = C(J, \overline{B}_\varepsilon(x_0))$ kein weiterer Fixpunkt von T liegt. Zum Beweis der Eindeutigkeit reicht es also zu zeigen, dass außerhalb von B kein Fixpunkt y liegen kann. Wir beweisen dies per Widerspruch: Angenommen, es existiert eine Fixpunktfunktion $y \notin B$ von T, d.h. es gilt $\|y(t) - x_0\| > \varepsilon$ für ein $t \in J$, für das wir o.B.d.A. $t > t_0$ annehmen. Dann existiert aus Stetigkeitsgründen ein $t^* \in J$ mit $\|y(t^*) - x_0\| = \varepsilon$ und $y(s) \in \overline{B}_\varepsilon(x_0)$ für $s \in [t_0, t^*]$. Damit folgt

$$\varepsilon = \|y(t^*) - x_0\| = \left\| \int_{t_0}^{t^*} f(s, y(s))ds \right\| \leq \int_{t_0}^{t^*} \|f(s, y(s))\|ds$$

$$\leq \quad (t^* - t_0)M \;<\; \delta M,$$

was wegen $\delta M \leq \varepsilon$ ein Widerspruch ist. Daher liegt jeder mögliche Fixpunkt $y \in C(J, \mathbb{R}^d)$ von T bereits in B, womit die Eindeutigkeit folgt.

Zusammenfassend liefert uns Teil 1 des Beweises also, dass *lokal* – also auf einem kleinen Intervall J um t_0 – eine eindeutige Lösung $x(t) = x(t; t_0, x_0)$ existiert. Dies ist die Aussage des *Satzes von Picard-Lindelöf*[1], der in vielen Büchern als eigenständiger Satz formuliert ist.

Teil 2: Wir zeigen als nächstes die Eindeutigkeit der Lösung auf beliebig großen Intervallen I. Seien dazu x und y zwei auf einem Intervall I definierte Lösungen des Anfangswertproblems. Wir beweisen $x(t) = y(t)$ für alle $t \in I$ per Widerspruch und nehmen dazu an, dass ein $t \in I$ existiert, in dem die beiden Lösungen nicht übereinstimmen, also $x(t) \neq y(t)$. O.b.d.A. sei $t > t_0$. Da beide Lösungen nach Teil 1 auf J übereinstimmen und stetig sind, existieren $t_2 > t_1 > t_0$, so dass

$$x(t_1) = y(t_1) \quad \text{und} \quad x(t) \neq y(t) \text{ für alle } t \in (t_1, t_2) \tag{1.4}$$

gilt. Offenbar lösen beide Funktionen das Anfangswertproblem mit Anfangsbedingung $(t_1, x(t_1)) \in D$. Aus Teil 1 des Beweises folgt die Eindeutigkeit der Lösungen dieses Problems auf einem Intervall $\widetilde{J}$ um t_1, also

$$x(t) = y(t) \text{ für alle } t \in \widetilde{J}.$$

Da $\widetilde{J}$ als Intervall um t_1 einen Punkt t mit $t_1 < t < t_2$ enthält, widerspricht dies (1.4), weswegen x und y für alle $t \in I$ übereinstimmen müssen.

Teil 3: Schließlich zeigen wir die Existenz des maximalen Existenzintervalls. Für J aus Teil 1 definieren wir dazu

$$t^+ := \sup\{s > t_0 \,|\, \text{es existiert eine Lösung auf } J \cup [t_0, s)\}$$

sowie

$$t^- := \inf\{s < t_0 \,|\, \text{es existiert eine Lösung auf } J \cup (s, t_0]\}$$

und setzen $I_{t_0,x_0} = (t^-, t^+)$. Sowohl t^- als auch t^+ existieren, da die Mengen, über die das Supremum bzw. Infimum genommen wird, nichtleer sind, da sie alle $s \in J$ enthalten. Per Definition von t^+ bzw. t^- kann es keine Lösung auf einem größeren Intervall $I \supset I_{t_0,x_0}$ geben, also ist dies das maximale Existenzintervall.

□

Am Rand des maximalen Existenzintervalls $I_{t_0,x_0} = (t^-, t^+)$ hört die Lösung auf zu existieren. Ist das Intervall in einer Zeitrichtung beschränkt, so kann dies nur zwei verschiedene Ursachen haben: Entweder die Lösung divergiert, oder sie konvergiert gegen einen Randpunkt von D. Formal ausgedrückt:

Falls $t^+ < \infty$ ist und die Lösung $x(t; t_0, x_0)$ für $t \nearrow t^+$ gegen ein $x^+ \in \mathbb{R}^d$ konvergiert, so muss $(t^+, x^+) \notin D$ gelten. Analog gilt die Aussage für $t \searrow t^-$. Hierbei steht $t \nearrow t^+$ kurz für $t \to t^+$ und $t < t^+$ und $t \searrow t^-$ für $t \to t^-$ und $t > t^-$.

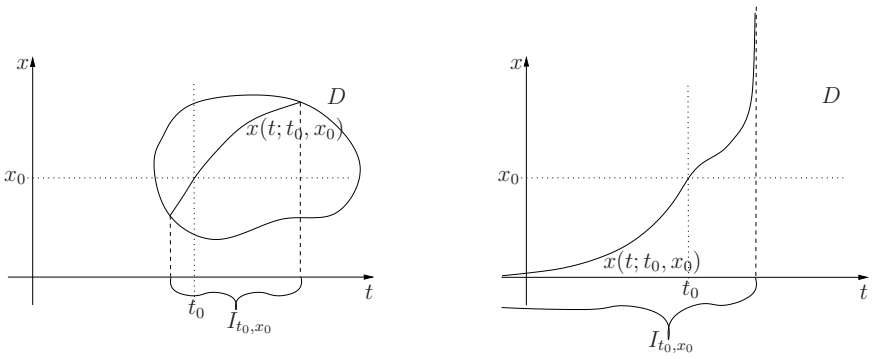

Abbildung 1.2: Lösungsverhalten am Rand des Existenzintervalls für eine beschränkte (links) und eine unbeschränkte Definitionsmenge D (rechts)

Anschaulich sind die zwei Möglichkeiten in Abbildung 1.2 dargestellt.

Die Begründung für dieses Verhalten ist wie folgt:

Wenn $x(t;t_0,x_0)$ für $t \nearrow t^+$, gegen $x^+ \in \mathbb{R}^d$ mit $(t^+,x^+) \in D$ konvergiert, so existiert eine Lösung $x(t;t^+,x^+)$ auf einem offenen Intervall I_{t^+,x^+} um t^+. Dann ist die zusammengesetzte Lösung

$$y(t) = \begin{cases} x(t;t_0,x_0), & t \in I_{t_0,x_0} \\ x(t;t^+,x^+), & t \in I_{t^+,x^+} \setminus I_{t_0,x_0} \end{cases}$$

stetig und erfüllt für alle $t \in I_{t_0,x_0} \cup I_{t^+,x^+}$ die Integralgleichung (1.3), damit nach Bemerkung 1.3 auch das Anfangswertproblem und ist folglich eine Lösung, die über t^+ hinaus definiert ist: ein Widerspruch zur Definition von t^+.

Im Fall $D = \mathbb{R} \times \mathbb{R}^d$ gilt daher für $t^+ < \infty$ bzw. $t^- > -\infty$ insbesondere, dass die Lösung $x(t;t_0,x_0)$ für $t \nearrow t^+$ bzw. $t \searrow t^-$ divergieren muss, da eine Konvergenz gegen $(t^+,x^+) \notin D$ bzw. $(t^-,x^-) \notin D$ nicht möglich ist. Beachte, dass dieser Fall tatsächlich auftreten kann: eine unbeschränkte Definitionsmenge D von f bedeutet nicht, dass auch die Lösungen auf einem unbeschränkten Intervall $I_{t_0,x_0} = \mathbb{R}$ existieren. Ein Beispiel dafür ist die Differentialgleichung $\dot{x}(t) = x(t)^2$ mit $x(t) \in \mathbb{R}$. Diese besitzt für Anfangsbedingung $x(0) = 1$ die Lösung $x(t) = 1/(1-t)$, die für $t \to 1$ gegen unendlich strebt. Es gilt also $t^+ = 1$, obwohl $D = \mathbb{R} \times \mathbb{R}$ unbeschränkt ist.

Zu beachten ist weiterhin, dass die Divergenz nicht wie im rechten Bild in Abbildung 1.2 skizziert bedeuten muss, dass die Lösung gegen unendlich (oder minus unendlich) strebt. Ein Beispiel dafür ist $\dot{x}(t) = -\cos(1/t)/t^2$ mit $D = \mathbb{R} \setminus \{0\} \times \mathbb{R}$. Für die Anfangsbedingung $x(-1) = \sin(-1)$ erhält man hier die Lösung $x(t) = \sin(1/t)$. Für $t \to t^+ = 0$, $t < 0$ konvergiert diese Lösung nicht, weil sie immer schneller zwischen -1 und -1 oszilliert; sie ist aber für alle $t < 0$ nach oben und unten beschränkt.

Wir werden im Folgenden immer annehmen, dass die Annahmen von Satz 1.4 erfüllt sind, auch ohne dies explizit zu erwähnen. Auch werden wir oft Mengen der Form $[t_1,t_2] \times K$

[1] Charles Picard, französischer Mathematiker, 1856–1941
Ernst Lindelöf, finnischer Mathematiker, 1870–1946

mit $K \subset \mathbb{R}^n$ betrachten, bei denen wir — ebenfalls ohne dies immer explizit zu erwähnen — annehmen, dass alle Lösungen $x(t; t_0, x_0)$ mit $x_0 \in K$ für alle $t_0, t \in [t_1, t_2]$ existieren.

Eine einfache Konsequenz aus Satz 1.4 ist die sogenannte *Kozykluseigenschaft* der Lösungen, die für $(t_0, x_0) \in D$ und zwei Zeiten $t_1, t \in \mathbb{R}$ gegeben ist durch

$$x(t; t_0, x_0) = x(t; t_1, x(t_1; t_0, x_0)), \tag{1.5}$$

vorausgesetzt natürlich, dass alle hier auftretenden Lösungen zu den angegebenen Zeiten auch existieren. Zum Beweis rechnet man nach, dass der linke Ausdruck in (1.5) das Anfangswertproblem (1.1), (1.2) zur Anfangsbedingung $(t_1, x(t_1; t_0, x_0))$ löst. Da der rechte dies ebenfalls tut, müssen beide übereinstimmen.

Unter den Voraussetzungen von Satz 1.4 ist die Lösungsabbildung $x(t; t_0, x_0)$ zudem stetig in all ihren Variablen, also in t, t_0 und x_0.

1.4 Grafische Darstellung der Lösungen

Zur grafischen Darstellung von Lösungen verwenden wir zwei verschiedene Methoden, die wir hier an der zweidimensionalen DGL

$$\dot{x}(t) = \begin{pmatrix} -0.1 & 1 \\ -1 & -0.1 \end{pmatrix} x(t)$$

mit $x(t) = (x_1(t), x_2(t))^T$ und Anfangsbedingung $x(0) = (1, 1)^T$ illustrieren wollen. Da jede Lösung einer Differentialgleichung eine Funktion von $\mathbb{R}$ nach $\mathbb{R}^n$ darstellt, kann man die Graphen der einzelnen Komponenten $x_i(t)$ der Lösung in Abhängigkeit von t darstellen. Für die obige DGL ist dies in Abbildung 1.3 dargestellt. Die durchgezogene Linie zeigt $x_1(t)$ während die gestrichelte Linie $x_2(t)$ darstellt.

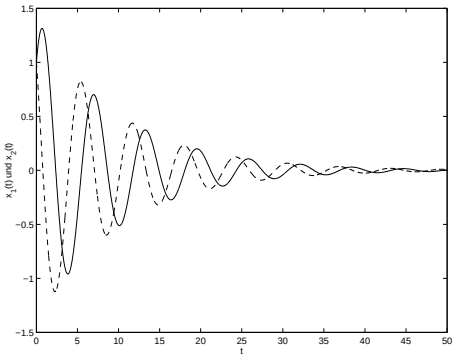

Abbildung 1.3: Darstellung von $x(t)$ mittels Graphen ($x_1(t)$ durchgezogen, $x_2(t)$ gestrichelt)

Eine alternative Darstellung, die speziell für zwei– und dreidimensionale Differentialgleichungen geeignet ist, ergibt sich, wenn man statt der Funktionsgraphen der Komponenten x_i die Kurve $\{x(t) \,|\, t \in [0, T]\} \subset \mathbb{R}^n$ darstellt. Hier geht in der Grafik die Information über die Zeit (sowohl über die Anfangszeit t_0 als auch über die laufende Zeit t) verloren.

Letzteres kann zumindest teilweise durch das Anbringen von Pfeilen, die die Zeitrichtung symbolisieren, ausgeglichen werden. Ein Beispiel für diese Darstellung zeigt Abbildung 1.4.

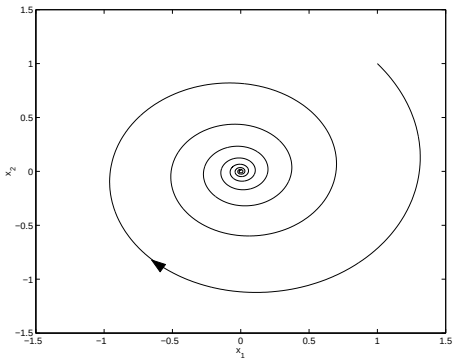

Abbildung 1.4: Darstellung von $x(t)$ als Kurve

Am Computer kann man die Darstellung als Kurve mit einer Animation verbinden, so dass man die Information über den zeitlichen Ablauf der Lösung über die Animation wieder zurück erhält. Ein MATLAB M–File, das sowohl die Abbildungen 1.3 und 1.4 sowie eine animierte Version von Abbildung 1.4 erstellt, findet sich auf der Vorlesungs–Homepage[2] unter dem Namen "darstellung.m".

Für autonome Differentialgleichungen ist der Verlust der Anfangszeit in der Grafik nicht weiter schlimm, da die Lösungen nicht wirklich von der Anfangszeit abhängen: man rechnet leicht nach, dass hier für die Anfangszeiten t_0 und $t_0 + t_1$ die Beziehung

$$x(t; t_0 + t_1, x_0) = x(t - t_1; t_0, x_0) \tag{1.6}$$

gilt. Die Lösung verschiebt sich also auf der t–Achse, verändert sich aber ansonsten nicht. Insbesondere ist die in Abbildung 1.4 dargestellte Kurve für autonome DGL für alle Anfangszeiten gleich.

[2] http://www.uni-bayreuth.de/departments/math/∼lgruene/numerik05/

Kapitel 2

Allgemeine Theorie der Einschrittverfahren

In diesem Kapitel werden wir eine wichtige Klasse von Verfahren zur Lösung gewöhnlicher Differentialgleichungen einführen und analysieren, die *Einschrittverfahren*.

2.1 Diskrete Approximationen

In der Numerik gewöhnlicher Differentialgleichungen wollen wir eine Approximation an die Lösungsfunktion $x(t; t_0, x_0)$ für $t \in [t_0, T]$ berechnen (wir nehmen hier immer an, dass die Lösungen auf den angegebenen Intervallen existieren). In der folgenden Definition definieren wir die Art von Approximationen, die wir betrachten wollen und einen Begriff der Konvergenzordnung.

Definition 2.1 (i) Eine Menge $\mathcal{T} = \{t_0, t_1, \ldots, t_N\}$ von Zeiten mit $t_0 < t_1 < \ldots < t_N = T$ heißt *Gitter* auf dem Intervall $[t_0, T]$. Die Werte

$$h_i = t_{i+1} - t_i$$

heißen *Schrittweiten*, der Wert

$$\bar{h} = \max_{i=0,\ldots,N-1} h_i$$

heißt *maximale Schrittweite*. Im Fall *äquidistanter Schrittweiten* $h_0 = h_1 = \ldots = h_{N-1}$ schreiben wir zumeist h statt h_i.

(ii) Eine Funktion $\tilde{x} : \mathcal{T} \to \mathbb{R}^n$ heißt *Gitterfunktion*.

(iii) Es seien $\tilde{x}_\mathcal{T}$ Gitterfunktionen zu Gittern $\mathcal{T}$ auf dem Intervall $[t_0, T] \subset I_{t_0,x_0}$ mit maximalen Schrittweiten $\bar{h}_\mathcal{T}$. Die Gitterfunktionen $\tilde{x}_\mathcal{T}$ bilden eine *(diskrete) Approximation* der Lösung $x(t; t_0, x_0)$ von (1.1), falls für jede kompakte Menge $K \subset D$ mit $[t_0, T] \subset I_{t_0,x_0}$ für alle $(t_0, x_0) \in K$ eine Funktion $\rho(h)$ mit $\rho(h) \to 0$ für $h \to 0$ existiert mit

$$\max_{t_i \in \mathcal{T}} \|\tilde{x}_\mathcal{T}(t_i) - x(t_i; t_0, x_0)\| \leq \rho(\bar{h}_\mathcal{T}).$$

Die diskrete Approximation hat die *Konvergenzordnung* $p > 0$, falls für jede kompakte Menge $K \subset D$ und alle $T > 0$ mit $[t_0, T] \subset I_{t_0,x_0}$ für alle $(t_0, x_0) \in K$ ein $C > 0$ existiert, so dass

$$\rho(h) = Ch^p$$

gewählt werden kann. In diesem Fall schreiben wir kurz $\tilde{x}(t_i; t_0, x_0) = x(t_i; t_0, x_0) + O(\bar{h}^p)$. □

Bemerkung 2.2 Wir haben in der Einführung in die Numerik verschiedene Methoden kennen gelernt, mit denen man Funktionen numerisch darstellen kann, z.B. Polynom– oder Splineinterpolation. Jede Gitterfunktion gemäß Definition 2.1 kann natürlich mit diesen Methoden zu einer "echten" Funktion erweitert werden. □

Ein Einschrittverfahren ist nun gegeben durch eine numerisch auswertbare Funktion Φ, mittels derer wir eine Gitterfunktion zu einem gegebenen Gitter berechnen können. Formal ist dies wie folgt definiert.

Definition 2.3 Ein *Einschrittverfahren* ist gegeben durch eine stetige Abbildung

$$\Phi : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n,$$

mit der zu jedem Gitter $\mathcal{T}$ und jedem Anfangswert x_0 mittels

$$\tilde{x}(t_0) = x_0, \;\; \tilde{x}(t_{i+1}) = \Phi(t_i, \tilde{x}(t_i), h_i) \text{ für } i = 0, 1, \ldots, N-1$$

rekursiv eine Gitterfunktion definiert werden kann.

Wenn die so erzeugten Gitterfunktionen die Bedingung aus Definition 2.1 (iii) erfüllen, so nennen wir das Einschrittverfahren *konvergent* bzw. *konvergent mit Konvergenzordnung p*. □

Der Name *Einschrittverfahren* ergibt sich dabei aus der Tatsache, dass der Wert $\tilde{x}(t_{i+1})$ nur aus dem direkten Vorgängerwert $\tilde{x}(t_i)$ berechnet wird. Wir werden später auch *Mehrschrittverfahren* kennen lernen, bei denen $\tilde{x}(t_{i+1})$ aus $\tilde{x}(t_{i-k}), \tilde{x}(t_{i-k+1}), \ldots, \tilde{x}(t_i)$ berechnet wird.

2.2 Erste einfache Einschrittverfahren

Bevor wir in die Konvergenztheorie einsteigen und mathematisch untersuchen, welche Bedingungen Φ erfüllen muss, damit die erzeugte Gitterfunktion eine Approximation darstellt, wollen wir in diesem Abschnitt zwei Einschrittverfahren heuristisch betrachten.

Die Idee der Verfahren erschließt sich am einfachsten über die Integralgleichung (1.3). Die exakte Lösung erfüllt ja gerade

$$x(t_{i+1}) = x(t_i) + \int_{t_i}^{t_{i+1}} f(\tau, x(\tau)) d\tau.$$

Die Idee ist nun, das Integral durch einen Ausdruck zu ersetzen, der numerisch berechenbar ist, wenn wir $x(\tau)$ für $\tau > t_i$ nicht kennen. Die einfachste Approximation ist die Rechteck–Regel (oder Newton–Cotes Formel mit $n = 0$, die wir in der Einführung in die Numerik wegen ihrer Einfachheit gar nicht betrachtet haben)

$$\int_{t_i}^{t_{i+1}} f(\tau, x(\tau))d\tau \approx (t_{i+1} - t_i)f(t_i, x(t_i)) = h_i f(t_i, x(t_i)). \tag{2.1}$$

Setzen wir also

$$\Phi(t, x, h) = x + hf(t, x), \tag{2.2}$$

so gilt

$$\tilde{x}(t_{i+1}) = \Phi(t_i, \tilde{x}(t_i), h_i) = \tilde{x}(t_i) + h_i f(t_i, \tilde{x}(t_i))$$

und wenn wir $\tilde{x}(t_i) \approx x(t_i)$ annehmen, so können wir fortfahren

$$\ldots \approx x(t_i) + h_i f(t_i, x(t_i)) \approx x(t_i) + \int_{t_i}^{t_{i+1}} f(\tau, x(\tau))d\tau.$$

Da $\tilde{x}(t_0) = x_0 = x(t_0)$ ist, kann man damit rekursiv zeigen, dass $\tilde{x}(t_{i+1})$ eine Approximation von $x(t_{i+1})$ ist. Wir werden dies im nächsten Abschnitt mathematisch präzisieren.

Das durch (2.2) gegebene Verfahren ist das einfachste Einschrittverfahren und heißt *Euler'sche Polygonzugmethode* oder einfach *Euler–Verfahren*. Es hat eine einfache geometrische Interpretation: In jedem Punkt $\tilde{x}(t_i)$ berechnen wir die Steigung der exakten Lösung durch diesen Punkt (das ist gerade $f(t_i, \tilde{x}(t_i))$) und folgen der dadurch definierten Geraden bis zum nächsten Zeitschritt. Das Prinzip ist in Abbildung 2.1 grafisch dargestellt.

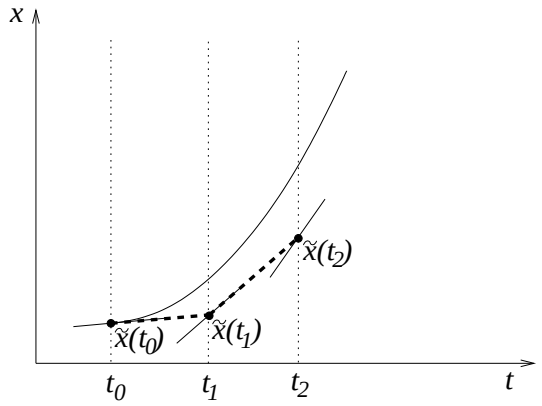

Abbildung 2.1: Grafische Veranschaulichung des Euler–Verfahrens

Das Euler–Verfahren liefert nur eine recht grobe Approximation der Lösung. Bessere Verfahren kann man erhalten, wenn man statt (2.1) eine genauere Approximation verwendet. Eine bessere Möglichkeit ist z.B.

$$\int_{t_i}^{t_{i+1}} f(\tau, x(\tau))d\tau \approx \frac{h_i}{2}\left(f(t_i, x(t_i)) + f\Big(t_{i+1}, x(t_i) + h_i f(t_i, x(t_i))\Big)\right). \tag{2.3}$$

Dies ist nichts anderes als die Trapez–Regel (oder Newton–Cotes Formel mit $n = 1$), bei der wir den unbekannten Wert $x(t_{i+1})$ durch die Euler–Approximation $x(t_{i+1}) \approx x(t_i) + h_i f(t_i, x(t_i))$ ersetzen. Das daraus resultierende Verfahren ist gegeben durch

$$\Phi(t, x, h) = x + \frac{h}{2}\left(f(t, x) + f\Big(t + h, x + hf(t, x)\Big)\right)$$

und heißt *Heun–Verfahren*. Es ist tatsächlich schon deutlich besser als das Euler–Verfahren.

Man kann sich leicht vorstellen, dass weitere bessere Verfahren sehr komplizierte Formeln benötigen. Wir werden deshalb später einen Formalismus kennen lernen, mit dem man auch sehr komplizierte Verfahren einfach aufschreiben und implementieren kann.

Ein Grundalgorithmus zur Approximation einer Lösung $x(t; t_0, x_0)$ auf $[t_0, T]$ mittels eines Einschrittverfahrens Φ lässt sich nun leicht angeben. Wir beschränken uns hierbei zunächst auf Gitter mit konstanter Schrittweite, also $h_i = h$ für alle $i = 0, 1, 2, \ldots, N$, wobei wir N als Parameter vorgeben.

Algorithmus 2.4 (Lösung eines Anfangswertproblems mit Einschrittverfahren)

Eingabe: Anfangsbedingung (t_0, x_0), Endzeit T, Schrittzahl N, Einschrittverfahren Φ

(1) Setze $h := (T - t_0)/N$, $\tilde{x}_0 = x_0$

(2) Berechne $t_{i+1} = t_i + h$, $\tilde{x}_{i+1} := \Phi(t_i, \tilde{x}_i, h)$ für $i = 0, \ldots, N - 1$.

Ausgabe: Werte der Gitterfunktion $\tilde{x}(t_i) = \tilde{x}_i$ in $t_0, \ldots, t_N$ □

2.3 Konvergenztheorie

Die Grundidee der Konvergenztheorie für numerische Methoden für Differentialgleichungen liegt in einem geschickten Trick, mit dem verschiedene Fehlerquellen separiert werden können. Wir schreiben hier kurz $x(t) = x(t; t_0, x_0)$. Um nun den Fehler

$$\|\tilde{x}(t_i) - x(t_i)\| = \|\Phi(t_{i-1}, \tilde{x}(t_{i-1}), h_{i-1}) - x(t_i)\|$$

abzuschätzen, schieben wir mittels der Dreiecksungleichung die Hilfsgröße

$$\Phi(t_{i-1}, x(t_{i-1}), h_{i-1})$$

ein. Wir erhalten so mit (1.5) die Abschätzung

$$\begin{aligned}
\|\tilde{x}(t_i) - x(t_i)\| \;&\leq\; \|\Phi(t_{i-1}, \tilde{x}(t_{i-1}), h_{i-1}) - \Phi(t_{i-1}, x(t_{i-1}), h_{i-1})\| \\
&\quad +\; \|\Phi(t_{i-1}, x(t_{i-1}), h_{i-1}) - x(t_i)\| \\
&=\; \|\Phi(t_{i-1}, \tilde{x}(t_{i-1}), h_{i-1}) - \Phi(t_{i-1}, x(t_{i-1}), h_{i-1})\| \\
&\quad +\; \|\Phi(t_{i-1}, x(t_{i-1}), h_{i-1}) - x(t_i; t_{i-1}, x_{i-1})\|
\end{aligned}$$

Statt also direkt den Fehler zur Zeit t_i abzuschätzen, betrachten wir getrennt die zwei Terme

(a) $\|\Phi(t_{i-1}, \tilde{x}(t_{i-1}), h_{i-1}) - \Phi(t_{i-1}, x(h_{i-1}), h_{i-1})\|$, also die Auswirkung des Fehlers bis zur Zeit t_{i-1} in Φ

(b) $\|\Phi(t_{i-1}, x(t_{i-1}), h_{i-1}) - x(t_i; t_{i-1}, x_{i-1})\|$, also den lokalen Fehler beim Schritt von $x(t_{i-1})$ nach $x(t_i)$

Die folgende Definition gibt die benötigten Eigenschaften an Φ an, mit denen diese Fehler abgeschätzt werden können.

Definition 2.5 (i) Ein Einschrittverfahren erfüllt die *Lipschitzbedingung* (oder *Stabilitätsbedingung*), falls für jede kompakte Menge $K \subset D$ des Definitionsbereiches der Differentialgleichung ein $L > 0$ existiert, so dass für alle Paare $(t_0, x_1), (t_0, x_2) \in K$ und alle hinreichend kleinen $h > 0$ die Abschätzung

$$\|\Phi(t_0, x_1, h) - \Phi(t_0, x_2, h)\| \leq (1 + Lh)\|x_1 - x_2\| \tag{2.4}$$

gilt.

(ii) Ein Einschrittverfahren Φ heißt *konsistent*, falls für jede kompakte Menge $K \subset D$ des Definitionsbereiches der Differentialgleichung eine Funktion $\varepsilon(h)$ mit $\lim_{h\to 0} \varepsilon(h) = 0$ existiert, so dass für alle $(t_0, x_0) \in K$ und alle hinreichend kleinen $h > 0$ die Ungleichung

$$\|\Phi(t_0, x_0, h) - x(t_0 + h; t_0, x_0)\| \leq h\varepsilon(h) \tag{2.5}$$

gilt. O.B.d.A. nehmen wir dabei an, dass $\varepsilon(h)$ monoton ist, ansonsten können wir $\varepsilon(h)$ durch $\sup_{h\in[0,h]} \varepsilon(h)$ ersetzen.
Das Verfahren hat die *Konsistenzordnung* $p > 0$, falls für jede kompakte Menge $K \subset D$ ein $E > 0$ existiert, so dass $\varepsilon(h) = Eh^p$ gewählt werden kann. In diesem Fall schreiben wir auch $\Phi(t_0, x_0, h) = x(t_0 + h; t_0, x_0) + O(h^{p+1})$. □

Offenbar garantiert (2.4), dass der Fehlerterm (a) nicht zu groß wird, während (2.5) dazu dient, den Term (b) abzuschätzen. Der formale Beweis folgt in Satz 2.7. Bevor wir diesen formulieren, wollen wir uns noch überlegen, ob die im vorherigen Abschnitt definierten Verfahren diese Bedingungen erfüllen.

Man rechnet leicht nach, dass das Euler– und das Heun–Verfahren die Lipschitzbedingung erfüllen. Die Konsistenzbedingung (2.5) ist allerdings nicht so leicht nachzuprüfen, da sie mit Hilfe der (unbekannten) Lösungen $x(t; t_0, x_0)$ formuliert ist. Das folgende Lemma stellt eine alternative und leichter nachprüfbare Formulierung der Bedingung vor.

Lemma 2.6 Gegeben sei ein Einschrittverfahren Φ der Form

$$\Phi(t, x, h) = x + h\varphi(t, x, h)$$

mit einer stetigen Funktion $\varphi : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. Dann ist das Verfahren genau dann konsistent, falls für alle $(t, x) \in D$ die Bedingung

$$\varphi(t, x, 0) = f(t, x) \tag{2.6}$$

gilt.

Beweis: Wir schreiben wieder kurz $x(t) = x(t; t_0, x_0)$. Es gilt

$$\begin{aligned}
&\frac{\Phi(t_0, x_0, h) - x(t_0 + h)}{h} \\
&= \frac{1}{h}\left(\Phi(t_0, x_0, h) - x_0 - \int_{t_0}^{t_0+h} f(\tau, x(\tau))d\tau\right) \\
&= \frac{1}{h}\left(\Phi(t_0, x_0, h) - x_0 - \int_{t_0}^{t_0+h} f(t_0, x_0)d\tau + \int_{t_0}^{t_0+h} f(t_0, x_0)d\tau - \int_{t_0}^{t_0+h} f(\tau, x(\tau))d\tau\right) \\
&= \frac{1}{h}\left(h\varphi(t_0, x_0, h) - \int_{t_0}^{t_0+h} f(t_0, x_0)d\tau\right) + \frac{1}{h}\left(\int_{t_0}^{t_0+h} f(t_0, x_0) - f(\tau, x(\tau))d\tau\right) \\
&= \varphi(t_0, x_0, h) - f(t_0, x_0) + \frac{1}{h}\left(\int_{t_0}^{t_0+h} f(t_0, x_0) - f(\tau, x(\tau))d\tau\right)
\end{aligned}$$

Sei nun $K \subset D$ gegeben. Die Funktion $f(t_0 + s, x(t_0 + s; t_0, x_0))$ ist stetig in s, t_0 und x_0, also gleichmäßig stetig für $(s, t_0, x_0) \in [0, h] \times K$ für hinreichend kleines $h > 0$ (so klein, dass die Lösungen $x(t_0 + s; t_0, x_0))$ für $s \in [0, h]$ existieren), da diese Menge kompakt ist. Also existiert eine Funktion $\varepsilon_1(h) \to 0$ mit

$$\|f(\tau, x(\tau)) - f(t_0, x(t_0))\| \leq \varepsilon_1(h)$$

für $\tau = t_0 + s \in [t_0, t_0 + h]$ und damit

$$\frac{1}{h}\left\|\int_{t_0}^{t_0+h} f(t_0, x_0) - f(\tau, x(\tau))d\tau\right\| \leq \frac{1}{h}\int_{t_0}^{t_0+h} \|f(t_0, x_0) - f(\tau, x(\tau))\|d\tau \leq \varepsilon_1(h). \quad (2.7)$$

Wir nehmen nun an, dass (2.6) gilt. Ebenfalls wegen gleichmäßiger Stetigkeit und wegen (2.6) existiert eine Funktion $\varepsilon_2(h) \to 0$ mit

$$\|\varphi(t_0, x_0, h) - f(t_0, x_0)\| \leq \varepsilon_2(h).$$

Damit folgt

$$\frac{\|\Phi(t_0, x_0, h) - x(t_0 + h)\|}{h} \leq \varepsilon_2(h) + \varepsilon_1(h),$$

also (2.5) mit $\varepsilon(h) = \varepsilon_1(h) + \varepsilon_2(h)$.

Gelte umgekehrt (2.5). Sei $(t, x) \in D$ gegeben und sei $K = \{(t, x)\} \subset D$. Mit (2.5) und (2.7), angewendet mit $(t_0, x_0) = (t, x)$, folgt aus der Gleichung vom Anfang des Beweises

$$\|\varphi(t, x, h) - f(t, x)\| \leq \varepsilon(h) + \varepsilon_1(h),$$

also

$$\lim_{h\to 0} \|\varphi(t, x, h) - f(t, x)\| = 0$$

und damit (2.6) wegen der Stetigkeit von φ. □

Mit Hilfe der Bedingung (2.6) prüft man leicht nach, dass das Euler– und das Heun–Verfahren konsistent sind. Die Konsistenzordnung kann man aus (2.6) allerdings nicht ableiten, da die Abschätzung von $\varepsilon(h)$ mittels $\varepsilon_1(h)$ und $\varepsilon_2(h)$ dafür zu grob ist, denn falls $f \not\equiv 0$ ist, gilt $\varepsilon_1(h) \geq O(h)$, so dass man maximal die Konsistenzordnung $p = 1$ nachweisen könnte. Wir werden später sehen, wie man die Konsistenzordnung berechnen kann.

Wir kommen nun zu unserem ersten wichtigen Satz, der besagt, dass Lipschitzbedingung und Konsistenz tatsächlich ausreichend für die Konvergenz sind.

Satz 2.7 Betrachte ein Einschrittverfahren Φ, das die Lipschitzbedingung erfüllt und konsistent ist. Dann ist das Verfahren konvergent. Falls das Verfahren dabei die Konsistenzordnung p besitzt, so besitzt es auch die Konvergenzordnung p.

Beweis: Wir müssen die Eigenschaft aus Definition 2.1(iii) nachprüfen. Sei dazu eine kompakte Menge $K \subset D$ und ein $T > 0$ mit $[t_0, T] \subset I_{t_0,x_0}$ für alle $(t_0, x_0) \in K$ gegeben. Die Menge

$$K_1 := \{(t, x(t; t_0, x_0)) \,|\, (t_0, x_0) \in K, t \in [t_0, T]\}$$

ist dann ebenfalls kompakt, da x stetig in allen Variablen ist und Bilder kompakter Mengen unter stetigen Funktionen wieder kompakt sind. Wir wählen ein $\delta > 0$ und betrachten die kompakte Menge

$$K_2 := \bigcup_{(t,x)\in K_1} \{t\} \times \overline{B}_\delta(x).$$

Die Menge K_2 ist also genau die Menge aller Punkte (t, x), deren x–Komponente einen Abstand $\leq \delta$ von einer Lösung $x(t; t_0, x_0)$ mit $x_0 \in K$ hat. Für hinreichend kleines $\delta > 0$ ist K_2 Teilmenge des Definitionsbereiches D von f, da D offen ist und $K_1 \subset D$ gilt. Das betrachtete Einschrittverfahren ist deswegen konsistent auf K_2 mit einer Funktion $\varepsilon(h)$, wobei $\varepsilon(h) = Eh^p$ im Falle der Konsistenzordnung p ist. Ebenfalls erfüllt Φ auf K_2 die Lipschitzbedingung mit einer Konstanten $L > 0$.

Wir beweisen die Konvergenz nun zunächst unter der folgenden Annahme, deren Gültigkeit wir später beweisen werden:

> Für alle hinreichend feinen Gitter $\mathcal{T}$ und alle Anfangsbedingungen $(t_0, x_0) \in K$ gilt für die gemäß Definition 2.3 erzeugte Gitterfunktion $\tilde{x}$ die Beziehung $(t_i, \tilde{x}(t_i)) \in K_2$ für alle $t_i \in \mathcal{T}$. (2.8)

Zum Beweis der Konvergenz wählen wir eine Anfangsbedingung $(t_0, x_0) \in K$ und schreiben wieder kurz $x(t) = x(t; t_0, x_0)$. Mit $\tilde{x}$ bezeichnen wir die zugehörige numerisch approximierende Gitterfunktion und mit

$$e(t_i) := \|\tilde{x}(t_i) - x(t_i)\|$$

bezeichnen wir den Fehler zur Zeit $t_i \in \mathcal{T}$. Dann gilt nach den Vorüberlegungen am Anfang dieses Abschnitts

$$\begin{aligned}
e(t_i) \;=\; \|\tilde{x}(t_i) - x(t_i)\| \;&\leq\; \|\Phi(t_{i-1}, \tilde{x}(t_{i-1}), \tau_{i-1}) - \Phi(t_{i-1}, x(t_{i-1}), \tau_{i-1})\| \\
&\quad +\; \|\Phi(t_{i-1}, x(t_{i-1}), h_{i-1}) - x(t_i)\| \\
&=\; \|\Phi(t_{i-1}, \tilde{x}(t_{i-1}), h_{i-1}) - \Phi(t_{i-1}, x(t_{i-1}), h_{i-1})\| \\
&\quad +\; \|\Phi(t_{i-1}, x(t_{i-1}), h_{i-1}) - x(t_i; t_{i-1}, x(t_{i-1}))\| \\
&\leq\; (1 + Lh_{i-1})\|\tilde{x}(t_{i-1}) - x(t_{i-1})\| \;+\; h_{i-1}\varepsilon(h_{i-1}) \\
&=\; (1 + Lh_{i-1})e(t_{i-1}) + h_{i-1}\varepsilon(h_{i-1})
\end{aligned}$$

wobei wir im vorletzten Schritt die Lipschitzbedingung und die Konsistenz sowie die Tatsache, dass $(t_{i-1}, \tilde{x}(t_{i-1})) \in K_2$ liegt, ausgenutzt haben. Wir erhalten also für den Fehler $e(t_i)$ die rekursive Gleichung

$$e(t_i) \leq (1 + Lh_{i-1})e(t_{i-1}) + h_{i-1}\varepsilon(h_{i-1})$$

gemeinsam mit der "Anfangsbedingung" $e(t_0) = 0$, da $\tilde{x}(t_0) = x_0 = x(t_0)$ ist.

Mittels Induktion zeigen wir nun, dass daraus die Abschätzung

$$e(t_i) \leq \varepsilon(\bar{h})\frac{1}{L}(\exp(L(t_i - t_0)) - 1)$$

folgt. Für $i = 0$ ist die Abschätzung klar. Für $i - 1 \to i$ verwenden wir

$$\exp(Lh_i) = 1 + Lh_i + \frac{L^2h_i^2}{2} + \ldots \geq 1 + Lh_i$$

und erhalten damit mit der Induktionsannahme

$$\begin{aligned}
e(t_i) &\leq (1 + Lh_{i-1})e(t_{i-1}) + h_{i-1}\varepsilon(h_{i-1}) \\
&\leq (1 + Lh_{i-1})\varepsilon(\bar{h})\frac{1}{L}(\exp(L(t_{i-1} - t_0)) - 1) + h_{i-1}\underbrace{\varepsilon(h_{i-1})}_{\leq\varepsilon(h)} \\
&= \varepsilon(\bar{h})\frac{1}{L}\Big(h_{i-1}L + (1 + Lh_{i-1})(\exp(L(t_{i-1} - t_0)) - 1)\Big) \\
&= \varepsilon(\bar{h})\frac{1}{L}\Big(h_{i-1}L + (1 + Lh_{i-1})\exp(L(t_{i-1} - t_0)) - 1 - Lh_{i-1}\Big) \\
&= \varepsilon(\bar{h})\frac{1}{L}\Big((1 + Lh_{i-1})\exp(L(t_{i-1} - t_0)) - 1\Big) \\
&\leq \varepsilon(\bar{h})\frac{1}{L}\Big(\exp(Lh_{i-1})\exp(L(t_{i-1} - t_0)) - 1\Big) \\
&= \varepsilon(\bar{h})\frac{1}{L}(\exp(L(t_i - t_0)) - 1).
\end{aligned}$$

Damit folgt die Konvergenz und im Falle von $\varepsilon(\bar{h}) \leq E\bar{h}^p$ auch die Konvergenzordnung mit $C = E(\exp(L(T - t_0)) - 1)/L$.

Es bleibt zu zeigen, dass unsere oben gemachte Annahme (2.8) tatsächlich erfüllt ist. Wir zeigen, dass (2.8) für alle Gitter $\mathcal{T}$ gilt, deren maximale Schrittweite $\bar{h}$ die Ungleichung

$$\varepsilon(\bar{h}) \leq \frac{\delta L}{\exp(L(T - t_0)) - 1}$$

erfüllt. Wir betrachten dazu eine Lösung $\tilde{x}$ mit Anfangswert $x_0 \in K$ und beweisen die Annahme per Induktion. Für $\tilde{x}(t_0)$ ist wegen $\tilde{x}(t_0) = x_0$ nichts zu zeigen. Für den Induktionsschritt $i - 1 \to i$ sei $(t_k, \tilde{x}(t_k)) \in K_2$ für $k = 0, 1, \ldots, i - 1$. Wir müssen zeigen, dass $(t_i, \tilde{x}(t_i)) \in K_2$ liegt. Beachte, dass die oben gezeigte Abschätzung

$$e(t_i) \leq \varepsilon(\bar{h})\frac{1}{L}(\exp(L(T - t_0)) - 1)$$

bereits gilt, falls $(t_k, \tilde{x}(t_k)) \in K_2$ liegt für $k = 0, 1, \ldots, i-1$. Mit der Wahl von h folgt damit $e(t_i) \leq \delta$, also

$$\|\tilde{x}(t_i) - x(t_i)\| \leq \delta.$$

Da $(t_i, x(t_i)) \in K_1$ liegt, folgt $(t_i, \tilde{x}(t_i)) \in \{t_i\} \times \overline{B}_\delta(x(t_i)) \subset K_2$, also die gewünschte Beziehung. □

Bemerkung 2.8 (i) Schematisch dargestellt besagt Satz 2.7 das Folgende:

Lipschitzbedingung + Konsistenz	$\Rightarrow$	Konvergenz
Lipschitzbedingung + Konsistenzordnung p	$\Rightarrow$	Konvergenzordnung p

(ii) Die Schranke für $e(T)$ wächst — sogar sehr schnell — wenn die Intervallgröße $T - t_0$ wächst. Insbesondere lassen sich mit dieser Abschätzung keinerlei Aussagen über das Langzeitverhalten numerischer Lösungen machen, z.B. über Grenzwerte $\tilde{x}(t_i)$ für $t_i \to \infty$. Tatsächlich kann es passieren, dass der "numerische Grenzwert" von $\tilde{x}(t_i)$ für $t_i \to \infty$ für beliebig feine Gitter $\mathcal{T}$ weit von dem tatsächlichen Grenzwert der exakten Lösung $x(t)$ entfernt ist. Wir werden später genauer auf dieses Problem eingehen.

(iii) Der Konsistenzfehler $\varepsilon(h)h$ wird auch als *lokaler Fehler* bezeichnet, während der im Beweis abgeschätzte Fehler $e(t)$ als *globaler Fehler* bezeichnet wird. Im Falle der Konsistenzordnung p gilt $\varepsilon(h)h = O(h^{p+1})$ und $e(t) = O(h^p)$. Man "verliert" also eine Ordnung beim Übergang vom lokalen zum globalen Fehler. Dies lässt sich anschaulich wie folgt erklären: Bis zur Zeit t muss man (bei äquidistantem Gitter) gerade ca. $N(t) = (t - t_0)/h$ Schritte machen, weswegen sich $N(t)$ lokale Fehler aufsummieren, was zu dem globalen Fehler $O(h^{p+1})N(t) = O(h^{p+1})/h = O(h^p)$ führt. □

2.4 Kondition

Wie bei allen numerischen Problemen sollte auch hier die Kondition des Problems "Berechne eine Lösung des Anfangswertproblems (1.1), (1.2)" betrachtet werden. Eine detaillierte Darstellung der hierfür nötigen Theorie würde den Rahmen dieser Vorlesung leider sprengen. Wir werden hier nur kurz (ohne Beweise) beschreiben, wie sich die Kondition bzgl. Störungen Δx_0 im Anfangswert x_0 berechnen lässt, d.h., wir wollen eine Abschätzung für den Ausdruck

$$\kappa := \max_{\Delta x_0 \in \mathbb{R}^n, \|\Delta x_0\| = 1} \left\| \frac{\partial}{\partial x_0} x(t; t_0, x_0) \Delta x_0 \right\|$$

berechnen. Dazu betrachtet man das Anfangswertproblem

$$\dot{y}(t) = f_x(t, x(t; t_0, x_0))y(t), \quad y(t_0) = \Delta x_0, \tag{2.9}$$

wobei $f_x(t, x) = \frac{\partial}{\partial x} f(t, x) \in \mathbb{R}^{n \times n}$ und $x(t; t_0, x_0)$ die Lösung von (1.1), (1.2) ist. Die Lösung von (2.9) lässt sich in der Form

$$y(t; t_0, \Delta x_0) = W(t; t_0)\Delta x_0$$

mit einer Matrix $W(t; t_0) \in \mathbb{R}^{n \times n}$ schreiben. Dieses W ist dann gerade gleich der obigen Ableitung $\frac{\partial}{\partial x_0} x(t; t_0, x_0)$, die Matrix–Norm $\|W(t; t_0)\|$ gibt also gerade die Kondition κ an.

Als Beispiel betrachte die eindimensionale DGL

$$\dot{x}(t) = \lambda x(t)$$

für $\lambda \in \mathbb{R}$. Für diese Gleichung ist $f(t,x) = \lambda x$, also $f_x(t,x) = \lambda$, weswegen (2.9) die Form

$$\dot{y}(t) = \lambda y(t)$$

hat. Die Lösungen sind durch $y(t;t_0,\Delta x_0) = e^{\lambda(t-t_0)}\Delta x_0$ gegeben, es gilt also $W(t;t_0) = e^{\lambda(t-t_0)}$. Die Matrixnorm dieser 1×1–Matrix ist gerade der Betrag, da $e^{\lambda(t-t_0)}$ positiv ist, gilt also

$$\kappa = e^{\lambda(t-t_0)}.$$

Für $t >> t_0$ und $\lambda > 0$ ist das Problem also schlecht konditioniert (κ wird sehr groß), während das Problem für $t >> t_0$ und $\lambda < 0$ sehr gut konditioniert ist, da $\kappa \approx 0$ ist.

Eine ausführliche Diskussion der Kondition für gewöhnliche Differentialgleichungen findet sich im Kapitel 3 des Buches [2].

Kapitel 3

Taylor–Verfahren

Wir werden in diesem Kapitel eine spezielle Klasse von Einschrittverfahren einführen, die in der numerischen Praxis zwar eher selten verwendet werden (wir werden später sehen, wieso), für das Verständnis der weiteren Einschrittverfahren aber sehr nützlich sind.

3.1 Definition

Die Taylor–Verfahren haben ihren Namen von der zu Grunde liegenden Taylor–Formel und gehen in direkter Weise aus diesen hervor. Allerdings wird die Taylor–Formel in zunächst etwas ungewohnt erscheinender Weise angewendet: Wir verwenden den Differentialoperator L_f^i, $i \in \mathbb{N}$, der für (hinreichend oft differenzierbare) Funktionen f, $g : D \to \mathbb{R}^n$ mit $D \subseteq \mathbb{R} \times \mathbb{R}^n$ mittels

$$L_f^0 g(t,x) := g(t,x), \quad L_f^1 g(t,x) := \frac{\partial g}{\partial t}(t,x) + \frac{\partial g}{\partial x}(t,x) f(t,x), \quad L_f^{i+1} g(t,x) = L_f^1 L_f^i g(t,x)$$

definiert ist. Beachte, dass $L_f^i g$ wieder eine Funktion von D nach $\mathbb{R}^n$ ist. Der folgende Satz stellt die hier benötigte Version der Taylor–Formel vor.

Satz 3.1 Gegeben sei eine Differentialgleichung (1.1) mit p–mal stetig differenzierbarem Vektorfeld f. Sei $x(t) = x(t; t_0, x_0)$ eine Lösung dieser Differentialgleichung. Dann gilt

$$x(t) = x_0 + \sum_{i=1}^{p} \frac{(t-t_0)^i}{i!} L_f^{i-1} f(t_0, x_0) + O((t-t_0)^{p+1}),$$

wobei das O–Symbol im Sinne von Definition 2.1(iii) verwendet wird.

Beweis: Aus der Theorie der gewöhnlichen Differentialgleichungen ist bekannt, dass die Lösung $x(t)$ unter der vorausgesetzten Differenzierbarkeitsbedingung an f $p+1$–mal stetig differenzierbar nach t ist. Nach der aus der Analysis bekannten Taylor–Formel für Funktionen von $\mathbb{R}$ nach $\mathbb{R}^n$ gilt demnach

$$x(t) = x_0 + \sum_{i=1}^{p} \frac{(t-t_0)^i}{i!} \frac{d^i x}{dt^i}(t_0) + O((t-t_0)^{p+1}).$$

Zum Beweis des Satzes werden wir nun nachweisen, dass

$$\frac{d^i x}{dt^i}(t) = L_f^{i-1} f(t, x(t)) \tag{3.1}$$

ist für alle $t \in I_{t_0,x_0}$, denn dann folgt die Behauptung aus

$$\frac{d^i x}{dt^i}(t_0) = L_f^{i-1} f(t_0, x(t_0)) = L_f^{i-1} f(t_0, x_0).$$

Wir zeigen (3.1) per Induktion über i. Für $i = 1$ gilt

$$\frac{dx}{dt}(t) = f(t, x(t)) = L_f^0 f(t, x).$$

Für $i \to i+1$ beachte, dass für je zwei differenzierbare Funktionen $g : D \to \mathbb{R}^n$ und $x : \mathbb{R} \to \mathbb{R}^n$ die Gleichung

$$\frac{d}{dt} g(t, x(t)) = \frac{\partial g}{\partial t}(t, x(t)) + \frac{\partial g}{\partial x}(t, x(t)) \frac{d}{dt} x(t)$$

gilt (man nennt dies auch die *totale Ableitung* von g entlang der Funktion $x(t)$). Mit $g(t, x) = L_f^{i-1} f(t, x)$ gilt damit

$$\begin{aligned}
\frac{d^{i+1} x}{d^{i+1} t}(t) &= \frac{d}{dt}\frac{d^i x}{d^i t}(t) = \frac{d}{dt} L_f^{i-1} f(t, x(t)) = \frac{d}{dt} g(t, x(t)) \\
&= \frac{\partial g}{\partial t}(t, x(t)) + \frac{\partial g}{\partial x}(t, x(t)) \frac{d}{dt} x(t) \\
&= \frac{\partial g}{\partial t}(t, x(t)) + \frac{\partial g}{\partial x}(t, x(t)) f(t, x(t)) \\
&= L_f^1 g(t, x(t)) = L_f^1 L_f^{i-1} f(t, x(t)) = L_f^i f(t, x(t)),
\end{aligned}$$

also gerade (3.1). □

Die Idee der Taylor–Verfahren ist nun denkbar einfach: Wir verwenden die Taylor–Formel und lassen den Restterm weg.

Definition 3.2 Das *Taylor–Verfahren der Ordnung* $p \in \mathbb{N}$ ist gegeben durch

$$\Phi(t, x, h) = x + \sum_{i=1}^{p} \frac{h^i}{i!} L_f^{i-1} f(t, x).$$

□

3.2 Eigenschaften

Der folgende Satz gibt die wesentlichen Eigenschaften der Taylor–Verfahren an.

Satz 3.3 Gegeben sei eine Differentialgleichung mit p–mal stetig differenzierbarem Vektorfeld $f : D \to \mathbb{R}^n$. Dann erfüllt das Taylor–Verfahren der Ordnung p die Lipschitzbedingung und ist konsistent mit Konsistenzordnung p.

Beweis: Wir zeigen zunächst die Lipschitzbedingung. Beachte, dass in der Formulierung der Taylor–Verfahren partielle Ableitungen von f bis zur Ordnung $p-1$ auftreten. Jede der auftretenden Funktionen $L_f^{i-1} f$ ist also ein weiteres mal stetig differenzierbar, woraus (mit dem Mittelwertsatz der Differentialrechnung) folgt, dass für jede kompakte Menge $K \subset D$ Lipschitz–Konstanten $L_i > 0$ existieren, so dass $L_f^{i-1} f$ Lipschitz in x mit dieser Konstante ist. Für die Funktion Φ gilt also für alle $h \leq 1$ die Abschätzung

$$\begin{aligned} \|\Phi(t, x_1, h) - \Phi(t, x_2, h)\| &\leq \|x_1 - x_2\| + \sum_{i=1}^{p} \frac{h^i}{i!} L_i \|x_1 - x_2\| \\ &\leq \|x_1 - x_2\| + \sum_{i=1}^{p} h L_i \|x_1 - x_2\| \;=\; (1 + Lh)\|x_1 - x_2\| \end{aligned}$$

mit

$$L = \sum_{i=1}^{p} L_i.$$

Dies ist gerade die gewünschte Lipschitz–Bedingung.

Die Konsistenz sowie die behauptete Konsistenzordnung folgt direkt aus Satz 3.1. □

Bemerkung 3.4 Wenn alle auftretenden Ableitungen auf ganz D beschränkt sind, so sind auch die Konstanten in den Lipschitz– und Konsistenzabschätzungen unabhängig von K gültig, man erhält also globale Fehlerabschätzungen. □

Beachte, dass das Taylor–Verfahren der Ordnung $p = 1$ durch

$$\Phi(t, x, h) = x + h L_f^0 f(t, x) = x + h f(t, x).$$

gegeben ist, also gerade das Euler–Verfahren ist. Dies führt sofort zu dem folgenden Korollar.

Korollar 3.5 Falls f einmal stetig differenzierbar ist, so ist das Euler–Verfahren konsistent mit Konsistenzordnung $p = 1$.

Beweis: Das Taylor–Verfahren der Ordnung $p = 1$ ist gerade das Euler–Verfahren, das also nach Satz 3.3 die Konsistenzordnung $p = 1$ besitzt. □

Bemerkung 3.6 Mit einem direkten Beweis kann man die Konsistenzordnung $p = 1$ für das Euler–Verfahren auch beweisen, wenn f nur Lipschitz–stetig (in x *und* t) ist. Die Beweisidee geht wie folgt: Zunächst zeigt man, dass $\|x(t + h) - x(t)\| \leq C_1 |h|$ für ein $C_1 > 0$ und alle hinreichend kleinen h ist; dies verwendet man dann, um

$$\int_t^{t+h} \|f(\tau, x(\tau)) - f(t, x(t))\| d\tau \leq C_2 h^2$$

für ein $C_2 > 0$ zu beweisen. Damit kann man schließlich die Konsistenzordnung zeigen. □

Das Euler–Verfahren ist das einzige Taylor–Verfahren, bei dem keine Ableitungen des Vektorfeldes f auftreten. Das Auftreten der Ableitungen ist tatsächlich der Hauptgrund dafür, dass Taylor–Verfahren in der Praxis eher selten verwendet werden, da man dort Verfahren bevorzugt, die ohne explizite Verwendung der Ableitung funktionieren (auch wenn symbolische Mathematikprogramme wie z.B. MAPLE heutzutage zur automatischen Berechnung der benötigten Ableitungen verwendet werden können). Trotzdem gibt es Spezialanwendungen, in denen Taylor–Verfahren verwendet werden: Für hochgenaue Numerik, bei der Verfahren sehr hoher Ordnung ($p \geq 15$) benötigt werden, sind Taylor–Verfahren nützlich, da sie systematisch für beliebige Konsistenzordnungen hergeleitet werden können und die auftretenden Konstanten (in der Lipschitzbedingung und der Konsistenzabschätzung) durch genaue Analyse der Ableitungen und Restterme exakt abgeschätzt werden können.

Eine der Hauptanwendungen der Taylor–Verfahren bzw. der Taylor–Entwicklung aus Satz 3.1 ist die Konsistenzanalyse beliebiger Einschrittverfahren. Hier gilt der folgende Satz.

Satz 3.7 Sei $f : D \to \mathbb{R}^n$ p–mal stetig differenzierbar. Gegeben sei ein Einschrittverfahren $\Phi : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$, das $p+1$–mal stetig differenzierbar ist. Dann besitzt Φ genau dann die Konsistenzordnung $p \in \mathbb{N}$, wenn die Bedingungen

$$\Phi(t,x,0) = x \quad \text{und} \quad \frac{\partial^i \Phi}{\partial h^i}(t,x,0) = L_f^{i-1} f(t,x) \text{ für } i = 1,\ldots,p \tag{3.2}$$

für alle $(t,x) \in D$ gelten.

Beweis: Es bezeichne $\Phi_{T,p}$ das Taylor-Verfahren der Ordnung p. Die Taylor–Entwicklung von Φ nach der Variablen h in $h = 0$ ist gegeben durch

$$\Phi(t,x,h) = \Phi(t,x,0) + \sum_{i=1}^{p} \frac{h^i}{i!} \frac{\partial^i \Phi}{\partial h^i}(t,x,0) + O(h^{p+1}).$$

Sei nun (3.2) erfüllt. Dann liefert der Koeffizientenvergleich mit $\Phi_{T,p}$

$$\Phi(t,x,h) = \Phi_{T,p}(t,x,h) + O(h^{p+1})$$

Aus Satz 3.3 folgt daher

$$x(t+h;t,x) = \Phi_{T,p}(t,x,h) + O(h^{p+1}) = \Phi(t,x,h) + O(h^{p+1}),$$

was die Konsistenz zeigt.

Falls (3.2) nicht erfüllt ist, so gibt es $(t,x) \in D$, so dass entweder $\Phi(t,x,0) \neq x$ gilt (in diesem Fall setzen wir $i^* = 0$) oder

$$\frac{\partial^{i^*} \Phi}{\partial h^{i^*}}(t,x,0) \neq L_g^{i^*-1} f(t,x)$$

für ein $i^* \in \{1, \ldots, p\}$ gilt. Wenn wir i^* minimal mit dieser Eigenschaft wählen, so folgt aus dem Koeffizientenvergleich mit $\Phi_{T,p}$, dass ein $C > 0$ existiert, so dass für alle hinreichend kleinen $h > 0$ die Ungleichung

$$\|\Phi(t, x, h) - \Phi_{T,p}(t, x, h)\| > Ch^{i^*}$$

gilt. Mit Satz 3.3 und der umgekehrten Dreiecksungleichung erhalten wir daher

$$\|x(t + h, t, x) - \Phi(t, x, h)\| > Ch^{i^*} - O(h^{p+1}) > \widetilde{C}h^{i^*}$$

für geeignetes $0 < \widetilde{C} < C$ und alle hinreichend kleinen $h > 0$, was der Konsistenz widerspricht. Also folgt die behauptete Äquivalenz. □

Mit diesem Satz können wir die Konsistenzordnung beliebiger Einschrittverfahren überprüfen. Beachte, dass die Aussage über die Ordnung nur stimmt, wenn das Vektorfeld f hinreichend oft differenzierbar ist. Verfahren mit hoher Konsistenzordnung verlieren diese typischerweise, wenn das Vektorfeld der zu lösendenden DGL nicht die nötige Differenzierbarkeit besitzt!

Ein wesentlicher Nachteil dieses Satzes ist, dass die Ausdrücke $L_f^i f(t, x)$ für große i sehr umfangreich und kompliziert werden. Hier können — wie bereits erwähnt — symbolische Mathematikprogramme wie MAPLE bei den Rechnungen helfen. Das folgende MAPLE Programm berechnet die Ableitungen $L_f^i f(t, x)$ für $i = 0, \ldots, p$. (Vor der Ausführung muss der Variablen `p` natürlich ein Wert zugewiesen werden.)

```
> L[0]:=f(t,x);
> for i from 1 to p do
>   L[i] := simplify(diff(L[i-1],t) + diff(L[i-1],x)*f(t,x));
> od;
```

Die Ausgabe für `p:=3` ist

$$\begin{aligned}
L_0 &:= \mathrm{f}(t,\, x)\\
L_1 &:= (\frac{\partial}{\partial t}\,\mathrm{f}(t,\, x)) + (\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x)\\
L_2 &:= (\frac{\partial^2}{\partial t^2}\,\mathrm{f}(t,\, x)) + 2\,(\frac{\partial^2}{\partial x\,\partial t}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x) + (\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))\,(\frac{\partial}{\partial t}\,\mathrm{f}(t,\, x))\\
&\quad + (\frac{\partial^2}{\partial x^2}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x)^2 + \mathrm{f}(t,\, x)\,(\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))^2\\
L_3 &:= (\frac{\partial^3}{\partial t^3}\,\mathrm{f}(t,\, x)) + 3\,(\frac{\partial^3}{\partial x\,\partial t^2}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x) + 3\,(\frac{\partial^2}{\partial x\,\partial t}\,\mathrm{f}(t,\, x))\,(\frac{\partial}{\partial t}\,\mathrm{f}(t,\, x))\\
&\quad + (\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))\,(\frac{\partial^2}{\partial t^2}\,\mathrm{f}(t,\, x)) + 3\,(\frac{\partial^3}{\partial x^2\,\partial t}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x)^2\\
&\quad + 3\,(\frac{\partial^2}{\partial x^2}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x)\,(\frac{\partial}{\partial t}\,\mathrm{f}(t,\, x)) + (\frac{\partial}{\partial t}\,\mathrm{f}(t,\, x))\,(\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))^2\\
&\quad + 5\,\mathrm{f}(t,\, x)\,(\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))\,(\frac{\partial^2}{\partial x\,\partial t}\,\mathrm{f}(t,\, x)) + (\frac{\partial^3}{\partial x^3}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x)^3\\
&\quad + 4\,(\frac{\partial^2}{\partial x^2}\,\mathrm{f}(t,\, x))\,\mathrm{f}(t,\, x)^2\,(\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x)) + \mathrm{f}(t,\, x)\,(\frac{\partial}{\partial x}\,\mathrm{f}(t,\, x))^3
\end{aligned}$$

Diese Ausdrücke gelten für den skalaren Fall $x \in \mathbb{R}$, für höhere Dimensionen muss das MAPLE–Programm erweitert werden.

Bemerkung 3.8 Man sieht, dass die Ausdrücke tatsächlich sehr unübersichtlich werden; ebenso ist das natürlich bei den entsprechenden Termen der Einschrittverfahren. Eine Hilfe hierfür bietet ein Formalismus, der von dem neuseeländischen Mathematiker J.C. Butcher in den 1960er Jahren entwickelt wurde, und bei dem die auftretenden Ableitungen mittels einer grafischen Repräsentierung in einer Baumstruktur übersichtlich strukturiert werden. □

Kapitel 4

Explizite Runge–Kutta–Verfahren

In diesem Kapitel kommen wir zu einer der wichtigsten Klassen von Einschrittverfahren, zu denen z.B. das Euler– und das Heun–Verfahren gehören.

