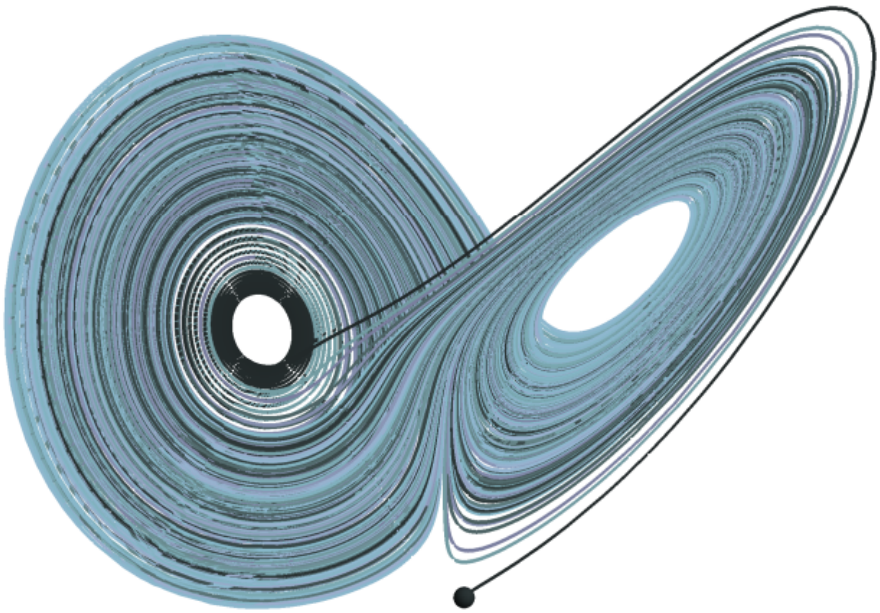
Julius-Maximilians-Universität Würzburg
Institut für Mathematik
Lehrstuhl IX (Wissenschaftliches Rechnen)

Modellierung und Wissenschaftliches Rechnen



Ausgearbeitete Mitschrift zur Vorlesung
im Wintersemester 2016/2017
von Professor Dr. Alfio Borzi,
erstellt von Richard Schmähl

Würzburg 2017

Inhaltsverzeichnis

1 Gewöhnliche Differentialgleichungen

Eine (gewöhnliche) ***Differentialgleichung*** (DGL, englisch: ode) ist eine Gleichung, welche eine Funktion sowie deren Ableitung(en), enthält. Anders ausgedrückt ist eine DGL eine Gleichung der Form

$$F\left(x, y, y', \ldots, y^{(n)}\right) = 0,$$

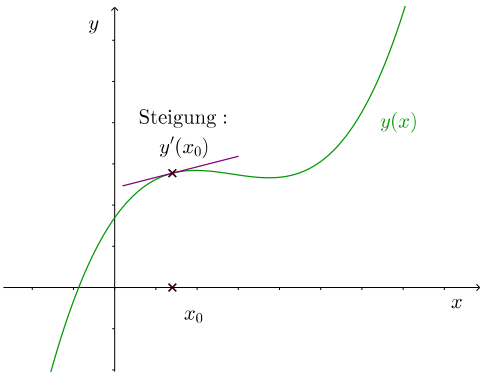
wobei $y = y(x)$ eine von x abhängige C^n-Funktion auf einem Intervall I ist (d. h. $y \in C^n(I)$) und $y', y'', \ldots, y^{(n)}$ dementsprechend die Ableitungen $y' = \frac{\mathrm{d}}{\mathrm{d}x}y, y'' = \frac{\mathrm{d}^2}{\mathrm{d}x^2}y, \ldots, y^{(n)} = \frac{\mathrm{d}^n}{\mathrm{d}x^n}y$ sind. n nennt man die ***Ordnung*** der Differentialgleichung. Hat eine Differentialgleichung die Gestalt

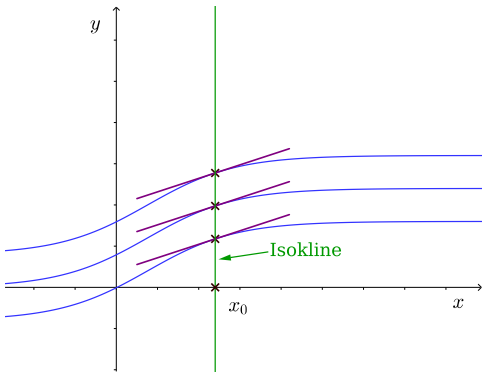
$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right),$$

so liegt sie in ***Normalform*** vor. Eine Funktion y, welche eine solche Differentialgleichung erfüllt, heißt naheliegenderweise ***Lösung*** der Differentialgleichung. Die Lösungskurven einer Differentialgleichung nennt man ***Trajektorien***. Es kann durch aus mehrere geben, welche dann eine ***Schar*** bilden. Da für Differentialgleichungen ***Anfangswerte*** $y(x_0) = y_0$, $y'(x_0) = y'_0$, ..., $y_0^{(n-1)}(x_0) = y_0^{(n-1)}$) vorgegeben sein können, bezeichnet man zur Unterscheidung die ***allgemeine Lösung*** (ohne diese) mittels $y(x; c_1, ..., c_n)$ mit freien, aber konstanten Parametern $c_1, ..., c_n$.
Als erste „Grobeinteilung“ für eine Differentialgleichung erster Ordnung in Normalform lassen sich folgende Fälle betrachten:

$$y'(x) = f(x, y(x))$$

$$y' = f(x) \Rightarrow y(x) = \int_{x_0}^{x} f(t)\,\mathrm{d}t + c$$

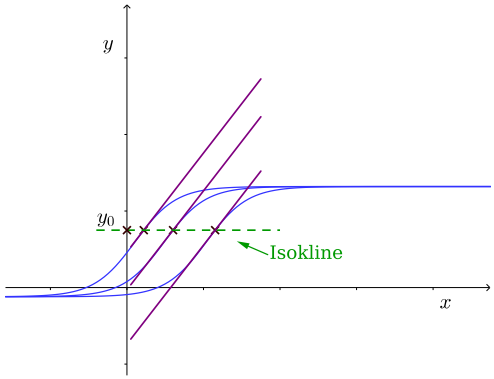




Dies ist eine Differentialgleichung erster Ordnung in ihrer allgemeinsten Form, y' hängt sowohl von y als auch von x ab.

Hieran sieht man, dass die allgemeine Lösung einer Differentialgleichung im Allgemeinen nicht eindeutig ist, sondern beispielsweise nur bis auf additive Konstanten.

$$y' = f(y)$$

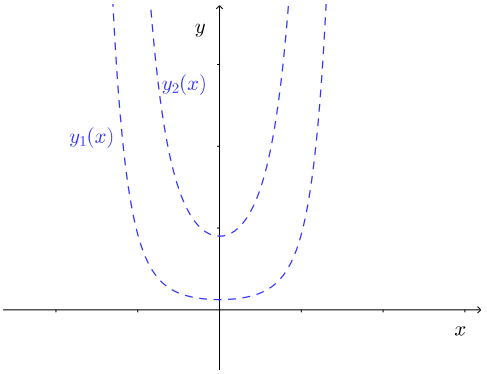


Differentialgleichungen dieser Form, bei denen f nicht von der freien Variablen abhängt, heißen ***autonom***.

Die sogenannten „Isoklinen“ sind bei einer solchen graphischen Darstellung Kurven, entlang denen die Steigung $y'(x)$ konstant bleibt. Das Tripel $(x, y, y') = (x, y, f(x, y))$ ist ein ***Linienelement***, die Menge aller Linienelemente heißt ***Richtungsfeld***. Richtungsfelder erlauben eine geometrische Lösung: In einem Punkt (x_0, y_0) wird die Lösungskurve durch diesen Punkt durch eine Gerade durch diesen Punkt mit Steigung $f(x_0, y(x_0))$ approximiert. Folgt man nun den durch das Richtungsfeld vorgegeben Richtungen, erhält man nicht unbedingt die exakte Lösungskurve, aber eine gute Vorstellung, wie diese ungefähr aussieht.

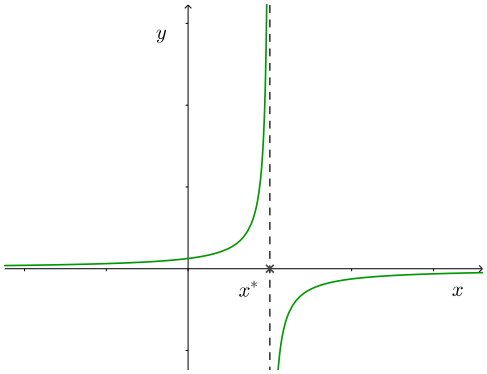
Die bisherigen Skizzen sind jedoch nur exemplarisch, bei einer qualitativen Analyse von Differentialgleichungen lassen sich verschiedenste Verhalten der Lösung(en) beobachten, wie folgende kleine Auswahl zeigt:

$$y' = xy$$



In diesem Falle wäre die allgemeine Lösung zum Beispiel $y(x; c) = c \cdot e^{\frac{x^2}{2}}$.

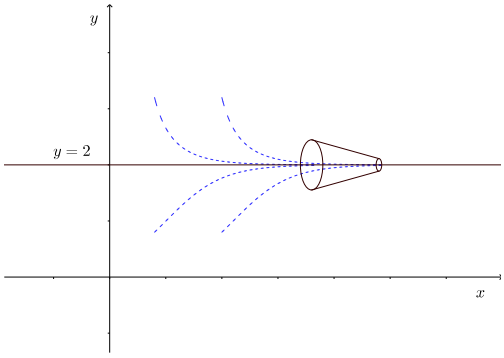
$$y' = ky^2, k > 0$$



Hier ist die allgemeine Lösung $y(x; c) = \frac{1}{c-kx}$, die Lösungskurve hat eine vertikale Asymptote bei $x = \frac{c}{k}$.

$$y' = 2y - y^2$$

$\rightarrow$ Nullstelle bei $y = 2$ (Ruhelage)



Eine solche Situation, bei der sich die Lösungskurven der Kurve einer Lösung $\bar{y}$ (hier $\bar{y} \equiv 2$) asymptotisch ($x \to \infty$) annähern wird Trichter genannt. Entfernen sie sich, liegt ein Antitrichter vor.

Erfüllt eine Funktion die Differentialgleichung $y'(x) = f(x, y(x))$, $x \in I$ (I Intervall) nicht, kann sie immer noch einer Art „Differentialungleichung“ genügen: Ist für eine Funktion α $\alpha(x) \leq f(x, \alpha(x))$, $x \in I$, so ist α ***Unterfunktion***. Eine Funktion β, welche $\beta(x) \geq f(x, \beta(x))$

erfüllt, heißt entsprechend ***Oberfunktion***. Gilt:

- $\alpha(x) < \beta(x)$ auf einem Intervall I, so bildet der Bereich $\alpha(x) \leq y \leq \beta(x)$ zusammen mit I einen Trichter.
- $\alpha(x) > \beta(x)$ auf einem Intervall I, so bildet der Bereich $\alpha(x) \geq y \geq \beta(x)$ zusammen mit I einen Antitrichter.
- $\alpha(x) < \beta(x)$ (Trichter) und ist $y'(x) = f(x, y(x))$ Lösung der DGL mit $\alpha(x^*) < y(x^*) < \beta(x^*)$ für ein x^*, so ist $\alpha(x) < y(x) < \beta(x)$ für alle $x \geq x^*$.
- $\alpha(x) > \beta(x)$ (Antitrichter), so existiert (mindestens) eine Lösung $y(x)$ von $y' = f(x, y)$, welche $\beta(x) \leq y(x) \leq \alpha(x)$ für alle $x \in I$ erfüllt.

Eine genauere Betrachtung des Verhalten von Lösungen, speziell von deren Stabilität, erfolgt jedoch erst später. Vorher soll erarbeitet werden, wie derartige Gleichungen gelöst werden können. Als Referenz-Literatur sei auf die Quellen [1]-[6] verwiesen.

1.1 Elementar lösbare Differentialgleichungen

Im Folgenden sollen verschiedene Klassen elementar lösbarer Differentialgleichungen sowie ihre (allgemeinen) Lösungen betrachtet werden.

(1) $y' = f(x)$: Der Lösungsansatz „$F(x) + c = \int f(x)\mathrm{d}x$“ führt auf $y(x; c) = F(x) + c$, wobei c konstant und $F(x)$ beliebige eine Stammfunktion zu $f(x)$ ist.

(2) $y' = f(y)$: Der Lösungsansatz „$\mathrm{d}y(x) = f(y(x))\mathrm{d}x$“ führt $\left(\text{„}\int \frac{1}{f(y)}\mathrm{d}y = \int \mathrm{d}x\text{“}\right)$ auf $F(y) = x + c$, wobei c erneut konstant sowie $F(y)$ beliebige Stammfunktion zu $\frac{1}{f(y)}$ ist.

(3) $y'(x) = f(x)\, g(y)$: Der Lösungsansatz „$\int \frac{1}{g(y)}\mathrm{d}y = \int f(x)\,\mathrm{d}x$“ führt auf $G(y) = F(x) + c$ mit beliebigen Stammfunktionen $F(x), G(x)$ von $f(x)$ beziehungsweise $\frac{1}{g(x)}$. c bezeichne wie gehabt eine Konstante.

(4) $y'(x) = f(a + by + c)$: Durch die Substitution $u = ax + by + c$ erhält man die Differentialgleichung $u' = a + by' = a + bf(u)$. Der Lösungsansatz „$\int \frac{1}{a+bf(u)}\mathrm{d}u = \int \mathrm{d}x$“ liefert (siehe (1)) $F(u) = x + k$ mit $F(u)$ Stammfunktion zu $\frac{1}{a+bf(u)}$, nach Rücksubstitution folglich $F(ax + by + c) = x + k$.

(5) $y' = f\left(\frac{y}{x}\right)$: Durch die Substitution $u = \frac{y}{x}$ (d. h. $y = ux$ und somit $y' = u' \cdot x + u$) erhält man für u die Differentialgleichung $f(u) = u'x + u$ beziehungsweise $u'x = f(u) - u$. Bezeichnet $F(u)$ eine Stammfunktion von $\frac{1}{f(u)-u}$, so erhält man analog zu (3) mittels „$\int \frac{1}{x}\mathrm{d}x = \int \frac{1}{f(u)-u}\mathrm{d}u$“ als Lösung $\log(x) + c = F(u)$ und somit $F(\frac{y}{x}) = \log(x) + c$.

Beispiele:

- (Zu (4)) Gegeben sei

$$y'(x) = (x + y)^2$$

Substitution: $u = x + y \quad \Rightarrow \quad u' = 1 + y' = 1 + u^2$

Ansatz:

$$\text{„}\int \frac{1}{1 + u^2}\mathrm{d}u = \int \mathrm{d}x\text{“}$$

Lösung (siehe (4)):

$$\arctan u = x + c \quad \Rightarrow \quad u = \tan(x + c) \quad \Rightarrow \quad \underline{\underline{y = \tan(x + c) - x}}$$

- (Zu (5)) Gegeben sei

$$y' = \frac{y}{x} - \frac{x^2}{y^2}$$

Substitution: $u = \frac{y}{x} \overset{\text{für } x\neq 0}{\Rightarrow} y = ux$
Folglich (wegen $y' = u'x + u$ und somit $u - \frac{1}{u^2} = u'x + u$):

$$-u^2u' = \frac{1}{x} \quad \text{bzw} \quad u' = -\frac{1}{xu^2}$$

Lösung (siehe (3)):

$$-\frac{u^3}{3} = \log x + c \quad \Rightarrow \quad u^3 = -3\log x - 3c \quad \Rightarrow$$

$$u = -\sqrt[3]{3\log x + 3c} \quad \Rightarrow \quad \underline{\underline{y = -x\sqrt[3]{3\log x + 3c}}}$$

Ein weiterer elementar lösbarer Differentialgleichungstyp ist der Folgende:

(6) $\underline{y' = p(x)y + q(x)}$: Eine Gleichung dieser Form nennt man ***inhomogene, lineare Differentialgleichung***. Ihre Lösung setzt sich aus zwei Komponenten zusammen:

a) Der Lösung der ***homogenen*** Gleichung $y' = p(x)y$: Der aus (3) bekannte Ansatz „$y_p(x) = \int \frac{1}{y}\mathrm{d}y = \int p(x)\mathrm{d}x$“ führt auf $\log y = \int_{x_0}^{x} p(x)\mathrm{d}x + c$. Somit lautet die Lösung

$$y = e^c e^{\int_{x_0}^{x} p(x)\mathrm{d}x}$$

beziehungsweise (nach „Umtaufen“ von c und Definieren von $P(x) := \int_{x_0}^{x} p(x)\mathrm{d}x$)

$$y_{\text{hom.}}(x; c) = c \cdot e^{P(x)}$$

Allgemeine Lösung der homogenen Gleichung

(Streng genommen wurde bisher nur gezeigt, dass Funktionen dieser Form Lösungen der homogenen DGL. darstellen. Betrachtet man jedoch zwei C^1-Funktionen u_1, u_2 mit $u_2 \neq 0$, so lässt sich durch Differentiation des Quotienten leicht verifizieren, dass tatsächlich jede Lösung diese Form hat.)

b) Einer partikulären Lösung y_p von

$$y_p' = p(x)y_p + q(x)$$

Da für zwei Lösungen y_1, y_2 die Differenz stets die entsprechende homogene Gleichung erfüllt, genügt es, eine spezielle Lösung y_p zu kennen, um alle Lösungen darstellen zu können. Eine solche findet man über die sogenannte ***Methode der Variation der Konstanten***: Der Ansatz hierfür lautet „$c(x) \cdot e^{\int p(x)\mathrm{d}x}$“. (In diesem Falle hängt die „Konstante“ c also von x ab.)
Durch Ableiten erhält man:

$$y_p'(x) = c'(x) \cdot e^{\int_{x_0}^{x} p(x)\mathrm{d}x} + c(x)p(x)e^{\int_{x_0}^{x} p(x)\mathrm{d}x} = \\ = p(x)y_p(x) + q(x)$$

Wegen $c(x)p(x)e^{\int_{x_0}^{x} p(x)\mathrm{d}x} = p(x)y_p(x)$ folgt hieraus $c'(x) = e^{-\int_{x_0}^{x} p(x)\mathrm{d}x}q(x)$ und somit:

$$c(x) = \int_{x_0}^{x} q(t)e^{-\int_{x_0}^{t} p(s)\mathrm{d}s}\mathrm{d}t$$

Kontrollrechnung:

$$y_p(x) = \int_{x_0}^{x} q(t) e^{-\int_{x_0}^{t} p(s)\mathrm{d}s} \mathrm{d}t \cdot e^{\int_{x_0}^{x} p(x)\mathrm{d}x}$$

$$\Downarrow$$

$$\begin{aligned} y_p'(x) &= \int_{x_0}^{x} q(t) e^{-\int_{x_0}^{t} p(s)\mathrm{d}s} \mathrm{d}t \cdot e^{\int_{x_0}^{x} p(x)\mathrm{d}x} \cdot p(x) + q(x) \cdot e^{-\int_{x_0}^{x} p(x)\mathrm{d}x} \cdot e^{\int_{x_0}^{x} p(x)\mathrm{d}x} \\ &= p(x) y_p(x) + q(x) \end{aligned}$$

Aus diesen beiden Teilen ergibt sich als allgemeine Lösung:

$$\begin{aligned} y_{\text{inh.}}(x;c) &= y_{\text{hom.}} + y_p = \\ &= c \cdot e^{\int_{x_0}^{x} p(x)\mathrm{d}x} + e^{\int_{x_0}^{x} p(x)\mathrm{d}x} \cdot \int_{x_0}^{x} q(t) e^{-\int_{x_0}^{t} p(s)\mathrm{d}s} \mathrm{d}t = \\ &= e^{\int_{x_0}^{x} p(x)\mathrm{d}x} \left(c + \int_{x_0}^{x} q(t) e^{-\int_{x_0}^{t} p(s)\mathrm{d}s} \mathrm{d}t \right) \end{aligned}$$

beziehungsweise, mit den Bezeichnungen $P(x) = \int_{x_0}^{x} p(x)\mathrm{d}x$ und $\tilde{Q}(x) = \int_{x_0}^{x} q(s) e^{-P(s)}\mathrm{d}s$,

Allgemeine Lösung der inhomogenen Gleichung
$y_{\text{inh.}}(x;c) = e^{P(x)} \cdot \left(c + \tilde{Q}(x) \right)$

Beispiel:

$$y' = \underbrace{a}_{=p(x)} y + \underbrace{b}_{=q(x)}, \quad a, b \text{ konstant}$$

Homogene Gleichung: $y' = ay$ mit allgemeiner Lösung $y(x) = ce^{ax}$ (siehe (1))
Partikuläre Lösung:

$$\begin{aligned} y_p(x) &= e^{\int_0^x p(x)\mathrm{d}x} \cdot \int_{x_0}^{x} q(t) e^{-\int_0^t p(s)\mathrm{d}s} \mathrm{d}t \\ &= e^{\int_0^x a\mathrm{d}x} \cdot \int_0^x b e^{-\int_0^t a\mathrm{d}s} \mathrm{d}t \\ &= e^{ax} \left(-\frac{b}{a} e^{-ax} + \frac{b}{a} \right) \end{aligned}$$

$$\Downarrow$$

$$\underline{\underline{y(x;c) = \left(c - \frac{b}{a} e^{-ax} \right) e^{ax}}}$$

(Der Summand $\frac{b}{a}$ kann in der allgemeinen Lösung weggelassen werden, da er als Teil der additiven Konstanten c gesehen werden kann.)

Der in (6) vorgestellte Lösungsweg ist zwar für Differentialgleichungen der Art $y'(x) = p(x)y + q(x)$ der am vielseitigsten Verwendbare, jedoch nicht immer der einfachste. Für einige Klassen von Fällen lässt sich die Lösung mithilfe der ***Methode der unbestimmten Koeffizienten*** schneller bestimmen, falls p konstant ist. Je nach $q(x)$ lautet das allgemeine Schema hierfür:

$q(x)$	$y_p(x)$
$\alpha \sin x + \beta \cos x$	$a \sin x + b \cos x$
$\alpha + \beta x + \gamma x^2 + ...$	$a + bx + cx^2 + ...$
$\alpha e^{\beta x}$	$\begin{cases} a \cdot e^{\beta x}, & \beta \neq p \\ \alpha x \cdot e^{px}, & \beta = p \end{cases}$

Die Koeffizienten lassen sich anschließend durch die Bedingung $y_p' = py_p + q(x)$ bestimmen. Einige Beispiele sind:

$y'(x) = ay(x) + q(x)$	$y_p(x)$
$y' = y + \sin x$	$-\frac{1}{2}(\sin x + \cos x)$
$y' = y + x^2$	$-(2 + 2x + x^2)$
$y' = e^{2x}$	$\frac{1}{2}e^{2x}$

1.2 Bernoulli- und Riccati-Differentialgleichung

Unter einer ***Bernoulli-Differentialgleichung*** versteht man eine Gleichung folgender Form:

$$y' + g(x)\, y + h(x)\, y^{\alpha} = 0, \alpha \neq 1, \alpha \neq 0$$

Multipliziert mit $(1-\alpha)\, y^{-\alpha}$ ergibt sich die Gleichung

$$(1-\alpha)\, y^{-\alpha} y' + g(x)\,(1-\alpha)\, y^{1-\alpha} + h(x)\,(1-\alpha) = 0$$

Diese kann wegen der Identität $(1-\alpha)y^{-\alpha}y' = (y^{1-\alpha})'$ mittels der Substitution $u = y^{1-\alpha}$ in eine inhomogene, lineare Differentialgleichung umgeschrieben werden:

$$u' + \underbrace{[g(x)\,(1-\alpha)]}_{-p(x)}\, u + \underbrace{[h(x)\,(1-\alpha)]}_{-q(x)} = 0$$

Hat man diese Gleichung auf einem der im vorherigen Kapitel beschriebenen Wege gelöst, erhält man durch anschließende Rücksubstitution von $y = u^{\frac{1}{1-\alpha}}$ eine (allgemeine) Lösung für y.

Beispiel 1.2.1:

$$y' + \left(\frac{1}{1+x}\right) y + (1+x)\, y^4 = 0$$

In diesem Falle ist $\alpha = 4$.
Multiplikation mit $-3y^{-4}$ liefert:

$$-3y^{-4}y' + \left(\frac{1}{1+x}\right) y + (1+x)y^4 = 0$$

Substitution $u = y^{-3}$:

$$u' - 3\left(\frac{1}{1+x}\right) u - 3\,(1+x) = 0$$

→ Dies lässt sich zu einer inhomogenen DGL umformulieren:

$$u' = 3\left(\frac{1}{1+x}\right) u + 3\,(1+x)$$

a) *Homogene Gleichung*: $u' = 3\left(\frac{1}{1+x}\right)u = 0$
Lösung (siehe (3)): $u = c\,(1+x)^3$

b) *Partikuläre Lösung*:
Der Ansatz $u_p(x) = c(x)\,(1+x)^3$ liefert (siehe (6)):

$$u_p(x) = -3(1+x)^2$$

Somit lautet die allgemeine Lösung für u : $u(x;c) = c(1+x)^3 - 3(1+x)^2$, die allgemeine Lösung für y entsprechend

$$\underline{\underline{y(x;c) = \frac{1}{\sqrt[3]{u(x;c)}} = \frac{1}{\sqrt[3]{c(1+x)^3 - 3(1+x)^2}}}}$$

Mit Hilfe dieser Gleichung kann man (häufig) auch Lösungen für eine weitere spezielle Gattung von Differentialgleichungen finden, die sogenannten ***Riccati-Differentialgleichungen***. Dies sind Gleichungen der Form

$$y' = g(x)y + h(x)y^2 = k(x)$$

Zu ihrer Lösung verfolgt man den Ansatz $y(x;c) = \varphi(x) + u(x;c)$, wobei u eine noch zu bestimmende Funktion ist und φ eine bereits bekannte, spezielle Lösung (welche man z. B. durch Raten gefunden hat). Da

$$\varphi' + u' + gy + hy^2 + gu + h\varphi^2 + hu^2 + 2\varphi uh = k$$

gelten soll, lässt sich das Problem durch Subtraktion von $\varphi' + g\varphi + h\varphi^2 = k$ auf das Lösen der Bernoulli-DGL

$$u' + gu + hu^2 + 2\varphi uh = u' + (g + 2\varphi h)u + hu^2 = 0$$

reduzieren. Lässt sich dies lösen, so hat man alle Lösungen gefunden.

Beispiel 1.2.2:

$$y' + \frac{2x+1}{x}y - \frac{1}{x}y^2 = x + 2$$

Eine Lösung ist die Funktion $\varphi(x) = x$
Ansatz: $y(x;c) = x + u(x;c)$. Reduziert auf eine Bernoulli-DGL mit $\alpha = 2$ ergibt sich:

$$u' + \left(\frac{2x+1}{x} + 2x\left(-\frac{1}{x}\right)\right)u - \frac{1}{x}u^2 =$$
$$u' + \frac{1}{x}u - \frac{1}{x}u^2 = 0$$

Die Substitution $z = \frac{1}{u}$ liefert: $z' - \frac{1}{x}z + \frac{1}{x} = 0$ Die allgemeine Lösung des homogenen Teils $z' - \frac{1}{x}z = 0$ ist $z(x) = cx$, eine partikuläre Lösung $z_p(x) = 1$. Somit lautet die allgemeine Lösung der inhomogenen DGL

$$z(x;c) = 1 + cx,$$

woraus sich durch Rücksubstitution $u(x;c) = \frac{1}{1+cx}$ ergibt. Die allgemeine Lösung der Riccati-Gleichung ist folglich

$$y(x;c) = x + \frac{1}{1+cx}$$

1.3 Implizite Differentialgleichungen erster Ordnung

Eine Differentialgleichung der Form

$$F(x, y, y') = 0$$

mit $F \in C(D), D \subseteq \mathbb{R}^3$ heißt ***implizite Differentialgleichung***. Jede Differentialgleichung in Normalform, d. h. der Gestalt $y' = f(x, y)$ (auch als explizite DGL bezeichnet), kann durch Subtraktion von $f(x, y)$ auf beiden Seiten in eine implizite Differentialgleichung umgewandelt werden. Umgekehrt ist dies jedoch nicht zwingend der Fall, wie das Beispiel

$$0 = F(x, y, y') = (y' - 1)^2 + y - x = 0 \tag{1.3.1}$$

zeigt. Sei nun $p \coloneqq y'$. Da $\nabla F(x, y, p) = (1, -1, 2(p - 1))$ gilt und somit $F_p(0, 0, 0) = -2$ und ferner $F \in C^\infty(\mathbb{R})$ ist, lässt sich der Satz über implizite Funktionen anwenden:

Satz 1.3.1 (Implizite Funktionen, $\mathbb{R}^n$):
Sei D offen $\mathbb{R}^m \times \mathbb{R}^n$ und $F \in C^k(D, \mathbb{R}^l)$. Ferner sei $(x_0, y_0) \in W$ $(x_0 \in \mathbb{R}^m, y_0 \in \mathbb{R}^n)$, so dass

(i) $F(x_0, y_0) = 0$

(ii) $\frac{\partial}{\partial y} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^l)$ ein Isomorphismus ist.

Dann existieren offene Umgebungen $U \subseteq \mathbb{R}^{m+n}$ von (x_0, y_0) und $V \subseteq \mathbb{R}^n$ von (x_0) sowie ein eindeutig bestimmte $g \in C^k(\mathbb{R}^m, \mathbb{R}^n)$, sodass

$$(x, y) \in U \wedge f(x, y) = 0$$

und

$$x \in V \wedge y = g(x)$$

äquivalente Aussagen sind. Ferner gilt

$$\frac{\partial}{\partial x} g(x) = -\left(\frac{\partial}{\partial y} F(x, g(x))\right)^{-1} \frac{\partial}{\partial x} F(x, g(x)), x \in V$$

Angewandt auf obiges Beispiel ($m = 2, n = 1, x \mathrel{\widehat{=}} (x, y), y \mathrel{\widehat{=}} p$) folgt wegen $F(0, 0, 0) = 0, F_p(0) = -2$, dass F in einem Ball $B_r(0, 0)$ $(r > 0)$ lokal mittels einer C^∞-Funktion g nach p auflösbar ist, d. h. dass $F(x, y, p) = 0$ eine Funktion $p = g(x, y)$ definiert. Aus der (globalen) Glattheit folgt insbesondere, dass g auf einem beliebigen Rechteck $R \coloneqq [-a, a] \times [-b, b]$, welches komplett in $B_r(0)$ enthalten ist, stetig differenzierbar ist. Wie in Kapitel 1.5 gezeigt werden wird, garantiert dies die Lösbarkeit der Differentialgleichung $F(x, y, y') = 0$ unter der Zusatzbedingung $y(0) = 0$ auf einem Intervall $J \subseteq [-a, a]$. Der Satz über implizite Funktionen hat jedoch ein kleines Manko: Weder er selbst noch sein Beweis sind konstruktiv. Durch ihn kann man unter gewissen Voraussetzungen nur die Existenz einer Lösung herleiten. Sollte y' nicht isolierbar sein, so wird diese Schwierigkeit, eine solche zu finden, durch ihn nicht behoben.

Nichtsdestotrotz heißt dies noch lange nicht, dass man keine Lösung der Differentialgleichung angeben kann. Ein methodischer Ansatz besteht darin, x und y in Abhängigkeit von p zu definieren, d. h. $x = x(p), y = y(p)$. Setzt man $\dot{x} \coloneqq \frac{\mathrm{d}}{\mathrm{d}p} x(p), \dot{y} \coloneqq \frac{\mathrm{d}}{\mathrm{d}p} y(p)$, so gilt: $\dot{y} = y' \cdot \dot{x} = p\dot{x}$. Ist F stetig differenzierbar, so ergibt Differentiation nach p

$$0 = F_x \dot{x} + F_y \dot{y} + F_p = F_x \dot{x} + F_y \dot{y} + F_p$$

Wegen $\dot{y} = p\dot{x}, \dot{x} = \frac{\dot{y}}{p}$ folgt hieraus durch Subtraktion von F_p und anschließender Division von $F_x + pF_y$

$$\dot{x} = -\frac{F_p}{F_x + pF_y}, \qquad \dot{y} = -\frac{pF_p}{F_x + pF_y},$$

vorausgesetzt, $F_x + pF_y \neq 0$. Aus diesen Gleichungen erhält man nun Lösungen $x(p), y(p)$, welche eine parametrisierte Form der Lösung darstellen. Unter Umständen lassen sie sich ferner derart auflösen, dass sich y in Abhängigkeit von x schreiben lässt.

Beispiel 1.3.2:

Ersetzt man in Gleichung (1.3.1) y' durch p, führt dies auf

$$0 = (p-1)^2 + y - x = 0$$

Hier ist $F_x = -1, F_y = 1$ sowie $F_p = 2(p-1)$. Somit ergibt Differentiation nach p:

$$0 = F_x\dot{x} + F_y\dot{y} + F_p = 2(p-1) + \dot{y} - \dot{(}x)$$

Mittels der Substitution $\dot{y} = p\dot{x}$ ergibt sich:

$$\dot{x} = -\frac{2(p-1)}{-1+p} = -2,\ \dot{y} = -2p$$

Die Lösungen dieser Differentialgleichungen lauten

$$x(p) = -2p + c_1,\ y(p) = -p^2 + c_2 \tag{1.3.2}$$

Aufgelöst nach p entspricht ersteres $p = c_1 - \frac{1}{2}x$, woraus sich

$$y(x) = -(c_1 - \frac{1}{2}x)^2 + c_2$$

ergibt. Wegen $(y'-1)^2 = ((c_1 - \frac{1}{2}x) - 1)^2 = (c_1 - \frac{1}{2}x)^2 - 2(c_1 - \frac{1}{2}x) + 1$ ergibt sich aus $(y'-1)^2 + y - x = 0$ die Bedingung $1 + c_2 - 2c_1 = 0$, also $c_2 = 2c_1 - 1$. Man erhält

$$y(x) = -(c_1 - \frac{1}{2}x)^2 + 2c_1 - 1$$

Manchmal ist es jedoch möglich und einfacher, y' zu isolieren, und die Differentialgleichung so in eine explizite Umzuwandeln:

Beispiel 1.3.3:

Gegeben sei

$$0 = F(x, y, y') = y'^2 - 4x^2$$

Hier ist $F_p = 2p$, die Gleichung also eindeutig auflösbar, falls $p \neq 0$. Dies lässt sich unmittelbar umstellen zu

$$y'(x) = p(x) = \pm\sqrt{4x^2} = \pm 2x$$

Die Lösungen der Differentialgleichung lauten folglich

$$y_1(x) = -x^2 + c_1$$
$$y_2(x) = x^2 + c_2$$

1.4 Totale Differentialgleichung

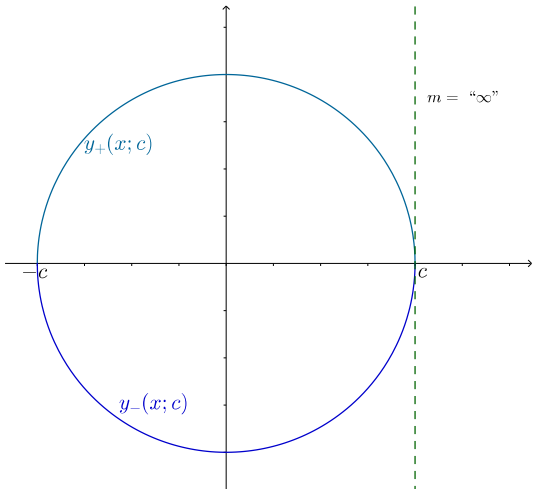
Zur Motivation des folgenden Kapitels soll folgendes Beispiel betrachtet werden:

$$y' + \frac{x}{y} = 0$$

Löst man diese Riccati-Gleichung auf dem im vorherigen Kapitel beschriebenen Wege, erhält man als allgemeine Lösung

$$y(x;c) = \pm\sqrt{c - x^2}$$

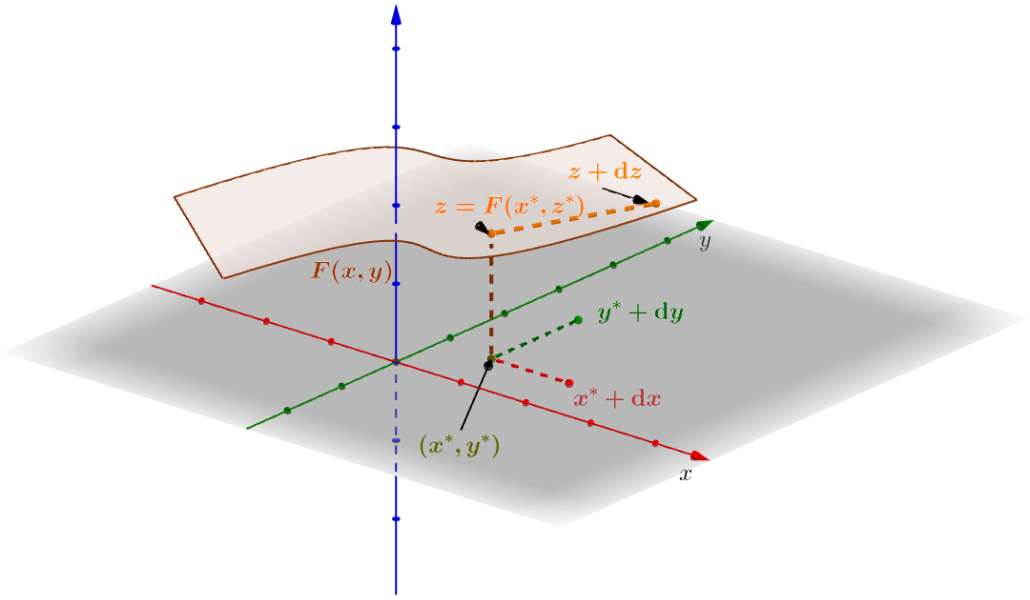
Das Problem hierbei: Diese Lösung existiert nur im Intervall $(-c, c)$, für $x = \pm\sqrt{c}$ ist y' nicht definiert, die Differentialgleichung folglich sinnfrei.



Um dieses Problem zu umgehen, kann man die Kurve in impliziter Form angeben, d. h. $F(x, y) \equiv C$, wobei man Darstellung der Differentialgleichung anpasst:

$$g(x, y)\mathrm{d}x + h(x, y)\mathrm{d}y = 0$$

Anders ausgedrückt, fasst man die Gleichung nicht mehr als DGL der Form $y' = f(x, y)$ auf, sondern als DGL der Form $\mathrm{d}z = \underbrace{\left(\frac{\partial F}{\partial x}\right)}_{g(x,y)}\mathrm{d}x + \underbrace{\left(\frac{\partial F}{\partial y}\right)}_{h(x,y)}\mathrm{d}y = 0$



Differentialgleichungen der Form

$$g(x, y)\mathrm{d}x + h(x, y)\mathrm{d}y = 0 \qquad (*)$$

nennt man ***exakt*** bzw. ***total***, sofern eine Funktion $F(x, y)$ mit $\frac{\partial F}{\partial x} = g(x, y)$ und $\frac{\partial F}{\partial y} = h(x, y)$ existiert.

Bemerkung: *Ist die DGL von der Form*

$$g(x)\mathrm{d}x + h(y)\mathrm{d}y$$

mit stetigen Funktionen $g : \mathbb{R} \to \mathbb{R}, h : \mathbb{R} \to \mathbb{R}$, so existiert für jedes kompakte Rechteck R eine Funktion F, sodass $F_x = g$ und $F_y = h$ in R gelten. Die allgemeine Lösung ist in diesem Falle durch $G(x)+H(y) = C$ gegeben, wobei $G(x)$ und $H(y)$ Stammfunktionen von g, h, C eine Konstante ist. ($F(x, y)$ wäre entsprechend durch $F(x, y) = G(x) + H(y)$ gegeben.)

Beispiel 1.4.1:

Gegeben sei die Differentialgleichung

$$\underbrace{2x}_{g(x)}\mathrm{d}x\underbrace{-9y^2}_{h(y)}\mathrm{d}y$$

Geeignete Stammfunktionen sind $G(x) = x^2$ und $H(y) = -3y^3$, somit lautet die allgemeine Lösung in impliziter Form $x^2 - 3y^3 = C$.

Beispiel 1.4.2:

Nun sei die Differentialgleichung

$$x\cos(y)\mathrm{d}x + \sqrt{x+1}\sin(y)\mathrm{d}y = 0$$

gegeben. Für $x > -1$ und $y \neq \pm\frac{k\pi}{2}, k \in \mathbb{N}_0$ lässt sich dies Umschreiben zu

$$\frac{x}{\sqrt{x+1}}\mathrm{d}x + \frac{\sin(y)}{\cos(y)}\mathrm{d}y = 0.$$

In imliziter Form lautet die allgemeine Lösung also

$$\frac{2}{3}(x-2)\sqrt{x+1} + \log(|\cos(y)|) = C$$

Satz 1.4.3:
Die Funktionen g, h seien im Gebiet $D \subset \mathbb{R}^2$ stetig und es gelte $g^2 + h^2 > 0$ in D. Sei $F \in C^1(D)$ eine geeignete ***Stammfunktion*** (d. h. $g = \frac{\partial F}{\partial x}, h = \frac{\partial F}{\partial y}$). Dann erhält man durch auflösen von $F(x, y) = c$ sämtliche Lösungen der exakten Differentialgleichung $(*)$.

Bemerkung: *Eine Differentialgleichung $y' = f(x, y)$ lässt sich in der Form $(*)$ schreiben, sofern $f(x, y) = -\frac{g(x,y)}{h(x,y)} = -\frac{F_x(x,y)}{F_y(x,y)}$ für geeignete g, h, F gilt. Die allgemeine Lösung lässt sich dann mittels $F(x, y) = c$ angeben. Sind Anfangswerte vorgegeben, so berechnet sich c durch $c = F(x_0, y_0)$.*

Beispiel 1.4.4:

Passt man im zur Motivation angeführten Beispiel die Darstellung von $y' + \frac{x}{y}$ an, so führt dies auf

$$x\mathrm{d}x + y\mathrm{d}y = 0$$

Eine passende Stammfunktion ist $F(x,y) = \frac{1}{2}\left(x^2+y^2\right)$. Die allgemeine Lösung in impliziter Form lautet folglich

$$\left(x^2+y^2\right) \equiv c.$$

(Der Faktor $\frac{1}{2}$ ist in der unbestimmten Konstanten c enthalten). Im Gegensatz zu $y(x) = \pm\sqrt{c-x^2}$ ergibt diese Darstellung auch für $x = \pm\sqrt{c}$ Sinn.

Beispiel 1.4.5:

Betrachtet werden soll die DGL

$$(2xy)\mathrm{d}x + \left(x^2-1\right)\mathrm{d}y = 0$$

In diesem Fall ist $g(x,y) = 2x$ und $h(x,y) = x^2-1$. Eine geeignete Stammfunktion ist folglich $F(x,y) = (x^2-1)\,y$. Somit ist

$$\underline{\underline{\left(x^2-1\right)y = c}}$$

die allgemeine Lösung der DGL $(2xy)\mathrm{d}x + (x^2-1) = 0$. Als Funktion $y(x)$ aufgefasst: $y(x) = \frac{c}{x^2-1}$ mit zugehöriger DGL $y' = -\left(\frac{2x}{x^2-1}\right)y$ (wegen „$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{2xy}{x^2-1}$“).