4.1 Definition

Bei der Konstruktion des Heun–Verfahrens haben wir das Euler–Verfahren verwendet, um einen Schätzwert für den unbekannten Wert $x(t_{i+1})$ zu erhalten. Es liegt nun nahe, diese Methode systematisch rekursiv anzuwenden, um zu Verfahren höherer Konsistenzordnung zu gelangen. Genau dies ist die Grundidee der Runge–Kutta–Verfahren.

Um die dabei entstehenden Verfahren übersichtlich zu schreiben, benötigen wir einen geeigneten Formalismus. Wir erläutern diesen am Beispiel des Heun–Verfahrens

$$\Phi(t,x,h) = x + \frac{h}{2}\Big(f(t,x) + f\Big(t+h, x+hf(t,x)\Big)\Big).$$

Wir schreiben dieses nun als

$$\begin{aligned} k_1 &= f(t,x) \\ k_2 &= f(t+h, x+hk_1) \\ \Phi(t,x,h) &= x + h\left(\frac{1}{2}k_1 + \frac{1}{2}k_2\right) \end{aligned}$$

Was zunächst vielleicht komplizierter als die geschlossene Formel aussieht, erweist sich als sehr günstige Schreibweise, wenn man weitere k_i–Terme hinzufügen will. Dies ist gerade die Schreibweise der expliziten Runge–Kutta–Verfahren.

Definition 4.1 Ein *s–stufiges explizites Runge–Kutta–Verfahren* ist gegeben durch

$$k_i = f\left(t + c_i h,\, x + h\sum_{j=1}^{i-1} a_{ij}k_j\right) \quad \text{für } i = 1,\ldots,s$$

$$\Phi(t,x,h) \quad = \quad x + h\sum_{i=1}^{s} b_i k_i.$$

Den Wert $k_i = k_i(t,x,h)$ bezeichnen wir dabei als i–te Stufe des Verfahrens. □

Die Koeffizienten eines Runge–Kutta–Verfahrens können wir mittels

$$b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_s \end{pmatrix} \in \mathbb{R}^s, \quad c = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_s \end{pmatrix} \in \mathbb{R}^s, \quad \mathcal{A} = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{s1} & \cdots & \cdots & a_{s,s-1} & 0 \end{pmatrix} \in \mathbb{R}^{s\times s}$$

kompakt schreiben. Konkrete Verfahren werden meist in Form des Butcher–Tableaus (oder Butcher–Schemas)

$$\begin{array}{c|cccc} c_1 & & & & \\ c_2 & a_{2\,1} & & & \\ c_3 & a_{3\,1} & a_{3\,2} & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s\,1} & a_{s\,2} & \cdots & a_{s\,s-1} \\ \hline & b_1 & b_2 & \cdots & b_{s-1} \quad b_s \end{array}$$

geschrieben, das wiederum auf J.C. Butcher zurückgeht.

Einfache Beispiele solcher Verfahren sind das Euler–Verfahren ($s = 1$), das Heun–Verfahren ($s = 2$) und das sogenannte *klassische Runge–Kutta–Verfahren* ($s = 4$), das von C. Runge[1] und M. Kutta[2] entwickelt wurde, und dem die ganze Verfahrensklasse ihren Namen verdankt. Diese Verfahren sind (von links nach rechts) gegeben durch die Butcher–Tableaus

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ 1 & 0 & 0 & 1 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} \quad \frac{1}{6} \end{array}$$

Beachte, dass das Euler–Verfahren sowohl das einfachste Runge–Kutta–Verfahren als auch das einfachste Taylor–Verfahren ist; es ist das einzige Verfahren, das in beiden Klassen liegt, da alle Runge–Kutta–Verfahren per Definition ohne Ableitungen von f auskommen, was gegenüber den Taylor–Verfahren einen großen Vorteil darstellt.

[1] deutscher Mathematiker, 1856–1927
[2] deutscher Mathematiker und Ingenieur, 1867–1944

Es ist in diesem Zusammenhang interessant, den Aufwand des Heun–Verfahrens und des Taylor–Verfahrens der Ordnung 2 z.B. für $x \in \mathbb{R}$ zu vergleichen, die ja die gleiche Konsistenzordnung besitzen. Beim Taylor–Verfahren der Ordnung 2 müssen in jedem Schritt $L_f^0 f(t,x) = f(t,x)$ und

$$L_f^1 f(t,x) = \frac{\partial f}{\partial t}(t,x) + \frac{\partial f}{\partial x}(t,x) f(t,x)$$

ausgewertet werden, also 3 Funktionsauswertungen; beim Heun Verfahren müssen $k_1 = f(t,x)$ und $f(t+h, x+hk_1)$, also 2 Funktionen ausgewertet werden. Der Aufwand ist folglich nur 2/3 so groß. Dieser geringere Aufwand, der bei höherer Konsistenzordnung noch deutlicher ausfällt, ist typisch für Runge–Kutta–Verfahren, ein weiterer Vorteil gegenüber den Taylor–Verfahren.

Beachte, dass Runge–Kutta–Verfahren immer die Lipschitz–Bedingung erfüllen, wenn das Vektorfeld f Lipschitz–stetig im Sinne des Eindeutigkeitssatzes 1.4 ist: Mittels Induktion sieht man leicht, dass jede Stufe k_i Lipschitz–stetig ist. Damit gilt dies auch für ihre Summe, weswegen Φ die gewünschte Bedingung erfüllt.

4.2 Konsistenz

Wir wollen nun untersuchen, wie sich die Konsistenzeigenschaften der Runge–Kutta–Verfahren über ihre Koeffizienten auszudrücken lassen. Das erste wichtige Resultat ist das folgende Lemma.

Lemma 4.2 Ein explizites Runge–Kutta–Verfahren ist genau dann konsistent, wenn die Bedingung

$$\sum_{i=1}^{s} b_i = 1$$

erfüllt ist.

Beweis: Beachte, dass ein Runge–Kutta–Verfahren von der Form

$$\Phi(t,x,h) = x + h\varphi(t,x,h)$$

mit

$$\varphi(t,x,h) = \sum_{i=1}^{s} b_i k_i(t,x,h)$$

ist. Nach Lemma 2.6 ist das Verfahren also genau dann konsistent, wenn

$$\varphi(t,x,0) = \sum_{i=1}^{s} b_i k_i(t,x,0) = f(t,x)$$

ist. Aus Definition 4.1 folgt sofort, dass $k_i(t,x,0) = f(t,x)$, also ist das Verfahren genau dann konsistent, falls $\sum_{i=1}^{s} b_i f(t,x) = f(t,x)$, was für beliebige f dann und nur dann der Fall ist, wenn $\sum_{i=1}^{s} b_i = 1$ ist. □

Etwas schwieriger wird die Sache, wenn wir Aussagen über die Konsistenzordnung machen wollen. Zunächst wollen wir eine obere Schranke für die Konsistenz beweisen.

Lemma 4.3 Für ein s–stufiges explizites Runge–Kutta–Verfahren Φ mit Konsistenzordnung p gilt die Ungleichung $p \leq s$, d.h. die Konsistenzordnung ist maximal so groß wir die Stufenzahl.

Beweis: Wir wenden das Verfahren auf das Anfangswertproblem

$$\dot{x}(t) = x(t), \;\; x(0) = 1$$

an. Für die exakte Lösung gilt hier

$$x(h;0,1) = e^h = 1 + h + \frac{h^2}{2!} + \cdots + \frac{h^s}{s!} + \frac{h^{s+1}}{(s+1)!} + O(h^{s+2}).$$

Andererseits sieht man durch Induktion über i, dass $k_i(0,1,\cdot) \in \mathcal{P}_{i-1}$ ist, also ein Polynom vom Grad $\leq i-1$ in h ist. Also ist $\Phi(0,1,\cdot) \in \mathcal{P}_s$, weswegen in $\Phi(0,1,h)$ kein Term der Form ah^{s+1} auftreten kann. Daher gilt für jede Konstante $C > 0$ und hinreichend kleines $h > 0$ die Abschätzung

$$\|x(h;0,1) - \Phi(0,1,h)\| \geq \frac{h^{s+1}}{(s+1)!} - O(h^{s+2}) \geq \left(\frac{1}{h(s+1)!} - \widetilde{C}\right) h^{s+2} \geq Ch^{s+2},$$

weswegen die Konsistenzordnung maximal s sein kann, also $p \leq s$ gilt. □

Um nun genauere Aussagen über die Konsistenzordnung zu machen, empfiehlt es sich, die zu betrachtenden Differentialgleichungen etwas zu vereinfachen: Wir wollen uns auf autonome DGL einschränken. Damit wir trotzdem Aussagen für allgemeine Probleme erhalten können, überlegen wir uns zuerst, dass dies keine echte Einschränkung ist. Tatsächlich kann man aus jeder Differentialgleichung

$$\dot{x}(t) = f(t, x(t)), \quad x(t_0) = x_0 \tag{4.1}$$

mittels

$$y = \begin{pmatrix} x \\ s \end{pmatrix}, \quad \hat{f}(y) = \begin{pmatrix} f(s,x) \\ 1 \end{pmatrix}$$

(mit $s \in \mathbb{R}$) eine *autonome* Differentialgleichung

$$\dot{y}(t) = \hat{f}(y(t)), \quad y(t_0) = y_0 = \begin{pmatrix} x_0 \\ t_0 \end{pmatrix} \tag{4.2}$$

machen, für deren Lösungen die Beziehung

$$y(t;t_0,y_0) = \begin{pmatrix} x(t;t_0,x_0) \\ t \end{pmatrix} \tag{4.3}$$

gilt. Die ursprüngliche Lösung $x(t;t_0,x_0)$ von (4.1) findet sich also gerade in den ersten n Komponenten der $n+1$–dimensionalen Lösung $y(t;t_0,y_0)$ der autonomen Gleichung (4.2) wieder. Mit anderen Worten kann jede DGL im $\mathbb{R}^n$ in eine autonome DGL im $\mathbb{R}^{n+1}$ umgewandelt werden, dieses Verfahren nennt man *Autonomisierung*. Beachte, dass die neue DGL die Bedingungen des Eindeutigkeitssatzes nur dann erfüllt, wenn f Lipschitz–stetig bezüglich x *und* t ist, was eine stärkere Forderung als die Lipschitz–Stetigkeit bzgl. x ist.

Da wir diese Bedingung für unsere numerischen Aussagen aber sowieso immer benötigen (meist nehmen wir ja sogar Differenzierbarkeit von f bzgl. x und t an), stellt diese Annahme für unsere numerischen Untersuchungen keine Einschränkung dar.

Wir betrachten nun die von einem Runge–Kutta–Verfahren Φ erzeugten approximativen Lösungen $\tilde{x}(t_i)$ und $\tilde{y}(t_i)$ der Gleichungen (4.1) und (4.2). Unser Ziel ist es, uns bei der folgenden Konsistenzordnungsanalyse auf autonome Gleichungen einzuschränken. Damit wir dabei trotzdem Resultate für allgemeine nichtautonome Gleichungen erhalten können, also die für (4.2) gültigen Resultate auf (4.1) übertragen können, muss hier die zu (4.3) analoge Beziehung

$$\tilde{y}(t_i) = \begin{pmatrix} \tilde{x}(t_i) \\ t_i \end{pmatrix} \tag{4.4}$$

gelten. Ein Runge–Kutta–Verfahren, das (4.4) erfüllt, wird *invariant unter Autonomisierung* genannt. Nicht jedes Runge–Kutta–Verfahren ist aber invariant unter Autonomisierung. Das folgende Lemma gibt die dafür notwendige und hinreichende Bedingung an.

Lemma 4.4 Ein explizites Runge–Kutta–Verfahren ist genau dann invarant unter Autonomisierung, wenn es konsistent ist und die Bedingung

$$c_i = \sum_{j=1}^{i-1} a_{ij}$$

für $i = 1, \ldots, s$ erfüllt ist.

Beweis: Wir bezeichnen das Verfahren für (4.1) mit Φ und das Verfahren für (4.2) mit $\widehat{\Phi}$, die zugehörigen Stufen bezeichnen wir mit k_i und $\hat{K}_i = (\hat{k}_i, \theta_i)^T$. Das Verfahren ist genau dann invariant unter Autonomisierung, wenn

$$\widehat{\Phi}(t, x, h) = \begin{pmatrix} \Phi(t, x, h) \\ t + h \end{pmatrix} \tag{4.5}$$

gilt, da sich (4.4) dann mittels Induktion über i ergibt. Wegen

$$\widehat{\Phi}(t, x, h) = \begin{pmatrix} x + h \sum_{i=1}^{s} b_i \hat{k}_i \\ t + h \sum_{i=1}^{s} b_i \theta_i \end{pmatrix} \quad \text{und} \quad \Phi(t, x, h) = x + h \sum_{i=1}^{s} b_i k_i$$

gilt (4.5) genau dann, wenn

$$\hat{k}_i = k_i \quad \text{und} \quad t + h \sum_{i=1}^{s} b_i \theta_i = t + h \tag{4.6}$$

erfüllt ist. Für $\hat{k}_i$ und θ_i gilt gerade

$$\begin{pmatrix} \hat{k}_i \\ \theta_i \end{pmatrix} = \begin{pmatrix} f\left(t + h \sum_{j=1}^{i-1} a_{ij} \theta_j,\ x + h \sum_{j=1}^{i-1} a_{ij} \hat{k}_j\right) \\ 1 \end{pmatrix}.$$

Wegen $k_i = f\left(t + c_i h,\ x + h \sum_{j=1}^{i-1} a_{ij} k_j\right)$ und $\theta_j = 1$ ergibt sich, dass die erste Gleichung in (4.6) genau dann gilt, wenn $c_i = \sum_{j=1}^{i-1} a_{ij}$ gilt. Wegen $\theta_i = 1$ gilt die zweite Gleichung

in (4.6) genau dann, wenn $t + h\sum_{i=1}^{s} b_i = t + h$ erfüllt ist, also wenn $\sum_{i=1}^{s} b_i = 1$ ist, was gerade äquivalent zur Konsistenz ist. □

Auf Basis dieses Lemmas können wir uns also im Folgenden auf autonome DGL einschränken, wenn wir Verfahren betrachten, die die Bedingung von Lemma 4.4 erfüllen. Dies hat den Vorteil, dass sich der Differentialoperator L_f^1 zu

$$L_f^1 g(x) := \left(\frac{d}{dx} g(x)\right) f(x)$$

vereinfacht, was die Taylorentwicklung deutlich übersichtlicher macht. Dies wird im folgenden Satz ausgenutzt.

Satz 4.5 Betrachte ein Runge–Kutta–Verfahren, das die Bedingung aus Lemma 4.4 erfüllt. Dann gilt für alle Vektorfelder $f \in C^p(D, \mathbb{R}^n)$:

(i) Das Verfahren besitzt genau dann die Konsistenzordnung $p = 1$, wenn die Gleichung

$$\sum_i b_i = 1$$

gilt.

(ii) Es besitzt genau dann die Konsistenzordnung $p = 2$, wenn zusätzlich zu (i) die Gleichung

$$\sum_i b_i c_i = 1/2$$

gilt.

(iii) Es besitzt genau dann die Konsistenzordnung $p = 3$, wenn zusätzlich zu (i), (ii) die Gleichungen

$$\sum_i b_i c_i^2 = 1/3, \quad \sum_{ij} b_i a_{ij} c_j = 1/6$$

gelten.

(iv) Es besitzt genau dann die Konsistenzordnung $p = 4$, wenn zusätzlich zu (i)–(iii) die Gleichungen

$$\sum_i b_i c_i^3 = 1/4, \qquad \sum_{ij} b_i a_{ij} c_i c_j = 1/8$$

$$\sum_{ij} b_i a_{ij} c_j^2 = 1/12, \qquad \sum_{ijk} b_i a_{ij} a_{jk} c_k = 1/24$$

gelten.

Hierbei laufen die Summations–Indizes in den Grenzen $i = 1, \ldots, s$, $j = 1, \ldots, i-1$ und $k = 1, \ldots, j-1$.

Beweis: Die Gleichung ergeben sich aus der Bedingung (3.2), wobei die für $p \in \mathbb{N}$ angegebenen Gleichungen gerade äquivalent zu der Bedingung

$$\frac{\partial^p \Phi}{\partial h^p}(x, 0) = L_f^{p-1} f(x) \tag{4.7}$$

aus (3.2) sind. Für $p = 1$ ergibt sich die angegebene Gleichung dabei aus den gleichen Rechnungen wie im Beweis von Lemma 4.2.

Wir zeigen die Behauptung hier exemplarisch für $p = 2$, die höheren Ordnungen folgen mit der gleichen Beweistechnik, allerdings mit aufwändigeren Rechnungen.

Wir zeigen also, dass die in (ii) angegebene Gleichung äquivalent zu (4.7) für $p = 2$ ist. Die zweite Ableitung von $\Phi = x + h\varphi$ nach h ist gerade

$$\begin{aligned}\frac{\partial^2\Phi}{\partial h^2} &= \frac{\partial}{\partial h}\frac{\partial}{\partial h}(x + h\varphi) = \frac{\partial}{\partial h}\left(\varphi + h\frac{\partial}{\partial h}\varphi\right)\\ &= \frac{\partial}{\partial h}\varphi + \frac{\partial}{\partial h}\varphi + h\frac{\partial^2}{\partial h^2}\varphi = 2\frac{\partial}{\partial h}\varphi + h\frac{\partial^2}{\partial h^2}\varphi\end{aligned}$$

In $h = 0$ ergibt sich damit

$$\frac{\partial^2\Phi}{\partial h^2}(x,0) = 2\frac{\partial}{\partial h}\varphi(x,0) = 2\sum_{i=1}^{s} b_i \sum_{j=1}^{i-1} a_{ij}\left(\frac{d}{dx}f(x)\right)f(x).$$

Andererseits ist die Ableitung $L_f^1 f(x)$ gerade durch

$$L_f^1 f(x) = \left(\frac{d}{dx}f(x)\right)f(x)$$

gegeben ist. Damit diese Ausdrücke für alle $f(x)$ übereinstimmen, muss also gerade

$$2\sum_{i=1}^{s} b_i \sum_{j=1}^{i-1} a_{ij} = 1$$

gelten, was wegen der angenommen Autonomieinvarianzbedingung

$$c_i = \sum_{j=1}^{i-1} a_{ij}$$

genau dann der Fall ist, wenn die Gleichung aus (ii) erfüllt ist. □

Diese Gleichungen an die Koeffizienten werden *Bedingungsgleichungen* genannt. Wie komplex das Problem des Aufstellens der Bedingungsgleichungen für große p wird, zeigt die folgende Tabelle, die die Anzahl der Gleichungen für gegebenes p angibt.

Konsistenzordnung p	1	2	3	4	5	6	7	8	9	10	20
Anzahl Bedingungsgl'en	1	2	4	8	17	37	85	200	486	1205	20247374

Nicht nur das Aufstellen, auch das Lösen dieser (nichtlinearen!) Gleichungssysteme wird ziemlich komplex. Hier kommt wieder das in Bemerkung 3.8 bereits erwähnte grafische Verfahren von Butcher ins Spiel. Mit diesem Verfahren können die einzelnen Terme der $L_f^i f$–Ableitungen ebenso wie die Terme der Ableitungen von Φ mittels einer Baumstruktur grafisch dargestellt werden. Dieses Verfahren erlaubt eine Einsicht in die Struktur dieser riesigen nichtlinearen Gleichungssysteme, womit es gelungen ist, die Gleichungen bis $p = 10$

(ohne Computerhilfe) zu lösen. Eine wichtige Rolle spielt dabei natürlich die Stufenzahl s der betrachtetem Verfahren. Insbesondere ist hierbei wichtig, wie viele Stufen s man zur Realisierung einer gegebenen Konsistenzordnung p benötigt. Die folgende Tabelle gibt die ebenfalls durch Butcher (in den Jahren 1964–1985) berechneten bekannten minimalen Schranken an.

Konsistenzordnung p	1	2	3	4	5	6	7	8	≥ 9
minimale Stufenzahl s	1	2	3	4	6	7	9	11	$\geq p+3$

Der Eintrag für $p \geq 9$ bedeutet nicht, dass für jedes $p \geq 9$ ein Verfahren mit $s = p + 3$ Stufen bekannt ist, sondern dass es kein Verfahren mit weniger Stufen geben kann. Für $p = 10$ wurde 1978 von E. Hairer ein Verfahren mit $s = 17$ Stufen angegeben, das sich im Guinness–Buch der Rekorde findet. Möglichst wenig Stufen zu verwenden ist allerdings nicht das einzige Qualitätsmerkmal für Runge–Kutta–Verfahren, oftmals spielen andere Kriterien eine wichtigere Rolle. Wir kommen später darauf zurück.

Kapitel 5

Implizite Runge–Kutta–Verfahren

5.1 Definition

Bisher haben wir Runge–Kutta–Verfahren betrachtet, bei denen die Koeffizientenmatrix die Form

$$A = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{s1} & \cdots & \cdots & a_{s,s-1} & 0 \end{pmatrix} \in \mathbb{R}^{s\times s}$$

hatte. Es stellt sich nun die Frage, was passiert, wenn wir hier "volle" Matrizen der Form

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & & \vdots \\ a_{s1} & \cdots & a_{ss} \end{pmatrix} \in \mathbb{R}^{s\times s}$$

zulassen. Zunächst einmal können wir auch mit solchen Koeffizienten ganz formal durch Erweiterung von Definition 4.1 wieder Runge–Kutta–Verfahren definieren.

Definition 5.1 Ein *s–stufiges implizites Runge–Kutta–Verfahren* ist gegeben durch

$$\begin{aligned} k_i &= f\left(t + c_i h,\, x + h\sum_{j=1}^{s} a_{ij}k_j\right) \quad \text{für } i = 1,\ldots,s \\ \Phi(t,x,h) &= x + h\sum_{i=1}^{s} b_i k_i. \end{aligned}$$

Den Wert $k_i = k_i(t,x,h)$ bezeichnen wir dabei als i–te Stufe des Verfahrens. □

Der Grund für den Namen *implizites Verfahren* liegt darin, dass die Definition der k_i nun keine "Zuweisung" mehr ist, sondern ein s–dimensionales nichtlineares Gleichungssystem

bildet, dessen Lösung gerade der Vektor $k^T = (k_1^T, \ldots, k_s^T) \in \mathbb{R}^{s \cdot n}$ ist. Die Werte $k_i \in \mathbb{R}^n$ sind also *implizit* definiert.

Das einfachste Verfahren dieser Klasse ist durch das Butcher–Tableau

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

gegeben. Ausgeschrieben lautet es

$$k_1 = f(t + h, x + hk_1), \quad \Phi(t, x, h) = x + hk_1,$$

die dadurch erzeugt Gitterfunktion ist rekursiv gegeben durch

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_i) + h_i f(t_{i+1}, \tilde{x}(t_{i+1})).$$

Dieses Verfahren heißt *implizites Euler–Verfahren* und besitzt genau wie sein explizites Gegenstück die Konsistenzordnung $p = 1$. Beachte, dass hier tatsächlich in jedem Schritt ein nichtlineares Gleichungssystem gelöst werden muss. Implizite Runge–Kutta–Verfahren mit Konsistenzordnung $p = 2$ sind z.B. die implizite Mittelpunktregel oder die implizite Trapezregel, die durch

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \quad \text{bzw.} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

gegeben sind.

Wir werden später sehen, dass implizite Verfahren für manche Differentialgleichungen gegenüber den expliziten Verfahren deutliche Vorteile besitzen. Zunächst wollen wir uns aber Gedanken darüber machen, wie solch ein implizites Verfahren implementiert werden kann, d.h., wie wir das nichtlineare Gleichungssystem zur Berechnung der k_i lösen können.

5.2 Lösbarkeit und Implementierung

Zunächst einmal gibt es manchmal die Möglichkeit, die entstehenden Gleichungen per Hand in explizite Form zu bringen. Betrachten wir z.B. das implizite Euler–Verfahren angewendet auf die eindimensionale DGL

$$\dot{x}(t) = \lambda x(t),$$

so erhalten wir

$$k_1 = f(t + h, x + hk_1) = \lambda(x + hk_1) = \lambda x + h\lambda k_1,$$

woraus für hinreichend kleine h die Gleichung

$$k_1 = \frac{\lambda x}{1 - h\lambda}$$

folgt.

Oft kommt man mit dieser Strategie aber nicht weiter, wir müssen das entstehende Gleichungssystem

$$k = F(k)$$

mit

$$k = \begin{pmatrix} k_1 \\ \vdots \\ k_s \end{pmatrix} \in \mathbb{R}^{s\cdot n} \quad \text{und} \quad F(k) = \begin{pmatrix} f\left(t + c_1 h,\, x + h\sum_{j=1}^{s} a_{1j}k_j\right) \\ \vdots \\ f\left(t + c_s h,\, x + h\sum_{j=1}^{s} a_{sj}k_j\right) \end{pmatrix}$$

also numerisch lösen.

Eine einfache Möglichkeit hierzu beruht auf der Tatsache, dass f nach Voraussetzung Lipschitz–stetig mit Konstante L ist. Hieraus folgt sofort, dass auch die Abbildung F Lipschitz–stetig mit Konstante hL ist. Falls $hL =: K < 1$ ist, folgt damit

$$\|F(k^1) - F(k^2)\| \leq K\|k^1 - k^2\|,$$

so dass F eine Kontraktion ist, weswegen der Vektor k mittels der aus der Einführung in die Numerik bekannten Fixpunktiteration

$$k^{(j+1)} = F(k^{(j)}) \tag{5.1}$$

berechnet werden kann. Als Startwert für diese Iteration empfiehlt es sich, im ersten Schritt $k_i^{(0)} = f(t+c_i h, x)$ und in den folgenden Schritten den Wert von k aus dem vorhergehenden Schritt zu verwenden. Ein geeignetes Abbruchkriterium ergibt sich wie in der Einführung in die Numerik diskutiert aus dem Banach'schen Fixpunktsatz: Die Iteration wird so lange durchgeführt, bis

$$\|k^{(j+1)} - k^{(j)}\| \leq \varepsilon$$

für eine vorgegebene Toleranz ε ist, damit ist dann die Genauigkeit

$$\|k^{(j+1)} - k^*\| \leq \frac{hL}{1-hL}\varepsilon$$

garantiert, wobei k^* die exakte Lösung bezeichnet. Als Letztes müssen wir uns noch überlegen, wie ε gewählt werden sollte. Damit das Verfahren

$$\Phi(t, x, h) = x + h\sum_{i=1}^{s} b_i k_i$$

den Konsistenzfehler $O(h^{p+1})$ einhält, sollte $\|k^{(j+1)} - k^*\| \leq \varepsilon_0 h^p$ für ein $\varepsilon_0 > 0$ gelten. Damit diese Schranke eingehalten wird, muss

$$\frac{hL}{1-hL}\varepsilon \leq \varepsilon_0 h^p$$

gelten, was für kleine h gerade durch die Wahl $\varepsilon \approx \varepsilon_0 h^{p-1}$ garantiert wird. Das Abbruchkriterium hängt für $p \geq 2$ also von der Schrittweite h ab.

Die Iteration (5.1) wird auch *Gesamtschrittiteration* genannt. Eine einfache Modifikation dieser Iteration ist die *Einzelschrittiteration*, die durch die Vorschrift

$$k_i^{(j+1)} = f\left(t + c_i h,\, x + h\sum_{l=1}^{i-1} a_{il} k_l^{(j+1)} + h\sum_{l=i}^{s} a_{il} k_l^{(j)}\right), \quad i = 1, \ldots, s \tag{5.2}$$

gegeben ist. Dies ist ein ähnlicher Trick, wie wir ihn in der Einführung in die Numerik beim Übergang vom Jacobi– zum Gauß–Seidel–Verfahren angewendet haben: Wir verwenden die bereits bekannten Werte $k_1^{j+1}, \ldots, k_{i-1}^{j+1}$ der $j+1$–ten Iteration bei der Berechnung von k_i^{j+1}. Im Allgemeinen konvergiert die Einzelschrittiteration (5.2) etwas schneller als die Gesamtschrittiteration (5.1).

Falls die Lipschitz–Konstante L des Vektorfeldes groß ist, werden bei diesen Fixpunktiterationen sehr kleine Zeitschrittweiten $h > 0$ benötigt, um die Kontraktionsbedingung $K = hL < 1$ sicher zu stellen. In diesem Falle können andere Verfahren vorteilhaft sein. So kann man das Problem $k = F(k)$ in ein geeignetes Nullstellenproblem umwandeln, z.B. mittels $0 = G(k) := k - F(k)$ (es gibt weitere, u.U. numerisch günstigere äquivalente Nullstellenprobleme, vgl. [2], Abschnitt 6.2.2). Wenn man nun die Ableitung DG ausrechnen kann, die sich aus der Ableitung $\partial/\partial x f(x)$ ergibt, so ist das Newton–Verfahren sehr gut geeignet, da man mit k aus dem vorhergehenden Schritt bzw. mit $k_i = f(t + c_i, x)$ einen guten Startwert für das (ja nur lokal konvergente) Newton–Verfahren besitzt.

Zusammenfassend führt dies auf den folgenden Algorithmus.

Algorithmus 5.2 (Lösung eines Anfangswertproblems mit implizitem Runge–Kutta–Verfahren)

Eingabe: Anfangsbedingung (t_0, x_0), Endzeit T, Schrittzahl N, Einschrittverfahren Φ

(1) Setze $h := (T - t_0)/N$, $\tilde{x}_0 = x_0$

(2) Für $i = 0, \ldots, N-1$:

(2a) Berechne $t_{i+1} = t_i + h$ und löse das nichtlineare Gleichungssystem $k = F(k)$

(2b) Berechne $\tilde{x}_{i+1} := \Phi(t_i, \tilde{x}_i, h) = \tilde{x}_i + h\sum_{j=1}^{s} b_j k_j$

Ausgabe: Werte der Gitterfunktion $\tilde{x}(t_i) = \tilde{x}_i$ in $t_0, \ldots, t_N$ □

Die Analyse impliziter Runge–Kutta–Verfahren ist im Vergleich zu den expliziten Verfahren komplizierter, da die Ableitungen von Φ (mit denen man sowohl die Konsistenz gemäß Satz 3.7 als auch die Lipschitz–Bedingung über die Ableitung nach x überprüfen kann) mit Hilfe des Satzes über implizite Funktionen berechnet werden müssen. Die Grundideen der Beweise sind aber gleich und die resultierenden Bedingungsgleichungen sind identisch zu denen in Satz 4.5, weswegen wir die technischen Details hier nicht vertiefen wollen.

Bemerkung 5.3 Für explizite Runge–Kutta–Verfahren haben wir in Lemma 4.3 gesehen, dass die Stufenanzahl s eine obere Schranke für die Konsistenzordnung p bildet, also immer $p \leq s$ gilt. Für implizite Verfahren ist die Schranke nicht ganz so strikt: Für ein s–stufiges implizites Runge–Kutta–Verfahren Φ mit Konsistenzordnung p gilt die Ungleichung $p \leq 2s$,

d.h. die Konsistenzordnung ist maximal zwei mal so groß wie die Stufenzahl. Zum Beweis dieser Aussage wenden wir das Verfahren wieder auf das Anfangswertproblem

$$\dot{x}(t) = x(t), \;\; x(0) = 1$$

mit exakter Lösung e^t an. Man kann nun zeigen, dass die numerische Lösung von der Form

$$\Phi(0, 1, h) = P(h)/Q(h)$$

für zwei Polynome $P, Q \in \mathcal{P}_s$ mit $Q \not\equiv 0$ ist. Falls nun $\Phi(0, 1, h) - e^h = O(h^{2s+2})$ gilt, so folgt auch $P(h) - Q(h)e^h = O(h^{2s+2})$. Mittels Induktion über s zeigt man dann, dass dies nur für $P \equiv Q \equiv 0$ gelten kann, was ein Widerspruch zu $Q \not\equiv 0$ ist. Also kann $\Phi(0, 1, h) - e^h = O(h^{2s+2})$ nicht gelten, weswegen im besten Fall $\Phi(0, 1, h) - e^h = O(h^{2s+1})$ sein kann, also $p \leq 2s$. □

Während es bei expliziten Runge–Kutta–Verfahren sehr schwierig ist, Verfahren für große p zu konstruieren, lässt sich die maximale Konsistenzordnung $p = 2s$ bei impliziten Verfahren relativ leicht realisieren. Wiederum auf Butcher geht nämlich die Familie der *Gauß–Verfahren* zurück, bei denen sich die Koeffizienten durch Nullstellen der Legendre–Polynome (ähnlich wie bei der Gauß–Quadratur) ermitteln lassen und die eine Familie von impliziten Verfahren mit $p = 2s$ bildet. Für Details siehe Abschnitt 9.2.

Kapitel 6

Steife Differentialgleichungen

Steife Differentialgleichungen sind eine Klasse von Differentialgleichungen, die mit expliziten Verfahren nur schwer zu lösen sind. Sie bilden die Hauptmotivation dafür, implizite Verfahren zu betrachten und zu verwenden. Leider ist es nicht ganz leicht, einer Differentialgleichung anzusehen, ob sie "steif" ist; es ist nicht einmal leicht, diese Eigenschaft formal zu definieren. Vielleicht ist die informelle Beschreibung "mit expliziten Verfahren schwer zu lösen" bereits die beste mögliche Definition. Wir wollen aber trotzdem versuchen, diese Eigenschaft etwas zu formalisieren und gewisse Kriterien herausarbeiten, an denen man erkennen kann, ob man es mit einer steifen DGL zu tun hat.

Wir wollen dazu zunächst den Begriff "schwer zu lösen" etwas genauer fassen. Aus Satz 2.7 wissen wir, dass für allgemeine Einschrittverfahren mit Konvergenzordnung $p > 0$ die Abschätzung der Form

$$\|\tilde{x}(t_i) - x(t_i)\| \leq CEh^p$$

für alle hinreichend kleinen $h > 0$ gilt, wobei $E > 0$ aus der Konsistenzbedingung stammt und

$$C = \frac{1}{L}(\exp(L(t_i - t_0)) - 1)$$

von der Konstanten L der Lipschitzbedingung sowie von der Größe des Zeitintervalls $T - t_0$ abhängt. Eine Differentialgleichung ist nun schwer zu lösen, wenn CE eine sehr große Konstante ist oder wenn die Abschätzung für $\|\tilde{x}(t_i) - x(t_i)\|$ nur für sehr kleine $h > 0$ gilt. Was "sehr groß" bzw. "sehr klein" in diesem Zusammenhang bedeutet, hängt im Wesentlichen davon ab, wieviel Zeit man in die Berechnung der Lösung investieren möchte und wie kleine Zeitschritte man noch zulassen möchte. Eine genaue Schranke kann man — ähnlich wie bei der Frage "wann ist ein Problem schlecht konditioniert?" — nicht angeben.

Sicherlich muss man damit rechnen, dass eine Differentialgleichung schwer zu lösen ist, wenn sie schlecht konditioniert ist. Steife Differentialgleichungen zeichnen sich nun dadurch aus, dass sie mit expliziten Verfahren schwer zu lösen sind, *obwohl* sie gut konditioniert sind. Dass dies tatsächlich passieren kann, wollen wir an einem bereits bekannten Beispiel illustrieren: Wir betrachten wieder die 1d DGL

$$\dot{x}(t) = \lambda x(t)$$

mit $\lambda \in \mathbb{R}$. Für diese Gleichung hatten wir gesehen, dass sie die Kondition

$$\kappa = e^{\lambda(t-t_0)}.$$

besitzt und deswegen für $t >> t_0$ und $\lambda < 0$ sehr gut konditioniert ist, da $\kappa \approx 0$ ist. Wir wollen die exakte Lösung $x(t; x_0) = e^{\lambda t} x_0$ dieser Gleichung für $\lambda << 0$ mit der numerischen Approximation durch das Euler–Verfahren vergleichen. Diese Approximation ist gegeben durch

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_i) + h\lambda\tilde{x}(t_i) = (1 + h\lambda)\tilde{x}(t_i).$$

Durch Induktion sieht man leicht, dass die Euler–Lösung für $t_i = hi$ damit gerade durch

$$\tilde{x}(t_i) = (1 + h\lambda)^i x_0$$

gegeben ist. Für kleine $\lambda < 0$ konvergiert die exakte Lösung z.B. mit Anfangswert $x_0 = 1$ sehr schnell gegen 0. Damit die Euler–Lösung eine vernünftige Approximation darstellt, sollte diese also auch gegen Null streben. Damit dies passiert, muss $|1 + h\lambda| < 1$ sein, was für negative λ genau dann der Fall ist, wenn $|h\lambda| < 2$, also

$$h < 2/|\lambda|$$

ist. Z.B. für $\lambda = -10000$ müssen wir den Zeitschritt $h < 1/5000$ wählen, um überhaupt eine halbwegs sinnvolle Approximation zu erhalten und das, obwohl die Gleichung sehr gut konditioniert ist.

Zum Vergleich betrachten wir nun das implizite Euler–Verfahren, das durch

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_i) + h\lambda\tilde{x}(t_{i+1}) \quad \Leftrightarrow \quad \tilde{x}(t_{i+1}) = \frac{\tilde{x}(t_i)}{1 - h\lambda}$$

gegeben ist. Die approximierte Lösung ist also

$$\tilde{x}(t_i) = \frac{1}{(1 - h\lambda)^i} x_0.$$

Hier strebt die Lösung genau dann gegen Null, wenn $|1/(1 - h\lambda)| < 1$ ist, also wenn $|1 - h\lambda| > 1$ ist. Da $\lambda < 0$ ist, ist diese Bedingung für sämtliche Zeitschritte $h > 0$ erfüllt, die Lösung konvergiert also für alle Zeitschritte gegen Null und stellt damit eine sinnvolle Approximation dar. Abbildung 6.1 zeigt die exakte Lösung sowie die numerischen Approximationen für $\lambda = -100$ für verschiedene Zeitschritte.

6.1 Stabilität

Für die 1d–Gleichung $\dot{x}(t) = \lambda x(t)$ können wir also sagen, dass sie steif ist, wenn $\lambda < 0$ und $|\lambda|$ groß ist. Wir wollen dieses Kriterium auf eine größere Klasse von Differentialgleichungen verallgemeinern.

Wir betrachten dazu die Klasse der *linearen zeitinvarianten* DGL, die gegeben ist durch

$$\dot{x}(t) = Ax(t), \tag{6.1}$$

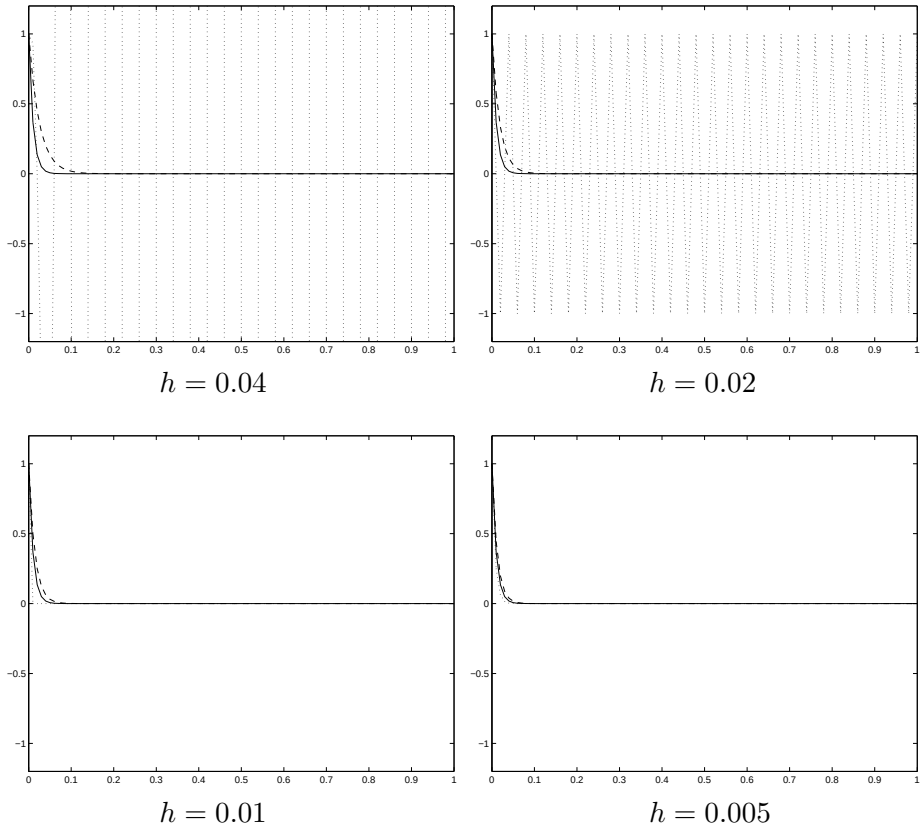

Abbildung 6.1: Exakte Lösung (durchgezogen), explizite Euler–Lösung (gepunktet) und implizite Euler–Lösung (gestrichelt) für $\dot{x}(t) = \lambda x(t)$, $x(0) = 1$, $\lambda = -100$

wobei $x(t) \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times n}$ ist. Die Idee, diese Klasse von Differentialgleichungen zu betrachten, geht auf Germund Dahlquist[1] zurück. Für solche Gleichungen sind die durch ein Runge–Kutta–Verfahren erzeugten approximativen Lösungen stets von der Form

$$\tilde{x}(t_{i+1}) = \widetilde{A}\tilde{x}(t_i) \tag{6.2}$$

für ein $\widetilde{A} \in \mathbb{R}^{n\times n}$. Wir beschränken uns in diesem Abschnitt auf den Fall äquidistanter Zeitschritte $h_i = h$ und $t_0 = 0$, woraus sich $t_i = hi$ ergibt. Eine Gleichung der Form (6.2) wird *lineare zeitinvariante Differenzengleichung* genannt. Wir bezeichnen die Lösungen von (6.2) mit $\tilde{x}(0) = x_0$ mit $\tilde{x}(t; x_0)$, wobei $t \in \mathbb{R}$ ein Vielfaches von h ist. Offenbar gilt gerade $\tilde{x}(hi; x_0) = \widetilde{A}^i x_0$.

Für das explizite Euler–Verfahren gilt z.B. $\widetilde{A} = \mathrm{Id} + hA$, während für das implizite Euler–Verfahren $\widetilde{A} = (\mathrm{Id} - hA)^{-1}$ gilt, wobei $\mathrm{Id} \in \mathbb{R}^{n\times n}$ die Einheitsmatrix bezeichnet. Genauer beschreibt das folgende Lemma, wie A und $\widetilde{A}$ zusammenhängen.

Lemma 6.1 Für jedes s–stufige Runge–Kutta–Verfahren lässt sich die Matrix $\widetilde{A}$ in (6.2) als

$$\widetilde{A} = R(hA)$$

[1] schwedischer Mathematiker, 1925-2005

schreiben, wobei R eine von h unabhängige Funktion ist. Für explizite Runge–Kutta–Verfahren ist R ein Polynom vom Grad $\leq s$, für implizite Verfahren ist R eine rationale Funktion, d.h. eine Funktion der Form $R(z) = P(z)Q(z)^{-1}$, wobei P und Q wieder Polynome vom Grad $\leq s$ sind.

Beweis: Es seien a_{ij} und b_i die Koeffizienten des Verfahrens. Dann gilt für die Stufen k_i bei Anwendung auf (6.1) die Beziehung

$$hk_i = hAx + \sum_{j=1}^{s} a_{ij} hAhk_j$$

wobei wir beim expliziten Verfahren die Konvention $a_{ij} = 0$ für $j \geq i$ machen. Im expliziten Fall folgt per Induktion, dass jedes hk_i ein Polynom in hA vom Grad $\leq i$ ist und linear in x ist. Damit ist $\Phi(t, x, h) = x + \sum b_i hk_i$ ein Polynom vom Grad $\leq s$ in hA und linear in x, also gerade von der behaupteten Form.

Im impliziten Fall erhalten wir

$$\left(hk_i - \sum_{j=1}^{s} a_{ij} hAhk_j \right) = hAx$$

für $i = 1, \ldots, s$. Der $n \cdot s$–dimensionale Vektor $k = (k_1^T, \ldots, k_s^T)^T$ ist also gerade die Lösung eines $n \cdot s$–dimensionalen linearen Gleichungssystems, dessen Matrix affin linear von A und dessen rechte Seite linear von A und x abhängt. Durch Auflösen dieses Gleichungssystems sieht man (nach länglicher Rechnung, die wir hier nicht durchführen wollen), dass sich die k_i als

$$hk_i = \hat{P}_i(hA)Q(hA)^{-1}x$$

schreiben lassen, wobei die $\hat{P}_i$ und Q Polynome vom Grad $\leq s$ sind. Damit ist auch Φ wegen

$$\begin{aligned} \Phi(t, x, h) &= x + \sum b_i hk_i \\ &= x + \sum b_i \hat{P}_i(hA)Q(hA)^{-1}x \\ &= \left(Q(hA) + \sum b_i \hat{P}_i(hA) \right) Q(hA)^{-1}x \\ &= P(hA)Q(hA)^{-1}x \end{aligned}$$

von der behaupteten Form. □

Bemerkung 6.2 (i) Das Wichtige an der soeben bewiesenen Struktur ist, dass die Abbildung R Eigenwerte von hA auf Eigenwerte von $R(hA)$ abbildet. Mit anderen Worten ist $\lambda \in \mathbb{C}$ genau dann ein Eigenwert von hA, wenn $R(\lambda) \in \mathbb{C}$ ein Eigenwert von $R(hA)$ ist. Dies werden wir im Beweis von Lemma 6.6 beweisen.

(ii) Aus dem Gleichungssystem des obigen Beweises kann man eine explizite Formel für $R(hA)$ berechnen, die aber recht kompliziert ist. Da wir später allerdings nur betrachten werden, wie Eigenwerte unter der Abbildung R abgebildet werden, reicht es aus, die Funktion R für komplexwertige Argumente $z \in \mathbb{C}$ explizit zu kennen. Wenn wir die Koeffizienten

des Verfahrens mit $\mathcal{A} = (a_{ij})_{i,j=1,\dots,s}$ und $b = (b_1, \dots, b_s)^T$ bezeichnen, so kann man hierfür den expliziten Ausdruck

$$R(z) = 1 + zb^T(\mathrm{Id} - z\mathcal{A})^{-1}\mathbf{e}$$

mit $\mathbf{e} = (1, \dots, 1)^T \in \mathbb{R}^s$ berechnen. Für komplexe Argumente wird die Funktion $R : \mathbb{C} \to \mathbb{C}$ als *Stabilitätsfunktion* des Verfahrens bezeichnet.

Z.B. ergeben sich für das explizite Euler–Verfahren $R(z) = 1 + z$, für das implizite Euler–Verfahren $R(z) = (1 - z)^{-1}$ und für die implizite Trapezregel aus Kapitel 5 $R(z) = (1 + z/2)/(1 - z/2)$. □

Wie im obigen eindimensionalen Fall wollen wir speziell Lösungen betrachten, die gegen Null streben und untersuchen, für welche Zeitschritte die numerische Approximation dieses Verhalten widerspiegelt. Dazu verwenden wir die folgende Definition.

Definition 6.3 Eine Differentialgleichung (6.1) bzw. eine Differenzengleichung (6.2) heißt *(global) exponentiell stabil*, falls Konstanten $c, \sigma > 0$ existieren, so dass für alle Anfangswerte $x_0 \in \mathbb{R}^n$ die Ungleichung

$$\|x(t; x_0)\| \leq ce^{-\sigma t}\|x_0\| \text{ für alle } t \geq 0$$

bzw.

$$\|\tilde{x}(t; x_0)\| \leq ce^{-\sigma t}\|x_0\| \text{ für alle } t = ih \geq 0$$

gilt. □

Für die obigen Gleichungstypen (6.1) und (6.2) kann man zeigen, dass sie genau dann exponentiell stabil sind, wenn alle Lösungen gegen Null konvergieren. Die spezielle exponentielle Abschätzung ergibt sich dann aus der Linearität der Gleichungen.

In Analogie zum eindimensionalen Fall nennen wir eine exponentiell stabile Differentialgleichung der Form (6.1) steif, wenn für explizite Verfahren ein sehr kleiner Zeitschritt nötig ist, damit die durch das Verfahren erzeugte Differenzengleichung (6.2) ebenfalls exponentiell stabil ist.

Um nun zu sehen, wie man anhand der Matrix A die Steifheit erkennen kann und zu verstehen, warum implizite Verfahren hier Vorteile haben, brauchen wir ein geeignetes Kriterium für exponentielle Stabilität. Glücklicherweise muss man nicht alle Lösungen kennen, um zu entscheiden, ob exponentielle Stabilität vorliegt; man kann diese Eigenschaft anhand der Matrizen A bzw. $\widetilde{A}$ erkennen, wie der folgende Satz zeigt. Hierbei bezeichnet $\Re(z) = a$ den Realteil und $|z| = \sqrt{a^2 + b^2}$ den Betrag einer komplexen Zahl $z = a + ib \in \mathbb{C}$.

Satz 6.4 (i) Die Differentialgleichung (6.1) ist genau dann exponentiell stabil, wenn für alle Eigenwerte λ_i von A die Ungleichung $\Re(\lambda_i) < 0$ gilt.

(ii) Die Differenzengleichung (6.2) ist genau dann exponentiell stabil, wenn für alle Eigenwerte $\tilde{\lambda}_i$ von $\widetilde{A}$ die Ungleichung $|\tilde{\lambda}_i| < 1$ gilt.

Beweisskizze: Wir beweisen Teil (ii) unter der Annahme, dass $\widetilde{A}$ diagonalisierbar ist (der Beweis von (ii) im nicht–diagonalisierbaren Fall funktioniert genauso, ist aber technischer; der Beweis von (i) ist ähnlich, verlangt aber weitere Kenntnisse über die Lösungsstruktur von (6.1), auf die wir hier nicht eingehen können).

Falls $\widetilde{A}$ diagonalisierbar ist, so existiert eine Koordinatentransformationsmatrix $T \in \mathbb{R}^{n\times n}$, so dass

$$T^{-1}\widetilde{A}T = \widetilde{\Lambda} = \begin{pmatrix} \tilde{\lambda}_1 & & & \\ & \tilde{\lambda}_2 & & \\ & & \ddots & \\ & & & \tilde{\lambda}_n \end{pmatrix}$$

ist, wobei die $\tilde{\lambda}_i$ gerade die Eigenwerte von $\widetilde{A}$ sind. Für die Lösung $\tilde{x}(hi; x_0)$ gilt dann gerade

$$\begin{aligned} \tilde{x}(ih; x_0) &= \widetilde{A}^i x_0 \\ &= (T\widetilde{\Lambda}T^{-1})^i x_0 \\ &= T\widetilde{\Lambda}^i T^{-1} x_0. \end{aligned}$$

Sei nun $\alpha = \max_i |\tilde{\lambda}_i| < 1$. Wenn wir $y = (y_1, \ldots, y_n)^T = T^{-1}x_0$ setzen, so folgt

$$\widetilde{\Lambda}^i y = \begin{pmatrix} \tilde{\lambda}_1^i y_1 \\ \vdots \\ \tilde{\lambda}_n^i y_n \end{pmatrix}$$

und damit $\|\widetilde{\Lambda}^i y\| \leq \alpha^i \|y\|$. Mit $\sigma = -\ln(\alpha)/h > 0$ und $t = hi$ folgt

$$\|\widetilde{\Lambda}^i y\| \leq e^{-\sigma t}\|y\|$$

und damit

$$\|\tilde{x}(t; x_0)\| \leq \|T\|e^{-\sigma t}\|T^{-1}x_0\| \leq e^{-\sigma t}\|T\|\|T^{-1}\|\|x_0\| = ce^{-\sigma t}\|x_0\|$$

mit $c = \|T\|\|T^{-1}\|$.

Sei umgekehrt $|\tilde{\lambda}_j| \geq 1$ für ein j und sei x_0 ein zugehöriger Eigenvektor. Dann gilt

$$\|\tilde{x}(t; x_0)\| = \|\widetilde{A}^i x_0\| = |\tilde{\lambda}_j^i|\|x_0\| \geq \|x_0\|$$

für alle $t = ih > 0$, weswegen (6.2) nicht exponentiell stabil ist. □

Wir bezeichnen mit

$$\Sigma(A) = \{\lambda_i \,|\, \lambda_i \text{ ist Eigenwert von } A\}$$

die Menge aller Eigenwerte, das sogenannte *Spektrum* von A.

Für die Differentialgleichung muss damit gerade

$$\Sigma(A) \subset \mathbb{C}^- := \{z \in \mathbb{C} \,|\, \Re(z) < 0\}$$

gelten, damit exponentielle Stabilität vorliegt. Ein Eigenwert $\lambda_i \in \mathbb{C}^-$ wird dabei als *stabiler Eigenwert* bezeichnet. Analog muss für die numerische Approximation (6.2)

$$\Sigma(\widetilde{A}) \subset B_1(0) := \{z \in \mathbb{C} \,|\, |z| < 1\}$$

gelten, damit exponentielle Stabilität vorliegt.

6.2 Stabilitätsgebiet und A-Stabilität

Zu klären bleibt die Frage, welche Bedingung A aus (6.1) erfüllen muss, damit (6.2) für die Matrix $\widetilde{A} = R(hA)$ exponentiell stabil ist. Sicherlich hängt dies vom verwendeten Verfahren und vom Zeitschritt ab. Hierzu verwenden wir die folgende Definition. Beachte dabei, dass

$$\Sigma(A) \subset \mathbb{C}^- \;\Leftrightarrow\; \Sigma(hA) \subset \mathbb{C}^- \text{ für alle } h > 0$$

gilt, da λ_i genau dann ein Eigenwert von A ist, wenn $h\lambda_i$ ein Eigenwert von hA ist.

Definition 6.5 (i) Das *Stabilitätsgebiet* $\mathcal{S} \subset \mathbb{C}$ eines Runge-Kutta-Verfahrens mit Stabilitätsfunktion R ist definiert als die maximale Teilmenge der komplexen Zahlen, für die für alle $A \in \mathbb{R}^{n\times n}$ und alle $h > 0$ die Folgerung

$$\Sigma(hA) \subset \mathcal{S} \;\Rightarrow\; \Sigma(R(hA)) \subset B_1(0)$$

gilt. Mit anderen Worten ist $\mathcal{S}$ gerade die Menge von Eigenwerten λ_i, die hA aus (6.1) annehmen darf, damit (6.2) mit $\widetilde{A} = R(hA)$ exponentiell stabil ist.

(ii) Ein Runge-Kutta-Verfahren heißt *A-stabil*, falls

$$\mathbb{C}^- \subseteq \mathcal{S}$$

gilt bzw., äquivalent dazu, falls die Folgerung

$$\Sigma(hA) \subset \mathbb{C}^- \;\Rightarrow\; \Sigma(R(hA)) \subset B_1(0)$$

gilt. □

Die Interpretation von (i) ist wie folgt: Zur korrekten numerischen Approximation einer exponentiell stabilen Gleichung muss die Schrittweite $h > 0$ so gewählt werden, dass die Eigenwerte von hA in $\mathcal{S}$ liegen. Je besser $\mathcal{S}$ die Menge $\mathbb{C}^-$ ausschöpft, desto geringer sind die Anforderungen an die Schrittweite; im Falle der A-Stabilität gibt es überhaupt keine Einschränkungen der Schrittweite, die exponentielle Stabilität von (6.1) wird für alle Zeitschritte $h > 0$ von (6.2) "geerbt".

Das folgende Lemma zeigt, wie der Stabilitätsbereich $\mathcal{S}$ berechnet werden kann.

Lemma 6.6 Gegeben sei ein Runge-Kutta-Verfahren mit Stabilitätsfunktion R aus Bemerkung 6.2. Dann ist der Stabilitätsbereich gegeben durch

$$\mathcal{S} = \{z \in \mathbb{C} \,|\, |R(z)| < 1\}.$$

Beweis: Zum Beweis der Behauptung zeigen wir zunächst, dass für alle Matrizen $B \in \mathbb{R}^{n\times n}$ gilt: $\lambda_i \in \mathbb{C}$ ist genau dann ein Eigenwert von B, wenn $R(\lambda_i)$ ein Eigenwert von $R(B)$ ist.

Sei $C \in \mathbb{R}^{n\times n}$ eine beliebige Matrix mit Eigenwerten λ_i, $i = 1, \ldots, p \le n$. Für ein Polynom

$$P(C) = \alpha_0 \mathrm{Id} + \alpha_1 C + \ldots + \ldots \alpha_s C^s$$

sind die Eigenwerte von $P(C)$ gerade die Eigenwerte $P(\lambda_i)$ von C, was man am einfachsten sieht, indem man P auf die Jordan–Normalform J von C anwendet. Ein Jordanblock J_i zum Eigenwert λ_i wird dabei auf eine obere Dreiecksmatrix mit $P(\lambda_i)$ in der Diagonalen abgebildet, die genau den einzigen Eigenwert $P(\lambda_i)$ besitzt (lediglich die Vielfachheiten können sich u.U. ändern). Hierbei sind Eigenvektoren von C wieder Eigenvektoren von $P(C)$.

Für die Inverse C^{-1} ist sind die Eigenwerte gerade $1/\lambda_i$ und für ein Produkt zweier Matrizen mit gleichen Eigenvektoren sind die Eigenwerte gerade die Produkte der Eigenwerte.

Also folgt, dass die Eigenwerte von $R(B) = P(B)Q(B)^{-1}$ gerade die Produkte der Eigenwerte $P(\lambda_i)$ und $Q(\lambda_i)^{-1}$, also $P(\lambda_i)Q(\lambda_i)^{-1}$ sind.

Weil R also Eigenwerte von hA auf Eigenwerte von $R(hA)$ abbildet, gilt $\Sigma(R(hA)) = R(\Sigma(hA))$ und damit

$$\begin{aligned} \Sigma(R(hA)) \subset B_1(0) &\Leftrightarrow R(\Sigma(hA)) \subset B_1(0) \\ &\Leftrightarrow |R(\lambda_i)| < 1 \text{ für alle Eigenwerte } \lambda_i \text{ von } hA. \end{aligned}$$

Für alle Matrizen hA mit $\Sigma(hA) \subset \{z \in \mathbb{C} \,|\, |R(z)| < 1\}$ gilt also $\Sigma(R(hA)) \subset B_1(0)$, woraus wegen der Maximalität von $\mathcal{S}$ die Inklusion $\{z \in \mathbb{C} \,|\, |R(z)| < 1\} \subseteq \mathcal{S}$ folgt. Andererseits gilt für jedes $z \in \mathbb{C}$ mit $|R(z)| \geq 1$ und die 1×1-Matrix $A = (z)$ sowie $h = 1$, dass $\{z\} = \Sigma(hA) \not\subset \mathcal{S}$. Also folgt die behauptete Gleichheit. □

Mit Hilfe dieses Satzes können wir die Stabilitätsbereiche nun bestimmen. Für das explizite Euler–Verfahren mit $R(z) = 1 + z$ gilt

$$|R(z)| < 1 \;\Leftrightarrow\; |1 + z| < 1$$

also ist $\mathcal{S} = \{z \in \mathbb{C} || 1 + z| < 1\} = B_1(-1)$, also gerade der offene Ball mit Radius 1 um -1. Der Zeitschritt muss also so klein gewählt werden, dass für alle Eigenwerte die Bedingung $h\lambda_i \in B_1(-1)$ erfüllt ist.

Abbildung 6.2 zeigt die Stabilitätsbereiche einiger expliziter Runge-Kutta-Verfahren mit den Ordnungen $p = 1, \ldots, 4$. Man sieht, dass der Stabilitätsbereich $\mathcal{S}$ für wachsende Konsistenz größer wird, allerdings die Menge $\mathbb{C}^-$ bei weitem nicht ausschöpft. Im Falle betragsmäßig großer Eigenwerte λ_i erhält man für all diese Verfahren starke Einschränkungen bei der Wahl der Zeitschritte.

Beachte, dass bei mehrdimensionalen Problemen nicht unbedingt der *Realteil* eines Eigenwertes betragsmäßig groß werden muss, damit der Betrag des Eigenwertes groß wird. Das folgende Beispiel illustriert dies.

Betrachte die zweidimensionale lineare DGL

$$\dot{x}(t) = \begin{pmatrix} -1 & \alpha \\ -\alpha & -1 \end{pmatrix} x(t). \tag{6.3}$$

Die zugehörige Matrix besitzt die Eigenwerte $\lambda_{1/2} = -1 \pm i\alpha$. Hier haben Realteil und Imaginärteil eine geometrische Bedeutung für die Lösung: Der Realteil gibt an, wie schnell die Lösung gegen Null konvergiert (diese Größe ist hier konstant gleich -1), während der Imaginärteil angibt, wie schnell die Lösung sich dabei dreht. Abbildung (6.3) zeigt die

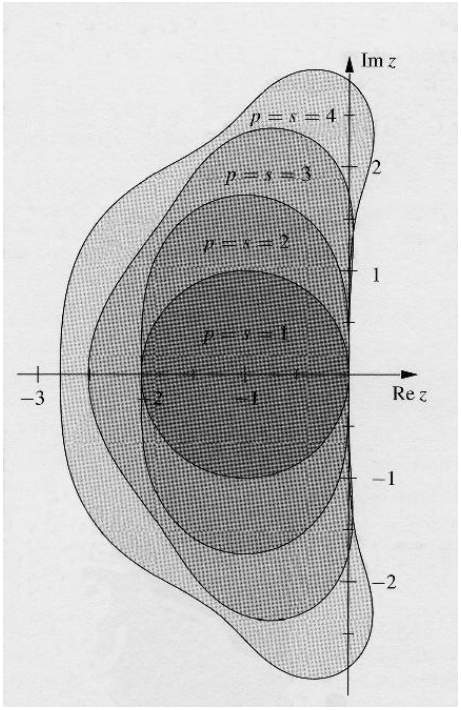

Abbildung 6.2: Stabilitätsbereiche expliziter Runge-Kutta-Verfahren, entnommen aus [2]

exakten Lösungen für $\alpha = 0,\ 1,\ 10$ sowie die zugehörigen Euler–Lösungen mit $h = 0.02$. Man sieht: Wenn der Eigenwert betragsmäßig größer wird, weil der Imaginärteil wächst, dann wird die Euler–Lösung instabil.

Wie verhalten sich nun implizite Verfahren? Für das implizite Euler–Verfahren z.B. berechnet man

$$|R(z)| < 1 \ \Leftrightarrow\ 1/|1-z| < 1 \ \Leftrightarrow\ |1-z| > 1 \ \Leftarrow\ \Re(z) < 0.$$

Folglich gilt $\mathbb{C}^- \subset \mathcal{S}$, das Verfahren ist also A-stabil.

Viele implizite Verfahren sind A-stabil, und von denjenigen, die es nicht sind, besitzen viele einen Stabilitätsbereich, der deutlich größer ist als bei expliziten Verfahren. Eine Übersicht über die Stabilitätsbereiche einiger impliziter Verfahren findet sich z.B. im Abschnitt IV.3 des Buchs [5].

Eine lineare DGL (6.1) kann auch dann steif sein, wenn sie nicht exponentiell stabil ist, aber zumindest einige stabile Eigenwerte besitzt, also solche mit negativem Realteil. Die Lösungskomponenten in den zugehörigen Eigenräumen (man nennt deren Vereinigung *stabilen Unterraum*) verhalten sich dann wie bei einer exponentiell stabilen Gleichung. Folglich treten bei betragsmäßig großen stabilen Eigenwerten exakt die gleichen Probleme auf, auch wenn die Gleichung insgesamt nicht exponentiell stabil ist. Dies führt uns auf die folgende Charakterisierung.

Bemerkung 6.7 Eine lineare zeitinvariante Differentialgleichung ist steif, falls die zugehörige Matrix A betragsmäßig große stabile Eigenwerte besitzt. □

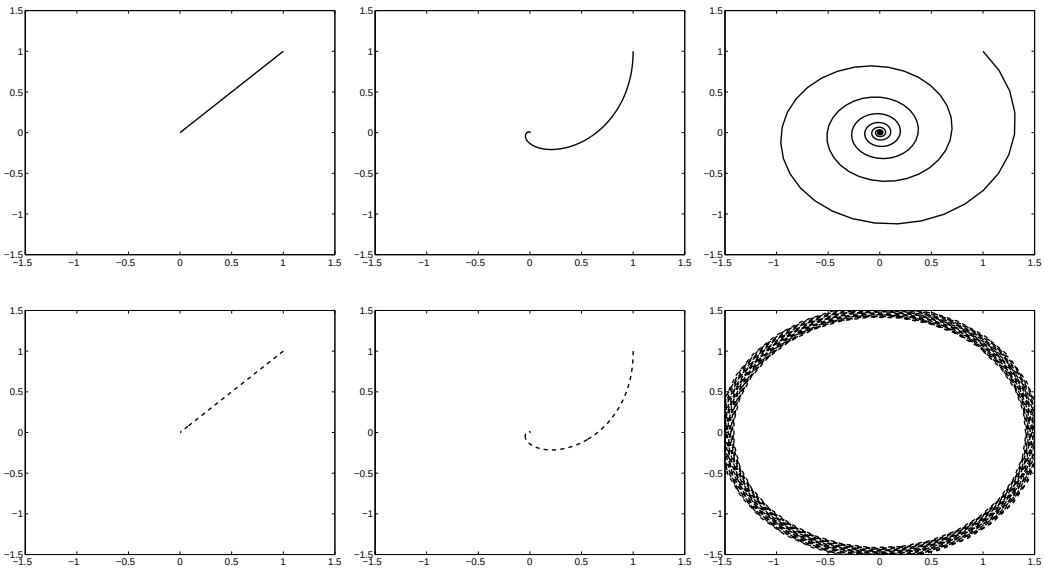

Abbildung 6.3: Exakte und Euler–Lösungen von (6.3) mit $\alpha = 0,\ 1,\ 10$, $h = 0.02$

Für nichtlineare DGL $\dot{x}(t) = f(t, x(t))$ gibt es viele weitere Phänomene, die zur Steifheit führen; meistens kann man diese nicht so einfach am Vektorfeld f ablesen. Im einfachsten Fall ist f autonom und besitzt ein Gleichgewicht x^*, in dem f stetig differenzierbar ist. In diesem Fall kann man $A = Df(x^*)$ betrachten; wenn diese Matrix betragsmäßig große stabile Eigenwerte besitzt, so wird auch die nichtlineare DGL typischerweise steif sein. Steifheit kann aber auch auftreten, wenn kein Gleichgewicht vorliegt, z.B. wenn die DGL eine exponentiell stabile periodische Lösung besitzt (also eine periodische Lösung, gegen die alle Lösungen exponentiell konvergieren, zumindest für nahe liegende Anfangswerte). In diesem Fall kann die Gleichung steif sein, wenn die anderen Lösungen sehr schnell gegen die periodische Lösung streben (dies entspricht betragsmäßig großen negativen Realteilen im linearen Fall) oder wenn sich die periodische Lösung sehr schnell bewegt (dies entspricht den großen Imaginärteilen.)

6.3 Weitere Stabilitätsbegriffe

Der Begriff der A-Stabilität wurde von G. Dahlquist in den 1960er Jahren eingeführt. A-Stabilität ist nützlich bei der Lösung steifer Differentialgleichungen, ist aber für sich genommen weder eine positive noch eine negative Eigenschaft: Zwar ist es zur numerischen Lösung steifer DGL vorteilhaft, wenn die exponentielle Stabilität von der numerischen Approximation geerbt wird. Allerdings kann es andererseits auch passieren, dass die numerische Approximation exponentiell stabil ist, obwohl die exakte Gleichung diese Eigenschaft *nicht* besitzt, was zu falschen Rückschlüssen auf das Verhalten der exakten Lösungen führen kann.

Eine stärkere Eigenschaft ist die *Erhaltung der Isometrie*, die verlangt, dass $\mathcal{S} = \mathbb{C}^-$ ist, d.h. für alle Zeitschritte $h > 0$ ist die numerische Approximation *genau dann* exponentiell

stabil, wenn die exakte Gleichung exponentiell stabil ist. Diese Eigenschaft besitzen z.B. die bereits erwähnten Gauß–Verfahren. Ein anderes Verfahren mit dieser Eigenschaft ist die implizite Mittelpunktregel, vgl. Abschnitt 5.1, für die wir dieses nachweisen wollen: Ausgeschrieben ist die zugehörige Iterationsvorschrift gegeben durch

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_i) + h_i f\left(t_i + \frac{h_i}{2}, \frac{1}{2}(\tilde{x}(t_i) + \tilde{x}(t_{i+1}))\right).$$

Angewendet auf die lineare Differentialgleichung $\dot{x}(t) = Ax(t)$ ergibt sich

$$\begin{aligned}
& \tilde{x}(t_{i+1}) &&= \tilde{x}(t_i) + \frac{h_i}{2}A\tilde{x}(t_i) + \frac{h_i}{2}A\tilde{x}(t_{i+1}) \\
\Leftrightarrow\quad & \tilde{x}(t_{i+1}) - \frac{h_i}{2}A\tilde{x}(t_{i+1}) &&= \tilde{x}(t_i) + \frac{h_i}{2}A(\tilde{x}(t_i) \\
\Leftrightarrow\quad & \left(\mathrm{Id} - \frac{h_i}{2}A\right)\tilde{x}(t_{i+1}) &&= \left(\mathrm{Id} + \frac{h_i}{2}A\right)\tilde{x}(t_i) \\
\Leftrightarrow\quad & \tilde{x}(t_{i+1}) &&= \left(\mathrm{Id} - \frac{1}{2}h_iA\right)^{-1}\left(\mathrm{Id} + \frac{1}{2}h_iA\right)\tilde{x}(t_i).
\end{aligned}$$

Die Stabilitätsfunktion ist daher gegeben durch

$$R(z) = \frac{1 + z/2}{1 - z/2}.$$

Wir wollen nun nachweisen, dass dieses Verfahren die Isometrie erhält. Dazu müssen wir die Äquivalenz $\Re(z) < 0 \Leftrightarrow |R(z)| < 1$ zeigen, wozu wir alternativ auch

$$\Re(z) < 0 \Leftrightarrow |R(z)|^2 < 1$$

nachprüfen können. Für $z = a + ib$ gilt wegen $|x + iy|^2 = x^2 + y^2$ nun

$$|R(z)|^2 = \frac{|1 + z/2|^2}{|1 - z/2|^2} = \frac{(1 + a/2)^2 + (b/2)^2}{(1 - a/2)^2 + (b/2)^2}.$$

Dieser Ausdruck ist nun genau dann < 1, wenn $(1 + a/2)^2 < (1 - a/2)^2$ gilt. Dies ist aber genau dann der Fall, wenn $a < 0$ gilt, womit die Erhaltung der Isometrie folgt.

Unglücklicherweise ist es aber so, dass diese Eigenschaft stets gemeinsam mit einer anderen — unerwünschten — Eigenschaft auftritt. Um diese zu illustrieren, betrachten wir die implizite Mittelpunktregel angewendet auf die Gleichung $\dot{x}(t) = \lambda x(t)$ mit $\lambda = -1000$ und Schrittweite $h = 0.01$.

Zwar ist die Lösung asymptotisch stabil, allerdings konvergiert sie nicht — wie die exakte Lösung — monoton sondern oszillierend gegen 0. Zudem konvergiert die numerische Approximation um so langsamer gegen 0, je größer $|\lambda|$ wird. Und dass, obwohl die exakte Lösung $e^{\lambda t}x_0$ ja für negative λ um so schneller gegen 0 konvergiert, je größer $|\lambda|$ ist. Dies kann man auch an der Stabilitätsfunktion sehen, die für betragsmäßig große negative λ Werte $R(\lambda) \approx 1$ und damit sehr langsame Konvergenz gegen 0 liefert. Die Approximation zeigt also das richtige Konvergenzverhalten, hat ansonsten aber nicht viel mit der exakten Lösung zu tun.

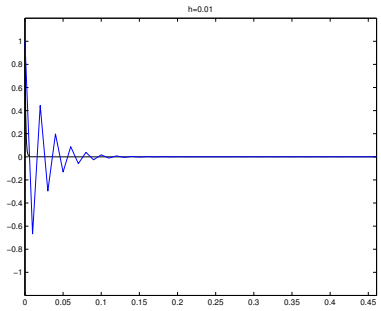

Abbildung 6.4: Oszillationen bei der impliziten Mittelpunktregel

Dies ist kein Zufall, denn jedes isometrieerhaltende Verfahren besitzt diese Eigenschaft. Der Grund liegt darin, dass für rationale Funktionen der Grenzwert

$$\lim_{k\to\infty} R(z_k)$$

— sofern er existiert — für alle komplexen Folgen $(z_k)_{k\in\mathbb{N}}$ mit $|z_k| \to \infty$ identisch ist, unabhängig von der Wahl der Folge z_k. Für $z_k = ib_k$ gilt aber nun $|R(z_k)| = 1$ (weil $|R|$ in der rechten komplexen Halbebene Werte > 1 und in der linken Halbebene Werte < 1 annimmt, muss aus Stetigkeitsgründen für rein imaginäre Zahlen $|R(z_k)| = 1$ gelten). Also gilt für $b_k \to \infty$ die Konvergenz $\lim_{k\to\infty} |R(z_k)| = \lim_{k\to\infty} |R(ib_k)| = 1$ und damit auch für alle anderen Folgen. Insbesondere gilt also $\lim_{a_k\to-\infty} |R(a_k + ib)| = 1$ für alle $b \in \mathbb{R}$ und damit $|R(a + ib)| \approx 1$ für betragsmäßig große negative a. Dies erklärt die langsame Konvergenz der isometrierhaltenden Verfahren bei sehr schnell konvergierenden Differentialgleichungen.

Was man stattdessen zur guten Approximation der Lösung haben möchte, ist die Konvergenz $R(a_k + ib) \to 0$ für $a_k \to -\infty$. Dies würde garantieren, dass mit der exakten auch die numerische Lösung immer schneller gegen 0 konvergiert.

Definition 6.8 Ein Runge-Kutta-Verfahren heißt L-stabil, wenn es A-stabil ist und zudem

$$\lim_{k\to\infty} R(z_k) = 0$$

gilt für alle komplexen Folgen $(z_k)_{k\in\mathbb{N}}$ mit $|z_k| \to \infty$. □

Wir schreiben diese Bedingung auch kurz als $R(\infty) = 0$. Beachte, dass $R(\infty)$ wohldefiniert ist, da der Grenzwert $\lim_{k\to\infty} R(z_k)$ für $|z_k| \to \infty$ nicht von der Wahl der z_k abhängt. Für diese Gleichung kann man eine hinreichende Bedingung an die Koeffizienten des Runge-Kutta-Verfahrens herleiten.

Satz 6.9 Wenn die Koeffizientenmatrix $\mathcal{A}$ eines impliziten Runge-Kutta-Verfahrens invertierbar ist und eine der beiden Bedingungen

$$a_{sj} = b_j \text{ für } j = 1, \ldots, s \quad \text{oder} \quad a_{i1} = b_1 \text{ für } i = 1, \ldots, s$$

gelten, so gilt $R(\infty) = 0$. Falls das Verfahren zusätzlich A-stabil ist, so ist es L-stabil.

Beweis: Da die Stabilitätsfunktion durch

$$R(z) = 1 + zb^T(\mathrm{Id} - z\mathcal{A})^{-1}\mathbf{e} = 1 + b^T\left(\frac{1}{z}\mathrm{Id} - \mathcal{A}\right)^{-1}\mathbf{e}$$

gegeben ist, folgt

$$R(\infty) = 1 - b^T\mathcal{A}^{-1}\mathbf{e}.$$

Im Fall der ersten Bedingung gilt $\mathcal{A}^T e_s = b$, wobei $e_s = (0, \ldots, 0, 1)^T \in \mathbb{R}^s$. Damit folgt $e_s^T\mathcal{A} = b^T$, folglich $e_s^T = b^T\mathcal{A}^{-1}$ und damit

$$R(\infty) = 1 - e_s^T\mathbf{e} = 1 - 1 = 0.$$

Im Fall der zweiten Bedingung gilt $\mathcal{A}e_1 = \mathbf{e}b_1$ und damit $\mathcal{A}^{-1}\mathbf{e} = (1/b_1, 0, \ldots, 0)^T$. Damit folgt

$$R(\infty) = 1 - b_1/b_1 = 1 - 1 = 0.$$

□

L-Stabilität spielt eine wichtige Rolle bei der Lösung sogenannter Differential-Algebraischer Gleichungen sowie bei singulär gestörten Problemen, da bei diesen Problemen typischerweise Eigenwerte mit betragsmäßig sehr großen negativen Realteilen auftreten. Ein Beispiel für ein L-stabiles implizites Runge-Kutta-Verfahren ist das Radau IIA-Verfahren (mit Ordnung $p = 5$) mit dem Butcher-Tableau

$$\begin{array}{c|ccc}
\frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\
\frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\
1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\
\hline
 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9}
\end{array}$$

Offenbar ist hier gerade die erste Bedingung von Satz 6.9 erfüllt. Für die Herleitung dieses Verfahrens siehe Abschnitt 9.2.

Das Problem mit der A-Stabilität ist, dass es viele numerische Verfahren gibt, die nicht A-stabil sind, für Eigenwerte und Schrittweiten in "sinnvollen" Bereichen aber trotzdem gute Lösungen liefern. Es ist daher zu einschränkend, prinzipiell A-Stabilität zu fordern, wenn man es mit steifen Differentialgleichungen zu tun hat. Nichtsdestotrotz ist es erstrebenswert, Stabilität für Eigenwerte mit beliebig kleinen Realteilen zu haben (also ein unbeschränktes Stabilitätsgebiet), allerdings nicht für beliebige Kombinationen von Real- und Imaginärteil.

Eine Stabilitätsbedingung, die dies mathematisch präzise definiert, ist die folgende $A(\alpha)$-Stabilität.

Definition 6.10 Ein Runge-Kutta Verfahren heißt $A(\alpha)$-stabil für ein $\alpha > 0$, falls der Sektor

$$S_\alpha := \{z \in \mathbb{C} \,|\, z \neq 0 \text{ und } |\arg(-z)| < \alpha\}$$

im Stabilitätsgebiet S enthalten ist. □

Ein Beispiel für eine Methode, die $A(\alpha)$-stabil aber nicht A-stabil ist ist die sogenannte $(0, 3)$-Padé-Approximation, mit $\alpha = 88.23°$. Für Details siehe [5, Abschnitt IV.3].

6.4 Nichtlineare A-Stabilität

Wie am Ende von Abschnitt 6.2 bereits erwähnt, gelten die bisher gemachten Stabilitätsaussagen auch für autonome nichtlineare Differentialgeichungen $\dot{x}(t) = f(x(t))$ in der Nähe von Gleichgewichten. Allerdings lässt sich mit der linearen Theorie basierend auf Jacobi-Matrizen und Eigenwerten prinzipiell keine Aussage über das Verhalten weit weg von den Gleichgewichten machen.

Um eine Eigenschaft wie die A-Stabilität nicht nur in der Nähe von Gleichgewichten nachzuweisen (also die Tatsache, dass die Approximationen für alle Schrittweiten asymptotisch stabil sind), benötigt man andere Methoden. Eine davon sind die sogenannten Lyapunov-Funktionen.

Definition 6.11 Eine stetig differenzierbare Funktion $V : \mathbb{R}^n \to \mathbb{R}$ heißt Lyapunov-Funktion für eine autonome gewöhnliche Differentialgleichung $\dot{x}(t) = f(x(t))$ an einem Gleichgewicht x^*, falls die folgenden Bedingungen gelten.

(a) $V(x^*) = 0$ und $V(x) > 0$ für alle $x \neq x^*$

(b) $V(x_n) \to \infty$ für alle Folgen $(x_n)_{n\in\mathbb{N}}$ mit $\|x_n\| \to \infty$ für $n \to \infty$

(c) $DV(x)f(x) < 0$ für alle $x \neq x^*$.

□

Satz 6.12 Betrachte eine autonome gewöhnliche Differentialgleichung $\dot{x}(t) = f(x(t))$ auf $D = \mathbb{R}^n$ mit einem Gleichgewicht x^*. Falls eine Lyapunov Funktion V existiert, so konvergieren alle Lösungen $x(t)$ der Differentialgleichung für $t \to \infty$ gegen x^*.

Beweisidee: Aus (c) folgt mit der Kettenregel die Ungleichung

$$\frac{d}{dt}V(x(t)) = DV(x(t))\dot{x}(t) = DV(x(t))f(x(t)) < 0,$$

falls $x(t) \neq x^*$. Daraus folgt, dass $t \mapsto V(x(t))$ streng monoton fällt, so lange die Lösung nicht bereits im Gleichgewicht x^* ist. Weil $V(x(t))$ wegen (a) zudem nach unten beschränkt ist, konvergiert $V(x(t))$ für $t \to \infty$ gegen einen Wert V_∞. Daraus folgt wiederum, dass die Ableitung $\frac{d}{dt}V(x(t))$ gegen 0 konvergieren muss. Dies kann wegen (c) nur passieren, wenn $x(t) \to x^*$ oder wenn $\|x(t)\| \to \infty$ konvergiert. Wegen (b) und der Beschränktheit von $V(x(t))$ ist aber nur ersteres möglich. □

Bemerkung 6.13 Tatsächlich folgt aus der Existenz einer Lyapunov Funktion mehr als nur die Konvergenz, nämlich die sogenannte *globale asymptotische Stabilität*. Neben der Konvergenz umfasst diese Eigenschaft auch die Tatsache, dass Lösungen die in der Nahe von x^* starten für alle positiven Zeiten in der Nähe von x^* bleiben. □

Die Idee einer nichtlinearen Verallgemeinerung der A-Stabilität liegt nun darin nachzuweisen, dass eine Lyapunov Funktion der Differentialgleichung für beliebige Schrittweiten auch eine Lyapunov Funktion für die numerische Approximation ist. Dazu genügt es, das Gegenstück zu Bedingung (c) aus Definition 6.11 nachzuweisen, nämlich

(c') $V(\Phi(x,h)) < V(x)$ für alle $x \neq x^*$.

Es ist nun leider zu optimistisch anzunehmen, dass es ein Verfahren gibt, mit dem (c') für alle $h > 0$, alle Lyapunov Funktionen und alle Differentialgleichungen gilt. Man kann aber beweisen, dass (c') für alle Schrittweiten $h > 0$ gilt

- für das implizite Euler-Verfahren, wenn V eine konvexe Funktion ist, wenn also für alle $x_1, x_2 \in \mathbb{R}^n$ und alle $\lambda \in [0,1]$ gilt

$$V(\lambda x_1 + (1-\lambda)x_2) \leq \lambda V(x_1) + (1-\lambda)V(x_2)$$

- für die implizite Mittelpunktregel, wenn V eine positiv definite quadratische Funktion ist, wenn also eine positiv definite Matrix $P \in \mathbb{R}^{n\times n}$ gibt, so dass V von der Form

$$V(x) = x^T P x$$

ist.

Da jede exponentiall stabile lineare zeitinvariante Differentialgleichung $\dot{x}(t) = Ax(t)$ eine quadratische Lyapunov Funktion besitzt und da eine solche stets konvex ist, bilden beide Aussagen eine echte Verallgemeinerung der linearen A-Stabilität.

Die Aussage für das implizite Euler-Verfahren folgt mit einigen (relativ einfachen) Argumenten aus der konvexen Analysis, die Aussage über die implizite Mittelpunktregel mit Hilfe einer geeigneten Taylor-Entwicklung unter Ausnutzung der Tatsache, dass die zweite Ableitung der quadratischen Funktion V konstant ist.

Kapitel 7

Schrittweitensteuerung

Nach den eher theoretischen Überlegungen des letzten Kapitels wollen wir uns jetzt wieder algorithmischen Aspekten widmen. Bisher sind wir davon ausgegangen, dass die Schrittweiten h_i gegeben sind, meistens haben wir sie als konstant $h_i \equiv h$ angenommen. In diesem Kapitel wollen wir uns überlegen, wie man die Schrittweiten automatisch so steuern kann, so dass dort, wo es nötig ist, kleine Schrittweiten gewählt werden, damit eine gewünschte Genauigkeit eingehalten wird und dort, wo es ohne Genauigkeitsverlust möglich ist, große Schrittweiten erlaubt werden, die eine schnellere Rechnung ermöglichen. Wir nehmen dabei durchgehend an, dass das Vektorfeld der betrachteten DGL hinreichend oft differenzierbar ist, so dass die Konsistenzordnungen der betrachteten Verfahren tatsächlich realisiert werden.

7.1 Fehlerschätzung

Zur Entscheidung darüber, ob die Schrittweite groß oder klein gewählt werden soll, ist es nötig, den Fehler zu kennen, den wir im aktuellen Schritt machen. Wir wollen uns zuerst überlegen, welcher Fehler hierfür wichtig ist. Hierbei müssen wir zunächst überlegen, wie wir die Schrittweite steuern wollen. Wie in der numerischen Praxis üblich wollen wir uns hier darauf beschränken, zur Zeit t_i eine gute Schrittweite h_i für den Schritt von t_i nach $t_{i+1} = t_i + h_i$ zu bestimmen und dabei auch einen "Schrittweitenvorschlag" h_{i+1} für den nächsten Schritt zu machen. Wir wollen aber nicht zum Zeitpunkt t_i die Schrittweiten in vorhergehenden Schritten t_j für $j < i$ nachträglich korrigieren, da die dadurch anfallenden Neuberechnungen algorithmisch sehr ineffizient wären.

Um ein gutes h_i zu bestimmen, müssen wir den Fehleranteil kennen, der durch den Schritt von t_i nach t_{i+1} hervorgerufen wird. Dieser Fehleranteil wird *lokaler Fehler* genannt. Wir haben in der Konvergenzanalyse in Abschnitt 2.3 verwendet, dass sich der Fehler zur Zeit t_{i+1} mittels

$$\begin{aligned} \|\tilde{x}(t_{i+1}) - x(t_{i+1})\| \quad &\leq \quad \|\Phi(t_i, \tilde{x}(t_i), h_i) - \Phi(t_i, x(t_i), h_i)\| \\ &+ \ \|\Phi(t_i, x(t_i), h_i) - x(t_{i+1}; t_i, x(t_i))\| \end{aligned}$$

zerlegen lässt. Diese Zerlegung war für unsere theoretischen Überlegungen nützlich, hier ist sie nicht so günstig, da wir den in diesem Schritt hinzukommenden Fehleranteil

$$\|\Phi(t_i, x(t_i), h_i) - x(t_{i+1}; t_i, x(t_i))\|$$

nicht berechnen können, da wir $x(t_i)$ nicht kennen. Statt also in der Dreiecksungleichung den Term $\Phi(t_i, x(t_i), h_i)$ einzuschieben, schieben wir den Term $x(t_{i+1}; t_i, \tilde{x}(t_i))$ und erhalten so

$$\begin{aligned}\|\tilde{x}(t_{i+1}) - x(t_{i+1})\| \quad &\leq \quad \|\Phi(t_i, \tilde{x}(t_i), h_i) - x(t_{i+1}, t_i, \tilde{x}(t_i))\| \\ &+ \ \|x(t_{i+1}, t_i, \tilde{x}(t_i)) - x(t_{i+1}, t_i, x(t_i))\|\end{aligned}$$

Der zweite Fehlerterm hängt hierbei im Wesentlichen von dem bis zum Zeitpunkt t_i gemachten Fehler ab, den wir nur durch Änderung der Zeitschritte h_j für $j < i$ beeinflussen können, was wir gerade nicht machen wollen. Der Fehlerterm, den wir mit der Wahl von h_i wirklich beeinflussen können, ist der erste.

Die Idee der Schrittweitensteuerung (man sagt auch "adaptive Wahl der Schrittweite") liegt nun darin, h_i so groß zu wählen, dass die Fehlerbedingung

$$\|\Phi(t_i, \tilde{x}(t_i), h_i) - x(t_{i+1}, t_i, \tilde{x}(t_i))\| \leq tol$$

für eine vorgegebene Größe $tol > 0$ gerade eingehalten wird. Dies ist natürlich so nicht möglich, da wir dafür die exakte Lösung $x(t_{i+1}, \tilde{x}(t_i), t_i)$ kennen müssten. Um dieses Problem zu lösen, verwendet man einen sogenannten *Fehlerschätzer*, der wie folgt definiert ist.

Definition 7.1 Eine numerisch berechenbare Größe $\bar{\varepsilon}$ heißt *Fehlerschätzer* für den tatsächlichen Fehler ε eines numerischen Verfahrens, falls von $\bar{\varepsilon}$ und ε unabhängige Konstanten $\kappa_1, \kappa_2 > 0$ existieren, so dass die Abschätzung

$$\kappa_1 \varepsilon \leq \bar{\varepsilon} \leq \kappa_2 \varepsilon$$

gilt. □

Wie können wir nun für unsere Einschrittverfahren einen solchen Fehlerschätzer bekommen? Die Idee besteht darin, den Schritt von t nach $t_{i+1} = t_i + h_i$ mit zwei Verfahren $\widehat{\Phi}$ und Φ verschiedener Konsistenzordnung $\hat{p}$ und p zu berechnen. Für

$$\hat{\eta}_i := \widehat{\Phi}(t_i, \tilde{x}(t_i), h_i) - x(t_{i+1}, t_i, \tilde{x}(t_i)) \quad \text{und} \quad \eta_i := \Phi(t_i, \tilde{x}(t_i), h_i) - x(t_{i+1}, t_i, \tilde{x}(t_i))$$

gilt damit

$$\hat{\varepsilon}_i := \|\hat{\eta}_i\| \leq \widehat{E} h_i^{\hat{p}+1} \quad \text{und} \quad \varepsilon_i := \|\eta_i\| \leq E h_i^{p+1}. \tag{7.1}$$

Wir nehmen hierbei an, dass $p \geq \hat{p} + 1$ gilt und dass $\hat{p}$ die maximale (oder echte) Konsistenzordnung von $\widehat{\Phi}$ ist. Damit ist Φ das genauere Verfahren, weswegen für alle hinreichend kleinen $h_i > 0$ die Ungleichung $\varepsilon_i < \hat{\varepsilon}_i$ bzw.

$$\theta = \frac{\varepsilon_i}{\hat{\varepsilon}_i} < 1 \tag{7.2}$$

gilt, da $\theta \to 0$ strebt, wenn $h_i \to 0$ geht.

Wir definieren den Fehlerschätzer nun als

$$\bar{\varepsilon} := \|\bar{\eta}\| \quad \text{mit} \quad \bar{\eta} = \widehat{\Phi}(t_i, \tilde{x}(t_i), h_i) - \Phi(t_i, \tilde{x}(t_i), h_i). \tag{7.3}$$

Der folgende Satz zeigt, dass diese Größe tatsächlich ein Fehlerschätzer im Sinne von Definition 7.1 ist.

Satz 7.2 Betrachte zwei Einschrittverfahren $\widehat{\Phi}$ und Φ mit Konsistenzordnungen $\hat{p}$ und p mit $p \geq \hat{p} + 1$. Dann ist die Größe $\bar{\varepsilon}$ aus (7.3) für alle hinreichend kleinen Schrittweiten $h_i > 0$ ein Fehlerschätzer für $\hat{\varepsilon}_i$ aus (7.1).

Beweis: Wir wählen h_i so klein, dass die Abschätzung (7.2) gilt und $\theta < \theta_0 < 1$ ist. Aus der Definition von $\bar{\eta}$ folgt $\bar{\eta} = \hat{\eta}_i - \eta_i$, also

$$\frac{\|\hat{\eta}_i - \bar{\eta}\|}{\|\hat{\eta}_i\|} = \frac{\|\eta_i\|}{\|\hat{\eta}_i\|} = \frac{\varepsilon_i}{\hat{\varepsilon}_i} = \theta.$$

Damit ergibt sich

$$(1-\theta)\hat{\varepsilon}_i = (1-\theta)\|\hat{\eta}_i\| = \left(1 - \frac{\|\hat{\eta}_i - \bar{\eta}\|}{\|\hat{\eta}_i\|}\right)\|\hat{\eta}_i\| = \|\hat{\eta}_i\| - \underbrace{\|\hat{\eta}_i - \bar{\eta}\|}_{\geq \|\hat{\eta}_i\| - \|\bar{\eta}\|} \leq \|\bar{\eta}\| = \bar{\varepsilon},$$

also die untere Abschätzung mit $\kappa_1 = 1 - \theta_0$ und

$$\bar{\varepsilon} = \|\bar{\eta}\| \leq \|\hat{\eta}_i\| + \|\hat{\eta}_i - \bar{\eta}\| = \left(1 + \frac{\|\hat{\eta}_i - \bar{\eta}\|}{\|\hat{\eta}_i\|}\right)\|\hat{\eta}_i\| = (1+\theta)\|\hat{\eta}_i\| = (1+\theta)\hat{\varepsilon}_i,$$

also die obere Abschätzung mit $\kappa_2 = 1 + \theta_0$. □

Beachte, dass die Gültigkeit des Fehlerschätzers entscheidend von (7.2) abhängt, also nur für bereits hinreichend kleine Schrittweiten gilt.

7.2 Schrittweitenberechnung und adaptiver Algorithmus

Wir wollen nun untersuchen, wie man aus dem geschätzten Fehler effektiv eine neue Schrittweite berechnen kann. Hierzu benötigen wir eine weitere Annahme, nämlich dass der Fehler $\hat{\varepsilon}_i$ für kleine h_i von der Form

$$\hat{\varepsilon}_i \approx c_i h_i^{\hat{p}+1} \tag{7.4}$$

ist. Für Runge–Kutta–Verfahren ist dies erfüllt, falls f $\hat{p}+2$–mal stetig differenzierbar ist, wobei sich die c_i gerade aus dem zu $h_i^{\hat{p}+1}$ gehörigen Koeffizienten der Taylor–Entwicklung ergeben. Allerdings ist der exakte Wert von c_i unbekannt bzw. kann nur mit unverhältnismäßig großem Aufwand berechnet werden.

Sei nun eine Fehlerschranke $tol > 0$ für den lokalen Fehler vorgegeben. Wir führen jeweils einen Schritt mit beiden Verfahren $\widehat{\Phi}$ und Φ zum Zeitschritt h_i durch. Sei $\bar{\varepsilon}$ der gemäß (7.3)

berechnete Fehlerschätzer. Für kleine Schrittweiten gilt $\kappa_1 \approx \kappa_2 \approx 1$, also $\bar{\varepsilon} \approx \hat{\varepsilon}_i \approx c_i h_i^{\hat{p}+1}$. Hieraus können wir einen Schätzwert

$$\bar{c}_i = \frac{\bar{\varepsilon}}{h_i^{\hat{p}+1}}$$

für c_i berechnen. Die gewünschte Fehlertoleranz wird damit (approximativ) für diejenige Schrittweite $h_{i,neu}$ eingehalten, für die die Gleichung

$$tol = \bar{c}_i h_{i,neu}^{\hat{p}+1} = \frac{\bar{\varepsilon}}{h_i^{\hat{p}+1}} h_{i,neu}^{\hat{p}+1}$$

bzw.

$$h_{neu} = \sqrt[\hat{p}+1]{\frac{tol}{\bar{\varepsilon}}}\, h$$

gilt. Da diese Gleichungen (wegen der verschiedenen "$\approx$") nur näherungsweise gelten, führt man in der Praxis noch einen "Sicherheitsfaktor" $fac \in (0,1)$ ein, um die Fehlerquellen bei der Fehlerschätzung zu kompensieren: man setzt

$$h_{i,neu} = \sqrt[\hat{p}+1]{fac\frac{tol}{\bar{\varepsilon}}}\, h_i.$$

Eine typische Wahl hierfür ist $fac = 0.9$.

Nach der Durchführung eines Schrittes mit Schrittweite h_i und der Schätzung des Fehlers $\bar{\varepsilon}$ können nun zwei Fälle auftreten:

(i) $\bar{\varepsilon} > tol$:
In diesem Fall wird der Schritt mit $h_i = h_{i,neu}$ erneut durchgeführt ("zurückweisen und wiederholen").

(ii) $\bar{\varepsilon} \leq tol$:
In diesem Fall wurde die gewünschte Genauigkeit tol erreicht. Der Schritt wird akzeptiert und die neue Schrittweite $h_{i,neu}$ wird als Schrittweite h_{i+1} für den nächsten Schritt verwendet ("akzeptieren").

Beachte, dass die Schrittweite in Schritt (i) immer verkleinert wird. Die Wahl von $h_{i,neu}$ als Schrittweitenvorschlag für h_{i+1} in (ii) ist also ein notwendiger Schritt, damit auch Vergrößerungen der Schrittweite ermöglicht werden und darf daher auf keinen Fall weggelassen werden.

Formal lassen sich unsere Überlegungen in dem folgenden Grundalgorithmus zusammenfassen.

Algorithmus 7.3 (Einschrittverfahren mit Schrittweitensteuerung)

Eingabe: Anfangsbedingung (t_0, x_0), Endzeit T, Toleranz $tol > 0$, Sicherheitsfaktor fac, Einschrittverfahren $\widehat{\Phi}$ und Φ mit unterschiedlichen Kosistenzordnungen $p \geq \hat{p} + 1$, Schrittweitenvorschlag h_0 für den ersten Schritt

(1) Setze $\tilde{x}_0 = x_0$, $i = 0$

(2) Falls $t_i = T$, beende den Algorithmus; falls $t_i + h_i > T$, setze $h_i = T - t_i$.

(3) Berechne $t_{i+1} = t_i + h_i$, $\tilde{x}^1_{i+1} = \Phi(t_i, \tilde{x}_i, h_i)$, $\tilde{x}^2_{i+1} = \widehat{\Phi}(t_i, \tilde{x}_i, h_i)$, den Fehlerschätzer $\bar{\varepsilon}$ und den Schrittweitenvorschlag $h_{i,neu}$

(4) Falls $\bar{\varepsilon} > tol$ setze $h_i = h_{i,neu}$ und gehe zu (3)

(5) Falls $\bar{\varepsilon} \leq tol$ setze $\tilde{x}_{i+1} := \tilde{x}^1_{i+1}$, $h_{i+1} := h_{i,neu}$, $i := i + 1$ und gehe zu (2)

Ausgabe: Werte der Gitterfunktion $\tilde{x}(t_i) = \tilde{x}_i$ in $t_0, \ldots, t_N = T$, □

Beachte, dass wir in (5) die genauere Lösung $\tilde{x}^1_{i+1}$ zum Weiterrechnen und für die Ausgabe verwenden. Diese Praxis wurde früher (und zum Teil noch heute) abgelehnt, da der Fehlerschätzer ja den Fehler in $\tilde{x}^2_{i+1}$ misst. Da das gesamte Verfahren aber auf der Annahme (7.2) beruht, die gerade besagt, dass Φ (also $\tilde{x}^1_{i+1}$) eine genauere Approximation ist, ist es durchaus gerechtfertigt, diesen Wert zu verwenden.

In der Praxis wird der Algorithmus in mehreren Punkten verfeinert:

(i) Statt in der euklidischen Norm wird $\bar{\varepsilon}$ in der Maximumsnorm

$$\bar{\varepsilon} = \|\bar{\eta}\|_\infty = \max_{i=1,\ldots,n} |\bar{\eta}_i|$$

berechnet, da diese schneller auszuwerten ist.

(ii) Der Bruch $tol/\bar{\varepsilon}$ in der Berechnung der neuen Schrittweite wird durch einen Wert ersetzt, in dem der absolute und der relative Fehler eingeht. Z.B. verwendet man statt $tol/\bar{\varepsilon}$ den Wert $1/err$ mit

$$err = \max_{j=1,\ldots,n} \frac{|\bar{\eta}_j|}{atol + |\widehat{\Phi}_j| \cdot rtol}$$

für absolute und relative Fehlertoleranzen $atol$ und $rtol > 0$; das Fehlerkriterium $\bar{\varepsilon} \leq tol$ wird dabei zu $err \leq 1$. Damit wird bei betragsmäßig großen Lösungskomponenten $|\widehat{\Phi}_j|$ ein größerer Fehler erlaubt, was Probleme mit Rundungsfehlern vermeidet, die bei sehr großen Komponenten ebenfalls groß werden können, weswegen eine rein absolute Fehlertoleranz in diesem Fall nicht einzuhalten wäre.

(iii) Die erlaubte Schrittweite wird durch Schranken h_{min} und h_{max} nach unten und oben beschränkt. Falls für die berechnete Schrittweite $h_{neu} < h_{min}$ gilt, so wird eine Warnung ausgegeben oder mit einer Fehlermeldung abgebrochen.

(iv) Der Variationsfaktor der Schrittweite, der durch

$$\sqrt[\hat{p}+1]{fac\frac{tol}{\bar{\varepsilon}}} \quad \text{bzw. allgemeiner durch} \quad \sqrt[\hat{p}+1]{fac\frac{1}{err}}$$

gegeben ist, wird durch Schranken fac_{min} und fac_{max} nach unten und oben beschränkt. Dadurch werden starke Schwankungen der Schrittweite vermieden.

(v) Im Falle eines Fehlberg–Verfahrens (vgl. Abschnitt 7.3) setzt man in Schritt (5) $h_{i+1} = h_i$ falls $h_{i,neu} \approx h_i$. Damit kann man Zwischenergebnisse aus dem i–ten Schritt effizient im $i+1$–ten Schritt verwenden.

Einige dieser Punkte werden in der Programmieraufgabe auf dem aktuellen Übungsblatt berücksichtigt; dort ist auch der oben angegebene Algorithmus noch einmal in etwas anderer Form dargestellt.

Abbildung 7.1 zeigt die Anwendung dieses Algorithmus auf das aus den Übungen bekannte restringierte Dreikörperproblem (Satellitenlaufbahn). Die Gitterpunkte t_i sind auf der in Kurvenform dargestellten Lösung markiert. Das Beispiel wurde mit der Routine `ode45` in MATLAB mit $atol = rtol = 10^{-7}$ gerechnet; die Routine verwendet zwei Runge–Kutta–Verfahren der Konsistenzordnung 4 und 5.

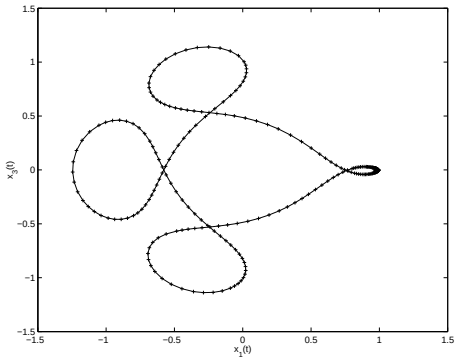

Abbildung 7.1: Adaptive Schrittweitensteuerung an einem Beispiel

7.3 Eingebettete Verfahren

Die in vielen Beispielen sehr effiziente Schrittweitensteuerung hat den Nachteil, dass man zur Berechnung des Fehlerschätzers zwei Einschrittverfahren $\widehat{\Phi}$ und Φ in jedem Schritt auswerten muss. Der Aufwand dieser Auswertungen kann allerdings beträchtlich reduziert werden, wenn man hierfür geschickt gewählte Verfahren verwendet, die sogenannten *eingebetteten Runge–Kutta–Verfahren.*

Wir betrachten zur Erläuterung zwei Verfahren $\widehat{\Phi}$ und Φ mit Konsistenzordnungen $\hat{p}$ und $p \geq \hat{p} + 1$. Bezeichnen wir die Stufen der Verfahren mit $\hat{k}_i$ bzw. k_i, so besteht die Idee der Einbettung einfach darin, dass die Verfahren so konstruiert werden, dass $\hat{k}_i = k_i$ für $i = 1, \ldots, s$ gilt. Für die Koeffizienten der Verfahren muss also $\hat{a}_{ij} = a_{ij}$ und $\hat{c}_i = c_i$ gelten, weswegen wir bei den alten Bezeichnugen a_{ij} und c_i bleiben. Lediglich $\hat{b}_i$ und b_i unterscheiden sich. Ein solches Paar $(\Phi, \widehat{\Phi})$ eingebetteter Verfahren wird mit RK$p(\hat{p})$ bezeichnet. Sie werden in einem Butcher–Tableau der Form

$$
\begin{array}{c|ccccc}
c_1 & & & & & \\
c_2 & a_{2\,1} & & & & \\
c_3 & a_{3\,1} & a_{3\,2} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s\,1} & a_{s\,2} & \cdots & a_{s\,s-1} & \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \cdots & \hat{b}_{s-1} & \hat{b}_s
\end{array}
$$

dargestellt. Um zu zeigen, dass eine solche Einbettung nicht ganz trivial ist, betrachten wir das klassische Runge–Kutta–Verfahren mit Ordnung 4, das durch die Koeffizienten

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

gegeben ist. Wir wollen dieses als Verfahren Φ der Ordnung $p = 4$ verwenden und versuchen, Koeffizienten $\hat{b}^T = (\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4)$ zu finden, so dass

$$\widehat{\Phi}(t, x, h) = x + h \sum_{i=1}^{4} \hat{b}_i k_i$$

ein Verfahren $\widehat{\Phi}$ der Ordnung $p = 3$ ergibt, womit wir ein RK4(3)–Verfahren erhalten würden. Wenn man die Bedingungsgleichungen aus Satz 4.5 (iii) löst, so stellt man fest, dass die einzige Lösung durch $\hat{b}^T = (1/6, 1/3, 1/3, 1/6)$ gegeben ist. Wir erhalten damit $\widehat{\Phi} = \Phi$, was keine sinnvolle Lösung ist, da sich die zwei Verfahren in der Konsistenzordnung echt unterscheiden müssen. So paradox es erscheinen mag: Um ein Verfahren niedrigerer Konsistenzordnung zu erhalten, müssen wir eine Stufe hinzunehmen, also s um 1 erhöhen.

Um die Berechnung der nötigen weiteren Stufe (nun wieder mit k_s bezeichnet) möglichst effizient zu gestalten, hilft ein Trick, den E. Fehlberg Ende der 1960er Jahre entwickelt hat: Wir wählen die letze Stufe gerade so, dass

$$k_s = k_1^* \tag{7.5}$$

gilt, wobei k_1^* die erste Stufe des nächsten Schritts des Verfahrens bezeichnet. Damit muss man trotzdem eine Stufe mehr berechnen, kann diese aber speichern und im nächsten Schritt des Verfahrens verwenden, wenn die Schrittweite $h_{i+1} = h_i$ gewählt werden kann (vgl. Punkt (v) in den praktischen Anmerkungen zu Algorithmus 7.3). Ein s–stufiges Verfahren mit diesem Trick ist also effektiv ein $s - 1$–stufiges Verfahren.

Der Fehlberg–Trick lässt sich in Bedingungen an die Koeffizienten der letzten Stufe s ausdrücken. Wegen Konsistenz und Autonomieinvarianz gilt $k_1 = f(t, x)$, also $k_1^* = f(t + h, \Phi(t, x, h))$. Damit ergibt sich (7.5) zu

$$\underbrace{f(t + c_s h, x + h \sum_{j=1}^{s-1} a_{sj} k_j)}_{=k_s} = \underbrace{f(t + h, x + h \sum_{j=1}^{s} b_j k_j)}_{=k_1^*},$$

was gerade dann der Fall ist, wenn für die Koeffizienten der s–ten Stufe die Bedingungen

$$c_s = 1, \quad b_s = 0, \quad a_{sj} = b_j \text{ für } j = 1, \ldots, s - 1 \tag{7.6}$$

gelten. Beachte dass es keine Garantie gibt, dass dieser Trick wirklich auf eine sinnvolle Lösung für $\hat{b}$ führt; wenn dies aber gelingt, so liefert er eine sehr effiziente Lösung.

Wir wollen diesen Trick auf das klassische Runge–Kutta–Verfahren anwenden. Dazu müssen wir die unbestimmten Koeffizienten in dem Butcher–Tableau

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & & \\ 1 & 0 & 0 & 1 & & \\ c_5 & a_{51} & a_{52} & a_{53} & a_{54} & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} & 0 \\ \hline & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & \hat{b}_4 & \hat{b}_5 \end{array}$$

bestimmen. Aus (7.6) erhalten wir

$$c_5 = 1, \quad a_{51} = a_{54} = \frac{1}{6}, \quad a_{52} = a_{53} = \frac{1}{3}$$

und aus den Bedingungsgleichungen von Satz 4.5 (iii) erhält man die Lösungen

$$\hat{b}^T = (1/6, 1/3, 1/3, 1/6, 0) \quad \text{und} \quad \hat{b}^T = (1/6, 1/3, 1/3, 0, 1/6).$$

Die erste führt wiederum auf $\widehat{\Phi} = \Phi$, die zweite hingegen führt tatsächlich auf ein Verfahren mit maximaler Konsistenzordnung 3. Zusammenfassend erhalten wir

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

Dieses von Fehlberg Ende der 1960er Jahre entwickelte Verfahren ist für anspruchsvolle numerische DGL–Probleme durchaus schon zu gebrauchen.

Die heutzutage gebräuchlichsten eingebetteten Runge–Kutta–Verfahren wurden allerdings erst Anfang der 1980er Jahren von J.R. Dormand und P.J. Prince entwickelt. Es handelt sich um ein effektiv 6–stufiges RK5(4)–Verfahren mit den Koeffizienten

$$\begin{array}{c|ccccccc} 0 & & & & & & & \\ \frac{1}{5} & \frac{1}{5} & & & & & & \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\ \frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\ \frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\ 1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\ 1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\ \hline & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \hline & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \end{array}$$

sowie um ein 13–stufiges RK8(7)–Verfahren, das sich z.B. im Abschnitt 5.4 des Buches von Deuflhard/Bornemann findet. Diese Verfahren sind deswegen besonders gut, weil der von f unabhängige Anteil der Konstanten E in der Konsistenzabschätzung für Φ sehr klein im Vergleich zu anderen Verfahren ist. Das Dormand–Prince–RK5(4)–Verfahren ist MATLABs "Standard–Löser" und ist dort unter dem Namen `ode45` implementiert. Im Internet finden sich MATLAB Implementierungen des RK8(7)–Verfahrens unter dem Namen `ode87.m` (zu finden mit Google mit dem Suchbegriff `ode87 matlab`).

Kapitel 8

Extrapolationsverfahren

Die Konstruktion von expliziten Runge–Kutta–Verfahren über die in Satz 4.5 angegebenen Bedingungsgleichungen an die Koeffizienten ist i.A. kompliziert und für Konsistenzordnungen $p \geq 10$ kaum durchführbar. Als Alternative gibt es Einschrittverfahren, die mit anderen Methoden hergeleitet und implementiert werden. Ein Beispiel hierfür sind die expliziten Extrapolationsverfahren, die wir in diesem Kapitel betrachten werden[1]. Tatsächlich liefert die Extrapolationsidee aber nicht etwa eine neue Verfahrensklasse, sondern wieder Runge-Kutta-Verfahren, die allerdings ganz anders implementiert werden. Der Zusammenhang wird auf dem aktuellen Übungsblatt genauer untersucht.

8.1 Theoretische Grundlagen

Die Extrapolationsverfahren für DGL beruhen auf der Idee, ein Polynom durch numerische Näherungen zu verschiedenen Zeitschritten $h_i > 0$ zu legen und dieses dann in $h = 0$ auszuwerten. Dies wird unten noch einmal genauer beschrieben. Die Grundlage dafür, dass dieses Verfahren funktioniert bildet der folgende Satz von Gragg (bewiesen im Jahre 1964, eine frühere Version wurde 1962 von Henrici bewiesen).

Satz 8.1 Betrachte ein Einschrittverfahren Φ mit Konvergenzordnung p. Wir bezeichnen die zugehörige approximative Lösung mit Anfangsbedingung (t_0, x_0) und äquidistantem Zeitschritt $h > 0$ zur Zeit $t > t_0$ als $\tilde{x}_h(t)$. Dann gilt: Falls das Vektorfeld f und die Abbildung Φ mindestens $p+k$–mal stetig differenzierbar sind, so existieren stetig differenzierbare Funktionen $e_0, \ldots, e_{k-1} : \mathbb{R} \to \mathbb{R}^n$, so dass die asymptotische Entwicklung

$$\tilde{x}_h(t) = x(t; t_0, x_0) + e_0(t)h^p + \ldots + e_{k-1}(t)h^{p+k-1} + O(h^{p+k}) \tag{8.1}$$

gilt.

Der Beweis, der auf einer geschickt gewählten Taylor–Entwicklung beruht, findet sich in [2, Satz 4.37].

[1] Ein anderes Beispiel sind die impliziten Kollokationsverfahren, die im nachfolgenden Kapitel behandelt werden.

Bemerkung 8.2 Diese Entwicklung muss für $k \to \infty$ nicht konvergieren, selbst wenn f und Φ analytisch sind. Für unsere Zwecke sind wir allerdings auch nicht am Verhalten für $k \to \infty$, sondern am Verhalten für festes k und $h \to 0$ interessiert. □

Wir werden uns bei der Beschreibung der Extrapolation auf einen Spezialfall von (8.1) einschränken, bei dem die Konstruktion besonders einfach wird. Wir nehmen dazu an, dass das Verfahren eine asymptotische Entwicklung der Form

$$\tilde{x}_h(t) = x(t; t_0, x_0) + e_0(t)h^p + e_p(t)h^{2p} + \ldots + e_{p(m-2)}(t)h^{p(m-1)} + O(h^{pm}) \tag{8.2}$$

besitzt. Diese Form folgt unter geeigneten Bedingungen aus (8.1), z.B. wenn $p = 1$ ist oder wenn $e_i = 0$ gilt für alle i mit $i \neq lp$ für alle $l \in \mathbb{N}$. Gleichung (8.2) gilt für das Euler-Verfahren mit $p = 1$; andere Verfahren, die diese Bedingung erfüllen, diskutieren wir am Ende dieses Abschnitts.

Die Idee der Extrapolation ist nun, aus einem weniger genauen Verfahren Φ und der asymptotischen Entwicklung (8.2) eine genauere Approximation zu erhalten. Die Grundidee verläuft dabei wie folgt:

- Für ein Verfahren Φ berechnen wir approximative Lösungen $\tilde{x}_{h_i}(t)$ für $x(t; t_0, x_0)$ zur Zeit $t > t_0$ und verschiedenen Schrittweiten $h_1 > \ldots > h_{k+1} > 0$ und erhalten so Wertepaare $(h_i^p, \tilde{x}_{h_i}(t))$, $i = 1, \ldots, k+1$.
- Durch diese Werte legen wir ein Interpolationspolynom $P(h^p)$ und werten dieses in $h^p = 0$ aus. Da $h^p = 0$ *außerhalb* der Stützstellen $h_1^p, \ldots, h_{k+1}^p$ des Polynoms liegt, spricht man von *Extra*polation.
- Mittels $\Phi_E(t_0, x_0, h_0) := P(0)$, wobei P mit $t = t_0 + h_0$ berechnet wird, kann man mit der Extrapolation ein neues Einschrittverfahren Φ_E erzeugen. Falls Φ ein explizites Runge-Kutta-Verfahren ist, so ist auch Φ_E ein explizites Runge-Kutta-Verfahren. Dieses Zusammenhang wird auf dem aktuellen Übungsblatt genauer untersucht.

Der folgende Satz zeigt, dass dieses Vorgehen tatsächlich eine Approximation höherer Genauigkeit liefert.

Satz 8.3 Betrachte ein Einschrittverfahren mit Konvergenzordnung p und asymptotischer Entwicklung (8.2). Dann liefert die oben beschriebene Extrapolation mit $k = m - 1$ eine Approximation $P(0)$ von $x(t; t_0, x_0)$ der Ordnung $O(h_1^{mp})$.

Beweis: Wir betrachten zwei Interpolationspolynome

$$\begin{aligned} Q(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k \\ \widetilde{Q}(x) &= \tilde{a}_0 + \tilde{a}_1 x + \tilde{a}_2 x^2 + \ldots + \tilde{a}_k x^k \end{aligned}$$

zu Daten (x_i, f_i) und $(x_i, \tilde{f}_i)$, $i = 0, \ldots, k$. Aus den Lagrange–Darstellungen

$$\begin{aligned} Q(x) &= \sum_{i=0}^{k} L_i(x) f_i \\ \widetilde{Q}(x) &= \sum_{i=0}^{k} L_i(x) \tilde{f}_i \end{aligned}$$

sieht man durch Ausmultiplizieren, dass die Koeffizienten a_i und $\tilde{a}_i$ die Abschätzung

$$|a_i - \tilde{a}_i| \leq C \max_{j=0,\ldots,k} |f_j - \tilde{f}_j|$$

für eine von den L_i abhängige Konstante $C > 0$ erfüllen.

Wir betrachten nun das durch die Daten $(h_i^p, \tilde{x}_{h_i}(t))$, $i = 1, \ldots, m$ definierte Interpolationspolynom

$$P(h^p) = a_0 + a_1 h^p + a_2 (h^p)^2 + \ldots + a_{m-1}(h^p)^{m-1}$$

und vergleichen dieses mit dem durch Abschneiden von (8.2) gewonnenen Polynom

$$\begin{aligned} \widetilde{P}(h^p) &= \tilde{a}_0 + \tilde{a}_1 h^p + \tilde{a}_2 (h^p)^2 + \ldots + \tilde{a}_{m-1}(h^p)^{m-1} \\ &= x(t; t_0, x_0) + e_0(t) h^p + e_p(t) h^{2p} + \ldots + e_{p(m-2)}(t) h^{p(m-1)}. \end{aligned}$$

An den Stützstellen h_i^p unterscheiden sich die Werte dieser Polynome (komponentenweise betrachtet, da $f_i \in \mathbb{R}^n$) um $O(h_i^{mp})$, weswegen sich auch die Komponenten der (vektorwertigen) Koeffizienten a_0 und $\tilde{a}_0$ um höchstens

$$C \max_{j=1,\ldots,m} O(h_j^{mp}) = O(h_1^{mp})$$

unterscheiden. Damit folgt

$$P(0) = a_0 = \tilde{a}_0 + O(h_1^{mp}) = x(t; t_0, x_0) + O(h_1^{mp}),$$

also die Behauptung. □

8.2 Algorithmische Umsetzung

Die im vorherigen Abschnitt skizzierte Extrapolationsidee kann als sogenanntes Diagonalschema implementiert werden. Dieses Schema ermöglicht die iterative Berechnung von $P(0)$ für eine wachsende Anzahl von Stützstellen, ohne dass wir diese Polynome explizit aufstellen müssen.

Wir wählen dazu eine aufsteigende Folge $n_i \in \mathbb{N}$ und setzen $h_i = (t - t_0)/n_i$ bzw. $h_i = h_0/n_i$ für die Schrittweite $h_0 > 0$ des neuen Einschrittverfahrens Φ_E. Wir bezeichnen die mit dem Verfahren Φ zur Anfangsbedingung (t_0, x_0) und Zeitschritt h_i erhaltenen Lösungen mit $T_{i,1} = \tilde{x}_{h_i}(t)$. Mit

$$P_{i,k}(h^p)$$

bezeichnen wir die durch die Stützstellen $(h_{i-k+1}^p, T_{i-k+1,1}), \ldots, (h_i^p, T_{i,1})$ definierten Interpolationspolynome in h^p und mit $T_{i,k} = P_{i,k}(0)$ ihre Werte in $h^p = 0$. Gemäß Satz 8.3 liefern die Diagonalwerte $T_{k,k}$ also eine Approximation der Ordnung $O(h_1^{kp})$ für die Lösung $x(t; t_0, x_0)$ zur Zeit t. Das folgende Lemma zeigt, wie die Werte $T_{i,k}$ iterativ berechnet werden können.

Lemma 8.4 Für die Werte $T_{i,k}$ gilt die Rekursionsformel

$$T_{i,k} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k+1}}{h_i}\right)^p - 1}, \quad k = 2, 3, \ldots; \;\; i = k, k+1, \ldots$$

Beweis: Durch Nachprüfen der Interpolationseigenschaft sieht man leicht, dass für die Interpolationspolynome $P_{i,k}$ die Gleichung

$$P_{i,k}(h^p) = \frac{(h^p_{i-k+1} - h^p)P_{i,k-1}(h^p) - (h^p_i - h^p)P_{i-1,k-1}(h^p)}{h^p_{i-k+1} - h^p_i}$$

gilt, die auch als *Lemma von Aitken* bekannt ist. Damit folgt die oben angegebene Formel durch Auswerten in $h^p = 0$ und Kürzen des Bruchs mit h^p_i. □

Die Struktur dieser Berechnung, die als *Extrapolationsschema* bezeichnet wird, ist in Abb. 8.1 grafisch dargestellt.

$$\begin{array}{ccccccccc}
T_{1,1} & & & & & & & & \\
 & \searrow & & & & & & & \\
T_{2,1} & \rightarrow & T_{2,2} & & & & & & \\
 & \searrow & & \searrow & & & & & \\
T_{3,1} & \rightarrow & T_{3,2} & \rightarrow & T_{3,3} & & & & \\
 & \searrow & & \searrow & & \searrow & & & \\
T_{4,1} & \rightarrow & T_{4,2} & \rightarrow & T_{4,3} & \rightarrow & T_{4,4} & & \\
\vdots & & \vdots & & \vdots & & \vdots & \ddots &
\end{array}$$

Abbildung 8.1: Illustration des Extrapolationsschemas

In der Praxis verwendet man zur Berechnung der h_i oft die naheliegendste Folge $n_i = (1, 2, 3, 4, 5, 6, \ldots)$. Eine weitere Möglichkeit ist die Halbierung der Schrittweite, also die Folge $n_i = (1, 2, 4, 8, 16, \ldots)$.

Natürlich will man auch mit Extrapolationsverfahren i.A. nicht nur *einen* Wert $\tilde{x}(t)$ sondern eine Gitterfunktion $\tilde{x}(t_i)$ auf einem Zeitgitter $(t_i)_{i\in\mathbb{N}}$ auf $[t_0, T]$ berechnen. Wenn wir eine gewünschte Extrapolationsordnung kp fixieren und das Verfahren mit $t = t_0 + h_0$ anwenden, ergibt sich das neue Einschrittverfahren als $\Phi_E(t_0, x_0, h_0) := T_{k,k}$, mit dem sich in üblicher Form die gewünschte Gitterfunktion berechnen lässt. Die Schrittweitensteuerung ist hier besonders effizient zu implementieren, da man für den Fehlerschätzer mit $T_{k,k-1}$ und $T_{k-1,k-1}$ gleich zwei Approximationen niedrigerer Ordnung zur Verfügung hat, ohne weitere Berechnungen durchführen zu müssen. In der Praxis wählt man üblicherweise $\widehat{\Phi}_E(t_0, x_0, h) := T_{k,k-1}$ für die Fehlerschätzung, da dieser Ausdruck eine genauere Approximation liefert.

Wir wollen abschließend untersuchen, welche Einschrittverfahren sich als Basisverfahren Φ der Extrapolation eignen. Betrachtet man die zu Grunde liegende asymptotische Entwicklung (8.2), so sieht man (aus Satz 8.1), dass das Euler–Verfahren die Voraussetzung für $p = 1$ erfüllt. Dieses Verfahren kann also als Basis der Extrapolation verwendet werden.

Effizienter wäre es aber sicherlich, wenn wir ein Verfahren Φ verwendeten, welches (8.2) für $p > 1$ erfüllt, da wir mit jedem Extrapolationsschritt die Ordnung um den Faktor p erhöhen. Leider kann man nachweisen, dass es kein explizites Runge–Kutta–Verfahren Φ gibt, für das dieses gilt.

Wir wollen den Fall $p = 2$ genauer untersuchen. Hier lässt sich ein Kriterium angeben, unter dem (8.2) gilt, wobei wir annehmen, dass das betrachtete Einschrittverfahren von der Form $\Phi(t, x, h) = x + h\varphi(t, x, h)$ ist, vgl. Lemma 2.6.

Satz 8.5 Falls das Einschrittverfahren *reversibel* ist, d.h. die Bedingung

$$\Phi(t + h, \Phi(t, x, h), -h) = x \tag{8.3}$$

erfüllt und die Konsistenz– und Konvergenzordnung $p = 2$ besitzt, so existiert für hinreichend glattes Vektorfeld f eine asymptotische Entwicklung der Form (8.2) mit $p = 2$.

Beweisskizze: Betrachte die exakte Lösung $x(t)$ und die $e_i(t)$, $i = 0, 1, 2, \ldots$ aus Satz 8.1. Für ein beliebiges $k \in \mathbb{N}$ definieren wir

$$x^*(t) = x(t) + e_0(t)h^2 + e_1(t)h^3 + \ldots + e_{k-1}(t)h^{2+k-1}.$$

Mittels Taylor–Entwicklung von $\Phi(t, x^*(t), h)$ nach x und h im Punkt $(t, x(t), 0)$ sieht man, dass dann eine differenzierbare Funktion $d_k(t)$ existiert, so dass die Gleichungen

$$\begin{aligned} x^*(t + h) - \Phi(t, x^*(t), h) &= d_k(t)h^{2+k+1} + O(h^{2+k+2}) \\ x^*(t) - \Phi(t + h, x^*(t + h), -h) &= d_k(t + h)(-h)^{2+k+1} + O(h^{2+k+2}) \end{aligned}$$

gelten. Durch Koeffizientenvergleich erhält man dabei die Gleichung

$$d_k(t) = e_k(t),$$

wobei e_k der Koeffizient aus Satz 8.1 für $k = i$ ist. Da d_k differenzierbar ist, folgt

$$d_k(t + h)h^{2+k+1} = d_k(t)h^{2+k+1} + O(h^{2+k+2})$$

und da Φ differenzierbar ist auch

$$\begin{aligned} &\Phi(t + h, x^*(t + h) + d_k(t)h^{2+k+1}, -h) \\ &\quad = x^*(t + h) + d_k(t)h^{2+k+1} - h\varphi(t + h, x^*(t + h) + d_k(t)h^{2+k+1}, -h) \\ &\quad = x^*(t + h) + d_k(t)h^{2+k+1} - h\varphi(t + h, x^*(t + h), -h) + O(h^{2+k+2}). \end{aligned}$$

Aus der Reversibilität von Φ folgt damit

$$\begin{aligned} x^*(t) &= \Phi(t, \Phi(t, x^*(t), h), -h) \\ &= \Phi(t, x^*(t + h) - d_k(t)h^{2+k+1} + O(h^{2+k+2}), -h) \\ &= \Phi(t, x^*(t + h), -h) - d_k(t)h^{2+k+1} + O(h^{2+k+2}) \\ &= x^*(t) - d_k(t + h)(-h)^{2+k+1} - d_k(t)h^{2+k+1} + O(h^{2+k+2}) \\ &= x^*(t) + ((-1)^{2+k} - 1)d_k(t)h^{2+k+1} + O(h^{2+k+2}). \end{aligned}$$

Da dies für alle $h > 0$ gelten muss, folgt $((-1)^{2+k} - 1)d_k(t)h^{2+k+1} = 0$. Falls k ungerade ist, folgt damit $d_k = 0$, also $e_k(t) = 0$. Damit erhalten wir (8.2) für $p = 2$. □

Leider ist Reversibilität eine Eigenschaft, die kein explizites Runge–Kutta–Verfahren besitzt. Wir haben im Beweis von Lemma 4.3 gesehen, dass jede explizite Runge–Kutta–Approximation der Gleichung $\dot{x}(t) = x(t)$ mit $x(0) = x_0$ für alle $t \in \mathbb{R}$ von der Form

$\Phi(t, x_0, h) = P(h)x_0$ für ein Polynom in h ist. Die Bedingung (8.3) würde $P(h)P(-h) = 1$ erzwingen, was für nichtkonstante Polynome unmöglich ist.

Zwar gibt es reversible implizite Runge-Kutta-Verfahren der Ordnung $p = 2$, wie z.B. die implizite Trapezregel (5.1), für die man die Reversibilität der Vorschrift

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_i) + \frac{h}{2}\Big(f(t_i, \tilde{x}(t_i)) + f(t_{i+1}, \tilde{x}(t_{i+1}))\Big)$$

leicht überprüfen kann. Allerdings ist dies ein implizites Verfahren, wodurch wir den Vorteil verlieren, dass die Extrapolation zur Konstruktion expliziter Verfahren verwendet werden kann.