Satz 1.4.6:
Gilt zusätzlich zu den Voraussetzungen des letzten Satzes, dass D einfach zusammenhängend ist sowie g, h stetig differenzierbar sind, so existiert eine Stammfunktion F mit der Eigenschaft

$$F_x(x,y) = g(x,y), F_y(x,y) = h(x,y)$$

genau dann, wenn

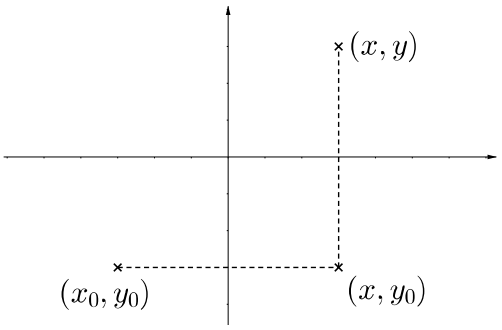
$$\frac{\partial g}{\partial y}(x,y) = \frac{\partial h}{\partial x}(x,y)$$

erfüllt ist. Man erhält eine solche über das Kurvenintegral

$$\iint_{(x_0,y_0)}^{(x,y)} g(\tau,y)\mathrm{d}\tau + h(x,\nu)\mathrm{d}\nu$$

Beispiel (Fortsetzung):

$$g(x,y) = 2xy \qquad h(x,y) = x^2-1$$
$$\frac{\partial g}{\partial y}(x,y) = 2x = \frac{\partial h}{\partial x}(x,y), (x,y) \in \mathbb{R}^2$$



Nutzt man die Wegunabhängigkeit des Kurvenintegrals aus und integriert entlang des links skizzierten Weges, so ergibt sich:

$$\begin{aligned} F(x,y) &= \iint_{(x_0,y_0)}^{(x,y)} (2\tau y)\mathrm{d}\tau + \left(x^2-1\right)\mathrm{d}\nu = \\ &= \int_{x_0}^{x} (2\tau y_0)\mathrm{d}\tau + \int_{y_0}^{y} (x^2-1)\mathrm{d}\nu = \\ &= \left(x^2-x_0^2\right)y_0 + \left(x^2-1\right)(y-y_0) = \\ &= \left(x^2-1\right)y - x_0^2 y_0 + y_0 \end{aligned}$$

Manchmal ist es auch möglich, eine DGL der Form $g(x,y)\mathrm{d}x + h(x,y)\mathrm{d}y = 0$ zu lösen, wenn diese nicht exakt ist (d. h. $g_y \neq h_x$), nämlich dann, wenn man eine Funktion $M(x,y)$ findet, sodass die DGL $(M(x,y)g(x,y))\,\mathrm{d}x + (M(x,y)h(x,y))\,\mathrm{d}y$ exakt ist. Ein solches M wird ***Eulerscher Multiplikator*** genannt.
Wegen $(Mg)_y = (Mh)_x$ ergibt sich die Bedingung

$$M_y g + M g_y = M_x h + M h_x \qquad (**)$$

Hierzu lassen sich folgende zwei Ansätze betrachten:

- *Man nimmt an, M hängt nur von y ab.*
  In diesem Falle vereinfacht sich die Bedingung $(**)$ auf

  $$M_y g + M g_y = M h_x$$

  Lässt sich die hieraus resultierende Gleichung

  $$\frac{\mathrm{d}}{\mathrm{d}y}(\log M) = \frac{M_y}{M} = \frac{h_x - g_y}{g}$$

  lösen, hat man einen geeigneten Multiplikator gefunden.

- *Man nimmt an, M hängt nur von x ab.*
  Analog zu obigem vereinfacht sich die Bedingung $(**)$ auf

  $$M g_y = M_x h + M h_x$$

  Auch hier hat man mit der Lösung der hieraus resultierenden Gleichung

  $$\frac{\mathrm{d}}{\mathrm{d}x}(\log M) = \frac{M_x}{M} = \frac{g_y - h_x}{h}$$

  einen geeigneten Multiplikator gefunden.

Beispiel 1.4.7:

Gegeben sei

$$\underbrace{y}_{=g(x,y)}\,\mathrm{d}x + \underbrace{2x}_{=h(x,y)}\,\mathrm{d}y = 0$$

Diese Gleichung ist wegen $\frac{\partial g}{\partial y} = 1 \neq 2 = \frac{\partial h}{\partial x}$ nicht exakt. Angenommen, es existiert ein Eulerscher Multiplikator $M = M(y)$. In diesem Falle führt dies auf die DGL

$$\frac{\mathrm{d}}{\mathrm{d}y}\log M = \frac{M_y}{M} = \frac{h_x - g_y}{g} = \frac{2-1}{y} = \frac{1}{y},$$

welche von $M(y) = y$ gelöst wird. Multiplikation der ursprünglichen DGL mit diesem Mul-

tiplikator ergibt:

$$\tilde{g}(x,y)\mathrm{d}x + \tilde{h}(x,y)\mathrm{d}y = 0$$

mit $\tilde{g}(x,y) = y^2$ und $\tilde{h}(x,y) = 2xy$. Diese Gleichung ist nun exakt, wie sich durch Nachrechnen leicht verifizieren lässt. Zur Bestimmung der Stammfunktion lässt sich der selbe „Integrationsweg“ wie im letzten Beispiel verwenden:

$$\begin{aligned}\tilde{F}(x,y) &= \iint_{(x_0,y_0)}^{(x,y)} y^2\mathrm{d}\tau + 2x\nu\mathrm{d}\nu = \int_{x_0}^{x} y_0^2\mathrm{d}\tau + \int_{y_0}^{y} 2x\nu\mathrm{d}\nu = \\ &= y_0^2\,(x - x_0) + x\left(y^2 - y_0^2\right) = xy_0^2 - x_0y_0^2 + xy^2 - xy_0^2 = \\ &= xy^2 + \underbrace{\tilde{c}}_{-x_0y_0^2}\end{aligned}$$

Die Lösungen der Gleichung erhält man nun durch auflösen von $F(x,y) = xy^2 = c$, wobei $F(x,y) = xy^2$ sei. (*Anm.*: Diese Funktion wird aufgrund ihrer Einfachheit verwendet, es ließe sich bspw. auch $F(x,y) = xy^2 + 10$ nutzen.)
Hieraus ergibt sich

$$\underline{\underline{y = \sqrt{\frac{c}{x}} = \frac{\bar{c}}{\sqrt{x}}}}$$

In Normalform sähe die DGL folgendermaßen aus: $y' = \left(-\frac{1}{2x}\right) y$

Dies sind zwei sehr einfache Ansätze, die jedoch nicht zwangsläufig erfolgreich sein müssen. Doch auch wenn sie nicht zum Erfolg führen, kann dennoch ein eulerscher Multiplikator existieren, welcher von x und y abhängt.

1.5 Lösung von Anfangswertproblemen

Unter einem ***Anfangswertproblem*** (AWP, auch *Cauchy-Problem* genannt) versteht man eine Differentialgleichung, bei der zusätzlich ein Anfangswert vorgegeben ist, den die gesuchte Lösung annehmen soll:

$$\begin{cases} y'(x) = f(x, y(x)) \\ y(x_0) = y_0 \end{cases} \qquad (*)$$

Für Differentialgleichungen höherer n. Ordnung (z .B. n.), werden darüberhinaus Anfangsbedingungen an die ersten $n-1$ Ableitungen gestellt:

$$\begin{cases} y^{(n)}(x) = F(x, y(x), y'(x), ..., y^{(n-1)}(x)) \\ \quad y(x_0) = y_0,\ y'(x_0) = y_1, ..., y^{(n-1)}(x_0) = y_{n-1} \end{cases}$$

Man löst ein solches Problem, in dem man zuerst die Differentialgleichung löst und anschließend in die allgemeine Lösung die Anfangsbedingung einsetzt, um einen konkreten Wert für die Konstante c zu erhalten. Hierbei stellen sich vor allem zwei Fragen:

1) Existiert eine Lösung?

2) Ist diese Lösung, sofern existent, eindeutig?

Um dies zu untersuchen, sei im Folgenden

- $R = \{(x,y) \in \mathbb{R}^2 : |x - x_0| \leq a, |y - y_0| \leq b\}$ ein kompaktes Rechteck
- $f : \mathbb{R}^2 \to \mathbb{R}$ eine auf R stetige Funktion

- $M \coloneqq \max\limits_{(x,y)\in R} |f(x,y)|$
- $\alpha \coloneqq \min\{a, \frac{b}{M}\}$
- $J \coloneqq [x_0 - \alpha, x_0 + \alpha]$

Satz 1.5.1 (Satz von Peano):
Sei $f : R \to \mathbb{R}$ stetig, dann existiert mindestens eine Lösung zum Anfangswertproblem $(*)$ in J.

Beispiel 1.5.2:

Gegeben sei das AWP

$$\begin{cases} y'' = 3y' - 2y \\ y(0) = 0, y'(0) = 1 \end{cases}$$

Eine (die) Lösung lautet $y(x) = -e^x + e^{2x}$.

(Wie die allgemeine Lösung einer Differentialgleichung n. Ordnung berechnet wird, wird im Folgenden Kapitel erklärt.)

Beispiel 1.5.3:

$$\begin{cases} y' = 2\sqrt{y} \\ y(0) = 0 \end{cases}$$

Diese DGL hat die Lösungen $y(x) = 0$ und $y(x) = x^2$.

Zum Beweis von diesem Existenzsatz (sowie anderen) wird folgendes Theorem benötigt:

Satz 1.5.4:
Sei $f : R \to \mathbb{R}$ stetig. Die Funktion $y : J \to \mathbb{R}$ ist genau dann eine Lösung des Anfangswertproblems, wenn sie folgende Integralgleichung erfüllt:

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$$

Beweis:
„⇒“: y sei Lösung des Anfangswertproblems. Dann ist $\int_{x_0}^{x} y'(t)\mathrm{d}t = \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$, also $y(x) - y_0 = \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$.
„⇐“: Sei $y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$. Da f stetig ist, folgt, dass $\int_{x_0}^{x} f(t, y(t))\mathrm{d}t$ stetig differenzierbar ist. Somit ist $y(x)$ ebenfalls stetig differenzierbar. Es gilt $y'(x) = 0 + f(x, y(x))$ und $y(x_0) = y_0 + 0$.

□

Motiviert durch diesen Satz sei folgender Operator definiert:

$$A : C(J) \to C(J), y(x) \mapsto (Ay)(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$$

Hierbei ist $C(J)$ die Menge der stetigen Funktionen auf J, welche versehen mit der Maximumsnorm $\|y\|_\infty = \max_{x\in J}|y(x)|$ zu einem Banachraum wird. Ferner sei folgende Teilmenge von $C(J)$ definiert:

$$K \coloneqq \{y \in C(J) : |y(x) - y_0| \leq b \,\forall x \in J\}$$

Diese ist abgeschlossen und konvex, wobei letzteres im Folgenden bewiesen werden soll.
Seien hierzu $y_1, y_2 \in K$, d. h.

$$(\circ)\, |y_1(x) - y_0| \leq b,\ |y_2(x) - y_0| \leq b \,\forall x \in J\ .$$

Ferner sei $\hat{\alpha} \in (0,1)$. Es gilt (für alle $x \in J$):

$$\begin{aligned}
&|\hat{\alpha} y_1(x) + (1-\hat{\alpha})\, y_2(x) - y_0| = \\
&= |\hat{\alpha} y_1(x) + (1-\hat{\alpha})\, y_2(x) - y_0 + \hat{\alpha} y_0 - \hat{\alpha} y_0| = \\
&= |\hat{\alpha} y_1(x) - \hat{\alpha} y_0 + (1-\hat{\alpha})\,(y_2(x) - y_0)| \overset{\Delta\text{-Ungl.}}{\leq} \\
&\leq \hat{\alpha}\,|y_1(x) - y_0| + (1-\hat{\alpha})\,|y_2(x) - y_0| \overset{(\circ)}{\leq} \\
&\leq \hat{\alpha} b + (1-\hat{\alpha})\, b = b
\end{aligned}$$

Somit folgt $\hat{\alpha} y_1 + (1-\hat{\alpha}) y_2 \in K \,\forall y_1, y_2 \in K, \hat{\alpha} \in (0,1)$, womit die Behauptung bewiesen ist.
Der Operator A hat nun folgende Eigenschaften:

- $A(K) \subseteq K$. *Beweis*: Sei $y \in K$. Dann gilt für beliebiges $x \in J$:

$$\begin{aligned}
|(Ay)(x) - y_0| &= \left| y_0 + \int_{x_0}^{x} f(t, y(t))\, dt - y_0 \right| \leq \int_{x_0}^{x} |f(t, y(t))|\, dt \leq \\
&\leq |x - x_0| \cdot \max_{x_0 \leq t \leq x} |f(t, y(t))| \leq \alpha \cdot M = \min\left\{a, \frac{b}{M}\right\} \cdot M \leq b
\end{aligned}$$

  Somit gilt $Ay \in K$.

- A ist stetig. *Beweis*: Nach Voraussetzung ist f stetig auf dem Rechteck R. Damit gilt:

$$\forall \varepsilon > 0 \exists \delta > 0 : |f(x, y_1) - f(x, y_2)| < \frac{\varepsilon}{2\alpha}$$

  für alle $|y_1 - y_2| < \delta$ $(y_1, y_2 \in [y_0 - b, y_0 + b])$. Für beliebige Funktionen $y, z \in K$ mit $\|y - z\|_\infty < \delta$ folgt somit

$$|f(x, y(x)) - f(x, z(x))| < \frac{\varepsilon}{2\alpha} \quad \forall x \in J$$

  Somit folgt für den Operator A mittels

$$|(Ay)(x) - (Az)(x)| \leq \underbrace{|x - x_0|}_{\leq \alpha} \cdot \max_{x_0 \leq t \leq x} |f(t, y(t)) - f(t, z(t))| \leq \frac{\varepsilon}{2} \forall x \in J$$

$$\|Ay - Az\|_\infty = \max_{x \in J} |(Ay)(x) - (Az)(x)| \leq \frac{\varepsilon}{2} < \varepsilon$$

  für alle $y, z \in K$ mit $\|y - z\|_\infty < \delta$.

- A ist beschränkt: Betrachtet man im Beweis zur ersten Eigenschaft anstatt $|(Ay)(x) - y_0|$ nur $|(Ay)(x)|$, ergibt sich mittels Dreiecksungleichung

$$|(Ay)(x)| \leq y_0 + \alpha M$$

- Ay ist sogar Lipschitz-stetig auf J für beliebiges $y \in K$: Für alle $x_1, x_2 \in J$ gilt:

$$|(Ay)(x_1) - (Ay)(x_2)| = \left| \int_{x_1}^{x_2} f(t, y(t)) \mathrm{d}t \right| \leq |x_1 - x_2| \cdot M$$

Aufgrund der letzten beiden Punkte lässt sich auf das Bild $A(K)$ der Satz von Arzelà-Ascoli anwenden:

Satz 1.5.5 (Arzelà-Ascoli):
Sei (X, d) kompakter, metrischer Raum, $M \subseteq C(X)$ (Raum der stetigen Funktionen von X nach $\mathbb{K}$), wobei $C(X)$ mit der Supremumsnorm versehen sei. Dann M ist relativ kompakt (d. h. $\bar{M}$ kompakt) genau dann, wenn Es die folgenden Bedingungen gelten:

a) M ist beschränkt.

b) M ist gleichgradig stetig: $\forall \varepsilon > 0$ existiert ein $\delta(\varepsilon) > 0$, sodass $|f(x) - f(y)| < \varepsilon$ für alle $x, y \in X$ mit $\mathrm{d}(x, y) < \delta$ und alle $f \in M$ existiert.

Nach diesem Satz ist $A(K)$ relativ kompakt. Da die dafür verwendeten Punkte für beliebige beschränkte Teilmengen von $C(J)$ gelten (und somit das entsprechende Resultat), ist A ein kompakter Operator. (Operator welcher, beschränkte Teilmengen von Banachräumen auf relativ kompakte Teilmengen von Banachräumen abbildet.) Der Satz von Peano folgt dann aus dem Fixpunktsatz von Schauder:

Satz 1.5.6 (Schauder'scher Fixpunktsatz):
Ist K eine nichtleere, abgeschlossene und konvexe Teilmenge eines Banachraums X und $A : K \to K$ ein stetiger, kompakter Operator, so hat A mindestens einen Fixpunkt in K.

Ist die Frage der Existenz positiv beantwortet, stellt sich die Frage, ob eine Lösung eindeutig ist. Wie an Beispiel 1.5.3 ersichtlich, reicht hierfür Stetigkeit als Voraussetzung nicht aus. Auch der Fixpunktsatz von Schauder trifft keine Aussage über die Eindeutigkeit. Aus diesem Grund soll folgender Fixpunktsatz betrachtet werden:

Satz 1.5.7 (Weissinger'scher Fixpunktsatz):
Sei $K \neq \emptyset$ eine abgeschlossene Teilmenge eines Banachraums mit Norm $\|\cdot\|$. Ferner sei $\sum_{n=1}^{\infty} \alpha_n$ eine konvergente Reihe (d. h. $\sum_{n=1}^{\infty} \alpha_n < \infty$) mit $\alpha_n > 0$, und es gelte

$$\|A^m y - A^m z\| \leq \alpha_m \|y - z\| \, \forall y, z \in K, m \in \mathbb{N}$$

für eine Abbildung $A : K \to K$. Dann hat A einen eindeutigen Fixpunkt $Ay = y$. Dieser Punkt ist Grenzwert der Folge $y_{n+1} = Ay_n$ mit beliebigen Anfangswert. Ferner gilt: $\|y - y_n\| \leq (\sum_{k=n}^{\infty} \alpha_k) \|y_1 - y_0\|$.

Beweis: Nach Voraussetzung gilt $\|y_{n+1} - y_n\| = \|A^n y_1 - A^n y_0\| \leq \alpha_n \|y_1 - y_0\|$. Somit ergibt sich für $k > 1$:

$$\begin{aligned}
\|y_{n+k} - y_n\| &= \|y_{n+k} - y_{n+k-1} + y_{n+k-1} \mp \ldots + y_{n+1} - y_n\| \leq \\
&\leq \|y_{n+k} - y_{n+k-1}\| + \|y_{n+k-1} - y_{n+k-2}\| + \ldots + \|y_{n+1} - y_n\| \leq \\
&\leq (\alpha_{n+k-1} + \alpha_{n+k-2} + \ldots + \alpha_n) \|y_1 - y_0\| = \\
&= \left(\sum_{l=n}^{n+k} \alpha_l \right) \cdot \|y_1 - y_0\| \, (\circ)
\end{aligned}$$

Für beliebiges $\varepsilon > 0$ existieren nun ein $N > 0$, sodass für alle $n \geq N, k \geq 1$ die Ungleichung

$$\sum_{l=n}^{n+k} \alpha_l < \frac{\varepsilon}{\max \{\|y_1 - y_0\|, 1\}}$$

erfüllt ist, da die Folge der Partialsummen einer konvergenten Reihe eine Cauchyfolge ist. Somit ist wegen $\|y_{n+k} - y_n\| < \varepsilon, \; n \geq N, k \geq 1$ die Folge y_n ebenfalls Cauchyfolge und somit

konvergent in K (abgeschlossene Teilmenge eines Banachraums). Es existiert also ein $y \in K$ mit $y_n \to y$. Es ergibt sich:

- $\|y_{n+1} - Ay\| = \|Ay_n - Ay\| \leq \alpha_1 \|y_n - y\| \overset{n\to\infty}{\to} 0$. Somit folgt aus $\lim\limits_{n\to\infty} y_{n+1} - y = 0$, $\lim\limits_{n\to\infty} y_{n+1} - Ay = 0$ aufgrund der Eindeutigkeit des Grenzwertes $Ay = y$, womit y Fixpunkt ist.
- y ist der einzige Fixpunkt: Sei $z \in K$ mit $Az = z$. Da infolge der Konvergenz der Reihe (α_n) Nullfolge ist, gilt

$$\|y - z\| = \|Ay - Az\| = \|A^n y - A^n z\| \leq \alpha_n \|y - z\| \overset{n\to\infty}{\to} 0.$$

  Somit folgt $y = z$.
- Lässt man in $(\circ)$ k gegen unendlich streben, so folgt die behauptete Fehlerabschätzung $\|y - y_n\| \leq (\sum_{k=n}^{\infty} \alpha_k) \|y_1 - y_0\|$.

□

Mit diesem Theorem lässt sich nun (unter stärkeren Einschränkungen) die Eindeutigkeit einer Lösung zeigen.

Satz 1.5.8 (Satz von Picard-Lindelöf):
Sei $f : \mathbb{R}^2 \to \mathbb{R}$ stetig in R. Darüber hinaus sei sie auf R stetig differenzierbar nach y oder erfülle bezüglich y eine Lipschitz-Bedingung

$$|f(x, y_1) - f(x, y_2)| \leq L |y_1 - y_2| \, \forall (x, y_1), (x, y_2) \in R$$

mit Lipschitz-Konstante L. Dann hat das Anfangswertproblem $(*)$ genau eine Lösung im Intervall J. Diese Lösung kann man durch die Iteration

$$y_n(x) \coloneqq (Ay_{n-1})(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t)) \, \mathrm{d}t$$

mit beliebigem Startwert $y_0(x)$ erreichen.

Beweis: Es gilt $|(A^n y)(x) - (A^n z)(x)| \leq \frac{|x-x_0|^n}{n!} L^n \|y - z\|_\infty \ \forall n \in \mathbb{N}, x \in J, y, z \in K$
Beweis (durch Induktion):

Induktionsanfang:

$$|(Ay)(x) - (Az)(x)| = \left| \int_{x_0}^{x} (f(t, y(t)) - f(t, z(t))) \, dt \right| \leq |x - x_0| \, L \|y - z\|_\infty$$

Induktionsvoraussetzung: Die Behauptung gelte für $n - 1$, d. h.

$$\left|\left(A^{n-1} y\right)(x) - \left(A^{n-1} z\right)(x)\right| \leq \frac{|x - x_0|^{n-1}}{(n-1)!} L^{n-1} \|y - z\|_\infty \, \forall x \in J, y, z \in K$$

Induktionsschluss: $n-1 \to n$ (Man beachte $A(K) \subseteq K$)

$$\begin{aligned}
|(A^n y)(x) - (A^n z)(x)| &= \left| \left(A \underbrace{\left(A^{n-1} y\right)}_{y_1 \in K} \right) x - \left(A \underbrace{\left(A^{n-1} z\right)}_{y_2 \in K} \right)(x) \right| = \\
&= |(Ay_1)(x) - (Ay_2)(x)| = \left| \int_{x_0}^{x} \left(f(t, y_1(t)) - f(t, y_2(t))\right) \mathrm{d}t \right| \leq \\
&\leq \int_{x_0}^{x} L\,|y_1(t) - y_2(t)|\,\mathrm{d}t = \int_{x_0}^{x} L\left|\left(A^{n-1} y\right)(t) - \left(A^{n-1} z\right)(t)\right| \mathrm{d}t \leq \\
&\overset{\text{IV}}{\leq} \int_{x_0}^{x} L \frac{(x-x_0)^{n-1}}{(n-1)!} L^{n-1} \|y-z\|_\infty \,\mathrm{d}t = \frac{|x-x_0|^n}{n!} L^n \|y-z\|_\infty
\end{aligned}$$

Somit folgt $|(A^n y)(x) - (A^n z)(x)| \leq \frac{|x-x_0|^n}{n!} L^n \|y-z\|_\infty$. Hieraus ergibt sich insbesondere: $\|A^n y - A^n z\|_\infty \leq \frac{\alpha^n}{n!} L^n$. Definiert man $\beta_n := \frac{\alpha^n}{n!} L^n$, so genügen A, K sowie die Reihe $\sum_{n=1}^{\infty} \beta_n$ den Voraussetzungen des Weissinger'schen Fixpunktsatzes. Daher existiert genau ein $y \in K$ mit $y = Ay$, also $y = y_0 + \int_{x_0}^{x} f(t, y(t))\mathrm{d}t$. Wie bereits gesehen, ist dies genau die Lösung des Anfangswertproblems

$$\begin{cases} y'(x) = f(x, y(x)) \\ y(x_0) = y_0 \end{cases}$$

□

Beispiel (Fortsetzung zu 1.5.3):

In Beispiel 1.5.3,

$$\begin{cases} y' = 2\sqrt{y} \\ y(0) = 0 \end{cases}$$

ist $f(x,y) = \sqrt{y}$ weder Lipschitz-stetig bezüglich der 2. Komponenten in $(0,0)$ noch stetig differenzierbar auf $[-a,a] \times [-b,b]$. Somit sind die Voraussetzungen des letzten Satzes nicht erfüllt, die Existenz zweier Lösungen stellt also keinen Widerspruch dar.

Beispiel 1.5.9:

$$\begin{cases} y' = y \\ y(0) = 1 \end{cases}$$

Hier ist $f(x,y) = y$.

$$\begin{aligned}
y_0(x) &= 1 \\
y_1(x) &= y_0 + \int_{x_0}^{x} y_0(t)\,\mathrm{d}t = 1 + \int_0^x 1 \mathrm{d}t = 1 + x \\
y_2(x) &= 1 + \int_0^x 1 + t \mathrm{d}t = 1 + x + \frac{1}{2}x^2 \\
y_3(x) &= 1 + \int_0^x 1 + t + \frac{1}{2}t^2 \mathrm{d}t = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3
\end{aligned}$$

Hieraus ergibt sich: $y(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$

Alternativer Startwert $y_0(x) = 2$:

$$y_1(x) = y_0 + \int_{x_0}^{x} y_0(t)\,\mathrm{d}t = 1 + 2x$$
$$y_2(x) = 1 + \int_0^x (1+2t)\,\mathrm{d}t = 1 + x + x^2$$
$$y_3(x) = 1 + \int_0^x \left(1 + t + t^2\right)\mathrm{d}t = 1 + x + \frac{x^2}{2} + \frac{x^3}{3}$$
$$\ldots \to y(x) = e^x$$

Satz 1.5.10 (Gronwall-Ungleichung):
Sei $I \subset \mathbb{R}$ ein Intervall, $x_0 \in I$ sowie $y, z : I \to \mathbb{R}$ stetige Funktionen, wobei z darüber hinaus nicht negativ sei. Ferner sei $c \geq 0$ so, dass

$$y(x) \leq c + \int_{x_0}^{x} z(t)\, y(t)\, dt$$

auf I gelte. Dann gilt:

$$y(x) \leq c \cdot \exp\left(\int_{x_0}^{x} z(t)\,\mathrm{d}t\right)$$

Anwendung dieses Satzes: Gilt $f(x, y(x)) \leq z(x)y(x)$ für alle $x \in I$, so sind sämtliche Lösungen $y(x)$ des AWPs

$$\begin{cases} y' = f(x,y) \\ y(x_0) = y_0 \end{cases}$$

nach oben durch $\max\{0, y_0\} \cdot \exp(\int_{x_0}^{x} z(t)\mathrm{d}t)$ beschränkt.

1.6 Differentialgleichungen im $\mathbb{R}^n$

Anstatt einzelner Differentialgleichungen der Form $y' = f(x, y)$ mit $y : I \to \mathbb{R}$ kann man auch Differentialgleichungssysteme der Form

$$\begin{cases} y_1'(x) = f_1(x, y_1(x), \ldots, y_n(x)) \\ y_2'(x) = f_2(x, y_1(x), \ldots, y_n(x)) \\ \qquad\qquad \ldots \\ y_m'(x) = f_m(x, y_1(x), \ldots, y_n(x)) \end{cases}$$

betrachten, bei welchen die einzelnen Lösungen y_k (u. U.) von den anderen abhängen. Fasst man die einzelnen Gleichungen zur Vereinfachung zu Vektoren zusammen, d. h.

$$\underline{y}(x) = \begin{pmatrix} y_1(x) \\ \ldots \\ y_n(x) \end{pmatrix},\ \underline{f}(x, \underline{y}) = \begin{pmatrix} f_1(x, y_1(x), \ldots, y_n(x)) \\ \ldots \\ f_n(x, y_1(x), \ldots, y_n(x)) \end{pmatrix},$$

so hat man erneut eine Gleichung der Form $y'(x) = f(x, y(x))$, jedoch diesmal mit Funktionen $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ und $y : \mathbb{R} \to \mathbb{R}^n$. Im Folgenden Kapitel soll zur Unterscheidung jedoch die unterstrichene Variante $\underline{f}, \underline{y}$ zur Abgrenzung vom Fall $n = 1$ beibehalten werden.

An ein solches System kann man ebenfalls Anfangsbedingungen stellen, sodass man wieder ein Anfangswertproblem erhält. Die Lösbarkeit wird unter gewissen Voraussetzungen durch folgenden Satz garantiert:

Satz 1.6.1:
Sei $\underline{f}(x, \underline{y})$ stetig auf einem Gebiet D und $\left(x_0, \underline{y}_0\right) \in D$, dann existiert (mindestens) eine (lokale) Lösung des des AWPs

$$\begin{cases} \underline{f}(x, \underline{y}) \\ \underline{y}(x_0) = \underline{y}_0 \end{cases}$$

Eine allgemeinere Variante des Satzes von Picard-Lindelöf liefert (ggf.) Eindeutigkeit:

Satz 1.6.2:
$\underline{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ sei stetig in einem Gebiet $G \subseteq \mathbb{R}^{n+1}$ mit $(x_0, \underline{y}_0) \in G$. Ferner genüge f dort einer Lipschitz-Bedingung der Form $\left\| \underline{f}(x, \underline{y}_1) - \underline{f}(x, \underline{y}_2) \right\| \leq L \cdot \left\| \underline{y}_1 - \underline{y}_2 \right\|$ mit beliebiger Norm $\|\cdot\|$. Dann hat obiges Anfangswertproblem genau eine Lösung.

Alternativ zu obigen Existenzsätzen kann auch der folgende herangezogen werden:

Satz 1.6.3 (Caratheodory):
Gegeben sei das Problem

$$\begin{cases} \underline{y}' = \underline{f}\left(x, \underline{y}\right) \\ \underline{y}\left(x_0\right) = \underline{y}_0 \end{cases},$$

wobei $f : [0, T] \times D \to \mathbb{R}^n$ ($D \subseteq \mathbb{R}^n$ offen) folgenden Bedingungen genügt:

- f ist messbar für festes y
- f ist stetig für jedes feste t
- Für jedes feste $\bar{y} \in D$ existiere $\beta : [0, T] \to \mathbb{R}^+$, sodass $\|f(t, \bar{y})\| \leq \beta(t)$ gilt.

Darüberhinaus existiere für jedes feste $\bar{y} \in D$ ein $\varrho \in \mathbb{R}^+$ sowie eine lokal integrierbare Funktion $\alpha : [0, T] \to \mathbb{R}^+$, sodass $B_\varrho(\bar{y}) \subseteq D$ sowie

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

für alle $t \in [0, T]$ und für alle $x, y \in B_\varrho(\bar{y})$ gilt. Dann existiert für jedes $(t_0, y_0) \in [0, T] \times D$ eine eindeutige Lösung.

Existieren Lösungen, lassen sich diese unter geeigneten Voraussetzungen wie folgt abschätzen:

Satz 1.6.4:
Es seien $A(x), \underline{f}(x)$ stetige Funktionen von einem Intervall $J \subset \mathbb{R}$ nach $\mathbb{R}^{n\times n}$ beziehungsweise $\mathbb{R}^n$. Sei I ein Teilintervall von J, $x_0 \in I$, $\underline{y}_0 \in \mathbb{R}^n, \gamma \in \mathbb{R}$ mit $\left\|\underline{y}_0\right\| \leq \gamma$, ferner seien $L, \delta \in \mathbb{R}$, sodass $\|A(x)\| \leq L, \left\|\underline{f}(x)\right\| \leq \delta$ in I gelten. Dann genügt die die Lösung $\underline{y}(x)$ des AWPs

$$\begin{cases} \underline{y}' = A\underline{y} \\ \underline{y}(x_0) = \underline{y}_0 \end{cases}$$

der Abschätzung

$$\left\|\underline{y}(x)\right\| \leq \gamma e^{L|x-x_0|} + \frac{\delta}{L}\left(e^{L|x-x_0|} - 1\right)$$

in I.

Bemerkung 1.6.5: *Die Normen können (fast) beliebig gewählt werden, wichtig ist, dass Matrixnorm und Vektornorm verträglich sind, dass heißt für alle $\tilde{A} \in \mathbb{R}^{n\times n}, \underline{\tilde{y}} \in \mathbb{R}^n$ die Ungleichung*

$$\left\|\tilde{A}\underline{\tilde{y}}\right\| \leq \left\|\tilde{A}\right\| \cdot \left\|\underline{\tilde{y}}\right\|$$

gilt. Für die Euklidische norm $\|\cdot\|_2$ könnte man beispielsweise die durch $\left\|\tilde{A}\right\|_F = \sqrt{\sum_{i,j=1}^n \tilde{a}_{i,j}^2}$ gegebene Frobeniusnorm verwenden,

Bemerkung 1.6.6: *(Existenzsatz im Komplexen)*
Im Komplexen lassen sich ähnliche Aussagen treffen: Ist die Funktion

$$\underline{f} : \mathbb{C} \times \mathbb{C}^n \to \mathbb{C}^n, (z, \underline{w}) \mapsto \underline{f}(z, \underline{w})$$

in einem Gebiet $D \subset \mathbb{C}^{n+1}$ holomorph (d. h. stetig komplex differenzierbar). Ferner sei $(x_0, \underline{w}_0) \in D$, so ist das AWP

$$\begin{cases} \underline{w}' = \underline{f}(z, \underline{w}) \\ \underline{w}(z_0) = \underline{w}_0 \end{cases}$$

in einem Ball $B_\alpha(z_0), \alpha > 0$ eindeutig lösbar.

1.6.1 Inhomogene, lineare Differentialgleichungssysteme mit konstanten Koeffizienten

Es bleibt die Frage zu klären, wie man eine Lösung einer Gleichung wie beispielsweise

$$\underline{y}'(x) = A(x)\,\underline{y}(x) + \underline{b}(x)$$

mit einer Matrix $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ erhält. Es liegt nahe, analog zu Differentialgleichungen in $\mathbb{R}$ den Ansatz

$$\underline{y}(x) = \underline{y}_{\text{hom.}}(x) + \underline{y}_p(x)$$

bestehend aus der allgemeinen Lösung des homogenen Teils und einer partikulären Lösung zu probieren.

Bemerkung 1.6.7: *Ist $A(x)$ stetig auf einem Intervall I, so existieren n linear unabhängige Lösungen der homogenen DGL $\underline{y}'(x) = A(x)\underline{y}(x), x \in I$, und jede Lösung kann als Linearkombination dieser „Basislösungen“ geschrieben werden.*

Der Nachteil hierbei ist: Für $\underline{y}'_{\text{hom.}}(x) = A(x)\underline{y}_{\text{hom.}}(x)$ ist im Allgemeinen keine Lösung bekannt. Hat man jedoch ein System von Differentialgleichungen mit konstanten Koeffizienten, d. h. $A(x) = A$ ist von x unabhängig, so kann man durchaus eine Lösung finden. Dazu kann der Ansatz

$$\underline{y}(x) = \underline{k}e^{\lambda x}$$

mit einem von x unabhängigen Vektor $\underline{k}$ und einer Konstanten λ betrachtet werden. Die Ableitung lautet

$$\underline{y}'(x) = \underline{k}\lambda e^{\lambda x}$$

Für eine Lösungskurve dieser Gestalt muss $\underline{k}\lambda e^{\lambda x} = A\underline{k}e^{\lambda x}$ beziehungsweise $(A\underline{k} - \lambda\underline{k})\, e^{\lambda x} = 0$ für alle x gelten. Somit hat man das Problem in ein Eigenwertproblem umgewandelt. Schreibt man λ_i für die Eigenwerte und $\underline{k}_i$ für die zugehörigen Eigenvektoren, so lautet, im Falle einer diagonalisierbaren Matrix mit paarweise verschiedenen Eigenwerten, die allgemeine Lösung

$$\underline{y}(x; c_1, \ldots, c_n) = c_1\underline{k}_1 e^{\lambda_1 x} + \cdots + c_n\underline{k}_n e^{\lambda_n x}.$$

Eine Matrix muss jedoch nicht zwangsläufig diagonalisierbar sein oder kann mehrfache Eigenwerte haben. Für die einzelnen Eigenwerte λ_i gibt es im Wesentlichen folgende Fälle zu unterscheiden:

1. $\lambda_i \in \mathbb{R}$ mit algebraischer Vielfachheit 1:
   In diesem Falle lautet die zugehörige Lösungskomponente, wie bereits verwendet, $\underline{y}_i = \underline{k}_i e^{\lambda_i x}$.

   Beispiel:
   $$\frac{\mathrm{d}}{\mathrm{d}x}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$
   Das charakteristische Polynom der Matrix A lautet
   $$P_A(\lambda) = \det(A - \lambda I) = (1-\lambda)(3-\lambda) - 15 = \lambda^2 - 4\lambda - 12$$
   Die Eigenwerte sind somit $\lambda_1 = -2$ und $\lambda_2 = 6$. Löst man $(A - \lambda_i)\underline{k}_i = 0$ so erhält man zum Beispiel $\underline{k}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ als Eigenvektor zum Eigenwert λ_1 sowie $\underline{k}_2 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$ als Eigenvektor zum Eigenwert $\lambda_2 = 6$. Somit lautet die allgemeine Lösung:
   $$\underline{y}\,(x; c_1, c_2) = c_1\begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-2x} + c_2\begin{pmatrix} 3 \\ 5 \end{pmatrix} e^{6x}$$

2. $\lambda_i \in \mathbb{C} \setminus \mathbb{R}$: In diesem Falle ist $\bar{\lambda_i} \in \mathbb{C}$ ebenfalls Eigenwert (A ist reell). Setzt man $\lambda_i = \mu + i\nu$ mit $\mu, \nu \in \mathbb{R}$, so führt obiger Ansatz auf komplexe Funktionen ($\lambda_1 = \lambda_i$, $\lambda_2 = \bar{\lambda_i}$):
   $$\underline{k}_1 e^{\lambda_1 x} = \underline{k}_1 e^{\mu x}(\cos(\nu x) + i\sin(\nu x))$$
   $$\underline{k}_2 e^{\lambda_2 x} = \underline{k}_2 e^{\mu x}(\cos(\nu x) - i\sin(\nu x))$$
   Die zugehörigen Eigenvektoren sind u. U. ebenfalls komplex. Schreibt man $\underline{k}_1 = \underline{a} + i\underline{b}$ (mit reellen Vektoren $\underline{a}, \underline{b}$), so lautet ein zweiter Eigenvektor $\underline{k}_2 = \underline{a} - i\underline{b}$. Man könnte bereits jetzt eine Lösung als Linearkombination der Lösungskomponenten
   $$\underline{z}_1(x) = (\underline{a}e^{\mu x}\cos(\nu x) - \underline{b}e^{\mu x}\sin(\nu x)) + i\,(\underline{b}e^{\mu x}\cos(\nu x) + \underline{a}e^{\mu x}\sin(\nu x))$$
   $$\underline{z}_2(x) = (\underline{a}e^{\mu x}\cos(\nu x) - \underline{b}e^{\mu x}\sin(\nu x)) - i\,(\underline{b}e^{\mu x}\cos(\nu x) + \underline{a}e^{\mu x}\sin(\nu x))$$

darstellen, jedoch ist diese noch nicht reell. Um dies zu erreichen, kann man obige Lösungskomponenten entsprechend linear kombinieren (wobei die Kombinationen linear unabhängig sein müssen, um den ganzen Lösungsraum aufzuspannen), zum Beispiel mittels:

$$\underline{y}_1(x) = \frac{1}{2}\left(\underline{z}_1(x) + \underline{z}_2(x)\right) = \underline{a}e^{\mu x}\cos(\nu x) - \underline{b}e^{\mu x}\sin(\nu x)$$
$$\underline{y}_2(x) = \frac{1}{2i}\left(\underline{z}_1(x) - \underline{z}_2(x)\right) = \underline{b}e^{\mu x}\cos(\nu x) + \underline{a}e^{\mu x}\sin(\nu x)$$

Die allgemeine, reelle Lösung lautet nun:

$$\underline{y}(x) = c_1\underline{y}_1(x) + c_2\underline{y}_2(x)$$

Beispiel:

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

Charakteristisches Polynom: $\det(A - \lambda I) = (1-\lambda)^2 + 1$
Die Eigenwerte sind somit (wegen $(1-\lambda)^2 \overset{!}{=} -1$): $\lambda_1 = 1 + i, \lambda_2 = 1 - i$ Löst man $(A - \lambda_1)\underline{k}_1 = 0$, erhält man beispielsweise $\underline{k}_1 = \begin{pmatrix} 1 \\ i \end{pmatrix}$ als Eigenvektor. Hier ist $\underline{a} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ sowie $\underline{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Somit ergibt sich

$$\underline{y}_1(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^x\cos(x) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^x\sin(x)$$
$$\underline{y}_2(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^x\cos(x) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^x\sin(x)$$

3. $\lambda_i \in \mathbb{R}$ mit algebraischer Vielfachheit $m > 1$, geometrischer Vielfachheit 1: In diesem Falle konstruiert man linear unabhängige Lösungen der Form

$$\underline{y}_1(x) = \underline{k}_1 e^{\lambda_i x},$$
$$\underline{y}_2(x) = \underline{k}_2 e^{\lambda_i x} + \underline{k}_3 x e^{\lambda_i x},$$
$$\dots$$
$$\underline{y}_m(x) = \underline{k}_l e^{\lambda_i x} + \underline{k}_{l+1} x e^{\lambda_i x} + \underline{k}_{l+2}\frac{x^2}{2}e^{\lambda_i x} + \dots + \underline{k}_{l+m-1}\frac{x^{m-1}}{(m-1)!}e^{\lambda_i x}$$

mit Lösungen $\underline{k}_j$ des Systems

$$1.\ (A - \lambda_i I)\,\underline{k}_1 = 0$$
$$2.\begin{cases} (A - \lambda_i I)\,\underline{k}_3 = 0 \\ (A - \lambda_i I)\,\underline{k}_2 = \underline{k}_3 \end{cases}$$
$$\dots$$
$$m.\begin{cases} (A - \lambda_i I)\underline{k}_{l+m-1} = 0 \\ (A - \lambda_i I)\underline{k}_{l+m-2} = \underline{k}_{l+m-1} \\ \dots \\ (A - \lambda_i I)\underline{k}_l = \underline{k}_{l+1} \end{cases}$$

(wobei $\underline{k}_{l+m-1} \neq 0$)

Beispiel:

$$y_1' = 3y_1 - 18y_2$$
$$y_2' = 2y_1 - 3y_2$$

$$A = \begin{bmatrix} 3 & -18 \\ 2 & -9 \end{bmatrix}$$

Charakteristisches Polynom: $\det(A - \lambda I) = (-3 - \lambda)(9 + \lambda) + 36$.
Wegen $(-3 - \lambda)(9 + \lambda) + 36 = -26 + 6\lambda + \lambda^2 + 36 = \lambda^2 + 6\lambda + 9 = (\lambda + 3)^2$ folgt $\lambda_1 = \lambda_2 = -3$. Eine Lösung von $(A - \lambda I)\underline{k}_1 = 0$ ist (z. B.) $\underline{k}_1 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$
$\Rightarrow \underline{y}_1(x) = \begin{pmatrix} 3 \\ 1 \end{pmatrix} e^{-3x}$ Sei nun $\underline{k}_3 = \underline{k}_1$. Lösen von $(A + 2I)\underline{k}_2 = \underline{k}_3$:

$$\begin{bmatrix} 6 & -18 \\ 2 & -6 \end{bmatrix} \begin{pmatrix} \underline{k}_2^{(1)} \\ \underline{k}_2^{(2)} \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

$$\Downarrow$$

$$6\underline{k}_2^{(1)} - 18\underline{k}_2^{(2)} = 3$$
$$2\underline{k}_2^{(1)} - 6\underline{k}_2^{(2)} = 1$$

$\Rightarrow$ Eine Lösung ist bspw. $\underline{k}_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$. $\Rightarrow \underline{y}_2(x) = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} e^{-3x} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} x e^{-3x}$ ist eine weitere Lösung.

4. $\lambda_i \in \mathbb{R}$ mit algebraischer Vielfachheit $m > 1$, geometrischer Vielfachheit $\nu = m$: In diesem Falle erhält man m linear unabhängige Lösungskomponenten durch $\underline{y}_{i+(l-1)} = \underline{k}_l e^{\lambda_i x}$, wobei die $\underline{k}_l$ für $l = 1, ..., m$ m linear unabhängige Eigenvektoren sind.

5. $\lambda_i \in \mathbb{R}$ mit algebraischer Vielfachheit $m > 1$, geometrischer Vielfachheit $1 < \nu < m$:
In diesem Falle erhält man ν linear unabhängige Lösungskomponenten durch $\underline{y}_{i+(l-1)} = \underline{k}_l e^{\lambda_i x}$, wobei die $\underline{k}_l$ ν linear unabhängige Eigenvektoren sind, ähnlich wie im vorherigen Schritt. Weitere Lösungen konstruiert man analog zu Fall 3, wobei es diesmal auch mehrere linear unabhängige Lösungskomponenten mit der Form $underline k_l e^{\lambda_i x} + \underline{k}_{l+1} x e^{\lambda_i x} + ... + + \underline{k}_{l+\alpha} \frac{x^\alpha}{\alpha!} e^{\lambda_i x}$ gibt, welche entsprechend in der allgemeinen Lösung berücksichtigt werden.

6. Mischformen: Für $n > 2$ sind auch Mischformen der obigen Fälle möglich, in diesem Falle kombiniert man die Lösungsmethoden entsprechend.