Als Ausweg müssen wir eine andere Klasse von Verfahren betrachten. Hier kann man z.B. ein Verfahren Φ verwenden, das als *explizite Mittelpunktregel* bekannt ist und für äquidistante Stützstellen $h_i = h$ durch

$$\tilde{x}(t_{i+2}) = \tilde{x}(t_i) + 2hf(t_{i+1}, \tilde{x}(t_{i+1}))$$

gegeben ist. Dieses Verfahren ist reversibel, wenn man es als Abbildung von $\tilde{x}(t_i)$ nach $\tilde{x}(t_{i+2})$ auffasst.

Allerdings ist dies kein Einschrittverfahren, da die rechte Seite von $\tilde{x}(t_{i+1})$ und $\tilde{x}(t_i)$ abhängt. Wir haben es hier mit einem sogenannten *Mehrschrittverfahren* zu tun, einer Klasse von Verfahren, die wir in Kapitel 12 systematisch untersuchen wollen. Um das Verfahren zu starten, benötigen wir neben dem Anfangswert $\tilde{x}(t_0) = x_0$ noch den Wert $\tilde{x}(t_1)$, der (um die Konvergenzordnung $p = 2$ zu erhalten) mindestens mit Genauigkeit $O(h^2)$ bestimmt werden muss. Dies kann durch einen einfachen Euler–Schritt, also mittels $\tilde{x}(t_1) = x_0 + hf(t_0, x_0)$ geschehen.

Kapitel 9

Kollokationsmethoden

Eine andere Art, Runge-Kutta-Verfahren hoher Konsistenzordnung zu konstruieren, ohne dabei direkt die Bedingungsgleichungen aus Satz 3.7 zu verwenden, ist die sogenannte Kollokation. Auch diese beruht auf der Idee der Polynominterpolation, allerdings wird hierbei die Lösung der Gleichung selbst durch ein Polynom angenähert. Die Idee liegt darin, ein Polynom zu konstruieren, welches die Differentialgleichung an einer vorgegebenen Menge von Zeiten $t+c_1h, \dots, t+c_sh$ exakt erfüllt. Die Kollokation führt dabei in der Regel auf implizite Verfahren.

Definition 9.1 Sei $s \in \mathbb{N}$ und $c_1, \dots, c_s \in [0,1]$. Das *Kollokationspolynom* $p(t)$ vom Grad s ist das Polynom $p \in \mathcal{P}_s$, welches für gegebene $x \in \mathbb{R}^n$, $t_0 \in \mathbb{R}$, $h > 0$ und $f : D \to \mathbb{R}^n$, $D \subseteq \mathbb{R} \times \mathbb{R}^n$ die Bedingungen

$$p(t_0) = x \quad \text{und} \quad \dot{p}(t_0 + c_ih) = f(t_0 + c_ih, p(t_0 + c_ih)) \quad \text{für alle } i = 1, \dots, s$$

erfüllt. Das *Kollokationsverfahren* ist dann gegeben durch

$$\Phi(t_0, x, h) = p(t_0 + h).$$

□

Beispiel 9.2 Für $s = 1$ ist p von der Form $p(t) = p_0 + p_1(t - t_0)$, also $\dot{p}(t) = p_1$. Aus der ersten Bedingung erhält man sofort $p_0 = x$. Für $c_1 = 0$ ergibt sich die zweite Bedingung zu

$$p_1 = f(t_0, p(t_0)) = f(t_0, x)$$

und man erhält $\Phi(t_0, x, h) = p(t_0 + h) = p_0 + hp_1 = x + hf(t_0, x)$, also das explizite Euler-Verfahren. Mit ähnlichen Rechnungen sieht man, dass man für $c_1 = 1$ das implizite Euler-Verfahren und für $c_1 = 1/2$ die Mittelpunktregel $\tilde{x}(t_0 + h) = \tilde{x}(t_0) + hf(t_0 + h/2, (\tilde{x}(t_0 + h) + \tilde{x}(t_0))/2)$ erhält.

Für $s = 2$ und $c_1 = 0$, $c_2 = 1$ erhält man die implizite Trapezregel. □

Bemerkung 9.3 Beachte, dass wir hier stillschweigend angenommen haben, dass ein Interpolationspolynom mit den angegebenen Bedingungen existiert und eindeutig ist. Letzteres ist nicht unbedingt der Fall, genauso wie implizite Runge-Kutta-Verfahren nicht unbedingt eine eindeutige Lösung besitzen müssen. Ein Gegenbeispiel werden wir in Bemerkung 9.6 betrachten. □

Der folgende Satz zeigt, dass die Kollokationsmethode tatsächlich wieder Runge-Kutta-Verfahren erzeugt.

Satz 9.4 Die Kollokationsmethode aus Definition 9.1 liefert das gleiche Einschrittverfahren Φ wie das s-stufige Runge-Kutta-Verfahren mit Koeffizienten $c_1, \ldots, c_s$,

$$a_{ij} = \int_0^{c_i} L_j(\tau) d\tau, \qquad b_i = \int_0^1 L_i(\tau) d\tau \tag{9.1}$$

mit den Lagrange-Polynomen

$$L_i(\tau) = \prod_{\substack{j=1 \\ j \neq i}}^{s} \frac{\tau - c_j}{c_i - c_j}.$$

Beweis: Sei p das Kollokationspolynom und definiere $k_i := \dot{p}(t + c_i h)$. Aus $L_i(t) = 0$ für $t = c_i$ und $L_i(t) = 1$ für $t = c_j$ mit $j \neq i$ sowie der Tatsache, dass sowohl die L_i als auch $\dot{p}$ Polynome vom Grad $s - 1$ sind, folgt

$$\frac{d}{dt}[p(t_0 + th)] = \dot{p}(t_0 + th)h = h \sum_{j=1}^{s} \dot{p}(t_0 + c_j h) L_j(t) = h \sum_{j=1}^{s} k_j L_j(t).$$

Integration dieser Gleichung für t von 0 bis c_i liefert

$$p(t_0 + c_i h) - p(t_0) = h \int_0^{c_i} \sum_{j=1}^{s} k_j L_j(t),$$

was wegen $p(t_0) = x$ äquivalent ist zu

$$p(t_0 + c_j h) = x + h \sum_{j=1}^{s} k_j \int_0^{c_i} L_j(t) = x + h \sum_{j=1}^{s} k_j a_{ij}. \tag{9.2}$$

Einsetzen in die Gleichung für $\dot{p}$ in Definition 9.1 liefert

$$k_i = p(t_0 + c_i h) = f(t_0 + c_i h, p(t + c_i h)) = f(t_0 + c_i h, x + h \sum_{j=1}^{s} k_j a_{ij}),$$

was genau die definierenden Gleichungen der Stufen k_i des Runge-Kutta-Verfahrens sind. Die erste Gleichung in (9.2) mit oberer Integrationsgrenze 1 statt c_i liefert

$$\Phi(t_0, x, h) = p(t_0 + h) = x + h \sum_{j=1}^{s} k_j \int_0^1 L_j(t) = x + h \sum_{j=1}^{s} k_j b_j,$$

und damit die Behauptung. □

Die Bedingungen in (9.1) kann man äquivalent auch als Gleichungssystem für die Koeffizienten ausdrücken: Da $\sum_{j=1}^{s} c_j^{k-1} L_j(\tau)$ für jedes $k = 1, \ldots, s$ gerade das Interpolationspolynom durch (c_j, c_j^{k-1}) ist, gilt $\sum_{j=1}^{s} c_j^{k-1} L_j(\tau) = \tau^{k-1}$. Aus der ersten Bedingung aus (9.1) folgt daher

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{für alle } k = 1, \ldots, q,\ i = 1, \ldots, s \tag{9.3}$$

mit $q = s$ und aus der zweiten Bedingung folgt

$$\sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k} \quad \text{für alle } k = 1, \ldots, p \tag{9.4}$$

mit $p = s$. Da die c_i paarweise verschieden sind, liefern diese Gleichungen lineare Gleichungssysteme mit invertierbaren Matrizen für die Koeffizienten a_{ij} und b_j, weswegen diese eindeutig bestimmt sind. Die Bedingungen (9.3) und (9.4) mit $p = q = s$ sind also äquivalent zu (9.1).

9.1 Konsistenz

Satz 9.5 Jede Kollokationsmethode von Grad s besitzt die Konsistenzordnung s, falls $f \in C^{s+1}(D, \mathbb{R}^n)$ mit $D = \mathbb{R}\times\mathbb{R}^n$ ist. Zudem liefert das Kollokationspolynom mit $p(t_0) = x_0$ für alle $t \in [t_0, t_0 + h]$ eine Approximation der exakten Lösung $x(t) = x(t; t_0, x_0)$ der Ordnung $s + 1$, d.h.,

$$p(t) = x(t) + O(h^{s+1}) \quad \text{für alle } t \in [t_0, t_0 + h].$$

Beweis: Es genügt, die zweite Aussage zu beweisen, da die erste daraus für $t = t_0 + h$ folgt. Wir beweisen die Aussage zunächst für den Fall, dass die Lipschitzkonstante L von f bzgl. x global, also unabhängig von t und x gewählt werden kann.

Betrachte das n-dimensionale Interpolationspolynom q durch die Stützstellen $(t_0+c_ih, f(t_0+c_ih, x(t_0 + c_ih)))$. Für $E(t, h) = f(t, x(t)) - q(t)$, $t \in [t_0, t_0 + h]$ gilt dann

$$\dot{x}(t) = q(t) + E(t, h) = \sum_{i=1}^{s} f(t_0 + c_ih, x(t_0 + c_ih))L_i(t) + E(t, h)$$

Andererseits gilt für das Kollokationpolynom

$$\dot{p}(t) = \sum_{i=1}^{s} f(t_0 + c_ih, p(t_0 + c_ih))L_j(t),$$

also zusammen

$$\dot{x}(t) - \dot{p}(t) = \sum_{i=1}^{s} \underbrace{\big(f(t_0 + c_ih, x(t_0 + c_ih)) - f(t_0 + c_ih, p(t_0 + c_ih))\big)}_{=:\Delta_i f} L_i(t) + E(t, h). \tag{9.5}$$

Auf Grund der Lipschitzannahme an f können wir $\|\Delta_i f\|$ für alle $i = 1, \ldots, s$ abschätzen durch

$$\|\Delta_i f\| \le L \max_{t\in[t_0,t_0+h]} \|x(t) - p(t)\|.$$

Nach dem Satz über den Interpolationsfehler bei der Polynominterpolation gilt zudem

$$\|E(t, h)\| \le h^s \max_{t\in[t_0,t_0+h]} \frac{\|x^{(s+1)}(t)\|}{s!}.$$

Integration von (9.5) von t_0 nach $t \le t_0 + h$ und Ausnutzen von $x(t_0) - p(t_0) = 0$ liefert

$$x(t) - p(t) = \sum_{i=1}^{s} \Delta_i f \int_{t_0}^{t} L_i(\tau) d\tau + \int_{t_0}^{t} E(\tau, h) d\tau \tag{9.6}$$

und damit

$$\max_{t \in [t_0, t_0+h]} \|x(t) - p(t)\| \le hC_1 L \max_{t \in [t_0, t_0+h]} \|x(t) - p(t)\| + C_2 h^{s+1},$$

woraus die Behauptung folgt wenn $hC_1 L < 1$.

Falls f nicht global Lipschitz stetig in x ist, betrachten wir eine kompakte Menge K von Anfangsbedingungen und die kompakte Menge K_2 aus dem Beweis von Satz 2.7 mit $T = t_0 + h$. Sei $B = B_R(0) \in \mathbb{R} \times \mathbb{R}^n$, wobei $R > 0$ so groß ist, dass $K_2 \subset B_R(0)$. Wir betrachten nun eine C^{s+1}-Funktion $\rho : \mathbb{R} \to \mathbb{R}$ mit $\rho(r) = 1$ für $r \le R$, $\rho(r) \in [0, 1]$ für $r \in [R, R + 1]$ und $\rho(r) = 0$ für $r \ge R$ und setzen $\tilde{f}(t, x) = \rho(\|(t, x)\|) f(t, x)$. Dann ist $\tilde{f} \in C^{s+1}(D, \mathbb{R}^n)$ und damit Lipschitz und weil die Lipschitz-Konstante L für $(t, x), (t, y) \notin B_{R+1}(0)$ offenbar gleich 0 ist, ist $\tilde{f}$ global Lipschitz. Auf $\tilde{f}$ kann dann der erste Teil des Beweises angewendet werden. Für hinreichend kleines $h > 0$ folgt daraus $p(t) \in K_2$ für alle $t \in [t_0, t_0 + h]$. Dort stimmt f aber mit $\tilde{f}$ überein, weswegen die Konsistenz auch für f gilt. □

Bemerkung 9.6 Der Satz wird i.A. falsch, wenn $D \neq \mathbb{R} \times \mathbb{R}^n$ ist. Betrachte z.B. die eindimensionale autonome Gleichung mit

$$f(x) = \frac{x}{1 - x}$$

und $D = \mathbb{R} \setminus \{1\}$. Für $(t_0, x_0) = (0, 0)$ lautet die Lösung offensichtlich $x(t; 0, 0) \equiv 0$. Kollokation mit $s = 1$ und $c_1 = 1$ (also mit dem impliziten Euler) liefert aber das Interpolationspolynom $p(t) = (1 - h)t/h$, denn es gilt

$$p(0) = 0 \quad \text{und} \quad \dot{p}(0 + h) = \frac{1 - h}{h} = \frac{1 - h}{1 - (1 - h)} = f(1 - h) = f(p(0 + h)).$$

Daraus folgt $\Phi(0, 0, h) = p(0 + h) = 1 - h$, was für $h \to 0$ gegen 1 und nicht wie für Konsistenz nötig gegen 0 konvergiert.

Tatsächlich ist das obige Polynom nicht das eindeutige Interpolationspolynom. Man prüft leicht nach, dass $p(t) \equiv 0$ ebenfalls die Bedingungen des Kollokationsverfahrens erfüllt; für dieses Polynom gilt dann auch die Konsistenz. In der Regel bewirkt ein gut gewählter Startwert, dass man beim iterativen Lösen der nichtlinearen Gleichungen zur Bestimmung von Φ gegen die "richtige", also die konsistente Lösung konvergiert. □

Der folgende Satz zeigt, dass man die Konsistenzordnung unter gewissen Bedingungen noch verbessern kann. Da wir mit diesem Satz eine Konsistenz- und damit auch Konvergenzordnung $p > s$ erhalten kann, spricht man auch von „Superkonvergenz".

Satz 9.7 Betrachte ein Kollokationsverfahren, welches die Bedingung (9.4) für ein $p \in \mathbb{N}$ mit $s \le p \le 2s$ erfüllt. Dann besitzt das Verfahren die Konsistenzordnung p falls $f \in C^{p+1}(D, \mathbb{R}^n)$ und $D = \mathbb{R} \times \mathbb{R}^n$.

Beweis: Wir betrachten das Kollokationspolynom p als Lösung der gestörten Gleichung

$$\dot{p}(t) = f(t, p(t)) + \delta(t)$$

mit $\delta(t) = \dot{p}(t) - f(t, p(t))$. Ziehen wir die exakte Gleichung von dieser Gleichung ab, so erhalten wir mit Taylor-Entwicklung der Ordnung 1

$$\dot{p}(t) - \dot{x}(t) = f(t, p(t)) + \delta(t) - f(t, x(t)) = \frac{\partial f}{\partial x}(t, x(t))\Big(p(t) - x(t)\Big) + \delta(t) + r(t)$$

mit $r(t) = O(\|p(t) - x(t)\|^2) = O(h^{2s+2})$ gemäß Satz 9.5. Aus der Lösungsformel für inhomogene lineare Differentialgleichungen („Variation der Konstanten“) und $p(t_0) - x(t_0) = 0$ folgt

$$p(t_0 + h) - x(t_0 + h) = \int_{t_0}^{t_0+h} \Theta(t_0 + h, \tau)\,(\delta(\tau) + r(\tau))\,d\tau,$$

wobei Θ die Fundamentallösung der homogenen Gleichung $\dot{y}(t) = \frac{\partial f}{\partial x}(t, x(t))y(t)$ bezeichnet. Das Integral über $\Theta(t_0 + h, \tau)r(\tau)$ ist von der Ordnung $O(h^{2s+3})$. Die Funktion $g(\tau) := \Theta(t_0 + h, \tau)\delta(\tau)$ besitzt gerade die Nullstellen $\tau = t_0 + hc_1, \ldots, t_0 + hc_s$. Wenden wir nun die Quadraturformel

$$\int_{t_0}^{t_0+h} g(\tau)d\tau \approx \sum_{j=1}^{s} b_j g(t_0 + hc_j)$$

an, so impliziert Bedingung (9.4), dass diese Quadraturformel Polynome vom Grad $p - 1$ exakt integriert, woraus mit Satz 5.1 aus der „Einführung in die Numerischen Mathematik“

$$\left\| \int_{t_0}^{t_0+h} g(\tau)d\tau - \sum_{j=1}^{s} b_j g(t_0 + hc_j) \right\| \leq Ch^{p+1}$$

und wegen $g(t_0 + hc_j) = 0$ also $\| \int_{t_0}^{t_0+h} g(\tau)d\tau\| \leq Ch^{p+1}$ folgt. Die Konstanten in dem O-Term hängen dabei von den Ableitungen von p ab und ähnlich wie im Beweis von Satz 9.5 kann man zeigen, dass diese durch eine von h unabhängigen Konstante beschränkt sind. Damit folgt die behauptete Konsistenz für $\Phi(t_0, x, h) = p(t_0 + h)$. □

Bemerkung 9.8 Der Beweis zeigt insbesondere, dass die Ordnung des Kollokationsverfahrens durch die Ordnung des Quadraturverfahrens mit Stützstellen c_j und Gewichten b_j bestimmt ist. □

9.2 Beispiele

In diesem Abschnitt geben wir einige Beispiele für Kollokationsverfahren an.

Gauß-Verfahren Wählt man $c_1, \ldots, c_s$ als die Nullstellen des s-ten verschobenen Legendre Polynoms

$$\frac{d^s}{dx^s}\left(x^s(x-1)^s\right),$$

so erhält man eine Quadraturformel mit Ordnung $2s$, vgl. Abschnitt 5.3 der „Einführung in die Numerische Mathematik“. Nach Bemerkung 9.8 besitzt die zugehörige Quadraturformel also die gleiche Ordnung. Ihre Butcher-Tableaus für $s = 2$ und $s = 3$, d.h. mit Ordnungen $p = 4$ und $p = 6$ sind gegeben durch

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

und

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Radau-Verfahren Bei den Radau-Methoden legt man entweder $c_1 = 0$ oder $c_s = 1$ fest und bestimmt die restlichen Koeffizienten dann so, dass die Ordnung maximal, d.h. gleich $2s - 1$ wird. Die Verfahren mit $c_s = 1$ werden Radau IIA-Methoden genannt. Für ein Butcher-Tableau siehe Abschnitt 6.3.

Lobatto IIIA-Verfahren Diese Verfahren besitzen die höchste Ordnung $p = 2s-2$ unter den Bedingungen $c_1 = 0$ und $c_s = 1$. Die Stützstellen $c_2, \ldots, c_{s-1}$ müssen dazu gerade die Nullstellen des Polynoms

$$\frac{d^{s-2}}{dx^{s-2}} \left(x^{s-1}(x-1)^{s-1} \right)$$

sein. Für $s = 2$ erhält man die implizite Trapezregel, für $s = 3$ und $s = 4$ ergeben sich die Butcher-Tableaus

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

und

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 \\ \frac{5-\sqrt{5}}{10} & \frac{11+\sqrt{5}}{120} & \frac{25-\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & \frac{-1+\sqrt{5}}{120} \\ \frac{5+\sqrt{5}}{10} & \frac{11-\sqrt{5}}{120} & \frac{25+13\sqrt{5}}{120} & \frac{25+\sqrt{5}}{120} & \frac{-1-\sqrt{5}}{120} \\ 1 & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} \\ \hline & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} \end{array}$$

9.3 Unstetige Kollokation

Die Klasse der Kollokationsverfahren enthält nicht alle in der Praxis gebräuchlichen impliziten Verfahren. Um die durch die Kollokation möglichen relativ einfachen Konsistenzbeweise auf größere Klassen von Verfahren anzuwenden, wurde die Idee der unstetigen Kollokation entwickelt. Dabei werden in der Kollokation $c_1 = 0$ und $c_s = 1$ gesetzt und die vier Bedingungen

$$p(t_0) = x, \;\; \Phi(t_0, x, h) = p(t_0 + h) \quad \text{und} \quad \dot{p}(t_0 + c_i h) = f(t_0 + c_i h, p(t + c_i h))$$

für $i = 1$ und $i = s$ kombiniert zu den zwei schwächeren Bedingungen

$$p(t_0) = x - h b_1 \Big(\dot{p}(t_0) - f(t_0), p(t_0) \Big)$$

und

$$\Phi(t_0, x, h) = p(t_0 + h) - h b_s \Big(\dot{p}(t + h) - f(t + h, p(t + h)) \Big).$$

Da damit nur noch $s - 1$ statt $s + 1$ Bedingungen an p gestellt werden, ist p nun aus $\mathcal{P}_{s-2}$.

Auch diese Klasse von Verfahren ist äquivalent zu s-stufigen impliziten Runge-Kutta-Verfahren und die Sätze 9.5 und 9.7 gelten weiterhin, allerdings wegen der geringeren Ordnung der Polynome mit $O(h^{s-1})$ bzw. für $s \leq p \leq 2s - 2$. Die geänderten Bedingungen wirken sich in den Beweisen in Abschätzung (9.6) aus, wo ein zusätzlicher Fehlerterm der Ordnung $O(h^{s-1})$ entsteht.

Zu den Methoden dieser Klasse gehören gewisse Radau und Lobatto-Verfahren, z.B. die **Lobatto IIIB-Verfahren**, bei denen $a_{i1} = b_1$ und $a_{is} = 0$ festgelegt wird und die restlichen Koeffizienten so gewählt werden, dass die Ordnung maximal, also $p = 2s - 2$ wird. Für $s = 3$ und $s = 4$ ergeben sich so die Tableaus

$$\begin{array}{c|ccc} 0 & \frac{1}{6} & -\frac{1}{6} & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 \\ 1 & \frac{1}{6} & \frac{5}{6} & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad \text{und} \qquad \begin{array}{c|cccc} 0 & \frac{1}{12} & \frac{-1-\sqrt{5}}{24} & \frac{2-1+\sqrt{5}}{24} & 0 \\ \frac{5-\sqrt{5}}{10} & \frac{1}{12} & \frac{25+\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & 0 \\ \frac{5+\sqrt{5}}{10} & \frac{1}{12} & \frac{25+13\sqrt{5}}{120} & \frac{25-\sqrt{5}}{120} & 0 \\ 1 & \frac{1}{12} & \frac{11-\sqrt{5}}{24} & \frac{11+\sqrt{5}}{24} & 0 \\ \hline & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} \end{array}$$

Kapitel 10

Partitionierte Runge–Kutta–Verfahren

Differentialgleichungen der allgemeinen Form $\dot{x}(t) = f(t, x(t))$ lassen sich durch Aufspalten des Zustandsvektors

$$x = \begin{pmatrix} y \\ z \end{pmatrix}$$

mit $x \in \mathbb{R}^n$, $y \in \mathbb{R}^{n_y}$, $z \in \mathbb{R}^{n_z}$ mit $n = n_y + n_z$ in zwei Teilgleichungen

$$\dot{y}(t) = g(t, y(t), z(t)), \qquad \dot{z}(t) = h(t, y(t), z(t)) \tag{10.1}$$

zerlegen. Wir nennen (10.1) auch *partitionierte Form* der Differentialgleichung.

Oft ergibt sich diese Partitionierung in kanonischer Weise, wie in den folgenden beiden Beispielen.

Beispiel 10.1 (a) **Hamilton'sche Systeme** Wie in Abschnitt B.2 hergeleitet, ist ein mechanisches System in Hamilton'scher Form beschrieben durch die Differentialgleichungen

$$\begin{aligned} \dot{q}(t) &= \frac{\partial H}{\partial p}(q(t), p(t), t) \\ \dot{p}(t) &= -\frac{\partial H}{\partial q}(q(t), p(t), t). \end{aligned}$$

Dies führt sofort auf die Zerlegung (10.1) mit $y = q$, $z = p$, $g = \frac{\partial H}{\partial p}$ und $h = -\frac{\partial H}{\partial q}$. Für die Pendelgleichung

$$\dot{q}(t) = \frac{p(t)}{m\rho^2}, \quad \dot{p}(t) = -mg\rho \sin q(t)$$

ergibt sich damit $g(t, y, z) = z/(m\rho^2)$ und $h(t, y, z) = -mg\rho \sin y$. Beachte, dass g hier nicht von y und h nicht von z abhängt, was bei mechanischen Systemen — z.B. auch in der Moleküldynamik — oft der Fall ist.

(b) **Gleichungen zweiter Ordnung** Das Umwandeln einer Gleichung zweiter Ordnung

$$\ddot{u}(t) = a(t, u(t), \dot{u}(t))$$

in eine Gleichung erster Ordnung wird normalerweise durch Einführung des erweiterten Zustands

$$x = \begin{pmatrix} u \\ \dot{u} \end{pmatrix}$$

erreicht. Alternativ können wir diese mit $y = u$, $z = \dot{u}$, $g(t, y, z) = z$ und $h = a$ in die Form (10.1) umwandeln. Beachte, dass die Abbildung h hier eine besonders einfache Form hat. Im Fall, dass a nicht von $\dot{u}$ abhängt, haben wir wieder den Fall, dass g nicht von y und h nicht von z abhängt. □

10.1 Definition

Definition 10.2 Gegeben seien zwei Runge-Kutta-Verfahren mit Koeffizienten (a_{ij}, b_j, c_i) bzw. $(\hat{a}_{ij}, \hat{b}_j, \hat{c}_i)$. Ein *partitioniertes Runge-Kutta-Verfahren* $\Phi(t, y, z, h)$ zur Lösung der partitionierten Differentialgleichung (10.1) ist gegeben durch die Gleichungen

$$\begin{aligned} k_i &= g\left(t + c_i h, y + h\textstyle\sum_{j=1}^s a_{ij} k_j, z + h\sum_{j=1}^s \hat{a}_{ij}\hat{k}_j\right),\ i = 1, \dots, s \\ \hat{k}_i &= h\left(t + c_i h, y + h\textstyle\sum_{j=1}^s a_{ij} k_j, z + h\sum_{j=1}^s \hat{a}_{ij}\hat{k}_j\right),\ i = 1, \dots, s \\ y_1 &= y + h\textstyle\sum_{i=1}^s b_i k_i, \qquad z_1 = z + h\sum_{i=1}^s \hat{b}_i \hat{k}_i \\ \Phi(t, y, z, h) &= (y_1, z_1) \end{aligned} \tag{10.2}$$

□

Zwei einfache Beispiele für ein solches Verfahren sind das *symplektische Euler-Verfahren*, gegeben durch die Tableaus

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

und das *Störmer/Verlet-Verfahren* gegeben durch

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

Das erste Verfahren kombiniert das implizite und das explizite Euler-Verfahren, das zweite die implizite Trapezregel mit der impliziten Mittelpunktregel.

Bemerkung 10.3 (i) Auch in dem Fall dass die beiden Verfahren explizit sind, stellen die $2s$ Gleichungen für $k_1, \dots, k_s, \hat{k}_1, \dots, \hat{k}_s$ ein nichtlineares Gleichungssystem dar, das i.A. numerisch (z.B. mit Fixpunktiteration oder dem Newton-Verfahren) gelöst werden muss. Unter geeigneten strukturellen Eigenschaften der zu lösenden Differentialgleichung lässt sich dies aber vermeiden, wie das Beispiel nach dieser Bemerkung zeigt.

(ii) Beachte, dass sich ein partitioniertes Runge-Kutta-Verfahren i.A. nicht als "normales" Runge-Kutta-Verfahren für $f = (g, h)$ schreiben lässt. In einem normalen Runge-Kutta-Verfahren wird jede Komponente der Lösung gleich behandelt, was hier nicht der Fall ist. Es handelt sich also um eine echt größere Klasse von Verfahren. □

Beispiel 10.4 Wir wenden das Störmer/Verlet-Verfahren auf die Pendelgleichung an, wobei wir zur Vereinfachung $\rho = 1$ und $m = 1$ setzen und dadurch $g(t, y, z) = z$ und $h(t, y, z) = -g \sin y$ erhalten. Die Gleichungen für die Stufen k_i und $\hat{k}_i$ lauten dann

$$\begin{aligned}
k_1 &= g(t, y, z + (h/2)\hat{k}_1) &&= z + (h/2)\hat{k}_1 \\
k_2 &= g(t + h, y + (h/2)k_1 + (h/2)k_2, z + (h/2)\hat{k}_1) &&= z + (h/2)\hat{k}_1 \\
\hat{k}_1 &= h(t, y, z + (h/2)\hat{k}_1) &&= -g \sin y \\
\hat{k}_2 &= h(t + h, y + (h/2)k_1 + (h/2)k_2, z + (h/2)\hat{k}_1) &&= -g \sin(y + (h/2)k_1 + (h/2)k_2)
\end{aligned}$$

Diese Gleichungen sind explizit lösbar, wenn man zuerst $\hat{k}_1$, dann k_1 und k_2 und am Schluss $\hat{k}_2$ berechnet. Wir haben dadurch ein explizites Verfahren erhalten, das aber nach wie vor die Vorteile der impliziten Bestandteile besitzt, insbesondere die Eignung für steife Differentialgleichungen. Die explizite Lösbarkeit der Gleichungen gilt beim Störmer/Verlet-Verfahren immer dann, wenn g nicht von y und h nicht von z abhängt. Dieses Verfahren wird daher z.B. in der Moleküldynamik oft eingesetzt. □

10.2 Konsistenz

Um Konsistenz und gegebenbenfalls eine gewisse Konsistenzordnung zu garantieren, muss sicherlich jedes der beiden definierenden Verfahren konsistent mit der gewünschten Ordnung sein. Um das zu sehen, reicht es, das Verfahren auf eine partitionierte Differentialgleichung mit $g \equiv 0$ oder $h \equiv 0$ anzuwenden. Man sieht zudem mit Lemma 2.6 leicht, dass für Konsistenz ohne höhere Ordnung die Bedingung $\sum b_i = \sum \hat{b}_i = 1$ hinreichend und notwendig ist. Mit Hilfe von Satz 3.7 sieht man zudem, dass diese Bedingung auch hinreichend und notwendig für die Konsistenzordnung $p = 1$ ist. Es reicht für die Konsistenzordnung $p = 1$ also, wenn die beiden definierenden Verfahren diese Ordnung haben. Dies ist z.B. beim symplektischen Euler-Verfahren der Fall.

Dies ist für höhere Konsistenzordnungen $p \geq 2$ nicht mehr der Fall. Hier ergeben sich aus Satz 3.7 durch die Kopplung der Verfahren über die üblichen Bedingungen an die Teilverfahren zusätzliche Kopplungsbedingungen.

Für $p = 2$ lauten diese

$$\sum_{i,j=1}^{s} b_i \hat{a}_{ij} = 1/2, \quad \sum_{i,j=1}^{s} \hat{b}_i a_{ij} = 1/2.$$

Diese Bedingung ist für autonomieinvariante Verfahren (also falls $\sum_j a_{ij} = c_i$) erfüllt, falls die einzelnen Verfahren Ordnung $p = 2$ haben und $c_i = \hat{c}_i$ für $i = 1, \ldots, s$; dies folgt aus Satz 4.5(ii). Sie kann aber auch erfüllt sein, wenn c_i und $\hat{c}_i$ nicht übereinstimmen, was z.B. beim Störmer/Verlet-Verfahren der Fall ist, welches folglich die Ordnung $p = 2$ besitzt.

Für $p = 3$ sind die allgemeinen Bedingungen bereits recht kompliziert. Falls $c_i = \hat{c}_i$ für alle $i = 1, \ldots, s$ vereinfachen sie sich aber zu

$$\sum_{i,j=1}^{s} b_i \hat{a}_{ij} c_j = 1/6, \qquad \sum_{i,j=1}^{s} \hat{b}_i a_{ij} c_j = 1/6.$$

Für $p \geq 4$ werden die Bedingungen schnell sehr kompliziert. Auch bei partitionierten Verfahren können wir diese aber mit Hilfe der Kollokation umgehen, wie das Beispiel im folgenden Abschnitt zeigt.

10.3 Beispiele

Lobatto IIIA/IIIB-Verfahren Diese Klasse von Verfahren verallgemeinert das Störmer/Verlet-Verfahren auf höhere Ordnungen. Die beiden Einzelverfahren wurden bereits in Kapitel 9 vorgestellt; hier wird als zusätzliche Bedingung bei der Herleitung noch die Gleichheit $c_j = \hat{c}_j$ und $b_i = \hat{b}_i$ verwendet. Für $s = 3$ erhalten wir so z.B.

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|ccc} 0 & \frac{1}{6} & -\frac{1}{6} & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 \\ 1 & \frac{1}{6} & \frac{5}{6} & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Es gilt der folgende Konsistenzsatz.

Satz 10.5 Die aus den Lobatto IIIA- und Lobatto IIIB-Verfahren zusammengesetzte partitionierte Runge-Kutta-Methode besitzt Konsistenzordnung $2s - 2$.

Beweisskizze: Aus der Konstruktion der beiden Bestandteile durch Kollokation bzw. unstetige Kollokation folgen für $\Phi(t_0, y_0, z_0, h) = (y_1, z_1)$ die Gleichungen

$$\begin{aligned} p(t_0) &= y_0 \\ q(t_0) &= z_0 - hb_1\Big(\dot{q}(t_0) - h(t_0, p(t_0), q(t_0))\Big) \\ \dot{p}(t_0 + c_i h) &= g(t_0 + c_i h, p(t_0 + c_i h), q(t_0 + c_i h)), \qquad i = 1, \ldots, s \\ \dot{q}(t_0 + c_i h) &= h(t_0 + c_i h, p(t_0 + c_i h), q(t_0 + c_i h)), \qquad i = 2, \ldots, s-1 \\ y_1 &= p(t_0 + h) \\ z_1 &= q(t_0 + h) - hb_s\Big(\dot{q}(t_0 + h) - h(p(t_0 + h), q(t_0 + h))\Big) \end{aligned}$$

mit $p \in \mathcal{P}_s$ und $q \in \mathcal{P}_{s-2}$. Ab hier verläuft der Beweis nun ähnlich wie der Beweis von Satz 9.7. □

Trotz der strukturellen Ähnlichkeit mit der Störmer/Verlet-Methode ist hier i.A. keine explizite Implementierung möglich, wenn g nicht von y und h nicht von z abhängt.

Nyström-Verfahren Diese Methoden eignen sich speziell für partitionierte Gleichungen der Form

$$\dot{y}(t) = z(t), \qquad \dot{z}(t) = h(t, y(t), z(t)) \tag{10.3}$$

vgl. Beispiel 10.1(b). Eine allgemeine partitionierte Runge-Kutta-Methode liefert angewendet auf diese Gleichung

$$\begin{aligned} k_i &= z + h\sum_{j=1}^{s} \hat{a}_{ij}\hat{k}_j,\ i = 1,\dots,s \\ \hat{k}_i &= h\left(t + c_i h, y + h\sum_{j=1}^{s} a_{ij}k_j, z + h\sum_{j=1}^{s} \hat{a}_{ij}\hat{k}_j\right),\ i = 1,\dots,s \\ y_1 &= y + h\sum_{i=1}^{s} b_i k_i, \qquad z_1 = z + h\sum_{i=1}^{s} \hat{b}_i\hat{k}_i \\ \Phi(t,y,z,h) &= (y_1, z_1) \end{aligned}$$

Setzen wir die Formeln für die k_i's in die anderen Gleichungen ein, so ergibt sich die folgende Definition mit

$$\bar{a}_{ij} = \sum_{k=1}^{s} a_{ik}\hat{a}_{kj}, \qquad \bar{b}_i = \sum_{k=1}^{s} b_k\hat{a}_{ki}.$$

Definition 10.6 Für reelle Koeffizienten c_i, $\bar{b}_i$, $\bar{a}_{ij}$, $\hat{b}_i$ und $\hat{a}_{ij}$, $i,j = 1,\dots,s$, ist das *Nyström-Verfahren* zur Lösung von (10.3) gegeben durch

$$\begin{aligned} \hat{k}_i &= h\left(t + c_i h, y + c_i h z + h^2\sum_{j=1}^{s} \bar{a}_{ij}\hat{k}_j, z + h\sum_{j=1}^{s} \hat{a}_{ij}\hat{k}_j\right), \quad i = 1,\dots,s \\ y_1 &= y + hz + h^2\sum_{i=1}^{s} \bar{b}_i\hat{k}_i, \qquad z_1 = z + h\sum_{i=1}^{s} \hat{b}_i\hat{k}_i \\ \Phi(t,y,z,h) &= (y_1, z_1) \end{aligned}$$

□

Im Spezialfall, dass h nicht von z abhängt, müssen die Koeffizienten $\hat{a}_{ij}$ nicht festgelegt werden. Wie auch bei allen anderen Runge-Kutta-Methoden gibt es Bedingungsgleichungen für höhere Konsistenzordnungen, vgl. [4, Abschnitt III.2.3]. Das Störmer/Verlet-Verfahren angewendet auf (10.3) ergibt gerade ein Nyström-Verfahren; die Koeffizienten ergeben sich aus den obigen Formeln.

Kapitel 11

Symplektische Runge-Kutta-Verfahren

Die im Anhang B.2 beschriebenen Hamilton'schen Systeme

$$\begin{aligned} \dot{q}(t) &= \frac{\partial H}{\partial p}(q(t), p(t), t) \\ \dot{p}(t) &= -\frac{\partial H}{\partial q}(q(t), p(t), t). \end{aligned} \tag{11.1}$$

besitzen eine wichtige Struktureigenschaft, die *Symplektizität*, die wir im Folgenden erläutern.

11.1 Symplektizität

Es sei

$$J = \begin{pmatrix} 0 & \mathrm{Id} \\ -\mathrm{Id} & 0 \end{pmatrix},$$

wobei $\mathrm{Id} \in \mathbb{R}^{n \times n}$ die d-dimensionale Einheitsmatrix bezeichnet. Beachte, dass für diese Matrix $J^{-1} = -J = J^T$ gilt.

Definition 11.1 Eine lineare Abbildung $A : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ heißt *symplektisch*, falls

$$A^T J A = J.$$

□

Im Fall $n = 1$ bedeutet dies, dass der Flächeninhalt eines von zwei Vektoren $x, y \in \mathbb{R}^2$ aufgespannten Parallelogramms P unter Transformation mit einer symplektischen linearen Abbildung A erhalten bleibt, denn es gilt

$$\mathrm{area}(P) = \det(x\ y) = x^T J y,$$

und damit

$$\mathrm{area}(AP) = \det(Ax\ Ay) = (Ax)^T J A y = x^T J y = \mathrm{area}(P).$$

Im Fall $n > 1$ gilt eine analoge geometrische Interpretation: schreiben wir $x = (x^q, x^p)^T \in \mathbb{R}^{2n}$ mit $x^{q,p} = (x_1^q, \ldots, x_n^q)^T$, $x^p = (x_1^p, \ldots, x_n^p)^T$ und ist $A \in \mathbb{R}^{2n\times 2n}$ symplektisch, dann erhält A die Größe

$$\omega(x,y) = \sum_{i=1}^{n} \det \begin{pmatrix} x_i^q & y_i^q \\ x_i^p & y_i^p \end{pmatrix} = x^T J y.$$

Symplektizität kann wie folgt auf nichtlineare Abbildungen verallgemeinert werden.

Definition 11.2 Eine (nichtlineare) Abbildung $g : U \to \mathbb{R}^{2n}$, $U \subset \mathbb{R}^{2n}$, heißt *symplektisch*, falls $Dg(x)$ für alle $x \in U$ symplektisch ist. □

Der folgende Satz zeigt, dass es eine charakteristische Eigenschaft der Lösungsabbildung eines Hamilton-Systems ist, symplektisch zu sein.

Satz 11.3 (Poincaré, 1899) Sei $H : \mathbb{R}^{2n} \to \mathbb{R}$ eine zweimal stetig differenzierbare Hamilton-Funktion. Dann ist die Lösung $x(t; t_0, x_0) = (q(t; t_0, q_0, p_0), p(t; t_0, q_0, p_0))$ des zugehörigen Hamilton-Systems (11.1) als Abbildung in $x_0 = (q_0, p_0)$ symplektisch für alle t, für die er definiert ist.

Beweis: Wir beobachten zunächst, dass wir das Hamilton-System (11.1) auch in der Form

$$\dot{x}(t) = J\nabla H(x(t), t)$$

mit $x = (q, p)$ und $\nabla H = (\frac{\partial H}{\partial q}, \frac{\partial H}{\partial p})$ schreiben können. Die Ableitung $A(t) = \frac{d}{dx_0} x(t; t_0, x_0)$ ist Lösung der zugehörigen Variationsgleichung

$$\dot{A}(t) = J\nabla^2 H(x(t; t_0, x_0), t)A(t) \quad \text{mit} \quad A(t_0) = \mathrm{Id}$$

(wobei $\nabla^2 H$ die Hesse-Matrix von H bzgl. x bezeichnet). Daher gilt, da $\nabla^2 H$ symmetrisch ist,

$$\begin{aligned} \frac{d}{dt}\left(A(t)^T J A(t)\right) &= \left(\frac{d}{dt}A(t)\right)^T JA(t) + A(t)^T J\left(\frac{d}{dt}A(t)\right) \\ &= (J\nabla^2 H A(t))^T JA(t) + A^T J(J\nabla^2 H A(t)) \\ &= A(t)^T \nabla^2 H \underbrace{J^T J}_{=\mathrm{Id}} A(t) + A(t)^T \underbrace{JJ}_{=-\mathrm{Id}} \nabla^2 H A(t) \\ &= 0. \end{aligned}$$

Also ist $t \mapsto A(t)$ konstant und es folgt $A(t)^T JA(t) = A(t_0)^T JA(t_0) = \mathrm{Id}J\mathrm{Id} = J$, d.h. $\frac{d}{dx_0} x(t; t_0, x_0) = A(t)$ ist für alle x_0 symplektisch und damit ist die Lösungsabbildung $x_0 \mapsto x(t; t_0, x_0)$ symplektisch (für alle relevanten t und t_0). □

Aus dem Beweis dieses Satzes folgt insbesondere die (hier zur Vereinfachung ohne Argumente geschriebene) Gleichung

$$(J\nabla^2 HA)^T JA + A^T J(J\nabla^2 HA) = 0, \tag{11.2}$$

die für die nachfolgenden numerischen Untersuchungen nützlich sein wird.

11.2 Veranschaulichung an der Pendelgleichung

Inzwischen haben wir ein recht weitgehendes Verständnis des Pendelmodells aus der Einleitung erreicht, das in der folgenden Abbildung 11.1 noch einmal zusammengefasst ist:

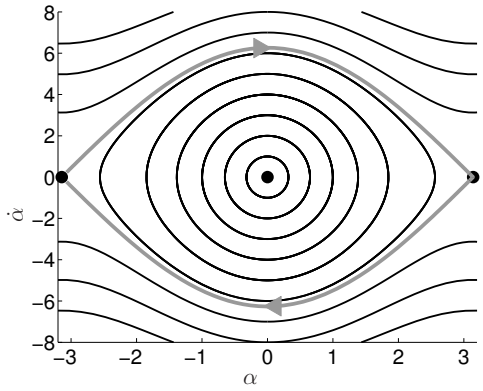

Abbildung 11.1: Überblick über qualitativ verschiedene Lösungen im Pendelmodell: Gleichgewichte, periodische Lösungen, homokline Orbits

Die Anwendung des expliziten Euler-Verfahrens zum Anfangswert $x_0 = (\alpha_0, \dot{\alpha}_0) = (0, 4)$ mit Schrittweite $h = 0.1$ ist in Abbildung 11.2 (links) graphisch dargestellt.

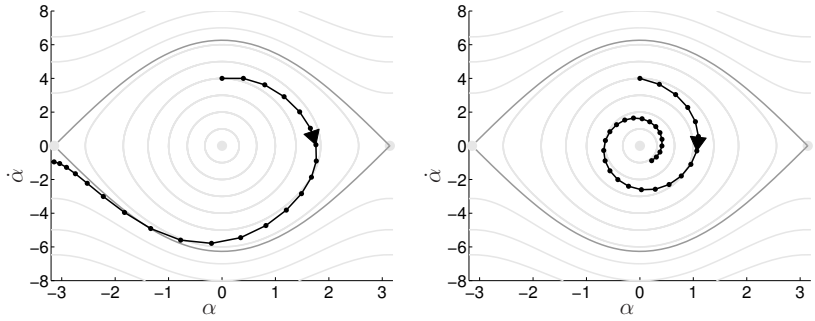

Abbildung 11.2: Numerische Integration des Pendelmodells zum Anfangswert $(0, 4)$ und Schrittweite $h = 0.1$ mit dem expliziten (links) und dem impliziten (rechts) Euler-Verfahren

Offensichtlich gibt diese Approximation das Verhalten der wahren Lösungen qualitativ falsch wieder: das Resultat der Rechnung ist nicht periodisch, die Amplitude der Schwingung steigt kontinuierlich an. Auch implizite Verfahren helfen in dieser Hinsicht nicht weiter, wie Abbildung 11.2 (rechts) zeigt, wo die numerische Lösung des impliziten Euler-Verfahrens zu denselben Daten dargestellt ist. Der Einsatz eines Verfahrens höherer Ordnung würde zwar den Approximationsfehler verkleinern, aber qualitativ an Verhalten der numerischen Lösung nichts ändern – es sein denn, man erwischt ein spezielles Verfahren:

Abbildung 11.3 zeigt das Ergebnis für die *implizite Mittelpunktregel*

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_i) + hf\left(\frac{\tilde{x}(t_i) + \tilde{x}(t_{i+1})}{2}\right). \tag{11.3}$$

Dieses Verfahren schneidet offenbar deutlich besser ab, die numerische Lösung scheint auf einem periodischen Orbit zu verbleiben. Tatsächlich ist die durch (11.3) definierte Abbil-

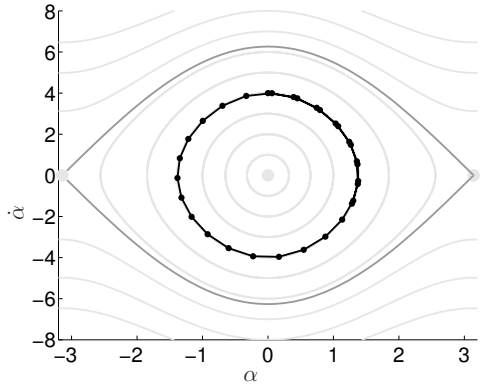

Abbildung 11.3: Numerische Integration des Pendelmodells zum Anfangswert $(0, 4)$ und Schrittweite $h = 0.1$ mit der impliziten Mittelpunktsregel

dung symplektisch (wenn f ein Hamilton-System ist) – und erbt damit die charakteristische Eigenschaft des Flusses eines Hamilton-Systems wie dem Pendel.

Tatsächlich folgt daraus nicht sofort, dass die numerische Lösung periodisch ist. Periodizität entspricht der Energieerhaltung, die für Hamilton'sche Systeme ebenfalls gilt während Symplektizität der Flächenerhaltung entspricht. Man kann aber zeigen, dass für symplektische Runge-Kutta-Verfahren eine Funktion $\widetilde{H} \approx H$ existiert, die entlang der numerischen Lösung konstant ist, siehe [4, Abschnitt IX.3]. Das bedeutet, dass die exakte Hamilton-Funktion entlang der Lösung "fast konstant" ist. Insbesondere kan H entlang der numerischen Lösung nicht mit der Zeit immer mehr anwachsen oder abfallen, weswegen die Energie bis auf kleine Schwankungen erhalten wird.

Die Erhaltung der Symplektizität ist also eine wünschenswerte Eigenschaft.

11.3 Symplektische Runge-Kutta-Verfahren

Definition 11.4 Ein Einschrittverfahren $\Phi(t, x, h)$ heißt *symplektisch*, falls die Abbildung $x \mapsto \Phi(t, x, h)$ für alle $h > 0$ symplektisch ist, wenn das Einschrittverfahren auf ein Hamilton-System angewendet wird. □

Im Fall von Runge-Kutta-Verfahren existiert ein einfaches Kriterium an die Koeffizienten des Verfahrens, das die Symplektizität sicherstellt.

Satz 11.5 Gilt für die Koeffizienten eines Runge-Kutta-Verfahrens

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \text{für alle } i, j = 1, \ldots, s, \tag{11.4}$$

dann ist das Verfahren symplektisch.

Zum Beweis dieses Satzes benötigen wir zwei vorbereitende Lemmata.

Lemma 11.6 Gegeben sei eine gewöhnliche Differentialgleichung $\dot{x}(t) = f(t, x(t))$ und die zugehörige Variationsgleichung $\dot{A}(t) = \frac{df}{dx}(t, x(t))A(t)$. Es seien $\tilde{x}(t_i)$, $\widetilde{A}(t_i)$ die durch ein Runge-Kutta- oder ein partitioniertes Runge-Kutta-Verfahren berechneten numerischen Lösungen zu den Anfangsbedingungen $\tilde{x}(t_0) = x_0$, $\widetilde{A}(t_0) = \mathrm{Id}$. Dann gilt für alle $t_i \in \mathcal{T}$

$$\widetilde{A}(t_i) = \frac{\partial}{\partial x_0}\widetilde{x}(t_i).$$

Beweis: Wir führen den Beweis für ein nicht-partitioniertes Runge-Kutta-Verfahren; der partitionierte Fall folgt analog mit etwas mehr Aufwand bei der Indexverwaltung.

Fassen wir die Stufen k_j im i-ten Schritt des Verfahrens als Funktionen in x_0 auf und leiten diese nach x_0 ab, so folgt mit der Kettenregel

$$\frac{\partial}{\partial x_0}k_j = \frac{df}{dx}\left(t_i + c_j h_i, \tilde{x}(t_i) + h_i \sum_{l=1}^{s} a_{jl}k_l\right)\left(\frac{\partial}{\partial x_0}\tilde{x}(t_i) + h_i \sum_{l=1}^{s} a_{jl}\frac{\partial}{\partial x_0}k_l\right).$$

Zudem gilt

$$\frac{\partial}{\partial x_0}\tilde{x}(t_{i+1}) = \frac{\partial}{\partial x_0}\tilde{x}(t_i) + h_i \sum b_j \frac{\partial}{\partial x_0}k_j.$$

Andererseits gilt für die Stufen $\hat{k}_j$ des Verfahrens angewendet auf die Variationsgleichung

$$\hat{k}_j = \frac{df}{dx}\left(t_i + c_j h_i, \tilde{x}(t_i) + h_i \sum_{l=1}^{s} a_{jl}k_l\right)\left(\widetilde{A}(t_i) + h_i \sum_{l=1}^{s} a_{jl}\hat{k}_l\right).$$

und es gilt

$$\widetilde{A}(t_{i+1}) = \widetilde{A}(t_i) + h_i \sum b_j \hat{k}_j.$$

Hieraus folgt $\hat{k}_j = \frac{\partial}{\partial x_0}k_j$ und damit die behauptete Identität. □

Lemma 11.7 Betrachte $\dot{x}(t) = f(t, x(t))$ und ein Runge-Kutta-Verfahren Φ, das die Bedingungen von Satz 11.5 erfüllt. Es sei $Q \in \mathbb{R}^{n\times n}$ eine Matrix, für die

$$x^T Q f(t, x) + f(t, x)^T Q x = 0 \tag{11.5}$$

gilt für $t \in \mathbb{R}$ und $x \in \mathbb{R}^n$, für die f definiert ist. Dann gilt

$$\Phi(t, x, h)^T Q \Phi(t, x, h) = x^T Q x \tag{11.6}$$

für alle $t \in \mathbb{R}$, $x \in \mathbb{R}^n$ und alle $h > 0$, für die Φ definiert ist.

Beweis: Für $x_1 = \Phi(t, x, h)$ gilt

$$x_1^T Q x_1 = x^T Q x + h \sum_{i=1}^{s} b_i k_i^T Q x + h \sum_{j=1}^{s} b_j x^T Q k_j + h^2 \sum_{i,j=1}^{s} b_i b_j k_i^T Q k_j.$$

Schreiben wir $k_i = f(t_i, X_i)$ mit $X_i = x + h\sum_{j=1}^{s} a_{ij}k_j$, setzen dies in die obige Gleichung ein und stellen die Terme um, so folgt

$$x_1^T Q x_1 = x^T Q x + h\sum_{i=1}^{s} b_i(Y_i^T Q f(t,Y_i) + f(t,Y_i)^T Q Y_i) + h^2 \sum_{i,j=1}^{s} (b_i b_j - b_i a_{ij} - b_j a_{ji}) k_i^T C k_j.$$

Aus der Bedingung an Q folgt, dass die erste Summe gleich Null ist und die Bedingung an die Koeffizienten des Schemas bewirkt, dass die zweite Summe gleich Null ist. Damit folgt die Behauptung. □

Beweis von Satz 11.5: Aus (11.2) folgt, dass

$$Q = \begin{pmatrix} 0 & 0 \\ 0 & J \end{pmatrix}$$

die Bedingung (11.5) für die Gleichung

$$\begin{aligned} \dot{x}(t) &= f(t, x(t)) \\ \dot{A}(t) &= J\nabla^2 H(t, x(t)) A(t) \end{aligned}$$

erfüllt. Wenden wir nun ein Runge-Kutta-Verfahren, das die Bedingungen von Satz 11.5 erfüllt, auf diese Gleichung an, so folgt aus Lemma 11.6 und Lemma 11.7[1]

$$\begin{aligned} \frac{d}{dx_0}\Phi(t,x,h)^T J \frac{d}{dx_0}\Phi(t,x,h) &= \begin{pmatrix} \Phi(t,x,h) \\ \frac{d}{dx_0}\Phi(t,x,h) \end{pmatrix}^T Q \begin{pmatrix} \Phi(t,x,h) \\ \frac{d}{dx_0}\Phi(t,x,h) \end{pmatrix} \\ &= \begin{pmatrix} x \\ \mathrm{Id} \end{pmatrix}^T Q \begin{pmatrix} x \\ \mathrm{Id} \end{pmatrix} = J. \end{aligned}$$

Dies ist gerade die Bedingung für die Symplektizität von Φ. □

Beispiel 11.8 Das Butcher-Schema der impliziten Mittelpunktsregel ist

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

Hier gilt also $b_1 = 1$ und $a_{11} = 1/2$, so dass (11.4) erfüllt ist. □

Leider gibt es neben der impliziten Mittelpunktregel nicht sehr viele Verfahren, die die Bedingung aus Satz 11.5 erfüllen. Weitere Verfahren kann man allerdings erhalten, wenn man partitionierte Runge-Kutta-Schemata betrachtet. Hier stellt man zwar zunächst fest, dass sich Lemma 11.7 nicht übertragen lässt. Betrachtet man statt (11.6) aber die Gleichung

$$y_1^T Q z_1 = y^T Q z \tag{11.7}$$

[1]Formal müsste man zur Anwendung von Lemma 11.7 diese Matrixgleichung als vektorwertige Gleichung umschreiben, was durch Ersetzen von Q durch eine passende Matrix höherer Dimension erreicht wird, auf deren explizite Berechnung wir hier verzichten.

mit $(y_1, z_1) = \Phi(t, y, z, h)$, so ist diese erfüllt, falls die Koeffizienten der beiden Verfahren die Bedingungen

$$b_i \hat{a}_{ij} + \hat{b}_j a_{ij} = b_i \hat{b}_i \quad \text{und} \quad b_i = \hat{b}_i \quad \text{für } i, j = 1, \ldots, s$$

erfüllen (der Beweis ist ähnlich wie der von Lemma 11.7). Weiterhin sieht man ähnlich wie im Beweis von Satz 11.5, dass (11.7) ausreicht, um Symplektizität des partitionierten Verfahrens zu beweisen. Dass dies gilt, sieht man, wenn man die Matrix A analog zu $x = (q, p)$ in A_q und A_p zerlegt. Die zu überprüfende Gleichung $J = A^T J A = A_q^T \mathrm{Id} A_p - A_p^T \mathrm{Id} A_q$ besteht dann aus (matrixwertigen) Termen der Form (11.7), weswegen es genügt, diese zu betrachten. Für Details siehe [4, Abschnitte IV.2 und VI.4].

Mit dem obigen Kriterium kann man dann überprüfen, dass jedes Lobatto IIIA/IIIB-Verfahren und damit insbesondere das Störmer/Verlet-Verfahren symplektisch ist.

Kapitel 12

Mehrschrittverfahren

Die Mehrschrittverfahren unterscheiden sich von den Einschrittverfahren dadurch, dass der Wert $\tilde{x}(t_{i+1})$ nicht nur von $\tilde{x}(t_i)$ sondern von einer ganzen Reihe von Vorgängerwerten $\tilde{x}(t_{i-k+1}), \ldots, \tilde{x}(t_i)$ abhängt. Wie schon bei den Einschrittverfahren gibt es explizite und implizite Mehrschrittverfahren; erstere geben einen expliziten Ausdruck für $\tilde{x}(t_{i+1})$, während bei letzteren noch eine Fixpunktgleichung zu lösen ist. Man hofft dabei, dass man – da ja durch die größere Anzahl von Punkten mehr Information zur Verfügung steht – im Vergleich zu Einschrittverfahren gleicher Konsistenzordnung mit weniger Auswertungen von f pro Schritt auskommt. Tatsächlich werden wir sehen, dass diese Hoffnung berechtigt ist.

Zur Motivation betrachten wir wieder Verfahren, die wir heuristisch aus numerischen Integrationsformeln ableiten. Wir nehmen dabei konstante Schrittweite $h_i = h$ an. Wenn wir in der Integralgleichung

$$x(t_{i+1}) = x(t_{i-1}) + \int_{t_{i-1}}^{t_{i+1}} f(t, x(t))dt$$

das Integral durch die Mittelpunktregel

$$\int_{t_{i-1}}^{t_{i+1}} f(t, x(t))dt \approx 2hf(t_i, x(t_i))$$

ersetzen, so erhalten wir die im letzten Kapitel bereits betrachtete *explizite Mittelpunktregel*

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_{i-1}) + 2hf(t_i, \tilde{x}(t_i)).$$

Wählen wir die Simpson–Regel

$$\int_{t_{i-1}}^{t_{i+1}} f(t, x(t))dt \approx \frac{h}{3}\Big(f(t_{i+1}, x(t_{i+1})) + 4f(t_i, x(t_i)) + f(t_{i-1}, x(t_{i-1}))\Big),$$

so erhalten wir das (implizite) *Milne–Simpson–Verfahren*

$$\tilde{x}(t_{i+1}) = \tilde{x}(t_{i-1}) + \frac{h}{3}(f(t_{i+1}, \tilde{x}(t_{i+1})) + 4f(t_i, \tilde{x}(t_i)) + f(t_{i-1}, \tilde{x}(t_{i-1}))).$$

Eine Verallgemeinerung, die diese beiden Verfahren umfasst, ist die folgende Klasse der *linearen Mehrschrittverfahren (MSV).*

Definition 12.1 Ein *k–stufiges lineares Mehrschrittverfahren (MSV)* ist gegeben durch die Gleichung

$$\begin{aligned} & a_k\tilde{x}(t_{i+k}) + a_{k-1}\tilde{x}(t_{i+k-1}) + \ldots + a_0\tilde{x}(t_i) \\ & \quad = \; h\Big(b_k\tilde{f}(t_{i+k}) + b_{k-1}\tilde{f}(t_{i+k-1}) + \ldots + b_0\tilde{f}(t_i)\Big) \end{aligned} \tag{12.1}$$

mit der Abkürzung $\tilde{f}(t_j) = f(t_j, \tilde{x}(t_j))$, wobei $a_k \neq 0$ ist □

Mit dieser Klasse von Verfahren wollen wir uns schwerpunktmäßig beschäftigen. Wenn $b_k = 0$ ist, so ist das Verfahren explizit, da es direkt nach $\tilde{x}(t_{i+k})$ aufgelöst werden kann. Falls $b_k \neq 0$ ist, so kann man die entstehenden Gleichungen analog zu den impliziten Einschrittverfahren lösen (algebraisch, Fixpunkt–Iteration, Newton–Verfahren,...). Wir beschränken uns zunächst auf den Fall äquidistanter Schrittweiten $h_i = h$ und gehen am Schluss dieses Kapitels (kurz) auf variable Schrittweiten und Schrittweitensteuerung ein.

Bemerkung 12.2 (i) Zum Start eines Mehrschrittverfahrens benötigt man neben dem Anfangswert $\tilde{x}(t_0)$ noch die Werte $\tilde{x}(t_1), \ldots, \tilde{x}(t_{k-1})$. Diese werden üblicherweise durch ein geeignetes Einschrittverfahren bestimmt. Details dazu besprechen wir etwas später.
(ii) Wenn man die $\tilde{f}$–Werte eines Schrittes zwischenspeichert, so muss in jedem Schritt lediglich der Wert $\tilde{f}(t_{i+k-1})$ neu berechnet werden. Ein explizites lineares MSV kommt also mit einer f–Auswertung pro Schritt aus. □

Zur Analyse von MSV hat sich der folgende (aus der Theorie der dynamischen Systeme stammende) Formalismus als sehr geeignet erwiesen.

Definition 12.3 Auf dem Raum der Gitterfunktionen $\Delta_{\mathcal{T}} := \{f : \mathcal{T} \to \mathbb{R}^n\}$ definieren wir den *Shift–Operator* $E : \Delta_{\mathcal{T}} \to \Delta_{\mathcal{T}}$ mittels

$$E(f)(t_i) = f(t_{i+1}).$$

Hierbei erweitern wir unser Gitter formal zu einem Gitter mit unendlich vielen Gitterpunkten $\mathcal{T} = \{t_0, t_1, t_2, \ldots\}$. □

Beispiel 12.4 Für eine Gitterfunktion mit $f(t_i) = a_i$ mit $a_i = (2, 4, 8, 16, 32, \ldots)$ gilt also $E(f) = \tilde{f}$ mit $\tilde{f}(t_i) = \tilde{a}_i$ mit $\tilde{a}_i = (4, 8, 16, 32, 64, \ldots)$. Die Wertefolge wird also um eine Stelle nach links verschoben, woraus sich der Name Shift–Operator (manchmal auch 'Linksshift' genannt) ergibt. □

Der Shift–Operator erlaubt die folgende, sehr kompakte Schreibweise von Mehrschrittverfahren: Mit den Polynomen

$$\begin{aligned} P_a(z) &= a_0 + a_1 z + \ldots + a_k z^k \\ P_b(z) &= b_0 + b_1 z + \ldots + b_k z^k \end{aligned}$$

kann man (12.1) als

$$P_a(E)(\tilde{x})(t_i) = hP_b(E)(\tilde{f})(t_i) \tag{12.2}$$

schreiben, wobei die Potenz E^j des Shift–Operators die j–malige Hintereinanderausführung des Operators bedeutet.

Wir wollen nun die Konvergenz von Mehrschrittverfahren untersuchen und dabei das für die Einschrittverfahren bewiesene Resultat "Konsistenz + Lipschitzbedingung ⇒ Konvergenz" verallgemeinern. Wir beginnen mit der Konsistenz.

12.1 Konsistenz

Bei der Untersuchung der Konsistenz bei Einschrittverfahren haben wir mittels

$$\varepsilon := \|\Phi(t, x, h) - x(t + h; t, x)\|$$

den Konsistenzfehler durch Vergleich des numerischen Verfahrens mit der exakten Lösung erhalten. Die Größe ε lässt sich aber auch anders interpretieren:
Für die numerisch berechnete Gitterfunktion gilt gerade die Gleichung

$$0 = \|\tilde{x}(t_{i+1}) - \Phi(t_i, \tilde{x}(t_i), h)\|$$

Setzen wir hier nun die exakte Lösungsfunktion $x(t) = x(t; t_0, x_0)$ ein, so erhalten wir

$$\|x(t_{i+1}) - \Phi(t_i, x(t_i), h)\| = \varepsilon,$$

also gerade wieder unseren Konsistenzfehler für $x = x(t_i)$. Beachte, dass jede Funktion $x : [t_0, T] \to \mathbb{R}^n$ auch eine Gitterfunktion auf den in $[t_0, T]$ liegenden Gitterpunkten ist.

Dieses Verfahren "Einsetzen der exakten Lösung in die numerische Gleichung" lässt sich auf viele numerische Verfahren anwenden, z.B. auf unsere Mehrschrittverfahren. In der kompakten Schreibweise (12.2) müssen wir also die Norm des Konsistenzfehlers

$$L(x, t, h) = P_a(E)(x)(t) - hP_b(E)(f)(t) = P_a(E)(x)(t) - hP_b(E)(\dot{x})(t)$$

bestimmen. Beachte, dass der Parameter x hier eine Funktion $x : [t_0, T] \to \mathbb{R}^n$ und dass L nur für solche Parametertripel (x, t, h) definiert ist, für die $[t, t + hk] \subset [t_0, T]$ gilt.

Definition 12.5 Ein lineares Mehrschrittverfahren besitzt die *Konsistenzordnung* p, falls für jede $p + 1$–mal stetig differenzierbare Lösung $x : [t_0, T] \to \mathbb{R}^n$ der Differentialgleichung (1.1) die Abschätzung

$$L(x, t, h) = O(h^{p+1})$$

gleichmäßig in t gilt für alle t, h, in denen $L(x, t, h)$ definiert ist. □

Interessanterweise hängt die Definition des Konsistenzfehlers L *nicht* von f ab, da wir die auftretenden Werte des Vektorfeldes f durch die Ableitungen $\dot{x}$ ersetzt haben. Dies nutzt der folgende Satz aus, der Bedingungen angibt, anhand derer man die Konsistenzordnung eines Mehrschrittverfahrens überprüfen kann.

Satz 12.6 Ein lineares Mehrschrittverfahren besitzt genau dann die Konsistenzordnung $p \in \mathbb{N}$, wenn eine der folgenden äquivalenten Bedingungen erfüllt ist.

(i) Für jede beliebige $p+1$–mal stetig differenzierbare Funktion $x : [t_0, T] \to \mathbb{R}^n$ gilt die Abschätzung

$$L(x, t, h) = O(h^{p+1})$$

gleichmäßig in t für alle t, h, in denen $L(x, t, h)$ definiert ist.

(ii) $L(Q, 0, h) = 0$ für alle Polynome $Q \in \mathcal{P}_p$.

(iii) Es gilt

$$\sum_{j=0}^{k} a_j = 0, \ \sum_{j=0}^{k} a_j j^l = l \sum_{j=0}^{k} b_j j^{l-1} \text{ für } l = 1, \ldots, p$$

mit der Konvention $0^0 = 1$.

Beweis: Wir zeigen die Äquivalenz durch die Implikationen

$$\text{(i)} \Rightarrow \text{Konsistenzordnung } p \Rightarrow \text{(ii)} \Rightarrow \text{(i)} \Rightarrow \text{(iii)} \Rightarrow \text{(i)}$$

"(i) $\Rightarrow$ Konsistenzordnung p": Dies folgt direkt, da mit jeder beliebigen Funktion auch jede Lösung die behauptete Abschätzung erfüllt.

"Konsistenzordnung $p \Rightarrow$ (ii)": Gegeben sei ein beliebiges Polynom $Q \in \mathcal{P}_p$. Mit $f(t, x) = \dot{Q}(t)$ erhalten wir eine "triviale" Differentialgleichung, deren Lösung Q ist. Nach Definition der Konsistenzordnung folgt also

$$L(Q, 0, h) = O(h^{p+1}).$$

Da Q ein Polynom vom Grad $\leq p$ ist, muss auch $L(Q, 0, h)$ ein Polynom vom Grad $\leq p$ in h sein, weswegen $L(Q, 0, h) = 0$ sein muss.

"(ii) $\Rightarrow$ (i)": Sei x eine beliebige $p+1$–mal differenzierbare Funktion und sei Sei $Q \in \mathcal{P}_p$ das Polynom, das durch die ersten p Terme der Taylorentwicklung von x in t^* definiert ist. Dann gilt

$$x(t) = Q(t) + O(h^{p+1}) \text{ für alle } t \in [t^* - h, t^* + h].$$

Aus der Struktur von L folgt damit sofort die Abschätzung

$$L(x, t, h) = L(Q, t, h) + O(h^{p+1}).$$

Diese Abschätzung ist gleichmäßig in $t \in [t_0, T]$, da das den $O(h^{p+1})$–Term bestimmende Taylor–Restglied gleichmäßig beschränkt auf kompakten Intervallen ist. Aus (ii) wissen wir, dass $L(Q, 0, h) = 0$ gilt, woraus (durch "Verschieben" des Polynoms) auch $L(Q, t, h) = 0$ folgt, was schließlich die Behauptung liefert.

"(i) $\Rightarrow$ (iii)": Die Implikation aus (i) gilt insbesondere für konstante Funktionen $x \equiv c$. Für diese gilt

$$O(h^{p+1}) = L(x, 0, h) = P_a(E)x(t) - \underbrace{hP_b(E)\dot{x}(t)}_{=0} = \sum_{j=0}^{k} a_j c.$$

Da die rechte Seite unabhängig von h ist, kann dies nur gelten, wenn die Summe der a_j gleich Null ist, was die erste Gleichung in (iii) zeigt.

Für die weiteren Gleichungen in (iii) betrachten wir (i) mit $x(t) = \exp(t)$. Wegen

$$E^j(\exp)(0) = \exp(jh) = \exp(h)^j \quad \text{und} \quad \frac{d}{dt}\exp(t) = \exp(t)$$

folgt

$$L(\exp, 0, h) = P_a(\exp(h)) - hP_b(\exp(h))$$

Wir betrachten die Taylorentwicklung dieses Ausdrucks in $h = 0$. Diese lautet

$$L(\exp, 0, h) = \sum_{l=0}^{p} \frac{1}{l!} \sum_{j=0}^{k} a_j j^l h^l - \sum_{l=0}^{p-1} \frac{1}{l!} \sum_{j=0}^{k} b_j j^l h^{l+1} + O(h^{p+1}).$$

Aus (i) wissen wir $L(\exp, 0, h) = O(h^{p+1})$, weswegen

$$\sum_{l=0}^{p} \frac{1}{l!} \sum_{j=0}^{k} a_j j^l h^l - \sum_{l=0}^{p-1} \frac{1}{l!} \sum_{j=0}^{k} b_j j^l h^{l+1} = O(h^{p+1})$$

sein muss. Dieser Summenausdruck ist ein Polynom vom Grad $\leq p$ in h, und kann daher nur von der Ordnung $O(h^{p+1})$ sein, wenn er bereits Null ist. Dies wiederum kann nur dann gelten, wenn sich die Koeffizienten zu gleichen Potenzen von h zu Null addieren, also

$$\frac{1}{l!} \sum_{j=0}^{k} a_j j^l - \frac{1}{(l-1)!} \sum_{j=0}^{k} b_j j^{l-1} = 0$$

gilt. Dies sind gerade die weiteren Gleichungen aus (iii).

"(iii) $\Rightarrow$ (i)": Die Taylorentwicklung von L für allgemeine x in $h = 0$ lautet

$$\begin{aligned} L(x, t, h) &= \sum_{l=0}^{p} \frac{1}{l!} \sum_{j=0}^{k} a_j j^l h^l x^{(l)}(t) \\ &\quad - h \left(\sum_{l=0}^{p-1} \frac{1}{l!} \sum_{j=0}^{k} b_j j^l h^l x^{(l+1)}(t) \right) + O(h^{p+1}). \end{aligned}$$

Wenn die Gleichungen aus (iii) gelten, so fallen alle diese Summanden weg, so dass nur $O(h^{p+1})$ übrig bleibt. Diese Abschätzung ist wegen der gleichmäßigen Beschränktheit des Taylor–Restgliedes gleichmäßig in $t \in [t_0, T]$, weswegen (i) folgt. □

Bemerkung 12.7 Der Fall $p = 1$ ist hierbei besonders interessant, da er die Frage beantwortet, wann ein Verfahren überhaupt konsistent ist. Für $p = 1$ erhalten wir aus (iii) die Bedingungen

$$\sum_{j=0}^{k} a_j = 0 \quad \text{und} \quad \sum_{j=0}^{k} a_j j = \sum_{j=0}^{k} b_j.$$

Beide Bedingungen lassen sich mit Hilfe der Polynome P_a und P_b ausdrücken, sie sind gerade äquivalent zu

$$P_a(1) = 0 \quad \text{und} \quad P_a'(1) = P_b(1).$$

Diese Bedingungen entsprechen der Bedingung $\sum b_i = 1$ bei den Runge–Kutta–Verfahren. Insbesondere muss für konsistente Verfahren die 1 eine Nullstelle von P_a sein. Wir werden im nächsten Teilabschnitt sehen, dass auch die weiteren Nullstellen von P_a eine wichtige Rolle bei der Konvergenzanalyse von Mehrschrittverfahren spielen. □

12.2 Stabilität

Wir wollen nun ein geeignetes Analogon der Lipschitzbedingung für Einschrittverfahren entwickeln. In der Konvergenztheorie der Einschrittverfahren haben wir diese Bedingung verwendet, um sicher zu stellen, dass sich die in vergangenen Schritten gemachten Fehler im aktuellen Schritt nicht zu sehr verstärken.

Sicherlich sollte die rechte Seite unseres Mehrschrittverfahrens (12.1) eine ähnliche Lipschitzbedingung erfüllen, diese erhalten wir aber "geschenkt", da wir ja nur Lipschitz–stetige Vektorfelder f betrachten. Leider reicht es aber nicht aus, wenn f Lipschitz–stetig ist. Diese Bedingung besagt ja nur, dass sich kleine Fehler in den vergangenen $\tilde{x}$ in der *rechten* Seite unseres Verfahrens wenig auswirken. Wir benötigen zusätzlich noch eine Bedingung, die uns garantiert, dass kleine Fehler auf der *linken* Seite von (12.1) auch nur kleine Fehler in $\tilde{x}(t_{i+k})$ hervorrufen.

Um zu sehen, dass dies ein nichttriviales Problem ist, betrachten wir zwei Mehrschrittverfahren, die wir auf das Anfangswertproblem

$$\dot{x}(t) = 0, \qquad x(0) = 0 \tag{12.3}$$

anwenden. Da die rechte Seite in (12.1) wegen $f \equiv 0$ verschwindet, reicht es, die Koeffizienten a_i anzugeben. Wir betrachten nun die Verfahren mit

$$a_2 = 1,\, a_1 = -3,\, a_0 = 2 \quad \text{und} \quad \tilde{a}_2 = 1,\, \tilde{a}_1 = -3/2,\, \tilde{a}_0 = 1/2. \tag{12.4}$$

Man sieht leicht, dass beide Verfahren wegen $\sum a_i = 0$ bzw. $\sum \tilde{a}_i = 0$ konsistent sind. Für die DGL (12.3) ergeben sich daraus die Iterationsvorschriften

$$\tilde{x}(t_{i+1}) = -a_1\tilde{x}(t_i) - a_0\tilde{x}(t_{i-1}) = 3\tilde{x}(t_i) - 2\tilde{x}(t_{i-1}) \tag{12.5}$$

und

$$\tilde{x}(t_{i+1}) = -\tilde{a}_1\tilde{x}(t_i) - \tilde{a}_0\tilde{x}(t_{i-1}) = 3/2\tilde{x}(t_i) - 1/2\tilde{x}(t_{i-1}). \tag{12.6}$$

Man sieht leicht, dass beide Verfahren für exakte Startwerte $\tilde{x}(t_0) = \tilde{x}(t_1) = 0$ die exakte Lösung $\tilde{x}(t_i) \equiv 0$ liefern. Wenn wir den Startwert $\tilde{x}(t_1)$ allerdings leicht stören, so unterscheidet sich das Verhalten der beiden Verfahren erheblich. Abbildung 12.1 zeigt das unterschiedliche Verhalten für $\tilde{x}(t_0) = 0$ und den (nur ganz leicht gestörten Wert) $\tilde{x}(t_1) = 10^{-12}$.

Offenbar reproduziert das zweite Verfahren die exakte konstante Lösung trotz der kleinen Störung in $\tilde{x}(t_1)$ gut, während das erste Verfahren nach nur etwa 35 Schritten riesige Fehler produziert.

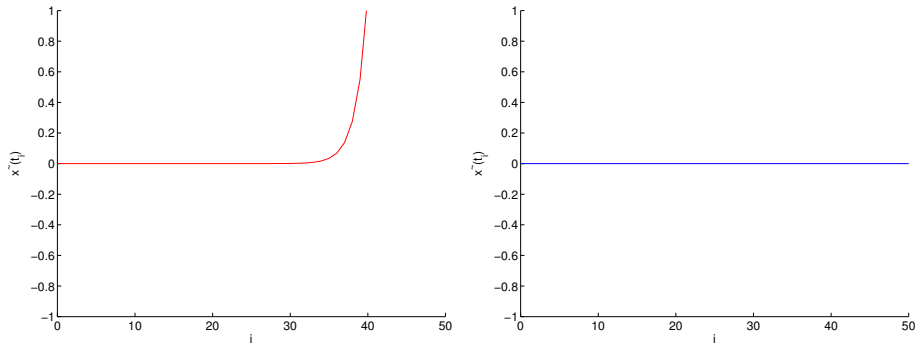

Abbildung 12.1: MSV (12.5) (links) und (12.6) (rechts) mit $\tilde{x}(t_0) = 0$, $\tilde{x}(t_1) = 10^{-12}$

Wir wollen nun untersuchen, warum dies so ist und wie man erkennen kann, ob ein Mehrschrittverfahren stabil gegenüber solchen kleinen Fehlern ist. Wegen der Linearität der linken Seite des Verfahrens genügt es, dazu das einfache Anfangswertproblem (12.3) zu betrachten (später im Beweis der Konvergenz werden wir genauer sehen, warum). Aus (12.1) folgt sofort, dass für (12.3) mit $\tilde{x}(t_0) = \ldots = \tilde{x}(t_{k-1}) = 0$ die Gleichung $\tilde{x} \equiv 0$ gilt, d.h. die exakte Lösung wird ohne Fehler reproduziert, falls die Startwerte exakt sind. Wie im obigen Beispiel betrachten wir nun den Fall, dass die bis zum Schritt $i^* \in \mathbb{N}$ erhaltenen Werte $\tilde{x}(t_i)$, $i = 0, \ldots, i^*$ durch Rechenfehler etwas gestört sind, wobei $\|\tilde{x}(t_i)\| \leq \varepsilon$ gelte. Für kleines $\varepsilon > 0$ sollten nun auch die nachfolgenden Werte $\tilde{x}(t_j)$, $j \geq i^*$ nur leicht gestört werden. Sicherlich kann man das nicht für alle Zeiten verlangen, aber doch zumindest auf vorgegebenen kompakten Zeitintervallen. Eine vernünftige Bedingung an das Verfahren für $f \equiv 0$ wäre also

$$\|\tilde{x}(t_i)\| \leq \varepsilon \text{ für } i = 0, \ldots, i^* \quad \Rightarrow \quad \|\tilde{x}(t_j)\| \leq C\varepsilon \text{ für alle } t_j \in [t_{i^*}, T].$$

Die wesentliche Beobachtung ist nun, dass zwar die Werte $\tilde{x}$ unabhängig von der Schrittweite h sind (dies ist gerade der entscheidende Unterschied zwischen der *linken* und der *rechten* Seite von (12.1)), nicht aber die Bedingung $t_j \in [t_{i^*}, T]$, die im Gegenteil stark von h abhängt: Je kleiner h wird, desto mehr Gitterpunkte t_j liegen in diesem Intervall. Da h beliebig klein werden kann, wird jeder t_j–Wert also für geeignetes h in $[t_{i^*}, T]$ liegen, weswegen man die Schranke $\|\tilde{x}(t_j)\| \leq C\varepsilon$ tatsächlich für alle $j \geq i^*$ fordern muss. Dies führt auf die folgende Definition, in der wir die jeweils die k Werte, die im Verfahren in Schritt i verwendet werden, gemeinsam betrachten.

Definition 12.8 Ein lineares Mehrschrittverfahren heißt *stabil*, falls ein $C > 0$ existiert, so dass für jeden Vektor $\tilde{x}^0 := (\tilde{x}(t_0), \ldots, \tilde{x}(t_{k-1}))^T$ von (reellen) Anfangswerten und alle $i \in \mathbb{N}$ die Ungleichung

$$\left\| \begin{pmatrix} \tilde{x}(t_i) \\ \vdots \\ \tilde{x}(t_{i+k-1}) \end{pmatrix} \right\| \leq C\|\tilde{x}^0\|$$

gilt. Hierbei ist die Folge $\tilde{x}(t_i)$ durch (12.1) bzw. (12.2) mit $\tilde{f}(t_i) = 0$ definiert, also kompakt geschrieben als

$$P_a(E)(\tilde{x})(t_i) = 0 \tag{12.7}$$

oder explizit ausgeschrieben als

$$\tilde{x}(t_{i+k}) = -\frac{a_{k-1}}{a_k}\tilde{x}(t_{i+k-1}) - \ldots - \frac{a_0}{a_k}\tilde{x}(t_i). \tag{12.8}$$

□

Wir werden nun ein einfaches Kriterium herleiten, das uns sagt, ob ein gegebenes Verfahren stabil ist. Hierzu stellen wir die Gleichung (12.8) zunächst in etwas anderer Form dar. Wir erinnern dazu an die linearen Differenzengleichungen (6.2), die durch eine Iterationsvorschrift der Form

$$x(t_{i+1}) = Ax(t_i)$$

mit einer Matrix $A \in \mathbb{R}^{k\times k}$ gegeben sind. Eine solche Gleichung heißt *stabil*, falls die Ungleichung

$$\|x(t_i)\| \leq C\|x(t_0)\|$$

für ein $C > 0$ und alle $i \in \mathbb{N}$ gilt. Das folgende Lemma zeigt, wie sich (12.8) als eine Matrix–Differenzengleichung schreiben lässt.

Lemma 12.9 Betrachte die lineare Differenzengleichung

$$x(t_{i+1}) = Ax(t_i) \tag{12.9}$$

mit

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -\frac{a_0}{a_k} & -\frac{a_1}{a_k} & -\frac{a_2}{a_k} & \cdots & \cdots & -\frac{a_{k-1}}{a_k} \end{pmatrix} \in \mathbb{R}^{k\times k}.$$

Dann gilt für die Lösungen von (12.8) mit $\tilde{x}^0 = x(t_0)$ die Gleichung

$$\begin{pmatrix} \tilde{x}(t_i) \\ \vdots \\ \tilde{x}(t_{i+k-1}) \end{pmatrix} = x(t_i).$$

Insbesondere ist das Mehrschrittverfahren genau dann stabil, wenn (12.9) stabil ist.

Beweis: Ausschreiben der Differenzengleichung (12.9) liefert für alle $i \in \mathbb{N}_0$ die Gleichungen

$$x_j(t_{i+1}) = x_{j+1}(t_i) \text{ für } j = 1, \ldots, k-1$$

und

$$x_k(t_{i+1}) = -\frac{a_0}{a_k}x_1(t_i) - \ldots - \frac{a_{k-1}}{a_k}x_k(t_i)$$

Hiermit folgt die Behauptung per Induktion über i. □

Um ein Stabilitätskriterium für (12.1) zu erhalten, genügt uns also ein Stabilitätskriterium für (12.9). Hier hilft der folgende Satz, der eine Erweiterung von Satz 6.4(ii) darstellt.

Hierbei nennen wir einen Eigenwert *halbeinfach*, wenn seine algebraische und geometrische Vielfachheit übereinstimmen. Dies ist genau dann der Fall ist, wenn er eine einfache Nullstelle des Minimalpolynoms m_A ist. Das Minimalpolynom m_A ist dabei das Polynom mit minimalem Grad $p \geq 1$, für das $m_A(A) = 0$ gilt. Das Minimalpolynom m_A teilt immer das charakteristische Polynom χ_A.

Satz 12.10 Eine lineare Differenzengleichung $x(t_{i+1}) = Ax(t_i)$ ist genau dann stabil, wenn alle Eigenwerte λ_i von A die Bedingung $|\lambda_i| \leq 1$ erfüllen und alle Eigenwerte λ_i mit $|\lambda_i| = 1$ halbeinfach sind.

Beweis: Wir nummerieren die Eigenwerte gemäß der Ordnung $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_d|$. Sei J die Jordan'sche Normalform von A mit Transformationsmatrix T, also $T^{-1}AT = J$. Wir schreiben kurz $x_i = x(t_i)$ und erinnern an die explizite Lösungsdarstellung $x_i = A^i x_0 = TJ^iT^{-1}x_0$. Wir schreiben $y_0 = T^{-1}x_0$ und $y_i = J^i y_0$.

Wir nehmen zunächst an, dass die Eigenwertbedingung erfüllt ist. Der Vektor y_0 lässt sich zerlegen in $y_0 = y_0^1 + y_0^2$ mit $y_0^1 = (y_1, \ldots, y_p, 0 \ldots, 0)^T$ und $y_0^2 = (0 \ldots, 0, y_{p+1}, \ldots, y_k)^T$, wobei $|\lambda_p| = 1$ und $|\lambda_{p+1}| < 1$ gilt. Mit $E_1 = \langle e_1, \ldots, e_p \rangle$ und $E_2 = \langle e_{p+1}, \ldots, e_k \rangle$ bezeichen wir die zugehörigen Unterräume. Für den Vektor y_i gilt nun

$$y_i = J^i y_0 = J^i(y_0^1 + y_0^2) = \underbrace{J^i y_0^1}_{=:y_i^1} + \underbrace{J^i y_0^2}_{=:y_i^2}.$$

Beachte, dass $y_i^1 \in E_1$ und $y_i^2 \in E_2$ liegt. Da die Einschränkung von J auf den Unterraum E_2 die Bedingung von Satz 6.4(ii) erfüllt (alle Eigenwerte im Betrag kleiner als 1), folgt die Existenz von $C_1 > 0$ und $\sigma > 0$ mit

$$\|y_i^2\| \leq C_1 \underbrace{e^{-\sigma(t_i - t_0)}}_{\leq 1} \|y_0^2\| \leq C_1 \|y_0^2\|.$$

Für y_i^1 gilt

$$\|y_i^1\| = \|J^i y_0^1\| \leq \|y_0\|,$$

wobei die letze Gleichung aus der Eigenwertstruktur folgt, denn J^i eingeschränkt auf E_1 ist wegen der Halbeinfachheit der Eigenwerte eine Diagonalmatrix mit Diagonalelementen λ_i mit $|\lambda_i| = 1$. Zusammen folgt also unter Verwendung der Definition der euklidischen Norm

$$\begin{aligned} \|x_i\| &\leq \|T\|\|y_i\| = \|T\|(\|y_i^1\| + \|y_i^2\|) \leq \|T\|(C_1\|y_0^2\| + |y_0^1\|) \\ &\leq (C_1 + 1)\|T\|\|y_0\| \leq (C_1 + 1)\|T\|\|T^{-1}\|\|x_0\| = C\|x_0\| \end{aligned}$$

für die Konstante $C = (C_1 + 1)\|T\|\|T^{-1}\|$.

Sei umgekehrt die Eigenwertbedingung nicht erfüllt. Falls ein Eigenwert λ_j mit $|\lambda_j| > 1$ existiert, so gilt für den zugehörigen Eigenvektor x_0

$$\|A^i x_0\| = |\lambda_j|^i \|x_0\| \to \infty \text{ für } i \to \infty,$$

was der Stabilität widerspricht. Falls ein nicht halbeinfacher Eigenwert λ_j mit $|\lambda_j| = 1$ existiert, so gibt es einen Eigenvektor x_0 sowie einen verallgemeinerten Eigenvektor x_1, für die die Gleichungen

$$Ax_0 = \lambda_j x_0 \quad \text{und} \quad Ax_1 = x_0 + \lambda_j x_1$$

gelten (dies folgt, da das Jordan–Kästchen zu dem nicht halbeinfachen Eigenwert λ_j eine 1 über der Diagonale besitzt). Per Induktion ergibt sich

$$A^i x_1 = i\lambda_j^{i-1} x_0 + \lambda_j^i x_1.$$

Da $\|\lambda_j^{i-1} x_0\| = \|x_0\|$ und $\|\lambda_j^i x_1\| = \|x_1\|$ (wegen $|\lambda_j| = 1$), folgt

$$\|A^i x_1\| \geq i\|x_0\| - \|x_1\| \to \infty \text{ für } i \to \infty,$$

was wiederum der Stabilität widerspricht. □

Zur Bestimmung der Stabilität genügt es also, die Eigenwerte der Matrix A zu bestimmen. Dies ist aber recht einfach, wie das folgende Lemma zeigt.

Lemma 12.11 Die Eigenwerte von A aus (12.9) sind genau die Nullstellen des Polynoms P_a aus (12.2). Ihre Vielfachheit im Minimalpolynom stimmt dabei mit ihrer Vielfachheit in P_a überein.

Beweis: Man rechnet nach, dass das charakteristische Polynom von A gerade durch

$$\chi_A(z) = z^k + \frac{a_{k-1}}{a_k} z^{k-1} + \ldots + \frac{a_1}{a_k} z + \frac{a_0}{a_k}$$

gegeben ist. Da die ersten Zeilen von $A^0, A^1, A^2, \ldots, A^{k-1}$ linear unabhängig sind (was aus der Verteilung der 0–Einträge leicht zu sehen ist), muss dies auch das Minimalpolynom m_A sein. Da $a_k \neq 0$ ist, stimmen die Nullstellen und Vielfachheiten von χ_A mit denen von

$$P_a(z) = a_0 + a_1 z + \ldots + a_k z^k = a_k \chi_A(z)$$

überein. □

Unsere Überlegungen führen nun direkt auf den folgenden Satz.

Satz 12.12 Ein lineares Mehrschrittverfahren (12.1) ist genau dann stabil, wenn alle Nullstellen λ_i von P_a die Bedingung $|\lambda_i| \leq 1$ erfüllen und alle Nullstellen λ_i von P_a mit $|\lambda_i| = 1$ einfache Nullstellen sind.

Beweis: Folgt sofort aus den vorangegangenen Aussagen.

Beachte, dass das Polynom P_a nach Bemerkung 12.7 für jedes konsistente Mehrschrittverfahren die Nullstelle 1 besitzen muss, also mindestens eine Nullstelle mit $|\lambda_i| = 1$ besitzt. Falls dies die einzige Nullstelle mit $|\lambda_i| = 1$ ist, nennt man das Verfahren *strikt stabil*. Falls es weitere Nullstellen λ_i mit $|\lambda_i| = 1$ gibt, so heißt das Verfahren *marginal stabil* oder *schwach stabil*. Obwohl sie theoretisch stabil sind, können solche Verfahren für bestimmte

Differentialgleichungen numerische Instabilitäten aufweisen, die z.B. durch Rundungsfehler hervorgerufen werden (vgl. das aktuelle Übungsblatt).

Für die explizite Mittelpunktregel z.B. berechnet man $P_a(z) = z^2 - 1$, das Polynom besitzt also die Nullstellen $z_{1/2} = \pm 1$ und ist damit stabil, genauer marginal stabil.

Für Einschrittverfahren, die als Spezialfall der Mehrschrittverfahren aufgefasst werden können, muss das Polynom P_a vom Grad $k = 1$ sein, denn nur x_{i+1} und x_i treten auf. Wegen $P_a(1) = 0$ kommt also nur $P_a(z) = z - 1$ in Frage, das als einzige Nullstelle $\lambda = 1$ besitzt. Also sind alle Einschrittverfahren stabil, weswegen wir die Stabilität dort nicht betrachten mussten. Dies ist auch der Grund, warum wir die Lipschitzbedingung für Einschrittverfahren nicht (wie in vielen Lehrbüchern) als Stabilitätsbedingung bezeichnet haben: Die Bedingungen bezeichnen verschiedene Sachverhalte, auch wenn sie den gleichen Zweck im Konvergenzbeweis erfüllen, nämlich zu garantieren, dass sich die in jedem Schritt gemachten lokalen Fehler nicht aufschaukeln können.

Auf Basis von Satz 12.12 können wir nun auch verstehen, warum die zwei Mehrschrittverfahren in dem einführenden Beispiel (12.4) so unterschiedliches Verhalten aufweisen. Für das Verfahren mit den Koeffizienten a_i ist das zugehörige Polynom $P_a(z) = z^2 - 3z + 2 = (z-1)(z-2)$, das gerade die Nullstellen 1 und 2 besitzt und das deswegen instabil ist. Für das zweite Verfahren mit den Koeffizienten $\tilde{a}_i$ gilt $P_{\tilde{a}}(z) = z^2 - 3/2z + 1/2 = (z-1)(z-1/2)$. Dieses Polynom hat die Nullstellen 1 und $1/2$, weswegen das Verfahren stabil ist.

12.3 Konvergenz

Ganz analog zu den Einschrittverfahren werden wir in diesem Abschnitt unser Hauptkonvergenzresultat

"Konsistenz (mit Ordnung p) + Stabilität $\Rightarrow$ Konvergenz (mit Ordnung p)"

formulieren und beweisen.

Zur Vorbereitung des Konvergenzsatzes benötigen wir noch ein Resultat über Lösungen von Differenzengleichungen, das im folgenden Lemma bereitgestellt wird.

Lemma 12.13 Betrachte die aus (12.7) hervorgehende *inhomogene Gleichung*

$$P_a(E)(y)(t_i) = c(t_i)$$

für eine Gitterfunktion $c : \mathcal{T} \to \mathbb{R}$ und ein stabiles Mehrschrittverfahren. Dann erfüllen die Lösungen dieser Gleichung die Abschätzung

$$|y(t_{i+k})| \leq C \left(\max_{l=0,\ldots,k-1} |y(t_l)| + \sum_{l=0}^{i} |c(t_l)| \right)$$

für eine geeignete Konstante $C > 0$.

Beweis: Für die vektorwertige Funktion $\hat{c}(t_i) = (0, \ldots, 0, c(t_i)/a_k)^T$ kann man die Gleichung in Matrixform

$$x(t_{i+1}) = Ax(t_i) + \hat{c}(t_i)$$

mit der Matrix A aus (12.9) und

$$\begin{pmatrix} y(t_i) \\ \vdots \\ y(t_{i+k-1}) \end{pmatrix} = x(t_i).$$

schreiben. Für diese Gleichung kann man die allgemeine Lösung per Induktion als

$$x(t_i) = A^i x(t_0) + \sum_{k=0}^{i-1} A^k \hat{c}(t_{i-k-1})$$

berechnen. Da A stabil ist, folgt aus der Definition der Matrixnorm sofort $\|A^k\|_\infty \leq \widetilde{C}$ für alle $k \in \mathbb{N}$ für ein $\widetilde{C} > 0$. Damit ergibt sich

$$\begin{aligned} |y(t_{i+k})| \;\leq\; \|x(t_{i+1})\|_\infty \;&\leq\; \widetilde{C}\|x(t_0)\|_\infty + \widetilde{C}\sum_{k=0}^{i} \|\hat{c}(t_{i-k})\|_\infty \\ &=\; \widetilde{C}\|x(t_0)\|_\infty + \widetilde{C}\sum_{k=0}^{i} |c(t_{i-k})/a_k| \\ &\leq\; \widetilde{C} \max_{l=0,\ldots,k-1} |y(t_l)| + \widetilde{C}/|a_k| \sum_{k=0}^{i} |c(t_i)|, \end{aligned}$$

also die Behauptung mit $C = \max\{\widetilde{C}, \widetilde{C}/|a_k|\}$. □

Wir kommen nun zum Konvergenzsatz. Wir formulieren das Resultat hier etwas schwächer als im Satz 2.7, da wir keine kompakte Menge von Anfangswerten, sondern nur einen einzelnen Anfangswert betrachten. Dies dient lediglich der Vermeidung allzu technischer Formulierungen in der Aussage und im Beweis des Satzes und hat keine prinzipiellen Gründe.

Satz 12.14 Gegeben sei ein Anfangswertproblem (1.1), (1.2) mit Anfangsbedingung (t_0, x_0) und p–mal stetig differenzierbarem Vektorfeld f. Gegeben seien weiterhin ein k–stufiges stabiles und konsistentes lineares Mehrschrittverfahren mit Ordnung $p \in \mathbb{N}$ und Näherungswerte $\tilde{x}(t_1), \ldots, \tilde{x}(t_{k-1})$ mit

$$\|\tilde{x}(t_i) - x(t_i; t_0, x_0)\| \leq \varepsilon_0 \;\text{ für }\; i = 1, \ldots, k-1.$$

Dann gilt für die durch das Verfahren auf dem Gitter $t_i = t_0 + hi$ zur Schrittweite h erzeugte Gitterfunktion $\tilde{x}(t_i)$ für alle Zeiten $t_i \in [t_0, T]$ und alle hinreichend kleinen $h > 0$ die Abschätzung

$$\|\tilde{x}(t_i) - x(t_i; t_0, x_0)\| \leq C(\varepsilon_0 + h^p)$$

für eine geeignete Konstante $C > 0$.

Beweis: Wir bezeichnen die exakte Lösung kurz mit $x(t)$ und wählen eine kompakte Umgebung $K \subset \mathbb{R} \times \mathbb{R}^n$ des exakten Lösungsgraphen $\{(t, x(t)) \,|\, t \in [t_0, T]\}$. Dann existiert ein $\delta_K > 0$, so dass für alle $t \in [t_0, T]$ die Folgerung $\|x - x(t)\| \leq \delta_K \Rightarrow (t, x) \in K$ gilt. Zudem existiert eine Konstante $L > 0$, so dass f auf K Lipschitz–stetig in x mit Konstante L ist. Mit N bezeichnen wir die größte ganze Zahl mit $N \leq (T - t_0)/h$.

Wie im Beweis von Satz 2.7 nehmen wir zunächst an, dass die numerische Lösung für alle $t_i \in [t_0, T]$ in K verläuft. Wir definieren den vektorwertigen Fehler als

$$\varepsilon_h(t_i) := x(t_i) - \tilde{x}(t_i).$$

Aus der Definition des Konsistenzfehlers folgt

$$P_a(E)(x)(t_i) = L(x, t_i, h) + hP_b(E)(\dot{x})(t_i) = L(x, t_i, h) + hP_b(E)(f)(t_i)$$

(wiederum mit der Abkürzung $f(t_i) = f(t_i, x(t_i))$). Von dieser Gleichung subtrahieren wir die Gleichung (12.2)

$$P_a(E)(\tilde{x})(t_i) = hP_b(E)(\tilde{f})(t_i).$$

Dies ergibt

$$P_a(E)(\varepsilon_h)(t_i) = L(x, t_i, h) + hP_b(E)\left(f(t_i) - \tilde{f}(t_i)\right).$$

Dies ist eine inhomogene (vektorwertige) Gleichung für ε_h. Indem wir Lemma 12.13 auf die einzelnen Komponenten von $\varepsilon_h(t_i)$ anwenden und $\|\varepsilon_h(t_j)\| \leq \varepsilon_0$ für $j = 0, \ldots, k-1$ ausnutzen, erhalten wir

$$\|\varepsilon_h(t_{i+k})\|_\infty \leq C\left(\varepsilon_0 + \sum_{l=0}^{i} \|L(x, t_l, h)\|_\infty + h\left\|P_b(E)\left(f(t_l) - \tilde{f}(t_l)\right)\right\|_\infty\right). \tag{12.10}$$

für alle $i = 0, \ldots, N - k$. Aus der Konsistenz folgt nun die Abschätzung

$$\|L(x, t_l, h)\|_\infty \leq C_p h^{p+1}$$

und aus der Lipschitz–Stetigkeit und der Definition von P_b und E folgt

$$\left\|P_b(E)\left(f(t_l) - \tilde{f}(t_l)\right)\right\|_\infty \leq L \sum_{m=0}^{k} |b_m| \, \|\varepsilon_h(t_{l+m})\|_\infty.$$

Setzen wir diese beiden Ungleichungen in (12.10) ein, so folgt

$$\begin{aligned}
\|\varepsilon_h(t_{i+k})\|_\infty &\leq C\left(\varepsilon_0 + \underbrace{\sum_{l=0}^{i} C_p h^{p+1}}_{\leq N C_p h^{p+1} \leq (T-t_0) C_p h^p} + h \sum_{l=0}^{i} L \sum_{m=0}^{k} |b_m| \, \|\varepsilon_h(t_{l+m})\|_\infty\right) \\
&\leq \widehat{C}_1 \varepsilon_0 + \widehat{C}_2 h^p + h\widehat{C}_3 \sum_{l=0}^{i+k} \|\varepsilon_h(t_l)\|_\infty
\end{aligned}$$

für geeignete Konstanten $\widehat{C}_q > 0$. Beschränken wir nun die Schrittweite durch $h \leq 1/(2\widehat{C}_3)$, so können wir nach $\|\varepsilon_h(t_{i+k})\|_\infty$ auflösen und erhalten mit $j = i + k$ die Ungleichung

$$\|\varepsilon_h(t_j)\|_\infty \leq C_1\varepsilon_0 + C_2h^p + hC_3 \sum_{l=0}^{j-1} \|\varepsilon_h(t_l)\|_\infty$$

mit $C_q = 2\widehat{C}_q$. Beachte, dass diese Ungleichung auch für $j = 1, \ldots, k-1$ stimmt wenn wir o.B.d.A. $C_1 \geq 1$ annehmen.

Per Induktion (wie im Beweis von Satz 2.7) ergibt sich daraus die Abschätzung

$$\|\varepsilon_h(t_j)\|_\infty \leq (C_1\varepsilon_0 + C_2h^p)e^{jhC_3} = (C_1\varepsilon_0 + C_2h^p)e^{(t_j - t_0)C_3},$$

also die gewünschte Behauptung.

Der induktive Beweis, dass die numerische Lösung für hinreichend kleine $h > 0$ tatsächlich in K liegt, verläuft für explizite Verfahren ganz analog zum Beweis von Satz 2.7. Für implizite Verfahren muss dieser Beweis in jedem Schritt um ein Fixpunktargument erweitert werden, das wir hier aber nicht ausführen wollen. □

Bemerkung 12.15 (i) Das Konvergenzresultat zeigt insbesondere, wie die Startwerte $\tilde{x}(t_1), \ldots, \tilde{x}(t_{k-1})$ bestimmt werden müssen. Um für das Mehrschrittverfahren die Konvergenzordnung p zu garantieren, müssen diese ebenfalls mit der Genauigkeit $O(h^p)$ bestimmt werden. Da es sich hier nur um endlich viele Werte handelt, deren Anzahl unabhängig von h ist, genügt es dazu, ein Einschrittverfahren mit Konsistenzordnung $p - 1$ zu verwenden. Der Beweis von Satz 2.7 zeigt nämlich, dass die ersten k Werte durch ein solches Verfahren immer die Genauigkeit $O(h^p)$ besitzen, falls k unabhängig von h ist. Der "Verlust" einer Ordnung beim Übergang von der Konsistenz– zur Konvergenzordnung ergibt sich erst dadurch, dass die Anzahl der nötigen Schritte von h abhängt.

(ii) Eine genauere Analyse zeigt, dass sogar die stärkere Aussage

$$\text{Konsistenz + Stabilität} \Leftrightarrow \text{Konvergenz}$$

gilt. Konsistenz und Stabilität sind also *notwendig und hinreichend* für die Konvergenz eines Verfahrens. □

12.4 Verfahren in der Praxis

In der Praxis haben sich zwei Klassen von Mehrschrittverfahren durchgesetzt. Beide Klassen haben gewisse Eigenschaften, die sie für gewisse Problemklassen besonders auszeichnen.

Adams–Verfahren

Historisch haben sich die Adams–Verfahren aus Quadraturformeln zur numerischen Integration entwickelt. Wir motivieren die Herleitung hier allerdings aus ihrer besonderen Eigenschaft, die ihre Vorteile in der Praxis begründet.

Wir haben gesehen, dass das Polynom P_a eines Mehrschrittverfahrens stabil sein muss, also — abgesehen von einer Nullstelle $= 1$ — nur Nullstellen mit Betrag $|\lambda_i| \leq 1$ besitzen darf. Je kleiner die Eigenwerte dabei im Betrag sind, desto "stabiler" wird das Verfahren. Bei den Adams–Verfahren wählt man P_a deswegen so, dass neben der $\lambda_1 = 1$ nur Nullstellen $\lambda_i = 0$ auftreten, also

$$P_a(z) = z^{k-1}(z-1) = z^k - z^{k-1}$$

ist. Beachte, dass damit auf der linken Seite von (12.1) nur die Werte $\tilde{x}(t_{i+k})$ und $\tilde{x}(t_{i+k-1})$ stehen bleiben.

Für jede beliebige Stufenanzahl k liefert Satz 12.6(iii) nun ein Gleichungssystem mit genau zwei Lösungen, nämlich

- genau ein *explizites* Adams–Verfahren der Konsistenzordnung $p = k$ (auch *Adams–Bashforth–Verfahren* genannt)
- genau ein *implizites* Adams–Verfahren der Konsistenzordnung $p = k + 1$ (auch *Adams–Moulton–Verfahren* genannt)

Z.B. lauten die Polynome P_b der ersten vier expliziten Adams–Verfahren

$$\begin{aligned} k=1: \quad P_b(z) &= 1 \\ k=2: \quad P_b(z) &= (3z-1)/2 \\ k=3: \quad P_b(z) &= (23z^2 - 16z + 5)/12 \\ k=4: \quad P_b(z) &= (55z^3 - 59z^2 + 37z - 9)/24 \end{aligned}$$

Interessanterweise ist das explizite Adams–Verfahren für $k = 1$ gerade das explizite Euler–Verfahren.

Für diese Verfahren hat sich ein Algorithmus durchgesetzt, der als *Prädiktor–Korrektor–Verfahren* bezeichnet wird. Ein Schritt dieses Algorithmus verläuft wie folgt:

Algorithmus 12.16 Prädiktor–Korrektor–Verfahren Gegeben seien das explizite und das implizite Adams–Verfahren der Stufe k.

1) Prädiktor–Schritt: Berechne $\tilde{x}(t_{i+k})$ mit dem expliziten Adams–Verfahren

2) Korrektor–Schritt: Führe *einen Schritt* der Fixpunktiteration zur Lösung des impliziten Verfahrens mit Startwert $\tilde{x}(t_{i+k})$ durch. □

Der Prädiktor–Schritt liefert hierbei eine Approximation mit Konsistenzfehler $O(h^{k+1})$. Für hinreichend kleine Schrittweite h ist die Kontraktionskonstante der Fixpunktiteration gleich Ch für ein $C > 0$. Also liefert der eine Iterationsschritt eine Approximation mit dem Konsistenzfehler

$$\frac{Ch}{1-Ch} O(h^{k+1}) = O(h^{k+2}).$$

Das Prädiktor–Korrektor–Verfahren besitzt also die Konsistenzordnung $k + 1$.

BDF–Verfahren

Obwohl die Familie der Adams–Verfahren implizite Verfahren enthält, sind diese (wegen ihrer recht kleinen Stabilitätsgebiete $\mathcal{S}$) schlecht für steife DGL geeignet.

Tatsächlich kann man beweisen, dass kein Mehrschrittverfahren der Ordnung $p > 2$ A–stabil ist. Die zur Lösung steifer DGL so nützliche Eigenschaft $\mathbb{C}^- \subseteq \mathcal{S}$ lässt sich also nicht erreichen. Es gibt allerdings eine Klasse impliziter Mehrschrittverfahren, die zumindest unendlich große Stabilitätsgebiete $\mathcal{S}$ besitzt, und die deswegen zur Lösung steifer DGL recht gut geeignet sind.

Dies ist die Klasse der BDF–Verfahren (BDF="backwards difference"). Hier wird gefordert, dass ein Kegel der Form $\{a + ib \in \mathbb{C}^- \,|\, |b| \leq c|a|\}$ für ein $c > 0$ in $\mathcal{S}$ liegt. Dies führt auf die Bedingung

$$P_b(z) = z^k.$$

Wiederum mit Satz 12.6(iii) erhält man dann Bedingungen, nun an die Koeffizienten von P_a, die die Konstruktion von Verfahren beliebig hoher Konsistenzordnung $p = k$ ermöglichen. Die ersten vier Polynome lauten hier

$$\begin{aligned}
k = 1: \quad & P_a(z) &&= z - 1 \\
k = 2: \quad & P_a(z) &&= \frac{3}{2}z^2 - 2z + \frac{1}{2} \\
k = 3: \quad & P_a(z) &&= \frac{11}{6}z^3 - 3z^2 + \frac{3}{2}z - \frac{1}{3} \\
k = 4: \quad & P_a(z) &&= \frac{25}{12}z^4 - 4z^3 + 3z^2 - \frac{4}{3}z + \frac{1}{4}
\end{aligned}$$

Für $k = 1$ ergibt sich gerade das implizite Euler–Verfahren. Die BDF–Verfahren sind allerdings nur bis $p = k = 6$ praktikabel, da die Verfahren für höhere Stufenzahlen instabil werden (beachte, dass die Bedingungen aus 12.6(iii) nur die Konsistenz, nicht aber die Stabilität sicher stellen).

Schrittweitensteuerung

Zuletzt wollen wir ganz kurz die Schrittweitensteuerung für Mehrschrittverfahren diskutieren. Sicherlich kann man die Fehlerschätzertheorie für Einschrittverfahren eins zu eins auf Mehrschrittverfahren übertragen und ebenso wie dort neue Schrittweiten berechnen und damit die Schrittweite adaptiv steuern.

Es ergibt sich aber ein technisches Problem, da die Schrittweite im aktuellen Schritt mit den Schrittweiten der $k-1$ vorangegangenen Schritte übereinstimmen muss, weil ansonsten die definierende Gleichung (12.1) nicht sinnvoll ausgewertet werden kann.

Abhilfe schafft hier eine alternative Darstellung, die wir für die Adams–Verfahren illustrieren: Wenn die Werte $\tilde{x}(t_i), \ldots, \tilde{x}(t_{i+k-1})$ eine Approximation der Ordnung p an die differenzierbare Funktion $x(t)$ in den Punkten $t_i, \ldots, t_{i+k-1}$ darstellen, so ist das durch die Daten

$$(t_i, \tilde{x}(t_i)), \ldots, (t_{i+k-1}, \tilde{x}(t_{i+k-1}))$$

definierte Interpolationspolynom $q(t)$ eine Approximation der Ordnung p an $x(t)$, und zwar für alle t aus einem vorgegebenen kompakten Intervall.

Für die Adams–Verfahren kann man nachrechnen, dass die Verfahren mit diesem Interpolationspolynom q gerade als

$$\tilde{x}(t_{i+k}) = \tilde{x}(t_{i+k-1}) + \int_{t_{i+k-1}}^{t_{i+k}} q(t)dt$$

gegeben sind (zum Beweis betrachtet man die Lagrange–Polynomdarstellung von q und integriert). Diese Gleichung ist nun unabhängig von der zur Berechnung von q verwendeten Schrittweite und kann daher für variable Schrittweiten ausgewertet werden.

Für die BDF–Verfahren ist ein ähnlicher Trick möglich, so dass auch hier die Schrittweitensteuerung anwendbar ist.

In MATLAB finden sich schrittweitengesteuerte Adams–Verfahren unter dem Namen `ode113` und BDF–Verfahren unter dem Namen `ode15s`.

Kapitel 13

Randwertprobleme

Bisher haben wir uns ausschließlich mit der Lösung von Anfangswertproblemen

$$\dot{x}(t) = f(t, x(t)), \quad x(t_0) = x_0$$

beschäftigt. In diesem Kapitel wollen wir eine weitere Problemstellung bei gewöhnlichen Differentialgleichungen betrachten, nämlich die sogenannten *Randwertprobleme*. Zur Einführung soll das folgende Beispiel dienen:

Beispiel 13.1 Betrachte die zweidimensionale Gleichung

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ -kx_2(t) - \sin x_1(t) \end{pmatrix},$$

die die Bewegung eines Pendels beschreibt, bei dem x_1 den Winkel des Pendels und x_2 die Winkelgeschwindigkeit des Pendels beschreibt. Die Konstante $k \geq 0$ gibt die Stärke der Reibung an, der das Pendel unterliegt.

Bei einem Anfangswertproblem gibt man nun eine Zeit t_0 und eine Anfangsbedingung $x_0 = (x_1^0, x_2^0)^T$ vor, was bedeutet, dass man Position und Geschwindigkeit des Pendels im Zeitpunkt t_0 festlegt und dann errechnet, wie sich das Pendel ausgehend von diese Anfangsbedingung in der Zukunft bewegt.

Bei einem Randwertproblem ist die Problemstellung anders: Hier gibt man sich zwei Zeitpunkte $t_0 < t_1$ vor, einen Anfangs- und einen Endzeitpunkt, und stellt zu beiden Zeitpunkten Bedingungen an die Lösung. Im Pendelmodell könnte man also zum Beispiel Winkel x_1^0 und x_1^1 vorgeben und nun eine Lösung $x^*(t) = (x_1^*(t), x_2^*(t))^T$ der Pendelgleichung berechnen wollen, für die $x_1^*(t_0) = x_1^0$ und $x_1^*(t_1) = x_1^1$ gilt. Gesucht ist also eine Pendelbewegung, die im Zeitpunkt t_0 den Winkel x_1^0 und im Zeitpunkt t_1 den Winkel x_1^1 annimmt. Die zugehörigen Geschwindigkeiten sind nicht festgelegt, sondern spielen hier vielmehr die Rolle freier Parameter, die während der numerischen Lösung so bestimmt werden müssen, dass die zugehörige Lösung die geforderten Bedingungen auch erfüllt. □

Allgemein formulieren wir das Randwertproblem wie folgt.

Definition 13.2 Ein *Randwertproblem* für eine gewöhnliche Differentialgleichung (1.1) im $\mathbb{R}^n$ besteht darin, eine Lösung $x^*(t)$ der Gleichung zu finden, die für Zeiten $t_0 < t_1$ und eine Funktion $r(x, y)$, $r : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ die Bedingung

$$r(x^*(t_0), x^*(t_1)) = 0$$

erfüllt. □

Für unser Pendelbeispiel 13.1 könnten wir die Funktion r z.B. als

$$r(x, y) = \begin{pmatrix} x_1 - x_1^0 \\ y_1 - x_1^1 \end{pmatrix}$$

definieren.

13.1 Lösbarkeit des Problems

Ob ein gegebenes Randwertproblem tatsächlich lösbar ist, ist im Allgemeinen sehr schwer zu überprüfen. Wir beschränken uns daher hier auf einen Existenzsatz für den speziellen Fall linearer Differentialgleichungen und beweisen im allgemeinen nichtlinearen Fall nur einen lokalen Eindeutigkeitssatz.

Aus der Theorie der Differentialgleichungen ist bekannt, dass die Lösungen linearer homogener Differentialgleichungen der Form

$$\dot{x}(t) = A(t)x(t) \tag{13.1}$$

als

$$x(t; t_0, x_0) = \Phi(t, t_0)x_0$$

geschrieben werden können, wobei die sogenannte *Fundamentalmatrix* $\Phi(t; t_0) \in \mathbb{R}^{n\times n}$ eine Lösung des matrixwertigen Anfangswertproblems

$$\dot{\Phi}(t) = A(t)\Phi(t), \qquad \Phi(t_0) = \mathrm{Id} \tag{13.2}$$

ist. Bezeichnet man die i-te Spalte dieser Matrix mit $\Phi_i(t; t_0)$, so sieht man leicht, dass Φ_i Lösung des Anfangswertproblems

$$\dot{\Phi}_i(t) = A(t)\Phi_i(t), \qquad \Phi_i(t_0) = e_i$$

ist, bei dem e_i den i-ten Einheitsvektor bezeichnet. Auf diese Weise kann man die Spalten der Matrix $\Phi(t; t_0)$ auch numerisch berechnen.

Satz 13.3 Gegeben sei eine inhomogene lineare Differentialgleichung der Form

$$\dot{x}(t) = A(t)x(t) + b(t) \tag{13.3}$$

und eine Randbedingung der Form

$$r(x, y) = Bx + Cy + d$$

für Matrizen $A(t), B, C \in \mathbb{R}^{n\times n}$ und Vektoren $b(t), d \in \mathbb{R}^n$. Es sei Φ die Fundamentalmatrix der zugehörigen homogenen Gleichung (13.1) und $x(t; t_0, x_0)$ eine Lösung von (13.3) mit beliebigem Anfangswert $x_0 \in \mathbb{R}^n$. Dann ist

$$x^*(t) = x(t; t_0, x_0^*)$$

genau dann eine Lösung des Randwertproblems, wenn der Anfangswert x_0^* eine Lösung des linearen Gleichungssystems

$$(B + C\Phi(t_1, t_0))(x_0^* - x_0) = -(Bx_0 + Cx(t_1; t_0, x_0) + d) \tag{13.4}$$

ist. Insbesondere existiert also genau dann eine eindeutige Lösung des Randwertproblems, wenn die Matrix $B + C\Phi(t_1, t_0)$ vollen Rang besitzt.

Beweis: Für zwei beliebige Anfangswerte $x_0, x_0^* \in \mathbb{R}^n$ gilt für die Differenz der zugehörigen Lösungen von (13.3)

$$\begin{aligned}\frac{d}{dt}(x(t; t_0, x_0^*) - x(t; t_0, x_0)) &= A(t)x(t; t_0, x_0^*) + b(t) - A(t)x(t; t_0, x_0) - b(t)\\ &= A(t)(x(t; t_0, x_0^*) - x(t; t_0, x_0))\end{aligned}$$

und damit

$$x(t; t_0, x_0^*) - x(t; t_0, x_0) = \Phi(t, t_0)(x_0^* - x_0)$$

und folglich auch

$$x(t; t_0, x_0^*) = x(t; t_0, x_0) + \Phi(t, t_0)(x_0^* - x_0).$$

Einsetzen in die Randbedingung ergibt

$$\begin{aligned}0 &= Bx(t_0; t_0, x_0^*) + Cx(t_1; t_0, x_0^*) + d\\ &= Bx_0^* + C\Big(x(t_1; t_0, x_0) + \Phi(t_1, t_0)(x_0^* - x_0)\Big) + d\\ &= \Big(B + C\Phi(t_1, t_0)\Big)(x_0^* - x_0) + Bx_0 + Cx(t_1; t_0, x_0) + d\end{aligned}$$

Die Randbedingung ist also genau dann erfüllt, wenn x_0^* eine Lösung des linearen Gleichungssystems (13.4) ist. □

Beachte, dass sich das Gleichungsystem (13.4) deutlich vereinfacht, wenn wir $x_0 = 0$ wählen. Wir werden später sehen, warum es dennoch nützlich ist, den Satz für allgemeine $x_0 \in \mathbb{R}^n$ zu formulieren.

Beispiel 13.4 Wenn wir die Pendelgleichung aus Beispiel 13.1 durch die lineare Pendelgleichung

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ -kx_2(t) - x_1(t) \end{pmatrix}$$

ersetzen, so erhalten wir ein Problem der Form aus Satz 13.3 mit

$$A = \begin{pmatrix} 0 & 1 \\ -1 & -k \end{pmatrix},\ b = 0,\ B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ C = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \text{ und } d = \begin{pmatrix} -x_1^0 \\ -x_1^1 \end{pmatrix}.$$

□

Für allgemeine nichtlineare Differentialgleichungen ist ein solcher Satz nicht beweisbar. Wir können aber, wenn wir annehmen, dass eine Lösung $x^*(t)$ des Randwertproblems existiert, zumindest Bedingungen für die lokale Eindeutigkeit der Lösung angeben und beweisen.

Dazu — aber auch für die numerische Lösung des Problems im nächsten Abschnitt — benötigen wir die partielle Ableitung

$$\frac{\partial}{\partial x_0} x(t; x_0, x_0)$$

der Lösung eines Anfangswertproblems. Wir hatten bereits im Abschnitt 2.4 über die Kondition verwendet, dass diese Ableitung über die sogenannte *Variationsgleichung*

$$\dot{y}(t) = A(t)y(t), \qquad A(t) = \frac{\partial f}{\partial x}(t, x(t; t_0, x_0)) \tag{13.5}$$

berechnet werden kann. Genauer gilt, dass die Fundamentalmatrix $\Phi(t, t_0)$ (vgl. (13.2)) der Variationsgleichung (13.5) gerade die (matrixwertige) Ableitung nach dem Anfangswert ist: Es gilt

$$\frac{\partial}{\partial x_0} x(t; x_0, x_0) = \Phi(t; t_0).$$

Diesen Zusammenhang nutzen wir in dem folgenden Satz.

Satz 13.5 Es sei $x^* : [t_0, t_1] \to \mathbb{R}^n$ eine Lösung des Randwertproblems aus Definition 13.2 mit $f \in C^1(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n)$ und $r \in C^1(\mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n)$. Es sei $\Phi^*(t, t_0)$ die Fundamentalmatrix (13.2) der Variationsgleichung (13.5) mit $x(t) = x^*(t)$. Zudem definieren wir die $n \times n$-Matrizen

$$B^* := \frac{\partial r}{\partial x}(x^*(t_0), x^*(t_1)) \text{ und } C^* := \frac{\partial r}{\partial y}(x^*(t_0), x^*(t_1))$$

über die Ableitungen der Randwertfunktion $r(x, y)$.

Dann gilt: Falls die *Sensitivitätsmatrix*

$$E^*(t) := B^*\Phi^*(t_0, t) + C^*\Phi^*(t_1, t)$$

für ein $t = \tau_0 \in [t_0, t_1]$ vollen Rang besitzt, so besitzt sie für alle $t \in [t_0, t_1]$ vollen Rang und x^* ist eine lokal eindeutige Lösung des Randwertproblems.

Beweis: Wir zeigen zunächst die lokale Eindeutigkeit. Definieren wir für eine beliebige Lösung $x(t; \tau_0, x_0)$ und die Randbedingungsfunktion r die Funktion

$$F(x_0) = r(x(t_0; \tau_0, x_0), x(t_1; \tau_0, x_0)),$$

so ist eine beliebige Lösung $x(t)$ der Differentialgleichung genau dann eine Lösung des Randwertproblems, wenn

$$F(x(\tau_0)) = 0 \tag{13.6}$$

gilt. Um die lokale Eindeutigkeit der Lösung zu zeigen, müssen wir also beweisen, dass eine Umgebung U um $x^*(\tau_0)$ existiert, so dass

$$F(x) \neq 0 \text{ für alle } x \in U \setminus x^*(\tau_0)$$

gilt.

Gleichung (13.6) ist ein nichtlineares Gleichungssystem mit n Gleichungen und n Unbekannten. Nach dem Satz über inverse Funktionen gibt es genau dann eine lokal eindeutige Lösung, wenn die Jacobi-Matrix

$$DF(x^*(\tau_0))$$

vollen Rang besitzt. Diese ist aber für $x_0 = x^*(\tau_0)$ gerade gegeben durch

$$\begin{aligned} DF(x_0) &= \frac{d}{dx_0} r(x(t_0;\tau_0,x_0), x(t_1;\tau_0,x_0)) \\ &= \frac{\partial r}{\partial x}(x(t_0;\tau_0,x_0), x(t_1;\tau_0,x_0))\frac{\partial}{\partial x_0} x(t_0,\tau_0,x_0) \\ &\quad + \frac{\partial r}{\partial y}(x(t_0;\tau_0,x_0), x(t_1;\tau_0,x_0))\frac{\partial}{\partial x_0} x(t_1,\tau_0,x_0) \\ &= B^*\Phi^*(t_0,\tau_0) + C^*\Phi(t_1,\tau_0) \end{aligned}$$

und besitzt daher vollen Rang. Daraus folgt die lokale Eindeutigkeit.

Würde nun ein $\tau_1 \in [t_0, t_1]$ existieren, für dass die Sensitivitätsmatrix keinen vollen Rang besitzt, so würden nach dem Satz über implizite Funktionen Werte x_0 beliebige nahe an $x^*(t)$ existieren, so dass $x(t;\tau_1,x_0)$ das Randwertproblem löst. Damit hätten wir Werte $x(\tau_0;\tau_1,x_0)$ gefunden, die in beliebig kleinen Umgebungen von $x^*(\tau_0)$ liegen und (13.6) lösen, was ein Widerspruch zur lokalen Eindeutigkeit ist. □

Auch wenn die Bedingungen dieses Satzes i.A. schwer zu überprüfen sind, so liefert er doch die Begründung dafür, dass eine numerische Berechnung der Lösung des Randwertproblems möglich ist, da das Problem zumindest lokal eine eindeutige Lösung besitzt und damit wohldefiniert ist. Zudem liefert er eine wichtige Einsicht in die Struktur des Problems, die wir im folgenden Abschnitt numerisch nutzen werden.

13.2 Schießverfahren

Der Beweis von Satz 13.5 zeigt bereits die Richtung auf, die wir bei der numerischen Lösung des Problems einschlagen können. Das Problem, eine Lösungsfunktion zu finden, die zwei vorgegebene Punkte verbindet, wurde dort reduziert auf das Problem, einen Anfangswert $x_0 \in \mathbb{R}^n$ zu finden, der das n-dimensionale nichtlineare Gleichungssystem (13.6) löst. Die dort definierte Abbildung F vereinfacht sich für $\tau_0 = t_0$ zu

$$F(x_0) = r(x_0, x(t_1;t_0,x_0)). \tag{13.7}$$

Diese Form wollen wir im Folgenden verwenden.

Unser Ziel ist nun, das Problem zu lösen, indem wir das Nullstellenproblem (13.7) numerisch lösen. Dieses Vorgehen — also die Lösung eines Randwertproblems durch die Lösung eines durch ein Anfangswertproblem bestimmtes Gleichungssystem — wird als *Schießverfahren* bezeichnet. Ursprung dieses etwas martialischen Namens ist tatsächlich das Schießen im militärischen Sinne, genauer die Artillerie. Auch hier hat man eine Endbedingung gegeben (nämlich ein zu treffendes Ziel) und variiert die Anfangsbedingung (Winkel des Geschützes oder Schussstärke), um die Endbedingung zu erfüllen.

Algorithmen zur Lösung nichtlinearer Gleichungssysteme kennen wir aus der Einführung in die Numerik, nämlich die Fixpunktiteration und das Newton-Verfahren. Während erstere nur unter relativ einschränkenden Bedingungen funktioniert (die wir hier realistischerweise nicht unbedingt annehmen können), funktioniert die zweite lokal immer, benötigt aber die Information über die Ableitung von F. Hier kommt als weitere Schwierigkeit hinzu, dass die Definition von F neben der — gegebenen — Abbildung r auch die — im Allgemeinen unbekannte — Lösung $x(t_1; t_0, x_0)$ enthält. Wie können F und die Ableitung DF aber numerisch auswerten, denn da $x(t_1; t_0, x_0)$ und $\Phi(t_1, t_0)$ ja gerade die Lösung von Anfangswertproblemen sind, können wir diese mit jedem der bisher behandelten Algorithmen berechnen.

Zunächst erinnern wir an das Newton-Verfahren im $\mathbb{R}^n$, vgl. die Einführung in die Numerik:

Gegeben sei eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$, ihre Ableitung $DF : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ sowie ein Startwert $x^{(0)} \in \mathbb{R}^n$ und eine gewünschte Genauigkeit $\varepsilon > 0$. Setze $i = 0$.

(1) Löse das lineare Gleichungssystem $DF(x^{(i)})\Delta x^{(i)} = F(x^{(i)})$
und berechne $x^{(i+1)} = x^{(i)} - \Delta x^{(i)}$

(2) Falls $\|\Delta x^{(i)}\| < \varepsilon$, beende den Algorithmus,
ansonsten setze $i = i + 1$ und gehe zu (1)

Um dies auf unser Problem anzuwenden, müssen wir nun klären, wie wir F und DF numerisch berechnen.