Begründung zu 3.-5.: Hat ein Eigenwert $\lambda \in \mathbb{R}$ algebraische Vielfachheit $m > 1$, so gibt es m linear unabhängige Hauptvektoren. Für einen Hauptvektor $\underline{k}$ der i. Stufe lässt sich eine Lösungskomponente $\underline{y}_i$ als

$$\underline{y}_i = e^{\lambda x} \cdot \left(\sum_{j=0}^{i-1} \frac{(A - \lambda I)^j x^j}{j!} \right) \cdot \underline{k}_{l(i)}$$

schreiben. Dass dies an die Exponentialreihe $\exp(z) = \sum_{j=0}^{\infty} \frac{z^j}{j!}$ erinnert, hat einen Grund. Betrachtet man die sogenannte Exponentialmatrix $e^{Ax} := \sum_{j=0}^{\infty} \frac{(Ax)^j}{j!}$ für beliebige $A \in \mathbb{R}^{n \times n}, x \in \mathbb{R}$, so erfüllt diese Matrix die Gleichung $\frac{\mathrm{d}}{\mathrm{d}x} e^{Ax} = A \cdot e^{Ax}$. Ferner gilt für $B \in \mathbb{R}^{n \times n}$ $e^{(A+B)x} = e^{Ax} e^{Bx}$, sofern A und B kommutieren. Für beliebige Vektoren $v \in \mathbb{R}^n$ gilt nun $\frac{\mathrm{d}}{\mathrm{d}x} e^{Ax} \underline{v} = A \cdot e^{Ax} \underline{v}$ (d. h. $\underline{y}(x) = e^{Ax} \underline{v}$ löst die DGL.) sowie $e^{Ax} \underline{v} = e^{\lambda I} \cdot e^{(A - \lambda I)x} \underline{v}$. Speziell für einen Hauptvektor i. Stufe, hier $\underline{k}$, ergibt sich somit wegen $(A - \lambda I)^j \underline{k} = 0, j \geq i$ sowie $e^{\lambda I x} = e^{\lambda x} I$:

$$e^{Ax} \underline{k} = e^{\lambda I x} \cdot e^{(A - \lambda I)x} \underline{k} = e^{\lambda x} \sum_{j=0}^{i-1} \frac{(A - \lambda I)^j x^j}{j!} \underline{k}.$$

Aus einem Eigenvektor mit algebraischer Vielfachheit m lassen sich somit m linear unabhängige Lösungskomponenten konstruieren.

Beispiel (zu 3.):

$$\begin{cases} y_1' = y_1 - y_2 \\ y_2' = 4y_1 - 3y_2 \end{cases}$$

In diesem Falle ist $A = \begin{bmatrix} 1 & -1 \\ 4 & -3 \end{bmatrix}$. Mit $\det(A - \lambda I) = \lambda^2 + 2\lambda + 1 = 0$ folgt, dass der einzige Eigenwert $\lambda = -1$ ist. Als Lösung von $(A - \lambda I)\underline{k}_1$ ergibt sich als ein Eigenvektor: $\underline{k}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. Lösen von $(A - \lambda I)\underline{k}_2 = \underline{k}_1$ ergibt $\underline{k}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Die allgemeine Lösung lautet somit:

$$\underline{y}(x; c) = c_1 e^{-x} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + c_2 e^{-x} \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} + x \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right)$$

Beispiel (zu 4.):

$$\begin{aligned} y_1' &= y_1 - 2y_2 \\ y_2' &= 2y_1 - y_3 \\ y_3' &= 4y_1 - 2y_2 - y_3 \end{aligned}$$

In diesem Falle ist

$$A = \begin{bmatrix} 1 & -2 & 0 \\ 2 & 0 & -1 \\ 4 & -2 & -1 \end{bmatrix}$$

Das charakteristische Polynom lautetet

$$\chi_A(\lambda) = \det(A - \lambda I) = (1 - \lambda) \cdot (\lambda^2 + \lambda + 2)$$

Die Eigenwerte sind folglich: $\lambda_{1,2} = -\frac{1}{2} \pm i\frac{\sqrt{7}}{2}$, $\lambda_3 = 1$ Lösen von

$$(A - \lambda_i I)\underline{k}_i = 0$$

liefert als Eigenvektoren beispielsweise:

$$\underline{k}_1 = \begin{pmatrix} \frac{3}{2} + i\frac{\sqrt{7}}{2} \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 2 \\ 4 \end{pmatrix} + i \begin{pmatrix} \frac{\sqrt{7}}{2} \\ 0 \\ 0 \end{pmatrix}, \quad \underline{k}_2 = \begin{pmatrix} \frac{3}{2} \\ 2 \\ 4 \end{pmatrix} - i \begin{pmatrix} \frac{\sqrt{7}}{2} \\ 0 \\ 0 \end{pmatrix},$$

$$\underline{k}_3 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

Somit ergeben sich als Lösungskomponenten der allgemeinen Lösung $\underline{y}(x; \underline{c}) = c_1 \underline{y}_1(x) +$

$c_2\underline{y}_2(x) + c_3\underline{y}_3(x)$:

$$\underline{y}_1(x) = e^{-\frac{1}{2}x}\begin{pmatrix}\frac{3}{2}\\2\\4\end{pmatrix}\cos\left(\tfrac{\sqrt{7}}{2}x\right) - e^{-\frac{1}{2}x}\begin{pmatrix}\frac{3}{2}\\2\\4\end{pmatrix}\sin\left(\tfrac{\sqrt{7}}{2}x\right)$$

$$\underline{y}_2(x) = e^{-\frac{1}{2}x}\begin{pmatrix}\frac{3}{2}\\2\\4\end{pmatrix}\sin\left(\tfrac{\sqrt{7}}{2}x\right) + e^{-\frac{1}{2}x}\begin{pmatrix}\frac{3}{2}\\2\\4\end{pmatrix}\cos\left(\tfrac{\sqrt{7}}{2}x\right)$$

$$\underline{y}_3(x) = e^{x}\begin{pmatrix}1\\0\\2\end{pmatrix}$$

Hat man nun auf beschriebenem oder anderem Wege ein System von n linear unabhängigen Lösungen, ein sogenanntes ***Hauptsystem*** (oder ***Fundamentalsystem***), gefunden, kann man dieses auch als Lösungsmatrix schreiben:

$$Y(x) = \begin{pmatrix}\underline{y}_1^{(1)}(x) & \underline{y}_2^{(1)}(x) & \cdots & \underline{y}_n^{(1)}(x)\\ \vdots & \vdots & \ddots & \vdots\\ \underline{y}_1^{(n)}(x) & \underline{y}_2^{(n)}(x) & \cdots & \underline{y}_n^{(n)}(x)\end{pmatrix}$$

Diese löst die Gleichung

$$Y'(x) = AY(x)$$

Die allgemeine Lösung der Differentialgleichung lässt sich hiermit auch in der Form $\underline{y}(x;\underline{c}) = Y(x)\cdot\underline{c}$ mit $\underline{c}\in\mathbb{R}^n$ ausdrücken.

Beispiel:

$$y_1' = y_1 + 3y_2$$
$$y_2' = 5y_1 + 3y_2$$

Hier ist $A = \begin{bmatrix}1 & 3\\5 & 3\end{bmatrix}$. Wegen $\chi_A(\lambda) = \lambda^2 - 4\lambda - 12$ sind die Eigenwerte $\lambda_1 = -2, \lambda_2 = 6$. Zugehörige Eigenvektoren sind $\underline{k}_1 = \begin{pmatrix}1\\-1\end{pmatrix}$, $\underline{k}_2 = \begin{pmatrix}3\\5\end{pmatrix}$. Somit ist mit $\underline{y}_1(x) = e^{-2x}\begin{pmatrix}1\\-1\end{pmatrix}$ und $\underline{y}_2(x) = e^{6x}\begin{pmatrix}3\\5\end{pmatrix}$ ein Fundamentalsystem gegeben. (Die hierzu nötige lineare Unabhängigkeit folgt durch Setzen von $x = 0$ direkt aus der linearen Unabhängigkeit von Eigenvektoren zu paarweise verschiedenen Eigenwerten.) Die allgemeine Lösung lautet somit:

$$\underline{y}(x;c) = c_1\underline{y}_1(x) + c_2\underline{y}_2(x) = \underbrace{\begin{pmatrix}e^{-2x} & 3e^{6x}\\-e^{-2x} & 5e^{6x}\end{pmatrix}}_{=Y(x)}\cdot\underbrace{\begin{pmatrix}c_1\\c_2\end{pmatrix}}_{=\underline{c}}$$

Definition 1.6.8:
Die Determinante der Matrix $Y(x) = \left(\underline{y}_1 \,|\, \underline{y}_2 \,|\, \ldots \,|\, \underline{y}_n\right)$ mit Lösungen $\underline{y}_1, \ldots, \underline{y}_n : \mathbb{R} \to \mathbb{R}^n$ des Systems $\underline{y}' = A(x)\underline{y}$,

$$\phi(x) = \det(Y(x)),$$

heißt ***Wronski-Determinante***.

Die Wronski-Determinante erfüllt die Gleichung

$$\phi'(x) = (\text{tr } A(x))\,\phi(x)$$

sowie

$$\phi(x) = \phi(x_0) \cdot \exp\left(\int_{x_0}^{x} (\text{tr } A(t))\,\mathrm{d}t\right)$$

Insbesondere gilt: Ist $\phi(x_0) \neq 0$ an einer beliebigen Stelle x_0, so ist $\phi(x) \neq 0$ für alle x.

Beispiel (Fortsetzung):

Für obiges Beispiel ergibt sich:

$$\phi(x) = \det Y(x) = \det \begin{pmatrix} e^{-2x} & 3e^{6x} \\ -e^{-2x} & 5e^{6x} \end{pmatrix} = 8e^{4x}$$

Setzt man $x_0 = 0$, stimmt dies wegen $\phi(0) = \det \begin{pmatrix} 1 & 3 \\ -1 & 5 \end{pmatrix} = 8$ und tr $A = \text{tr } \begin{pmatrix} 1 & 3 \\ 5 & 3 \end{pmatrix} = 4$ mit

$$\phi(x) = \phi(x_0) e^{\int_{x_0}^{x} \text{tr } A(t)\mathrm{d}t},\ \phi'(x) = (\text{tr } A(x))\,\phi(x)$$

überein.

Mit Hilfe der letzten beiden Definitionen kann nun auch eine partikuläre Lösung des inhomogenen Teils der DGL

$$\underline{y}_h'(x) = A\underline{y}(x) + \underline{b}(x)$$

gefunden werden, erneut über den bereits bekannten Ansatz der Variation der Konstanten. Hierzu setzt man

$$\underline{y}_p(x) = Y(x) \cdot \underline{c}(x)$$

mit einer noch zu bestimmenden vektorwertigen Funktion $c : \mathbb{R} \to \mathbb{R}^n$ sowie einer Fundamentalmatrix $Y(x)$. Ableiten ergibt:

$$Y'(x) \cdot \underline{c}(x) + Y(x) \cdot \underline{c}'(x) \stackrel{!}{=} AY(x) \cdot \underline{c}(x) + b(x)$$

Hieraus folgt:

$$\begin{gathered} (Y'(x) - AY(x))\,\underline{c}(x) + Y(x)\,\underline{c}'(x) = \underline{b}(x) \\ \Downarrow \\ Y(x)\,\underline{c}'(x) = \underline{b}(x) \\ \Downarrow \\ \underline{c}'(x) = Y(x)^{-1} \cdot \underline{b}(x) \end{gathered}$$

(Da die Spalten linear unabhängig sind, existiert die Inverse.) Nun lässt sich $\underline{c}(x)$ mittels Integration bestimmen:

$$\underline{c}(x) = \int_{x_0}^{x} (Y(t))^{-1}\,\underline{b}(t)\,\mathrm{d}t$$

Somit ergibt sich mit der partikulären Lösung

$$\underline{y}_p(x) = Y(x) \cdot \underline{c}(x) = Y(x) \cdot \int_{x_0}^{x} (Y(t))^{-1}\,\underline{b}(t)\,\mathrm{d}t$$

als allgemeine Lösung der inhomogenen Gleichung:

$$\underline{y}\,(x;\underline{c}) = Y\,(x)\,\underline{c} + Y\,(x)\int_{x_0}^{x} (Y(t))^{-1}\,\underline{b}\,(t)\,\mathrm{d}t$$

$\underline{c} \in \mathbb{R}^n$ ist hierbei ein Vektor, der die konstanten Vorfaktoren $c_1, c_2, \ldots$ der Lösungskomponenten des homogenen Teils als Einträge besitzt. Für ein Anfangswertproblem

$$\begin{cases} \underline{y}' = A\underline{y} + \underline{b} \\ \underline{y}(x_0) = \underline{y}_0 \end{cases}$$

lautet die Lösung:

$$\underline{y}\,(x) = Y\,(x)\left[(Y(x_0))^{-1}\cdot\underline{y}_0 + \int_{x_0}^{x} (Y(t))^{-1}\,\underline{b}\,(t)\,\mathrm{d}t\right]$$

1.6.2 Die Eliminationsmethode

Für kleine Systeme linearer Differentialgleichungen mit konstanten Koeffizienten lässt sich als Alternative auch die sogenannte ***Eliminationsmethode*** anwenden. Bei dieser wird der Differentialoperator $\mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}x} = \frac{\partial}{\partial x}$ als Koeffizient betrachtet, mit welchem man beliebig Addieren und Multiplizieren kann, mit den für die Differentiation geltenden Regeln: Multipliziert man bspw. D mit einer Konstanten, so ergibt dies 0, multipliziert man D mit einer Funktion der Form $e^{\lambda x}$, erhält man $\lambda e^{\lambda x}$ (usw.). Durch geschickte Kombination dieser Operationen angewandt auf das System von Differentialgleichungen versucht man, ***gekoppelte*** Differentialgleichungen zu entkoppeln. Gekoppelt bedeutet in diesem Zusammenhang, dass in einem System von Differentialgleichungen für $y_1', ..., y_n'$ mindestens ein y_i' in irgendeiner Form von mindestens einem y_j mind $j \neq i$ abhängt. Entkoppeln bedeutet entsprechend, das System so umzuschreiben, dass jede dieser Gleichungen für sich genommen eine gewöhnliche Differentialgleichung ist, welche von keinem anderen y_j abhängt. Das Eliminationsverfahren soll an folgenden Beispielen veranschaulicht werden:

Beispiel 1.6.9:

Das homogene System

$$\begin{cases} y' - 2z = \dfrac{\mathrm{d}}{\mathrm{d}x}y - 2z = 0 \\ z' - 3y = \dfrac{\mathrm{d}}{\mathrm{d}x}z - 3y = 0 \end{cases}$$

wird durch die Substitution $\mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}x}$ zu

$$\begin{cases} \mathrm{D}y - 2z = 0 & \text{(I)} \\ \mathrm{D}z - 3y = 0 & \text{(II)} \end{cases}$$

Multipliziert man (I) mit D, (II) mit 2, erhält man das System

$$\begin{cases} \mathrm{D}^2y - 2\mathrm{D}z = 0 & \text{(I)} \\ 2\mathrm{D}z - 6y = 0 & \text{(II)} \end{cases}$$

Mit Hilfe der Addition (I) $\oplus$ (II) ergibt sich nun $\mathrm{D}^2y - 6y = 0$. Die allgemeine Lösung der Differentialgleichung $y'' - 6y = 0$ ist $y(x; c_1, c_2) = c_1e^{\sqrt{6}x} + c_2e^{-\sqrt{6}x}$. Auf analoge Weise führt die Addition $3 \times$ (I) $\oplus$ D $\times$ (II) ((I),(II) im ursprünglichen Zustand) auf die Differentialgleichung $z'' - 6z = 0$, welche $z(x; c_3, c_4) = c_3e^{\sqrt{6}x} + c_4e^{-\sqrt{6}x}$ als allgemeine Lösung hat. Zur

Elimination der überschüssigen Variablen setzt man diese Lösungen in das ursprüngliche System ein:

$$\begin{cases} 0 = y'(x;c_1,c_2) - 2z(x;c3,c4) = \left(\sqrt{6}c_1 - 2c_3\right)e^{\sqrt{6}x} + \left(-\sqrt{6}c_2 - 2c_4\right)e^{-\sqrt{6}x} \\ 0 = z'(x;c_3,c_4) - 3y(x;c_1,c_2)\left(-3c_1 + \sqrt{6}c_3\right)e^{\sqrt{6}x} + \left(-3c_2 - \sqrt{6}c_4\right)e^{-\sqrt{6}x} \end{cases}$$

Aus $\sqrt{6}c_1 - 2c_3 = 0$ folgt $c_3 = \frac{\sqrt{6}}{2}$, aus $-\sqrt{6}c_2 - 2c_4 = 0$ folgt $c_4 = -\frac{\sqrt{6}}{2}$. Somit lautet die allgemeine Lösung des Systems

$$\begin{pmatrix} y(x;c_1,c_2) \\ z(x;c_1,c_2) \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ \frac{\sqrt{6}}{2} \end{pmatrix} e^{\sqrt{6}x} + c_2 \begin{pmatrix} 1 \\ -\frac{\sqrt{6}}{2} \end{pmatrix} e^{-\sqrt{6}x}$$

Beispiel 1.6.10:

Gegeben sei nun das inhomogene System

$$\begin{cases} y' - y - z = 4e^x \\ z' - 3z + y = -1 \end{cases}$$

Mit $\mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}x}$ wird aus

$$\begin{cases} \mathrm{D}y - y - z = 4e^x & | \times -1 \quad \text{(I)} \\ \mathrm{D}z - 3z + y = -1 & | \times (\mathrm{D}-1) \quad \text{(II)} \end{cases}$$

$$\begin{cases} -(\mathrm{D}-1)y + z = -4e^x & \text{(I)} \\ (\mathrm{D}-1)(\mathrm{D}-3)z + (\mathrm{D}-1)y = 1 & \text{(II)} \end{cases}$$

Hieraus ergibt sich

$$(\mathrm{I}) \oplus (\mathrm{II}) : (\mathrm{D}^2 - 4\mathrm{D} + 3)z + z = 1 - 4e^x$$

Da eine partikuläre Lösung der Gleichung $z'' - 4z' + 4z = 1 - 4e^x$ durch $\frac{1}{4} - 4e^x$ gegeben ist, ist die allgemeine Lösung für z

$$z(x;c_3,c_4) = c_3e^{2x} + c_4xe^{2x} + \frac{1}{4} - 4e^x$$

Analog führt die Elimination von y mittels $(\mathrm{D}-3)(\mathrm{I}) \oplus (\mathrm{II})$ auf die Differentialgleichung $y'' - 4y' + 4y = -1 - 8e^x$, deren allgemeine Lösung

$$y(x;c_1,c_2) = c_1e^{2x} + c_2xe^{2x} - \frac{1}{4} - 8e^x$$

ist. Rücksubstitution in das System ergibt

$$\begin{cases} (c_1 + c_2 - c_3)\,e^{2x} + (c_2 - c_4)\,xe^{2x} = 0 \\ (c_1 + c_2 - c_3)\,e^{2x} + (c_2 - c_4)\,xe^{2x} = 0 \end{cases},$$

was zuerst auf $c_2 = c_4$ und anschließend auf $c_3 = c_1 + c_2$ führt. Somit lautet die allgemeine Lösung des Systems

$$\begin{pmatrix} y(x;c_1,c_2) \\ z(x;c_1,c_2) \end{pmatrix} = \begin{pmatrix} c_1 \\ c_1 + c_2 \end{pmatrix} e^{2x} + \begin{pmatrix} c_2 \\ c_2 \end{pmatrix} xe^{2x} - \begin{pmatrix} \frac{1}{4} + 8e^x \\ -\frac{1}{4} + 4e^x \end{pmatrix}$$

Im letzten Beispiel ist die partikuläre Lösung vom Himmel gefallen. Bei linearen Differenti-

algleichungen mit konstanten Koeffizienten lässt sich hierbei jedoch ein systematischer Ansatz anwenden. Ersetzt man wie gehabt $\frac{\mathrm{d}}{\mathrm{d}x}$ durch D, so lässt sich eine Differentialgleichung der Form $y^{(n)} + \sum_{=0}^{n-1} a_{n-1} y^{(k)} = f(x)$ auch als $P(\mathrm{D})y = g$ schreiben, wobei P ein Polynom vom Grad n ist (nämlich $P(x) = x^n + \sum_{k=0}^{n-1} a_{n-1} x^k$). Gibt es ein Polynom G (ausgenommen $G \equiv 0$), sodass der Operator $G(\mathrm{D})$ $f(x)$ eliminiert, lässt sich eine partikuläre Lösung wie folgt finden:

1. Bestimmung der allgemeinen Lösung $y_p(x; c_1, ...)$ der durch $G(\mathrm{D})z = 0$ beschriebene homogene Differentialgleichung
2. Einsetzen der allgemeinen Lösung $y_p(x; c_1, ...)$ in die Gleichung $P(\mathrm{D})y_p = f(x)$, um die Koeffizienten zu bestimmen.

Der zweite Schritt ist stets durchführbar: Für ein lineares System von m Differentialgleichungen ($m = \text{grad } G$) kann eine Lösung stets als Linearkombination m linear unabhängiger (existenter!) Lösungen geschrieben werden und eine Differentialgleichung m. Ordnung kann stets in ein solches transformiert werden (vgl. 1.6.7, 1.7).

Beispiel 1.6.11:

$$y'' + 3y' + 2y = 4x^2 \qquad (*)$$

wird zu

$$(\mathrm{D} + 2)(\mathrm{D} + 1)y = 4x^2$$

Ein geeignetes G ist durch $G(\mathrm{D}) = \mathrm{D}^3$ gegeben: $\mathrm{D}^3 4x^2 = 0$. Die allgemeine Lösung von $y''' = 0$ ist $c_1x^2 + c_2x + c_3$, womit man den Ansatz $y_p(x; c_1, c_2, c_3)$ erhält. Eingesetzt in $(*)$ ergibt sich:

$$4x^2 = y_p'' + 3y_p' + 2y_p = 2c_1x^2 + (6c_1 + 2c_2)\,x + (2c_1 + 3c_2 + 2c_3)$$

Durch Vorwärtssubstitution erhält man für das LGS

$$\begin{bmatrix} 2 & & \\ 6 & 2 & \\ 2 & 3 & 2 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$$

die Lösung $c_1 = 2, c_2 = -6$, $c_3 = 7$. Addiert man allgemeinen Lösung $y(x; \tilde{c}_1, \tilde{c}_2) = \tilde{c}_1 e^{-2x} + \tilde{c}_2 e^{-x}$ der DGL $y'' + 3y' + 2y = 0$, so ergibt sich als allgemeine Lösung von $(*)$

$$y(x; \tilde{c}_1, \tilde{c}_2) = \tilde{c}_1 e^{-2x} + \tilde{c}_2 e^{-x} + 2x^2 - 6x + 7$$

1.6.3 Anwendung: Kompartimentmodelle

An dieser Stelle soll - als Ausblick auf Kapitel 2 - exemplarisch die praktische Anwendung von Differentialgleichungen vorgestellt werden: Die sogenannten ***Kompartimentmodellen***.
Diese Modelle sind aus Behältern oder ***Kompartiments*** $K_1, K_2, ..., K_n$ aufgebaut, welche beispielsweise räumlich getrennt sind und/oder sich in ihrer Funktion unterscheiden. Diesen werden zu jedem Zeitpunkt t $y_i(t)$ Masseneinheiten zugewiesen (z. B. Stoffmenge, Konzentration einer Lösung). Diese Behälter tauschen ihren Inhalt untereinander aus. Hierfür bezeichne $k_{i,j} > 0, i \neq j$ die Übertragungsrate von K_i nach K_j. Die Änderung von y_i in einem Zeitintervall $\mathrm{d}t$ setzt sich zusammen aus Zufluss und Abfluss. Erster ergibt sich durch $\sum_{\substack{j=1 \\ j \neq i}}^{n} k_{j,i} y_j \cdot \mathrm{d}t$, letzterer

durch $\sum_{\substack{j=1\\j\neq i}}^{n} k_{i,j}y_i \cdot \mathrm{d}t$. Der Einfachheit halber sei der Abfluss aus y_i in $k_{i,i} := \sum_{\substack{j=1\\j\neq i}}^{n} k_{i,j}$ zusammengefasst. Division durch $\mathrm{d}t$ und Grenzwertübergang führen letztendlich auf das homogene Differentialgleichungssystem

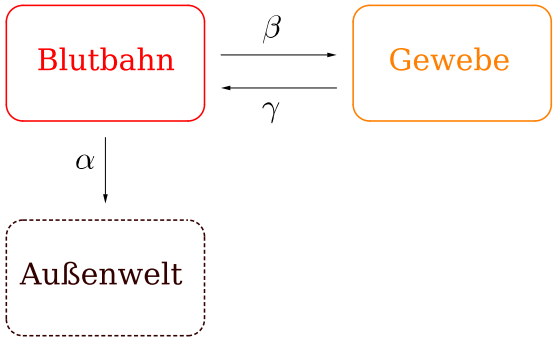
$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{pmatrix} = \begin{bmatrix} -k_{1,1} & k_{2,1} & \dots & k_{n,1} \\ k_{2,1} & -k_{2,2} & \ddots & k_{n,2} \\ \vdots & \ddots & \ddots & \vdots \\ k_{n,1} & \dots & \dots & -k_{n,n} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Unter Umständen besteht noch die Möglichkeit, dass den Behältern von außen Masse zugeführt wird, was durch einen „externen“ Zufluss der Rate $\delta_i(t)$ beschrieben werden kann. Hierdurch wird obiges System inhomogen:

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{pmatrix} = \begin{bmatrix} -k_{1,1} & k_{2,1} & \dots & k_{n,1} \\ k_{2,1} & -k_{2,2} & \ddots & k_{n,2} \\ \vdots & \ddots & \ddots & \vdots \\ k_{n,1} & \dots & \dots & -k_{n,n} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix}$$

Anwendung findet dieses Modell unter anderem in der Pharmakokinetik, da sich mit ihm der Abbau eines Medikaments beschreiben lässt. Betrachtet wird nun eine einfache Version des Modells.

Das Kompartiment K_1 repräsentiert die Blutbahn, K_2 das menschliche Gewebe und K_3 die Außenwelt. Die Übergangsrate vom Blut ins Gewebe sei β, die vom Gewebe ins Blut γ und α entspreche der Ausscheidungsrate, d. h. der Rate, mit der das Medikament vom Blut in die Außentwelt übertragen wird (z. B. durch die Niere).



Als System erhält man

$$\begin{cases} \dot{y}_1 = -(\alpha+\beta)y_1 + \gamma y_2 \\ \dot{y}_2 = \beta y_1 - \gamma y_2 \\ \dot{y}_3 = \alpha y_1 \end{cases} , \alpha, \beta, \gamma > 0$$

Lässt man nun die dritte, uninteressante Gleichung weg und substituiert $\mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}t}$, führt dies auf

$$\begin{cases} \mathrm{D}y_1 + (\alpha+\beta)y_1 - \gamma y_2 = 0 & \text{(I)} \\ \mathrm{D}y_2 - \beta y_1 + \gamma y_2 = 0 & \text{(II)} \end{cases}$$

Multipliziert man (II) mit D, ersetzt $\mathrm{D}y_1$ durch $-(\alpha+\beta)y_1 + \gamma y_2$ und anschließend y_1 durch $\frac{\dot{y}_2+\gamma y_2}{\beta}$, so erhält man die Gleichung

$$\mathrm{D}^2 y_2 - \beta\left(-(\alpha+\beta)y_1 + \gamma y_2\right) + \gamma \mathrm{D}y_2 = \mathrm{D}^2 y_2 + (\alpha+\beta+\gamma)\mathrm{D}y_2 + \alpha\gamma y_2 = 0$$

Diese Gleichung hat als allgemeine Lösung

$$y_2(t; c_3, c_4) = c_3 e^{\lambda_1 t} + c_4 e^{\lambda_2 t}$$

mit $\lambda_1 = \frac{-(\alpha+\beta+\gamma)-\sqrt{(\alpha+\beta+\gamma)^2-4\alpha\gamma}}{2}$, $\lambda_2 = \frac{-(\alpha+\beta+\gamma)+\sqrt{(\alpha+\beta+\gamma)^2-4\alpha\gamma}}{2}$. Wegen $\alpha, \beta, \gamma > 0$ gilt einerseits

$$(\alpha+\beta+\gamma)^2 - 4\alpha\gamma = (\alpha+\gamma)^2 - 4\alpha\gamma + 2\beta(\alpha+\gamma) + \beta^2 = (\alpha-\gamma)^2 + 2\beta(\alpha+\gamma) + \beta^2 > 0$$

und somit $\lambda_1, \lambda_2 \in \mathbb{R}$, andererseits gilt

$$\sqrt{(\alpha+\beta+\gamma)^2 - 4\alpha\gamma} < \sqrt{(\alpha+\beta+\gamma)^2},$$

woraus $\lambda_1, \lambda_2 < 0$ folgt.
Analog ergibt sich für y_1 als allgemeine Lösung

$$y_1(t; c_1, c_2) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

Setzt man beide Lösungen in die Gleichung $\dot{y}_1 = -(\alpha+\beta)y_1 + \gamma y_2$ ein, so erhält man durch Koeffizientenvergleich

$$c_3 = \frac{\lambda_1 + \alpha + \beta}{\gamma} c_1, \; c_4 = \frac{\lambda_2 + \alpha + \beta}{\gamma} c_2$$

Die Lösung des betrachteten (Teil-)Systems ist folglich

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = c_1 \cdot \begin{pmatrix} 1 \\ \frac{\lambda_1+\alpha+\beta}{\gamma} \end{pmatrix} e^{\lambda_1 t} + c_2 \cdot \begin{pmatrix} 1 \\ \frac{\lambda_2+\alpha+\beta}{\gamma} \end{pmatrix} e^{\lambda_2 t}$$

Seien nun die Anfangskonzentration im Blut, $y_1(0) = b_0$, sowie die im Gewebe, $y_2(0) = g_0$ gegeben. Diese Bedingungen führen auf das LGS

$$\begin{bmatrix} 1 & 1 \\ \frac{\lambda_1+\alpha+\beta}{\gamma} & \frac{\lambda_2+\alpha+\beta}{\gamma} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix},$$

dessen Lösung durch

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{\gamma}{\lambda_2 - \lambda_1} \begin{bmatrix} \frac{\lambda_2+\alpha+\beta}{\gamma} & -1 \\ -\frac{\lambda_1+\alpha+\beta}{\gamma} & 1 \end{bmatrix} \begin{pmatrix} b_0 \\ g_0 \end{pmatrix}$$

gegeben ist. Folglich:

$$y_1(t) = \frac{(\lambda_2+\alpha+\beta)\, b_0 - \gamma g_0}{\lambda_2 - \lambda_1} e^{\lambda_1 t} + \frac{-(\lambda_1+\alpha+\beta)\, b_0 + \gamma g_0}{\lambda_2 - \lambda_1} e^{\lambda_2 t}$$
$$y_2(t) = \frac{(\lambda_2+\alpha+\beta)\, b_0 - \gamma g_0}{\lambda_2 - \lambda_1} \frac{\lambda_1+\alpha+\beta}{\gamma} e^{\lambda_1 t} + \frac{-(\lambda_1+\alpha+\beta)\, b_0 + \gamma g_0}{\lambda_2 - \lambda_1} \frac{\lambda_2+\alpha+\beta}{\gamma} e^{\lambda_2 t}$$

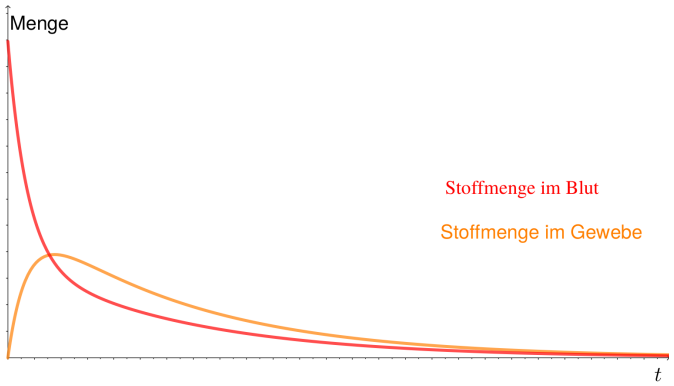


Abbildung 1.6.1: Beispielhafte Stoffmengen, aufgetragen gegen die Zeit.

1.6.4 Das Reduktionsverfahren von d'Alembert

Abschließend soll untersucht werden, wie im allgemeineren Falle einer Differentialgleichung der Form $\underline{y}'(x) = A(x)\underline{y}(x)$ mit nichtkonstanten Koeffizienten evtl. doch noch eine Lösung berechnet werden kann. Hat mein eine spezielle Lösung $\underline{y}_b = (y_{b,1}, y_{b,2}, ..., y_{b,n})^T$ gefunden - bspw. durch Raten - so lässt mittels des ***Reduktionsverfahrens von d'Alembert*** das System mit n Gleichungen auf ein System mit $n-1$ Gleichungen zurückführen. Der entsprechende Ansatz lautet wie folgt:

$$\underline{y}(x) = \psi(x) \cdot \underline{y}_b(x) + \underline{z}(x), \quad \underline{z}(x) = \begin{pmatrix} 0 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}, \ \psi \in C^1(\mathbb{R}) \tag{1.6.1}$$

Angenommen, $\underline{y}$ ist eine weitere Lösung. Differentiation liefert einerseits

$$\underline{y}' = \psi \cdot \underline{y}_b' + \psi' \cdot \underline{y}_b + \underline{z}' = \psi \cdot A\underline{y}_b + \psi' \cdot \underline{y}_b + \underline{z}',$$

andererseits gilt n. V.

$$\underline{y}' = A\underline{y} = A\left(\psi \cdot \underline{y}_b\right) + A\underline{z},$$

woraus

$$\underline{z}' = A\underline{z} - \psi'\underline{y}_b \tag{1.6.2}$$

folgt. Gilt umgekehrt (1.6.2), so ist (1.6.1) eine weitere Lösung. Für die erste Komponenten von $\underline{z}'$ folgt

$$\sum_{j=2}^{n} a_{1,j} z_j = \psi' y_{b,1},$$

für die restlichen $(2 \leq i \leq n)$):

$$z_i' = \sum_{j=2}^{n} a_{i,j} z_j - \psi' y_{b,i}.$$

Ist $y_{b,1}(x) \neq 0$, so lässt sich ψ' hierbei durch erstere Gleichung ersetzen, man erhält:

$$z_i' = \sum_{j=2}^{n} \left(a_{i,j} - \frac{y_{b,i}}{y_{b,1}} a_{1,j} \right) z_j, \ 2 \leq i \leq n$$

Dies entspricht einem homogenen System von $n-1$ Differentialgleichungen. Kennt man Lösungen $z_2, ..., z_n$, so lässt sich mit Hilfe des Ansatzes

$$\text{„}\psi(x) = \int \left(\frac{1}{y_{b,1}} \sum_{j=2}^{n} a_{1,j} z_j \right) \mathrm{d}x\text{“}$$

$\psi(x)$ und damit $\underline{y}(x)$ berechnen.

Beispiel:

Gegeben sei das System

$$\begin{cases} y_1' = \dfrac{1}{x} y_1 - y_2 \\ y_2' = \dfrac{1}{x^2} y_1 + \dfrac{2}{x} y_2 \end{cases}$$

In diesem Falle ist also $A(x) = \begin{bmatrix} \frac{1}{x} & -1 \\ \frac{1}{x^2} & \frac{2}{x} \end{bmatrix}$. Eine Lösung ist $\underline{y}_b(x) = \begin{pmatrix} x^2 \\ -x \end{pmatrix}$. Sei nun $\underline{z}(x) =$

$\begin{pmatrix} 0 \\ z_2 \end{pmatrix}, \psi \in C^1(\mathbb{R})$, $\underline{y}(x) = \psi(x)\underline{y_b}(x) + \underline{z}(x)$. Mit obigem ergibt sich

$$\begin{pmatrix} 0 \\ z_2'(x) \end{pmatrix} = \begin{bmatrix} \frac{1}{x} & -1 \\ \frac{1}{x^2} & \frac{2}{x} \end{bmatrix} \cdot \begin{pmatrix} 0 \\ z_2(x) \end{pmatrix} - \psi'(x) \cdot \begin{pmatrix} x^2 \\ -x \end{pmatrix} = \begin{pmatrix} -z_2(x) - \psi'(x)x^2 \\ \frac{2}{x} z_2(x) + \psi'(x) \cdot x \end{pmatrix}$$

Aus der ersten Komponenten ergibt sich $\psi'(x) = -\frac{z_2(x)}{x^2}$. Setzt man dies in die zweite ein, erhält man $z_2'(x) = \left(\frac{2}{x} - \frac{x}{x^2}\right) = \frac{1}{x} z_2(x)$. Man sieht leicht, dass $z_2(x) = x$ eine Lösung ist. Für diese folgt $\psi'(x) = -\frac{x}{x^2} = -\frac{1}{x}$, was von $\psi(x) = -\log(x)$ gelöst wird. Als weitere Lösung des Systems ergibt sich somit

$$\underline{y}(x) = \psi(x) \cdot \underline{y_b}(x) + \underline{z}(x) = -\log(x) \cdot \begin{pmatrix} x^2 \\ -x \end{pmatrix} + \begin{pmatrix} 0 \\ x \end{pmatrix} = \begin{pmatrix} -x^2 \log(x) \\ x \log(x) + x \end{pmatrix}$$

.

1.7 Differentialgleichungen höherer Ordnung

Bisher wurden nur Differentialgleichungen erster Ordnung betrachtet. Nun sei jedoch eine DGL der Ordnung n, d. h. der Form

$$y^{(n)}(x) + a_{n-1}(x) y^{(n-1)}(x) + \ldots + a_0(x) y(x) = b(x) \tag{1.7.1}$$

gegeben. Eine solche DGL lässt sich in ein äquivalentes Differentialgleichungssystem erster Ordnung Umwandeln. Setzt man

$$\begin{aligned} y_1(x) &:= y(x) \\ y_2(x) &:= y'(x) \\ y_3(x) &:= y''(x) \\ &\vdots \\ y_n(x) &:= y^{(n-1)}(x) \end{aligned},$$

so ergibt Ableiten:

$$\begin{aligned} y_1'(x) &= y_2(x) \\ y_2'(x) &= y_3(x) \\ &\vdots \\ y_{n-1}'(x) &= y_n(x) \\ y_n'(x) &= b(x) - a_{n-1}(x) y_n(x) - a_{n-2}(x) y_{n-1}(x) - \ldots - a_0(x) y_1(x) \end{aligned}$$

Dargestellt in Vektor-Form:

$$\frac{\mathrm{d}}{\mathrm{d}x} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \\ -a_0(x) & -a_1(x) & \cdots & \cdots & \cdots & \cdots & -a_{n-1}(x) \end{bmatrix}}_{A} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(x) \end{pmatrix}}_{\underline{b}}$$

Wegen $\operatorname{tr} A = -a_{n-1}$ lautet die Wronski-Determinante

$$\phi(x) = \phi(x_0) \cdot e^{\int_{x_0}^{x} -a_{n-1}(s)\mathrm{d}s}.$$

Sei

$$Y(x) = \begin{bmatrix} y_1(x) & y_2(x) & \cdots & y_n(x) \\ y_1'(x) & y_2'(x) & \cdots & y_n'(x) \\ \vdots & \ddots & \ddots & \vdots \\ y_1^{(n-1)}(x) & y_2^{(n-1)} & \cdots & y_n^{(n-1)}(x) \end{bmatrix}$$

eine Lösungsmatrix für das homogene System $\underline{y}' = A\underline{y}$. Die partikuläre Lösung $\underline{y}_p$ wird erneut mittels Variation der Konstanten, d. h. $\underline{y}_p(x) = Y(x)c(x)$ bestimmt. Wie bereits gesehen, ergibt sich die Gleichung $Y(x)\underline{c}'(x) = \underline{b}$. Nach Cramer lautet die i. Komponente $c_i'(x) = b(x)\frac{\det Y_i(x)}{\det Y(x)}$ mit

$$Y_i(x) = \begin{bmatrix} y_1(x) & \ldots & y_{i-1}(x) & 0 & y_{i+1}(x) & \ldots & y_n(x) \\ y_1'(x) & \ldots & y_{i-1}'(x) & 0 & y_{i+1}'(x) & \ldots & y_n'(x) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_1^{(n-1)}(x) & \ldots & y_{i-1}^{(n-1)}(x) & b(x) & y_{i+1}^{(n-1)}(x) & \ldots & y_n^{(n-1)}(x) \end{bmatrix}$$

Entwickeln nach der i. Spalte liefert $c_i'(x) = \frac{(-1)^{n+i} b(x)\phi_i(x)}{\phi(x)}$ mit

$$\phi_i \coloneqq \det \widetilde{Y}_i = \det \begin{bmatrix} y_1(x) & \cdots & y_{i-1}(x) & y_{i+1}(x) & \cdots & y_n(x) \\ y_1'(x) & \ddots & y_{i-1}'(x) & y_{i+1}'(x) & \cdots & y_n'(x) \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ y_1^{(n-2)}(x) & \cdots & y_{i-1}^{(n-2)}(x) & y_{i+1}^{(n-2)}(x) & \cdots & y_n^{(n-2)}(x) \end{bmatrix}.$$

Damit erhält man wegen $c(x) = \int_{x_0}^{x} (Y(t))^{-1} \underline{b}(t)\mathrm{d}t$ als Lösung für $\underline{c}(x) = (c_1(x), c_2(x), \ldots, c_n(x))^T$:

$$c_i(x) = (-1)^{n+i} \int_{x_0}^{x} b(t) \frac{\phi_i(t)}{\phi(t)} \mathrm{d}t$$

Beispiel (System zweiter Ordnung):

$$y'' + a_1 y' + a_0 y = f(x)$$

Sei

$$Y(x) = \begin{bmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{bmatrix}$$

eine Lösungsmatrix. Hier ist $\widetilde{Y}_1(x) = y_2(x)$ und $\widetilde{Y}_2(x) = y_1(x)$. Als partikuläre Lösung

$$\begin{pmatrix} y_p \\ y_p' \end{pmatrix} = \underline{y}_p(x) = Y(x)\,\underline{c}(x)$$

erhält man somit:

$$\underline{y}_p(x) = y_1(x) \left(-\int_{x_0}^{x} f(t) \frac{y_2(t)}{\phi(t)} \mathrm{d}t \right) + y_2(x) \left(\int_{x_0}^{x} f(t) \frac{y_1(t)}{\phi(t)} \mathrm{d}t \right)$$

Sind die Koeffizienten konstant, so kann man zur Bestimmung einer Lösungsmatrix erneut den auf Euler zurückgehenden Exponentialansatz $y(x) = e^{\lambda x}$ verwenden. Eingesetzt in den

homogenen Teil der DGL (1.7.1) ergibt dies:

$$\left(\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_0\right) e^{\lambda x} = 0$$

Da $e^{\lambda x}$ stets ungleich 0 ist, muss λ Nullstelle des Polynoms

$$p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$$

sein. Ähnlich zu linearen Differentialgleichungen im $\mathbb{R}$ gibt es, je nach Art und Vielfachheit der Nullstelle λ, vier Fälle zu unterscheiden:

1. $\lambda \in \mathbb{R}$, Vielfachheit 1: $y(x) = e^{\lambda x}$
2. $\lambda \in \mathbb{C} \setminus \mathbb{R}$, Vielfachheit 1: $\lambda = \mu \pm i\nu, y(x) = e^{\mu x}\cos(\nu x), y(x) = e^{\mu x}\sin(\nu x)$
3. $\lambda \in \mathbb{R}$, Vielfachheit m: $y(x) = e^{\lambda x}, xe^{\lambda x}, \ldots, x^{m-1}e^{\lambda x}$
4. $\lambda \in \mathbb{C} \setminus \mathbb{R}$, Vielfachheit m: Lösungen für $y(x)$:

$$\begin{aligned} e^{\mu x}\cos(\nu x)\ , &\quad e^{\mu x}\sin(\nu x) \\ xe^{\mu x}\cos(\nu x)\ , &\quad xe^{\mu x}\sin(\nu x) \\ &\vdots \\ x^{m-1}e^{\mu x}\cos(\nu x)\ , &\quad x^{m-1}e^{\mu x}\sin(\nu x) \end{aligned}$$

Bemerkung: *Ähnlich wie bei Differentialgleichungen im $\mathbb{R}^n$ lässt sich auch bei Differentialgleichungen höherer Ordnung das Reduktionsverfahren von d'Alembert anwenden. (Dies ist nicht weiter verwunderlich, bedenkt man, dass man Gleichungen des letzteren Typs ein ein DGL-System im $\mathbb{R}^n$ umwandeln kann.) Kennt man eine Lösung $y_1(x) \neq 0$ der (homogenen!) Differentialgleichung $y^{(n)}+a_{n-1}y^{(n-1)}+...+a_0y = 0$, so führt der Produktansatz $y(x) = u(x)y_1(x)$ mittels Differentiation unter Verwendung der Leibnizregel auf*

$$\sum_{k=0}^{n}\binom{n}{k}u^{(k)}y_1^{(n-k)}+a_{n-1}\sum_{k=0}^{n-1}\binom{n-1}{k}u^{(k)}y_1^{(n-1-k)}+a_{n-2}\sum_{k=0}^{n-2}\binom{n-2}{k}u^{(k)}y_1^{(n-2-k)}+...+a_0uy_1 = 0$$

Für die linke Seite der Gleichung ergibt sich

$$\begin{aligned} &\sum_{k=0}^{n}\binom{n}{k}u^{(k)}y_1^{(n-k)} + a_{n-1}\sum_{k=0}^{n-1}\binom{n-1}{k}u^{(k)}y_1^{(n-1-k)} + ... + a_0uy_1 = \\ &= \sum_{k=1}^{n}\binom{n}{k}u^{(k)}y_1^{(n-k)} + a_{n-1}\sum_{k=1}^{n-1}\binom{n-1}{k}u^{(k)}y_1^{(n-1-k)} + ... + a_1u'y_1 \\ &\qquad + \underbrace{\left(uy_1^{(n)} + a_{n-1}uy_1^{(n-1)} + ... + a_1uy_1' + a_0uy_1\right)}_{=0} \end{aligned}$$

Substituiert man $u' = w$ und dividiert beide Seiten durch y_1, so entspricht dies nun einer DGL $(n-1)$. Ordnung, welche erneut die Form

$$w^{(n-1)} + \tilde{a}_{n-2}w^{(n-2)} + ... + \tilde{a}_0w = 0$$

hat. Bilden die Funktionen $w_2, ..., w_n$ ein Hauptsystem dieser DGL und sind $W_2, ..., W_n$ zugehörige Stammfunktionen, so hat man mit $\{y_1, y_1W_2, ..., y_1W_n\}$ ein Hauptsystem der ursprünglichen Gleichung gefunden. Im Falle einer inhomogenen DGL. lässt sich dieses Verfahren ebenfalls anwenden. Kennt man zum Beispiel zusätzlich zu einer Lösung y_1 der homogenen Gleichung eine partikuläre Lösung der Gleichung (1.7.1), so hat man die allgemeine Lösung mit

$$y(x; c_1, ..., c_n) = c_1y_1 + c_2y_1W_2 + ... + c_ny_1W_n + y_p$$

gefunden.