Die Berechnung von F aus (13.7) stellt dabei kein größeres Problem dar: Für gegebenes $x^{(i)}$ berechnen wir numerisch die Lösung $\tilde{x} = x(t_1; t_0, x^{(i)})$ mittels eines Ein- oder Mehrschrittverfahrens und berechnen damit

$$F(x^{(i)}) \approx r(x^{(i)}, \tilde{x})$$

Komplizierter ist die Berechnung von DF. Zunächst gilt nach der Rechnung im Beweis von Satz 13.5 mit $\tau = t_0$

$$DF(x^{(i)}) = B + C\Phi(t_1, t_0)$$

mit Matrizen B und C gegeben durch

$$B = \frac{\partial r}{\partial x}(x^{(i)}, x(t_1; t_0, x^{(i)})) \text{ und } C = \frac{\partial r}{\partial y}(x^{(i)}, x(t_1; t_0, x^{(i)}))$$

für die Randwertfunktion $r(x, y)$. Die i-te Spalte der Matrix $\Phi(t_1, t_0)$ ist nun gerade die Lösung des Anfangswertproblems

$$\dot{y}_i(t) = \frac{\partial f}{\partial x}(t, x(t; t_0, x^{(i)}))y_i(t), \quad y_i(t_0) = e_i,$$

wobei e_i der i-te Einheitsvektor ist.

Die numerische Berechnung von F und DF kann also wie folgt geschehen. Vorab berechnen wir (analytisch) die Ableitungen

$$\frac{\partial f}{\partial x}(t, x), \quad \frac{\partial r}{\partial x}(x, y), \quad \frac{\partial r}{\partial y}(x, y).$$

In jedem Schritt des Newton-Verfahrens approximieren wir dann numerisch die Lösung $z(t_1)$ des $n(n+1)$–dimensionalen Anfangswertproblems

$$\dot{z}(t) = g(t, z(t)), \quad z(t_0) = z_0$$

mit

$$z(t) = \begin{pmatrix} x(t) \\ y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix}$$

und

$$g(t, z(t)) = \begin{pmatrix} f(t, x(t)) \\ \frac{\partial f}{\partial x}(t, x(t)) y_1(t) \\ \vdots \\ \frac{\partial f}{\partial x}(t, x(t)) y_n(t) \end{pmatrix}, \quad z_0 = \begin{pmatrix} x^{(i)} \\ e_1 \\ \vdots \\ e_n \end{pmatrix}.$$

Mit Hilfe der numerischen Approximation

$$\tilde{z} = \begin{pmatrix} \tilde{x} \\ \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{pmatrix} \approx z(t_1)$$

berechnen wir dann die Approximationen

$$F(x^{(i)}) \approx r(x^{(i)}, \tilde{x})$$

und

$$DF(x^{(i)}) \approx \widetilde{B} + \widetilde{C}\widetilde{\Phi}(t_1, t_0)$$

mit

$$\widetilde{B} = \frac{\partial r}{\partial x}(x^{(i)}, \tilde{x}), \;\; C = \frac{\partial r}{\partial y}(x^{(i)}, \tilde{x}) \;\text{ und }\; \widetilde{\Phi}(t_1, t_0) = \Big(\tilde{y}_1, \ldots, \tilde{y}_n\Big).$$

Damit kann das Newton-Verfahren nun vollständig implementiert werden.

In Beispiel 13.1 lautet das zu lösende Differentialgleichungssystem also

$$\dot{z}(t) = \begin{pmatrix} z_2(t) \\ -kz_2(t) - \sin(z_1(t)) \\ z_4(t) \\ -kz_4(t) - \cos(z_1(t)) z_3(t) \\ z_6(t) \\ -kz_6(t) - \cos(z_1(t)) z_5(t) \end{pmatrix} \quad \text{mit} \quad z(t_0) = z_0 = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Interessant ist, was im Falle eines linearen Problems im Sinne von Satz 13.3 passiert. In diesem Fall ist ein Newton-Schritt ausgehend von einem beliebigen Startwert $x^{(0)}$ gerade äquivalent zu dem linearen Gleichungssystem (13.4). Wir erhalten die (bis auf numerische Diskretisierungsfehler) exakte Lösung des Randwertproblems also nach genau einem Schritt des Newton-Verfahrens. Auf die genauen Auswirkungen der Diskretisierungsfehler im Linearen und im Nichtlinearen können wir hier aus Zeitgründen nicht genauer eingehen.

13.3 Mehrzielmethode

Die oben beschriebene Methode funktioniert theoretisch gut, hat aber in der Praxis den nicht zu unterschätzenden Nachteil, dass die Lösung $x(t_1; t_0, x_0)$ in vielen Beispielen sehr sensitiv vom Anfangswert x_0 abhängt.

Als Beispiel betrachten wir das Randwertproblem

$$\dot{x}(t) = x^2,\ t_0 = 0,\ t_1 = 1,\ r(x, y) = y - 9.$$

Gesucht ist also eine Lösung $x^*(t)$ dieser Gleichung mit $x(1) = 9$. Da die allgemeine Lösung hier leicht als

$$x(t; 0, x_0) = \frac{x_0}{1 - x_0 t}$$

ausgerechnet werden kann, sieht man, dass die gesuchte Lösung gerade

$$x^*(t) = \frac{0.9}{1 - 0.9t}, \quad \text{also} \quad x^*(0) = 0.9$$

lautet. Liegen wir mit unserem Anfangswert nur um 10% oberhalb dieses Wertes, also $x_0 = 0.99$, so erhalten wir

$$x(1; 0, 0.99) = 99 \quad \text{und damit} \quad r(0.9, x(1; 0, 0.99)) = 90.$$

Für $x_0 = 1$ ist die Sache noch schlimmer, da die Lösung dann zum Zeitpunkt $t_1 = 1$ gar nicht mehr existiert.

Die Schätzlösung $r(x^{(0)}, x(t_1; t_0, x^{(0)}))$ im Newton-Verfahren kann also selbst bei einer relativ guten Startschätzung $x^{(0)} \approx x^*(t_0)$ weit von $r(x^*(t_0), x^*(t_1)) = 0$ abweichen oder sogar undefiniert sein. Es ist leicht einzusehen, dass dies große numerische Konvergenzprobleme im Newton-Verfahren nach sich zieht und der Bereich der lokalen Konvergenz des Verfahrens dadurch sehr klein wird.

Eine Abhilfe ist die in den 1960er Jahren zuerst vorgeschlagene und in den 1970er Jahren vor allem durch Roland Bulirsch[1] weiterentwickelte *Mehrzielmethode* (auch *Mehrfachschießverfahren*). Die Idee dabei ist, das Intervall $[t_0, t_1]$ in $d \in \mathbb{N}$ Teilintervalle $[\tau_i, \tau_{i+1}]$ zu zerlegen mit

$$t_0 = \tau_0 < \tau_1 < \ldots < \tau_d = t_1.$$

Statt die Lösung $x(t_0; t_1, x_0)$ für einen Anfangswert x_0 auf dem gesamten Intervall $[t_0, t_1]$ zu berechnen, wählt man nun d Anfangswerte $x_0, \ldots, x_{d-1}$ und berechnet separat die Lösungen

$$x(\tau_k; \tau_{k-1}, x_{k-1}), \quad k = 1, \ldots, d$$

auf den Teilintervallen $[\tau_{k-1}, \tau_k]$. Damit sich diese Lösungen zu einer Gesamtlösung auf dem Intervall $[t_0, t_1]$ zusammensetzen lassen, müssen die Stetigkeitsbedingungen

$$x(\tau_k; \tau_{k-1}, x_{k-1}) = x_k, \quad k = 1, \ldots, d-1$$

[1] deutscher Mathematiker, geb. 1932

gelten und damit diese Gesamtlösung eine Lösung des Randwertproblems ist, muss zusätzlich noch die ursprüngliche Randbedingung

$$r(x_0, x(\tau_d; \tau_{d-1}, x_{d-1})) = 0$$

gelten.

Definieren wir nun eine neue Randbedingungsfunktion $R : \mathbb{R}^{2dn} \to \mathbb{R}^{dn}$ mittels

$$R(x_0, x_1, x_1', x_2, x_2' \ldots, x_{d-1}, x_{d-1}', x_d') = \begin{pmatrix} x_1 - x_1' \\ \vdots \\ x_{d-1} - x_{d-1}' \\ r(x_0, x_d') \end{pmatrix},$$

so liefert die Lösung des Nullstellenproblems

$$F(x_0, \ldots, x_{d-1}) = 0$$

mit $F : \mathbb{R}^{dn} \to \mathbb{R}^{dn}$ definiert durch

$$\begin{aligned} F(x_0, \ldots, x_{d-1}) \quad = \quad & R\Big(x_0, x_1, x(\tau_1; \tau_0, x_0), x_2, x(\tau_2; \tau_1, x_1), \ldots, \\ & x_{d-1}, x(\tau_{d-1}; \tau_{d-2}, x_{d-2}), x(\tau_d; \tau_{d-1}, x_{d-1})\Big) \end{aligned}$$

eine Lösung des ursprünglichen Randwertproblems. Da die Lösungen der Differentialgleichung hier nur auf kurzen Intervallen $[\tau_{k-1}, \tau_k]$ berechnet werden müssen, sind sie deutlich weniger sensitiv gegenüber Änderungen in den Anfangswerten. Dies überträgt sich auf die Randwertfunktion, weswegen die Mehrzielmethode einen deutlich größeren Konvergenzbereich besitzt.

Der Preis dafür ist natürlich die Erhöhung der Dimension des Nullstellenproblems von n auf dn, die sich insbesondere bei der Lösung der linearen Gleichungssysteme im Newton-Verfahren bemerkbar macht (beachte, dass die numerische Berechnung der höheren Anzahl von Differentialgleichungslösungen i.A. kaum mehr Aufwand verursacht, weil diese auf entsprechend kürzeren Intervallen zu lösen sind). Hier kann insbesondere durch Ausnutzen der speziellen Bandstruktur der entstehenden Gleichungssysteme viel Rechenzeit gespart werden, für Details der algorithmischen Umsetzung verweisen wir z.B. auf das Buch von Deuflhard und Bornemann [2, Abschnitt 8.2.2].

Anhang A

Biologische Modelle

Mathematische Modelle werden in vielen verschiedenen Bereichen der Biologie verwendet. Klassische Anwendungen sind z.B. die Untersuchung von Wachstumsprozessen und biochemischen Reaktionen oder die Ausbreitung von Epidemien, neuere Anwendungen finden sich z.B. in vielen Teilgebieten der Gentechnik oder in der Immunologie. Wir werden hier bei den klassischen Bereichen bleiben und uns (ausführlich) mit der Populationsdynamik sowie deren technischer Anwendung im Chemostat-Modell und (kürzer) mit Epidemien beschäftigen.

A.1 Populationsdynamik für eine Art

Populationsdynamik bezeichnet die Analyse des Wachstums einer oder mehrerer Arten oder Spezies in einem (meist sehr einfach modellierten) Ökosystem. In diesem Abschnitt wollen wir mit Modellen für eine Art beginnen.

A.1.1 Differenzen– und Differentialgleichungen

In der mathematischen Modellierung von Wachstumsprozessen stellt sich zunächst die Frage, ob gewöhnliche Differentialgleichungen überhaupt das richtige mathematische Modellierungswerkzeug sind. Tatsächlich "lebt" eine Differentialgleichung immer auf kontinuierlichen Räumen, während die in der Populationsdynamik auftretenden Größen zunächst einmal diskreter Natur sind: Die Größe einer Population wird üblicherweise in der Anzahl der Individuen gemessen, die selbstverständlich eine natürliche Zahl ist. Dieses Problem wird in praktisch allen Modellen dadurch gelöst, dass man die Größe der Population nicht anhand der diskreten Anzahl der Individuen sondern anhand ihrer Biomasse x misst, und genau so wollen wir es hier halten. Die Biomasse x ist eine (nichtnegative) reelle Zahl, deren zeitliche Entwicklung man durch eine Differentialgleichung modellieren kann.

Das nächste Problem ist die richtige Wahl der Zeitachse. Biologische Messungen (z.B. zum Bestand einer Population) werden niemals kontinuierlich für $t \in [t_0, t_1]$ durchgeführt, sondern zu diskreten Zeiten $t_1 < t_2 < t_3 < \ldots$. Der Zuwachs oder die Abnahme einer

Population wird dementsprechend oft bezüglich diskreter Zeitpunkte ausgedrückt. Ein allgemeines diskretes Modell einer Populationsdynamik für x in einem festgelegten Gebiet ist gegeben durch

$$x(t_{i+1}) = x(t_i) + \Delta G(t_i) - \Delta S(t_i) + \Delta M(t_i). \tag{A.1}$$

Hierbei bezeichnet

$\Delta G(t_i)$: Anzahl der Geburten im Intervall $[t_i, t_{i+1}]$ (≥ 0)

$\Delta S(t_i)$: Anzahl der Sterbefälle im Intervall $[t_i, t_{i+1}]$ (≥ 0)

$\Delta M(t_i)$: Migration (Zu– und Abwanderung) im Intervall $[t_i, t_{i+1}]$ $(\geq 0$ oder $\leq 0)$

Gleichungen von Typ (A.1) nennt man *Differenzengleichungen* und tatsächlich kann man mit solchen zeitdiskreten Modellen arbeiten und es gibt viele (auch aktuelle) Forschungsarbeiten, die sich mit der Theorie von Differenzengleichungen beschäftigen.

Wir werden hier nicht mit Differenzengleichungen arbeiten, sondern statt dessen ein Differentialgleichungsmodell herleiten. Der Grund dafür, Differentialgleichungen vorzuziehen, liegt im Wesentlichen darin, dass es für Differentialgleichungen viele mathematische Analysemethoden gibt, die für Differenzengleichungen entweder komplizierter sind oder gar nicht zur Verfügung stehen, zum Teil aus prinzipiellen mathematischen Gründen (da Lösungen von Differenzengleichungen i.A. ein sehr viel komplexeres Verhalten aufweisen als Lösungen von Differentialgleichungen) oder einfach, weil noch niemand versucht hat sie herzuleiten und zu beweisen. Als Modellierungwerkzeug sind Differenzengleichungen den Differentialgleichungen sicherlich ebenbürtig.

Wie kommt man nun von (A.1) zu einer Differentialgleichungsformulierung? Nehmen wir an, dass die Zeitpunkte t_i äquidistant verteilt sind, dass also $t_{i+1} - t_i =: \Delta t$ für ein von i unabhängiges Δt gilt. Dann kann man (A.1) für $t = t_i$ umformulieren als

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} = \frac{\Delta G(t)}{\Delta t} - \frac{\Delta S(t)}{\Delta t} + \frac{\Delta M(t)}{\Delta t}.$$

Beachte, dass ΔG, ΔS und ΔM von Δt abhängen, auch wenn dies in der Notation nicht explizit klar wird. Für $\Delta t \to 0$ erhält man so

$$\frac{d}{dt}x(t) = g(t) - s(t) + m(t).$$

Man könnte versuchen, die Funktionen g, s und m mittels

$$g(t) = \lim_{\Delta t \to 0} \frac{\Delta G(t)}{\Delta t},\ s(t) = \lim_{\Delta t \to 0} \frac{\Delta S(t)}{\Delta t} \text{ und } M(t) = \lim_{\Delta t \to 0} \frac{\Delta M(t)}{\Delta t}$$

aus ΔG, ΔS und ΔM zu bestimmen, was sinnvoll wäre, wenn wir ΔG, ΔS und ΔM definiert hätten. Diesen Umweg wollen wir nicht gehen, statt dessen werden wir g und s direkt aus geeigneten Modellannahmen ableiten. Migration werden wir in unseren Modellen nicht betrachten, weswegen m immer gleich 0 sein wird.

A.1.2 Einfache Modelle

Die Herleitung eines Modells geschieht typischerweise in zwei Schritten: Im ersten Schritt werden geeigete *strukturelle Annahmen* an die rechte Seite der Differentialgleichung gemacht, was mathematisch bedeutet, dass wir eine gewisse Form des Vektorfeldes f festlegen, die aus bekannten Gesetzmäßigkeiten oder aus heuristischen Überlegnungen folgt. In dieser Form finden sich dann eine Reihe von freien Parametern, die im zweiten Schritt — der *Parameterschätzung* — bestimmt werden, um die Ergebnisse des Modells in Übereinstimmung mit realen Daten zu bringen. Wir werden uns in dieser Vorlesung vorwiegend mit dem ersten Schritt befassen, für unser einfachstes Modell wollen wir aber auch den zweiten Schritt durchführen, um damit ein numerisches Verfahren zu illustrieren, mit dem man dies durchführen kann.

Das einfachste Modell der Populationsdynamik für eine Art macht die folgenden Annahmen:

(i) $g(t)$ ist linear proportional zum aktuellen Bestand der Population:

$$g(t) = \gamma x(t) \text{ für ein } \gamma \in \mathbb{R}$$

(ii) $s(t)$ ist linear proportional zum aktuellen Bestand der Population:

$$s(t) = \sigma x(t) \text{ für ein } \sigma \in \mathbb{R}$$

(iii) Migration findet nicht statt: $m(t) \equiv 0$

Dies führt auf die Differentialgleichung

$$\dot{x}(t) = \lambda x(t) \tag{A.2}$$

wobei γ *Geburtenrate*, σ *Sterberate* und λ mit $\lambda = \gamma - \sigma \in \mathbb{R}$ *Wachstumsrate* genannt wird. Man rechnet leicht nach, dass die Lösungen von (A.2) mit Anfangsbedingung $x(t_0) = x_0$ durch

$$x(t; x_0) = x_0 e^{\lambda(t-t_0)}$$

gegeben sind. Beachte, dass $x(t)$ hier — wie in allen Wachstumsmodellen — die Größe einer Population beschreibt, so dass in diesen Modellen nur $x \geq 0$ und damit insbesondere $x_0 \geq 0$ sinnvoll ist. Wir schreiben hier $\mathbb{R}^+ = \{x \in \mathbb{R} \,|\, x > 0\}$ und $\mathbb{R}_0^+ = \mathbb{R}^+ \cup \{0\}$.

Auch wenn dies ein sehr einfaches Modell ist, so beschreibt es doch manche realen Wachstumsphänomene relativ gut. Abbildung A.1 zeigt z.B. die Größe der Weltbevölkerung zwischen 1950 und 2000 (in Milliarden Menschen) mit einer Lösung von (A.2). Die Werte x_0 und λ wurden hier über ein nichtlineares Ausgleichsproblem geschätzt, vgl. Abschnitt 6.6 des Skripts zur Vorlesung "Numerische Mathematik 1"[1], das zugehörige MATLAB M–File `weltbev.m` ist im E-Learning erhältlich. Mit $t_0 = 1950$ erhalten wir hier $\lambda = 0.0173456$ und $x_0 = 2.605331$. Es sollte erwähnt werden, dass die Ermittlung geeigneter Parameter für ein Modell (man spricht von *Parameterschätzung* oder *Parameteridentifikation*) ein eigenständiges anspruchsvolles mathematisches (meist numerisches) Problem ist, das wir in dieser Vorlesung nicht weitergehend betrachten können. Beachte, dass die Parameter von

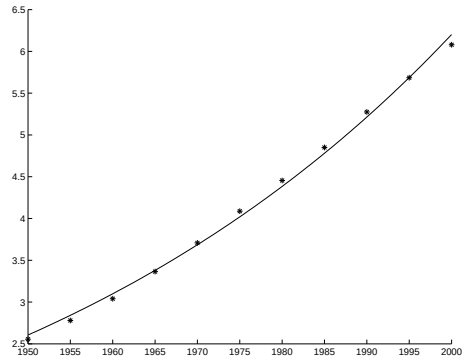

Abbildung A.1: Wachstum der Weltbevölkerung und Lösung von (A.2)

den verwendeten Einheiten abhängen. Hier haben wir x in Milliarden Menschen und t in Jahren angegeben.

Das derzeitige Weltbevölkerungswachstum wird also durch (A.2) offenbar recht gut beschrieben, andere reale Wachstumsprozesse hingegen werden durch dieses Modell überhaupt nicht gut beschrieben, beispielsweise das Wachstum der Bevölkerung in Europa, welches in den letzten Jahrzehnten praktisch zum Stillstand gekommen ist. Der Grund dafür ist aus der Struktur der Lösungen sofort ersichtlich: Aus $\lambda > 0$ folgt $e^{\lambda t} \to \infty$ für $t \to \infty$, für $x_0 > 0$ wächst die modellierte Population also über alle Grenzen; die Wahl $\lambda < 0$ (d.h., die Sterberate ist größer als die Geburtenrate) schafft hier keine brauchbare Abhilfe: in diesem Fall folgt $e^{\lambda t} \to 0$ für $t \to \infty$, was das reale Verhalten sicherlich auch nicht korrekt widerspiegelt — zumindest derzeit nicht.

Um sich verlangsamendes Wachstum modellieren zu können, werden wir (A.2) um eine "Wachstumsgrenze" erweitern, die wir hier durch eine obere Schranke $K > 0$ für die Größe der Population modellieren; K steht für die *Kapazität* des Lebensraums. Diese ergibt sich aus den zur Verfügung stehenden Ressourcen, wie z.B. Nahrung, Trinkwasser etc. Wir fügen dazu einen Faktor $w(x)$ mit den folgenden Eigenschaften in die Gleichung (A.2) ein.

(i) Falls $x < K$ ist, soll $w(x) > 0$ sein, da noch "Platz" für Wachstum vorhanden ist.

(ii) Falls $x > K$ ist, soll $w(x) < 0$ sein, um "negatives Wachstum" zu erzwingen.

Die einfachste Funktion, die dieses leistet, ist die lineare Funktion $w(x) = K - x$. Wir erhalten damit die Gleichung

$$\dot{x}(t) = \lambda(K - x(t))x(t), \tag{A.3}$$

die als *logistisches Wachstum* bezeichnet wird. Der Ausdruck $\lambda(K - x)$ ist hier die — nun nichtlineare — Wachstumsrate. Auch für diese DGL ist die explizite Lösung bekannt, sie ist gegeben durch

$$x(t; t_0, x_0) = \frac{K}{1 + \left(\frac{K}{x_0} - 1\right) e^{-\lambda K(t-t_0)}}.$$

[1] `http://www.uni-bayreuth.de/departments/math/~lgruene/numerik0405/`

Man kann das Verhalten der Lösung nun an diesem expliziten Ausdruck untersuchen. Wir wollen hier aber — zur Einübung — einen anderen Weg gehen und die dadurch erhaltenen Resultate an der expliziten Lösung überprüfen.

Hierzu definieren und betrachten wir zunächst einige wichtige Begriffe für Differentialgleichungen, und zwar allgemein im $\mathbb{R}^n$.

Definition A.1 Ein Punkt $x^* \in \mathbb{R}^n$ heißt *Gleichgewicht* (auch *Ruhelage*, *Fixpunkt* oder *Equilibrium*) für eine DGL (1.1), falls $x(t; t_0, x^*) = x^*$ ist für alle $t, t_0 \in \mathbb{R}$. □

Man sieht leicht, dass ein Punkt x^* genau dann ein Gleichgewicht ist, wenn $f(t, x^*) = 0$ ist für alle $t \in \mathbb{R}$. Für unser Modell (A.3) sind die Nullstellen von $f(x) = \lambda(K - x)x$ leicht zu bestimmen, es ergeben sich die Gleichgewichte $x^* = 0$ und $x^{**} = K$.

Gleichgewichte sind vor allem deswegen interessant, weil sie Aufschluss über das Langzeitverhalten der Lösungen geben können. Im Modell (A.3) sieht man, dass die Lösungen $x(t)$ zwischen den Gleichgewichten streng monoton wachsen, falls $x(t) \in (0, K)$ liegt (da die Ableitung $\dot{x}(t)$ dann positiv ist), während sie für $x(t) > K$ streng monoton fallen. Da die Lösungen in positiver Zeit durch die Gleichgewichtslösung $x(t) \equiv x^{**} = K$ beschränkt sind (wegen des Eindeutigkeitssatzes können sie diese nicht schneiden), sind sie also monoton und beschränkt, und damit konvegent.

Mit Hilfe des folgenden Satzes (der ein Spezialfall des sogenannten *Barbalat-Lemmas* ist) können wir mögliche Grenzwerte genau charakterisieren.

Satz A.2 Betrachte eine DGL (1.1) mit autonomem f. Sei $x(t; t_0, x_0)$ eine Lösung, die für $t \to \infty$ oder $t \to \infty$ gegen einen Punkt $x^* \in \mathbb{R}^n$ konvergiert. Dann ist x^* ein Gleichgewicht.

Beweis: Wir beweisen den Fall $t \to \infty$, der Fall $t \to -\infty$ folgt analog. Betrachte dazu die Lösung $x(t) = x(t; t_0, x_0)$. Da diese Lösung gegen x^* konvergiert, folgt aus der Stetigkeit von f die Konvergenz $f(x(t)) \to f(x^*)$. Sei nun für ein gegebenes $\varepsilon > 0$ die Zeit $t^* > 0$ so groß gewählt, dass die Ungleichungen

$$\|x(t) - x^*\| \le \varepsilon \quad \text{und} \quad \|f(x(t)) - f(x^*)\| \le \varepsilon$$

für alle $t \ge t^*$ gelten. Dann folgt für alle $t \ge t^*$ aus (1.3) die Ungleichung

$$\|x(t) - x(t^*)\| = \left\| \int_{t_*}^{t} f(x(\tau)) d\tau \right\| \ge \left\| \int_{t_*}^{t} f(x^*) d\tau \right\| - \left\| \int_{t_*}^{t} f(x(\tau)) - f(x^*) d\tau \right\|$$

und daraus

$$\begin{aligned}
(t - t^*)\|f(x^*)\| &= \left\| \int_{t_*}^{t} f(x^*) d\tau \right\| \\
&\le \|x(t) - x(t^*)\| + \left\| \int_{t_*}^{t} f(x(\tau)) - f(x^*) d\tau \right\| \\
&\le \underbrace{\|x(t) - x(t^*)\|}_{\le \|x(t) - x^*\| + \|x^* - x(t^*)\|} + \int_{t_*}^{t} \|f(x(\tau)) - f(x^*)\| d\tau \le 2\varepsilon + (t - t^*)\varepsilon.
\end{aligned}$$

Diese Ungleichung gilt für alle $t > t^*$, insbesondere also für $t = t^* + 1$. Mit dieser Wahl folgt

$$\|f(x^*)\| \leq 3\varepsilon,$$

also, da $\varepsilon > 0$ beliebig war, $\|f(x^*)\| = 0$ und damit $f(x^*) = 0$. Folglich ist x^* ein Gleichgewicht der DGL. □

Satz A.2 hat eine wichtige Konsequenz für die Analyse von Differentialgleichungen. Er besagt nämlich, dass wir mit den Gleichgewichten im autonomen Fall bereits alle möglichen Grenzwerte von Lösungstrajektorien kennen.

In unserem Modell (A.3) können wir auf Grund der Monotonie also schließen, dass alle Lösungen mit $x(t_0) > 0$ für $t \to \infty$ gegen $x^{**} = K$ konvergieren. In Rückwärtszeit folgt ebenfalls auf Grund der Monotonie, dass alle Lösungen mit $x(t_0) \in [0, K)$ für $t \to -\infty$ gegen 0 konvergieren, während die Lösungen mit $x(t_0) > K$ für $t \to -\infty$ gegen $+\infty$ divergieren: würden sie konvergieren, müsste wegen der Monotonie und auf Grund von Satz A.2 ein weiteres Gleichgewicht $x^{***} > K$ existieren, was aber nicht der Fall ist.

Im eindimensionalen Fall kann man leicht mit der Monotonie argumentieren um Grenzwerte von Lösungen zu ermitteln, für höherdimensionale Systeme geht dies i.A. nicht mehr, wir brauchen also andere Techniken. Grundlage dafür ist die folgende Definition, die für allgemeine DGL im $\mathbb{R}^n$ mögliche Konvergenzsituationen in einer Umgebung eines Gleichgewichts beschreibt.

Definition A.3 (i) Ein Gleichgewicht x^* einer DGL (1.1) heißt *(lokal) exponentiell stabil*, falls eine Umgebung N von x^* sowie Parameter $\sigma, \lambda > 0$ existieren, so dass für alle $x_0 \in N$, alle $t_0 \in \mathbb{R}$ und alle $t \geq t_0$ die Ungleichung

$$\|x(t; t_0, x_0) - x^*\| \leq \sigma e^{-\lambda(t-t_0)} \|x_0 - x^*\|$$

gilt.

(ii) Ein Gleichgewicht x^* einer DGL (1.1) heißt *exponentiell instabil*, falls Parameter $\sigma, \lambda > 0$ und eine Umgebung N von x^* existieren, so dass in jeder Umgebung N_0 von x^* ein Punkt $x_0 \in N_0$ existiert, für den für alle $t_0 \in \mathbb{R}$ die Ungleichung

$$\|x(t; t_0, x_0) - x^*\| \geq \sigma e^{\lambda(t-t_0)} \|x_0 - x^*\|$$

gilt für alle $t \geq t_0$ für die $x(t; t_0, x_0) \in N$ gilt.

(iii) Ein Gleichgewicht x^* einer DGL (1.1) heißt *exponentiell antistabil*, falls Parameter $\sigma, \lambda > 0$ und eine Umgebung N von x^* existieren, so dass für alle $x_0 \in N$ mit $x_0 \neq x^*$ und alle $t_0 \in \mathbb{R}$ die Ungleichung

$$\|x(t; t_0, x_0) - x^*\| \geq \sigma e^{\lambda(t-t_0)} \|x_0 - x^*\|$$

gilt für alle $t \geq t_0$ für die $x(t; t_0, x_0) \in N$ gilt. □

Für $t \to \infty$ konvergieren also im Fall (i) alle Lösungen aus einer Umgebung des Gleichgewichtes gegen das Gleichgewicht x^*. Im Fall (iii) laufen alle Lösungen für wachsendes t weg von x^*, Konvergenz gegen x^* ist nicht möglich. Im Fall (ii) gibt es beliebig nahe an

x^* startende Lösungen die von x^* weg laufen, es ist aber nicht ausgeschlossen, dass ein Anfangswert $x_0 \neq x^*$ existiert, für den $x(t; t_0, x_0)$ gegen x^* konvergiert. Wir werden später Beispiele dafür kennen lernen.

Beachte, dass (i)–(iii) keineswegs alle möglichen Szenarien beschreiben. So könnte z.B. eine Funktion $\beta(\|x_0 - x^*\|, t)$ existieren, die langsamer als $\sigma e^{-\lambda t}\|x_0 - x^*\|$ gegen Null konvergiert und für die statt (i) die Ungleichung

$$\|x(t; t_0, x_0) - x^*\| \leq \beta(\|x_0 - x^*\|, t)$$

gilt.

Der Grund dafür, in diesen Definitionen die (doch recht speziellen) exponentiellen Abschätzungen zu verwenden, liegt darin, dass sich für diese Definitionen einfache nachprüfbare Kriterien beweisen lassen — zumindest falls die DGL autonom ist.

Satz A.4 Sei x^* ein Gleichgewicht einer DGL (1.1) mit autonomem Vektorfeld $f : \mathbb{R}^n \to \mathbb{R}^n$. Sei f in einer Umgebung von x^* stetig differenzierbar und sei $Df(x^*) \in \mathbb{R}^{n\times n}$ die Ableitung (also die Jacobi–Matrix) von f an der Stelle x^*. Dann gilt:

(i) Das Gleichgewicht x^* ist genau dann (lokal) exponentiell stabil, wenn alle Eigenwerte $\lambda_i \in \mathbb{C}$ von $Df(x^*)$ negativen Realteil haben.

(ii) Das Gleichgewicht x^* ist genau dann exponentiell instabil, wenn es einen Eigenwert $\lambda_i \in \mathbb{C}$ von $Df(x^*)$ gibt, der positiven Realteil besitzt.

(iii) Das Gleichgewicht x^* ist genau dann exponentiell antistabil, wenn alle Eigenwerte $\lambda_i \in \mathbb{C}$ von $Df(x^*)$ positiven Realteil haben.

Ein Beweis für (i) findet sich z.B. als Korollar 7.6 in meinem Skript zur Vorlesung "Stabilität und Stabilisiernug linearer Systeme"[2]. Beweise für (ii) und (iii) finden sich in Büchern über gewöhnliche Differentialgleichungen. Die Jacobi–Matrix $Df(x^*)$ wird oft *Linearisierung* von (1.1) in x^* genannt.

Wir wollen dieses Resultat an unserem Modell (A.3) illustrieren und testen, ob sie mit den aus der Monotoniebetrachtungen erhaltenen Resultate übereinstimmen. Wie bereits erwähnt gilt hier

$$f(x) = \lambda(K - x)x$$

und die Gleichgewichte sind gegeben durch $x^* = 0$ und $x^{**} = K$. Da die DGL eindimensional ist, ist die Ableitung Df von f reellwertig. Nach Produktregel gilt

$$Df(x) = f'(x) = \lambda(K - x) - \lambda x \quad \Rightarrow \quad Df(x^*) = \lambda K \text{ und } Df(x^{**}) = -\lambda K.$$

Die Eigenwerte dieser "1×1–Matrizen" sind natürlich gerade $\lambda K > 0$ für $x^* = 0$ und $-\lambda K < 0$ für $x^{**} = K$. Das Gleichgewicht x^* ist also exponentiell antistabil, während x^{**} exponentiell stabil ist. Dies stimmt mit den bisherigen Beobachtungen überein: $x^{**} = K$ ist ein möglicher Grenzwert für $t \to \infty$, $x^* = 0$ nicht.

Hat man ein lokal exponentiell stabiles Gleichgewicht gefunden (hier also x^{**}), so besteht der nächste Analyseschritt darin, zu ermitteln, für welche Anfangswerte die Lösungen gegen

[2] `http://www.uni-bayreuth.de/departments/math/~lgruene/linstab0203/`

x^{**} konvergieren. Dies ist die Frage nach dem *Einzugsbereich* des Gleichgewichtes x^{**}. Allgemein ist der Einzugsbereich eines lokal exponentiell stabilen Gleichgewichtes x^* für eine autonome DGL gegeben als

$$\mathcal{D}(x^*) := \{x_0 \in \mathbb{R}^n \mid \lim_{t\to\infty} x(t; x_0) = x^*\}$$

und für die Umgebung N aus Definition A.3(i) gilt

$$\mathcal{D}(x^*) = \{x_0 \in \mathbb{R}^n \mid x(t; x_0) \in N \text{ für ein } t \geq 0\},$$

da alle Lösungen, die nach N laufen wegen (1.5) gegen x^* konvergieren müssen und umgekehrt alle Lösungen, die gegen x^* konvergieren, durch N laufen müssen.

Im $\mathbb{R}^n$ ist die Ermittlung von $\mathcal{D}$ eine schwierige, oft unlösbare Aufgabe. Im eindimensionalen Fall ist die Sache einfacher, da man mit der Monotonie der Lösungen argumentieren kann, wie wir oben bereits gesehen haben. Tatsächlich haben wir die Einzugsbereiche für (A.3) bereits in der Diskussion nach Satz A.2 schon fast vollständig bestimmt. Dort haben wir gesehen, dass alle Lösungen mit $x(t_0) > 0$ gegen x^{**} konvergieren, es gilt also $\mathcal{D}(x^{**}) \subseteq (x^*, \infty) = (0, \infty)$. Tatsächlich gilt hier sogar Gleichheit, da die Lösungen mit $x(t_0) \leq x^* = 0$ sicherlich nicht gegen x^{**} konvergieren, da sie die Gleichgewichtslösung $x(t) \equiv 0$ nicht verlassen bzw. nicht schneiden können.

Wir fassen unsere Analyse der Modells (A.3) noch einmal zusammen:

(1) Es gibt zwei Gleichgewichte, $x^* = 0$ und $x^{**} = K$, dabei ist x^* exponentiell antistabil und x^{**} exponentiell stabil.

(2) Genau die Lösungen mit Anfangswert $x_0 \in (0, \infty)$ konvergieren gegen x^{**}.

(3) Alle Lösungen mit Anfangswert $x_0 \in [0, x^{**})$ konvergieren in Rückwärtszeit (also für $t \to -\infty$) gegen x^*, alle Lösungen mit Anfangswert $x_0 > x^{**}$ divergieren in Rückwärtszeit gegen $+\infty$.

(Anfangswerte $x_0 < 0$ ergeben im Modell keinen Sinn, weswegen wir sie nicht betrachten).

In Abbildung A.2 sind die oben angegebenen expliziten Lösungen mit $K = \lambda = 1$ für die Anfangswerte $x_0 = 0, 1/100, 1$ und 2 dargestellt. Es ergibt sich genau das beschriebene Verhalten.

Auch das logistische Wachstum kann an reale Daten zur Weltbevölkerung angepasst werden, vgl. das MATLAB M-File `weltbevlog.m` im E-Learning, das zeigt, dass das Modell für die zukünftige Entwicklung ausgesprochen gut mit den Vorhersagen des US-Census Büros übereinstimmt.

Bemerkung A.5 Das logistische Wachstum (A.3) ist nicht das einzige Modell für beschränktes Wachstum. Zur Modellierung von Zellwachstum z.B. wird oft die DGL

$$\dot{x}(t) = \lambda x(t) \ln\left(\frac{K}{x(t)}\right) \tag{A.4}$$

verwendet, die als *Gompertz–Wachstum* bezeichnet wird und mit deren Lösungen sich klinische Ergebnisse gut nachvollziehen lassen. Hier sind die expliziten Lösungen unbekannt; mit ähnlichen Methoden wie oben kann man aber nachweisen, dass das qualitative Lösungsverhalten dem von (A.3) entspricht. □

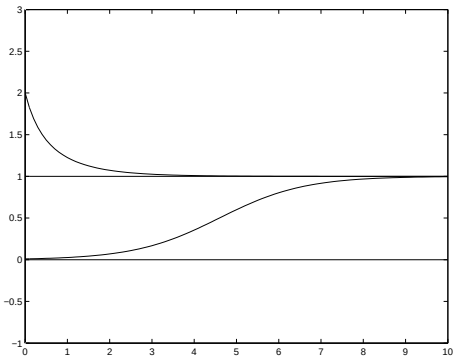

Abbildung A.2: Lösungen der logistischen Wachstumsgleichung (A.3)

A.1.3 Eine Anwendung des Modells

Mathematische Modelle werden oft zur Beschreibung und Erklärung realer Situationen eingesetzt. Sie dienen aber auch als Teilsysteme in der mathematischen Untersuchung komplexerer Phänomene. Wir wollen dieses Prinzip an einem Beispiel illustrieren.

Wir wollen ein Modell für den Fischfang aufstellen, bei dem $x(t)$ eine Fischpopulation beschreibt. Dazu ergänzen wir das Modell um eine *Fangstrategie* $u(t)$, welche die Intensität, mit der der Bestand befischt wird und damit die Abnahme der Population durch den Fischfang beschreibt. Als Modellannahme verwenden wir, dass sich das Fischwachstum durch das logistische Wachstum (A.3) beschreiben lässt, falls keine Fische gefangen werden.

Die sich daraus ergebende DGL

$$\dot{x}(t) = \lambda(K - x(t))x(t) - u(t). \tag{A.5}$$

wird *Schäfers Modell* genannt.

Beachte, dass dies nun eine nichtautonome DGL ist. Zudem ist es — je nach Wahl von $u(t)$ — möglich, dass die Lösungen mit positiven Anfangswert x_0 negativ werden, was aber vom Modell her nicht sinnvoll ist, weswegen wir $x(t) > 0$ durch die Wahl von $D = \mathbb{R} \times \mathbb{R}^+$ als Definitionsbereich sicher stellen. Die Wahl von D ist also nicht mathematisch sondern aus Modellgesichtspunkten motiviert.

Eine Fangstrategie auf einem Intervall $[t_0, t_1]$ ist nun einfach eine stetige Funktion $u : [t_0, t_1] \to \mathbb{R}_0^+$. Für einen Anfangswert $x_0 > 0$ nennen wir u *zulässig*, falls die zugehörige Lösung $x(t; t_0, x_0, u)$ auf dem ganzen Intervall $[t_0, t_1]$ existiert. Wir schreiben die Funktion u hier als zusätzlichen Parameter in die Lösung, um die Abhängigkeit der Lösungen von u zu betonen.

Die Anzahl M der gefangenen Fische (natürlich wieder als Biomasse ausgedrückt) ergibt sich nun als Integral über $u(t)$, also

$$M = \int_{t_0}^{t_1} u(t)dt.$$

Ziel des Fischers könnte es nun sein, diese Größe M zu maximieren. Dies würde jedoch unausweichlich zur Ausrottung der Fische führen: Wären zum Zeitpunkt t_1 noch Fische

da, so könnte man $u(t)$ erhöhen und würde trotzdem noch eine zulässige Fangstrategie erhalten. Dies wäre zwar auf dem betrachteten Intervall optimal, nach der Zeit t_1 wäre der Fischer aber arbeitslos, weswegen dies auf lange Sicht keine gute Strategie ist. Selbst wenn der Fischer sich nach der Zeit t_1 zur Ruhe setzen will, wäre dies keine gute Lösung, in jedem Fall aus ökologischer Sicht aber auch aus ökonomischer Sicht, da dies zur Vernichtung der Bestände führen würde.

Man muss also das Überleben der Fische in die Optimierung einbeziehen. Dies führt auf ein Optimierungsproblem unter Nebenbedingungen:

$$\text{maximiere } \int_{t_0}^{t_1} u(t)dt$$

unter den Nebenbedingungen

(i) u ist zulässig für den Anfangswert x_0

(ii) $x(t_1; t_0, x_0, u) \geq x_1$ für einen vorgegebenen Wert $x_1 > 0$

Dies ist ein sogenanntes *optimales Steuerungsproblem*, für dessen Lösung es eine Vielzahl von analytischen und numerischen Techniken gibt. (Weiterführende Vorlesungen in diesem Gebiet werden an der Uni Bayreuth regelmäßig angeboten.)

Hier können wir dieses Problem nicht lösen, statt dessen betrachten wir einen alternativen Ansatz, den wir mit unseren Methoden behandeln können. Wir wählen die Fangstrategie $u(t)$ proportional zur Menge der vorhandenen Fische: $u(t) = cx(t)$ für eine *Fangrate* $c > 0$. Dies vereinfacht nicht nur die Analyse, sondern liefert auch ein Modell für die Tatsache, dass man bei gleichbleibender Befischung (z.B. durch Auslegen von Netzen) in der Regel immer eine zu $x(t)$ proportionale Menge von Fischen fangen wird. Die Fangrate c ergibt sich dabei z.B. aus Anzahl und Größe der Netze und der Dauer des Auslegens. Mit dieser Wahl von u ergibt sich (A.5) zu

$$\dot{x}(t) = \lambda \left(K - \frac{c}{\lambda} - x(t)\right) x(t).$$

Dadurch verschiebt sich das Gleichgewicht x^{**} aus der obigen Analyse, genauer kann man leicht die Gleichgewichte

$$x^* = 0 \quad \text{und} \quad x^{**} = K - \frac{c}{\lambda}$$

berechnen. Für die Ableitung gilt

$$Df(x^*) = \lambda K - c = \lambda x^{**} \quad \text{und} \quad Df(x^{**}) = c - \lambda K = -\lambda x^{**}.$$

Jetzt muss man drei Fälle unterscheiden.

1. Fall: $x^{**} = K - \frac{c}{\lambda} > 0$. In diesem Fall bleibt alles wie oben, x^{**} ist lokal exponentiell stabil und jede Lösung mit Anfangswert $x_0 > 0$ konvergiert gegen x^{**}.

2. Fall: $x^{**} = K - \frac{c}{\lambda} < 0$. In diesem Fall wird $x^* = 0$ lokal exponentiell stabil und es gilt $f(x) < 0$ für alle $x > 0$. Alle Lösungen fallen also monoton und konvergieren schließlich gegen $x^* = 0$, für $t \to \infty$ werden die Fische also ausgerottet.

3. Fall: $x^{**} = K - \frac{c}{\lambda} = 0$. In diesem Fall vereinfacht sich die DGL zu $\dot{x}(t) = -\lambda(x(t))^2$, also ist jede Lösung monoton fallend. Zudem gilt $x^* = x^{**} = 0$. Alle Lösungen mit $x_0 > 0$ konvergieren gegen $x^* = 0$: sie können gegen keinen größeren Wert konvergieren, da kein größeres Gleichgewicht existiert; andererseits können sie die konstante Lösung $x(t; x^*) \equiv 0$ aber auch nicht schneiden. Beachte, dass das Gleichgewicht $x^* = 0$ weder lokal exponentiell stabil noch exponentiell instabil ist. Wie in Fall 2 werden die Fische für $t \to \infty$ ausgerottet.

Aus dieser Analyse kann man nun versuchen zu berechnen, wie $c > 0$ gewählt werden muss, damit der Ertrag maximiert wird. Auf beliebigen endlichen Intervallen ist das nicht so einfach, da aber alle Lösungen gegen eines der Gleichgewichte konvergieren, können wir zumindest approximativ den Ertrag für die Zeiten bestimmen, in denen die Lösung bereits nahe am Gleichgewicht liegt. Wir betrachten den Ertrag in einem Zeitintervall $[t_1, t_1 + 1]$ der Länge 1, wobei wir annehmen, dass t_1 so groß ist, dass wir uns bereits in der Nähe des Gleichgewichtes befinden. Im Fall 1 erhalten wir so

$$M = \int_{t_1}^{t_1+1} cx(t)dt \approx \int_{t_1}^{t_1+1} cx^{**}dt = cx^{**} = cK - \frac{c^2}{\lambda} > 0$$

und im Fall 2 und 3 ergibt sich analog

$$M \approx cx^* = 0.$$

Offensichtlich ist Fall 1 vorzuziehen, da nur dort (auf lange Sicht) ein positiver Ertrag erzielt wird. Zur Maximierung des Fangergebnisses muss man nun den Ausdruck $M(c) = cK - \frac{c^2}{\lambda}$ in c maximieren. Ableiten liefert die notwendige Bedingung

$$M'(c^*) = K - 2\frac{c^*}{\lambda} = 0 \quad \Leftrightarrow \quad c^* = \lambda K/2,$$

und da die zweite Ableitung $M''(c) = -2/\lambda < 0$ ist, ist dies tatsächlich ein lokales Maximum, sogar ein globales, da es das einzige ist. Der maximale Ertrag ergibt sich also zu

$$M(c^*) = cK - \frac{c^2}{\lambda} = \frac{\lambda K^2}{2} - \frac{\lambda^2 K^2}{4\lambda} = \frac{K^2\lambda}{4}.$$

Welchen Wert haben solche Folgerungen aus einem Modell? Zunächst einmal muss man sich die möglichen Unzulänglichkeiten des Basismodells vergegenwärtigen; für das hier zu Grunde liegende Modell (A.3) machen wir dies im folgenden Abschnitt. Wenn man nun annimmt (oder experimentell belegen kann), dass das Modell Aussagekraft besitzt, so erlauben solche Rechnungen Einsicht in die Struktur des modellierten Phänomens. Hier zum Beispiel beobachtet man, dass man auf lange Sicht den maximalen Ertrag nicht erzielt, indem man die Fangrate beliebig erhöht, denn oberhalb des Wertes c^* wird der langfristig erzielbare Ertrag wieder sinken. In unserem Fall erlaubt dies durchaus gerechtfertigte *qualitative* Folgerungen für das modellierte Fischfangproblem. Eine zuverlässige *quantitative* Berechnung der realen optimalen Fangrate dürfte auf Basis eines so einfachen Modells allerdings nahezu unmöglich sein.

A.1.4 Abschließende Diskussion

Wir wollen das Modell (A.3) noch einmal abschließend diskutieren:

- Das Modell eignet sich gut zur Beschreibung von Wachstum unter idealen Bedingungen; die Ergebnisse von Laborversuchen lassen sich damit gut reproduzieren

In der realen Anwendung gibt es allerdings eine Reihe von weiteren Einflüssen, die hier nicht berücksichtigt werden:

- Naturbedingungen sind in der Regel variabel, z.B. durch Jahreszeiten bedingt. Im Modell ist alles konstant (realistischere Modelle verwenden hier zeitabhängige bzw. stochastische Parameter, wie wir sie im Kapitel über Finanzmathematik kennen lernen werden).
- Die räumliche Verteilung sowohl der Population als auch der Ressourcen wird nicht modelliert (dies könnte z.B. eine partielle Differentialgleichung leisten, mit der vom Ort abhängige Populationsdichten modelliert werden können).
- Die Geburts– und Sterberate hängen unmittelbar von der Größe der Population ab. Faktoren wie z.B. die Altersverteilung werden nicht berücksichtigt (hier können Delay–Differentialgleichungen Abhilfe schaffen, die wir später betrachten werden).
- Der Einfluss anderer Arten ist nicht im Modell enthalten.

Im nächsten Abschnitt werden wir uns mit Modellen beschäftigen, in denen der letzte Punkt berücksichtigt wird.

A.2 Populationsdynamik für mehrere Arten

In diesem Abschnitt werden wir die Modelle (A.2) und (A.3) auf den Fall mehrerer Arten verallgemeinern. Wir werden dabei zunächst auf den Fall von zwei Arten eingehen, wobei die erste Art (Beute) die Nahrung der zweiten Art (Räuber) darstellt.

A.2.1 Das Räuber–Beute Modell mit unbeschränkten Ressourcen

Dieser Abschnitt behandelt die Erweiterung des sehr einfachen Modells A.2 auf den Fall von zwei Arten, und zwar Beutetiere (z.B. Ziegen) und Räubertiere (z.B. Wölfe).

Es bezeichne also x_1 die Größe der Beutepopulation und x_2 die Größe der Räuberpopulation. Wir machen die folgenden Modellannahmen:

(i) Die Beutepopulation x_1 verhält sich gemäß (A.2) mit $\lambda = \gamma - \sigma$, wobei γ konstant ist und $\sigma = \tilde{\sigma} + bx_2$. Für die Beutetiere gibt es also unbegrenzte Ressourcen und die Sterberate σ besteht aus einem konstanten Term $\tilde{\sigma} \in (0, \gamma)$ (natürlicher Tod) und einem zu x_2 proportionalen Term bx_2 (Tod durch Räuber). Für $x_2 = 0$ wächst die Population exponentiell. Wir setzen $a = \gamma - \tilde{\sigma} > 0$.

(ii) Die Räuberpopulation verhält sich ebenfalls gemäß (A.2) mit $\lambda = \gamma - \sigma$. Hier ist die Sterberate σ konstant und $\gamma = \tilde{\gamma} + dx_1$ für $\tilde{\gamma} \in (0, \sigma)$ und $d > 0$, d.h. die Geburtenrate hängt affin linear von der Anzahl der zur Verfügung stehenden Beute x_1 ab; für $x_1 = 0$ stirbt die Räuberpopulation wegen $\sigma > \tilde{\gamma}$ aus. Wir setzen $c = \sigma - \tilde{\gamma} > 0$.

Zusammen erhalten wir so die zweidimensionale Differentialgleichung

$$\begin{array}{rcl} \dot{x}_1(t) & = & ax_1(t) - bx_1(t)x_2(t) \\ \dot{x}_2(t) & = & -cx_2(t) + dx_1(t)x_2(t) \end{array} \tag{A.6}$$

mit den Parametern $a, b, c, d > 0$. Dieses Modell wird als *Lotka–Volterra Modell* bezeichnet. V. Volterra[3] hat dieses Modell im biologischen Kontext eingeführt (vgl. dazu Abschnitt A.3.1), A.J. Lotka[4] hat das Modell unabhängig von Volterra zur Beschreibung einer hypothetischen chemischen Reaktion entwickelt.

Um die Analyse von (A.6) zu vereinfachen wollen wir die Zahl der Parameter reduzieren. Dazu führt man die Koordinatentransformation $x_1 \to \frac{d}{c}x_1$ und $x_2 \to \frac{b}{a}x_2$ durch. Dies führt auf die neuen Gleichungen

$$\begin{array}{rcl} \dot{x}_1(t) & = & ax_1(t) - ax_1(t)x_2(t) \quad = \quad ax_1(t)(1 - x_2(t)) \\ \dot{x}_2(t) & = & -cx_2(t) + cx_1(t)x_2(t) \quad = \quad -cx_2(t)(1 - x_1(t)) \end{array} \tag{A.7}$$

Beachte, dass die Lösungen $\tilde{x}(t; t_0, \tilde{x}_0)$ von (A.6) und $x(t; t_0, x_0)$ von (A.7) mittels

$$x(t; t_0, x_0) = A\tilde{x}(t; t_0, A^{-1}x_0) \text{ und } \tilde{x}(t; t_0, \tilde{x}_0) = A^{-1}x(t; t_0, A\tilde{x}_0) \quad \text{für} \quad A = \begin{pmatrix} \frac{d}{c} & 0 \\ 0 & \frac{b}{a} \end{pmatrix}$$

zusammenhängen; alle Lösungen von (A.6) lassen sich also aus (A.7) berechnen und umgekehrt. Man nennt die zwei Gleichungen auch *äquivalent*.

Wir bestimmen zunächst die Gleichgewichte von (A.7), also die Nullstellen des Vektorfeldes

$$f(x) = \begin{pmatrix} ax_1(1 - x_2) \\ -cx_2(1 - x_1) \end{pmatrix}.$$

Hier sieht man leicht, dass die Punkte

$$x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{und} \quad x^+ = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

die einzigen Gleichgewichtspunkte sind. Zur Bestimmung der Stabilität dieser Gleichgewichte berechnen wir

$$Df(x^*) = \begin{pmatrix} a(1 - x_2^*) & -ax_1^* \\ cx_2^* & -c(1 - x_1^*) \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & -c \end{pmatrix} \quad \text{und} \quad Df(x^+) = \begin{pmatrix} 0 & -a \\ c & 0 \end{pmatrix}$$

Als Eigenwerte dieser Matrizen ergeben sich a und $-c$ in x^* sowie $\pm\sqrt{-ca}$ in x^+. Aus Satz A.4 folgt damit exponentielle Instabilität (aber nicht Antistabilität) von x^*. Dies ist gut zu erklären: Für Anfangswerte der Form $x_0 = (x_1, 0)^T$ (also keine Räuber) mit $x_1 \neq 0$ wächst der Betrag der Lösung exponentiell, sie läuft also exponentiell von $x^* = 0$ weg. Die Menge aller Punkte, die exponentiell weglaufen, heißt *instabile Mannigfaltigkeit* $M_i(x^*)$ von x^*, hier ist das einfach der Unterraum $M_i(x^*) = \langle (1, 0)^T \rangle$.) Umgekehrt konvergieren alle Lösungen zu Anfangswerten der Form $x_0 = (0, x_2)^T$ (also keine Beute) mit $x_2 \in \mathbb{R}$

[3] italienischer Physiker und Mathematiker, 1860–1940
[4] US–amerikanischer Chemiker und Mathematiker, 1880–1949

exponentiell gegen $x^* = 0$, dies ist die sogenannte *stabile Mannigfaltigkeit* $M_s(x^*)$, hier wiederum ein Unterraum, nämlich $M_s(x^*) = \langle (0,1)^T \rangle$.

Auf x^+ trifft keiner der Fälle in Satz A.4 zu, da hier beide Eigenwerte wegen $ca > 0$ offenbar die Realteile 0 besitzen. Wir wissen also, dass Lösungen weder exponentiell konvergieren noch weglaufen können. Was aber passiert statt dessen? Um sich einen Überblick über das Verhalten dieses Systems zu verschaffen, empfiehlt sich hier die numerische Lösung und Darstellung in Kurvenform, die in Abbildung A.3 mit $a = c = 1$ zu sehen ist.

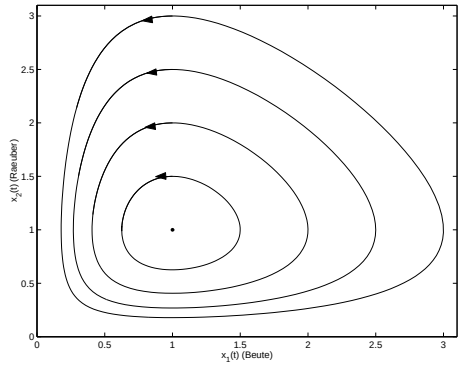

Abbildung A.3: Lösungen des Räuber–Beute Modells (A.7) mit $a = c = 1$

Man erkennt in dieser Grafik gut, warum die Ruhelage $x^+ = (1,1)^T$ weder exponentiell stabil noch exponentiell instabil ist: Alle Lösungen, die nicht auf $M_s(x^*)$ oder $M_i(x^*)$ liegen, laufen auf periodischen Bahnen um dieses x^+ herum, weder konvergieren sie noch laufen sie weg. Formal nennt man eine Lösung $x(t; t_0, x_0)$ *periodisch*, falls ein $T > 0$ existiert, so dass

$$x(t; t_0, x_0) = x(t + T; t_0, x_0)$$

gilt für alle $t \in \mathbb{R}$. Die Zeit $T > 0$ heißt *Periode* der Lösung. (Wir verlangen hier i.A. nicht, dass T die minimale Periode ist.) Beachte, dass eine Lösung $x(t)$ einer autonomen Gleichung genau dann periodisch ist, wenn es zwei Zeiten $t_1 < t_2 \in \mathbb{R}$ gibt, so dass $x(t_1) = x(t_2) =: x_P$ gilt. Aus dieser Gleichheit folgt nämlich sowohl $x(t) = x(t; t_1, x_P)$ als auch $x(t) = x(t; t_2, x_P)$. Aus (1.6) folgt damit $x(t+t_2-t_1) = x(t)$ für alle $t \in \mathbb{R}$, also Periodizität für $T = t_2 - t_1$.

Wir wollen diese numerische Erkenntnis nun mathematisch rigoros beweisen. Dazu betrachten wir den Quotienten

$$\frac{\dot{x}_2(t)}{\dot{x}_1(t)} = \frac{-cx_2(t)(1 - x_1(t))}{ax_1(t)(1 - x_2(t))}.$$

Aus dieser Gleichung folgt

$$ax_1(t)\dot{x}_2(t) - ax_1(t)x_2(t)\dot{x}_2(t) = -cx_2(t)\dot{x}_1(t) + cx_2(t)x_1(t)\dot{x}_1(t)$$

und damit

$$c\dot{x}_1(t) - c\frac{\dot{x}_1(t)}{x_1(t)} + a\dot{x}_2(t) - a\frac{\dot{x}_2(t)}{x_2(t)} = 0.$$

Beachte, dass alle diese Gleichungen nur gelten, wenn alle Nenner ungleich Null sind, also nur für Lösungen $x(t) = (x_1(t), x_2(t))$, die sich in $\mathbb{R}^+ \times \mathbb{R}^+$ befinden und keine Gleichgewichte sind.

Integrieren wir diese Gleichung nun von 0 bis t, so erhalten wir

$$cx_1(t) - c\ln x_1(t) + ax_2(t) - a\ln x_2(t) = k(x(0))$$

mit $k(x(0)) = cx_1(0) - c\ln x_1(0) + ax_2(0) - a\ln x_2(0)$. Die auf $D_V = \mathbb{R}^+ \times \mathbb{R}^+$ definierte Funktion

$$V(x) = cx_1 - c\ln x_1 + ax_2 - a\ln x_2 \tag{A.8}$$

ist also konstant entlang von Lösungen; es gilt

$$V(x(t; x_0)) = V(x_0) \text{ für alle } t \geq 0$$

bzw.

$$\frac{d}{dt}V(x(t; x_0)) = 0.$$

V heißt *erstes Integral* oder auch *Konstante der Bewegung* für (A.7). Die Lösungen von (A.7) mit Anfangswert $x_0 \in D_V$ laufen also entlang der Höhenlinien $V^{-1}(l) := \{x \in D_V \,|\, V(x) = l\}$ von V, die in Abbildung A.4 gemeinsam mit dem Graphen von V skizziert sind. Man sagt, die Höhenlinien $V^{-1}(l)$ sind *invariante Mengen* bezüglich (A.7). Beachte, dass V ein globales Minimum in x^+ mit $V(x^+) = c + a$ besitzt.

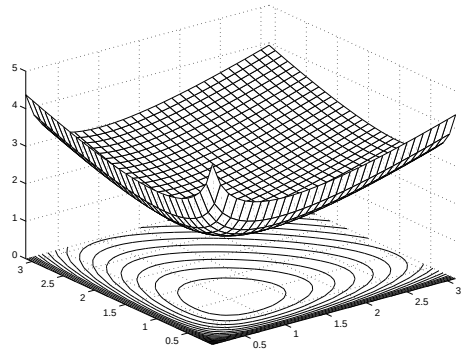

Abbildung A.4: Graph und Höhenlinien von V aus (A.8) mit $a = c = 1$

Dass die Lösungen tatsächlich periodisch sind, folgt aus der Analyse des Vektorfeldes auf den Höhenlinien. Wir betrachten eine Höhenlinie $V^{-1}(l)$ für ein $l > V(x^+)$ und teilen $V^{-1}(l)$ in die vier Segmente

$$\begin{aligned}
S_1 &= \{x \in V^{-1}(l) \,|\, x_1 \leq x_2 \leq 2 - x_1\} \\
S_2 &= \{x \in V^{-1}(l) \,|\, x_2 \leq x_1 \leq 2 - x_2\} \\
S_3 &= \{x \in V^{-1}(l) \,|\, x_1 \geq x_2 \geq 2 - x_1\} \\
S_4 &= \{x \in V^{-1}(l) \,|\, x_2 \geq x_1 \geq 2 - x_2\}
\end{aligned}$$

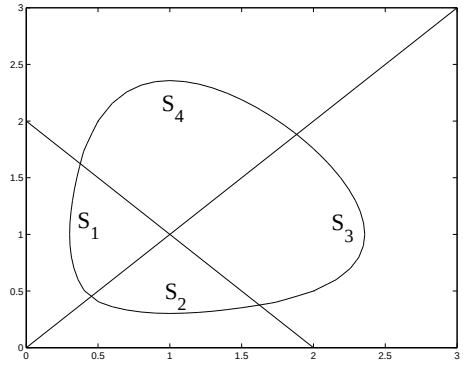

Abbildung A.5: Segmente S_1, S_2, S_3 und S_4

ein, vgl. Abbildung A.5.

Aus der Form der Höhenlinien folgt jetzt, dass ein $\alpha > 0$ existiert, so dass $|x_1 - 1| \geq \alpha$ gilt für alle $x \in S_1$ und $x \in S_3$ und $|x_2 - 1| \geq \alpha$ gilt für alle $x \in S_2$ und alle $x \in S_4$. Desweiteren existiert ein $\beta > 0$ mit $x_1 > \beta$ und $x_2 > \beta$ für alle $x \in V^{-1}(l)$. Aus Gleichung (A.7) folgert man damit die Ungleichungen

$$\begin{array}{rcrl} \dot{x}_2(t) & < & -c\beta\alpha, & \text{falls } x(t) \in S_1 \\ \dot{x}_1(t) & > & a\beta\alpha, & \text{falls } x(t) \in S_2 \\ \dot{x}_2(t) & > & c\beta\alpha, & \text{falls } x(t) \in S_3 \\ \dot{x}_1(t) & < & -a\beta\alpha, & \text{falls } x(t) \in S_4 \end{array}$$

In jedem Sektor ist also eine der beiden Komponenten $x_1(t)$ oder $x_2(t)$ streng monoton wachsend oder fallend mit von 0 (gleichmäßig in t) verschiedener Steigung. Deswegen muss jeder Sektor nach einer endlichen Zeit verlassen werden, und zwar in der Reihenfolge $S_1 \to S_2 \to S_3 \to S_4 \to S_1$. Die Lösung ist also tatsächlich periodisch.

Um die Aussagen des Modells für die modellierten Populationsgrößen zu interpretieren, ist es sinnvoll, eine beispielhafte Lösung in Abhängigkeit von t darzustellen. In Abbildung A.6 ist dies gemacht.