1.7.1 Eulersche Differentialgleichung

Einen Spezialfall von Differentialgleichungen höherer Ordnung stellen ***Eulersche Differentialgleichungen*** dar, welche die Form

$$a_n x^n y^{(n)} + a_{n-1} x^{n-1} y^{(n-1)} + ... + a_0 y = 0$$

haben. ($a_i, i = 0, ..., n$ konstant.) Ist y eine Lösung der Gleichung für $x > 0$, so ergibt die Kettenregel für $\tilde{y}(x) \coloneqq y(-x)$

$$a_n x^n \tilde{y}^{(n)}(x) + a_{n-1} x^{n-1} \tilde{y}^{(n-1)}(x) ... + a_0 \tilde{y}(x) = a_n (-x)^n y^{(n)}(-x) + ... + a_0 y(-x) = 0,$$

wenn $x < 0$ und somit $-x > 0$. Folglich genügt es, den Fall $x > 0$ zu betrachten. Die Substitution $x(t) = e^t, u(t) = y(x(t)) = y(e^t)$ führt auf eine Differentialgleichung der Form

$$\tilde{a}_n u^{(n)}(t) + \tilde{a}_{n-1} u^{(n-1)}(t) ... + \tilde{a}_0 u(t) = 0$$

Hat man für diese eine Lösung $u(t)$ gefunden, so ist $y(x) = u(\log(x))$ eine Lösung der ursprünglichen DGL. Man beachte, dass diese Art der Umformung auch dann funktioniert, wenn auf der rechten Seite eine von x abhängige Funktion ungleich der Nullfunktion steht.

Beispiel 1.7.1 ($n = 2$):

Gegeben sei die Differentialgleichung

$$a_2 x^2 y''(x) + a_1 x y'(x) + a_0 y(x) = 0$$

Mit $x = e^t$, $u(t) = y(e^t)$ ergibt sich $\dot{u}(t) = y'(x)x$, $\ddot{u}(t) = y'(x)x + y''(x)x^2$. Addiert man daher auf der linken Seite der Gleichung $0 = a_2 y'(x)x - a_2 y'(x)x$, so erhält man:

$$0 = a_2 \left(x^2 y''(x) + x y'(x)\right) + (a_1 - a_2) x y'(x) + a_0 y(x) = a_2 \ddot{u}(t) + (a_1 - a_2)\dot{u}(t) + a_0 u(t) = 0$$

Für $n = 2$ ist also $\tilde{a}_2 = a_2, \tilde{a}_1 = a_1 - a_2$ sowie $\tilde{a}_0 = a_0$.

Beispiel 1.7.2:

Gegeben sei die Differentialgleichung

$$x^2 y''(x) + x y'(x) + 2y(x) = \sin(\log(x)), x > 0$$

Dank des vorherigen Beispiels lässt sich die Gleichung für u direkt angeben:

$$\ddot{u}(t) + 2u(t) = \sin(t)$$

Die allgemeine Lösung der homogenen Differentialgleichung ist $u_h(t; c_1, c_2) = c_1 \cos(\sqrt{2}t) + c_2 \sin(\sqrt{2}t)$, eine partikuläre Lösung ist $u_p(t) = \sin(t)$. Somit lautet die allgemeine Lösung der ursprünglichen Gleichung

$$y(x; c_1, c_2) = u_h(\log(x); c_1, c_2) + u_p(\log(x)) = c_1 \cos(\sqrt{2}\log(x)) + \sin(\sqrt{2}\log(x)) + \sin(\log(x))$$

1.8 Stabilität

Im Folgenden sollen die Auswirkungen kleiner Störungen in den Anfangsbedingungen auf die Lösung analysiert werden. Gegeben sei hierzu die DGL $y'(x) = \lambda y$. Für $\lambda \neq 0$ hat das zugehörige Anfangswertproblem

$$\begin{cases} y' = \lambda y \\ y(x_0) = y_0 \end{cases}$$

die Lösung $y(x) = y_0 e^{\lambda x}$. Sei nun $\varepsilon > 0$. Das gestörte Anfangswertproblem

$$\begin{cases} z' = \lambda z \\ z(x_0) = y_0 + \varepsilon \end{cases}$$

hat als Lösung $z(x) = (y_0 + \varepsilon)e^{\lambda x}$. Für den Fehler $|y(x) - z(x)| = \varepsilon e^{\lambda x}$ gibt es zwei Fälle zu unterscheiden:

a) $\underline{\lambda < 0}$: In diesem Falle wird der Fehler mit steigendem x stets geringer, es gilt:

$$\lim_{x \to \infty} \varepsilon e^{\lambda y} = 0$$

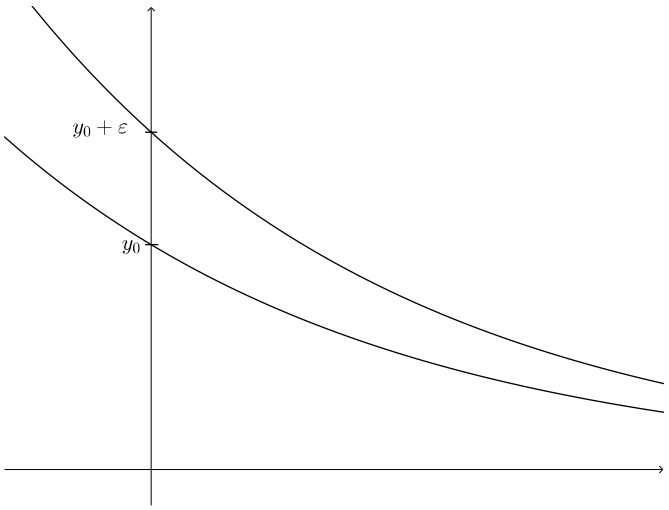


Abbildung 1.8.1: Die Lösungskurven nähern sich immer weiter an.

b) $\underline{\lambda > 0}$: In diesem Falle wächst der Fehler mit steigendem x, es gilt:

$$\lim_{x \to \infty} \varepsilon e^{\lambda y} = \infty$$

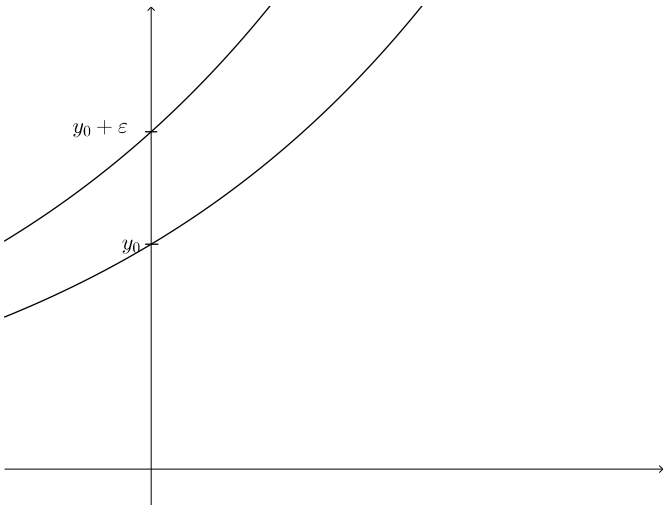


Abbildung 1.8.2: Die Lösungskurven entfernen sich voneinander.

Definition 1.8.1:
Die Lösung $y(x)$ eines Anfangswertproblems

$$\begin{cases} y' = f(x, y) \\ y(x_0) = y_0 \end{cases}$$

heißt ***stabil***, wenn zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, sodass für alle $\bar{y}_0$ mit $\|y_0 - \bar{y_0}\| < \delta$ für alle Lösungen des Anfangswertproblems

$$\begin{cases} \bar{y}' = f(x, \bar{y}) \\ \bar{y}(x_0) = \bar{y}_0 \end{cases}$$

die Ungleichung $\|y(x) - \bar{y}(x)\| < \varepsilon \; \forall x_0 \leq x < \infty$ gilt. (Kleine Störungen in den Eingangsdaten bewirken nur kleine Abweichungen von der eigentlichen Lösung).
Sie heißt ***asymptotisch stabil***, wenn zusätzlich ein $\tilde{\delta}$ existiert, sodass für alle $\bar{y}(x_0)$ die Implikation

$$\|y_0 - \bar{y}_0\| < \tilde{\delta} \quad \Rightarrow \quad \lim_{x \to \infty} \|y(x) - \bar{y}(x)\| = 0$$

gilt. Ist eine Lösung nicht stabil, so nennt man sie ***instabil***.

Satz 1.8.2 (Stabilität):
Seien $\lambda_1, \ldots, \lambda_p \, (p \leq n)$ die Eigenwerte von $A \in \mathbb{R}^{n \times n}$ und sei

$$\gamma := \max \{\operatorname{Re} \lambda_i \mid i \in \{1, \ldots, n\}\}$$

Dann ist die Lösung $\underline{y}(x) \equiv 0$ des Systems $\underline{y}'(x) = A\underline{y}(x)$

- asymptotisch stabil genau dann wenn $\gamma < 0$.
- stabil oder instabil, wenn $\gamma = 0$.
- instabil, wenn $\gamma > 0$.

Definition 1.8.3:
Eine Ruhelage des Systems

$$\underline{y}' = A\underline{y}$$

ist eine Lösung $\underline{y}^*$, für die $A\underline{y}^* = 0$ gilt.

Betrachtet man eine Differentialgleichung der Form

$$\underline{y}' = \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{A} \underline{y},$$

so lassen sich diese Ruhelagen anhand des Verhaltens der Lösungen in der Umgebung der jeweiligen Ruhelage in verschiedene Klassen einteilen. Maßgeblich für das Verhalten sind hierbei die Eigenwerte. Bestimmt man diese über die Nullstellen des charakteristischen Polynoms

$$\det(A - \lambda I) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - (a + d)\lambda + (ad - bc),$$

so ergeben sich mittels der Bedingung $\det(A - \lambda I) = 0 \Leftrightarrow \lambda^2 - \operatorname{tr} A \cdot \lambda + \det(A)$ die Eigenwerte

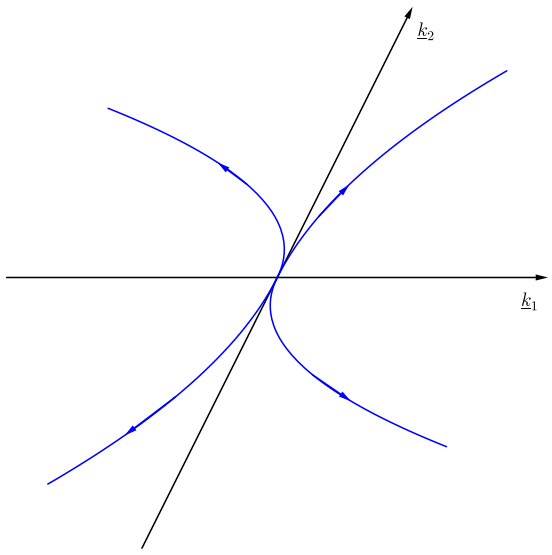
$$\lambda_{1,2} = \frac{\operatorname{tr} A \pm \overbrace{\sqrt{(\operatorname{tr} A)^2 - 4\det(A)}}^{=:\Delta}}{2}$$

Im Wesentlichen gibt es drei Fälle zu unterscheiden:

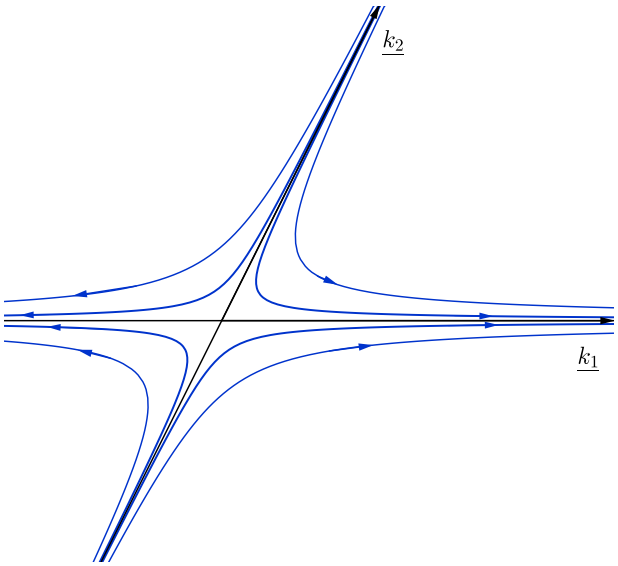
1. $\Delta > 0$: In diesem Falle gilt $\lambda_1, \lambda_2 \in \mathbb{R}, \lambda_1 \neq \lambda_2$.
2. $\Delta = 0$: Hier gilt $\lambda_1, \lambda_2 \in \mathbb{R}, \lambda_1 = \lambda_2$.
3. $\Delta < 0$: Hier sind $\lambda_1, \lambda_2 \in \mathbb{C} \setminus \mathbb{R}, \lambda_2 = \bar{\lambda}_1$.

Diese „Grobunterscheidung“ wurde bereits in den vorangegangenen Kapiteln behandelt. Jedoch lassen sich die Strukturen der Lösungen noch feiner unterscheiden, sofern man zusätzlich zu Δ die Lagen von λ_1 und λ_2 berücksichtigt. Die Lösungskurven werden hierbei bezüglich der Basis $\underline{k}_1, \underline{k}_2$ (Benennung gemäß Kapitel zu linearen DGL-Systemen) dargestellt.

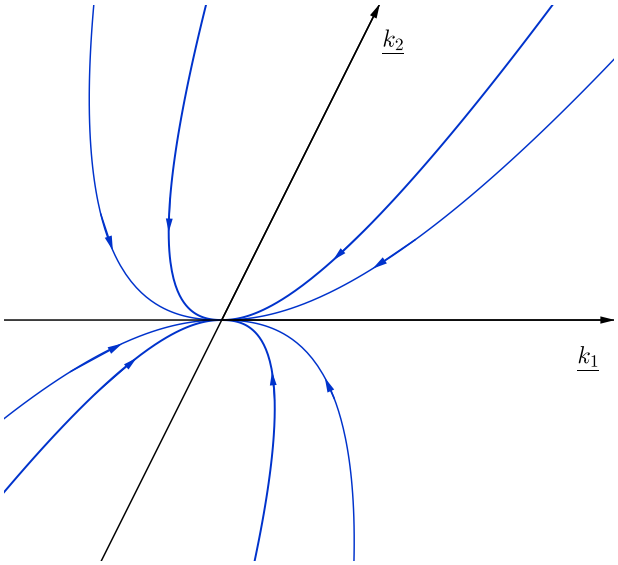
1. $\lambda_1 > \lambda_2 > 0$: Der Punkt $(0, 0)$ ist ein ***instabiler Knoten***, die entsprechende Lösung ist instabil:



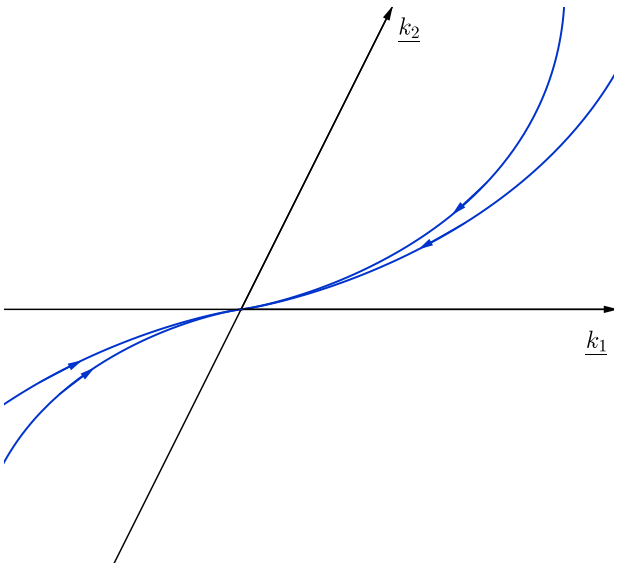
2. $\underline{\lambda_1 > 0 > \lambda_2}$: In $(0,0)$ liegt ein sogenannter ***Sattelpunkt*** vor, die entsprechende Lösung ist instabil:



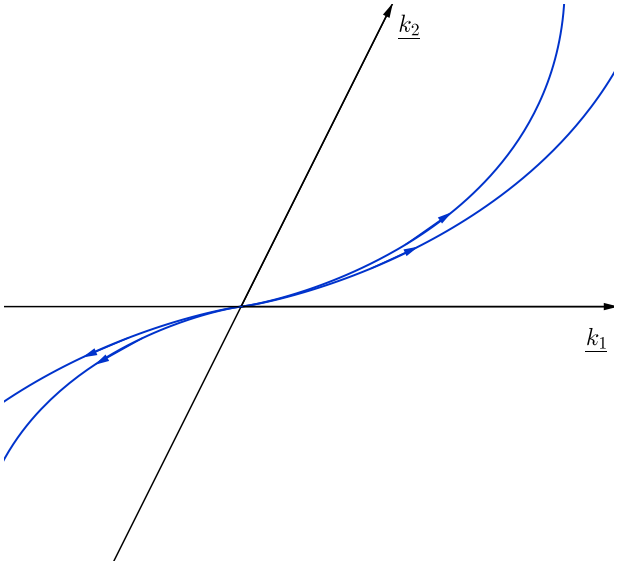
3. $\underline{\lambda_1 < \lambda_2 < 0}$: Hier ist $(0,0)$ ein ***stabiler Knoten***, die entsprechende Lösung asymptotisch stabil:



4. $\underline{\lambda_1 = \lambda_2 < 0}$: $(\Delta = 0)$ In diesem Falle liegt erneut ein stabiler Knoten in $(0,0)$ vor:

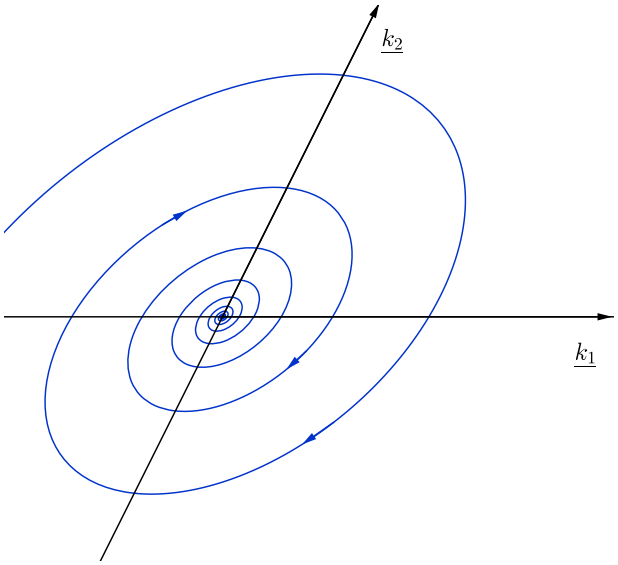


5. $\underline{\lambda_1 = \lambda_2 > 0}$: Die Ruhelage $(0,0)$ ist ein instabiler Knoten:

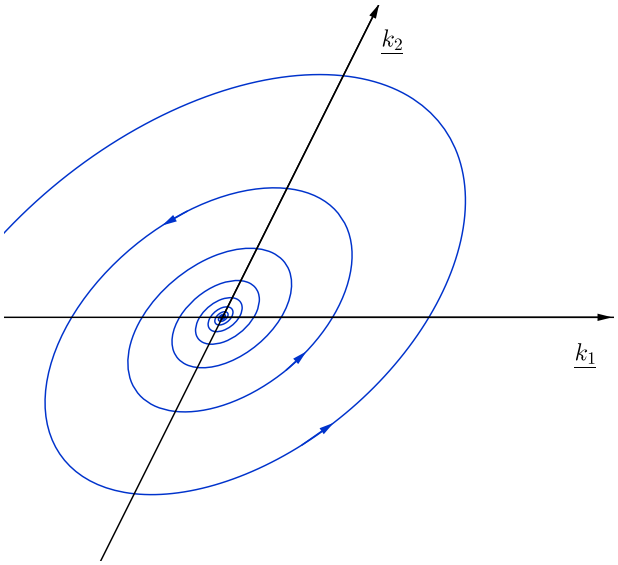


Im Falle $\Delta < 0$, d. h. $\lambda_{1,2} = \mu \pm i\nu \in \mathbb{C} \setminus \mathbb{R}$ hängt Art der Stabilität entscheidend vom Realteil μ ab:

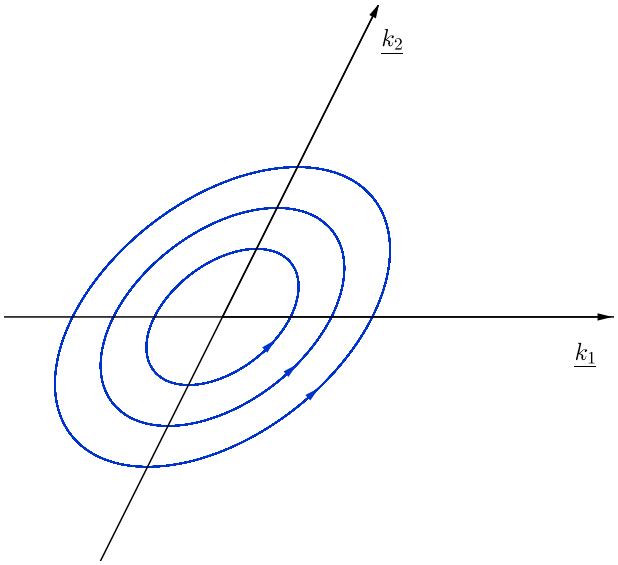
6. $\mu < 0$: In diesem Falle liegt ein ***stabiler Strudel*** vor, die entsprechende Lösung ist asymptotisch stabil:



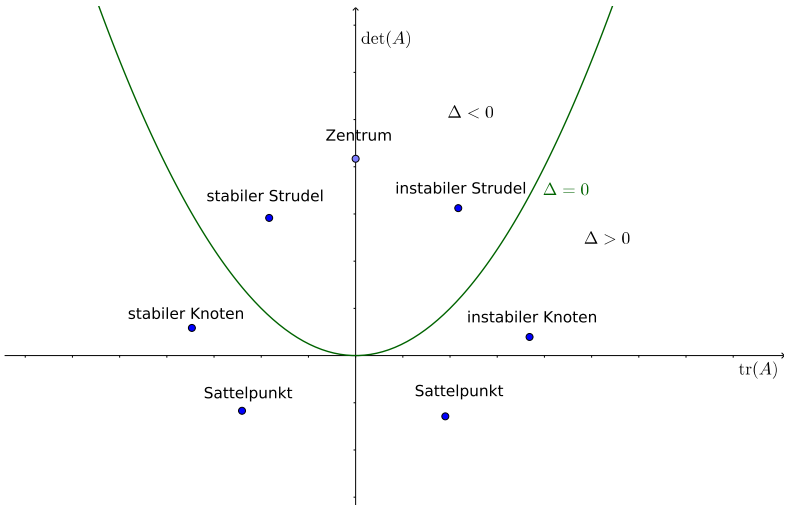
7. $\mu > 0$: Hier liegt ein ***instabiler Strudel*** vor, die entsprechende Lösung ist instabil:



8. $\underline{\mu = 0}$: In diesem Falle spricht man von einem ***Zentrum***, die entsprechende Lösung ist stabil (jedoch nicht asymptotisch):



Trägt man die Determinante von A gegenüber der Spur von A auf, ergibt sich folgender Gesamtüberblick:



1.8.2 Quasi-lineare Systeme

Gegeben sei folgende Differentialgleichung:

$$\underline{y}'(x) = A\underline{y}(x) + \underline{g}\left(x, \underline{y}(x)\right)$$

Auch in diesem Falle lassen sich unter bestimmten Voraussetzungen aus den Eigenwerten von A Aussagen über die Stabilität gewinnen.

Satz 1.8.4 (Theorem):
Sei g stetig für $x \geq x_0$ und $\left\|\underline{y}\right\| \leq b$ für ein $b < \infty$ sowie $\underline{g}(x, \underline{0}) = \underline{0}$. Ferner gelte

$$\lim_{\left\|\underline{y}\right\| \to 0} \frac{\left\|\underline{g}(x, \underline{y})\right\|}{\left\|\underline{y}\right\|} = 0$$

gleichmäßig für $x \geq x_0$. Sei $A \in \mathbb{R}^{n \times n}$ mit Eigenwerten $\lambda_1, \ldots, \lambda_p$, $p \leq n$ und $\gamma = \max\limits_{1 \leq i \leq p} \operatorname{Re} \lambda_i$
Dann ist die Ruhelage $\underline{y} \equiv \underline{0}$ des Systems $\underline{y}'(x) = A\underline{y}(x) + \underline{g}\left(x, \underline{y}(x)\right)$

- Asymptotisch stabil genau dann wenn $\gamma < 0$.
- Stabil oder instabil für $\gamma = 0$
- Instabil, wenn $\gamma > 0$

1.8.3 Stabilität für nichtlineare Systeme

Bisher wurde Stabilität nur für lineare Systeme betrachtet. In diesem Kapitel sollen Systeme der Form

$$\begin{aligned} y' &= f(y, z) \\ z' &= g(y, z) \end{aligned}$$

betrachtet werden, wobei f, g nichtlineare Funktionen sind.
Ist $\underline{f} = \underline{f}\left(\underline{y}\right)$ mit $\underline{f}(\underline{0}) = \underline{0}$ hinreichend glatt, so liefert Taylorentwicklung

$$\underline{f}\left(\underline{y}\right) = \underline{f}(\underline{0}) + \nabla \underline{f}(\underline{0}) \cdot \underline{y} + \ldots = 0 + \underbrace{A}_{:= \nabla \underline{f}(\underline{0})} \underline{y} + O\left(\left\|\underline{y}\right\|^2\right)$$

Dies soll im Folgenden genutzt werden.

Analog zu linearen System seien Lösungen (y^*, z^*) mit

$$\begin{aligned} f(y^*, z^*) &= 0 \\ g(y^*, z^*) &= 0 \end{aligned}$$

als ***Ruhelagen*** definiert. Um die Stabilität in diesen zu untersuchen, verschiebt man die Ruhelagen in den Nullpunkt (des entsprechenden Funktionenraumes) und linearisiert anschließend. Hierzu setzt man

$$\begin{aligned} \xi(x) &= y(x) - y^* \\ \eta(x) &= z(x) - z^* \end{aligned}$$

Anschließendes Ableiten liefert

$$\begin{aligned} \xi'(x) = y'(x) - 0 = f(y, z) - f(y^*, z^*) &\approx \frac{\partial f}{\partial y}(y^*, z^*)(y - y^*) + \frac{\partial f}{\partial z}(y^*, z^*)(z - z^*) \\ \eta'(x) = z'(x) - 0 = g(y, z) - g(y^*, z^*) &\approx \frac{\partial g}{\partial y}(y^*, z^*)(y - y^*) + \frac{\partial g}{\partial z}(y^*, z^*)(z - z^*) \end{aligned}$$

Näherungsweise ergibt sich also das System

$$\begin{cases} \xi' = \frac{\partial f}{\partial y}(y^*, z^*)\,\xi + \frac{\partial f}{\partial z}(y^*, z^*)\,\eta \\ \eta' = \frac{\partial g}{\partial y}(y^*, z^*)\,\xi + \frac{\partial g}{\partial z}(y^*, z^*)\,\eta \end{cases}$$

Ist das lineare System

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix}' = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \\ \frac{\partial g}{\partial y} & \frac{\partial g}{\partial z} \end{bmatrix}_{(y^*, z^*)}}_{=:A} \cdot \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

stabil (bzw. instabil) in der Ruhelage $(\bar{y}, \bar{z})$, dann ist dort auch das nichtlineare System stabil (bzw. instabil).

Beispiel 1.8.5:

$$\begin{cases} y_1' = y_2 \\ y_2' = -\sin(y_1) - y_2 \end{cases}$$

Aus

$$\begin{cases} y_2 = 0 \\ -\sin(y_1) - y_2 = 0 \end{cases}$$

ergeben sich die Ruhelagen $(k\pi, 0), \quad k \in \mathbb{Z}$. Für A gilt entsprechend:

$$A = \begin{bmatrix} \frac{\partial f}{\partial y_1} & \frac{\partial f}{\partial y_2} \\ \frac{\partial g}{\partial y_1} & \frac{\partial g}{\partial y_2} \end{bmatrix}_{(k\pi,0)} = \begin{bmatrix} 0 & 1 \\ -\cos(k\pi) & -1 \end{bmatrix}$$

Fallunterscheidung:

1. $\underline{k \in 2\mathbb{Z}}$:

   $$A_0 = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$$

   Bestimmung der Eigenwerte ergibt wegen $\chi_{A_0}(\lambda) = -\lambda(-1-\lambda)+1 = \lambda^2+\lambda+1 \stackrel{!}{=} 0$

   $$\lambda_{1,2} = \frac{-1 \pm i\sqrt{3}}{2}$$

   Aus $\gamma = \operatorname{Re}\lambda_{1,2} = -\frac{1}{2} < 0$ folgt, dass das System in der Ruhelage stabil ist, es ergibt sich ein stabiler Strudel. Der Imaginärteil beschreibt hierbei eine Rotation mit Frequenz $\frac{\sqrt{3}}{2}$.

2. $\underline{k \in 2\mathbb{Z}+1}$:

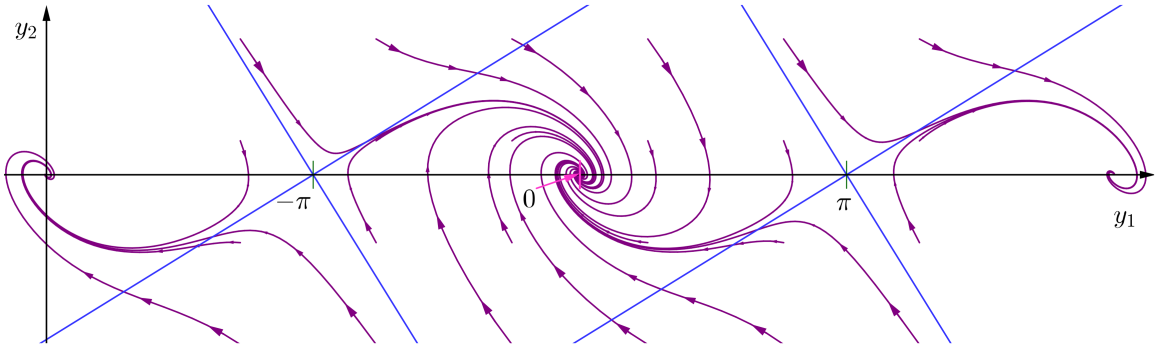
   $$A_\pi = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$$

   Lösen von $0 = \chi_{A\pi}(\lambda) = \lambda^2 + \lambda - 1$ ergibt als Eigenwerte:

   $$\lambda_{1,2} = \frac{-1 \pm \sqrt{5}}{2}$$

   Wegen $\lambda_1 < 0 < \lambda_2$ liegt hier ein Sattelpunkt vor.

Es ergibt sich folgendes Bild (welches sich aufgrund der Periodizität „links“ und „rechts“ unendlich oft wiederholt):



Beispiel 1.8.6:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} -y_2 \\ y_1 \end{pmatrix} + \varepsilon \left(y_1^2 + y_2^2\right) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \qquad (\varepsilon \neq 0)$$

Dies kann als quasi-lineares System

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \varepsilon \left(y_1^2 + y_2^2\right) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

oder als nichtlineares System

$$y_1' = \underbrace{-y_2 + \varepsilon \left(y_1^2 + y_2^2\right) y_1}_{f(y_1, y_2)} \tag{I}$$

$$y_2' = \underbrace{y_1 + \varepsilon \left(y_1^2 + y_2^2\right) y_2}_{g(y_1, y_2)} \tag{II}$$

mit Ruhelage $(0, 0)$ betrachtet werden. In beiden Fällen gilt für die zu betrachtende Matrix A:

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

sowie

$$\chi_A(\lambda) = \lambda^2 + 1$$

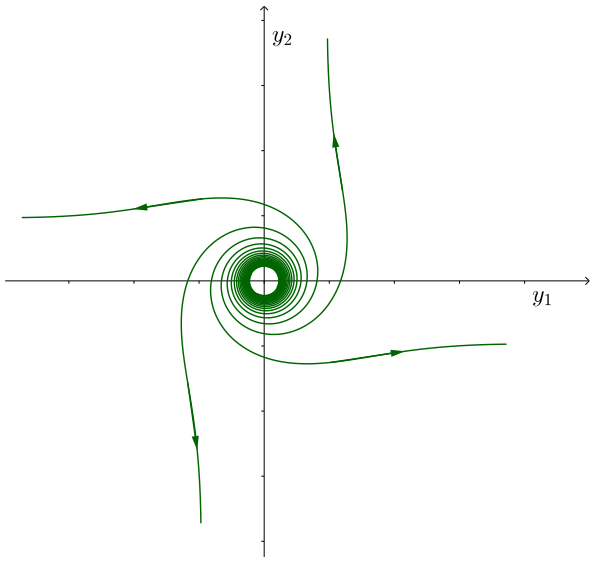
Die Eigenwerte sind somit $\lambda_{1,2} = \pm i$. Da der Realteil 0 ist, kann gemäß obigem Theorem zu quasi-linearen Systemen die Ruhelage des Systems stabil oder instabil sein. Um Aussagen über die Stabilität zu gewinnen, kann man Gleichung (I) mit y_1 sowie Gleichung (II) mit y_2 multiplizieren und anschließend addieren. Es ergibt sich:

$$y_1 y_1' + y_2 y_2' = \varepsilon \left(y_1^2 + y_2^2\right) \left(y_1^2 + y_2^2\right)$$

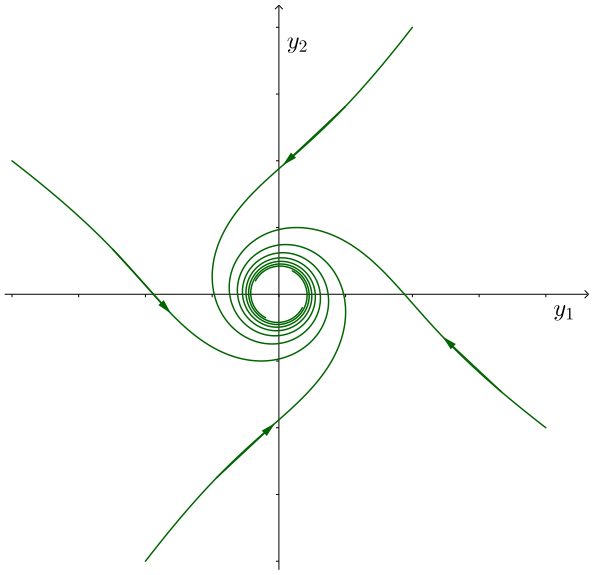
bzw.

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \left(y_1^2 + y_2^2\right) = \varepsilon \left(y_1^2 + y_2^2\right)^2$$

Substituiert man $u = y_1^2 + y_2^2$, erhält man die DGL $u' = 2\varepsilon u^2$. Hiermit lässt sich nun bestimmen, ob sich die Trajektorien mit der Zeit der Ruhelage annähern oder entfernen:
u beschreibt den Abstand (genauer dessen Quadrat) zur Ruhelage in der y_1-y_2-Ebene. Wird er kleiner, ist diese asymptotisch stabil, wird er größer, ist sie instabil. Da u^2 im Reellen außer in der Ruhelage immer positiv ist, ist ebendiese folglich asymptotisch stabil genau dann, wenn $\varepsilon < 0$ ist.



$\varepsilon = 1$: Ruhelage $(0,0)$ instabil



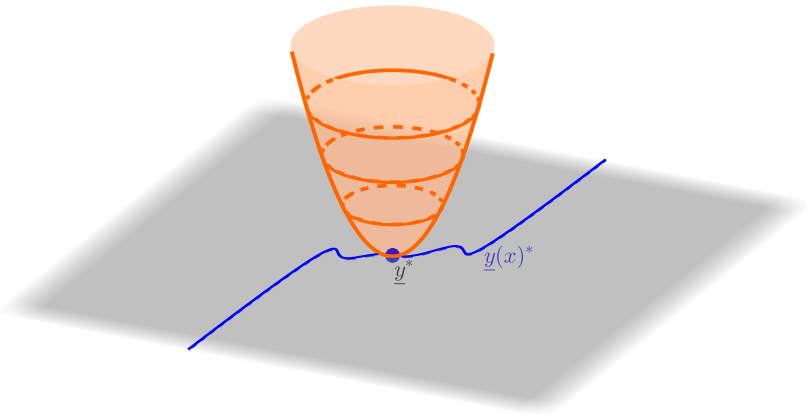
$\varepsilon = -1$: Ruhelage $(0,0)$ stabil

Definition 1.8.7:
Eine stetig differenzierbare Funktion $F : \mathbb{R}^n \to \mathbb{R}$ heißt ***Lyapunov-Funktion*** für das System $\underline{y}' = \underline{f}\left(\underline{y}\right)$ mit Ruhelage $\underline{y}^*$, wenn sie folgende Eigenschaften erfüllt:

1. $F\left(\underline{y}^*\right) = 0$
2. $F\left(\underline{y}\right) > 0$ in einer Umgebung $U \setminus \left\{\underline{y}^*\right\}$ von $\underline{y}^*$.
3. $\frac{\mathrm{d}}{\mathrm{d}x} F\left(\underline{y}(x)\right) \cdot \underline{f}\left(\underline{y}(x)\right) \leq 0 \quad \forall \underline{y}(x) \in U$

Satz 1.8.8:
Existiert eine Lyapunov-Funktion F für $\underline{y}' = \underline{f}\left(\underline{y}\right)$, dann ist die Ruhelage $\underline{y}^*$ stabil. Falls F eine strikte Lyapunov-Funktion ist, d. h. $\frac{\mathrm{d}}{\mathrm{d}x} F\left(\underline{y}\left(x\right)\right) < 0$, dann ist das System in der Ruhelage asymptotisch stabil.

Seitenansicht einer Möglichen Lyapunov-Funktion((orange)):

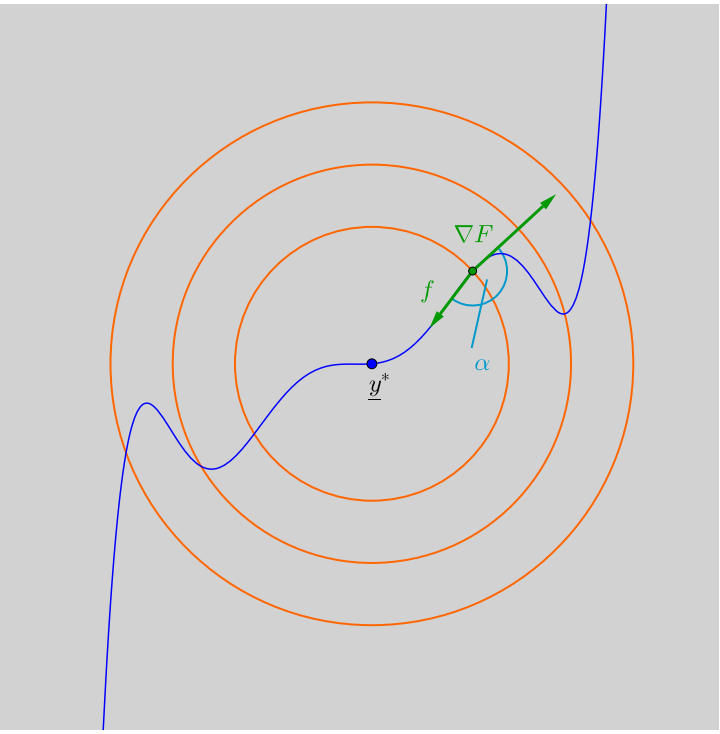


Ansicht von „oben“:

Die 3. Bedingung,

$$\frac{\mathrm{d}}{\mathrm{d}x} F\left(\underline{y}\left(x\right)\right) = \\ = \nabla F\left(\underline{y}\left(x\right)\right) \cdot \underline{f}\left(\underline{y}\left(x\right)\right) \leq 0,$$

impliziert $\alpha \geq 90°$, α ist hierbei der Winkel zwischen dem Gradienten der Lyapunov-Funktion und dem Richtungsvektor $\underline{f}$ (im entsprechenden Punkt).



Beispiel:

$$y_1' = -y_1^3 - 2y_2$$
$$y_2' = y_1 - y_2^3$$

Die Ruhelage ist $(y_1^*, y_2^*) = (0, 0)$. Es muss gelten:

$$\frac{\partial F}{\partial y_1}(y_1, y_2) \cdot \left(-y_1^3 - 2y_2\right) + \frac{\partial F}{\partial y_2}(y_1, y_2) \cdot \left(y_1 - y_2^3\right) \leq 0$$

Gilt

$$\frac{\partial F}{\partial y_1} = y_1, \frac{\partial F}{\partial y_2} = 2y_2,$$

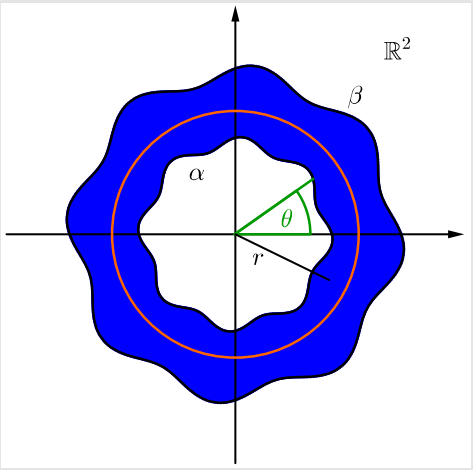
ist dies wegen $-\left(y_1^4 + 2y_2^4\right) \leq 0$ erfüllt. Da die Funktion $F(\underline{y}) = \frac{1}{2}y_1^2 + y_2^2$ dieser Bedingung sowie den anderen, im vorherigen Satz genannten, genügt, ist sie eine Lyapunov-Funktion.

Satz 1.8.9 (Folgerung aus Poincaré-Bendixson, Spezialfall):

Seien $0 \leq \alpha(\theta) < \beta(\theta)$ 2π-periodische (hinreichend glatte) Funktionen. Ferner sei $\mathfrak{V}$ der durch

$$\mathfrak{V} = \left\{ (r \cdot \cos(\theta), r \cdot \sin(\theta))^T : \alpha(\theta) \leq r \leq \beta(\theta) \right\}$$

definierte Ring, $D \subset \mathbb{R}^2$ ein Gebiet mit $\mathfrak{V} \subseteq \text{int}(D)$. Sei eine nichtlineare Gleichung $\underline{y}' = \underline{f}(\underline{y})$, $\underline{y} \in \mathbb{R}^2, \underline{f} \in C^1(D, \mathbb{R}^2)$ gegeben, so dass für alle $\underline{y} \in \partial\mathfrak{V}$ der Vektor $\underline{f}(\underline{y})$ nach innen zeigt, d. h. $\underline{n}(\underline{y}) \cdot \underline{f}(\underline{y}) < 0$, ferner existiere kein Punkt $x \in \mathfrak{V}$, sodass $\underline{f}(\underline{y}(x)) = 0$ gilt. Dann existiert ein ***Grenzzyklus*** in $\mathfrak{V}$, also - geometrisch gesprochen - eine periodische, geschlossene Trajektorie, welchem sich die anderen Trajektorien annähern.



Ring $\mathfrak{V}$ mit Grenzzyklus (orange)

Bemerkung 1.8.10:

1. *Das der Nullpunkt in obigem Satz Teil des Ringes sein muss, stellt keine Einschränkung dar: Ein Ring um einen Punkt $\underline{z}$ lässt sich durch eine Translation $\underline{y} + \underline{z}$ aus einem Ring um den Ursprung gewinnen. Auf die Gleichung $\underline{y}' = \tilde{\underline{f}}(\underline{y})$ mit $\tilde{\underline{f}}(\underline{y}) := \underline{f}(\underline{y} + \underline{z})$ lässt sich obiger Satz anwenden, die Lösungen der ursprünglichen DGL erhält man durch Translation $\underline{y} + \underline{z}$.*

2. *Das Theorem ist ebenfalls anwendbar, wenn der Richtungsvektor $\underline{f}$ in allen Randpunkten nach außen zeigt, sprich $\underline{n}(\underline{y}) \cdot \underline{f}(\underline{y}) > 0$: Löst $\underline{y}(x)$ die Gleichung $\underline{y}' = \underline{f}(\underline{y}(x))$, so löst $\underline{z}(x) := \underline{y}(-x)$ nach Kettenregel die DGL $\underline{z}'(x) = -\underline{f}(\underline{y}(-x)) = -\underline{f}(\underline{z}(x))$ (und vice versa). Für diese Gleichung zeigen alle Gradienten nach innen (auf dem Rand). Für den Grenzzyklus sowie die andere Trajektorien ändert sich nur die Richtung, in die sie durchlaufen werden. (Hierbei ist zu beachten, dass unter den genannten Voraussetzungen sowohl $\underline{y}(x)$ als auch $\underline{z}(x)$ auf dem Intervall $(-\infty, \infty)$ definiert sind.)*

3. *Zeigen die Gradienten nach innen, so nähern sich die Trajektorien dem Zyklus an, er ist **stabil**. Entfernen sie sich so, ist der Zyklus **instabil**.*

4. *Das Theorem gilt analog, wenn $\mathfrak{V}$ der Abschluss eines Gebietes ist, so dass der Fluss $\varphi_t(\mathfrak{V}) \subseteq \mathfrak{V}$ für alle $t > 0$ erfüllt.*

Beispiel:

Gegeben sei das System

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \alpha \left(1 - y_1^2 - y_2^2\right) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \qquad (*)$$

mit $\alpha \neq 0$ (sonst linear). Betrachtet man nun den Ring

$$\mathfrak{V} = \left\{ (r \cdot \cos(\theta), r \cdot \sin(\theta))^T : \frac{1}{2} \leq r \leq \frac{3}{2} \right\} = \overline{B_{\frac{3}{2}}(0)} \setminus B_{\frac{1}{2}}(0),$$

so gilt einerseits für $\underline{y} = (y_1, y_2)^T \in \partial B_{\frac{1}{2}}(0)$:

$$\begin{aligned}\underline{n}\left(\underline{y}\right) \cdot \underline{f}\left(\underline{y}\right) &= -2\left(y_1, y_2\right)\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix} - 2\alpha\left(1 - y_1^2 - y_2^2\right)\left(y_1, y_2\right)\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \\ &= -2\alpha\left(1 - y_1^2 - y_2^2\right)\left(y_1^2 + y_2^2\right) = \\ &= -2\alpha\left(1 - \frac{1}{4}\right)\frac{1}{4} = -\frac{3}{8}\operatorname{sign}\alpha\end{aligned} \quad ,$$

andererseits ergibt sich für $\underline{y} \in \partial B_{\frac{3}{2}}(0)$:

$$\begin{aligned}\underline{n}\left(\underline{y}\right) \cdot \underline{f}\left(\underline{y}\right) &= \frac{2}{3}\left(y_1, y_2\right)\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix} + \frac{2}{3}\alpha\left(1 - y_1^2 - y_2^2\right)\left(y_1, y_2\right)\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \\ &= \frac{2}{3}\alpha\left(1 - y_1^2 - y_2^2\right)\left(y_1^2 + y_2^2\right) = \\ &= \frac{2}{3}\alpha\left(1 - \frac{9}{4}\right)\frac{9}{4} = -\frac{15}{8}\operatorname{sign}\alpha\end{aligned} \quad ,$$

Das bedeutet, $\operatorname{sign}\left(\underline{n}\left(\underline{y}\right) \cdot \underline{f}\left(\underline{y}\right)\right) = \operatorname{sign}\alpha$ für alle $\underline{y} \in \partial\mathfrak{V}$. Da $\alpha \neq 0$, existiert also ein Grenzzyklus in diesem Ring. Im vorliegenden Fall lässt sich dieser sogar recht einfach bestimmen: Skalarmultipliziert man $(*)$ mit $\underline{y}$, so erhält man

$$y_1 y_1' + y_2 y_2' = \alpha\left(1 - y_1^2 - y_2^2\right)\left(y_1^2 + y_2^2\right),$$

durch Substitution $u^2 = y_1^2 + y_2^2$ folglich:

$$2uu' = 2y_1 y_1' + 2y_2 y_2'$$

Setzt man ferner $\theta = \arctan\left(\frac{y_2}{y_1}\right)$, ergibt Differentiation:

$$\theta' = \frac{1}{1 + \left(\frac{y_2}{y_1}\right)^2}\frac{y_2' y_1 - y_1' y_2}{y_1^2} = -1,$$

da

$$\begin{gathered}y_1 y_2' - y_1' y_2 = \\ = y_1\left[-y_1 + \alpha\left(1 - y_1^2 - y_2^2\right)y_2\right] + y_2\left[y_2 + \alpha\left(1 - y_1^2 - y_2^2\right)y_1\right] = \\ = -y_1^2 - y_2^2\end{gathered}$$

gilt. Hiermit lässt sich das System auch in Polarkoordinaten angeben:

$$\boxed{\begin{gathered}uu' = \alpha\left(1 - u^2\right)u^2 \\ \theta' = -1\end{gathered}} \quad ,$$

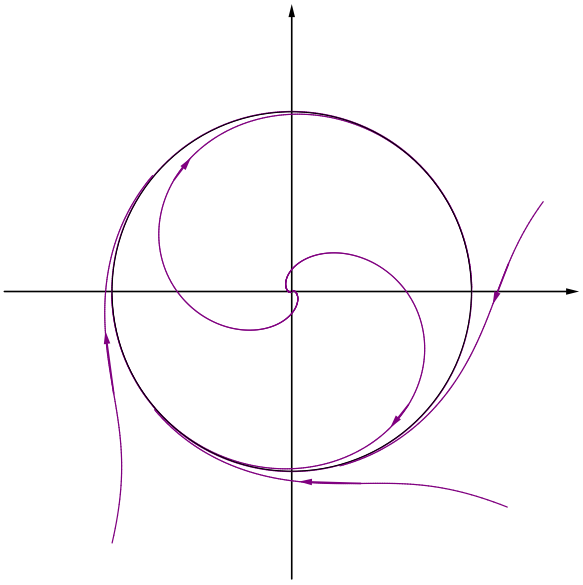
wobei die erste Gleichung die Änderung des Quadratabstands vom Ursprung beschreibt ($uu' = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}(u^2)$). Bei $u = 0$ hat man eine Ruhelage, für $u = 1$ einen Grenzzyklus, welcher u. U. stabil ist. Für die Stabilitätsanalyse betrachtet man kleine Störungen dieser Ruhelagen. Sei hierfür $\varepsilon > 0$. Wegen $u' = \alpha(1 - u^2)u$ verringert sich der Abstand zum Ursprung genau dann, wenn $\alpha < 0$.

Betrachtet man hingegen eine Störung der Ruhelage $u = 1$, so ergibt sich aus

$$\alpha\left(1-(1\pm\varepsilon)^2\right)(1\pm\varepsilon)$$

eine Vergrößerung des Abstands für $\alpha > 0$, für $\alpha < 0$ entsprechend eine Verkleinerung. Zusammengefasst ergibt sich für die Stabilität:

	Ruhelage, $u = 0$	Grenzzyklus, $u = 1$
$\alpha > 0$	instabil	stabil
$\alpha < 0$	stabil	instabil



Stabiler Grenzzyklus für $\alpha > 0$

1.9 Gradienten- und Hamilton-Systeme

An dieser Stelle folgt ein kleines Kapitel über zwei besondere Arten von linearen Differentialgleichungssystem sowie deren Eigenschaften. Die erste Kategorie, welche insbesondere in der Physik (vgl. Kapitel 2.7) Anwendung findet, bilden sogenannte Hamilton-Systeme.

Definition 1.9.1:
Sei $U = \mathbb{R}^d \times \mathbb{R}^d$, $H \in C^2(U)$ eine Funktion $H(q,p)$, $q,p \in C^1(\mathbb{R},\mathbb{R}^d)$. Das System

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial q} \end{aligned} \tag{1.9.1}$$

wird ***Hamiltonsystem*** genannt.

Hamilton-Systeme sind ***symplektisch***. Um diese Eigenschaft zu erklären, seien zwei Vektoren $\xi := \begin{pmatrix}\xi^p \\ \xi^q\end{pmatrix}, \eta := \begin{pmatrix}\eta^p \\ \eta^q\end{pmatrix}$ im $(p,q)-$Raum gegeben, wobei $\xi^p, \xi^q, \eta^p, \eta^q \in \mathbb{R}^d$. Ist $d = 1$, so ist die ***orientierte Fläche*** von $P = \{t\xi + s\eta | 0 \leq t \leq 1,\ 0 \leq s \leq 1\}$ gegeben durch

$$\text{or.area}(P) = \det\left(\begin{bmatrix}\xi^p & \eta^p \\ \xi^q & \eta^q\end{bmatrix}\right) = \xi^p\eta^q - \xi^q\eta^p,$$

Überträgt man dies auf den Fall $d > 1$ mittels

$$\omega(\xi,\eta) = \sum_{i=1}^{d} \det\left(\begin{bmatrix}\xi_i^p & \eta_i^p \\ \xi_i^q & \eta_i^q\end{bmatrix}\right) = \sum_{i=1}^{d} \xi_i^p\eta_i^q - \xi_i^q\eta_i^p,$$

so entspricht dies der Summe der orientierten Flächen der Projektionen von P auf die Koordinatenebenen. In Matrixnotation geschrieben ist

$$\omega(\xi,\eta) = \xi^T J \eta$$

mit $J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$, wobei I die Identität der Dimension d ist.

Definition 1.9.2:
Eine lineare Abbildung $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ wird symplektisch genannt, wenn $A^T JA = J$ gilt oder äquivalent, falls $\omega(A\xi, A\eta) = \omega(\xi, \eta)$ erfüllt ist.

Definition 1.9.3 (symplektische Abbildungen):
Eine differenzierbare Abbildung $g : U \to \mathbb{R}^{2d}$ ($U \subset \mathbb{R}^{2d}$ offen) wird symplektisch genannt, falls die Jacobi-Matrix $g'(q,p)$ überall symplektisch ist, d. h. falls

$$g'(q,p)^T J g'(q,p) = J \quad \forall (q,p) \in U$$

bzw.

$$\omega(g'(q,p)\,\xi, g'(q,p)\,\eta) = \omega(\xi,\eta) \quad \forall (q,p) \in U$$

gilt.

Anschaulich gesprochen, erhalten symplektische Abbildungen die Summe der orientierten Flächen.

Mithilfe der Notationen $y = (p,q)$ sowie $\nabla H(y) = \left(\frac{\partial}{\partial y} H(y)\right)^T$ lässt sich das Hamilton-System (1.9.1) schreiben als

$$\dot{y} = J^{-1} \nabla H(y) \tag{1.9.2}$$

Der ***Fluss*** eines Hamilton-Systems ist eine Abbildung $\varphi_t : U \to \mathbb{R}^{2d}$, die $(*)$ erfüllt, d. h.

$$\varphi_t(p_0, q_0) = (p(t, p_0, q_0), q(t, p_0, q_0)),$$

wobei $(p(t, p_0, q_0), q(t, p_0, q_0)\,(*))$ Lösung des Systems $(*)$ zu den Anfangswerten $p(0) = p_0$ und $q(0) = q_0$ ist. Durch die Darstellung (1.9.2) des Hamilton-Systems motiviert ist folgende Definition:

Definition 1.9.4:
Gegeben sei eine Differentialgleichung $\dot{y} = f(y)$. Dann wird $\dot{y} = f(y)$ ***lokal hamiltonisch*** genannt, falls jedes $y_0 \in U$ eine Umgebung hat, in der $f(y) = J^{-1}\nabla H(y) \quad (y \in U)$ für eine Funktion H gilt.

Satz 1.9.5:
Sei $f : U \to \mathbb{R}^{2d}$ stetig differenzierbar. Dann ist das Differentialgleichungssystem $\dot{y} = f(y)$ lokal hamiltonisch genau dann, wenn dessen Fluss $f_t(y)$ symplektisch ist für alle $y \in U$ und alle hinreichend kleinen t.

Neben diesen sind auch Gradientensysteme von Bedeutung. Gegeben sei hierzu eine offene Teilmenge E des $\mathbb{R}^n$ und $V \in C^2(E)$. Das System $\dot{x} = -\text{grad}\, V(x)$ mit

$$\text{grad}\, V(x) = \left(\frac{\partial V}{\partial x_1}, \ldots, \frac{\partial V}{\partial x_n}\right)^T$$

wird ***Gradientensystem*** genannt, V wird als ***Potenzial*** bezeichnet.

Satz 1.9.6:
Sei V wie oben, x eine Funktion $\mathbb{R} \to U$ gegeben. Es gilt $\frac{\mathrm{d}}{\mathrm{d}t} V(x) \leq 0$ mit Gleichheit genau dann, wenn x ein Gleichgewichtspunkt ist.

Da V nach Voraussetzung zweimal stetig differenzierbar ist, ist die Hessematrix $\mathrm{D}^2 V(x) = \left[\frac{\partial^2 V}{\partial x_i \partial x_j}\right]_{i,j}$ symmetrisch (Satz von Schwarz). Hieraus ergibt sich folgender Satz:

Satz 1.9.7:
An einem kritischen Punkt eines Gradientensystems sind alle Eigenwerte der Hessematrix $\mathrm{D}^2 V$ reell. Sind sie ferner alle strikt positiv, so ist die in diesem Punkt vorliegende Ruhelage asymptotisch stabil.

Beispiel:

Gegeben sei

$$V(x) = \frac{1}{4}x_1^4 - x_1^3 + x_1^2 + x_2^2$$

sowie das Gradientensystem

$$\dot{x} = f(x) = -\operatorname{grad} V(x) = \begin{pmatrix} -x_1^3 + 3x_1^2 - 2x_1 \\ -2x_2 \end{pmatrix}$$

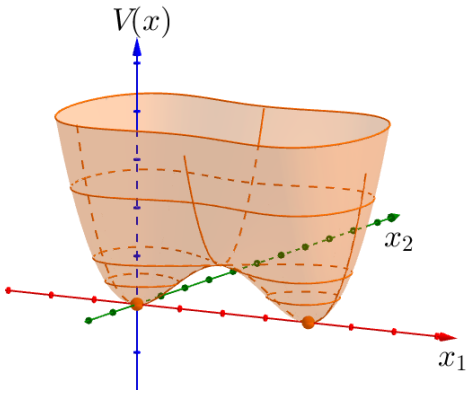


Abbildung 1.9.1: Potenzial $V(x)$

Wegen $-x_1^3 + 3x_1^2 - 2x_1 = -x_1(x_1 - 1)(x_1 - 2)$ sind die kritischen Punkte des Systems gegeben durch 1) $x^{(1)} = (0,0)$, 2) $x^{(2)} = (1,0)$ und 3) $x^{(3)} = (2,0)$. Die Jacobi-Matrix von f ist gegeben durch

$$\mathrm{D}f(x) = -\begin{bmatrix} \frac{\partial^2}{\partial x_1^2} V(x) & \frac{\partial^2}{\partial x_1 \partial x_2} V(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} V(x) & \frac{\partial^2}{\partial x_2^2} V(x) \end{bmatrix} = \begin{bmatrix} -3x_1^2 + 6x_1 - 2 & 0 \\ 0 & -2 \end{bmatrix},$$

die zugehörigen Eigenwerte sind $\lambda_1 = -3x_1^2 + 6x_1 - 2$ und $\lambda_2 = -2$, wie sich direkt ablesen lässt. Da λ_2 unabhängig von (x_1, x_2) stets kleiner 0 ist, hängt die Stabilität der Ruhelagen einzig von x_1 ab. Für diese gilt im Einzelnen:

1) $x^{(1)} = (0,0)$: Hier ist $x_1^{(1)} = 0$ und somit $\lambda_1 = -x_1^{(1)} + 6x_1^{(1)} - 2 = -2 < 0$, die Ruhelage ist folglich stabil.

2) $x^{(2)} = (1,0)$: Hier ist $x_1^{(2)} = 1$ und somit $\lambda_1 = 1 > 0$, es liegt ein Sattelpunkt vor.

3) $x^{(3)} = (2,0)$: Hier ist $x_1^{(3)} = 2$ und somit $\lambda_1 = -2 < 0$, die Ruhelage ist asymptotisch stabil.

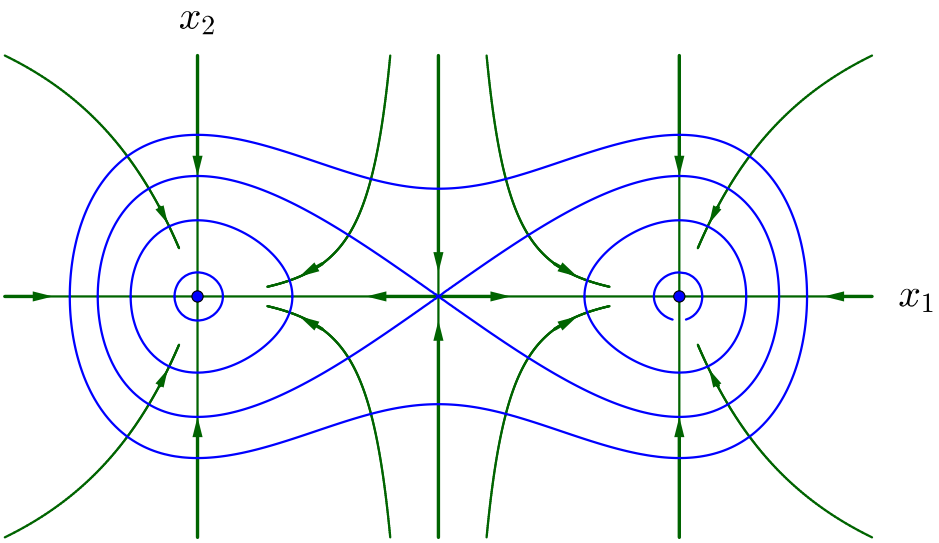


Abbildung 1.9.2: Phasenraumportät. Die Trajektorien (grün) schneiden die Niveaulinien $V(x) = \text{const.}$ (blau) rechtwinklig.

Definition 1.9.8:

Gegeben sei das planare System (Gradientensystem im $\mathbb{R}^2$) $\begin{matrix}\dot{x} = P(x,y) \\ \dot{y} = Q(x,y)\end{matrix}$. Das ***orthogonale System*** ist definiert durch

$$\begin{aligned}\dot{x} &= Q(x,y) \\ \dot{y} &= -P(x,y)\end{aligned}$$

.

Aus dieser Definition ergibt sich sofort, dass das orthogonale System eines Hamilton-Systems ein Gradientensystem ist.

1.10 Randwertprobleme

Bisher wurden Anfangswertprobleme betrachtet, d. h. Problemstellungen der Form $y'' = f(x, y, y')$ mit vorgegebenen Werten $y(x_0) = a, y'(x_0) = b$. Anstatt sämtliche Werte zum Anfangspunkt x_0 festzulegen, kann man jedoch auch Werte am Intervallende vorgeben, z. B.:

$$\begin{aligned}y'' &= f(x, y, y') \\ y(0) &= a, y(L) = b\end{aligned}$$

Ein solches Problem heißt ***Randwertproblem*** (RWP). Der Versuch, ein solches zu lösen führt auf das Problem globaler Existenz. Zur Erläuterung von ebendiesem sei das Randwertproblem $y'' = y^2 + y'^2$, $y'(0) = 1$, $y(1) = 0$ mit (gewünschtem) Existenzintervall $[0, 1]$ gegeben.
Sei y eine Lösung. Da $y'' = y^2 + y'^2 \geq 0$ und somit y' monoton wachsend ist, d. h. $y'(x) \geq y'(0) > 0 \, \forall x \in [0, 1]$, folgt für $x \in [0, 1]$:

(i) $y^2(x) + y'^2(x) \geq 1$ und somit $\frac{y''(x)}{y'^2(x)} \geq \frac{y''(x)}{y^2(x)+y'^2(x)} = 1$

(ii) $\frac{1}{y'(0)} - \frac{1}{y'(x)} < \frac{1}{y'(0)}$

Hieraus ergibt sich für $x \in (0,1]$:

$$0 < x = \int_0^x 1 \mathrm{d}t \overset{\text{(i)}}{\leq} \int_0^x \frac{y''(t)}{y'(t)^2} \mathrm{d}t = \left[\frac{-1}{y'(t)}\right]_0^x = \frac{1}{y'(0)} - \frac{1}{y'(x)} \overset{\text{(ii)}}{<} \frac{1}{y'(0)} = 1$$

Speziell für $x = 1$ ergibt sich ein Widerspruch, eine Lösung kann a priori nicht auf ganz $[0,1]$ existieren.
Im Weiteren sollen nur noch lineare Randwertprobleme 2. Ordnung betrachtet werden.

- Man kann zeigen, dass die Lösung auf jedem Intervall existiert, sofern sie auf einem Intervall existiert. Im Allgemeinen ist eine solche jedoch nicht eindeutig oder gar nicht erst existent.

Beispiel 1.10.1:

$$y'' + y = 0, y(0) = 0, y(\pi) = 1$$

Die allgemeine Lösung der DGL ist:

$$y(x) = c_1 \sin x + c_2 \cos x$$

Aus $y(0) = 0$ folgt $c_2 = 0$, aber $y(\pi) = 1$ ist für kein c_1 erfüllt.

Beispiel 1.10.2:

$$y'' + y = 0, y(0) = -1, y(\pi) = 1$$

Wie im vorherigen Beispiel ist die allgemeine Lösung:

$$y(x) = c_1 \sin x + c_2 \cos x$$

Für $y(0) = -1$ ist $c_2 = -1$ und die Randbedingung wird von jeder Funktion

$$y(x) = c_1 \sin x - \cos x,\ c_1 \in \mathbb{R}$$

erfüllt.

Sei nun

$$\begin{cases} Sy := y'' + a_1(x) y' + a_0(x) y = b(x) \\ R_1 y := \alpha_1 y(a) + \alpha_2 y'(a) = \varrho_1 \\ R_2 y := \beta_1 y(b) + \beta_2 y'(b) = \varrho_2 \end{cases} \tag{1.10.1}$$

Mit stetigen Funktionen $a_0, a_1, b : [a,b] \to \mathbb{R}$, $(\alpha_1, \alpha_2) \neq (0,0), (\beta_1, \beta_2) \neq (0,0)$.

Satz 1.10.3:
Gegeben sei das Randwertproblem (1.10.1) mit obigen Voraussetzungen. Sei $\{y_1, y_2\}$ Fundamentalsystem der homogenen Gleichung $Sy = 0$. Dann sind äquivalent:

1. (1.10.1) ist eindeutig lösbar.
2. $$\det\left(\begin{bmatrix} R_1y_1 & R_1y_2 \\ R_2y_1 & R_2y_2 \end{bmatrix}\right) \neq 0$$
3. Das homogene Randwertproblem $Sy = 0, R_1y = 0, R_2y = 0$ besitzt nur die triviale Lösung $y = 0$

Beweis: Sei y_p partikuläre Lösung von $Sy = b$ und $y(x) = c_1y_1(x) + c_2y_2(x) + y_p(x)$ Dann gilt:

$$\begin{gathered} R_1y = c_1R_1y_1 + c_2R_1y_2 + R_1y_p = \varrho_1 \\ R_2y = c_1R_2y_1 + c_2R_2y_2 + R_2y_p = \varrho_2 \\ \Leftrightarrow \\ \begin{bmatrix} R_1y_1 & R_1y_2 \\ R_2y_1 & R_2y_2 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \varrho_1 - R_1y_p \\ \varrho_2 - R_2y_p \end{pmatrix} \end{gathered}$$

□

Es lässt sich zeigen, dass (1.10.1) immer auf das Problem

$$S\tilde{y} = f, R_1\tilde{y} = R_2\tilde{y} = 0 \tag{1.10.2}$$

reduziert werden kann, wobei $f = b(x) - Su(x)$ ist und u eine beliebige zweimal stetig differenzierbarer Funktion u auf $[a, b]$ mit $R_1u = \varrho_1, R_2u = \varrho_2$ ist. Für eine Lösung y von (1.10.1) gilt dann $y = u + z$, wobei z (1.10.2) löst. ($Sy = Sz + Su = b(x) - Su + Su = b(x)$).
Man kann (1.10.1) auf seine ***selbst-adjungierte*** Form,

$$Ly := (py')' + qy = g,$$

bringen, wobei L ***Sturm-Liouville-Operator*** genannt wird: Man setzt $p(x) = e^{A(x)}$, wobei $A(x)$ Stammfunktion zu a_1 ist, sowie $q(x) = a_0(x)e^{A(x)}$ und $g(x) = b(x)e^{A(x)}$. Es gilt:

$$\begin{gathered} p'y' + py'' + qy = g \\ \Leftrightarrow \\ A'(x)e^{A(x)}y' + e^{A(x)}y'' + a_0(x)e^{A(x)}y = b(x)e^{A(x)} \\ \Leftrightarrow \\ y'' + a_1(x)y' + a_0(x)y = b(x) \end{gathered}$$

Ein Randwertproblem der Form

$$\begin{cases} Ly = (py')' + qy = g \\ R_1y = \varrho_1 \\ R_2y = \varrho_2 \end{cases} \tag{1.10.3}$$

heißt ***Sturm-Liouvillesches Randwertproblem***.
Mit Hilfe der sogenannten ***Greenschen Funktion*** lassen sich Lösungen von Randwertproblemen kompakt angeben:

Satz 1.10.4:
Gegeben sei das Randwertproblem

$$\begin{cases} Ly = g \\ R_1 y = 0 \\ R_2 y = 0 \end{cases}$$

Das zugehörige homogene Problem (d. h. $Ly = 0$) habe nur die triviale Lösung $y = 0$, ferner sei $\{y_1, y_2\}$ ein Fundamentalsystem der homogenen Gleichung. Dann ist ist die eindeutige Lösung des Randwertproblems durch

$$y(x) = \int_a^b G(x,\xi)\, g(\xi)\, d\xi$$

mit Greenscher Funktion

$$G(x,\xi) := \begin{cases} \frac{y_2(x)y_1(\xi)}{p(a)W(a)}, a \leq \xi \leq x \leq b \\ \frac{y_1(x)y_2(\xi)}{p(a)W(a)}, a \leq x < \xi \leq b \end{cases}$$

gegeben. Hierbei ist W ist die Wronski-Determinante des entsprechenden Fundamentalsystems.

Bemerkung 1.10.5:

- *Das dieser Satz nur für Probleme mit homogenen Randbedingungen $R_1 y = R_2 y = 0$ gilt, ist irrelevant. Analog zum bereits gesehen lassen sich beliebige Probleme der Form* (1.10.3) *auf diese „halbhomogene" Form bringen. Hat man nun eine beliebige C^2-Funktion y, welche die Randbedingungen erfüllt, so lässt sich die Lösung des Randwertproblems* (1.10.3) *mittels*

  $$y(x) = \tilde{y}(x) + \int_a^b G(x,\xi)\,(f(\xi) - L\tilde{y}(\xi))\,\mathrm{d}\xi$$

  schreiben.

- *Offensichtlich hängt G stark von den Koeffizientenfunktionen p und q ab. Die gute Nachricht ist jedoch: G hängt letztendlich nur von diesen Funktionen (sowie implizit dem Intervall) ab, insbesondere ist die Greensche Funktion unabhängig von den Randbedingungen und g.*

Beispiel 1.10.6:

Gegeben sei die Differentialgleichung

$$\begin{cases} -y''(x) = g(x),\ x \in (0,1) \\ y(0) = y(1) = 0 \end{cases}$$

Hier ist $p(x) \equiv -1$, $q(x) \equiv 0$ Ein Fundamentalsystem der homogenen Gleichung ist $y_1(x) = 1$, $y_2(x) = x$, die Wronski-Determinante lautet $W(x) = y_1(x)y_2'(x) - y_1'(x)y_2(x) = -1$. Somit ist die Greensche Funktion durch

$$G(x,t) = \begin{cases} x\,(1-t)\,, 0 \leq x \leq t \leq 1 \\ t\,(1-x)\,, 0 \leq t < x \leq 1 \end{cases}$$

gegeben. Das Ergebnis lässt sich auch mittels partieller Integration erhalten, verwendet man

den Lösungsansatz $y(x) = \int_a^b G(x,\xi)g(\xi)\mathrm{d}\xi$ sowie die Identität $y(x) = y(a) + \int_a^x y'(\xi)\mathrm{d}\xi = y(b) + \int_x^b y'(\xi)\mathrm{d}\xi$ (und analoges für $y''(x)$).

In einem engen Zusammenhang zu Sturm-Liouvilleschen Randwertproblemen steht das ***Sturm-Liouvillesche Eigenwertproblem***. Es hat die Form

$$\begin{aligned} Ly + \lambda r(x)y &= 0 \\ R_1 y &= 0 \\ R_2 y &= 0 \end{aligned}$$

mit positiver Gewichtsfunktion $r \in C^0([a,b])$. Analog zur linearen Algebra heißt λ ***Eigenwert*** von L, wenn eine Lösung $y \not\equiv 0$ existiert. Da der lineare Operator L selbstadjungiert ist, sind alle solchen Eigenwerte reell. Die zu einem Eigenwert λ_n gehörende Lösung y_n nennt man ***Eigenfunktion***. Unter den genannten Voraussetzungen existieren unendlich viele Eigenwerte $\lambda_1 < \lambda_2 < \ldots$ mit $\lim\limits_{n\to\infty} \lambda_n = \infty$. Aufgrund der Linearität sowie der Selbstadjungiertheit von L existiert jedoch ein Orthonormalsystem von Eigenfunktionen, d. h. eine Menge von Eigenfunktionen y_n mit

$$\int_a^b r(x)y_m(x)y_n(x)\mathrm{d}x = \delta_{mn}$$

Hierbei bezeichnet δ_{mn} das Kronecker-Delta, definiert durch $\delta_{mn} \coloneqq \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases}$. Ferner lassen sich alle Funktionen $\phi(x) \in C^1([a,b])$, die den Randbedingungen $R_1\phi = R_2\phi = 0$ genügen, in der Form

$$\phi(x) = \sum_{n=1}^{\infty} c_n y_n(x)$$

mit $c_n = \int_a^b r(x)\phi(x)y_n(x)\mathrm{d}x$ darstellen.
Abschließend soll noch ein anwendungsbezogenes Beispiel gegeben werden, in welchem Kontext Randwertprobleme auftreten können.

Beispiel (Wärmeleitungsgleichung):

Gegeben sei ein dünner Stab der Länge L. Physikalisch dünn bedeutet, dass $\frac{r}{L} \ll 1$ gilt, wobei r der Radius des Querschnitts ist. Die Temperaturverteilung $y(x,t)$ kann mit Hilfe der Wärmeleitungsgleichung

$$\begin{aligned} \frac{\partial}{\partial t} y(x,t) &= \alpha \frac{\partial^2}{\partial x^2} y(x,t) \\ y(0) = T_0, y(L) &= T_L \end{aligned}$$

beschrieben werden. Für den stationären Fall ist $\frac{\partial}{\partial t} y(x,t) = 0$. Also ist ein Randwertproblem durch

$$\begin{aligned} 0 = \frac{\partial^2}{\partial x^2} y(x,t) &= \frac{\partial^2}{\partial x^2} y(x) \\ y(0) = T_0, y(L) &= T_L \end{aligned}$$

gegeben. Die allgemeine Lösung lautet $y(x) = Ax + B$. Für $y(0) = T_0$, folgt, dass $B = T_0$ ist.
Für $y(L) = T_L$ folgt $T_L = AL + T_0$. Somit ergibt sich aus $A = \frac{T_L - T_0}{L}$ die Lösung

$$\underline{\underline{y(x,t) = \frac{T_L - T_0}{L} x + T_0}}$$

1.11 Numerik gewöhnlicher Differentialgleichungen: Einschrittverfahren

Betrachtet werden soll die numerische Lösung für das folgende AWP

$$\begin{cases} y'(x) = f(x,y) \\ y(\bar{x}_0) = \bar{y}_0 \end{cases}, \tag{1.11.1}$$

wobei f eine stetige, reellwertige Funktion in zwei Variablen x, y ist. Nichtsdestotrotz können die folgenden Betrachtungen auf den Fall, in dem y und f vektorwertig sind, verallgemeinert werden. Der Satz von Peano und der Satz Satz von Picard-Lindelöf garantieren die Existenz einer Lösung von (1.11.1) im Rechteck $R = \left\{(x,y) \in \mathbb{R}^2 | \, |x - \bar{x}_0| \leq \bar{a}, |y - \bar{y}_0| \leq \bar{b}\right\}$, und insbesondere im Intervall $\bar{J} = [\bar{x}_0 - \alpha, \bar{x_0} + \alpha]$, mit

$$\alpha = \min\left\{\bar{a}, \frac{\bar{b}}{M}\right\}, \qquad M = \max_{(x,y)\in\mathbb{R}^2} |f(x,y)|,$$

sowie Eindeutigkeit der Lösung, sofern f der Lipschitz-Bedingung

$$|f(x,y) - f(x,z)| \leq L\,|y - z| \tag{1.11.2}$$

für alle $(x,y),(x,z) \in R$ genügt. Nimmt man an, dass $f(x,\cdot) \in C^1(\left[\bar{y}_0 - \bar{b}, \bar{y}_0 + \bar{b}\right])$, wäre eine hinreichende Bedingung für (1.11.2), dass die Ableitung bezüglich y beschränkt ist, d. h. dass ein $L > 0$ mit

$$\left|\frac{\partial f}{\partial y}(x,y)\right| \leq L \; \forall (x,y) \in R$$

existiert. Der erste Schritt der numerischen Lösung besteht darin, ein Gitter mit Gitterpunkten $x_i \in J \coloneqq [x_0, x_0 + \alpha], 1 \leq 0 \leq N$ zu konstruieren. Seien nun $a, b \in \mathbb{R}, b > a$ sowie die Gitterpunkte derart, dass

$$a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$$

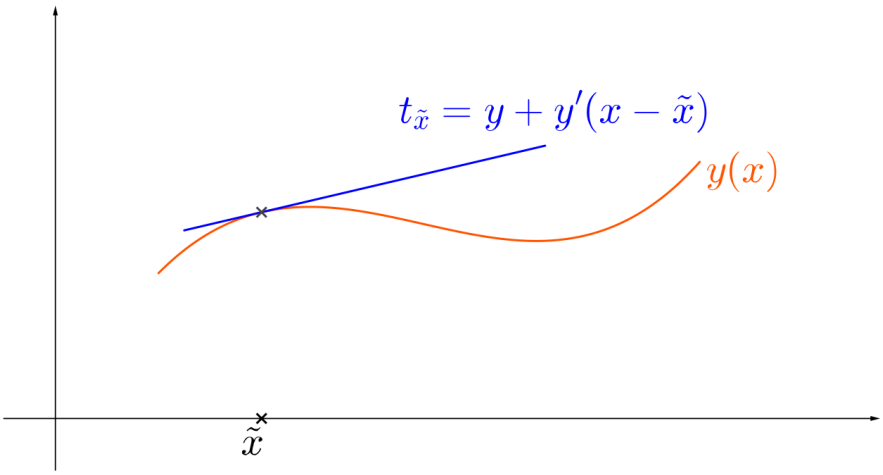
wobei $N \geq 2$ ist. Mit $h_i = x_{i+1} - x_i, i = 0, ..., N-1$ wird die Schrittweite (oder Gitterweite) bezeichnet.

Das Problem (1.11.1) numerisch zu lösen, bedeutet nun, eine Gitterfunktion mit Werten $y_i, i = 0, ..., N$ zu finden, wobei y_i eine numerische Näherung der exakten Lösung an der Stelle x_i ist, d. h. $y_i \approx y(x_i)$. Um diese zu bestimmen, soll das folgende Einschrittverfahren betrachtet werden:

$$y_{i+1} = y_i + h_i \phi(x_i, y_i; h_i), \; i = 0, ..., N-1 \tag{1.11.3}$$

mit Anfangsbedingung $y_0 = \bar{y}_0$ (ϕ stetige Funktion).

Um dies zu motivieren, sei nun das Gitter uniform, d. h. $h_i = h, i = 0, ..., N-1$ mit $h = \frac{b-a}{N}$. (Insbesondere gilt dann $x_i = a + ih$). Für festes $\tilde{x}$ wird durch $t_{\tilde{x}}(x) = y(\tilde{x}) + y'(\tilde{x}) \cdot (x - \tilde{x}) = y(\tilde{x}) + f(\tilde{x}, y(\tilde{x})) \cdot (x - \tilde{x})$ eine lineare Approximation (die Tangente) an die Lösung in $\tilde{x}$ definiert (s. Abbildung unten)



Andererseits kann das durch h dividierte Inkrement $y_{i+1} - y_i$ als $O(h)$-Approximation an $y'(x_i)$ interpretiert werden. Immerhin gilt $\frac{y(x_{i+1})-y(x_i)}{h} = y'(x_i) + O(h)$. (1.11.3) lässt sich auch durch folgenden Umstand anregen:

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y(x))\mathrm{d}x$$

Wendet man nun verschiedene Quadraturformeln zur Berechnung des Integral an, so erhält man verschiedene ϕ. Einige wichtige Quadraturformeln sind:

$$\left.\begin{array}{ll}
1. & \displaystyle\int_{x_i}^{x_{i+1}} f(x, y(x))\mathrm{d}x \approx h f(x_i, y_i) \\
2. & \displaystyle\int_{x_i}^{x_{i+1}} f(x, y(x))\mathrm{d}x \approx h f(x_{i+1}, y_{i+1}) \\
3. & \displaystyle\int_{x_i}^{x_{i+1}} f(x, y(x))\mathrm{d}x \approx \frac{h}{2}\left(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right) \text{ (Trapezregel)} \\
4. & \displaystyle\int_{x_i}^{x_{i+1}} f(x, y(x))\mathrm{d}x \approx h f\left(\frac{x_i + x_{i+1}}{2}, \frac{y_i + y_{i+1}}{2}\right) \text{ (Mittelpunktsregel)}
\end{array}\right\}$$

Diese Quadraturformeln beschreiben somit folgende Einschrittverfahren:

$$\left.\begin{array}{ll}
1. & \phi(x_i, y_i; h) = f(x_i, y_i) \rightarrow \text{ expliziter Euler} \\
2. & \phi(x_i, y_i; h) = f(x_i + h, y_i + h\phi(x_i, y_i; h)) \rightarrow \text{ impliziter Euler} \\
3. & \phi(x_i, y_i; h) = \frac{1}{2}\left(f(x_i, y_i) + f(x_i + h, y_i + h\phi(x_i, y_i; h))\right) \rightarrow \text{ Crank-Nicolson} \\
4. & \phi(x_i, y_i; h) = f(x_i + \frac{h}{2}, y_i + \frac{h}{2}\phi(x_i, y_i; h)) \rightarrow \text{ Mittelpunktsverfahren}
\end{array}\right\} \quad (1.11.4)$$

Im Folgenden soll nun untersucht werden, wie gut die Approximationen von Einschrittverfahren an die Lösung von (1.11.1) sind. Zu diesem Zweck sei der ***lokale Fehler*** (engl. solution error) als die durch

$$e_i = y(x_i) - y_i, \; i = 0, ..., N \quad (1.11.5)$$

gegebene Gitterfunktion definiert. Ferner bezeichne

$$T_i = \frac{y(x_{i+1}) - y(x_i)}{h} - \phi(x_i, y(x_i); h) \quad (1.11.6)$$

den ***Diskretisierungsfehler*** (engl. truncation error).

Definition 1.11.1:
Das Einschrittverfahren (1.11.3) ist ***konsistent***, wenn für alle $\varepsilon > 0$ eine Schrittweite $h(\varepsilon) > 0$ existiert mit $|T_i| < \varepsilon$ für alle $(x_i, y(x_i)) \in R, 0 \leq i \leq N-1$ und alle Schrittweiten $0 < h < h(\varepsilon)$.

Sei nun $x \in J$ dergestalt, dass für hinreichend großes N_0 ein i mit $x_i = x$ existiert. Für alle $N = 2^k \cdot N_0$ existiert dann ebenfalls ein i, sodass $x = x_i$ ist. Folglich kann man N so gegen unendlich laufen lassen, dass $x = x_i$ fixiert ist. In diesem Falle gilt notwendigerweise $i \to \infty, h \to 0$ sowie

$$\lim_{i\to\infty} T_i = y'(x) - \phi(x, y(x); 0).$$

Andererseits ist $y'(x) = f(x, y(x))$. Somit ist das Verfahren genau dann konsistent, wenn

$$\phi(x, y(x); 0) = f(x, y(x))$$

gilt.
Für endliche Schrittweite h ist der Diskretisierungsfehler im Allgemeinen ungleich 0, wie das folgende Beispiel zeigt. Hierzu wird der Fall (1.11.4).1 betrachtet, wobei f hinreichend glatt sei. Taylorentwicklung ergibt:

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + hy'(x_i) + \frac{1}{2}h^2 y''(x_i) + ... \\ &= y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2}\left(\frac{\partial f}{\partial x}(x_i, y(x_i)) + \frac{\partial f}{\partial y}(x_i, y(x_i)) \cdot f(x_i, y(x_i))\right) + ... \end{aligned}.$$

Somit gilt

$$T_i = \frac{h}{2}\left(\frac{\partial f}{\partial x}(x_i, y(x_i)) + \frac{\partial f}{\partial y}(x_i, y(x_i)) \cdot f(x_i, y(x_i))\right) + O(h^2),$$

was auf die folgende Definition führt:

Definition 1.11.2:
Das Einschrittverfahren (1.11.3) hat ***Konsistenzordnung p***, wenn p die größte positive Zahl ist, sodass für jede hinreichend Glatte Lösungskurve $(x, y(x))$ in R des Problems (1.11.1) Konstanten K und h_0 existieren mit

$$|T_i| \leq Kh^p \qquad \forall 0 < h < h_0, i = 0, ..., N-1 \tag{1.11.7}$$

Nun soll ein ein Theorem bewiesen werden, welches den Zusammenhang zwischen Lösungsfehler (1.11.5) und Abschneidefehler (1.11.6) beschreibt.
Hierzu muss jedoch angenommen werden, dass das Einschrittverfahren (1.11.3) Lipschitz-stetig in der zweiten Komponenten ist, d. h. es gibt eine Lipschitz-Konstante $L_\phi \in \mathbb{R}^+$ mit

$$|\phi(x, y; h) - \phi(x, z; h)| \leq L_\phi |y - z| \tag{1.11.8}$$

für alle $(x, y), (x, z) \in R$ und alle Schrittweiten $h > 0$.

Satz 1.11.3:
Das Einschrittverfahren (1.11.3) sei stetig in allen Komponenten und erfülle die Lipschitz-Bedingung aus (1.11.8). Dann gilt

$$|e_i| \leq \frac{T}{L_\phi}\left(e^{L_\phi |x_i - x_0|} - 1\right) \qquad \forall i = 0, ..., N-1, \tag{1.11.9}$$

wobei $T = \max\limits_{0\leq i\leq N-1} |T_i|$ gilt.