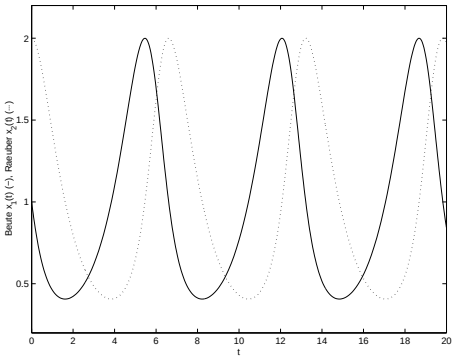

Abbildung A.6: Lösung von (A.7) mit $x_0 = (1, 2)^T$ und $a = c = 1$

Beide Populationen schwanken also periodisch. Wenn (wie am Anfang) viele Räuber und wenig Beute vorhanden sind, nehmen beide Populationen ab. Wenn die Zahl der Räuber unter einer gewissen Schwelle liegt, nimmt die Beutepopulation wieder zu. Wenn genügend Beute vorhanden ist, beginnt auch die Räuberpopulation wieder zuzunehmen und wenn diese eine kritische Marke überschritten hat, nimmt die Zahl der Beute wieder ab, usw. Ein solches Verhalten ist in der Natur durchaus zu beobachten.

A.2.2 Das Räuber–Beute Modell mit beschränkten Ressourcen

Modell (A.6) hat die (unrealistische) Eigenschaft, dass sich die Beutepopulation in Abwesenheit der Räuber gemäß (A.2) verhält, also unbeschränkt wächst. Wir wollen dies durch das realistischere Modell (A.3) ersetzen, das wir hier mit $\mu = \lambda K$ und $e = \lambda$ als

$$\dot{x}(t) = \mu x_1(t) - e x_1(t)^2 \tag{A.9}$$

schreiben. Wir ändern damit die Modellannahme (i) wie folgt ab.

(i') Die Beutepopulation x_1 verhält sich gemäß (A.9) mit $\mu = \gamma - \sigma$ und $e > 0$, wobei e und γ konstant sind und $\sigma = \tilde{\sigma} + bx_2$. Für die Beutetiere gibt es also begrenzte Ressourcen und die Sterberate σ besteht aus einem konstanten Term $\tilde{\sigma} \in (0, \gamma)$ (natürlicher Tod) und einem zu x_2 proportionalen Term bx_2 (Tod durch Räuber). Für $x_2 = 0$ konvergiert die Populationsgröße gegen $K = a/e$ mit $a = \gamma - \tilde{\sigma} > 0$.

Damit erhalten wir die Gleichung

$$\begin{array}{rcl} \dot{x}_1(t) & = & ax_1(t) - bx_1(t)x_2(t) - ex_1(t)^2 \\ \dot{x}_2(t) & = & -cx_2(t) + dx_1(t)x_2(t) \end{array} \tag{A.10}$$

Analog zu (A.6) können wir diese Gleichung durch eine lineare Koordinatentransformation vereinfachen. Hier transformieren wir $x_1 \to \frac{d}{c}x_1$, $x_2 \to \frac{bd}{da-ec}x_2$ und erhalten so

$$\begin{array}{rcl} \dot{x}_1(t) & = & \alpha x_1(t)(1 - x_2(t)) + \beta x_1(t)(1 - x_1(t)) \\ \dot{x}_2(t) & = & -cx_2(t)(1 - x_1(t)) \end{array} \tag{A.11}$$

mit $\alpha = a - ec/d$ und $\beta = ec/d$. Hier muss man aufpassen, dass bei dieser Transformation positive x_1, x_2 wieder auf positive x_1, x_2 abgebildet werden. Da $a, b, c, d, e > 0$ sind, ist dies genau dann der Fall, wenn $\frac{bd}{da-ec} > 0$ ist, also wenn $ad > ec$ gilt. Wir wollen uns auf diesen Fall einschränken, nicht nur aus formalen Gründen, sondern auch aus Modellierungsgründen: Für $ad \leq ec$ kann man zeigen, dass die Räuberpopulation für $t \to \infty$ für alle Anfangswerte ausstirbt, wir wollen hier aber den Fall der langfristigen Koexistenz beider Arten betrachten, für den $ad > ec$ eine notwendige Bedingung ist.

Als Gleichgewichte erhält man hier $x^* = (0,0)^T$, $x^{**} = ((\alpha+\beta)/\beta, 0)^T$ und $x^+ = (1,1)^T$. Nur x^+ liegt in $\mathbb{R}^+ \times \mathbb{R}^+$, weswegen wir dieses Gleichgewicht genauer untersuchen wollen.

Die Linearisierung ergibt sich zu

$$Df(x) = \begin{pmatrix} \alpha(1 - x_2) + \beta(1 - 2x_1) & -\alpha x_1 \\ cx_2 & -c(1 - x_1) \end{pmatrix}$$

also

$$Df(x^+) = \begin{pmatrix} -\beta & -\alpha \\ c & 0 \end{pmatrix}.$$

Die Eigenwerte dieser Matrix sind

$$\lambda_{1/2} = -\frac{\beta}{2} \pm \sqrt{\frac{\beta^2}{4} - \alpha c}.$$

Falls die Wurzel komplex ist, sind die Realteile $-\beta/2$ negativ, falls die Wurzel reell ist, sind auch $\lambda_{1/2}$ reell und es gilt

$$\lambda_{1/2} \leq -\frac{\beta}{2} + \sqrt{\frac{\beta^2}{4} - \alpha c} < -\frac{\beta}{2} + \sqrt{\frac{\beta^2}{4}} = 0,$$

also erhalten wir in beiden Fällen negative Realteile, weswegen x^+ lokal exponentiell stabil ist. Wir wissen also insbesondere, dass es eine Umgebung von x^+ gibt, so dass alle Lösungen in dieser Umgebung gegen x^+ konvergieren. Was ist aber nun der Einzugsbereich $\mathcal{D}(x^+)$?

Dieser lässt sich hier analytisch ermitteln: Betrachte dazu das erste Integral (A.8)

$$V(x) = cx_1 - c\ln x_1 + \alpha x_2 - \alpha \ln x_2.$$

Im Gegensatz zu (A.7) ist diese Funktion für (A.11) nicht mehr konstant entlang von Lösungen, statt dessen gilt für jede Lösung $x(t)$ in D_V die Gleichung

$$\begin{aligned} \frac{d}{dt}V(x(t)) &= c\dot{x}_1(t) - c\frac{\dot{x}_1(t)}{x_1(t)} + \alpha\dot{x}_2(t) - \alpha\frac{\dot{x}_2(t)}{x_2(t)} \\ &= \Big(c\alpha x_1(t)(1-x_2(t)) + c\beta x_1(t)(1-x_1(t))\Big)\left(1 - \frac{1}{x_1(t)}\right) \\ &\quad - \alpha c x_2(t)(1-x_1(t))\left(1 - \frac{1}{x_2(t)}\right) \\ &= c\beta(1-x_1(t))(x_1(t)-1) = -c\beta(x_1(t)-1)^2 \end{aligned}$$

Die Funktion $V(x(t))$ fällt also monoton in t, für $x_1(t) \neq 1$ sogar streng monoton. Beachte, dass $V(x)$ in $x = x^+$ ein globales Minimum besitzt; weitere lokale Minima existieren nicht. Eine solche Funktion wird in der Stabilitätstheorie auch *Lyapunovfunktion*[5] genannt. Hier haben wir den Sonderfall einer *semidefiniten* Lyapunovfunktion, da die Ableitung entlang der Lösungen nicht strikt kleiner als Null ist (wie meist für eine Lyapunovfunktion verlangt) sondern nur ≤ 0.

Wir beweisen nun $x(t) \to x^+$ für $t \to \infty$. Da $V(x(t))$ monoton fällt und nach unten beschränkt ist, konvergiert $V(x(t))$ gegen einen Wert V_∞. Ähnlich wie im Beweis von Satz A.2 sieht man nun, dass $\frac{d}{dt}V(x(t)) \to 0$ für $t \to \infty$ gilt, also muss $x_1(t) \to 1$ für $t \to \infty$ gelten. Dies ist aber nur dann möglich, falls $x_2(t) \to 1$ konvergiert: Wäre $|x_2(t) - 1| \geq \delta$ so würde aus (A.11) für $x_1(t)$ in einer Umgebung der 1 entweder $\dot{x}_1(t) > \varepsilon$ oder $\dot{x}_1(t) < -\varepsilon$ folgen, was der Konvergenz $x_1 \to 1$ widersprechen würde. Also gilt $x_2(t) \to 1$ und damit $x(t) \to x^+$. Alle Lösungen mit Anfangswerten in D_V konvergieren also gegen x^+, weswegen

[5] A.M. Lyapunov, russischer Mathematiker, 1857–1918

$\mathcal{D}(x^+) = \mathbb{R}^+ \times \mathbb{R}^+$ ist. Die numerischen Ergebnisse in Abbildung A.7 bestätigen dieses Ergebnis.

Die Argumentation, die wir hier verwendet haben, ist als *Lasalles Invarianzprinzip* bekannt und lässt sich auch allgemein als Satz formulieren, was wir hier aber nicht vertiefen wollen.

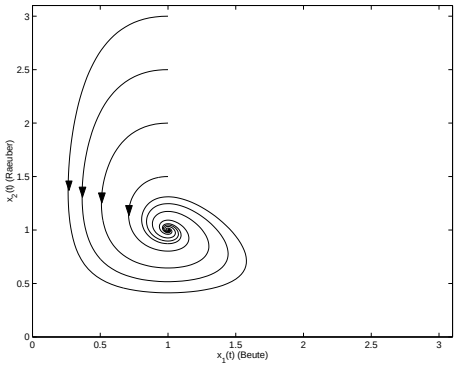

Abbildung A.7: Lösungen des Räuber–Beute Modells (A.11) mit $a = c = 1$, $\beta = 0.5$

Zur Interpretation des Modells ist wieder die Darstellung einer Lösung in Abhängigkeit von t nützlich, wie sie in Abbildung A.8 gegeben ist.

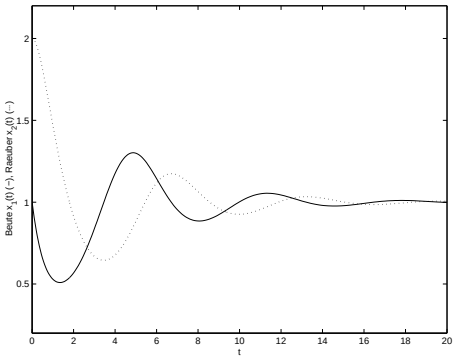

Abbildung A.8: Lösung von (A.11) mit $x_0 = (1, 2)^T$ und $a = c = 1$, $\beta = 0.5$

Die Lösung zeigt zwar ein ähnliches Schwankungsverhalten wie in Abbildung A.6, konvergiert aber mit wachsendem t gegen ein Gleichgewicht, Solche Gleichgewicht verschiedener gemeinsam existierender Arten sind in realen Ökosystemen oft zu beobachten, ebenso wie die charakteristischen Schwankungen, die auftreten, wenn das System durch äußere Einflüsse “aus dem Gleichgewicht” gebracht wurde.

Auf eine abschließende Diskussion dieses Modells können wir hier verzichten, da hier exakt die gleichen Unzulänglichkeiten wie in der Diskussion im Abschnitt A.1.4 bestehen, mit Ausnahme des letzten Punktes natürlich.

A.2.3 Verallgemeinerung auf n Arten

Wir wollen in diesem Abschnitt abschließend auf die Verallgemeinerung des Modells (A.10) auf n verschiedene Arten $x_1, \ldots, x_n$ eingehen. Wenn wir für alle Arten die gleichen Modellannahmen treffen, nämlich, dass die Dynamik durch (A.9) gegeben ist, wobei die jeweilige Wachstumsrate μ affin linear von allen anderen Arten abhängt, so erhalten wir das Modell

$$\dot{x}_i(t) = k_i x_i(t) + b_i^{-1} \sum_{j=1}^{n} a_{ij} x_i(t) x_j(t), \qquad i = 1, \ldots, n. \tag{A.12}$$

mit $k_i \neq 0$, $a_{ii} \leq 0$, $a_{ij} \in \mathbb{R}$ und $b_i > 0$. Wir definieren mittels der a_{ij} die Matrix $A = (a_{ij})$. Der Koeffizient a_{ii} entspricht für jede Art gerade dem e aus (A.9), er modelliert also die Ressourcenbeschränkung, während die a_{ij} für $i \neq j$ die Interaktion der Arten beschreibt. Für Beute x_i und Räuber x_j muss die Bedingung $a_{ij} < 0$ und $a_{ji} > 0$ gelten. Die etwas seltsam anmutende Notation mit b_i^{-1} ergibt sich aus der ursprünglichen, etwas anderen Schreibweise des Modells. Beachte, dass die Modelle (A.6) und (A.10) Spezialfälle dieses Modells sind.

Der Spezialfall $a_{ii} = 0$ und $a_{ij} = -a_{ji}$ wird als *Volterra–Ökologie* bezeichnet. In diesem Fall ist die Matrix $A = (a_{ij})$ antisymmetrisch, d.h. $x^T A x = 0$ für alle $x \in \mathbb{R}^n$.

Wenn wir nach Gleichgewichten x^+ suchen, für die alle Arten koexistieren, so gilt $x_i^+ > 0$, also

$$k_i x_i^+ + b_i^{-1} \sum_{j=1}^{n} a_{ij} x_i^+ x_j^+ = 0 \quad \Rightarrow \quad b_i k_i + \sum_{j=1}^{n} a_{ij} x_j^+ = 0, \tag{A.13}$$

diese Gleichgewichte sind also als Lösungen eines linearen Gleichugssystems gegeben. Wenn A invertierbar ist, existiert also höchstens ein solches Gleichgewicht: es gibt genau eine Lösung x^* des linearen Gleichungssystems, für die aber nicht $x_i^* > 0$ gelten muss.

Die Konstruktion des ersten Integrals V lässt sich auf dieses Modell verallgemeinern. Wenn ein Gleichgewicht x^+ mit $x_i^+ > 0$ für $i = 1, \ldots, n$ existiert, so kann man nachrechnen, dass die Funktion

$$V(x) = \sum_{i=1}^{n} b_i (x_i - x_i^+ \ln x_i) \tag{A.14}$$

die Gleichung

$$\frac{d}{dt} V(x(t)) = (x(t) - x^+)^T A (x(t) - x^+)$$

erfüllt. Falls A negativ semidefinit ist, so ist diese Ableitung ≤ 0 und wir können die obige Argumentation auf das n–dimensionale Modell übertragen. Im Falle einer Volterra–Ökologie ist A antisymmetrisch, weswegen $\frac{d}{dt} V(x(t)) = 0$ ist. Hier erhalten wir also wieder das Phänomen periodischer Lösungen.

A.3 Anwendungen der Populationsdynamik

A.3.1 Auswirkungen der Befischung

Die Volterra–Ökologie und speziell das Lotka–Volterra Modell gilt in der Biologie i.A. als zu stark vereinfacht, da hierbei in dem sowieso schon einfachen Modell (A.12) noch weitere

strukturelle Vereinfachungen gemacht werden. Man muss aber berücksichtigen, dass dieses Modell zur Erklärung eines speziellen Sachverhaltes entwickelt wurde, für den es tatsächlich gut funktioniert. Wir wollen diese Anwendung nun erläutern.

In den 1920er Jahren beobachtete der italienische Biologe D'Ancona, dass der Anteil der Raubfische (Haie, Rochen, ...) am Gesamtfischfang während des 1. Weltkrieges im Mittelmeer deutlich höher als vorher und nachher war. Im Hafen Fiume in Italien wurden die folgenden Anteile der Raubfische am Gesamtfang festgestellt:

Jahr	1914	1915	1916	1917	1918
Raubfischanteil	11,9%	21,4%	22,1%	21,2%	36,4%
Jahr	1919	1920	1921	1922	1923
Raubfischanteil	27,3%	16,0%	15,9%	14,8%	10,7%

Natürlich war D'Ancona klar, dass während des Krieges weniger gefischt wurde, aber warum sollte dies die Raubfische mehr begünstigen?

Das Volterra–Modell wurde zur Erklärung dieses Phänomens entwickelt. Tatsächlich handelt es sich hier nur um zwei (Gruppen von) Arten, so dass sich (A.12) zu (A.6) vereinfacht, wenn man $a = k_1$, $c = -k_2$, $b = -b_1^{-1}a_{12}$, $d = b_2^{-1}a_{21} = -b_2^{-1}a_{12}$ setzt (dies zeigt insbesondere, dass (A.6) ein Spezialfall der Volterra–Ökologie ist). Wie kann dieses Modell mit den bekannten periodischen Lösungen aus den Abbildungen A.3 und A.6 das Phänomen beschreiben? Die Werte in der obigen Tabelle sind Jahresmittelwerte, weswegen es sich anbietet, auch die vom Modell gegebenen Werte zu mitteln. Hier gilt das folgende Lemma.

Lemma A.6 Sei $x(t)$ eine periodische Lösung von (A.6) mit Periode T. Dann gilt

$$\overline{x}_1 := \frac{1}{T}\int_0^T x_1(t)dt = \frac{c}{d} \quad \text{und} \quad \overline{x}_2 := \frac{1}{T}\int_0^T x_2(t)dt = \frac{a}{b}.$$

Beweis: Es gilt

$$\frac{\dot{x}_1(t)}{x_1(t)} = a - bx_2(t).$$

Für diesen Ausdruck gilt

$$\frac{1}{T}\int_0^T \frac{\dot{x}_1(t)}{x_1(t)}dt = \frac{1}{T}\int_0^T a - bx_2(t)dt.$$

Andererseits gilt

$$\frac{1}{T}\int_0^T \underbrace{\frac{\dot{x}_1(t)}{x_1(t)}}_{=\frac{d}{dt}\ln x_1(t)} dt = \frac{1}{T}(\ln x(T) - \ln x(0)) = 0,$$

da die Lösung periodisch mit $x(T) = x(0)$ ist. Also folgt

$$0 = \frac{1}{T}\int_0^T a - bx_2(t)dt = a - b\overline{x}_2$$

und damit die Behauptung für $\overline{x}_2$. Analog berechnet man den Wert für $\overline{x}_1$. □

Die zunächst vielleicht etwas überraschende Erkenntnis dieses Lemmas ist, dass die Mittelwerte über eine Periode nicht vom Anfangswert abhängen. Der Anteil der Raubfische an der Gesamtmenge ist im Mittel also gegeben durch

$$\overline{x}_2^A = \frac{\overline{x}_2}{\overline{x}_1 + \overline{x}_2} = \frac{\frac{a}{b}}{\frac{a}{b} + \frac{c}{d}} = \frac{ad}{ad + cb}$$

Um die veränderten Anteile während des 1. Weltkrieges zu erklären, müssen wir den Fischfang in (A.6) berücksichtigen. Nimmt man hier proportionale Fangraten px_1 und px_2 an, so ergibt sich das Modell mit Fischfang zu

$$\begin{aligned} \dot{x}_1(t) &= (a-p)x_1(t) - bx_1(t)x_2(t) \\ \dot{x}_2(t) &= -(c+p)x_2(t) + dx_1(t)x_2(t) \end{aligned}$$

Der mittlere Raubfischanteil bei Fangrate p ist demnach

$$\overline{x}_2^A(p) = \frac{(a-p)d}{(a-p)d + (c+p)b}$$

oder als Kehrwert ausgedrückt

$$\overline{x}_2^A(p)^{-1} = \frac{(a-p)d + (c+p)b}{(a-p)d} = \frac{(c+p)b}{(a-p)d} + 1.$$

Wenn also die Fangrate p abnimmt, so verringert sich der Bruch $\frac{(c+p)b}{(a-p)d}$ ebenfalls (der Zähler wird kleiner und der Nenner größer), womit auch $\overline{x}_2^A(p)^{-1}$ kleiner wird, weswegen der Raubfischanteil $\overline{x}_2^A(p)$ zunimmt. Das Modell liefert also eine Erklärung dafür, warum bei geringerer Befischung der Raubfischanteil zunimmt.

A.3.2 Selektion gleichartiger Spezies

Ähnlich wie wir das beim Räuber–Beute Modell (A.10) gemacht haben, wollen wir hier wieder die Interaktion zweier Populationen beschreiben, die dem logistischen Wachstum (A.3) unterliegen. Diesmal wollen wir aber nicht Räuber und Beute beschreiben, sondern zwei friedlich koexistierende Arten x_1 und x_2 modellieren, die teilweise um die gleichen Ressourcen konkurrieren.

Beide Arten sollen also durch die Gleichung

$$\dot{x}_i(t) = \lambda_i x(t)(K_i - x_i(t))$$

beschrieben werden. Um die Konkurenz zu modellieren, ersetzen wir K_i durch $K_i - m_i$ und machen dazu die Modellannahme, dass m_1 der Anteil der Ressourcen K_1 von x_1 ist, der von x_2 in Anspruch genommen wird, und umgekehrt. Die einfachste Art, dies zu modellieren, ist die Wahl

$$m_1 = \alpha x_2 \quad \text{und} \quad m_2 = \beta x_1.$$

Das Gesamtmodell ergibt sich so zu

$$\begin{aligned} \dot{x}_1(t) &= \lambda_1 x_1(t)(K_1 - x_1(t) - \alpha x_2(t)) \\ \dot{x}_2(t) &= \lambda_2 x_2(t)(K_2 - x_2(t) - \beta x_1(t)) \end{aligned} \tag{A.15}$$

Dieses Modell ist wieder ein Spezialfall des Modells (A.12), mit $k_i = \lambda_i K_i$, $b_i = 1$, $a_{ii} = -\lambda_i$, und $a_{12} = -\lambda_1\alpha$ und $a_{21} = -\lambda_2\beta$.

Wir analysieren hier zunächst den Fall $\alpha = \beta = 1$ und betrachten, wie das Verhalten der Lösungen von K_1 und K_2 abhängt. Beachte zunächst, dass Lösungen, die auf der x_1– bzw. x_2–Achse starten, für alle Zeiten dort bleiben. Umgekehrt bedeutet dies, da sich Lösungen nicht schneiden können, dass Lösungen in $\mathbb{R}^+ \times \mathbb{R}^+$ für alle Zeiten in $\mathbb{R}^+ \times \mathbb{R}^+$ bleiben. Eine Art kann also nicht in endlicher Zeit aussterben, dies kann aber durchaus für $t \to \infty$ passieren. Genau dieser Fall tritt hier ein; es gilt:

Falls $\alpha = \beta = 1$ und $K_1 > K_2$ ist, so konvergiert jede Lösung $x(t; t_0, x_0)$ mit $x_0 \in \mathbb{R}^+ \times \mathbb{R}^+$ gegen $x^* = (K_1, 0)^T$ für $t \to \infty$. Mit anderen Worten: Unabhängig von den Wachstumsraten λ_i überlebt nur die Art mit den größeren Ressourcen K_i (denn für $K_2 > K_1$ ergibt sich wegen der Symmetrie des Modells gerade das umgekehrte Verhalten).

Wir wollen den Beweis dieser Behauptung skizzieren. Zunächst rechnet man nach, dass die Gleichung (A.15) genau die Gleichgewichte $(0,0)^T$, $(K_1,0)^T$ und $(0,K_2)^T$ besitzt. Da kein Gleichgewicht $x^+ \in \mathbb{R}^+ \times \mathbb{R}^+$ existiert, können wir V aus (A.14) nicht zur Analyse verwenden, wir müssen das Modell also direkt analysieren.

Dazu teilt man den positiven Quadranten $\mathbb{R}^+ \times \mathbb{R}^+$ auf in die drei Bereiche

$$\begin{aligned} A &:= \{(x_1, x_2) \,|\, x_1 > 0,\, x_2 > 0,\, K_2 \geq x_1 + x_2\} \\ B &:= \{(x_1, x_2) \,|\, x_1 > 0,\, x_2 > 0,\, K_2 \leq x_1 + x_2 \leq K_1\} \\ C &:= \{(x_1, x_2) \,|\, x_1 > 0,\, x_2 > 0,\, K_1 \leq x_1 + x_2\}, \end{aligned}$$

vgl. Abbildung A.9.

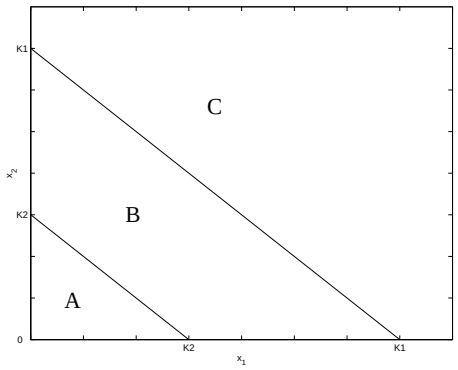

Abbildung A.9: Bereiche von $\mathbb{R}^+ \times \mathbb{R}^+$

Nun unterscheidet man die folgenden Fälle:

(1) In int A ist $\dot{x}_1(t) > 0$ und $\dot{x}_2(t) > 0$, die Lösungen wachsen also in beiden Komponenten streng monoton. Für $x_0 \in \mathbb{R}^+ \times \mathbb{R}^+$ muss die Lösung also entweder nach int B laufen, oder

gegen ein Gleichgewicht $x^* \in A$ konvergieren, für das $x_1^* > x_{0,1} > 0$ und $x_2^* > x_{0,1} > 0$ gilt. Da ein solches Gleichgewicht nicht existiert, müssen die Lösungen also nach $\operatorname{int} B$ laufen.

(2) In $\operatorname{int} C$ gilt $\dot{x}_1(t) < 0$ und $\dot{x}_2(t) < 0$, die Lösungen fallen also in beiden Komponenten streng monoton. Folglich muss die Lösung hier entweder nach $\operatorname{int} B$ laufen oder gegen ein Gleichgewicht $x^* \in C$ konvergieren. Dies kann nur $x^* = (K_1, 0)^T$ sein, so dass in diesem Fall die Behauptung gezeigt ist.

(3) In $\operatorname{int} B$ gilt $\dot{x}_1(t) > 0$ und $\dot{x}_2(t) < 0$, $x_1(t)$ wächst und $x_2(t)$ fällt also streng monoton. Man rechnet nach, dass eine Lösung $x(t)$, die für ein t^* in $\operatorname{int} B$ liegt, für alle zukünftigen Zeiten $t \geq t^*$ auch in $\operatorname{int} B$ liegt (dies leitet man aus den Nullstellen der Ableitungen $\dot{x}_1$ bzw $\dot{x}_2$ an den Übergängen zwischen den Mengen ab). Also müssen beide Komponenten $x_1(t)$ und $x_2(t)$ konvergieren, weswegen die Gesamtlösung auch konvergieren muss, und zwar gegen ein Gleichgewicht $x^* \in B$. Da $x_1(t)$ wächst und $x_2(t)$ fällt, kann dies nur $x^* = (K_1, 0)^T$ sein.

Abbildung A.10 zeigt verschiedene numerisch berechnete Trajektorien von (A.15), die diese Analyse bestätigen.

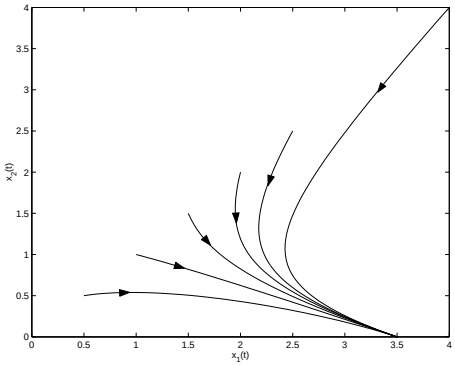

Abbildung A.10: Lösungen von (A.15) mit $\alpha = \beta = 1$, $\lambda_1 = \lambda_2 = 1$, $K_1 = 3.5$, $K_2 = 1.5$

Die Situation ändert sich deutlich, wenn wir die Annahme $\alpha = \beta = 1$ wegfallen lassen. Wenn also die Art x_1 z.B. auf andere Nahrungsressourcen ausweichen kann, die von x_2 nicht beansprucht werden, so würde sich K_1 vergrößern und β verkleinern. Tatsächlich reicht es aus, β zu verkleinern, um eine langfristige Koexistenz der Arten zu erreichen. Für $\beta < K_2/K_1 < 1$ liegt das mittels (A.13) errechnete Gleichgewicht

$$x^+ = \begin{pmatrix} \frac{K_1 - \alpha K_2}{1 - \alpha\beta} \\ \frac{K_2 - \beta K_1}{1 - \alpha\beta} \end{pmatrix}$$

in $\mathbb{R}^+ \times \mathbb{R}^+$. Zudem ist $A = (a_{ij})$ negativ definit, so dass V aus (A.14) entlang aller Lösungen streng monoton fällt, weswegen alle Lösungen in $\mathbb{R}^+ \times \mathbb{R}^+$ gegen x^+ konvergieren müssen. Mit $\beta = 0.1$ ergeben sich die in Abbildung A.11 dargestellten Lösungen, die dieses Verhalten bestätigen.

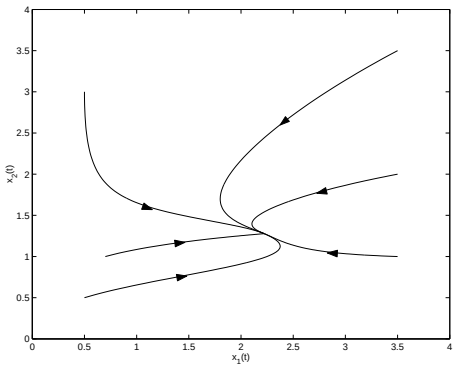

Abbildung A.11: Lösungen von (A.15) mit $\alpha = 1$, $\beta = 0.1$, $\lambda_1 = \lambda_2 = 1$, $K_1 = 3.5$, $K_2 = 1.5$

A.3.3 Der Chemostat

Eine konkrete technische Anwendung von Räuber–Beute–Modellen ist der sogenannte *Chemostat*, eine Apparatur zur Züchtung von Mikroorganismen, die sowohl in der Forschung als auch der technischen Anwendung eine Rolle spielt, z.B. bei der Herstellung von Insulin. Schematisch besteht ein Chemostat aus drei Gefäßen, vgl. Abbildung A.12: Ein Vorratsgefäß, in dem eine Nährlösung bereitgestellt wird, der eigentliche Chemostat, in dem sich die Mikroorganismen befinden und ein Auffanggefäß, in dem die entstehenden Organismen gesammelt werden.Im eigentlichen Chemostat wird dabei durch Rühren sicher gestellt, dass die enthaltenen Organismen und Nährstoffe homogen verteilt sind.

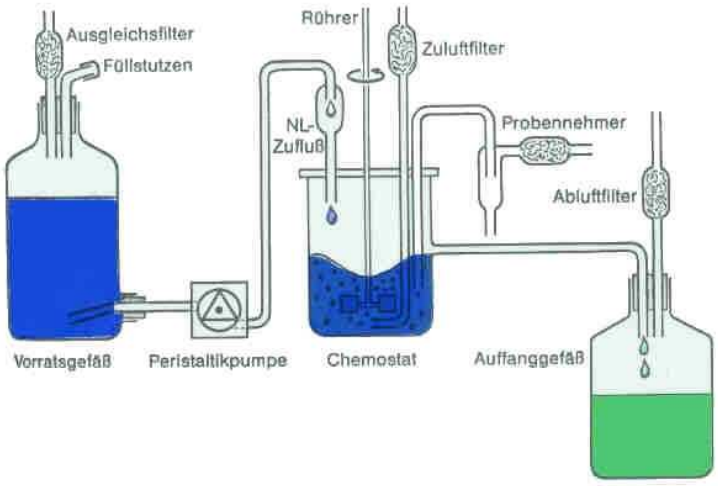

Abbildung A.12: Chemostat, vgl. `www.wb.fh-heilbronn.de/test/bionet/6_4.html`

Für den einfachsten Fall mit einer Art Mikroorganismen ist die Idee der Modellierung nun relativ einfach: Wir modellieren die Nährlösung als Beute S und die Mikroorganismen x_1 als Räuber. Hierbei ergeben sich allerdings einige Änderungen gegenüber unseren bisherigen

Modellen, die wir nun diskutieren werden.

Für die Nährlösung S entsprechen die "Geburten" nun der Menge der Zufuhr aus dem Vorratsbehälter. Im Gegensatz zu unserem bisherigen Modell hängt diese Größe nun aber nicht von der Anzahl der bereits vorhandenen Nährlösung S ab, sie wird daher durch einen konstanten positiven Term $k \cdot D > 0$ modelliert, der sich aus der Konzentration k der Lösung und der Menge der eingeleiteten Lösung D (Durchflussrate) ergibt. Die "Sterbefälle" setzen sich aus zwei Komponenten zusammen, nämlich aus dem Anteil der Nährstoffe, die in den Auffangbehälter gespält werden — dieser Anteil ist gerade gleich DS — und dem Anteil, der von den Mikroorganismen als Nahrung aufgenommen wird. Aus experimentellen Daten hat sich herausgestellt, dass der dafür bisher verwendete Term bSx_1 die experimentelle Realität nicht gut genug beschreibt. Für große Mengen an Nährlösung $S >> 1$ steigt die Aufnahme nämlich nicht proportional zur Nahrungsmenge S, weil die Organismen natürlich nicht beliebig viel Nahrung aufnehmen können, selbst wenn diese zur Verfügung steht. Als realistischer hat sich hier ein Term der Form

$$\frac{mS}{a+S}\frac{x_1}{\gamma}$$

herausgestellt. Insgesamt kommen wir damit auf die Gleichung

$$\dot{S}(t) = (k - S(t))D - \frac{mS(t)}{a+S(t)}\frac{x_1(t)}{\gamma}.$$

Die Population x_1 verhält sich nun wie im klassischen Lotka–Volterra Modell mit dem Unterschied, dass der von x_1 abhängige Term in der Wachstumsrate gleich $\frac{mS(t)}{a+S(t)}$ gewählt wird, was bewirkt, dass die Wachstumsrate bei sehr großem Nahrungsangebot nicht ins Unendliche steigt. Wir erhalten also

$$\dot{x}_1(t) = x_1(t)\left(\frac{mS(t)}{a+S(t)} - D\right).$$

Beachte, dass die Sterberate hier nicht dem natürlichen Tod entspricht (dieser taucht im Modell nicht auf), sondern dem Anteil der Organismen, die durch die nachströmende Flüssigkeit in das Auffanggefäß gespült werden. Durch die Koordinatentransformation $S \to \frac{S}{k}$ und $x_1 \to \frac{x_1}{k\gamma}$, die Parameter-Skalierung $m \to \frac{m}{D}$ und $a \to \frac{a}{K}$ sowie die Wahl einer geeigneten Zeiteinheit $t \to tD^{-1}$ vereinfacht sich das Modell zu dem normierten Chemostat–Modell

$$\begin{aligned} \dot{S}(t) &= (1 - S(t)) - \tfrac{mS(t)}{a+S(t)}x_1(t) \\ \dot{x}_1(t) &= x_1(t)\left(\tfrac{mS(t)}{a+S(t)} - 1\right) \end{aligned} \tag{A.16}$$

Für dieses erste einfache Modell ergeben sich die Gleichgewichte

$$x^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{und} \quad x^+ = \begin{pmatrix} \frac{a}{m-1} \\ \frac{m-1-a}{m-1} \end{pmatrix}$$

Beachte, dass das Gleichgewicht x^+ für $a > 0$ nur für $m > 1 + a$ im positiven Quadranten liegt. Die Eigenwerte der Jacobi–Matrix in x^+ sind $\lambda_1 = -1$ und $\lambda_2 = (am - m^2 + 2m - a - 1)/(am)$. Man rechnet nach, dass diese für $m > a + 1$ und $a > 0$ negativ sind, das

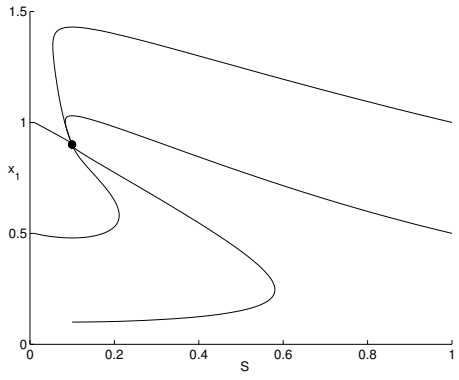

Abbildung A.13: Lösungen von (A.16) mit $a = 0.1$, $m = 2$

Gleichgewicht x^+ ist also lokal exponentiell stabil. Abbildung A.13 zeigt einige ausgewählte numerische Lösungen mit $a = 0.1$, $m = 2$

Die in Abbildung A.13 dargestellten Lösungen legen nahe, dass der Einzugsbereich $\mathcal{D}(x^+)$ tatsächlich der ganze positive Quadrant ist. Ein rigoroser Nachweis dieser Eigenschaft kann ähnlich wie in Abschnitt A.2.2 mittels einer geeigneten Lyapunovfunktion V durchgeführt werden.

Das Modell (A.16) lässt sich auf d Mikroorganismenkulturen $x_1, \ldots, x_d$ verallgemeinern, indem man weitere Gleichungen der gleichen Struktur hinzugefügt und je nach den Abhängigkeiten entsprechende Kopplungsterme hinzugügt. Stellt z.B. x_j Nahrung von x_i dar, so fügt man zu den Gleichungen von x_j und x_i die Terme

$$\pm x_j(t) \frac{m_i x_i(t)}{a_i + x_j(t)}$$

hinzu, mit Vorzeichen '$-$' für x_j und Vorzeichen '$+$' für x_i.

Beispielsweise ist ein Modell für $d = 3$ Kulturen gegeben durch

$$\begin{aligned}
\dot{S}(t) &= (1 - S(t)) - \frac{m_1 S(t)}{a_1 + S(t)} x_1(t) \\
\dot{x}_1(t) &= x_1(t) \left(\frac{m_1 S(t)}{a_1 + S(t)} - 1 - \frac{m_2 x_2(t)}{a_2 + x_1(t)} \right) \\
\dot{x}_2(t) &= x_2(t) \left(\frac{m_2 x_1(t)}{a_2 + x_1(t)} - 1 - \frac{m_3 x_3(t)}{a_3 + x_2(t)} \right) \\
\dot{x}_3(t) &= x_3(t) \left(\frac{m_3 x_2(t)}{a_3 + x_2(t)} - 1 \right)
\end{aligned} \tag{A.17}$$

In diesem Beispiel stellt S Nahrung für x_1 dar, während x_1 Nahrung für x_2 und x_2 wiederum Nahrung für x_3 ist.

Durch die geschickte Skalierung der Parameter ergibt sich eine interessante Eigenschaft des Modells, die man zur Vereinfachung der entstehenden Gleichungen ausnutzt. Definieren wir

die Variable $\Sigma(t) = 1 - S(t) - \sum_{k=1}^{d} x_k(t)$, so sieht man, dass für diese die Differentialgleichung

$$\dot{\Sigma}(t) = -\Sigma(t)$$

gilt, da sich die Kopplungsterme gerade gegenseitig aufheben. Es gilt also

$$\Sigma(t) = e^{-t}\Sigma(0).$$

Mit anderen Worten konvergieren alle Lösungen $(S(t), x_1(t), \ldots, x_d(t))^T$ gegen die Menge

$$\Omega = \{(S, x_1, \ldots, x_d)^T \in \mathbb{R}^{d+1} \mid S + \sum_{k=1}^{d} x_k(t) = 1\}.$$

Diese Menge wird *Omega–Limesmenge* des Systems genannt. Wenn wir also am Langzeitverhalten der Lösungen interessiert sind, genügt es die Lösungen auf Ω zu betrachten, da sich Lösungen in der Nähe von Ω aus Stetigkeitsgründen wie Lösungen auf Ω verhalten[6].

Die Gleichungen auf Ω erhält man nun einfach, indem man $S = 1 - \sum_{k=1}^{d} x_k(t)$ setzt und diese Größe in die Gleichungen für $x_1, \ldots, x_d$ einsetzt. Für unser einfaches Modell (A.16) ergibt sich damit

$$\dot{x}_1(t) = x_1(t)\left(\frac{m(1-x_1(t))}{a+1-x_1(t)} - 1\right) = x_1(t)\left(\frac{m-1}{1+a-x_1(t)}\right)(1-\lambda-x_1(t)) \qquad \text{(A.18)}$$

mit $\lambda = \frac{a}{m-1}$. Die zweite Form der Gleichung ist deswegen nützlich, da man hier die Gleichgewichte $x_1^* = 0$ und $x_1^+ = 1 - \lambda$ sofort ablesen kann. Tatsächlich stimmen diese mit den oben bestimmten Gleichgewichten überein.

Für das Modell mit drei Organismen ergibt sich

$$\begin{aligned}
\dot{x}_1(t) &= x_1(t)\left(\frac{m_1(1-x_1(t)-x_2(t)-x_3(t))}{a_1+1-x_1(t)-x_2(t)-x_3(t)} - 1 - \frac{m_2 x_2(t)}{a_2+x_1(t)}\right) \\
\dot{x}_2(t) &= x_2(t)\left(\frac{m_2 x_1(t)}{a_2+x_1(t)} - 1 - \frac{m_3 x_3(t)}{a_3+x_2(t)}\right) \qquad \text{(A.19)} \\
\dot{x}_3(t) &= x_3(t)\left(\frac{m_3 x_2(t)}{a_3+x_2(t)} - 1\right)
\end{aligned}$$

In diesem Modell stellt sich heraus, dass die Gleichgewichte leider keine wesentlichen Informationen über das Langzeitverhalten des Systems liefern, da sie antistabil sind und damit keine möglichen Grenzwerte sind. Tatsächlich existieren in diesem Modell kompliziertere Grenzlösungen, gegen die die Lösungen aus einer Umgebung streben. Eine theoretische Analyse ist hier zwar ebenfalls möglich, erfordert allerdings tiefliegende Resultate aus der Theorie der dynamischen Systeme, die wir hier nicht behandeln können. Wir begnügen uns daher mit numerischen Ergebnissen, und zwar für die Parameter $m_1 = 10$, $a_1 = 0.08$, $m_2 = 4.0$, $a_2 = 0.23$, $m_3 = 3.5$ und a_3 zwischen 0.2 und 0.4. Abbildung A.14 zeigt die zugehörigen Lösungen.

[6]Obwohl diese Eigenschaft intuitiv anschaulich ist, ist der formale Beweis nicht trivial und nur unter geeigneten Annahmen an die Lösungen erfüllt.

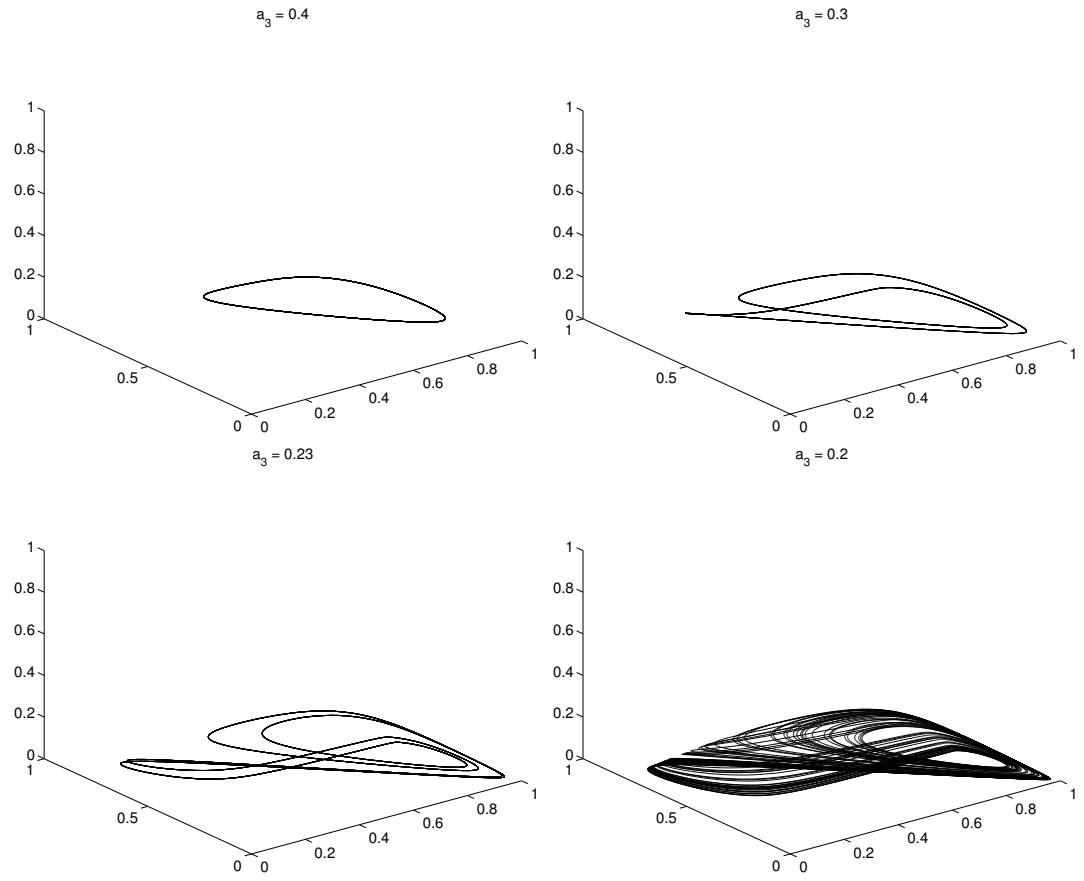

Abbildung A.14: Lösungen von (A.19) mit verschiedenen Werten von a_3

Man sieht, das mit kleiner werdendem a_3 die Perioden der Lösungen immer länger werden, man spricht von *Periodenverdopplung*. Tatsächlich ist für $a = 0.2$ keine Periodizität mehr feststellbar, die Lösung zeigt scheinbar unvorhersehbare Oszillationen. Man spricht hier von chaotischem Verhalten oder kurz Chaos. In allen vier Fällen ist es so, dass Lösungen aus einer Umgebung gegen die dargestellten Lösungen konvergieren, die Mengen sind also "anziehend" oder attrahierend und heißen deswegen *Attraktor*.

Bemerkung A.7 Eine Variante des Modells entsteht, wenn man (z.B. durch einen geeigneten Regelmechanismus) sicher stellt, dass die vorhandenen Nährstoffe $S(t)$ konstant gehalten werden, also $S(t) \equiv S_0 > 0$ sind. In diesem Fall kann die S–Gleichung ebenfalls weggelassen werden, dafür muss aber wieder ein Kapazitätsterm eingeführt werden, um das unbeschränkte Wachstum zu vermeiden. □

A.4 Ausbreitung von Epidemien (2010 nicht behandelt)

Zum Abschluss dieses Kapitels wollen wir ein Modell für die Ausbreitung von Epidemien betrachten, das zur Modellierung eine andere Art von Differentialgleichungen verwendet.

Als Beispiel für eine Epidemie betrachten wir hier eine Pflanzenkrankheit, nämlich die Kartoffelfäule.

Wir machen zunächst die Modellannahme, dass sich die Masse x der infizierten Pflanzen gemäß dem logistischen Wachstum mit $K = 1$ verhält, also

$$\dot{x}(t) = \lambda x(t)(1 - x(t)). \tag{A.20}$$

Die Kapazität $K = 1$ ist hierbei gerade die normierte Größe des Gesamtbestandes der Pflanzen, die befallen werden können. Natürlich muss man hierbei annehmen, dass alle diese Pflanzen so gleichmäßig stehen, dass der Erreger sich konstant mit Infektionsrate λ ausbreiten kann. Der Faktor $(1-x(t))$ modelliert in (A.20) die Kapazität des Lebensraumes, während der Faktor $\lambda x(t)$ die Ausbreitung der Infektion bestimmt.

Für Epidemien ist dies aber ein zu einfaches Modell, da wir aus der Analyse des Modells ja bereits wissen, dass die Lösungen gegen 1 konvergieren. Insbesondere würde eine mittels (A.20) modellierte Epidemie immer den gesamten Bestand befallen. In einem realistischeren Modell sollten also weitere aus der Biologie bekannte Tatsachen einbezogen werden. Wir werden hier nun den zeitlichen Verlauf einer Infektion berücksichtgen. Für die Kartoffelfäule ist bekannt, dass sich die Krankheit nach erfolgter Infektion zum Zeitpunkt t^* in zwei Stadien entwickelt:

- Das Latenzstadium $[t^*, t^* + p]$, in dem sich der Erreger nicht ausbreiten kann
- Das Infektionsstadium $[t^* + p, t^* + p + q]$, in dem sich der Erreger verbreiten kann

Nach der Zeit $t^* + p + q$ ist eine weitere Ausbreitung nicht möglich.

Mit $x(t)$ bezeichnen wir weiterhin die Masse der infizierten Pflanzen. Wir wollen nun eine Differentialgleichung für $x(t)$ aufstellen. Wir nehmen an, dass die Kapazität des "Lebensraumes" der Infektion von den Stadien der Krankheit nicht abhängt, so dass der Faktor $(1 - x(t))$ in (A.20) unverändert bleibt. Der Wachstumsfaktor $\lambda x(t)$ muss aber geändert werden: Die Zunahme der Infektion ist nun proportional zur Größe des Anteils der infizierten Population, die sich zur Zeit t im Infektionsstadium befindet. Diese Größe ist gegeben durch die Menge aller Infektionen, die älter als p sind, also $x(t - p)$, minus der Anzahl der Infektionen, die älter als $p + q$ sind, also $x(t - p - q)$. Wir ersetzen $\lambda x(t)$ also durch $\lambda(x(t - p) - x(t - p - q))$ und erhalten so die Gleichung

$$\dot{x}(t) = \lambda(1 - x(t))(x(t - p) - x(t - p - q)). \tag{A.21}$$

Dies ist jetzt keine gewöhnliche Differentialgleichung im üblichen Sinne mehr, da die rechte Seite nicht nur von $x(t)$ sondern auch von $x(t - p)$ und $x(t - p - q)$ abhängt. Eine solche Gleichung nennt man *Delay–Differentialgleichung*, auf deutsch auch *verzögerte Differentialgleichung*.

Allgemein kann man diese Gleichungen in der Form

$$\dot{x}(t) = f(x(t), x(t - \tau_1), \ldots, x(t - \tau_k))$$

für ein $f : (\mathbb{R}^n)^{k+1} \to \mathbb{R}^n$ schreiben, wobei wir $\tau_k > \tau_{k-1} > \ldots > \tau_1$ anehmen. Auch für diese Gleichungen gibt es einen Existenz– und Eindeutigkeitssatz, der dem Satz 1.4

sehr ähnlich ist (man benötigt wieder eine Lipschitz–Bedingung etc.). Ein wesentlicher Unterschied besteht aber bei der Wahl der Anfangsbedingung: Es genügt hier nicht, nur die Zeit t_0 und den Wert $x(t_0)$ festzulegen. Tatsächlich reicht es auch nicht, zusätzlich die Werte $x(t_0 - \tau_i)$ für $i = 1, \ldots, k$ festzulegen, denn für jeden Zeitpunkt $t > t_0$ benötigt man zur Berechnung von $\dot{x}(t)$ ja insbesondere die Werte $x(t-\tau_k)$. Da $t-\tau_k$ das gesamte Intervall $[t_0 - \tau_k, t_0]$ durchläuft, müssen wir als Anfangs"wert" also zusätzlich zu $x(t_0) = x_0$ noch eine Funktion $\Psi : [t_0 - \tau_k, t_0) \to \mathbb{R}^n$ festlegen. Für die Existenz– und Eindeutigkeitsaussage reicht es dabei aus, Ψ als stetige Funktion zu wählen, wobei es nicht nötig ist, dass Ψ in t_0 durch x_0 stetig fortgesetzt wird.

Wir wollen nun das Langzeitverhalten der Lösungen von (A.21) untersuchen um damit zu ermitteln, wie groß der für $t \to \infty$ befallene Pflanzenbestand bei der durch (A.21) modellierten Epidemie ist und wie dieser Wert von p und q abhängt. Wir wählen dabei die Anfangsfunktion $\Psi \equiv 0$ und einen Anfangswert $x(0) = x_0 \in (0, 1)$; die Infektion gelangt also zum Zeitpunkt $t_0 = 0$ von außen in den Pflanzenbestand.

Unter dieser Annahme sieht man per Induktion über $n = 1, 2, 3, \ldots$ aus (A.21), dass auf jedem Intervall $[(n-1)(p+q), n(p+q)]$ die Ungleichungen $\dot{x}(t) \geq 0$ und $x(t) \in [0, 1)$ gelten. Also ist $x(t)$ monoton wachsend und durch 1 beschränkt und konvergiert damit gegen einen Wert $\beta \in (0, 1]$. Mit $g(t) := x(t-p) - x(t-p-q)$ können wir (A.21) als

$$\dot{x}(t) = \lambda g(t)(1 - x(t))$$

schreiben. Dies ist nun wieder eine klassische gewöhnliche Differentialgleichung, für die man (mit einer Technik, die in Lehrbüchern unter dem Namen *Trennung der Variablen* zu finden ist) die explizite Lösung

$$x(t) = 1 - (1 - x_0) \exp\left(-\lambda \int_0^t g(\tau) d\tau\right)$$

berechnen kann. Durch Ableiten nach t prüft man leicht nach, dass dies tatsächlich die Lösung ist. Für $g(t)$ gilt nun wegen $x(\sigma) = \Psi(\sigma) = 0$ für $\sigma < 0$ die Gleichung

$$\begin{aligned} \int_0^t g(\tau) d\tau &= \int_0^t x(\tau - p) d\tau - \int_0^t x(\tau - p - q) d\tau \\ &= \int_0^{t-p} x(\sigma) d\sigma - \int_0^{t-p-q} x(\sigma) d\sigma = \int_{t-p-q}^{t-p} x(\sigma) d\sigma \end{aligned}$$

für alle $t \geq 0$. Wir erhalten somit

$$x(t) = 1 - (1 - x_0) \exp\left(-\lambda \int_{t-p-q}^{t-p} x(\sigma) d\sigma\right),$$

(was man wiederum durch Ableiten nach t überprüfen kann) und damit

$$\beta = \lim_{t\to\infty} x(t) = 1 - (1 - x_0) \exp\left(-\lambda \lim_{t\to\infty} \int_{t-p-q}^{t-p} x(\sigma) d\sigma\right).$$

Mit dem Mittelwertsatz der Integralrechnung und wegen $\lim_{t\to\infty} x(t) = \beta$ folgt

$$\lim_{t\to\infty} \int_{t-p-q}^{t-p} x(\sigma) d\sigma = \lim_{t\to\infty} q x(t - p - \theta q) = q\beta.$$

Der Limes β ist also bestimmt durch die Gleichung

$$\beta = 1 - (1 - x_0)e^{-q\lambda\beta}.$$

Leider erlaubt diese Gleichung keine explizite Lösung. In Abbildung A.15 ist $\beta(q, x_0)$ für $\lambda = 1$ und $x_0 = 0.1, 0.2, \ldots, 0.9$ in Abhängigkeit von q dargestellt. Die Graphen wurden numerisch berechnet. Beachte, dass $\beta(0, x_0) = x_0$ gilt.

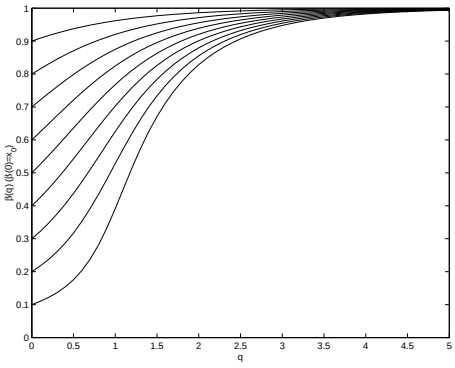

Abbildung A.15: Die Funktion $\beta(q, x_0)$ für verschiedene x_0 in Abhängigkeit von q

Die Graphen geben also den aus dem Modell (A.21) berechneten Anteil befallener Pflanzen für $t \to \infty$ in Abhängigkeit vom Anfangsbefall x_0 und der Infektionszeit q an. Für wachsendes q nähert sich die Größe der 1 an, d.h. fast der gesamte Bestand wird befallen. Für kleinere Infektionszeiten q hingegen wird nur ein Teil des Bestandes befallen. Für q gegen 0 nähert sich dieser Wert dem Anfangsbefall x_0 an. Der Grund für dieses Verhalten liegt darin, dass zur Ausbreitung der Krankheit eine gewisse Anzahl (relativ) frischer Infektionen vorliegen muss. Wenn die Wachstumsrate $\dot{x}(t)$ abnimmt, so fehlt der "Nachschub" an frischen Infektionen, dadurch nimmt $\dot{x}(t)$ weiter ab usw., weswegen die Lösung gegen $\beta < 1$ konvergiert.

Auch dieses Modell ist sicherlich für viele praktische Zwecke zu einfach, weil viele wichtige Einflüsse unberücksichtigt bleiben, z.B. Resistenzen gegen die Krankheit oder die räumliche Verteilung der Pflanzen. Trotzdem kann es zum Verständnis der Abhängigkeiten zwischen Infektionszeiten und Ausbreitungen von Epidemien beitragen und hierbei insbesondere die Komplexität der möglichen Abhängigkeiten illustrieren.

A.5 Literaturhinweise

Eine umfassende Einführung in die mathematische Biologie bietet das (in der ersten und zweiten Auflage einbändige, in der dritten Auflage zweibändige) Buch

J.D. Murray, Mathematical Biology, Springer–Verlag, 2002 (dritte Auflage).

Die Theorie des Chemostat ist in einer Reihe von Büchern beschrieben, z.B. in

H.L. Smith and P. Waltman, The Theory of the Chemostat, Cambridge University Press, 2003 (zweite Auflage)

Anhang B

Mechanische Modelle

Die mathematische Modellierung der klassischen Mechanik geht im Wesentlichen auf die Arbeiten von Isaac Newton[1], Jean Baptiste le Rond d'Alembert[2], Joseph–Louis Lagrange[3] und William R. Hamilton[4] zurück. Newton entwickelte die elementaren Bewegungsgleichungen (und nebenbei die Differentialrechnung), während Lagrange und Hamilton weiterführende mathematische Modellierungs– und Analysemethoden entwickelten, die wir im zweiten Abschnitt dieses Kapitels kennen lernen werden.

B.1 Mechanisch–technischer Ansatz

In diesem ersten Abschnitt wollen wir uns zunächst mit einem auch als d'Alembertsches Prinzip bezeichneten "modularen" Ansatz zur Modellierung mechanischer Systeme beschäftigen, der auf der Kombination verschiedener Elemente und der dazu gehörigen DGLs beruht. Für jedes Element werden wir ein grafisches Symbol und die dazugehörige Bewegungsgleichung (die nicht immer eine Differentialgleichung ist) betrachten. Man unterscheidet dabei zwischen verschiedenen Arten von Bewegungen, die wir der Reihe nach einführen wollen. Der Ansatz ist konstruktiv und erlaubt mit sehr wenig mathematischem Aufwand die Modellierung von (im Prinzip) beliebig komplizierten mechanischen Systemen. Wir werden allerdings auch sehen, dass der Ansatz für komplizierte Systeme unpraktikabel wird, was die Einführung mathematisch anspruchsvollerer Techniken im nachfolgenden zweiten Abschnitt rechtfertigt.

B.1.1 Translationale Bewegungselemente

Wir betrachten in diesem Teilabschnitt Bewegungselemente, die sich in eine Richtung auf einer Geraden bewegen können, also 1–dimensionale Bewegungen.

Wir verwenden dabei die folgenden Bezeichnungen:

[1] englischer Mathematiker und Physiker, 1642–1727

[2] französischer Mathematiker und Physiker, 1717–1783

[3] französischer Mathematiker, 1736–1813 (geboren als Giuseppe Lodovico Lagrangia in Turin, deshalb manchmal — vor allem in italienischen Büchern — auch als italienischer Mathematiker bezeichnet)

[4] irischer Mathematiker, 1805–1865

Variable	Bedeutung	Maßeinheit	
y	Ort, Ausdehnung	m	[Meter]
$v = \dot{y}$	Geschwindigkeit	m/s	[Meter pro Sekunde]
$a = \ddot{y}$	Beschleunigung	m/s^2	[Meter pro Sekunde zum Quadrat]
F	Kraft	$N = kg\,m/s^2$	[Newton]

a) Das Trägheitselement (Masse)

Das Trägheitselement besteht aus einer (zeitlich konstanten) Masse m auf die eine Kraft F wirkt und die sich mit einer Geschwindigkeit v bewegt. Das Symbol für das Trägheitselement ist in Abbildung B.1 dargestellt.

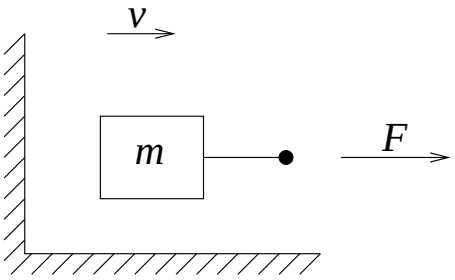

Abbildung B.1: Symbol für das Trägheitselement

Die Differentialgleichung für das Trägheitselement ist nach dem 2. Newton'schen Gesetz gegeben durch

$$F(t) = ma(t) = m\dot{v}(t). \tag{B.1}$$

Beachte, dass hier die Kraft F und die Geschwindigkeit v in die selbe Richtung zeigen müssen, ansonsten muss F durch $-F$ ersetzt werden; dies ist eine beliebte Quelle für Vorzeichenfehler! Die "Wände" im Symbol symbolisieren das verwendete Koordinatensystem, was wichtig sein kann, wenn mehrere Massen in einem System verbunden werden.

Eine Masse speichert *Energie*: Wenn die Masse in Bewegung ist, so besitzt sie die kinetische Energie

$$E_k(t) = \frac{m}{2}v(t)^2$$

und wenn sie sich in einem Schwerefeld befindet, so besitzt sie potentielle Energie, auf der Erde nahe der Erdoberfläche gerade

$$E_p(t) = mgh(t),$$

wobei $g \approx 9,80665 m/s^2$ die Erdbeschleunigung und h die Höhe über der Erdoberfläche bezeichnet.

b) Das Elastizitätselement (Feder)

Das Elastizitätselement wird ganz allgemein als deformierbares Objekt definiert, bei dem die Größe y der Deformation eine Funktion der einwirkenden Kraft F ist. Zwei gebräuchliche Symbole sind in Abbildung B.2 dargestellt.

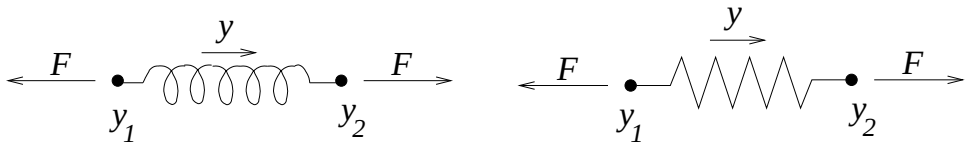

Abbildung B.2: Symbole für das Elastizitätselement

Beim *linearen Modellansatz* wird zur Beschreibung des Elastizitätselementes das *Hook'sche Gesetz* verwendet. Mit $y = y_2 - y_1$ ist dies durch

$$ky(t) = F(t) \tag{B.2}$$

gegeben, wobei $k > 0$ die *Federkonstante* ist. Per Konvention ist y_2 der Angriffspunkt in positiver Koordinatenrichtung und y_1 der Angriffspunkt in negativer Koordinatenrichtung.

Dieses Modell beschreibt eine reale Feder bei kleinen Auslenkungen i.A. hinreichend gut. Realistischere Ansätze verwenden einen nichtlinearen Zusammenhang zwischen y und F, worauf wir hier aber nicht näher eingehen wollen. Unabhängig von der Modellierung dieses Zusammenhangs sind reine Elastizitätselemente prinzipiell eine Idealisierung, da in der Realität keine Federn ohne Masse und Dämpfung (s.u.) existieren. Beachte, dass bei der obigen Wahl der Punkt $y = 0$ gerade die Feder in Ruhelage bezeichnet, die Ausdehnung kann in diesem mathematischen Modell also positiv (gedehnte Feder) oder negativ (gestauchte Feder) sein.