Beweis: Schreibt man (1.11.6) um zu

$$y(x_{i+1}) = y(x_i) + h\phi(x_i, y(x_i); h) + hT_i$$

und subtrahiert (1.11.3), erhält man

$$e_{i+1} = e_i + h \cdot (\phi(x_i, y(x_i); h) - \phi(x_i, y_i; h)) + hT_i.$$

Aus der Lipschitz-Bedingung folgt somit

$$|e_{i+1}| \leq |e_i| + hL_\phi |e_i| + h |T_i| = (1 + hL_\phi) |e_i| + h |T_i|$$

Für $i \in \{0, 1\}$ ergibt sich $|e_0| = 0, |e_1| \leq hT$. Anschließend erhält man induktiv $|e_{i+1}| \leq hT \sum_{j=0}^{i} (1 + hL_\phi)^j$. Nutzt man $\frac{q^{i+1}-q}{q-1} = 1 + q + ... + q^i$ aus, wobei q im vorliegenden Fall $1 + hL_\phi$ entspricht, führt dies auf $|e_{i+1}| \leq \frac{T}{L_\phi} \left((1 + hL_\phi)^{i+1} - 1 \right), i = 0, ..., N - 1$. Wegen $1 + hL_\phi \leq \exp(hL_\phi)$ sowie $x_{i+1} = ih + a = ih + x_0$ folgt (1.11.9). □

Es bleibt zu klären, ob sich mit einer Verfeinerung des Gitters auch die numerische Lösung $(y_i)_0^N$ gegen die exakte Lösung von (1.11.1) konvergiert, d. h. für jedes feste $x \in J$

$$\lim_{i \to \infty} y_i = y(x) \qquad \text{für} \quad x_i = ih + a \to x, \tag{1.11.10}$$

wenn $i \to \infty$.

Satz 1.11.4:
Angenommen, das Anfangswertproblem (1.11.1) genügt den Bedingungen von Satz 1.5.8, das Einschrittverfahren (1.11.3) genügt der Lipschitz-Bedingung (1.11.8) und für alle $0 < h \leq h_0$ liegt die numerische Lösung ebenfalls in R. Dann konvergiert die numerische Lösung $(y_i)_{i=0}^N$ für steigendes N mit $\frac{1}{N} = h < h_0$ gegen die Lösung des AWPs.

Beweis: Sei N hinreichend groß, sodass $h \leq h_0$. Aus Theorem 1.11.3 folgt

$$|y(x_i) - y_i| \leq \frac{e^{L_\phi(b-a)} - 1}{L_\phi} \cdot \max_{0 \leq k \leq N} |T_k|, \quad i = 0, ..., N \tag{1.11.11}$$

Schreibt man nun

$$T_i = \frac{y(x_{i+1}) - y(x_i)}{h} - f(x_i, y(x_i)) + \phi(x_i, y(x_i); 0) - \phi(x_i, y(x_i); h)$$

so existiert nach dem Mittelwertsatz ein $\xi_i \in [x_i, x_{i+1}]$ mit $\frac{y(x_{i+1})-y(x_i)}{h} = y'(\xi_i) = f(\xi_i, y(\xi_i))$. Folglich:

$$T_i = y'(\xi_i) - y'(x_i) + \phi(x_i, y(x_i); 0) - \phi(x_i, y(x_i); h)$$

Da y' stetig ist, existiert $h_1(\varepsilon)$ mit

$$|y'(\xi_i) - y'(x_i)| \leq \frac{1}{2}\varepsilon \quad \forall 0 < h < h_1(\varepsilon)$$

Da ferner ϕ stetig ist, existiert $h_2(\varepsilon)$ mit

$$|\phi(x_i, y(x_i); 0) - \phi(x_i, y(x_i); h)| \leq \frac{1}{2}\varepsilon \quad \forall 0 < h < h_2(\varepsilon)$$

Somit gilt ür alle $0 < h < h(\varepsilon) \coloneqq \min\{h_1(\varepsilon), h_2(\varepsilon)\}$:

$$|T_i| \overset{\Delta-\text{Ungl.}}{\leq} |y'(\xi_i) - y'(x_i)| + |\phi(x_i, y(x_i); 0) - \phi(x_i, y(x_i); h)| \leq \varepsilon$$

Aus (1.11.11) folgt nun

$$|y(x) - y_i| \leq |y(x) - y(x_i)| + |y(x_i) - y_i| \leq$$
$$\leq |y(x) - y(x_i)| + \varepsilon \frac{e^{L_\phi(b-a)} - 1}{L_\phi}$$

Gilt nun $x_i \to x$ für $h \to 0, i \to \infty$, (vgl. (1.11.10)), so garantiert die Stetigkeit von y nun für jedes $\varepsilon' > 0$ die Existenz eines $h_3(\varepsilon')$ mit $|y(x) - y(x_i)| < \varepsilon'$ für alle $0 < h < h_3(\varepsilon')$. □

Nun lässt sich auch der globale Fehler der Lösung untersuchen:

Satz 1.11.5:
Es seien die Voraussetzungen von Satz 1.11.3 erfüllt und das Einschrittverfahren (1.11.3) habe Konsistenzordnung p. Dann hat das Verfahren ***Konvergenzordnung*** p, d. h. es existieren Konstanten $c, h_0 > 0$, sodass der ***globale Fehler*** $e = \max\limits_{0 \leq i \leq N} |e_i|$ die Ungleichung

$$e \leq ch^p \quad 0 < h < h_0$$

erfüllt.

Beweis: Betrachtet man (1.11.7) und (1.11.11), so existieren Konstanten $\tilde{c}, h_0 > 0$, sodass

$$|y(x_i) - y_i| \leq \frac{e^{L_\phi(b-a)} - 1}{L_\phi} \cdot \tilde{c}h^p \quad \forall 0 < h < h_0$$

für alle $0 \leq i \leq N$ und somit

$$\max_{0 \leq i \leq N} |y(x_i) - y_i| \leq ch^p, \quad 0 < h < h_0$$

gilt. □

Betrachtet man den ersten Fall in (1.11.4), so gilt $L_\phi = L$, wobei L die Lipschitz-Konstante von f ist. Diese Gleichheit gilt jedoch im Allgemeinen nicht. Trotzdem ist es naheliegend, einen funktionalen Zusammenhang zwischen diesen Konstanten zu vermuten. Dieser soll nun für das Crank-Nicolson-Verfahren hergeleitet werden:

$$y_{i+1} = y_i + \frac{h}{2}\left(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right)$$

In diesem Falle gilt

$$\phi(x_i, y_i; h) = \frac{1}{2}f(x_i, y_i) + \frac{1}{2}f(x_i + h, y_i + h\phi(x_i, y_i; h))$$

Es gilt:

$$\begin{aligned}
|\phi(x_i, y; h) - \phi(x_i, z; h)| = \Big| &\frac{1}{2}f(x_i, y) + \frac{1}{2}f(x_i + h, y + h\phi(x_i, y; h)) \\
&-\frac{1}{2}f(x_i, z) - \frac{1}{2}f(x_i + h, z + h\phi(x_i, z; h))\Big| \leq \\
\leq \frac{1}{2}&\left|f(x_i, y) - f(x_i, z)\right| + \frac{1}{2}\left|f(x_i + h, y + h\phi(x_i, y; h))\right. \\
&\left. - f(x_i + h, z + h\phi(x_i, z; h))\right| \leq \\
\leq \frac{1}{2}&L\left|y - z\right| + \frac{1}{2}L\left|y + h\phi(x_i, y; h)) - z - h\phi(x_i, z; h)\right| \leq \\
\leq L&\left|y - z\right| + \frac{h}{2}L\left|\phi(x_i, y; h) - \phi(x_i, z; h)\right|
\end{aligned}$$

Daraus ergibt sich

$$\left(1-\frac{h}{2}L\right)|\phi(x_i,y;h)-\phi(x_i,z;h)|\leq L\,|y-z|\,,$$

weshalb man

$$L_\phi=\frac{L}{1-\frac{1}{2}hL} \tag{1.11.12}$$

wählen kann, vorausgesetzt, $\frac{1}{2}hL<1$. Weitere Berechnungen zeigen, dass der Diskretisierungsfehler der Trapezregel der Ungleichung

$$|T_i|\leq\frac{1}{12}h^2M,\quad M=\max_{x\in J}|y'''(x)|$$

genügt. Das Verfahren hat also Konsistenzordnung 2 und für h hinreichend klein ist dies auch die Konvergenzordnung.

Im zweiten Fall von (1.11.4) (impliziter Euler) gilt ebenfalls $L_\phi\neq L$. ϕ ist hier gegeben durch

$$\phi(x_i,y_i;h)=f(x_i+h,y_i+h\phi(x_i,y_i;h))$$

Analoge Vorgehensweise wie im vorherigen Fall liefert

$$\begin{aligned}|\phi(x_i,y;h)-\phi(x_i,z;h)|&=|f(x_i+h,y+h\phi(x_i,y;h))-f(x_i+h,z+h\phi(x_i,z;h))|\leq\\&\leq L\,|y-z+h\,(\phi(x_i,y;h)-\phi(x_i,z;h))|\leq\\&\leq L\,|y-z|+hL\,|\phi(x_i,y;h)-\phi(x_i,z;h)|\,,\end{aligned}$$

woraus sich

$$L_\phi=\frac{L}{1-hL}$$

ergibt (sofern $hL<1$). Ferner erhält man durch Taylorentwicklung der exakten Lösung,

$$y(x_i)=y_(x_{i+1})-hy'(x_{i+1})+\frac{1}{2}h^2y''(\xi_{i+1}),\quad \xi_{i+1}\in[x_i,x_{i+1}]\,,$$

für den Diskretisierungsfehler die Abschätzung

$$|T_i|\leq\frac{h}{2}M,\qquad M=\max_{x\in J}|y''(x)|$$

Analog findet man, dass das explizite Eulerverfahren ebenfalls Konsistenzordnung 1 hat.

Nun soll noch das Mittelpunktsverfahren auf diese Weise untersucht werden. Hier gilt (s. (1.11.4))

$$\phi(x_i,y_i;h)=f(x_i+\frac{h}{2},y_i+\frac{h}{2}\phi(x_i,y_i;h)). \tag{1.11.13}$$

Man erhält aus

$$\begin{aligned}|\phi(x_i,y;h)-\phi(x_i,z;h)|&=\left|f(x_i+\frac{h}{2},y+\frac{h}{2}\phi(x_i,y;h))-f(x_i+\frac{h}{2},z+\frac{h}{2}\phi(x_i,z;h))\right|\leq\\&\leq L\left|y-z+\frac{h}{2}\,(\phi(x_i,y;h)-\phi(x_i,z;h))\right|\leq\\&\leq L\,|y-z|+\frac{h}{2}L\,|\phi(x_i,y;h)-\phi(x_i,z;h)|\end{aligned}$$

als geeignete Lipschitz-Konstante (unter den selben Voraussetzungen wie in (1.11.12))

$$L_\phi=\frac{L}{1-\frac{1}{2}hL}.$$

Der Diskretisierungsfehler wird erneut durch Taylorentwicklung (nun in $x_i + \frac{h}{2}$) abgeschätzt. Dies führt auf

$$|T_i| \leq \frac{1}{24}h^2 L M_3 + \frac{1}{4}h^2 M_2,$$

wobei

$$M_2 = \max_{x \in J} |y''(x)| \qquad M_3 = \max_{x \in J} |y'''(x)|$$

ist.

Zur Erinnerung: L_ϕ war die Lipschitz-Konstante aus (1.11.8). Wie bereits gesehen, lies sich mit der Lipschitz-Stetigkeit Abschätzung (1.11.9) zeigen. Diese Abschätzung erlaubt jedoch, dass der Fehler mit steigendem $|x_i - x_0|$ exponentiell wächst. Für Probleme mit asymptotisch stabilen Lösungen, die für $x \to \infty$ gegen eine Ruhelage(-lösung) konvergieren, ist diese daher wenig befriedigend. Ist die DGL in (1.11.1) autonom und ferner y skalarwertig , so ist eine Ruhelage $\bar{y}$ durch $f(\bar{y}) = 0$ charakterisiert. Wie in Kapitel 1.8 gesehen, ist das Modell $y' = f(y)$ asymptotisch stabil genau dann wenn

$$\lambda = \frac{\partial f}{\partial y}(\bar{y}) < 0 \tag{1.11.14}$$

gilt. Im allgemeinen Fall, wenn f vektorwertig ist, lautet die Bedingung, dass die die Reilteile der Eigenwerte der Jacobi-Matrix von f bzgl. y negativ sind.

Hat man daher nun ein asymptotisch stabiles System gegeben, sollte man erwarten, dass der Fehler nicht wächst (zumindest nicht exponentiell). Für den expliziten Euler lässt sich leicht zeigen, dass die Bedingung (1.11.14) sich auf ϕ übertragt:

$$\frac{\partial \phi}{\partial y}(\bar{y}; h) = \frac{\partial f}{\partial y}(\bar{y}) < 0.$$

In den anderen Fällen gilt analoges für hinreichend kleine h. Im Folgenden soll nun die Abschätzung (1.11.9) aus Satz 1.11.3 verbessert werden unter der Voraussetzung. dass $\frac{\partial \phi}{\partial y}(y; h) < 0$ für alle y in einer Nachbarschaft $\bar{Y}$ von $\bar{y}$ gilt. Im Beweis von Theorem 1.11.3 wurde folgende Gleichheit verwendet:

$$e_{i+1} = e_i + h \cdot (\phi(y(x_i); h) - \phi(y_i; h)) + hT_i$$

Angenommen, ϕ ist stetig differenzierbar in der zweiten Komponente. Dann existiert ein ξ_i zwischen $y(x_i)$ und y_i, sodass

$$e_{i+1} = e_i + h\frac{\partial}{\partial y}\phi(\xi_i; h) + hT_i$$

gilt. Seien nun $y_i, y(x_i) \in \bar{Y}$ und $c = \sup_{\xi \in \bar{Y}} \left| \frac{\partial}{\partial y}\phi(\xi; h) \right| < \infty$. Unter der Annahme $0 < 1 - ch < 1$ ergibt sich nun die Gleichung

$$|e_{i+1}| = |e_i| - ch\,|e_i| + h\,|T_i|$$

Analog wie im Beweis von Theorem 1.11.3 ergibt sich

$$|e_{i+1}| \leq \frac{T}{c}\left(1 - (1 - ch)^i\right),$$

was einem viel geringeren Fehlerzuwachs entspricht. Hieraus wird ersichtlich, dass asymptotisch stabile Systeme weiterführender Untersuchung bedürfen. Insbesondere ist zu klären, unter welchen Bedingungen ein Einschrittverfahren das Verhalten von stabilen Lösungen reproduziert. Zu diesem Zweck sei das Testproblem

$$y' = \lambda y, \qquad \lambda \in \mathbb{C}, \operatorname{Re}\lambda \leq 0 \tag{1.11.15}$$

gegeben. Der Fokus liegt hierbei auf der Untersuchung der absoluten Stabilität:

Definition 1.11.6:
Ein Einschrittverfahren ist ***A-stabil***, wenn für alle Schrittweiten h die generierten Näherungslösungen für (1.11.15) (betragsmäßig) monoton fallen, d. h. $|y_{i+1}| \leq |y_i|$ für alle $i \geq 0$.

Da (1.11.15) linear ist, hat die von einem Einschrittverahren generierte Approximationsfolge $(y_i)_i$ die Gestalt

$$y_{i+1} = R(z)y_i,$$

wobei z eine Funktion in h und λ ist. (R wird auch ***Stabilitätsfunktion*** genannt). Im Falle des expliziten Eulers gilt

$$y_{i+1} = y_i + h\lambda y_i = (1 + h\lambda)y_i$$

Daher ist $R(z) = 1 + z, z = h\lambda$. Da jedoch - vorausgesetzt $\lambda \neq 0$ - $h > 0$ stets so gewählt werden kann, dass $|R(z)| = |1 + h\lambda| > 1$ gilt, kann dies in einer (betragsmäßig) streng monoton steigenden Folge der approximierten Lösungswerte resultieren: Das explizite Eulerverfahren ist nicht A-stabil.

Definition 1.11.7:
Das ***Stabilitätsgebiet*** eines Einschrittverfahrens ist die Menge

$$S = \{z \in \mathbb{C}|\ |R(z)| \leq 1\}$$

Das Stabilitätsgebiet des expliziten Eulerverfahrens ist in Abbildung 1.11.1 dargestellt. Für das implizite gilt hingegen

$$y_{i+1} = y_i + h\lambda y_{i+1} \Rightarrow (1 - h\lambda)\, y_{i+1} = y_i$$

Daher gilt

$$R(z) = \frac{1}{1 - z}, \quad z = h\lambda$$

Eine Darstellung des entsprechenden Stabilitätsgebietes befindet sich in Abbildung 1.11.2.

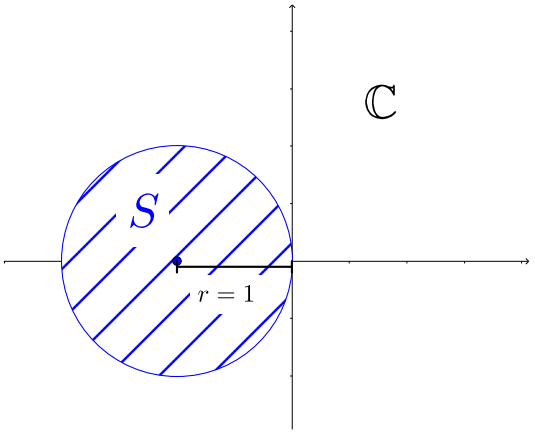


Abbildung 1.11.1

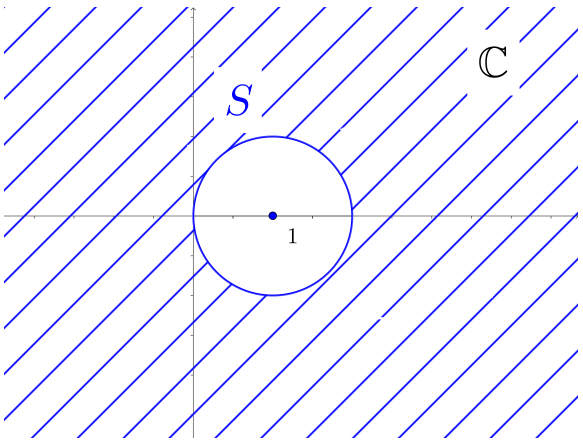


Abbildung 1.11.2

Für das Crank-Nicolson-Verfahren gilt

$$y_{i+1} = y_i + \frac{h\lambda}{2}\left(y_i + y_{i+1}\right),$$

was zu

$$R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}, \qquad z = h\lambda$$

sowie

$$|R(z)| \leq 1 \quad \forall z \in \{w \in \mathbb{C} | \mathrm{Re}\, w \leq 0\}$$

führt.
Exakt die selben Resultate lassen sich für das (implizite) Mittelpunktsverfahren erzielen. Beide Verfahren sind somit A-stabil.

Zum Abschluss des Kapitels sollen einige geometrische Eigenschaften von Differentialgleichungen und verwandten numerischen Themen untersucht werden. Sei hierzu eine vektorwerti-

ge, autonome Differentialgleichung

$$y' = f(y) \tag{1.11.16}$$

gegeben.

Definition 1.11.8:
Ein ***erstes Integral*** (oder ***Invariante***) von (1.11.16) ist eine stetig differenzierbare Funktion $I = I(y)$, welche für jede Lösung von (1.11.16) konstant ist, d. h.

$$\frac{\mathrm{d}}{\mathrm{d}x} I(y(x)) = I'(y(x)) f(y(x)) = 0$$

erfüllt.

Eine spezielle Klasse von Modellen, welche ein erstes Integral haben, sind (vgl. Kapitel 2.7) Hamilton-Systeme, gegeben durch

$$p' = -\frac{\partial}{\partial q} H(p,q), \quad q' = \frac{\partial}{\partial p} H(p,q),$$

wobei H die entsprechende Hamiltonfunktion ist, welche im Folgenden zweimal stetig differenzierbar sei. Es gilt

$$\frac{\mathrm{d}}{\mathrm{d}x} H(p(x), q(x)) = \frac{\partial H}{\partial p} p' + \frac{\partial H}{\partial q} q' = 0$$

Diese und ähnliche Modelle haben eine weitere Invariante: Das Volumen.
Zu diesem Zweck soll der Fluss ϕ_x betrachtet werden. Zur Erinnerung: Für festen Anfangspunkt x_0 ist

$$\psi_x(y_0) = y(x; x_0, y_0),$$

wobei $y(x; x_0, y_0)$ die Lösung des APWs

$$\begin{cases} y' = f(y) \\ y(x_0) = y_0 \end{cases}$$

ist. Ist die Anfangsbedingung y_0 aus einer Menge Y, so sein nun für x die folgende Menge definiert:

$$\psi_x(Y) \coloneqq \{y | y = y(x; x_0, y_0), y_0 \in Y\}$$

Das Volumen von Y ist durch

$$\mathrm{Vol}\,(Y) = \int_Y \mathrm{d}y$$

gegeben.

Definition 1.11.9:
Der Fluss das ψ_x (bzgl. (1.11.16)) ist ***volumenerhaltend***, wenn

$$\mathrm{Vol}\,(\phi_x(Y)) = \mathrm{Vol}\,(Y) \quad \forall x \geq x_0 \tag{1.11.17}$$

gilt.

Es gilt das folgende Resultat:

$$\begin{aligned} \mathrm{Vol}\,(\psi_x(Y)) &= \int_{\psi_x(Y)} \mathrm{d}y = \int_Y \left| \det\left(\frac{\partial y}{\partial y_0}(x; x_0, y_0) \right) \right| \mathrm{d}y_0 = \\ &= \int_Y \exp\left(\int_{x_0}^{x} \mathrm{tr}\, \left(\frac{\partial f}{\partial y}(y(s; x_0, y_0)) \right) \mathrm{d}s \right) \mathrm{d}y_0 \end{aligned},$$

wobei $\frac{\partial f}{\partial y}$ die Jacobi-Matrix von f bezeichnet. Gilt nun tr $\left(\frac{\partial f}{\partial y}(y)\right) = 0$, so ist (1.11.17) erfüllt. Im Falle eines Hamilton-Systems entspricht die Jacobi-Matrix der Matrix

$$\begin{bmatrix} -H_{pq} & -H_{qq} \\ H_{pp} & H_{pq} \end{bmatrix},$$

weshalb die Spur 0 ist.

Ein Spezialfall von Volumen bilden Parallelogramme, welche von zwei Vektoren aufgespannt werden. Zu dem Hamilton-System (1.11.16) seien einmal die Anfangsbedingungen $p(0) = \xi^p$ und $q(0) = \xi^q$ gegeben, zum anderen die Bedingungen $p(0) = \eta^0, q(0) = \eta^q$. Die Vektoren $\xi = \begin{pmatrix} \xi^p \\ \xi^q \end{pmatrix}$ und $\eta = \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}$ definieren ein Parallelogramm im (p,q)-Raum (vgl. Abbildung 1.11.3).

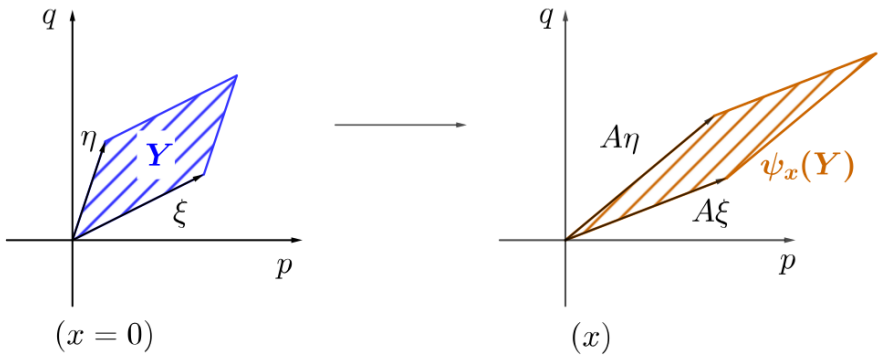


Abbildung 1.11.3

Das (orientierte) Volumen (besser gesagt, die Fläche) ist in diesem Falle durch

$$V_o = \det \begin{pmatrix} \xi^p & \eta^p \\ \xi^q & \eta^q \end{pmatrix} = \xi^p \eta^q - \xi^q \eta^p$$

gegeben. Für eine symplektische Abbildung A gilt dann

$$\omega(A\xi, A\eta) = \omega(\xi, \eta) = V_0, \tag{1.11.18}$$

was bedeutet, dass A die Fläche des Parallelogramms erhält. Dies gilt auch für den Fluss φ_x: Ist dieser symplektisch, d. h. gilt $\varphi_x'^T J \varphi_x' = J$, wie z. B. im Falle von Hamilton-Systemen, so ist er auch Volumenerhaltend. Es stellt sich die Frage, inwiefern Einschrittverfahren dies widerspiegeln. Sei hierfür y' lokal hamiltonisch, es gelte also $y' = f(y) = J^{-1}H'(y)^T$ in einer Umgebung von y_0 sowie einer geeigneten, hinreichend glatten Funktion H. Für das implizite Mittelpunktsverfahren gilt:

$$y_{i+1} = y_i + hJ^{-1}H'\left(\frac{y_{i+1} + y_i}{2}\right)^T$$

Sei nun $\nabla H = H'(\frac{y_{i+1}+y_i}{2})^T$ sowie $y_{i+1} = y_{i+1}(y_i)$. Dann ergibt Differentiation ($\nabla^2 H$ Hessematrix):

$$\frac{\partial y_{i+1}}{\partial y_i} = I + \frac{h}{2}J^{-1}\nabla^2 H \cdot \left(\frac{\partial y_{i+1}}{\partial y_i} + I\right) = \left(I - J^{-1}\frac{h}{2}\nabla^2 H\right)^{-1}(I + J^{-1}\frac{h}{2}\nabla^2 H)$$

Soll das Mittelpunktsverfahren daher die Symplektizität für die numerische Lösung erhalten, genügt es, dass $\frac{\partial y_{i+1}}{\partial y_i}^T J \frac{\partial y_{i+1}}{\partial y_i} = J$ oder - gemäß der letzten Rechnung -

$$\left(I + \frac{h}{2}J^{-1}\nabla^2 H\right) J \left(I + \frac{h}{2}J^{-1}\nabla^2 H\right)^T = \left(I - \frac{h}{2}J^{-1}\nabla^2 H\right) J \left(I - J^{-1}\frac{h}{2}\nabla^2 H\right)^T$$

gilt. Beachtet man, dass $(J^{-1})^T = -J^{-1} = J$ sowie $\nabla^2 H = \nabla^2 H^T$ gelten, ergibt sich mit $(I + \frac{h}{2} J^{-1}\nabla^2 H)^T = I + \frac{h}{2}\nabla^2 H J$ für die linke Seite

$$\left(I + \frac{h}{2}J^{-1}\nabla^2 H\right) J \left(I + \frac{h}{2}J^{-1}\nabla^2 H\right)^T = \left(I - \frac{h}{2}J\nabla^2 H\right) J \left(I + \frac{h}{2}\nabla^2 H J\right) =$$
$$= J - \frac{h}{2}J\nabla^2 H J + \frac{h}{2}J\nabla H^2 J - \frac{h^2}{4}J\nabla^2 H J \nabla^2 H J$$

und analog für die rechte Seite mit $(I - \frac{h}{2} J^{-1}\nabla^2 H)^T = I - \frac{h}{2}\nabla^2 H J$

$$\left(I - \frac{h}{2}J^{-1}\nabla^2 H\right) J \left(I - \frac{h}{2}J^{-1}\nabla^2 H\right)^T = \left(I + \frac{h}{2}J\nabla^2 H\right) J \left(I - \frac{h}{2}\nabla^2 H J\right) =$$
$$= J + \frac{h}{2}J\nabla^2 H J - \frac{h}{2}J\nabla H^2 J - \frac{h^2}{4}J\nabla^2 H J \nabla^2 H J$$

Da beide Seiten den selben Wert haben, ist das Mittelpunktsverfahren somit symplektisch.

Hat $y' = f(y)$ ein quadratisches erstes Integral, existiert also ein erstes Integral der Form $I(y) = y^T C y$ mit $y^T C f(y) = 0, C^T = C$, so erhält das Mittelpunktsverfahren auch diese Quadratische Invariante. Multipliziert man beide Seiten von

$$\frac{y_{i+1} - y_i}{h} = f\left(\frac{y_{i+1} + y_i}{2}\right)$$

mit $\frac{1}{2}\left(\frac{y_{i+1}+y_i}{2}\right)^T C$, so ergibt sich

$$\frac{1}{2h}\left(I(y_{i+1}) - I(y_i)\right) = \left(\frac{y_{i+1} + y_i}{2}\right)^T C f\left(\frac{y_{i+1} + y_i}{2}\right) = 0$$

Analoges gilt für lineare Invarianten $I(y) = y^T d$, wobei $d\ d^T f(y) = 0$ erfüllt:

$$d^T y_{i+1} = d^T y_i + h d^T f(\frac{y_{i+1} + y_i}{2}) = d^T y_i$$

Letzteres gilt für alle oben vorgestellten Verfahren. Jedoch sind weder das explizite, noch das implizite Euler- oder das Crank-Nicolson-Verfahren symplektisch.

Abschließende Bemerkungen

- Obiges Prinzip, die Lösung mittels Quadraturformeln zu approximieren, führt auf die sogenannten ***Runge-Kutta-Verfahren***. Die allgemeine Berechnungsvorschrift lautet lautet:

  $$y_{n+1} = y_n + \sum_{i=1}^{s} b_i K_i$$

  mit

  $$K_i = F(x_n + c_i h, y_n + h \cdot \sum_{j=1}^{s} a_{ij} K_j)$$

  s heißt ***Stufenzahl***, $b_i, i = 1, ..., s$ sind die Gewichte und $c_i, i = 1...s$ die ***Knoten*** des Verfahrens. Die Koeffizienten $b_i, c_i, a_{ij}, i, j = 1, ..., s$ werden hierbei in sogenannten ***Butcher-Tableaus*** zusammengefasst:

  $$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

Gilt hierbei $a_{ij} = 0$ für alle $j \geq i$, so ist das Runge-Kutta-Verfahren (RKV) ***explizit***, andernfalls ***implizit***. Da bei expliziten Verfahren für die Berechnung der K_i nur bereits berechnete K_j $(j < i)$ relevant sind, benötigen diese weniger Zeit, während bei impliziten Verfahren ein (im Allgemeinen nichtlineares) Gleichungssystem für ebendiese gelöst werden muss. Jedoch sind die Abweichungen der implizit berechneten Näherungen von der exakten Lösung im Allgemeinen (teilweise wesentlich) geringer als die explizit berechneten. Durch geschickte Wahl der Koeffizienten lässt sich hierbei die Konsistenzordnung für den jeweiligen Fall maximieren.

- Manchmal sind auch Verfahren mit Schrittweitensteuerung sinnvoll. Diese führen in jedem Schritt einen Kontrollschritt aus, d. h. berechnen eine Lösung mit einem Verfahren höherer Ordnung (aber akzeptablen Mehraufwand), um einen Fehlerschätzer zu bekommen. Erfüllt dieser eine vorgegebene Genauigkeit nicht, wird der Schritt mit kleinerer Schrittweite wiederholt, ansonsten der nächste mit einer mäßig größeren durchgeführt. Auf diese Weise wird versucht, eine vorgegebene Präzesion mit möglichst wenig Rechenaufwand zu erreichen.

- Weitere wichtige Stabilitätsklassen sind L-Stabilität und Isometrieerhaltung. Für ersteres muss $\lim\limits_{|z|\to\infty} |R(z)| = 0$ gelten, für letzteres $|R(z)| = 1$, sofern $\mathrm{Re}\,(z) = 0$. Beide Eigenschaften schließen sich gegenseitig aus. Je nach Anwendungsfall muss man sich daher (ggf.) für Isometrieerhaltung oder L-Stabilität entscheiden. Für den Fall der eindimensionalen Wärmeleitungsgleichung z. B. ist letzteres sinnvoll.

2 Modellierung

Dieses Kapitel behandelt die (näherungsweise) Beschreibung realer Situationen mittels mathematischer Modelle und orientiert sich hierbei in den ersten drei Kapiteln an [7]. Als Ausblick auf die Schwierigkeiten, die sich bei einer solchen Modellbildung auftun können, soll folgendes, einführendes Beispiel betrachtet werden:

Beispiel 2.0.1:

Ein Bauer habe 200 Kühe. Diese vermehren sich, sodass er nach einem Jahr 230 Kühe hat. Gesucht ist nun die Zeit (in Jahren), nach der er 500 Kühe besitzt. Hierfür kann man zwei Ansätze betrachten.

1. Konstanter Wachstumsfaktor: In einem Jahr ist die Kuhpopulation bzw., der Größenordnung angepasst, Kuhherde auf das r-fache Angwachsen, wobei $r = \frac{230}{200} = 1.15$ ist. Nimmt man an, dass dies jedes Jahr so ist, so erhält man für die Anzahl an Kühe, mit x bezeichnet, das Modell
$$x\left(t_{(n+1)}\right) = r \cdot x_n\left(t_n\right)$$
Hierbei bezeichnet $t_n = t_0 + n\Delta t$ den betrachteten Zeitpunkt, Δt ein Zeitinkrement, in diesem Falle ein Jahr. Induktiv ergibt sich:
$$\boxed{x\left(t_n\right) = r^n \cdot x\left(t_0\right)}$$
Setzt man $x\left(t_n\right) = 500, x\left(t_0\right) = 200$, so erhält man mittels $n \log r = \log\left(\frac{500}{2}\right)$ als Lösung der Ausgangsfrage $n \approx 6.6$ Jahre.

2. Konstanter jährlicher Zuwachs: Aufgrund der geringen Anzahl an Messwerten ist genauso legitim, den jährlichen Zuwachs als konstant anzunehmen. Im vorliegendem Falle ergibt sich als Zuwachsrate $p = \frac{230-200}{200} = 0.15$ pro Jahr. Der Wachstumsfaktor lautet $r = p \cdot \Delta t$, wobei $r = 1.15$ für $\Delta t = 1$ Jahr. Allgemein ergibt sich mit den Bezeichnungen x, t_n wie gehabt:
$$\boxed{x\left(t_n\right) = \left(1 + p \cdot t_n\right) x(t_0)}$$
Hier ist $x\left(t_0\right) = 200$, womit nach zwei Jahren folglich $r = 1 + 0.3$, gilt, entsprechend nach 7 Jahren also $r = 1 + 1.05$. Wegen $(1 + 1.05) \cdot 200 = 410 < 500$ liefert dieser Ansatz offensichtlich ein anderes Ergebnis als der erste.

Obige, diskrete Modelle lassen sich auch zu kontinuierlichen Modellen erweitern. Hierdurch werden diese unabhängig von der der Wahl des Zeitinkrements zugrunde liegenden Willkür, auch wenn im Sachzusammenhang $4,23$ Kühe interpretatorisch schwierig sind. Dazu betrachtet man den Übergang
$$\lim_{\Delta t \to 0^+} \frac{x\left(t + \Delta t\right) - x\left(t\right)}{\Delta t}, \qquad (*)$$
durch welchen sich folgende Modellmöglichkeiten ergeben:

Ansatz 1:

$$x'(t) = (*) = \rho x(t)$$

mit $\rho = \log(r)$, bzw. streng genommen $\rho = \log(r)/J$; im vorliegenden Falle also $\rho = \log(1.15) \approx 0.1398(1/J)$. Mit Kapitel 1 ergibt sich folglich:

$$\boxed{x(t) = x(t_0) \cdot e^{\rho(t-t_0)}}$$

Ansatz 2:

$$x'(t) = (*) = p \cdot x(t_0)$$

Diesmal ergibt sich durch Integration:

$$\leftrightarrow \qquad \boxed{x(t) = (1 + p \cdot (t - t_0))\, x(t_0)}$$

2.1 Dimensionsanalyse und Skalierung

Wie im Einführungsbeispiel gesehen, liegen bei einem Modell häufig ***physikalische Dimensionen*** (Kraft, Zeit, Helligkeit, ...) vor. Im „Populationsmodell" des genannten Beispiels sind dies beispielsweise die Zeit und die Anzahl an Kühen. Hierdurch motiviert definiert man für jede Dimension eine Art Grundeinheit, welche ***charakteristische Größe*** genannt wird. Mit ihrer Hilfe lassen sich Modelle ***entdimensionalisieren*** und wesentlich vereinfachen.

Vereinbarung: Im Folgenden bezeichne

- $\bar{f}$ die charakteristische Größe
- $[f]$ die physikalische Dimension

der Größe f.

Beispiel (Fortsetzung zu obigem Wachstumsmodell):

Die physikalischen Dimensionen sind:

$$[t] = T \text{ (für „time")}$$

$$[x(t)] = A \text{ (für „amount")}$$

Weitere Dimensionen:

$$[x'(t)] = \frac{A}{T}$$

$$[p] = [\rho] = \frac{1}{T}$$

Mit t_0 als Startzeitpunkt ist

$$\tau = \frac{t - t_0}{\bar{t}}$$

ein geeignetes, dimensionsloses Zeitmaß. $\bar{t}$ wird später problemangepasst spezifiziert.
Als dimensionslose, abhängige Variable lässt sich

$$y = \frac{x}{\bar{x}},$$

mit $\bar{x} = x_0$ wählen.

Drückt man y als Funktion $y(\tau)$ aus, führt dies auf

$$y(\tau) = \frac{x(\bar{t} \cdot \tau + t_0)}{\bar{x}},$$

mit $\bar{x} = x_0$. (Sinnvollerweise sei $x_0 > 0$ vorausgesetzt.) Hieraus ergibt sich

$$y'(\tau) = \frac{\bar{t}}{\bar{x}} \cdot x'(t)$$

und somit gilt (wegen $x'(t) = \rho x(t)$) $\frac{\bar{x}}{\bar{t}} y'(\tau) = \rho \bar{x} y(\tau)$. Dividiert man beide Seiten durch $\bar{x}$, erhält man das AWP

$$\begin{cases} \frac{1}{\bar{t}} y'(\tau) = \rho y(\tau) \\ y(0) = 1 \end{cases}$$

Für $\bar{t} = \frac{1}{\rho}$ wird dieses besonders einfach: Die (eindeutige, vgl. Kapitel 1.4) Lösung ist die Exponentialfunktion. Will man nun alle Lösungen haben (mit Berücksichtigung von Startzeitpunkt und Anfangspopulation), so erhält man diese durch Rücksubstitution ($\tau = \frac{t-t_0}{\bar{t}} = \rho(t - t_0)$):

$$x(t) = \bar{x} \cdot y(\tau) = x_0 \cdot e^{\rho(t-t_0)}$$

Für große t wird dieses Modell unbrauchbar, da - realistisch gesehen - die Kuhherde nicht unbegrenzt wachsen kann. Um eine (durch Weideplatz, Futtermenge bedingte) Wachstumsschranke oder Kapazität x_M zu berücksichtigen, kann man den Ansatz $\rho = \rho(x)$ verfolgen, wobei $\rho(x)$ geschickterweise so gewählt wird, dass $\rho(x) > 0$ für $x \in [0, x_M)$ und $\rho(x) < 0$ für $x > x_M$ gilt. Wählt man hierfür $\rho(x) = q \cdot (x_M - x)$ mit positiver Konstante q, so erhält man die ***logistische Gleichung***

$$x'(t) = q\,(x_M - x(t))\,x(t) \tag{2.1.1}$$

bzw.

$$x'(t) = (qx_M)\,x(t) - q\,(x(t))^2$$

Löst man diese Bernoulli-DGL, ergibt sich:

$$\boxed{x\,(t) = \frac{x_M x_0}{x_0 + (x_M - x_0)\,e^{-x_M q(t-t_0)}}}$$

In diesem Falle wären geeignete charakteristische Größen $\bar{x} = x_M$ und $\bar{t} = \frac{1}{q}$. Die Ruhelagen des Modells sind $x \equiv 0$ (instabil) und $x \equiv x_M$ (asymptotisch stabil).

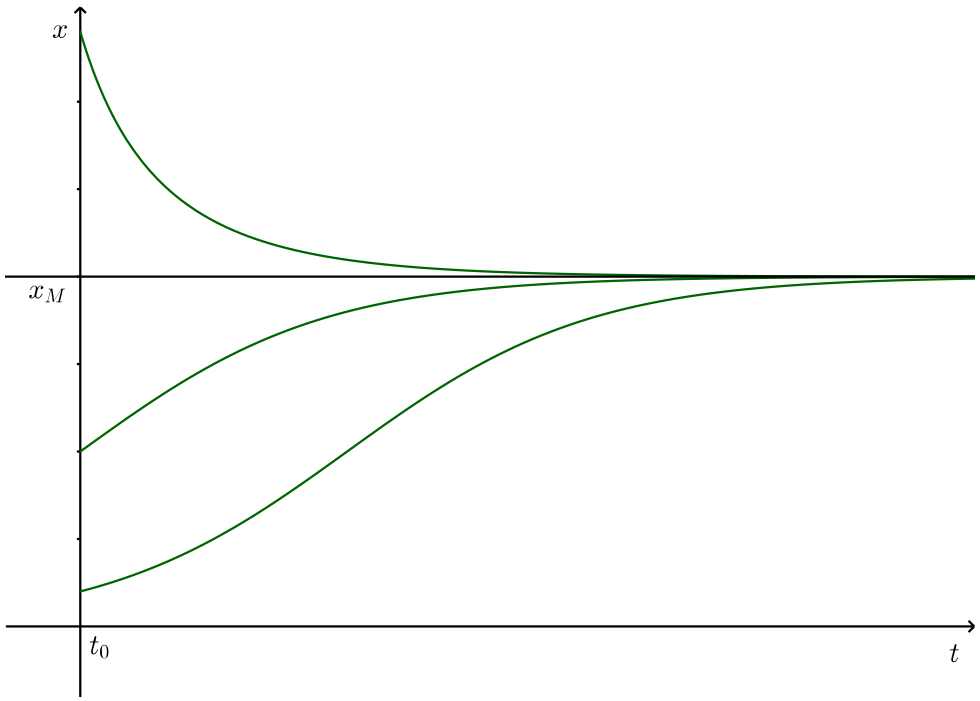


Abbildung: Trajektorien zu unterschiedlichen Anfangswerten

Welche Auswirkungen eine ungeschickte Wahl der charakteristischen Größen haben kann, zeigt folgendes Beispiel.

Beispiel 2.1.1 (Gravitation):

Es soll ein Sprung auf der Erde modelliert werden. Zur Vereinfachung des Modells sei der Luftwiderstand vernachlässigt und die Erde eine Kugel.
Nachdem Absprung ist die einzige auf den Springenden wirkende Kraft die Gravitationskraft, welche gemäß dem Newtonschen Gesetz (näherungsweise) durch

$$F = -G\frac{m_{Erde} \cdot m}{(x+R)^2}$$

gegeben ist. Hierbei bezeichne m die Masse des Springenden, m_{Erde} entsprechend die Erdmasse, R den Erdradius, x den Abstand zum Boden und G die Gravitationskonstante. Wegen $F = m \cdot a$, d. h. $a = \frac{F}{m}$ (a Beschleunigung) führt dies mittels Substitution $g := \frac{Gm_{Erde}}{R^2}$ und somit $F = -\frac{gR^2}{(x+R)^2}m$ auf das Modell

$$x''(t) = -\frac{-gR^2}{(x(t)+R)^2} \qquad (2.1.2)$$

Versehen mit den Anfangsbedingungen $x(0) = 0$, $x'(0) = v_0$ (v_0 bezeichne hierbei die Geschwindigkeit beim Absprung) - ist dies ein AWP mit einer Differentialgleichung 2. Ordnung. (Stark) gerundete Werte für die auftretenden Konstanten sind hierbei $g \approx 10\frac{m}{s^2}$ und $R \approx 10^7 m$, ferner sei $v_0 = 10\frac{m}{s}$. Die im Problem auftretenden Dimensionen sind Länge L, Zeit T und Masse M, für vorkommenden Größen gilt:

- $[m] = M$ für die Masse

- $[t] = T$ für die Zeit
- $[x] = L$ für die Höhe
- $[v_0] = \frac{L}{T}$ für die Geschwindigkeit
- $[g] = \frac{L}{T^2}$ für die Beschleunigung

Zur Bestimmung der charakteristischen Größen verfolgt man den Ansatz

$$\Pi = m^a \cdot v_0^b \cdot g^c \cdot R^d$$

mit nochzubestimmenden Parametern a, b, c, d. Für die Dimension gilt entsprechend:

$$[\Pi] = M^a \cdot \left(\frac{L}{T}\right)^b \cdot \left(\frac{L}{T^2}\right)^c \cdot L^d = M^a \cdot L^{b+c+d} \cdot T^{-b-2c}$$

Für die Analyse der DGL genügt es, folgende drei Fälle zu unterscheiden:

1) $a = 0, b + c + d = 0, -b - 2c = 0$: In diesem Falle muss $b = -2c$ und $d = c$ gelten, somit $\Pi = \left(\frac{gR}{v_0^2}\right)^c$. Im Folgenden sei

$$\varepsilon := \frac{v_0^2}{gR}$$

als ***dimensionsloser Universalparameter*** gewählt.

2) $a = 0, b + c + d = 1, b + 2c = 0$: Hieraus folgt $a = 0, b = -2c, d = 1 + c$. Die charakteristische Länge ist also

$$\bar{x} = v_0^{-2c} g^c R^{1+c} = \left(\frac{gR}{v_0^2}\right)^c R = R \cdot \frac{1}{\varepsilon^c}$$

3) $a = 0, b + c + d = 0, b + 2c = -1$: Hier gilt $b = -1 - 2c, d = 1 + c$. Als charakteristische Zeit ergibt sich:

$$\bar{t} = v_0^{-1-2c} g^c R^{1+c} = \left(\frac{gR}{v_0^2}\right)^c \frac{R}{v_0} = \frac{R}{v_0} \frac{1}{\varepsilon^c}$$

Entdimensionalisiert man nun das Problem (2.1.2) mittels $x(t) = \bar{x} y(\frac{t}{\bar{t}})$ und $\tau = \frac{t}{\bar{t}}$, ergibt sich:

$$\frac{\bar{x}}{\bar{t}^2} y''(\tau) = -\frac{gR^2}{\left(\bar{x} y(\tau) + R\right)^2}$$

beziehungsweise

$$\frac{\bar{x}}{\bar{t}^2 g} y''(\tau) = -\frac{1}{\left(1 + \left(\frac{\bar{x}}{R}\right) y(\tau)\right)^2}$$

sowie für die Anfangsbedingungen:

$$y(0) = 0, y'(0) = \frac{\bar{t}}{\bar{x}} v_0$$

Da das Modell besonders einfach wird, wenn möglichst viele Parameter gleich 1 sind, resultieren hieraus folgende Skalierungsmöglichkeiten:

1) $\frac{\bar{x}}{\bar{t}^2 g} = 1, \frac{\bar{x}}{R} = 1$: Hieraus folgt $\bar{x} = R$ und $\bar{t} = \sqrt{\frac{R}{g}}$. Somit vereinfacht sich das Problem wegen

$$\frac{\bar{t}}{\bar{x}} v_0 = \frac{v_0}{\sqrt{Rg}} = \sqrt{\frac{v_0^2}{Rg}} = \sqrt{\varepsilon}$$

auf das AWP

$$\begin{cases} y''(\tau) = -\dfrac{1}{(1+y(\tau))^2} \\ y(0) = 0,\ y'(0) = \sqrt{\varepsilon} \end{cases}$$

2) $\frac{\bar{x}}{R} = 1, \frac{\bar{t}}{\bar{x}} v_0 = 1$: Hier gilt $\bar{x} = R, \bar{t} = \frac{R}{v_0}$, es ergibt sich

$$\frac{\bar{x}}{\bar{t}^2 g} = \frac{R}{\frac{R^2}{v_0^2} g} = \frac{v_0^2}{Rg} = \varepsilon$$

und folglich

$$\begin{cases} \varepsilon y''(\tau) = -\dfrac{1}{(1+y(\tau))^2} \\ y(0) = 0,\ y'(0) = 1 \end{cases}$$

3) $\frac{\bar{x}}{\bar{t}^2 g} = 1, \frac{\bar{t}}{\bar{x}} v_0 = 1$: Es folgt $\bar{t} = \frac{v_0}{g}, \bar{x} = \frac{v_0^2}{g}$, daher gilt

$$\frac{\bar{x}}{R} = \frac{v_0^2}{Rg} = \varepsilon$$

und somit

$$\begin{cases} y''(\tau) = -\dfrac{1}{(1+\varepsilon y(\tau))^2} \\ y(0) = 0,\ y'(0) = 1 \end{cases}$$

Mit obigen Beispieldaten, $g = 10\frac{m}{s},\ R = 10^7 m, v_0 = 10\frac{m}{s}$, gilt $\varepsilon = \frac{v_0^2}{Rg} = 10^{-6}$. Vernachlässigt man ε infolgedessen, ergibt sich in den obigen Fällen:

1)

$$\begin{cases} y''(\tau) = -\dfrac{1}{(1+y(\tau))^2} \\ y(0) = 0,\ y'(0) = 0 \end{cases}$$

Aus $y''(0) < 0, y'(0) = 0$ folgt für $\tau > 0$ hinreichend klein $y'(\tau) < 0$ und somit analog wegen $y(0) = 0$ auch $y(\tau) < 0$. Das Modell ist unbrauchbar. Dies liegt in der ungeeigneten Skalierung begründet: Die charakteristische Länge ist $\bar{x} = 10^7 m$, die charakteristische Zeit $\bar{t} = 10^3 s$, viel zu große Werte für Sprünge.

2)

$$\begin{cases} 0 = -\dfrac{1}{(1+y(\tau))^2} \\ y(0) = 0,\ y'(0) = 1 \end{cases}$$

Dieses Problem ist sogar unlösbar, ebenfalls infolge zu großer Skalen. (Die Werte sind hier $\bar{x} = 10^7 m$, $\bar{t} = 10^6 s$.)

3)

$$\begin{cases} y''(\tau) = -1 \\ y(0) = 0,\ y'(0) = 1 \end{cases}$$

Hieraus ergibt sich $y'(\tau) = -\tau + c$ und somit folgt wegen $y(\tau) = -\frac{1}{2}\tau^2 + c\tau$ sowie den Anfangsbedingungen:

$$y(\tau) = -\frac{1}{2}\tau^2 + \tau$$

Somit folgt für $x(t) = \bar{x}y\left(\frac{t}{\bar{t}}\right)$ wegen $\bar{x} = \frac{v_0^2}{g}$, $\bar{t} = \frac{v_0}{g}$:

$$\boxed{x(t) = v_0 t - \frac{1}{2}gt^2}$$

Dies entspricht dem aus der Schule bekannten Modell einer konstant beschleunigten Bewegung. Für die Beispieldaten gilt $\bar{x} = 10m$, $\bar{t} = 1s$, die charakteristischen Größen haben somit vernünftige Werte.

2.2 Asymptotische Entwicklung

Aus obigen Beispiel geht hervor, dass die Vernachlässigung eines scheinbar unbedeutenden ε gravierende Auswirkungen haben kann. Daher soll im Folgenden der Ansatz betrachtet werden, ebendieses nicht zu vernachlässigen, sondern eine Reihenentwicklung zu verwenden, welche am Folgenden Beispiel veranschaulicht werden soll.

Beispiel 2.2.1:

Gegeben sei die Gleichung

$$x^2 + 2\varepsilon x - 1 = 0$$

mit $0 < \varepsilon \ll 1$. Diese soll durch eine ***asymptotische Entwicklung*** bis zur 2. Ordnung approximiert werden. Man setzt

$$x = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots = \sum_{i=0}^{\infty} \varepsilon^i x_i$$

mit noch zu bestimmenden Koeffizienten x_i, $i \in \mathbb{N}_0$. *Bis zur 2. Ordnung* bedeutet hierbei, dass man nur die Koeffizienten x_0, x_1, x_2 bestimmt, d. h. die Näherungslösung $x = x_0 + \varepsilon x_1 + \varepsilon^2 x_2$ berechnet.
Idee dahinter: Wegen $\sum_{i=0}^{\infty} \varepsilon^i x_i - \sum_{i=0}^{n} \varepsilon^i x_i = O(\varepsilon^{(n+1)})$ erwartet man bei einer asymptotischen Entwicklung bis zur n. Ordnung für $n > 1$ geringere Abweichungen von der exakten Lösung als bei einer Vernachlässigung von ε, was einer Entwicklung bis zur 0. Ordnung entspräche.
Zur Bestimmung der Koeffizienten setzt man den Reihenansatz in die usprüngliche Gleichung ein:

$$\begin{aligned} 0 &= \left(x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \ldots\right)^2 + 2\varepsilon\left(x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \ldots\right) - 1 = \\ &= x_0^2 + 2\varepsilon x_0 x_1 + \varepsilon^2 x_1^2 + \ldots + 2\varepsilon x_0 + 2\varepsilon^2 x_1 + \ldots - 1 \end{aligned}$$

Mittels Koeffizientenvergleich ergibt sich für

- ε^0: $x_0^2 - 1 = 0$

- ε^1: $2x_0x_1 + 2x_0 = 0$
- ε^2: $x_1^2 + 2x_0x_2 + 2x_1 = 0$

Hieraus folgt $x_0 = \pm 1$. Somit wäre ein erster Ansatz für die Lösung der Gleichung $x = x_0 = 1$ bzw. $x = -1$.
Unabhängig vom Vorzeichen von x_0 folgt aus aus $2x_0\,(x_1 + 1) = 0$ $x_1 = -1$. Es ergeben sich die Näherungslösungen $x = -1 - \varepsilon$ und $x = 1 - \varepsilon$.
Für x_2 ergibt sich wegen $1 + 2x_0x_2 - 2$ aus obigem die Bedingung $x_0x_2 = \frac{1}{2}$ und somit $x_2 = \pm\frac{1}{2}$. Somit lauten die Reihentwicklungen bis zur 2. Ordnung:

$$x = -1 - \varepsilon - \frac{\varepsilon^2}{2} \quad \text{und} \quad x = 1 - \varepsilon + \frac{\varepsilon^2}{2}$$

Nun soll dieses Prinzip auf die Gravitationsgleichung (2.1.2),

$$\begin{cases} y''(\tau) = \dfrac{-1}{(1 + \varepsilon y(\tau))^2} \\ y(0) = 0,\ y'(0) = 1 \end{cases}$$

angewandt werden. Hierzu sei

$$y_\varepsilon(\tau) = y_0(\tau) + \varepsilon y_1(\tau) + \varepsilon^2 y_2(\tau) + \dots$$

Die Taylorentwicklung von $\frac{1}{(1+z)^2}$ um 0 lautet:

$$\frac{1}{(1+z)^2} = 1 - 2z + 3z^2 - 4z^3 \pm \dots$$

Aus $y_\varepsilon''(\tau) = -1 + 2\varepsilon y_\varepsilon(\tau) - 3\varepsilon^2 y_\varepsilon^2(\tau) \pm \dots$ ergibt sich:

$$\begin{gathered} y_0''(\tau) + \varepsilon y_1''(\tau) + \varepsilon^2 y_2''(\tau) + \dots = \\ = -1 + 2\varepsilon\left(y_0(\tau) + \varepsilon y_1(\tau) + \varepsilon^2 y_2(\tau) + \dots\right) - 3\varepsilon^2\left(y_0(\tau) + \varepsilon y_1(\tau) + \dots\right)^2 \pm \dots \end{gathered}$$

Die Koeffizienten für eine Reihenentwicklung bis zur 2. Ordnung erhält man erneut sukzessive mittels Koeffizientenvergleich (man beachte $y_\varepsilon(0) \stackrel{!}{=} 0$, $y_\varepsilon(1) \stackrel{!}{=} 1$):

- ε^0: $y_0''(\tau) = -1 \to$ Die (eindeutige) Lösung des AWPs

  $$\begin{cases} y_0'' = -1 \\ y_0(0) = 0,\ y_0'(0) = 1 \end{cases}$$

  ist

  $$y_0(\tau) = \tau - \frac{1}{2}\tau^2$$

- ε^1: $y_1''(\tau) = 2y_0(\tau) \to$ Einsetzen des bisherigen liefert $y_1''(\tau) = 2\tau - \tau^2$.
  Die (eindeutige) Lösung des AWPs

  $$\begin{cases} y_1'' = 2\tau - \tau^2 \\ y_1(0) = 0,\ y_1'(0) = 0 \end{cases}$$

  lautet

  $$y_1(\tau) = \frac{1}{3}\tau^3 - \frac{1}{12}\tau^4$$

- ε^2: $y_2''(\tau) = 2y_1(\tau) - 3y_0^2(\tau) \rightarrow$ Einsetzen des Bisherigen liefert $y_2''(\tau) = \frac{2}{3}\tau^3 - \frac{1}{6}\tau^4 - 3\tau^2 + 3\tau^3 - \frac{3}{4}\tau^4 = -3\tau^2 + \frac{11}{3}\tau^3 - \frac{11}{12}\tau^4$.
  Die (eindeutige) Lösung des AWPs

$$\begin{cases} y_2'' = -3\tau^2 + \dfrac{11}{3}\tau^3 - \dfrac{11}{12}\tau^4 \\ y_2(0) = 0,\ y_2'(0) = 0 \end{cases}$$

  ist

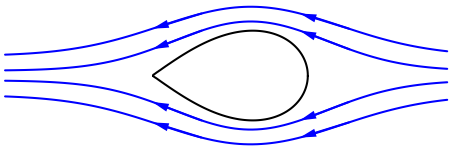
$$y_2(\tau) = -\frac{1}{4}\tau^4 + \frac{11}{60}\tau^5 - \frac{11}{360}\tau^6$$

Es folgt:

$$y_\varepsilon(\tau) = \left(\tau - \frac{1}{2}\tau^2\right) + \varepsilon\left(\frac{1}{3}\tau^3 - \frac{1}{12}\tau^4\right) + \varepsilon^2\left(-\frac{1}{4}\tau^4 + \frac{11}{60}\tau^5 - \frac{11}{360}\tau^6\right) + O(\varepsilon^3)$$

2.3 Beispiel aus der Strömungsmechanik

Gegeben sei ein von einem Fluid (Gas, Flüssigkeit) umströmter Körper.



skizzenhafter Querschnitt des Körpers mit
umgebendem Fluid

Nun soll die Strömungsgeschwindigkeit $v(x,t) \in \mathbb{R}^3$ $(t \geq 0)$ bestimmt werden. Als Zusatzannahme sei v für $|x| \rightarrow \infty$ konvergent, d. h. es existiere $V \in \mathbb{R}^3$ mit $\lim\limits_{|x|\rightarrow\infty} v(x,t) = V$.
Vernachlässigt man äußere Kräfte, so ergeben sich für ein inkompressibles Fluid mit konstanter Dichte ϱ_0 - unter bestimmten Voraussetzungen an das Fluid - die sogenannten ***Navier-Stokes-Gleichungen***:

$$\varrho_0\left(\partial_t v + (v\cdot\nabla)\cdot v\right) = -\nabla\rho + \mu\Delta v$$
$$\nabla v = 0$$

Hierbei bezeichnet

- ρ den Druck
- μ die dynamische Viskosität des Fluids
- $\nabla \cdot v = \sum_{i=1}^3 \frac{\partial}{\partial x_i} v_i$ die Divergenz des Vektorfeldes v
- $\Delta v = \sum_{i=1}^3 \frac{\partial^2}{\partial x_i^2} v_i$ den Laplace-Operator

Ferner gilt:

$$(v\cdot\nabla)\,v = \left(\sum_{i=1}^{3} v_i \frac{\partial}{\partial x_i} v_j\right)_{j=1,2,3}$$

Mit den Bezeichnungen L für Länge, T für Zeit und M für Masse ergeben sich folgende Dimensionen:

Parameter/Größe	Dimension
Geschwindigkeit v	$[v] = \frac{L}{T}$
Dichte ϱ_0	$[\varrho_0] = \frac{M}{L^3}$
Druck ρ	$[\rho] = M \cdot \frac{L}{T^2} \cdot \frac{1}{L^2} = \frac{M}{T^2 L}$
Dynamische Viskosität μ	$[\mu] = \frac{M}{LT}$

Sei nun $\tau = \frac{t}{\bar{t}}$, $y = \frac{x}{\bar{x}}$, $u(y,\tau) = \frac{v(x,t)}{\|V\|_2}$ sowie $q(y,\tau) = \frac{\rho(x,t)}{\bar{\rho}}$, wobei $\bar{\cdot}$ wie gehabt die (noch zu bestimmende) charakteristische Größe bezeichne. Setzt man die Identitäten $\partial_t = \frac{1}{\bar{t}}\partial_\tau$ und $\nabla_x = \frac{1}{\bar{x}}\nabla_y$ in die Navier-Strokes-Gleichungen ein, so erhält man:

$$\frac{1}{\bar{t}}\varrho_0 \|V\|_2 \left(\partial_\tau u + \frac{\|V\|_2 \bar{t}}{\bar{x}} (u\cdot\nabla)\cdot u\right) = -\frac{1}{\bar{x}}\bar{\rho}\nabla q + \mu\frac{\|V\|_2}{\bar{x}^2}\Delta u$$

$$\Updownarrow$$

$$\partial_\tau u + \frac{\|V\|_2 \bar{t}}{\bar{x}} (u\cdot\nabla)\cdot u = -\frac{\bar{\rho}}{\varrho_0}\frac{\bar{t}}{\bar{x}\|V\|_2}\nabla q + \frac{\mu}{\varrho_0}\frac{\bar{t}}{\bar{x}^2}\Delta u$$

Mit $\frac{\|V\|_2 \bar{t}}{\bar{x}} = 1$, $\bar{\rho} = \bar{v}^2\varrho_0$ und ***kinematischer Viskosität*** $\eta = \frac{\mu}{\varrho_0}$ ergeben sich wegen

a) $\bar{t} = \frac{\bar{x}}{\|V\|_2}$

b) $\frac{\bar{t}\bar{\rho}}{\varrho_0 \bar{x}\|V\|_2} = 1$

als dimensionslose Gleichungen:

$$\partial_\tau u + (u\cdot\nabla)\,u = -\nabla q + \frac{1}{\mathrm{Re}}\Delta u$$
$$\nabla u = 0$$

$\mathrm{Re} = \frac{\bar{x}\|V\|_2}{\eta}$ ist hierbei die Reynoldszahl. Es bleibt die Frage zu klären, was die Größe der Reynoldszahl für über das Modell, z. B. die Abhängigkeiten/Gestalt der Reibungskraft, aussagt.

Kleine Reynoldszahl (z. B. hohe Viskosität, geringe Strömungsgeschwindigkeit): In diesem Falle sind die charakteristischen Größen die Geschwindigkeit v des Körpers (relativ zur Strömung), eine charakteristische Länge $\bar{x}$ des Körpers sowie die dynamische Viskosität des Fluids. Wegen $[v] = \frac{L}{T}$, $[\bar{x}] = L$, $[\mu] = \frac{FT}{L^2}$ (F Dimension einer Kraft) ergibt sich für Kombinationen dieser Größen:

$$\left[v^a \bar{x}^b \mu^c\right] = \left(\frac{L}{T}\right)^a L^b \left(\frac{FT}{L^2}\right)^c = L^{a+b-2c} T^{-a+c} F^c$$

Für eine Kraft muss folglich $c = 1$, $a + b - 2c = 0$ sowie $-a + c = 0$, mithin also $a = b = 1$ gelten. Somit muss das Gesetz für den Reibungswiderstand von der Form

$$F_R = c_R \mu \bar{x} v$$

sein. Hierbei bezeichnet c_R einen Reibungskoeffizienten. Für eine Kugel ist dieser beispielsweise 6π. Wählt man als charakteristische Länge $\bar{x}$ den Kugelradius r, so ergibt dies das ***Stokessche Gesetz***:

$$F_R = 6\pi r \mu v$$

Große Reynoldszahl (geringe Viskosität, hohe Strömungsgeschwindigkeit): In diesem Falle ist die zur Verdrängung der Flüssigkeit benötigte Kraft die für die Reibung ausschlaggebendere. Bezeichnet A die Querschnittsfläche des Körpers, dann ergibt sich für ein Zeitintervall Δt als

- pro Zeitintervall verdrängtes Volumen ΔVol: $\Delta Vol \sim Av\Delta t$
- pro Zeitintervall verdrängte Masse Δm: $\Delta m = \varrho \cdot \Delta Vol \sim \varrho Av\Delta t$
- pro Zeitintervall hinzugeführte kinetische Energie ΔE_{kin}: $\Delta E_{kin}\frac{1}{2}\Delta m v^2 \sim \frac{1}{2}\varrho Av^3\Delta t$

Für die Reibung ergibt sich folglich

$$F_R v\Delta t = \Delta E_{kin}$$

beziehungsweise

$$F_R \sim \frac{1}{2}\varrho Av\Delta t$$

In diesem Falle wird die Proportionalitätskonstante mit c_W bezeichnet und es ergibt sich

$$F_R = \frac{1}{2}c_w\varrho Av\Delta t$$

2.4 Populationsdynamik

Anhand des Einführungsbeispiels dieses Kapitels wurden drei Modellierungsmöglichkeiten für biologisches Populationswachstum aufgezeigt. Im Folgenden sollen weitere Wachstumsmodelle vorgestellt und (exemplarisch) analysiert werden, insbesondere solche, die sich mit gegenseitiger Beeinflussung unterschiedlicher Spezies beschäftigen. Hierbei werden Resultate aus [8] und [9] zusammengefasst.

Ein sehr simples Modell, welches bereits im Jahre 1798 von *Thomas Robert Malthus* veröffentlicht wurde, basiert auf der Annahme eines unbeschränkten Wachstums mit konstantem Wachstumsfaktor R. Das Wachstum wird folglich durch die DGL $\dot{x} = Rx$ beschrieben. Wie im genannten Beispiel lautet die Lösung für gegebene Anfangsgröße $x_0 = x(t_0 = 0)$ also

$$x(t) = x_0 \cdot e^{Rt}$$

Für lange Zeitfenster ist ein derartiges Modell jedoch nur bedingt geeignet, da dem Wachstum von Populationen, wie beim Beispiel der Kuhherde erwähnt, durch verschiedene Faktoren wie z. B. Nahrungsvorkommen Grenzen gesetzt sind. Ein Modell, das solche Wachstumsgrenzen berücksichtigt, ist das Modell von *Pierre-François Verhulst* aus dem Jahre 1848. In diesem ist der Wachstumsfaktor R nicht mehr konstant, sondern eine von der Populationsgröße x und der (festen) Kapazität K abhängige Funktion $R(x) = r \cdot \left(1 - \frac{x}{K}\right)$. Die entsprechende DGL ist die logistische Gleichung (2.1.1) $\dot{x} = rx\left(1 - \frac{x}{K}\right)$, die Lösung ist, analog zu oben (man beachte $q = \frac{r}{K}$, $x_M = K$ sowie $t_0 = 0$):

$$x\left(t\right) = \frac{Kx_0}{x_0 + \left(K - x_0\right)e^{-r(t-t_0)}}$$

Jacques Monod entwickelte ein Modell basierend auf biologischen Studien, welches das beschränkte Wachstum von Bakterien besser beschreibt. In seinem Modell lautet die Wachstumsrate

$$R(x) = \frac{\mu_0 x}{K + x},$$

wobei μ_0 die maximale Wachstumsrate und K die Substratkonzentration ist, für die die Wachstumsrate die Hälfte der maximalen ist.
Eines der ersten Modelle zur Beschreibung von Populationswachstum bei gegenseitiger Beeinflussung zweier Spezies sind die nach ihren Urhebern/Entdeckern *Alfred James Lotka* und *Vito Volterra* benannten Lotka-Volterra-Gleichungen:

$$\begin{aligned} \frac{\mathrm{d}N}{\mathrm{d}t} &= N\,(a - bP) \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= P\,(cN - d) \end{aligned} \tag{2.4.1}$$

N beschreibt eine Beutetierpopulation, deren Wachstum durch die Räuberpopulation P negativ beinflusst wird, während diese wiederum durch große Beutepopulationen positiv beeinflusst wird. a ist hierbei eine Art „natürliche“ Wachstumsrate, d entsprechend eine natürliche Todesrate, b beschreibt die Todesrate der Beutepopulation pro Raubtier, c entsprechend die Wachstumsrate der Räuberpopulation pro Beutetier.
Letzteres soll im Folgenden untersucht und verfeinert werden. Hierzu sei $\tau = at$ und $\alpha = \frac{d}{a}$. Die Substitutionen $u(\tau) = \frac{c}{d}N(t)$ und $v(\tau) = \frac{b}{a}P(t)$ liefern:

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}\tau} = u \cdot (1 - v) \\ \dfrac{\mathrm{d}v}{\mathrm{d}\tau} = \alpha v \cdot (u - 1) \end{cases} \tag{2.4.2}$$

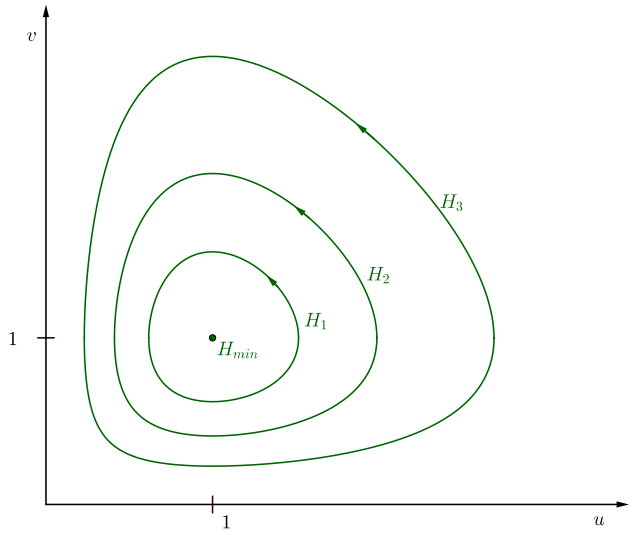
Setzt man $\dot{u} = \frac{\mathrm{d}u}{\mathrm{d}\tau}$ und $\dot{v} = \frac{\mathrm{d}v}{\mathrm{d}\tau}$, ergibt sich wegen $\frac{\dot{v}}{\dot{u}} = \alpha\frac{v(u-1)}{u(1-v)}$ die Bedingung

$$u(1 - v)\dot{v} + \alpha(1 - u)v\dot{u} = 0$$

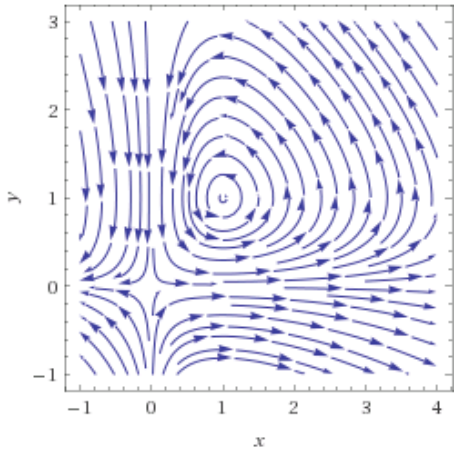
beziehungsweise (dividiert durch uv):

$$\left(\frac{1}{v} - 1\right)\dot{v} + \alpha\left(\frac{1}{u} - 1\right)\dot{u} = 0 \tag{2.4.3}$$

Sei $G(u, v) = \alpha u + v - \log(u^\alpha v)$. Dann gilt $\frac{\partial G}{\partial u}(u, v) = \alpha\left(1 - \frac{1}{u}\right)$ und $\frac{\partial G}{\partial v}(u, v) = 1 - \frac{1}{v}$ und die Lösung der DGL (2.4.3) ist implizit durch $H = G(u, v) = \alpha u + v - \log(u^\alpha v)$ mit konstantem H gegeben. Den minimalen Wert nimmt $G(u, v)$ in der Ruhelage $u = 1$, $v = 1$ ein, folglich muss $H \geq H_{min} = 1 + \alpha$ sein.



Phasenporträt mit Richtung derÄnderung für $\alpha = 1$



Mit Mathematica erstellter Streamplot des Systems (2.4.2)

Das System (2.4.2) hat 2 Ruhelagen:

1. $(0,0)$: Linearisiert man dieses wie in Kapitel 1.8.3 gezeigt, ergibt sich mittels $x = u-0$, $y = v-0$:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & -\alpha \end{bmatrix}}_{=:A} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

   Es ist $A-\lambda I = \begin{bmatrix} 1-\lambda & \\ 0 & -\alpha-\lambda \end{bmatrix}$, die Eigenwerte sind somit $\lambda_1 = 1 > 0$ und $\lambda_2 = -\alpha < 0$. Somit liegt hier ein Sattelpunkt vor

2. $(1,1)$: Mit $x = u-1$ und $y = v-1$ lautet das linearisierte System

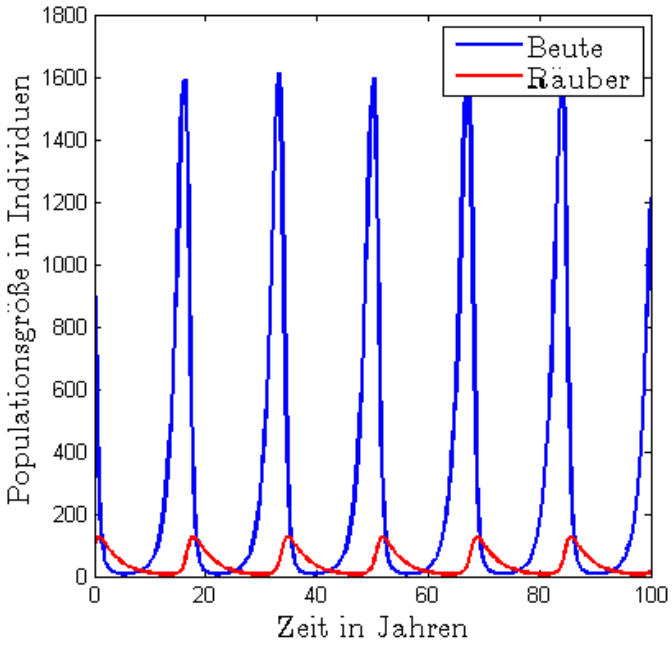
$$\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & -1 \\ \alpha & 0 \end{bmatrix}}_{=:A} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

   Aus $\det(A-\lambda I) = \det\left(\begin{bmatrix} -\lambda & -1 \\ \alpha & -\lambda \end{bmatrix}\right) = \lambda^2+\alpha$ folgt: Die Eigenwerte sind $\lambda_1 = i\sqrt{\alpha}$, $\lambda_2 = -i\sqrt{\alpha}$. Sind $\underline{k}_1$ und $\underline{k}_2$ zugehörige Eigenvektoren, so ist die allgemeine Lösung des linearisierten Systems durch

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \cdot \cos\left(\sqrt{\alpha}\tau\right) + \begin{pmatrix} -\frac{c_2}{\sqrt{\alpha}} \\ \frac{c_1}{\sqrt{\alpha}} \end{pmatrix} \cdot \sin\left(\sqrt{\alpha}\tau\right)$$

   gegeben (vgl. Kapitel 1.6). Die Lösungen sind also periodisch mit Periode $\frac{2\pi}{\sqrt{\alpha}} = 2\pi\sqrt{\frac{a}{d}}$

Simulation 2.4.1:
(Der Matlab-Code für diese sowie folgende Simulationen befindet sich, sofern nicht anders vermerkt im Anhang.) Für diese numerische Simulation wurden die Werte $a = 0,8$, $b = 0,02$, $c = 0,001$ sowie $d = 0,3$ gewählt. Der simulierte Zeitraum betrug 100 Jahre, wobei zum Anfang 1000 Beutetiere und 100 Räuber vorhanden waren.



Beobachtung und Interpretation der Ergebnisse: Anfangs existiert eine hohe Zahl an Beutetieren, welche jedoch stark von den Räubern dezimiert werden. In Folge mangelnder Nahrung sterben mehr Räuber als durch den Beutefang begünstigt geboren werden, sodass auch diese Anzahl sinkt. Hierdurch überleben mehr Beutetiere, die Population steigt wieder, wodurch sich nach einiger Zeit auch die Räuberpopulation erholen kann. Da hierdurch wieder mehr Beute gerissen wird, beginnt der Kreislauf von neuem.

Dieses Modell (2.4.1) hat jedoch denselben Nachteil, den das Einführungsbeispiels (zumindest anfangs) auch hatte: Das Wachstum ist nicht beschränkt, was sich vor allem bei Verzicht auf die Räuberpopulation bemerkbar macht. (In diesem Falle stimmen beide Modelle überein.) Daher soll nun der Ansatz

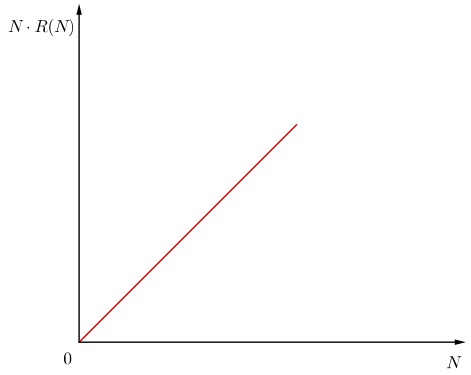
$$\begin{aligned} \frac{\mathrm{d}N}{\mathrm{d}t} &= N \cdot F\left(N, P\right) \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= P \cdot G\left(N, P\right) \end{aligned}$$

betrachtet werden, wobei F und G Funktionen sind, welche solche Wachstumsschranken berücksichtigen. Eine Möglichkeit für F wäre zum Beispiel

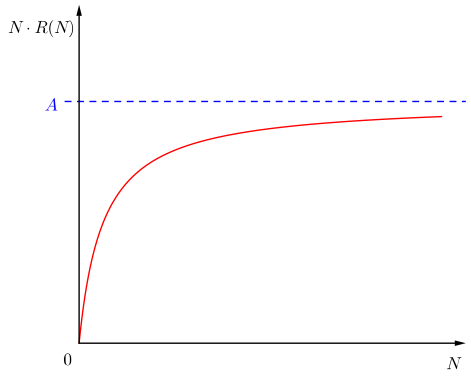
$$F\left(N, P\right) = r\left(1 - \frac{N}{K}\right) - P \cdot R\left(N\right)$$

Hierbei bezeichne K die maximale Größe, die die Beutepopulation bei Fehlen der Räuberspezies erreichen kann. (In diesem Falle entspräche die Gleichung der logistischen.) Je nach Wahl von $R(N)$ ergibt sich ein anderer Einfluss der Räuber- auf die Beutepopulation, wie folgende Beispiele zeigen:

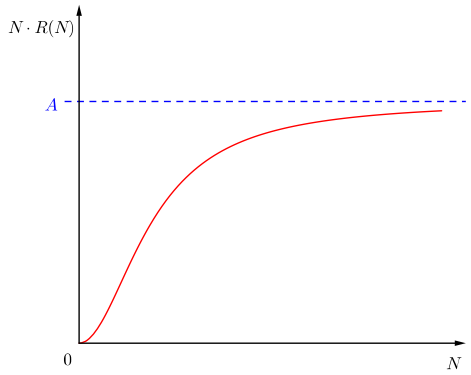
a) $R(N) = A$ mit positiver Konstante A.



b) $R(N) = \frac{A}{N+B}$ mit positiven Konstanten A, B



c) $R(N) = \frac{AN}{N^2+B^2}$ mit positiven Konstanten A, B



Für den Zuwachs $P \cdot G(N, P)$ der Räuberpopulation kann man anstelle von $G(N, P) = -c+dN$ zum Beispiel $G(N, P) = k\left(1 - h\frac{P}{N}\right)$ mit positiven Konstanten h, k wählen. Im Folgenden sei

nun $F(N,P) = r\left(1-\frac{N}{K}\right) - \frac{kP}{N+D}$ und $G(N,P) = s\left(1-h\frac{P}{N}\right)$ (D, K, h, k, r, $s > 0$). Das betrachtete Modell lautet also:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = N\left[r\left(1-\frac{N}{K}\right) - \frac{kP}{N+D}\right]$$
$$\frac{\mathrm{d}P}{\mathrm{d}t} = P\left[s\left(1-h\frac{P}{N}\right)\right]$$

Die Substitutionen $u(\tau) = \frac{N(t)}{K}$, $v(\tau) = h\frac{P(t)}{K}$, $\tau = rt$, $a = \frac{k}{hr}$ und $b = \frac{s}{r}, d = \frac{D}{K}$ führen auf das Modell

$$\begin{cases} \frac{\mathrm{d}u}{\mathrm{d}\tau} = u(1-u) - a\frac{uv}{u+d} =: f(u,v) \\ \frac{\mathrm{d}v}{\mathrm{d}\tau} = bv\left(1-\frac{v}{u}\right) =: g(u,v) \end{cases}$$

Zur Bestimmung der Ruhelagen muss man das System

$$f(u^*,v^*) = 0$$
$$g(u^*,v^*) = 0$$

lösen. Aus $bv^*\left(1-\frac{v^*}{u^*}\right) = 0$ folgt $u^* = v^*$. (Oder $v^* = 0$, aber in diesem Falle wäre u^* ebenfalls gleich 0 und das Populationswachstum nicht weiter interessant). Wegen

$$f(u^*,v^*) = u^*(1-v^*) - a\frac{u^*v^*}{u^*+d} = u^{*^2} + (a+d-1)u^* - d$$

folgt (Populationsgrößen sind stets nicht negativ):

$$u^* = \frac{(1-a-d) + \sqrt{(1-a-d)^2 + 4d}}{2} \tag{2.4.4}$$

Linearisiert man mit Hilfe der Substitutionen $x(\tau) = u(\tau) - u^*$ und $y(\tau) = v(\tau) - v^*$, ergibt sich das System

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{pmatrix} x \\ y \end{pmatrix} = A\begin{pmatrix} x \\ y \end{pmatrix}$$

mit

$$A = \begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}_{(u^*,v^*)} = \begin{bmatrix} u^*\cdot\left(\frac{au^*}{(u^*+d)^2} - 1\right) & -\frac{au^*}{u^*+d} \\ b & -b \end{bmatrix}$$

Aus der Bedingung $\det(A-\lambda I) = \lambda^2 - (\mathrm{tr}\ A)\lambda + \det A \stackrel{!}{=} 0$ ergeben sich die Eigenwerte

$$\lambda_{1,2} = \frac{\mathrm{tr}\ A \pm \sqrt{(\mathrm{tr}\ A)^2 - 4\det A}}{2}$$

Damit die Ruhelage stabil ist, sei $\mathrm{Re}\,\lambda_{1,2} < 0$, womit $\det A > 0$ sowie $\mathrm{tr}\ A < 0$ gelten muss, d. h.

$$\det A = \left(1 - \frac{au^*}{(u^*+d)^2}\right)bu^* + b\cdot\frac{au^*}{u^*+d} = \left(1 + \frac{ad}{(u^*+d)^2}\right)bu^* > 0$$
$$\mathrm{tr}\ A = u^*\left(\frac{au^*}{(u^*+d)^2} - 1\right) - b < 0$$

Ersteres ist immer erfüllt, für letzteres folgt au $u^*\left(\frac{au^*}{(u^*+d)^2} - 1\right) < b$ mittels Einsetzen von (2.4.4)

$$b > \left(a - \sqrt{(1-a-d)^2 + 4d}\right)\cdot\frac{1+a+d-\sqrt{(1-a-d)^2+4d}}{2a}$$

Für die Funktionen

$$\gamma(d) = 1 + a + d - \sqrt{(1-a-d)^2 + 4d}$$

und

$$\beta(d) = a - \sqrt{(1-a-d)^2 + 4d}$$

gilt $\gamma'(d) < 0$ und $\beta'(d) < 0$, d. h. beide fallen monoton. Ferner gilt $\gamma(d) > 0$ sowie $\max \beta(d) = \beta(0)$.
Somit folgt mittels

$$b_{d=0} > (a - |1-a|)\,\frac{(1 + a - |1-a|)}{2a}$$

$$b_{d=0} = \begin{cases} > 2a - 1,\ 0 < a \leq 1 \\ > \dfrac{1}{a},\ 1 \leq a \end{cases}$$

Hieraus folgt: Für $0 < a < \frac{1}{2}$ ist das System für beliebige $b > 0$, $d > 0$ in der Ruhelage stabil.
Sei nun $b = 0$. Am Rand des stabilen Systems gilt dann wegen $b \geq \beta(d) \cdot \gamma(d)$ sowie $\gamma(d) > 0$ $\beta(t) = 0$, d. h.

$$a = \sqrt{(1-a-d)^2 + 4d}$$

$$\Downarrow$$

$$a^2 = (1-a-d)^2 + 4d =$$

$$= (1-a)^2 + d^2 - 2(1-a)\,d + 4d$$

Hiermit gilt

$$d^2 + (4 - 2(1-a))\,d + (1-2a) = 0$$

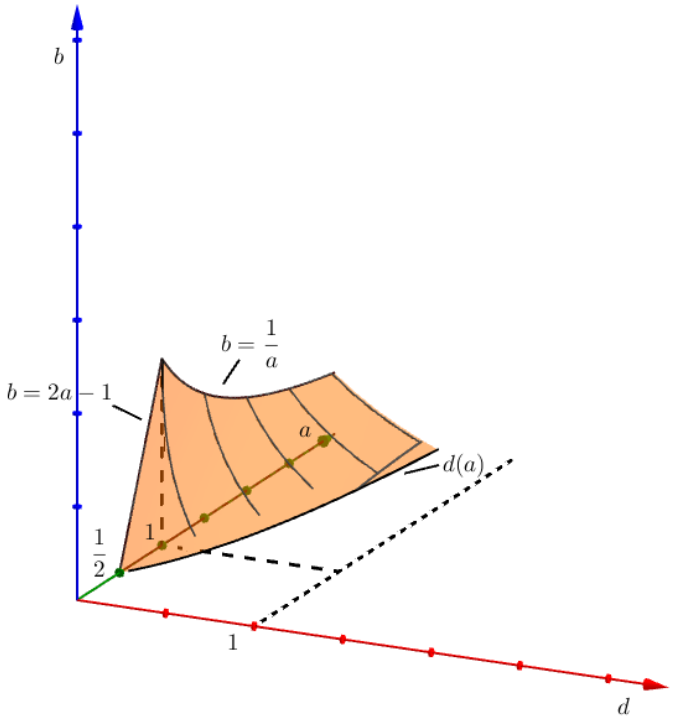


Abbildung: Für die Punkte (d, a, b), welche im gekennzeichneten Volumen liegen, ist die Ruhelage (u^*, v^*) instabil, für alle außerhalb ist sie stabil.

Für $b = 0$ gilt folglich (a, $d > 0$):

$$d(a) = -(2 - (1-a)) + \sqrt{(2-(1-a))^2 - (1-2a)} =$$
$$= -(1+a) + \sqrt{a^2 + 4a}$$

Da ferner $\gamma = 1$ die Ungleichung

$$a^2 + 4a < (1+\gamma)^2 + 2(1+\gamma)\,a + a^2$$

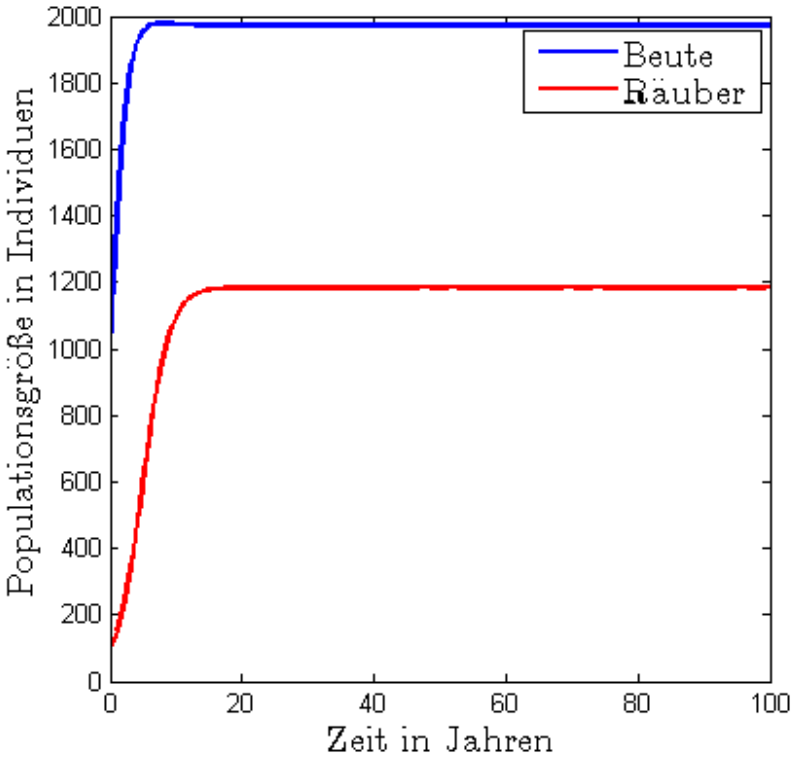
und somit $\sqrt{a^2+4a} < (1+\gamma) + a$ erfüllt, folgt

$$d(a) = \sqrt{a^2 + 4a} - (1+a) < 1$$

Simulation 2.4.2:
Für diese Simulation wurden die Werte

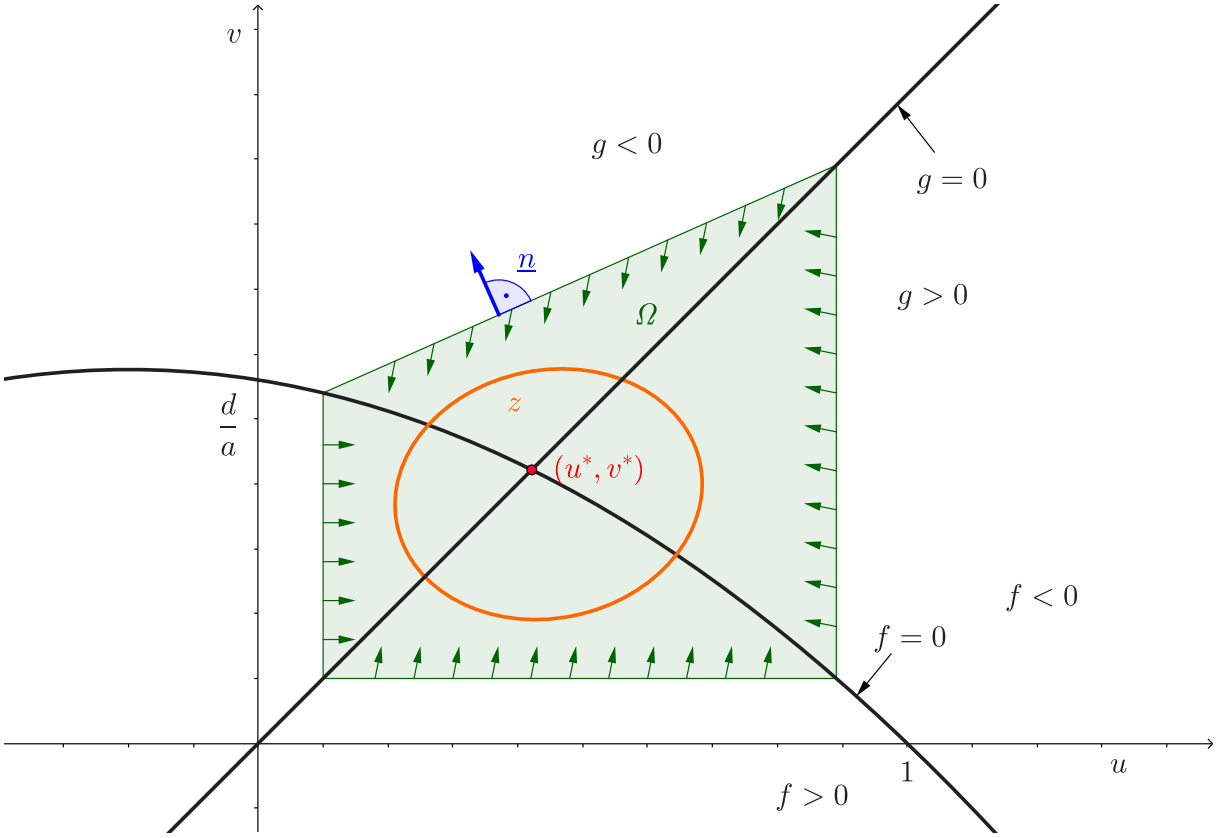
- $r = 0,8$
- $K = 2000$
- $k = 0,02$
- $D = 200$
- $s = 0,5$
- $h = 5/3$

gewählt. Die Anfangswerte waren erneut $N_0 = 1000$ und $P_0 = 100$. Auch der modellierte Zeitraum betrug erneut 100 Jahre.