Auch Federn speichern potentielle Energie, falls man (B.2) annimmt, ist diese durch

$$E_p(t) = \frac{k}{2} y(t)^2$$

gegeben.

c) Das Dämpfungselement (Dämpfer, Reibung)

Die allgemeine Definition ist hier gegeben durch ein mechanisches Element, das keinerlei Energie speichert, sondern die aufgenommene Energie in Wärme umwandelt und abgibt (Dissipator). Das Symbol für das Dämpfungselement ist in Abbildung B.3 angegeben.

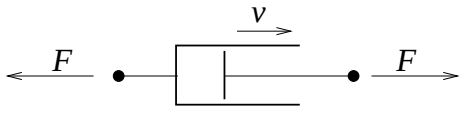

Abbildung B.3: Symbol für das Dämpfungselement

Wiederum betrachten wir hier nur das lineare Modell, das durch

$$F(t) = cv(t) \tag{B.3}$$

gegeben ist, wobei v die relative Geschwindigkeit zweier (in Abbildung B.3 durch Kolben und Zylinder symbolisierten) Körper ist, F die wirkende Kraft bezeichnet und $c > 0$ eine

Dämpfungskonstante ist: wenn die Kraft F wirkt, so wird die Geschwindigkeit cv erreicht. Die Relativgeschwindigkeit v berechnet sich hierbei als $v = v_+ - v_-$, wobei v_+ die Geschwindigkeit des Endpunktes in positiver und v_- die Geschwindigkeit des Endpunktes in negativer Koordinatenrichtung ist.

Dieses Modell nennt man *viskose Reibung.* Andere Modelle sind z.B. die *trockene Reibung*, bei der die Kraft F bei niedriger Geschwindigkeit größer wird (diese Funktion ist unstetig in $v = 0$) oder die *Strömungsreibung* (z.B. der Luftwiderstand), bei der $F = c\,|v|\,v$ ist, also quadratisch von v abhängt. Noch komplizierter ist die Haftreibung, die sich als klassische Funktion zwischen F und v nicht modellieren lässt, sondern nur mit sogenannten *Hysteresemodellen* beschrieben werden kann.

Die von dem Dämpfungselement zur Zeit t absorbierte Leistung ist gerade das Produkt $F(t)v(t)$, die im Intervall $[t_0, t_1]$ absorbierte Energie ergibt sich als Integral über die Leistung, also

$$E_{\text{abs}} = \int_{t_0}^{t_1} F(t)v(t)dt.$$

B.1.2 Einfache translationale Modelle

Die im letzten Abschnitt beschriebenen drei Elemente bilden die Grundbausteine für translationale mechanische Systeme. Der Ansatz, um kompliziertere Systeme beschreiben zu können funktioniert nun wie folgt.

(1) Modelliere ein translationales mechanisches System als Verbindung von Massen, Federn und Dämpfern.

(2) Stelle die zugehörigen Bewegungsgleichungen auf.

(3) Formuliere die Verbindungsgesetze (oder Kontaktkräfte).

Grundlegend für (3) ist Newtons 3. Gesetz *actio = reactio*: In jeder Masse ist die Summe der Kräfte (mit Berücksichtigung ihrer Vorzeichen) gleich Null. Falls zusätzlich eine externe Kraft wirkt, so ist die Summe der (systeminternen) Kräfte gleich der externen Kraft, wobei auch hier die Richtung der Kraft über das Vorzeichen berücksichtigt werden muss.

Wir illustrieren dies an zwei Beispielen.

Beispiel B.1 Mechanischer Oszillator (oder Schwinger)

Eine Masse ist an einer (realen) Feder an der Decke aufgehängt, auf die Masse wirkt eine Kraft $F(t)$, die nach unten gerichtet ist und die aus der Schwerkraft und einer externen Kraft zusammen gesetzt ist. Da eine reale Feder immer auch Dämpfung bewirkt, modellieren wir sie durch eine Kombination aus Elastizitäts– und Dämpfungselement. Abbildung B.4 stellt diese Kombination dar.

Mit $y(t)$ und $v(t)$ bezeichnen wir Position und Geschwindigkeit der Masse. Da Feder und Dämpfer mit der Masse verbunden sind, gilt in diesen Elementen $y_1 = y$ und $v_1 = v$ Der Aufhängepunkt der Feder und des Dämpfers sei $y_2 = 0$. Da dieser Punkt unbeweglich ist, folgt $v_2 = 0$. In diesen Elementen gilt also $y_2 - y_1 = -y$ und $v_2 - v_1 = -v$.

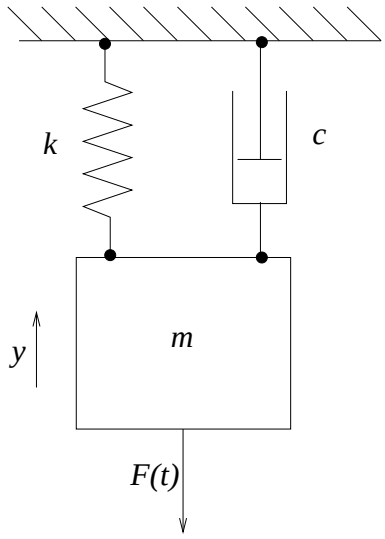

Abbildung B.4: Mechanischer Oszillator

Die Gleichungen der drei Elemente lauten damit

$$F_1(t) = m\dot{v}(t), \quad F_2(t) = c(v_2(t) - v_1(t)) = -cv(t), \quad F_3(t) = k(y_2(t) - y_1(t)) = -ky(t).$$

F_1 beschreibt die Kraft im Masseelement, F_2 die im Dämpfungselement und F_3 die im Elastizitätselement.

Wir müssen nun die Richtungen der Kräfte F_1, F_2, F_3 in den Elementen bestimmen. Da y durch den Pfeil angedeutet nach oben zunimmt, zeigt auch F_1 nach oben. Gemäß den Elementarmodellen zeigen F_2 und F_3 nach unten, da wir uns am unteren Ende der Elemente befinden. Ebenfalls zeigt die Kraft F gemäß dem eingezeichneten Pfeil nach unten; alle Kräfte sind also F_1 entgegengerichtet. Nach Newtons drittem Gesetz muss die Summe aller Kräfte (versehen mit Vorzeichen gemäß ihrer Richtungen) daher gleich $-F$ sein, also $F_1 - F_2 - F_3 = -F$. Daraus erhalten wir die Gesamtgleichung des Systems:

$$-F(t) = F_1(t) - F_2(t) - F_3(t) = m\dot{v} + cv(t) + ky(t) = m\ddot{y} + c\dot{y}(t) + ky(t).$$

Dies ist noch keine Differentialgleichung in der Form (1.1), denn hier treten höhere Ableitungen der unbekannten Funktion $y(t)$ auf. Man nennt diese Gleichungen *gewöhnliche Differentialgleichungen höherer Ordnung*, hier haben wir eine DGL zweiter Ordnung. In der Mechanik lässt man die Gleichungen meist in der obigen Form stehen, wir wollen sie hier in die Form (1.1) bringen, um sie in unseren abstrakten Rahmen einzupassen. Man kann jede DGL höherer Ordnung formal in eine Gleichung erster Ordnung der Form (1.1) bringen, indem man Hilfsgrößen für die Ableitungen einführt. Dazu definiert man $x_1(t) := y(t)$ und $x_2(t) := \dot{y}(t)$ (und so weiter, falls nötig) und erhält so das System

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -\frac{c}{m}x_2(t) - \frac{k}{m}x_1(t) - \frac{1}{m}F(t) \end{aligned}$$

in Form (1.1). In unserem Fall haben wir bereits die Symbole $y(t) = x_1(t)$ und $v(t) = x_2(t)$, so dass die Schreibweise

$$\begin{aligned} \dot{y}(t) &= v(t) \\ \dot{v}(t) &= -\frac{c}{m}v(t) - \frac{k}{m}y(t) - \frac{1}{m}F(t) \end{aligned}$$

mit der zweidimensionalen unbekannten Funktion $(y(t), v(t))^T$ aussagekräftiger ist.

Der Nullpunkt $y = v = 0$ ist gerade das einzige Gleichgewicht dieser Gleichung und entspricht dem Gleichgewichtspunkt der Masse wenn keine Kraft wirkt, also $F(t) \equiv 0$ ist. Dies ist i.A. keine gute Wahl für unser Modell, da (unabhängig von weiteren äußeren Kräften) immer die Schwerkraft $F_G = mg$ auf die Masse wirkt. Man kann die Schwerkraft aber aus dem Modell eliminieren, wenn man den Nullpunkt für y anders wählt. Wir zerlegen dazu $F(t) = F_G + F_e(t)$ in die Schwerkraft und eine (gegebenfalls wirkende) weitere externe Kraft $F_e(t)$. Sicherlich besitzt die Gleichung ein Gleichgewicht $(y_G, 0)^T$ für $F_e(t) \equiv 0$, also wenn nur die Schwerkraft F_G wirkt. Für dieses gilt

$$0 = -\frac{k}{m}y_G - \frac{1}{m}F_G \quad \Leftrightarrow \quad y_G = -\frac{F_G}{k} = -\frac{mg}{k}.$$

Mit der Koordinatentransformation $\tilde{y} = y - y_G$ wird dies der neue Nullpunkt, und in den neuen Koordinaten gilt nun

$$\dot{\tilde{y}}(t) = \dot{y}(t) = v(t)$$

und

$$\begin{aligned} \dot{v}(t) &= -\frac{c}{m}v(t) - \frac{k}{m}y(t) + \frac{1}{m}F(t) \\ &= -\frac{c}{m}v(t) - \frac{k}{m}\tilde{y}(t) \underbrace{-\frac{k}{m}y_G - \frac{1}{m}F_G}_{=0} - \frac{1}{m}F_e(t) \\ &= -\frac{c}{m}v(t) - \frac{k}{m}\tilde{y}(t) - \frac{1}{m}F_e(t) \end{aligned}$$

Also ergibt sich die Gleichung

$$\begin{aligned} \dot{\tilde{y}}(t) &= v(t) \\ \dot{v}(t) &= -\frac{c}{m}v(t) - \frac{k}{m}\tilde{y}(t) - \frac{1}{m}F_e(t) \end{aligned} \tag{B.4}$$

in der die Schwerkraft nicht mehr auftaucht. Wenn man den Nullpunkt für y also von vornherein in das Gleichgewicht unter Schwerkraft legt, so braucht man F_G gar nicht berücksichtigen oder mit anderen Worten: Wenn wir F_G nicht ins Modell aufnehmen, so beschreibt der Nullpunkt "automatisch" das Gleichgewicht unter Schwerkraft.

Abbildung B.5 zeigt eine Lösung dieser Gleichung mit $F_e(t) \equiv 0$ und Anfangswert $(y(0), v(0)) = (-1, 0)$. Die Masse wird also nach unten in den Punkt $y = -1$ "gezogen" und zum Zeitpunkt $t = 0$ losgelassen.

Die Grafik zeigt das zu erwartende Verhalten: Die Masse schwingt etwas und pendelt sich, bedingt durch die Reibung, auf das Gleichgewicht $(0, 0)$ ein. □

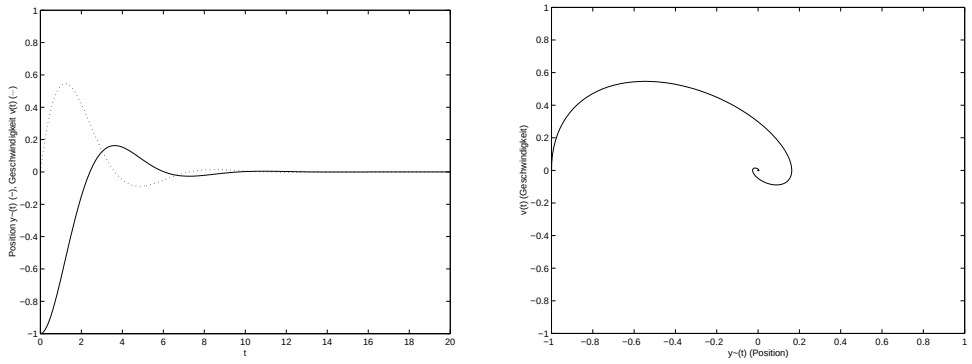

Abbildung B.5: Lösungen von (B.4) mit $m = k = c = 1$

Beispiel B.2 Ein einfaches Fahrzeug–Federungsmodell

Abbildung B.6 zeigt ein einfaches Modell für eine Fahrzeugfederung. Die Modellannahmen hier sind:

- Es werden nur vertikale Bewegungen berücksichtigt (keine Drehungen)
- Es wird nur ein Rad modelliert
- Die Karosserie wird als Masse m_1 mit Position y_1 modelliert, die Federung mittels Elastizitäts– und Dämpfungselement k_1, c_1
- Rad und Achse wird als Masse m_2 mit Position y_2 modelliert, der Reifen mittels Elastizitäts– und Dämpfungselement k_2, c_2
- Die Straßenunebenheiten werden durch eine "Straßenhöhenfunktion" $u(t)$ modelliert

Aus den Bewegungsgleichungen erhalten wir die Gleichungen für die Einzelkräfte

$$m_i \dot{v}_i^m(t) = F_i^m(t), \quad c_i v_i^c(t) = F_i^c(t), \quad k_i y_i^k(t) = F_i^k(t)$$

für $i = 1, 2$, wobei die Indizes angeben, für welches Element die Kräfte bzw. Positionen gelten. Hierbei gelten die Beziehungen

$$\begin{aligned} v_1^m(t) &= \dot{y}_1(t), & v_1^c(t) &= \dot{y}_1(t) - \dot{y}_2(t), & y_1^k(t) &= y_1(t) - y_2(t), \\ v_2^m(t) &= \dot{y}_2(t), & v_2^c(t) &= \dot{y}_2(t) - \dot{u}(t), & y_2^k(t) &= y_2(t) - u(t). \end{aligned}$$

Um die obigen Gleichungen zusammenzusetzen müssen wir nun die Einzelkräfte an den Massen bestimmen. Hierbei muss man wieder auf die Richtungen aufpassen, die von der betrachteten Masse abhängen. In m_1 ist die Kraft F_1^m (entsprechend der Richtung von y_1) nach oben gerichtet, ebenso zeigen F_1^c und F_1^k nach oben, da dies das obere Ende der Elemente ist. In m_1 ergibt sich so

$$F_1^m + F_1^c + F_1^k = 0.$$

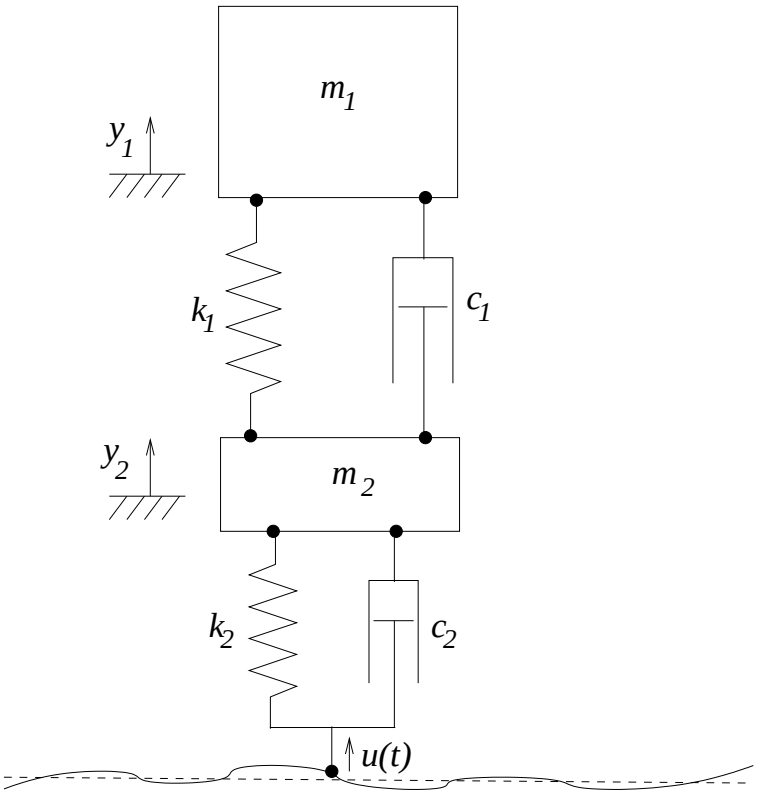

Abbildung B.6: Federungsmodell

In m_2 zeigen F_1^k und F_1^c nach unten und alle anderen Kräfte nach oben, also erhalten wir in m_2 die Gleichung

$$F_2^m - F_1^k - F_1^c + F_2^k + F_2^c = 0.$$

Zusammen ergibt dies die Gleichungen

$$\begin{aligned} 0 &= F_1^m(t) + F_1^c(t) + F_1^k(t) \\ &= m_1\dot{v}_1^m(t) + c_1 v_1^c(t) + k_1 y_1^k(t) \\ &= m_1\ddot{y}_1(t) + c_1(\dot{y}_1(t) - \dot{y}_2(t)) + k_1(y_1(t) - y_2(t)) \end{aligned}$$

und

$$\begin{aligned} 0 &= F_2^m(t) - F_1^c(t) - F_1^k(t) + F_2^c(t) + F_2^k(t) \\ &= m_2\dot{v}_2^m(t) - c_1 v_1^c(t) - k_1 y_1^k(t) + c_2 v_2^c(t) + k_2 y_2^k(t) \\ &= m_2\ddot{y}_2(t) - c_1(\dot{y}_1(t) - \dot{y}_2(t)) - k_1(y_1(t) - y_2(t)) + c_2(\dot{y}_2(t) - \dot{u}(t)) + k_2(y_2(t) - u(t)) \end{aligned}$$

Analog zum vorherigen Beispiel kann diese Gleichung nun in die Form (1.1) mit der vierdimensionalen unbekannten Funktion $(y_1(t), v_1(t), y_2(t), v_2(t))^T$ umgeformt werden. □

B.1.3 Rotationale Bewegungselemente

Bisher können wir nur Bewegungen in gerader Richtung beschreiben. Nun lernen wir drei analoge Elemente kennen, die Drehbewegungen darstellen können.

Wir verwenden dabei die folgenden Bezeichnungen:

Variable	Bedeutung	Maßeinheit	
θ	Winkel	rad	[Radiant]
$\omega = \dot{\theta}$	Winkelgeschwindigkeit	rad/s	[Radiant pro Sekunde]
$\alpha = \ddot{\theta}$	Winkelbeschleunigung	rad/s^2	[Radiant pro Sek. zum Quadrat]
τ	Drehmoment	$Nm = kg\,m^2/s^2$	[Newtonmeter]

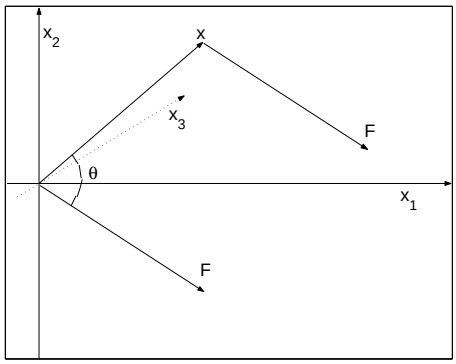

Abbildung B.7: Schematische Darstellung des Drehmoments

Das Drehmoment beschreibt dabei die auf einen rotierenden Körper wirkenden Kraft: Sei $F = (F_1, F_2, 0)$ eine gerichtete Kraft die im Punkt $x = (x_1, x_2, 0)$ an einem um die x_3–Achse rotierenden Körper angreift, vgl. Abbildung B.7: Man kann sich den Vektor x als Hebel an dem (nicht dargestellten) Körper vorstellen. Dann ist das resultierende Drehmoment gegeben durch

$$\tau = x_1 F_2 - x_2 F_1 = \|x\| \|F\| \sin\theta, \tag{B.5}$$

wobei θ den Winkel zwischen x und F beschreibt. Hierbei ist wie immer auf das Vorzeichen zu achten, die positive Drehrichtung ist so zu wählen, dass die beiden Ausdrücke in (B.5) übereinstimmen.

Zu beachten ist, dass wir die Kraft F nun als Kraftvektor im Koordinatensystem schreiben, die Richtungsinformation ist — im Gegensatz zu den translationalen Modellen — hier also bereits in F enthalten, so dass wir beim Bestimmen der Kontaktkräfte keine Richtungen mehr berücksichtigen müssen.

a) Das Trägheitselement (Trägheitsmoment)

Das Trägheitselement der Rotation besteht aus einer Masse, die um eine Achse rotiert. Die Gleichung lautet

$$\tau(t) = J\alpha(t) = J\dot{\omega}(t). \tag{B.6}$$

Das Trägheitsmoment J des Elementes wird durch die Masse des Elementes und ihre Verteilung bzgl. der Rotationsachse bestimmt.

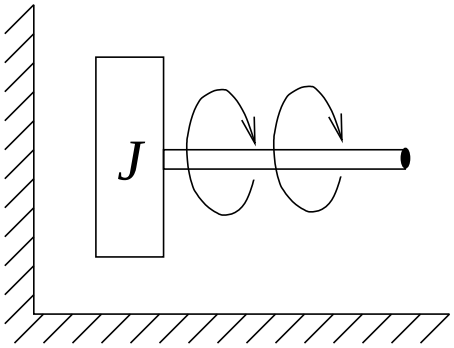

Abbildung B.8: Symbol für das Rotations–Trägheitselement

Für einen rotierenden Körper $B \subset \mathbb{R}^3$ mit Masse m und Massendichte $\rho : B \to \mathbb{R}_0^+$ gilt allgemein

$$J = \int_B r^2(x)\rho(x)dx.$$

In Spezialfällen kann man explizite Formeln angeben:

Ein rotierender Massepunkt mit Masse m und Abstand r von der Rotationsachse besitzt gerade das Trägheitsmoment

$$J = mr^2.$$

Für eine Menge von N Massepunkten x^i mit Koordinaten (x_1^i, x_2^i, x_3^i), die um die x_3–Achse rotieren, gilt

$$J = \sum_{i=1}^{N} m_i r_i^2 = \sum_{i=1}^{N} m_i((x_1^i)^2 + (x_2^i)^2).$$

Für einen dünnen Stab mit Länge l und einer eindimensionalen Massenverteilung $\rho : [0, l] \to \mathbb{R}_0^+$ gilt bei Rotation um den Endpunkt 0

$$J = \int_0^l x^2\rho(x)dx$$

und bei Rotation um den Mittelpunkt $l/2$

$$J = \int_0^l (x - l/2)^2\rho(x)dx.$$

Falls die Masse m im Stab gleichverteilt ist, gilt $\rho(x) = m/l$ und damit

$$J = \int_0^l x^2 m/l dx = \frac{ml^2}{3}$$

bei Rotation um den Endpunkt und

$$J = \int_0^l (x - l/2)^2 m/l dx = \int_{-l/2}^{l/2} x^2 m/L dx = \frac{ml^2}{12}$$

bei Rotation um den Mittelpunkt.

Allgemein gilt der *Parallelachsensatz*, auch *Steiner'scher Satz* genannt.

Satz B.3 Gegeben ein Körper $B \subset \mathbb{R}^3$ der Masse m mit Massenverteilung $\rho : B \to \mathbb{R}_0^+$ und Schwerpunkt

$$\overline{x} = \frac{1}{m} \int_B x\rho(x)dx \in \mathbb{R}^3.$$

Sei J das Trägheitsmoment des Körpers um eine beliebige Achse, J' das Trägheitsmoment des Körpers um die durch den Schwerpunkt verlaufende parallele Achse. Dann gilt

$$J = J' + mR^2,$$

wobei R den Abstand der beiden Achsen bezeichnet.

b) Das Elastizitätselement (Torsionsfeder)

Das Elastizitätselement und das nachfolgende Dämpfungselement sind völlig analog zu ihren translationalen Gegenstücken. Wir schon bei diesen betrachten wir hier nur die linearen Bewegungsmodelle. Für das rotationale Elastizitätselement its die entsprechende Gleichung durch

$$k\theta(t) = \tau(t) \tag{B.7}$$

gegeben.

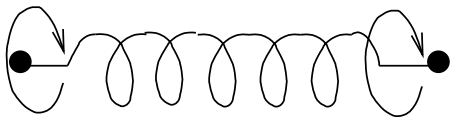

Abbildung B.9: Symbol für das Rotations–Elastizitätselement

c) Das Dämpfungselement (Rotationsdämpfer)

Wiederum analog zum translationalen Fall gibt es das rotationale Dämpfungselement, dessen Gleichung

$$c\omega(t) = \tau(t) \tag{B.8}$$

lautet.

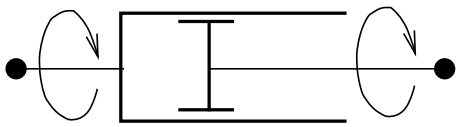

Abbildung B.10: Symbol für das Rotations–Dämpfungselement

B.1.4 Das Pendel

Wir wollen die besprochenen Elemente nun zu einem Modell eines Pendels zusammensetzen. Wir machen zuerst die folgenden vereinfachenden Modellannahmen

- Das Pendel ist eine Punktmasse m, die an einem masselosen Stab der Länge l befestigt ist
- Es gibt keine Reibung

Das Modell ist in Abbildung B.11 schematisch dargestellt.

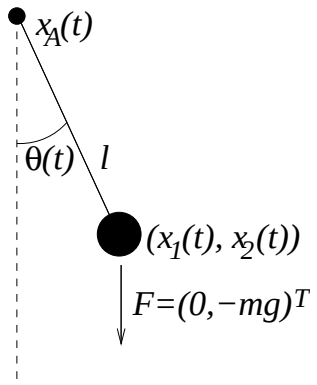

Abbildung B.11: Pendelmodell

Es sei $x(t) = (x_1(t), x_2(t))^T$ der Endpunkt des Pendels. Der Aufhängepunkt bzw. die Position der Drehachse $x_A(t)$ sei zunächst konstant gleich 0. Wie üblich im Koordinatensystem nehmen x_1 und x_2 nach rechts bzw. oben zu. Der Punkt $x(t)$ lässt sich mittels der Länge l und des Winkels $\theta(t)$ als

$$x(t) = (l \sin\theta(t),\ -l\cos\theta(t))^T$$

schreiben.

Der in x angreifende Kraftvektor F ist durch die Erdbeschleunigung gegeben als $F = (0, -mg)^T$ (er zeigt nach unten, deswegen '$-$'). Gemäß (B.5) gilt für das erzeugte Drehmoment also

$$\tau_F(t) = x_1(t) \cdot (-mg) - x_2(t) \cdot 0 = -mgx_1(t) = -mgl\sin\theta(t).$$

Andererseits gilt für das Trägheitselement die Gleichung

$$\tau_J(t) = J\ddot{\theta}(t) = ml^2\ddot{\theta}(t),$$

da wir das Pendel ja als Punktmasse modelliert haben. Gleichsetzen des externen Drehmoments τ_F mit τ_J ergibt

$$\tau_J(t) = \tau_F(t),$$

also

$$ml^2\ddot{\theta}(t) = -mgl\sin\theta(t).$$

Wiederum erhalten wir eine DGL zweiter Ordnung, die wir über die Gleichung $\omega(t) = \dot{\theta}(t)$ als

$$\begin{aligned} \dot{\theta}(t) &= \omega(t) \\ \dot{\omega}(t) &= -\frac{g}{l}\sin\theta(t) \end{aligned} \tag{B.9}$$

in der Form (1.1) schreiben können. Abbildung B.12 zeigt einige Lösungen dieser Gleichung.

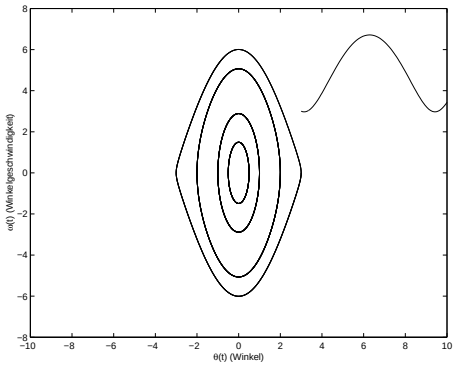

Abbildung B.12: Lösungen des Pendelmodells (B.9) mit $l = 1$

Die periodischen Lösungen gehören hierbei zu Anfangswerten, für die das Pendel hin– und her schwingt. Da keine Reibung modelliert wurde, wird das Pendel nicht gebremst und die Pendel schwingt für alle Zeiten mit der gleichen Bewegung, daher die Periodizität. Die Lösung, die rechts aus dem Bild hinausläuft wurde mit größerer Anfangsgeschwindigkeit gestartet. Hier überschlägt sich das Pendel, und zwar — da keine Reibung vorhanden ist — nicht nur einmal sondern immer wieder. Beachte, dass die Winkel θ und $\theta + k2\pi$ für alle $k \in \mathbb{Z}$ die gleiche Pendelposition bedeuten, aber in in unserem Modell unterschieden werden. Die Gleichung besitzt übrigens genau die Gleichgewichte $(\theta_k^*, \omega_k^*) = (k\pi, 0)$ für $k \in \mathbb{Z}$. Für gerades k ist dies gerade das herunterhängende Pendel, für ungerades k ist dies das aufrecht stehende Pendel. Die aufrechten Gleichgewichte sind exponentiell instabil (aber nicht antistabil), die herabhängenden sind weder exponentiell stabil noch instabil, denn die Realteile der Eigenwerte der Linearisierung sind gleich 0.

Wir wollen unser Modell nun realistischer machen. Zunächst fügen wir Reibung hinzu, Wir machen die Annahme:

- Es gibt viskose (d.h. lineare) Reibung mit Koeffizient c.

Wenn wir ein Dämpfungselement an der Achse hinzunehmen, so erhalten wir für das zusätzliche Drehmoment $\tau_c(t)$ die Gleichung

$$\tau_c(t) = c\omega_c(t) = c\dot{\theta}(t).$$

Wie bei den Translationselementen müssen wir nun die Drehmomente addieren und gleich der externen Kraft setzen. Da $\omega_c = \omega$ ist, erhalten wir also

$$\tau_J(t) + \tau_c(t) = \tau_F(t) \tag{B.10}$$

und damit die neue Gleichung

$$ml^2\ddot{\theta}(t) = -c\dot{\theta}(t) - mgl\sin\theta(t).$$

Nun wollen wir noch die Punktmassenannahme verallgemeinern. Wir nehmen nun an:

- Das Pendel ist ein starrer Körper mit Masseverteilung $\rho(x_1, x_2)$ und Trägheitsmoment J.

Die Frage ist jetzt: Wie berechnen wir das durch F erzeugte Drehmoment $\tau_F(t)$? Wir leiten die Lösung heuristisch her. In jedem Rechteck der Form $[x_1(t), x_1(t)+\Delta x_1] \times [x_2(t), x_2(t)+\Delta x_2]$ mit Masse Δm erzeugt F das Drehmoment

$$\Delta\tau_F(t) = -x_1(t)\Delta mg \approx -x_1(t)\rho(x(t))\Delta x_1 \Delta x_2 g.$$

Aufsummieren über alle Rechtecke und Grenzübergang $\Delta x_i \to 0$ liefert dann

$$\tau_F(t) = -\int_{B(t)} x_1\rho(x)gdx = -mg\overline{x}_1(t),$$

wobei $\overline{x}(t) = (\overline{x}_1(t), \overline{x}_2(t))^T$ die Position des Schwerpunktes des Körpers zur Zeit t bezeichnet.

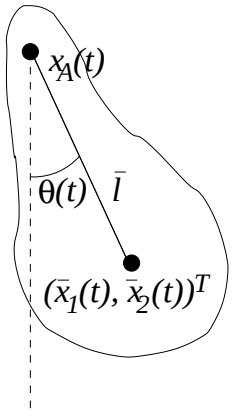

Abbildung B.13: Pendelmodell mit starrem Körper

Wenn wir mit $\bar{l} = \|\overline{x}(t)\|$ den Abstand des Schwerpunktes zur Rotationsachse (die zunächst weiterhin im Nullpunkt liegt) bezeichnen, und mit $\theta(t)$ den Winkel von $\overline{x}(t)$ mit der x_2–Achse bezeichnen (siehe Abbildung B.13), so gilt

$$\overline{x}(t) = (\bar{l}\sin\theta(t), -\bar{l}\cos\theta(t)).$$

Aus der schon bekannten Gleichung (B.10)

$$\tau_J(t) + \tau_c(t) - \tau_F(t) = 0$$

erhalten wir damit

$$J\ddot{\theta}(t) = -c\dot{\theta}(t) - mg\bar{l}\sin\theta(t)$$

bzw. in der Form (1.1) die Gleichung

$$\begin{aligned} \dot{\theta}(t) &= \omega(t) \\ \dot{\omega}(t) &= -\frac{c}{J}\omega(t) - \frac{mg\bar{l}}{J}\sin\theta(t) \end{aligned} \tag{B.11}$$

Abbildung B.14 zeigt die Lösungen dieser Gleichung mit den gleichen Anfangswerten wie in Abbildung B.12.

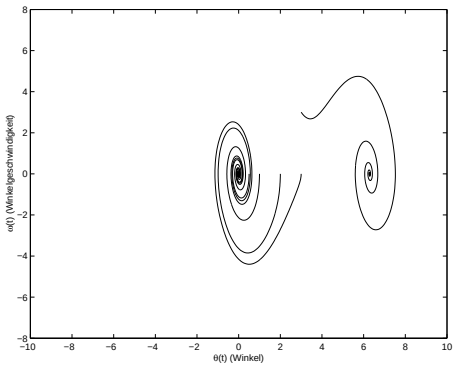

Abbildung B.14: Lösungen des Pendelmodells (B.11) mit $c/J = 1$ und $m\bar{l}/J = 1$

Hier streben alle Lösungen dem Gleichgewicht $(0, 0)$, bzw. nach einmaligem Überschlagen dem Gleichgewicht $(2\pi, 0)$ zu. Tatsächlich kann man zeigen, dass die Gleichgewichte der Form $(2k\pi, 0)$ mit $k \in \mathbb{Z}$ (herabhängendes Pendel) nun lokal exponentiell stabil sind. Die Gleichgewichte der Form $((2k+1)\pi, 0)$ mit $k \in \mathbb{Z}$ (aufrechtes Pendel) bleiben exponentiell instabil. Beachte, dass sich (B.9) und (B.11) in diesen Simulationen wegen der Wahl der Parameter nur durch den Reibungsterm unterscheiden.

Bemerkung B.4 Das oben heuristisch hergeleitete Prinzip gilt allgemein: Wenn Kräfte auf einen starren Körper wirken, so bewegt sich dessen Schwerpunkt genau so, wie sich die im Schwerpunkt konzentrierte Gesamtmasse unter Einfluss der Summe der Kräfte bewegen würde (ohne Beweis). □

Wir wollen nun untersuchen, wie sich das Pendel bei Bewegung des Aufhängepunktes x_A verhält. Wir betrachten dabei horizontale und vertikale Bewegung. Wenn der Aufhängepunkt $x_A(t)$ horizontal und vertikal bewegt wird, so gilt für die Kraft $F_A = (F_{A,1}, F_{A,2})^T$ nach (B.1)

$$F_{A,1}(t) = m\ddot{x}_{A,1}(t), \quad F_{A,2}(t) = m\ddot{x}_{A,2}(t) \tag{B.12}$$

Diese Kraft greift im Punkt x_A an und erzeugt gemäß Bemerkung B.4 und Gleichung (B.5) (beachte, dass der Vektor x in (B.5) hier gerade gleich $\overline{x} - x_A$ ist) nun das Drehmoment

$$\tau_A(t) = (\overline{x}_1(t) - x_{A,1}(t))F_{A,2}(t) - (\overline{x}_2(t) - x_{A,2}(t))F_{A,1}(t).$$

Für das Pendel gilt nun als Erweiterung der Gleichung (B.10)

$$\tau_J(t) + \tau_c(t) + \tau_A(t) = \tau_F(t).$$

Das Drehmoment τ_A wird hierbei als internes Drehmoment — also auf der linken Seite — eingesetzt, da es sich um die Auswirkung der Bewegung des (modellinternen) Aufhängepunktes auf die Masse handelt. Dass diese wiederum von einer externen Kraft F_A hervorgerufen wird, haben wir bereits in (B.12) berücksichtigt.[5]

Wegen

$$\overline{x}(t) - x_A(t) = (\bar{l}\sin\theta(t), -\bar{l}\cos\theta(t))^T$$

erhalten wir

$$\tau_A(t) = F_{A,2}(t)(\bar{l}\sin\theta(t)) - F_1(t)(-\bar{l}\cos\theta(t)) = m(\ddot{x}_{A,2}(t)\bar{l}\sin\theta(t) + \ddot{x}_{A,1}(t)\bar{l}\cos\theta(t)$$

und damit

$$J\ddot{\theta}(t) = -c\dot{\theta}(t) - mg\bar{l}\sin\theta(t) - m\ddot{x}_{A,2}(t)\bar{l}\sin\theta(t) - m\ddot{x}_{A,1}(t)\bar{l}\cos\theta(t)$$

bzw. wiederum in der Form (1.1) die Gleichung

$$\begin{aligned} \dot{\theta}(t) &= \omega(t) \\ \dot{\omega}(t) &= -\frac{c}{J}\omega(t) - \frac{m\bar{l}}{J}\Big(g\sin\theta(t) + \ddot{x}_{A,2}(t)\sin\theta(t) + \ddot{x}_{A,1}(t)\cos\theta(t)\Big) \end{aligned} \tag{B.13}$$

Wir wollen die Lösungen für eine spezielle Wahl der Beschleunigungen $\ddot{x}_A$ veranschaulichen. Nehmen wir an, dass nur vertikale Bewegungen vorliegen, wobei der Aufhängepunkt mit konstanter Frequenz Ω und Amplitude a cosinusförmig auf– und abschwingt, also $x_{A,2}(t) = a\cos\Omega t$ gilt. Dann folgt

$$\ddot{x}_{A,2}(t) = -a\Omega^2\cos\Omega t,$$

also

$$\begin{aligned} \dot{\theta}(t) &= \omega(t) \\ \dot{\omega}(t) &= -\frac{c}{J}\omega(t) - \frac{m\bar{l}}{J}\Big(g - a\Omega^2\cos\Omega t\Big)\sin\theta(t) \end{aligned} \tag{B.14}$$

Mit den Parametern

$$\Omega = 1.57,\ \frac{c}{J} = 0.2,\ \frac{m\bar{l}g}{J} = 1 \text{ und } \frac{m\bar{l}a\Omega^2}{J} = 1.42$$

erhält man das in Abbildung B.15 dargestellte Verhalten.

[5] Wie Sie sicherlich in der Vorlesung bemerkt haben, ist die Bestimmung der richtigen Vorzeichen der zusammenwirkenden Kräfte und Drehmomente ein subtiler Punkt und eine häufige Fehlerquelle. Es ist daher immer ratsam, das gewonnene Modell z.B. durch numerische Simulation auf seine Plausibilität zu prüfen.

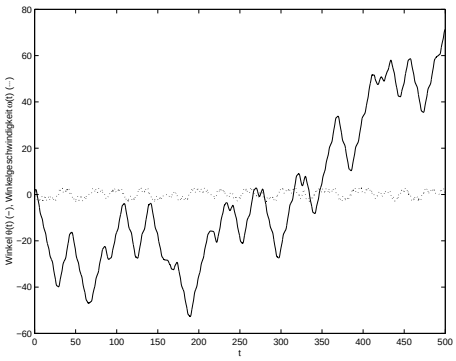

Abbildung B.15: Lösungen des Pendelmodells (B.14) in Abhängigkeit von t

Das Ganze sieht recht "chaotisch" aus und tatsächlich ist dies ein Beispiel für eine Differentialgleichung mit sogenanntem *chaotischen* Verhalten. Die Lösungen der Gleichung zeigen einen quasi zufälligen Verlauf: es ist nicht vorhersagbar, wann sich das Pendel überschlägt, die Richtung ändert etc. Trotzdem lassen sich in diesem chaotischen Verhalten Gesetzmäßigkeiten erkennen. Hierzu muss man zunächst die vorhandenen Periodizitäten berücksichtigen: Man kann in der Gleichung alle Winkel θ_1, θ_2 mit $\theta_1 = \theta_2 + 2k\pi$ für ein $k \in \mathbb{Z}$ identifizieren, da diese Werte die gleichen Positionen darstellen. Mittels $\tilde{\theta}(t) = \theta(t) - 2k(t)\pi$ (für das richtige $k(t)$) kann man die θ–Komponente der Lösung in das Intervall $[-\pi, \pi]$ "projizieren". Zusätzlich kann man die Periodizität der Beschleunigung berücksichtigen: Für alle $t \in \mathbb{R}$ gilt $\cos \Omega t = \cos \Omega(t + kT)$ für $T = 2\pi/\Omega$ und alle $k \in \mathbb{Z}$). Aufgrund dieser Beobachtung ist es sinnvoll, die Lösungen jeweils nach einer Periode der Beschleunigung darzustellen. Es zeigt sich, dass für jedes $t_0 \in \mathbb{R}$ eine (recht komplizierte) Menge A_{t_0} existiert, so dass jede "periodisch ausgewertete" Lösung, also jede Folge der Form

$$x(t_0 + kT; t_0, x_0) \text{ für } k = 0, 1, 2, \ldots$$

gegen diese Menge konvergiert. Diese Menge A_{t_0} heißt *Attraktor*. Durch Darstellung der Punkte $x(t_0 + kT; t_0, x_0)$ für $k = k_0, \ldots, k_1$ mit hinreichend großen $k_1 >> k_0 >> 0$ kann man einen Eindruck von dieser Menge gewinnen[6]. Abbildung B.16 zeigt die Punkte $x(t_0 + kT; t_0, x_0)$ für $x_0 = (1, 1)^T$, $t_0 = T/2$ und $k = 101, \ldots, 10000$.

Diese numerischen Ergebnisse lassen sich durch reale Experimente bestätigen. Im Buch *R.W. Leven, B. Koch und B. Pompe, "Chaos in dissiptiven Systemen", Akademie Verlag 1989 (1. Auflage) und 1994 (2. Auflage)* finden sich in Kapitel 1 der Versuchsaufbau und experimentelle Resultate.

B.2 Lagrange–Gleichungen und Hamilton–Formalismus

Die im vorherigen Abschnitt vorgestellte Methode hat den Nachteil, dass das Zusammensetzen der elementaren Gleichungen für große Systeme sehr kompliziert wird. Man muss

[6]Es gibt auch spezielle Algorithmen zur genaueren Berechnung von Attraktoren, diese werden in der Vorlesung "Numerik dynamischer Systeme" im kommenden Wintersemester behandelt.

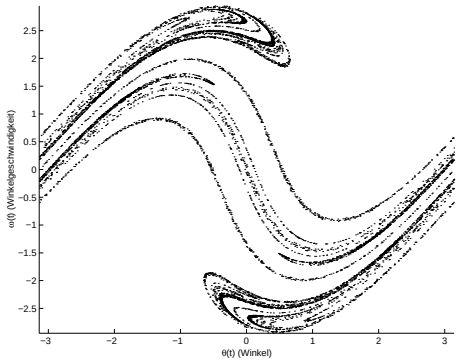

Abbildung B.16: Approximation des Attraktors des Pendelmodells (B.14)

für große Systeme und die Verbindungsgesetze bzw. Kontaktkräfte vieler Einzelgleichungen berücksichtigen, was zu sehr großen Gleichungssystemen führt, die dann nur schwer aufzulösen sind.

Die Alternative, die wir in diesem Abschnitt (aus Zeitgründen nur skizziert) vorstellen werden, ist die *Energie–basierte* Methode mit Hilfe der *Lagrange–Gleichungen.*

B.2.1 Lagrange–Gleichungen

Die Idee der Lagrange–Gleichungen basiert auf der Betrachtung der Energie des Systems. Wir beschränken uns hierbei auf ein System s von N Massepunkten mit Positionen $r_i = (x_i, y_i, z_i)^T$ und Massen m_i, $i = 1, \ldots, N$, dessen kinetische Energie gerade durch

$$E_{kin} = \sum_{i=1}^{N} \frac{m_i}{2} \|v_i\|^2$$

gegeben ist.

Zur Modellierung der Tatsache, dass sich ein mechanisches System — bedingt durch die mechanische Struktur — nur auf vorgegebenen Bahnen bewegen kann, verwenden wir Nebenbedingungen der Form

$$c_j(r_1, \ldots, r_N, t) = 0, \quad \text{für } j = 1, \ldots, J,$$

wobei die $r_i = (x_i, y_i, z_i)^T \in \mathbb{R}^3$ die Positionen der Massepunkte beschreiben.

Beispiel: Wir betrachten ein im Nullpunkt aufgehängtes und in der xy–Ebene schwingendes starres Pendel mit Punktmasse m im Punkt $r(t) = (x(t), y(t), z(t))^T$ und Länge ρ. Die möglichen Positionen von $r(t)$ werden dann genau durch die Gleichungen

$$c_1(r) = \|r\|^2 - \rho^2 \text{ und } c_2(r) = z$$

beschrieben.

Wir nehmen nun an, dass die durch

$$M = \{(r_1, \ldots, r_N)^T \mid c_j(r_1, \ldots, r_N, t) = 0, \quad \text{für } j = 1, \ldots, J\}$$

implizit definierte *Mannigfaltigkeit der verträglichen Konfigurationen* durch Koordinaten $q = (q_1, \ldots, q_l) \in Q$ mit einer offenen Menge $Q \subset \mathbb{R}^l$ parametrisieren lässt, d.h. dass stetig differenzierbare Funktionen $r_i(q,t)$ existieren mit

$$M = \{(r_1(q,t), \ldots, r_N(q,t)^T \mid q \in Q\}.$$

Wir nehmen weiterhin an, dass die partiellen Ableitungen

$$\frac{\partial r}{\partial q_i}(q,t) \in \mathbb{R}^{3N}$$

für $i = 1, \ldots, l$ linear unabhängig sind. Die Größen $q_1, \ldots, q_l$ heißen *verallgemeinerte Koordinaten*.

Beispiel: Für das Pendel gilt

$$r(q) = \begin{pmatrix} \rho \sin q \\ -\rho \cos q \\ 0 \end{pmatrix}$$

mit $q = q_1 \in Q = (-\varepsilon, 2\pi) \subset \mathbb{R}$ für beliebiges $\varepsilon > 0$. Beachte, dass q hier gerade den Winkel des Pendels beschreibt, also gerade gleich dem θ in unserem Pendelmodell aus Abschnitt B.1.4 ist.

Wir können das System nun vollständig mittels $q(t)$ beschreiben. Mittels der Kettenregel kann man die Geschwindigkeit über $q(t)$ ausdrücken. Es gilt

$$v_i(t) = \frac{d}{dt} r_i(q(t),t) = \sum_{j=1}^{l} \frac{\partial r_i}{\partial q_j}(q(t),t)\dot{q}_j(t) + \frac{\partial r_i}{\partial t}(q(t),t), \quad i = 1, \ldots, N.$$

Diese Gleichung kann wegen der linearen Unabhängigkeit der partiellen Ableitungen nach $\dot{q}_j$ aufgelöst werden, was i.A. aber nicht explizit durchgeführt werden muss. Die Größen $\dot{q}_1, \ldots, \dot{q}_l$ heißen *verallgemeinerte Geschwindigkeiten*.

Beispiel: Für das Pendel gilt

$$v(t) = \begin{pmatrix} \rho \cos q(t) \\ \rho \sin q(t) \\ 0 \end{pmatrix} \dot{q}(t),$$

Ebenso kann die kinetische Energie mittels q und $\dot{q}$ als

$$E_{kin} = \sum_{i=1}^{N} \frac{m_i}{2} \|v_i\|^2 = \sum_{i=1}^{N} \frac{m_i}{2} \left\| \sum_{j=1}^{l} \frac{\partial r_i}{\partial q_j}(q(t),t)\dot{q}_j(t) + \frac{\partial r_i}{\partial t}(q(t),t) \right\|^2 =: \mathcal{T}(q(t), \dot{q}(t), t)$$

geschrieben werden.

Beispiel: Für das Pendel gilt

$$\mathcal{T}(q(t), \dot{q}(t), t) = \frac{m}{2} \rho^2 \dot{q}(t)^2$$

Im Folgenden werden wir q und $\dot{q}$ oft als unabhängige Variablen auffassen. Wir lassen in dem Fall das zeitliche Argument weg, schreiben also z.B. $\mathcal{T}(q, \dot{q}, t)$.

Für Kräfte $f^i \in \mathbb{R}^3$, $i = 1, \ldots, N$, die jeweils auf den i–ten Massepunkt wirken, definiert man die *verallgemeinerten Kräfte*

$$F_j = \sum_{i=1}^{N} \left\langle f^i, \frac{\partial r_i}{\partial q_j} \right\rangle, \quad j = 1, \ldots, l.$$

Wir nennen das mechanische System *konservativ*, falls eine reelle Funktion $W(r_1, \ldots, r_N, t)$ existiert, so dass

$$f^i = -\frac{\partial W}{\partial r_i} =: -\nabla_i W$$

gilt. Für die verallgemeinerten Kräfte berechnet man dann

$$F_j = -\frac{\partial W(q,t)}{\partial q_j}$$

mit $W(q,t) = W(r(q),t)$. In Vektorform schreiben wir $F = -\nabla_q W(q,t)$. Die Funktion W kann physikalisch als die potentielle Energie des Systems interpretiert werden, weswegen man üblicherweise durch Addition einer geeigneten Konstanten die Bedingung $\min_q W(q,t) = 0$ sicher stellt. Beachte, dass die Addition einer Konstanten an $\nabla_q W$ nichts ändert.

Beispiel: Beim Pendel ohne Reibung wirkt auf den Massenpunkt die Kraft $f = (0, -mg, 0)^T$, die sich als $f = -\nabla W$ mit $W(r) = mgy$ schreiben lässt. Mit der oben eingeführten Darstellung $r(q) = (\rho \sin q, -\rho \cos q, 0)^T$ gilt $W(q,t) = -mg\rho \cos q$. Um $\min_q W(q,t) = 0$ zu gewährleisten, addieren wir $mg\rho$, d.h. wir setzen $W(q,t) = -mg\rho \cos q + mg\rho$.

Definition B.5 Die Funktion

$$L(q, \dot{q}, t) = \mathcal{T}(q, \dot{q}, t) - W(q,t)$$

heißt *Lagrange–Funktion* des konservativen mechanischen Systems. □

Die Variablen q und $\dot{q}$ werden hier als (formal) unabhängige Variablen aufgefasst.

Aus der Lagrange–Funktion kann man nun die Bewegungsgleichungen des Systems herleiten: Es gelten die *Lagrange–Gleichungen*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}(q(t), \dot{q}(t), t)\right) - \frac{\partial L}{\partial q_j}(q(t), \dot{q}(t), t) = 0, \quad j = 1, \ldots, l. \tag{B.15}$$

Die Herleitung dieser Gleichungen ergibt sich aus der physikalischen Bedingung, dass das *Wirkungsfunktional*

$$I(q) = \int_{t_0}^{t_1} L(q(t), \dot{q}(t), t)dt$$

entlang von Lösungen q minimal sein muss bzgl. aller differenzierbarer Funktionen, die die Punkte $(t_0, q(t_0))$ und $(t_1, q(t_1))$ verbinden. Setzt man $g(\alpha) = I(q + \alpha z)$ für $\alpha \in \mathbb{R}$ und eine

beliebige differenzierbare Funktion z mit $z(t_0) = z(t_1) = 0$, so muss $g'(0) = 0$ gelten. Mit etwas Rechnung sieht man, dass

$$g'(0) = \sum_{j=1}^{l} \int_{t_0}^{t_1} \left(\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_j}(q(t), \dot{q}(t), t) \right) - \frac{\partial L}{\partial q_j}(q(t), \dot{q}(t), t) \right) z(t) dt$$

ist, was schließlich auf die Lagrange–Gleichungen führt.

Beispiel: Für das Pendel ohne Reibung erhalten wir aus den obigen Überlegungen

$$L(q, \dot{q}, t) = \frac{m}{2} \rho^2 \dot{q}^2 + mg\rho \cos q - mg\rho,$$

also

$$\frac{\partial L}{\partial \dot{q}}(q, \dot{q}, t) = \frac{m}{2} \rho^2 2\dot{q} = m\rho^2 \dot{q}$$

und

$$\frac{\partial L}{\partial q}(q, \dot{q}, t) = -mg\rho \sin q.$$

Damit erhalten wir die Bewegungsgleichung

$$\begin{aligned} 0 &= \frac{d}{dt}(m\rho^2 \dot{q}(t)) + mg\rho \sin q(t) \\ &= m\rho^2 \ddot{q}(t) + mg\rho \sin q(t). \end{aligned}$$

Da $\rho > 0$ und $m > 0$ ist, vereinfacht sich diese zu

$$0 = \rho \ddot{q}(t) + g \sin q(t),$$

was gerade die in Abschnitt B.1.4 hergeleitete Gleichung (mit $q = \theta$) ist.

B.2.2 Dissipative Systeme

Wir haben die Lagrange–Gleichungen unter der Annahme hergeleitet, dass das mechanische System konservativ ist. Tatsächlich bedeutet dies, dass die wirkenden Kräfte nur von den Positionen r_i abhängen; weder externe noch geschwindigkeiteabhängige Kräfte (wie die Reibung) können hiermit modelliert werden.

Externe Kräfte können — wenn sie in verallgemeierter Form F_j^e also mittels Ihrer Wirkung auf die q ausgedrückt sind — einfach durch Ersetzen der "0" durch F_j^e auf der rechten Seite von (B.15) eingeführt werden. Die Umrechnung von physikalischen externen Kräften f_e^i auf verallgemeinerte externe Kräfte F_j^e erfolgt dabei analog zu den konservativen Kräften mittels

$$F_j^e(q, t) = \sum_{i=1}^{N} \left\langle f_e^i(r(q), t), \frac{\partial r_i}{\partial q_j}(q) \right\rangle, \quad j = 1, \ldots, l.$$

Bei der Reibung beschränken wir uns auf den einfachen Fall viskoser Reibung und bezeichnen mit $f^i \in \mathbb{R}^3$ nun die Reibungskräfte, die durch $f^i = -C^i v_i$ mit einer Diagonalmatrix $C^i = \mathrm{diag}(c_x^i, c_y^i, c_z^i) \in \mathbb{R}^{3\times 3}$ und den Geschwindigkeitsvektoren $v_i = (v_x^i, v_y^i, v_z^i)^T$ allgemein beschrieben werden können.

Wir wollen die Reibungskräfte analog zu den konservativen Kräften als Ableitung einer reellwertigen Funktion darstellen. Dazu definiert man die sogenannte *Ragleigh'sche Dissipationsfunktion*

$$\mathcal{D}(v_1, \ldots, v_N) = \frac{1}{2} \sum_{i=1}^{N} (c_x^i {v_x^i}^2 + c_z^i {v_z^i}^2 + c_z^i {v_z^i}^2)$$

und definiert die Reibungskräfte f^i als

$$f^i = -\nabla_i \mathcal{D},$$

wobei $\nabla_i \mathcal{D}(v) \in \mathbb{R}^3$ den Gradienten von $\mathcal{D}$ nach v^i bezeichnet.

Die Ragleigh–Funktion lässt sich als infinitesimale Arbeit des i–ten Partikels gegen die Reibungskraft interpretieren. Die von den Reibungskräften absorbierte Leistung ist gerade $2\mathcal{D}$ und wird als *Dissipationsrate* bezeichnet. Analog zu den verallgemeinerten Kräften lassen sich die *verallgemeinerten Reibungskräfte* als

$$F_j = -\frac{\partial \mathcal{D}(q, \dot{q}, t)}{\partial \dot{q}}$$

mit $\mathcal{D}(q, \dot{q}, t) = \mathcal{D}(v(q, \dot{q}, t))$ berechnen. In Kurzform schreiben wir

$$F = -\nabla_{\dot{q}} \mathcal{D}(q, \dot{q}, t).$$

Die unter der Berücksichtigung der externen und Reibungskräfte erhaltenen *verallgemeinerten Lagrange–Gleichungen* lauten damit

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}(q(t), \dot{q}(t), t)\right) - \frac{\partial L}{\partial q_j}(q(t), \dot{q}(t), t) + \frac{\partial \mathcal{D}}{\partial \dot{q}_j}(q(t), \dot{q}(t), t) = F_j^e(t), \quad j = 1, \ldots, l \quad \text{(B.16)}$$

Beispiel: Wir wollen unser Pendelmodell um einen Reibungsterm ergänzen und setzen $\mathcal{D}(v) = \frac{1}{2}(cv_x^2 + cv_y^2)$. Damit erhalten wir

$$\mathcal{D}(q, \dot{q}, t) = \frac{c}{2}\rho^2(\cos^2 q + \sin^2 q)\dot{q}^2 = \frac{c}{2}\rho^2\dot{q}^2$$

und folglich

$$\frac{\partial \mathcal{D}}{\partial \dot{q}}(q(t), \dot{q}(t), t) = c\rho^2\dot{q}.$$

Die Bewegungsgleichung ergibt sich damit zu

$$0 = m\rho^2\ddot{q}(t) + mg\rho \sin q(t) + c\rho\dot{q}.$$

Wir erhalten also wieder das bereits bekannte Modell, bei dem die Reibungskonstante c nun allerdings mit der Länge ρ multipliziert ist. Dies liegt daran, dass die Reibung hier an der Punktmasse wirkt, während sie im früheren Modell an der Drehachse wirkt.

B.2.3 Die Hamilton'sche Methode

Die Lagrange–Gleichungen führen in natürlicher Weise auf eine Differentialgleichung zweiter Ordnung, also eine Gleichung, in der q und $\dot{q}$ auftreten. Der Hamilton–Formalismus, den wir abschließend kurz behandeln wollen, ermöglicht es, für konservative mechanische Systeme direkt ein System erster Ordnung herzuleiten, in dem die dabei verwendete "Hilfsfunktion" H eine wohldefinierte physikalische Interpretation besitzt.

Hierzu definieren wir das *verallgemeinerte Moment* als

$$p = \frac{\partial L}{\partial \dot{q}}(q, \dot{q}, t) \in \mathbb{R}^l$$

und nehmen an, dass eine stetig differenzierbare Funktion $\dot{q}(q, p, t)$ existiert, so dass die Gleichung

$$p = \frac{\partial L}{\partial \dot{q}}(q, \dot{q}(q, p, t), t)$$

gilt. Die Abbildung

$$(q, p, t) \mapsto (q, \dot{q}(q, p, t), t)$$

heißt dabei *Legendre–Transformation.*

Definition B.6 Die reellwertige Funktion

$$H(q, p, t) = p^T \dot{q}(q, p, t) - L(q, \dot{q}(q, p, t), t)$$

heißt *Hamilton–Funktion* eines konservativen mechanischen Systems. □

Beispiel: Für das Pendel gilt

$$p = \frac{\partial L}{\partial \dot{q}} L(q, \dot{q}, t) = m\rho^2 \dot{q},$$

also ist

$$\dot{q}(q, p, t) = \frac{p}{m\rho^2}$$

wegen

$$\frac{\partial L}{\partial \dot{q}} L(q, \dot{q}(q, p, t), t) = m\rho^2 \dot{q}(q, p, t) = m\rho^2 \frac{p}{m\rho^2} = p$$

die gesuchte Abbildung mit Legendre–Transformation

$$(q, p, t) \mapsto (q, p/(\rho^2 m), t).$$

Die Hamilton–Funktion unseres Pendels lautet demnach

$$H(q, p, t) = \frac{p^2}{m\rho^2} - \frac{m}{2}\rho^2 \left(\frac{p}{m\rho^2}\right)^2 - mg\rho\cos q + mg\rho = \frac{1}{2}\frac{p^2}{m\rho^2} - mg\rho\cos q + mg\rho.$$

Sei nun $q(t)$ eine Lösung von (B.15) mit

$$p(t) = \frac{\partial L}{\partial \dot{q}}(q(t), \dot{q}(t), t).$$

Dann folgt aus der Definition von $\dot{q}(q,p,t)$ die Identität

$$\dot{q}(q(t),p(t),t) = \dot{q}(t).$$

Für H gilt nun

$$\begin{aligned}\frac{\partial H}{\partial p_j}(q(t),p(t),t) &= \dot{q}_j(t) + p^T \frac{\partial \dot{q}}{\partial p_j}(q(t),p(t),t) - \sum_{i=1}^{l} \underbrace{\frac{\partial L}{\partial \dot{q}_i}(q(t),\dot{q}(t),t)}_{=p_i(t)} \frac{\partial \dot{q}_i}{\partial p_j}(q(t),p(t),t) \\ &= \dot{q}_j(t)\end{aligned}$$

für $j = 1,\ldots,l$ und

$$\begin{aligned}\frac{\partial H}{\partial q_j}(q(t),p(t),t) &= p^T \frac{\partial \dot{q}}{\partial q_j}(q(t),p(t),t) - \frac{\partial L}{\partial q_j}(q(t),\dot{q}(t),t) \\ &\quad - \underbrace{\sum_{i=1}^{l} \frac{\partial L}{\partial \dot{q}_i}(q(t),\dot{q}(t),t)}_{=p_i(t)} \frac{\partial \dot{q}_i}{\partial q_j}(q(t),p(t),t) \\ &= -\frac{\partial L}{\partial q_j}(q(t),\dot{q}(t),t) = -\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}(q(t),\dot{q}(t),t)\right) \\ &= -\dot{p}_j(t)\end{aligned}$$

Also erfüllt die Funktion $(q(t),p(t))$ die DGL erster Ordnung

$$\begin{aligned}\dot{q}(t) &= \frac{\partial H}{\partial p}(q(t),p(t),t) \\ \dot{p}(t) &= -\frac{\partial H}{\partial q}(q(t),p(t),t)\end{aligned}$$

das sogenannte *Hamilton–System*.

Umgekehrt kann man nachweisen, dass jede Lösung des Hamilton–Systems eine Lösung der Lagrange–Gleichungen induziert. Die zwei Systeme sind also äquivalent.

Die Hamilton–Funktion ist deswegen eine schöne Form der Gleichung, da sie (in vielen Fällen) eine explizite physikalische Interpretation besitzt: $H(q,p,t)$ ist gerade die Gesamtenergie des Systems, die — aufgrund der Konservativität des Systems — entlang von Lösungen konstant ist.

Beispiel: Für das Pendel gilt

$$\frac{\partial H}{\partial p}(q,p,t) = \frac{p}{m\rho^2}$$

und

$$\frac{\partial H}{\partial q}(q,p,t) = mg\rho \sin q.$$

Wir erhalten also das Hamilton–System

$$\begin{aligned}\dot{q}(t) &= \frac{p(t)}{m\rho^2} \\ \dot{p}(t) &= -mg\rho \sin q(t)\end{aligned}$$

Dies ist genau das bekannte Modell (B.9), wenn wir $\theta = q$ und $\omega = p/(m\rho^2)$ setzen. Der Vorteil der hier erhaltenen Skalierung liegt darin, dass die Hamilton–Funktion

$$H(q,p,t) = \frac{1}{2}\frac{p^2}{m\rho^2} - mg\rho\cos q + mg\rho$$

hier tatsächlich die Gesamtenergie des Systems beschreibt, die gerade die Summe der kinetischen Energie $\frac{1}{2}\frac{p^2}{m\rho^2}$ und der potentiellen Energie $-mg\rho\cos q + mg\rho$ ist.

Literaturverzeichnis

[1] Aulbach, B.: *Gewöhnliche Differenzialgleichungen.* 2. Auflage. Elsevier–Spektrum Verlag, Heidelberg, 2004

[2] Deuflhard, P. ; Bornemann, F.: *Numerische Mathematik. II: Integration gewöhnlicher Differentialgleichungen.* 4. Auflage. de Gruyter, Berlin, 2013

[3] Grüne, L. ; Junge, O.: *Gewöhnliche Differentialgleichungen. Eine Einführung aus der Perspektive der Dynamischen Systeme.* Vieweg + Teubner Verlag, 2009

[4] Hairer, E. ; Lubich, C. ; Wanner, G.: *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations.* 2nd edition. Springer-Verlag, Berlin, 2006

[5] Hairer, E. ; Wanner, G.: *Solving ordinary differential equations. II. Stiff and differential-algebraic problems.* 2nd edition. Springer-Verlag, Berlin, 1996