Beobachtung und Interpretation der Ergebnisse: Bei den gewählten Werten streben die Populationsgrößen offenbar gegen ein stabiles Gleichgewicht. Dies ist auf mehrere Gründe zurückzuführen. Zum einen kann wegen $h > 1$ die Räuberpopulation nur wachsen, wenn ihre Größe einen Bruchteil der Beutepopulation ausmacht. Hierdurch kann es auch bei großen Beutepopulationen zu keinen „Wachstumsexplosionen“ wie im ersten Modell kommen, da ein Überhang an Räubern zu einer Verkleinerung der Population führen würde. Zum anderen ist infolge des Faktors $\frac{1}{N+D}$ der Einfluss der Anzahl an getöteten Beutetieren auf die Vermehrung der Beute umso geringer, je größer die Population ist. Da der Anteil der gerissenen Tiere wegen $k = 0,02$ ohnehin relativ gering ist, kommt es zu keinen starken Wachstumseinbrüchen der Beutepopulation. Dies führt in Verbindung mit dem vergleichsweiße gemäßigten Wachstum der Räuberpopulation dazu, dass die Beutepopulation die durch ihr Wachstum hervorgerufene höhere Anzahl an Räubern stets verkraften kann und somit wächst. Die Gleichgewichtslage ergibt sich infolge dessen durch $K = 2000$, da die Beutepopulation nicht unbegrenzt wachsen kann.

Ist die Ruhelage instabil, so existiert ein Gebiet Ω, so dass für alle $(u, v) \in \partial\Omega$ der Gradient $\begin{pmatrix} \frac{\mathrm{d}u}{\mathrm{d}t} \\ \frac{\mathrm{d}v}{\mathrm{d}t} \end{pmatrix}$ nach innen zeigt. $\underline{\omega} \cdot \underline{n} < 0$ mit $\underline{\omega} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}\tau}u \\ \frac{\mathrm{d}}{\mathrm{d}\tau}v \end{pmatrix}$ und äußerem Normalenvektor $\underline{n}$ gilt. Somit ist Theorem 1.8.9 anwendbar, es existiert folglich ein (in diesem Falle) stabiler Grenzzyklus z.

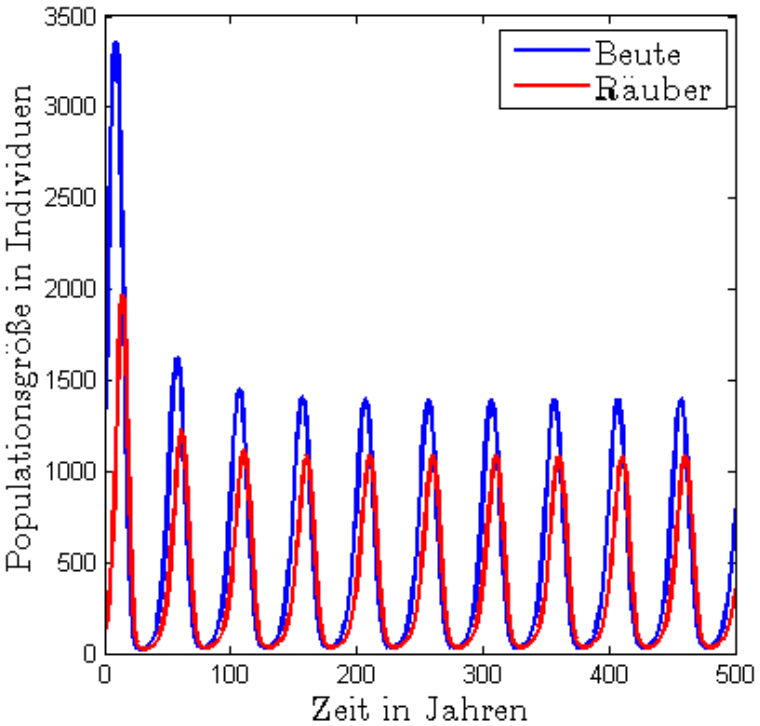


Stabiler Grenzzyklus bei instabiler Ruhelage (u^*, v^*)

Simulation 2.4.3:
Für diese Simulation wurden die Werte

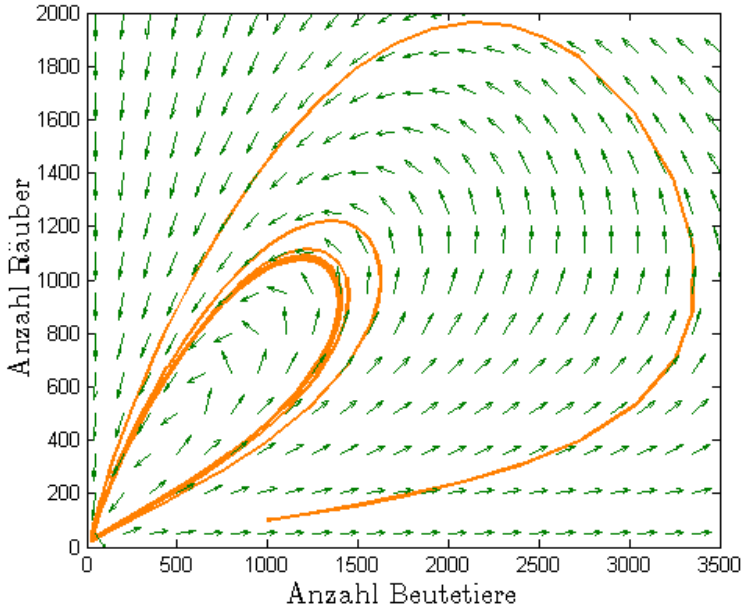
- $r = 0,4$
- $K = 5000$
- $k = 0,44$
- $D = 10$
- $s = 0,32$
- $h = 1,1$

gewählt. Wie in den vorherigen Modellen waren Anfangs 1000 Beutetiere und 100 Räuber vorhanden. Der simulierte Zeitraum umfasste 500 Jahre.



Beobachtung und Interpretation der Ergebnisse: Hier kommt es zu ähnlichen Schwankungen in den Populationsgrößen wie in der Simulation des klassischen Lotka-Volterra-Modells. Wie in der vorherigen Simulation ist wegen $h > 1$ auch hier positives Wachstum der Räuberpopulation nur für $P < N$ möglich. Ebenso verringert sich der Einfluss der Anzahl an gerissenen Beutetieren mit steigender Populationsgröße N, jedoch nicht mehr im selben Maße wie vorher. Dies macht sich vor allem bei kleineren Beutepopulation bemerkbar, auch wegen des relativ hohen Faktors k. Hierdurch wird die Beutepopulation sehr stark dezimiert, wodurch auch die Größe Räuberpopulation einbricht und der Beutepopulation die Möglichkeit der Erholung bietet. Die „Spitzen“ am anfang sind auf die willkürliche Wahl der Anfangsbedingungen zurückzu-

führen: Der Unterschied in den Populationsgrößen war zu groß, der entsprechende Punkt im Phasenraum war der (periodischen) Gleichgewichtslage (Grenzzyklus) nicht nahe genug, wie das zugehörige Phasenporträts veranschaulicht:



2.4.1 Kompetitives Modell

Nun soll das Modell

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = r_1 N_1 \left[1 - \frac{N_1}{K_1} - b_{12}\frac{N_2}{K_1}\right]$$
$$\frac{\mathrm{d}N_2}{\mathrm{d}t} = r_2 N_2 \left[1 - \frac{N_2}{K_2} - b_{21}\frac{N_1}{K_2}\right]$$

betrachtet werden. Dieses beschreibt Situationen, in welchen keine Spezies die Nahrungsgrundlage der anderen ist, beide Arten jedoch das Wachstum der anderen negativ beeinflussen, z. B. da sie um die selbe, knappe Nahrung konkurrieren.
Entdimensionalisiert man dieses mithilfe der Substitutionen $u_1 = \frac{N_1}{K_1}$, $u_2 = \frac{N_2}{K_2}$, $\tau = r_1 t$, $\rho = \frac{r_2}{r_1}$, $a_{12} = b_{12}\frac{K_2}{K_1}$ und $a_{21} = b_{21}\frac{K_1}{K_2}$, erhält man das vereinfachte Modell

$$\frac{\mathrm{d}u_1}{\mathrm{d}\tau} = u_1\left(1 - u_1 - a_{12}u_2\right) =: f_1\left(u_1, u_2\right)$$
$$\frac{\mathrm{d}u_2}{\mathrm{d}\tau} = \rho u_2\left(1 - u_2 - a_{21}u_1\right) =: f_2\left(u_1, u_2\right)$$

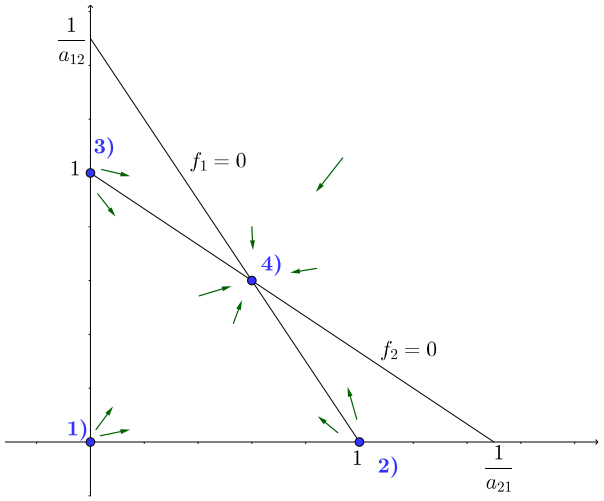
Dieses DGL-System hat drei bis vier Ruhelagen:

1) $u_1^* = 0$, $u_2^* = 0$
2) $u_1^* = 1$, $u_2^* = 0$
3) $u_1^* = 0$, $u_2^* = 1$
4) $u_1^* = \frac{1-a_{12}}{1-a_{12}a_{21}}, u_2^* = \frac{1-a_{21}}{1-a_{12}a_{21}} \rightarrow$ Diese existiert jedoch nur für $a_{12}a_{21} \neq 1$ und ist für das Modell nur dann relevant, wenn zusätzlich $u_1^*, u_2^* \geq 0$ gilt.

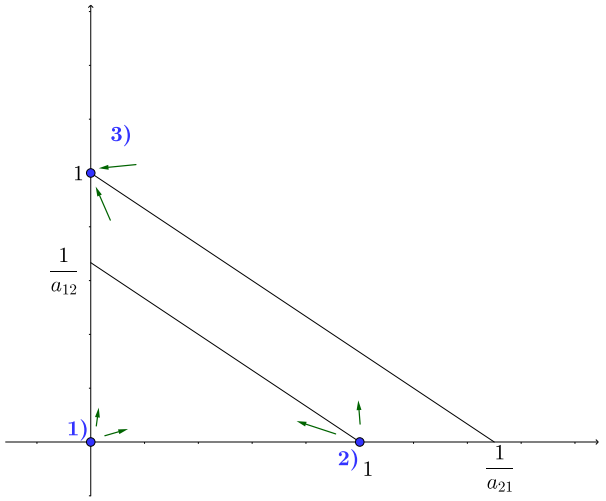
Die Frage nach der Existenz der vierten Ruhefrage lässt sich als Frage nach der Möglichkeit für eine Koexistenz der Spezies interpretieren: Diese besteht **höchstens** dann, wenn ebendiese

Ruhelage existiert. Bei genauerer Unterscheidung der Werte für a_{12} und a_{21} ergeben sich im Wesentlichen vier verschiedene Typen von Phasenporträts:

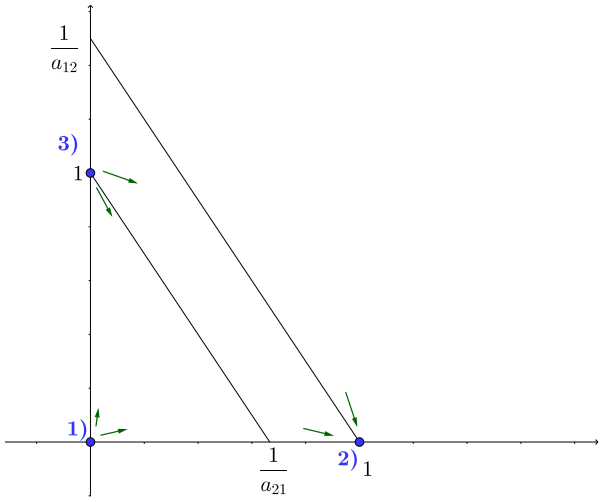
$a_{12} < 1,\ a_{21} < 1$



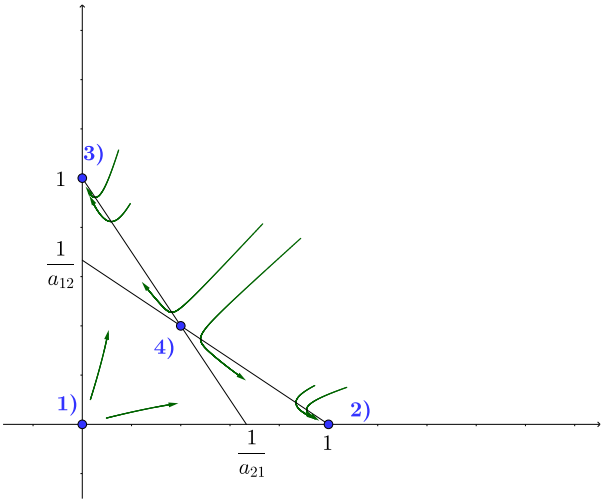
$a_{12} > 1,\ a_{21} < 1$



$a_{12} < 1,\ a_{21} > 1$



$a_{12} > 1,\ a_{21} > 1$



Diese ergeben sich mittels Linearisierung um die Ruhelagen: Für

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{pmatrix} u_1 - u_1^* \\ u_2 - u_2^* \end{pmatrix} \approx A \cdot \begin{pmatrix} u_1 - u_1^* \\ u_2 - u_2^* \end{pmatrix}$$

ist

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \frac{\partial f_1}{\partial u_2} \\ \frac{\partial f_2}{\partial u_1} & \frac{\partial f_2}{\partial u_2} \end{bmatrix}_{(u_1^*, u_2^*)} = \begin{bmatrix} 1 - 2u_1^* - a_{12}u_2^* & -a_{12}u_1^* \\ -\rho a_{21}u_2^* & \rho\left(1 - 2u_2^* - a_{21}u_1^*\right) \end{bmatrix}$$

Bestimmt man das Stabilitätsverhalten über die Eigenwerte, so ergibt sich:

1) $(u_1^*, u_2^*) = (0, 0)$: Wegen $A = \begin{bmatrix} 1 & 0 \\ 0 & \rho \end{bmatrix}$ sind die Eigenwerte $\lambda_1 = 1, \lambda_2 = \rho$. Wegen $\lambda_1,\ \lambda_2 > 0$ ist diese Ruhelage stets instabil.

2) $(u_1^*, u_2^*) = (1, 0)$: In diesem Falle ist $A = \begin{bmatrix} -1 & -a_{12} \\ 0 & \rho\left(1 - a_{21}\right) \end{bmatrix}$. Die Eigenwerte sind somit $\lambda_1 = -1$ und $\lambda_2 = \rho\left(1 - a_{21}\right)$. Wegen $\rho > 0$ ist diese Ruhelage für $a_{21} > 1$ stabil und für $a_{21} < 1$ instabil.

3) $(u_1^*, u_2^*) = (0, 1)$: Aus $A = \begin{bmatrix} 1 - a_{12} & 0 \\ -\rho a_{21} & -\rho \end{bmatrix}$ folgt: Die Eigenwerte sind $\lambda_1 = 1 - a_{12}$ und $\lambda_2 = -\rho$. Wegen $-\rho < 0$ ist die Ruhelage folglich stabil für $a_{12} > 1$ und instabil für $a_{12} < 1$.

4) $(u_1^*, u_2^*) = \left(\frac{1-a_{12}}{1-a_{12}a_{21}}, \frac{1-a_{21}}{1-a_{12}a_{21}}\right)$: Es gilt:

$$A = \frac{1}{1 - a_{12}a_{21}} \begin{bmatrix} a_{12} - 1 & a_{12}\,(a_12 - 1) \\ \rho a_{21}\,(a_{21} - 1) & \rho\,(a_{21-1}) \end{bmatrix}$$

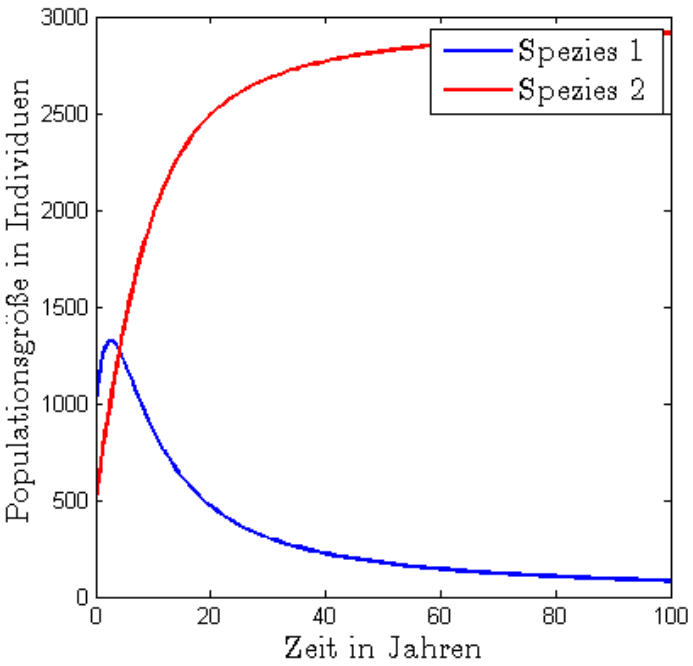
Die Eigenwerte hier sind

$$\lambda_{1,2} = \frac{(a_{12} - 1) + \rho\,(a_{21} - 1)}{2\,(1 - a_{12}a_{21})} \pm \frac{\sqrt{((a_{12} - 1) + \rho\,(a_{21} - 1))^2 - 4\rho\,(1 - a_{12}a_{21})\,(a_{12} - 1)(a_{21} - 1)}}{2\,(1 - a_{12}a_{21})}$$

Für $a_{12} < 1$, $a_{21} < 1$ ist diese Ruhelage stabil, für $a_{12} > 1$, $a_{21} > 1$ ergibt sich ein Sattelpunkt.

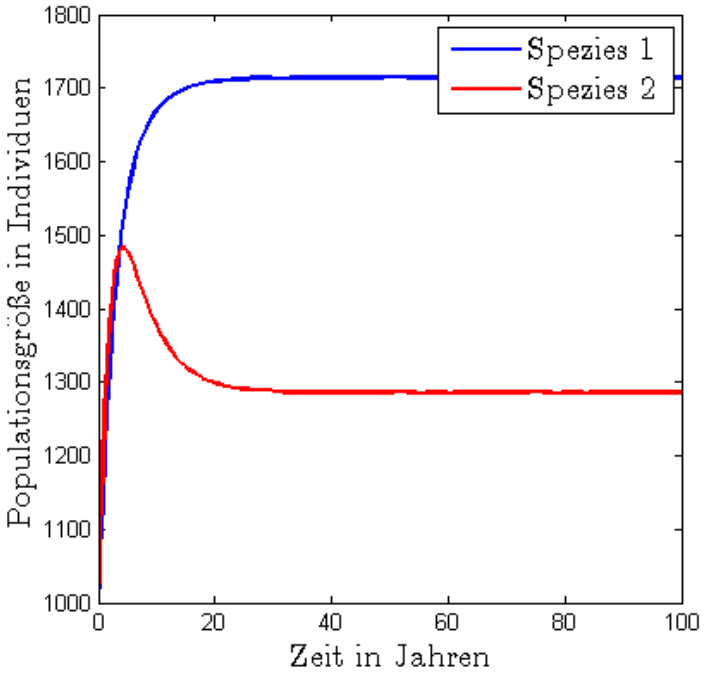
Aus letzterem folgt, dass die beiden Arten auf lange Zeit nur dann koexistieren können, wenn a_{12}, $a_{21} < 1$ gilt, d. h. wenn die gegenseitige Beeinflussung gering genug ist, sodass keine Spezies durch evtl. höhere Kapazität die andere durch zahlenmäßige überlegenheit verdrängt oder aber das Verhältnis der Kapazitäten gering genug ist, sodass eine zu starke Beinflussung der einen Spezies aufgrund der engen Limitation der anderen Spezies nicht zur Ausrottung ersterer führt. In den anderen Fällen stirbt auf lange Sicht eine der beiden Populationen aus, was sich auch in numerischen Simulationen wiederspiegelt.

Simulation 2.4.4:
Die Kapazitätsgrenzen dieser Simulation waren $K_1 = 2000$ und $K_2 = 1000$, als „Beeinflussungsfaktoren“ wurden $b_{12} = b_{21} = 1$ gewählt. Die Wachstumsraten der Populationen waren $r_1 = r_2 = 0,8$. Spezies 1 bestand aus anfangs 1000 Individuen, Spezies zwei aus 500. Simuliert wurde ein Zeitraum von 100 Jahren.



Beobachtung und Interpretation der Ergebnisse: Hier ist $a_{12} = \frac{3}{2}$ sowie $a_{21} = \frac{2}{3}$. Die vierte Ruhelage existiert also nicht. Es ist $a_{12} > 1$, d. h. die Beeinträchtigung des Wachstums von Spezies 1 durch Spezies 2 ist nicht gering genug, um den Einfluss der höheren Aufnahmekapazität von Spezies 2 zu dämpfen, sodass trotz größerer Anfangspopulation und gleichen Bedingungen in allen anderen Faktoren (Wachstumsrate, Beeinflussung) Spezies 1 schlussendlich durch Spezies 2 verdrängt wird.

Simulation 2.4.5:
Die Kapazitätsgrenzen dieser Simulation waren erneut $K_1 = 2000$ und $K_2 = 3000$, als „Beinflussungsfaktoren“ wurden diesmal $b_{12} = \frac{1}{3}$ und $b_{21} = 1$ gewählt. Die Wachstumsrate der Population von Spezies 1 war $r_1 = 0,5$, die von Spezies 2 war $r_2 = 0,8$. Beide Spezies hatten eine Anfangsgröße von 1000 Individuen. Der simulierte Zeitraum umfasste erneut 100 Jahre.



Beobachtung und Interpretation der Ergebnisse: Hier ist $a_{12} = \frac{1}{2}$ und $a_{21} = \frac{2}{3}$. Die gegenseitige Beeinträchtigung reicht also nicht aus, damit eine der beiden Spezies durch höhere Kapazitäten einen Vorteil hätte. Ferner ist die Wachstumsbeinflussung von Spezies 1 seitens Speziens 2 relativ gering, sodass trotz geringerer Wachstumsrate, Kapazität und gleicher Anfangsgröße der Populationen Spezies 1 die dominantere wird. Jedoch reicht der Einfluss auf das Wachstum von Spezies 2 nicht aus, um deren Kapazitätsvorteil auszugleichen und diese vollständig zu verdrängen.

2.4.2 Symbiose-Modell

Häufig beeinflussen sich Populationen jedoch auch positiv. Eine einfaches Modell, um solches symbiotisches Verhalten zu beschreiben ist das Folgende:

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = r_1 N_1 \left[1 - \frac{N_1}{K_1} + b_{12}\frac{N_2}{K_1}\right]$$
$$\frac{\mathrm{d}N_2}{\mathrm{d}t} = r_2 N_2 \left[1 - \frac{N_2}{K_2} + b_{21}\frac{N_1}{K_2}\right]$$

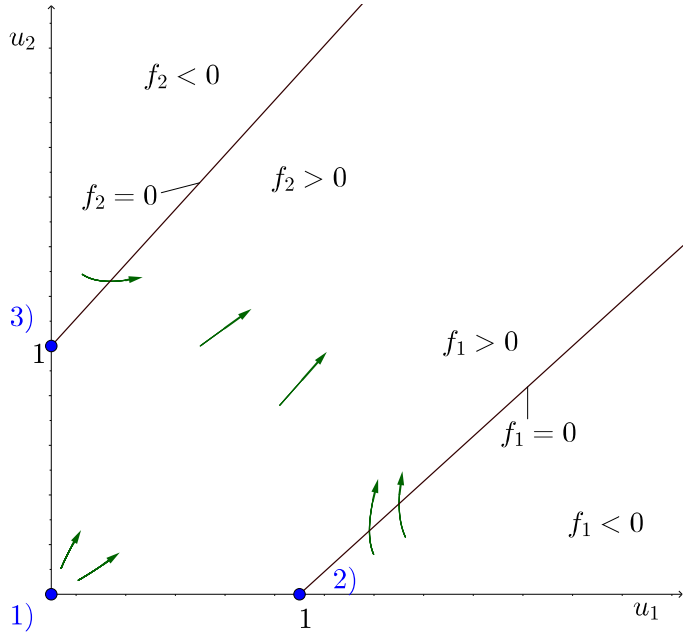
Wählt man die gleichen Substitutionen wie im vorherigen Modell, so erhält man:

$$\frac{\mathrm{d}u_1}{\mathrm{d}\tau} = u_1\left(1 - u_1 + a_{12}u_2\right) =: f_1\left(u_1, u_2\right)$$
$$\frac{\mathrm{d}u_2}{\mathrm{d}\tau} = \rho u_2\left(1 - u_2 + a_{21}u_1\right) =: f_2\left(u_1, u_2\right)$$

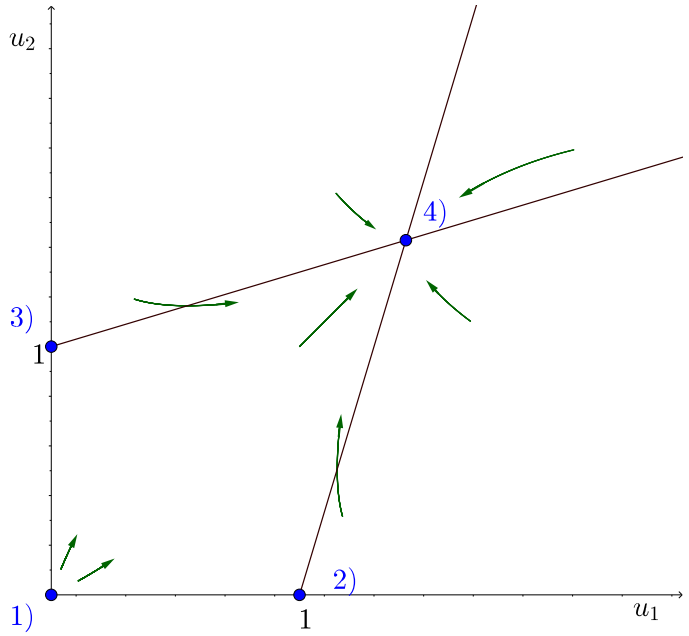
Die Ruhelagen (u_1^*, u_2^*) dieses Modells sind:

1) $(0, 0)$
2) $(1, 0)$
3) $(0, 1)$
4) $\left(\frac{1+a_{12}}{1-a_{12}a_{21}}, \frac{1+a_{21}}{1-a_{12}a_{21}}\right)$

Letztere ergibt jedoch nur dann Sinn, wenn $1-a_{12}a_{21} > 0$. Gilt $a_{12}a_{21} > 1$, so ist unbeschränktes Wachstum zu beobachten:

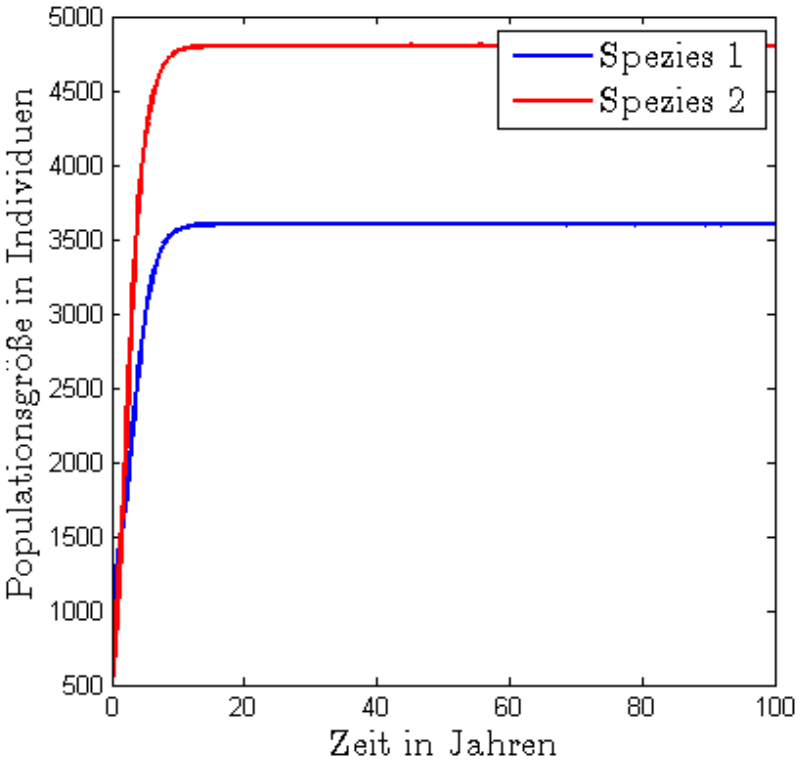


Für $a_{12}a_{21} < 1$ ist das Wachstum beschränkt, die Lösungen streben gegen die vierte Ruhelage:



Der positive gegenseitige Einfluss spiegelt sich hierin insofern wieder, dass in der vierten Ruhelage $u_1^* > 1$ sowie $u_2^* > 1$ gilt, d. h. beide Populationen haben (stabile!) größere Populationen, als ihre jeweiligen Kapazitätsgrenzen zulassen würden. Selbstverständlich lässt sich die Kapazitätsgrenze auch im Falle des beschränkten Wachstums einer Spezies überschreiten, in dem man nur die Anfangsgröße entsprechend setzt. Allerdings nimmt in diesem Falle, bei dem die andere Spezies fehlt, die Anzahl der Individuen mit der Zeit bis zur Kapazitätsgrenze stetig ab, für $t \to \infty$ konvergiert die Anzahl gegen die entsprechende Kapazitätsgrenze.

Simulation 2.4.6:
Die Kapazitätsgrenzen dieser Simulation waren erneut $K_1 = 2000$ und $K_2 = 3000$, die Beinflussungsfaktoren waren $b_{12} = b_{21} = \frac{1}{2}$. Die Wachstumsrate der Population von Spezies 1 war $r_1 = 0,5$, die von Spezies 2 war $r_2 = 0,8$. Beide Spezies hatten eine Anfangsgröße von 1000 Individuen. Auch hier betrug der simulierte Zeitraum 100 Jahre.



Beobachtung und Interpretation der Ergebnisse: Hier ist $a_{12} = \frac{3}{4}$ und $a_{21} = \frac{1}{3}$, weshalb das Wachstum beschränkt ist. Jedoch überschreiten die erreichten Populationsgrößen die Kapazitäten deutlich (Spezies 1 hat einen überhang von ca. 1500 Individuen, Spezies 2 von ca. 1700 Individuen). Hierdurch wird, wie bereits erwähnt, der Vorteil deutlich, den in Symbiose lebende Spezies sich gegenseitig verschaffen.

2.5 Das Lorenz-Modell: Ein Beispiel für Chaos

Das Gebiet der mathematischen Modellierung beschränkt sich, wie bereits an den Beispielen zur Gravitation oder Strömungsmechanik gesehen, nicht nur auf Populationswachstum, sondern ist sehr vielseitig. Mathematische Modelle können beispielsweise auch zur Klimaforschung dienen dienen. Das wohl bekannteste solcher Modelle ist das ***Lorenz-Modell***:

$$\begin{cases} x' = \sigma\,(y - x) \\ y' = rx - y - xz \\ z' = xy - bz \end{cases} \qquad (\sigma,\, r,\, b : \text{ positive Konstanten}) \tag{2.5.1}$$

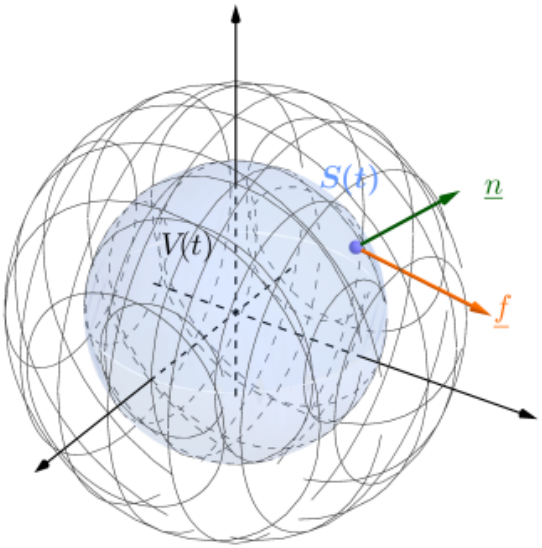
Es wurde 1963 von Edward N. Lorenz veröffentlicht und ist eine stark vereinfachte Beschreibung der Konvektionsströmung. Auch wenn es heute in der Klimaforschung nicht mehr relevant ist, so ist es dennoch insofern von besonderer Bedeutung, als dass Lorenz bei der Untersuchung der drei Gleichungen eine Art von Verhalten der Lösungen entdeckte, welche als ***Chaos*** bezeichnet wird. Dies soll im Folgenden erklärt und nachvollzogen werden (ähnlich zu [11]).
Eine erste Klassifizierung führt zu der Beobachtung, dass das System autonom, nichtlinear und ***symmetrisch*** ist: Ist $(x(t), y(t), z(t))$ eine Lösung, so auch $(-x(t), -y(t), -z(t))$. Darüber hinaus führt es zu einer Volumenkontraktion.

Um dies zu zeigen, sei $\underline{x}' = \underline{f}(\underline{x})$ ein allgemeines, dreidimensionales System, $S(t)$ eine geschlossene Fläche mit Volumen $V(t)$ im Phasenraum. Für einen Zeitpunkt t gilt nach einem kleinen Zeitintervall $\mathrm{d}t$ für das Volumen:

$$V(t+\mathrm{d}t) = V(t) + \int_{S(t)} \underline{f}(\underline{x}) \bullet \underline{n}\mathrm{d}t\mathrm{d}A$$

($\underline{n}$ bezeichnet hierbei den entsprechenden Normalenvektor). Somit gilt

$$V'(t) = \int_{S(t)} \underline{f} \bullet \underline{n}\mathrm{d}A = \int_V \nabla \bullet \underline{f}\mathrm{d}V$$

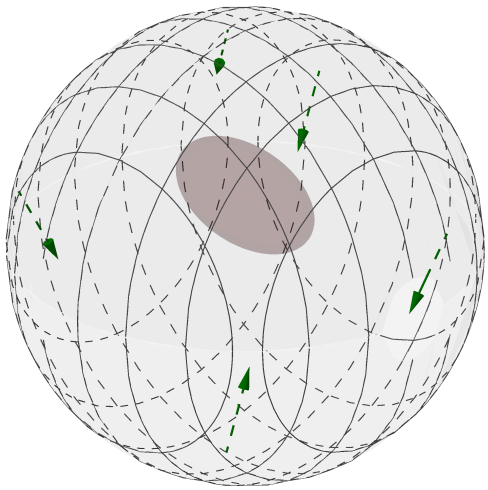


Angewandt auf das Lorenz-System:

$$\begin{aligned}\nabla \bullet \underline{f} &= [\sigma(y-x)]_x + [rx - y - xz]_y + [xy - bz]_z \\ &= -\sigma - 1 - b < 0\end{aligned}$$

Wegen $V'(t) = -(\sigma + 1 + b)\, V(t)$ und somit $V(t) = V(0)e^{-(\sigma+1+b)t}$ folgt die Behauptung.
Ferner sind alle Trajektorien beschränkt: Sei hierzu S_R eine sphärische Oberfläche, gegeben durch $S_R = x^2 + y^2 + (z - r - \sigma)^2 = R^2$. Für eine Trajektorie, welche auf der Oberfläche startet, gilt:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(x(t)^2 + y(t)^2 + (z(t) - r - \sigma)^2\right) &= 2x(t)\,\dot{x}(t) + 2y(t)\,\dot{y}(t) + 2(z(t) - r - \sigma)\,\dot{z}(t) \\ &\overset{(2.5.1)}{=} -2\left[\sigma x(t)^2 + y(t)^2 + b\left(z - \frac{r+\sigma}{2}\right)^2 - \frac{b(r+\sigma)^2}{4}\right]\end{aligned}$$



Kugel mit vollständig enthaltenem Ellipsoid
(dunkel gefärbt)

Ist R so groß, sodass das durch $\sigma x^2 + y^2 + b\left(z - \frac{r+d}{2}\right)^2 = \frac{b(r+\sigma)^2}{4}$ gegebene Ellipsoid komplett enthalten ist , so gilt auf der Fläche $\frac{\mathrm{d}}{\mathrm{d}t}\left[x^2 + y^2 + (z - r - \sigma)^2\right] < 0$, d. h. alle auf dieser Fläche startenden Traketorien werden in die entsprechende Sphäre „hineingezogen“ und können diese auch nicht mehr verlassen. Somit sind, wie behauptet, alle Trajektorien tatsächlich beschränkt (sofern der Anfangspunkt innerhalb der Sphäre liegt).

Im Folgenden soll die Stabilität der Ruhelagen untersucht werden. Die erste Ruhelage ist der Ursprung $(0, 0, 0)$. Weitere Ruhelagen erhält man aus

$$\begin{cases} y - x = 0 \\ rx - y - xz = 0 \\ xy - bz = 0 \end{cases}$$

wegen $x = y$ mithilfe der Gleichungen $(r-1-z)x = 0$ und $x^2 - bz = 0$. Folglich sind diese Ruhelagen $C^- = \left(-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1\right)$ und $C^+ = \left(\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1\right)$ und existieren nur für $r > 1$. Im Folgenden sei, analog zu Kapitel 1.8.3, $\xi = x - \bar{x}$, $\eta = y - \bar{y}$, $\zeta = z - \bar{z}$, wobei $(\bar{x}, \bar{y}, \bar{z})$ die jeweilige Ruhelage ist. Linearisieren ergibt dann:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\xi\\ \eta\\ \zeta\end{pmatrix} = \underbrace{\begin{bmatrix} \frac{\partial}{\partial x}(\sigma(y-x)) & \frac{\partial}{\partial y}(\sigma(y-x)) & \frac{\partial}{\partial z}(\sigma(y-x)) \\ \frac{\partial}{\partial x}(rx-y-xz) & \frac{\partial}{\partial y}(rx-y-xz) & \frac{\partial}{\partial z}(rx-y-xz) \\ \frac{\partial}{\partial x}(xy-bz) & \frac{\partial}{\partial y}(xy-bz) & \frac{\partial}{\partial z}(xy-bz) \end{bmatrix}_{(\bar{x},\bar{y},\bar{z})}}_{=:A} \begin{pmatrix}\xi\\ \eta\\ \zeta\end{pmatrix}$$

$$= \begin{bmatrix} -\sigma & \sigma & 0 \\ r-z & -1 & -x \\ y & x & -b \end{bmatrix}_{(\bar{x},\bar{y},\bar{z})} \begin{pmatrix}\xi\\ \eta\\ \zeta\end{pmatrix}$$

- Ruhelage $(0,0,0)$: Hier ist

  $$A = \begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{bmatrix}$$

  Für eine Lösungen (ξ, η, ζ) des Anfangswertproblems

  $$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\xi\\ \eta\\ \zeta\end{pmatrix} = A\begin{pmatrix}\xi\\ \eta\\ \zeta\end{pmatrix} \\ \begin{pmatrix}\xi\\ \eta\\ \zeta\end{pmatrix}(0) = \begin{pmatrix}\xi_0\\ \eta_0\\ \zeta_0\end{pmatrix} \end{cases}$$

  gilt

  $$\left\| \begin{pmatrix} \xi(t)-0 \\ \eta(t)-0 \\ \zeta(t)-0 \end{pmatrix} \right\|_2 = \sqrt{\xi(t)^2 + \eta(t)^2 + \zeta(t)^2}.$$

  Wegen $\dot{\zeta} = -b\zeta$ hat die entsprechende Komponente der Lösung des linearisierten Systems die Form $\zeta(t) = \zeta_0 \cdot e^{-bt}$, wegen $b > 0$ gilt $|\zeta(t)| = |\zeta_0|\, e^{-bt} \leq |\zeta_0|$ für alle $0 \leq t \leq \infty$. Ferner gilt $\lim\limits_{t\to 0} |\zeta(t)| = 0$, unabhängig von ζ_0. Somit ist das System in der Ruhelage bezüglich der ζ-Komponente stets asymptotisch stabil, für die (asymptotische) Stabilitität des gesamten Systems ist also das Verhalten der Komponenten ξ, η ausschlaggebend. In der (x,y)-Ebene gilt:

  $$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\xi\\ \eta\end{pmatrix} = \underbrace{\begin{bmatrix} -\sigma & \sigma \\ r & -1 \end{bmatrix}}_{=:A'} \begin{pmatrix}\xi\\ \eta\end{pmatrix}$$

  Wegen $\det(A - \lambda I) = \lambda^2 + (1+\sigma)\lambda + \sigma(1-r)$ sind die Eigenwerte $\lambda_{1,2} = \frac{-(1+\sigma) \pm \sqrt{(\sigma-1)^2 + 4\sigma r}}{2}$ Ist $r < 1$, so gilt $(\sigma-1)^2 + 4\sigma r < \sigma^2 - 2\sigma + 1 + 4\sigma = (\sigma+1)^2$, in diesem Falle ist wegen $\lambda_2 < \lambda_1 < 0$ die Ruhelage $(0,0,0)$ ein stabiler Knoten. Für $r > 1$ gelten die umgekehrten Ungleichungen, wegen $\lambda_2 < 0 < \lambda_1$ liegt somit ein Sattelpunkt vor, die Ruhelage ist instabil. *Bemerkung*: Mithilfe der Lyapunov-Funktion $V(x,y,z) = \frac{1}{\sigma}x^2 + y^2 + z^2$ lässt sich im Falle $r < 1$ sogar globale (asymptotische) Stabilität zeigen. Für $f = 1$ ist diese Ruhelage stabil.

- Ruhelagen C^+, C^-: Hier gilt

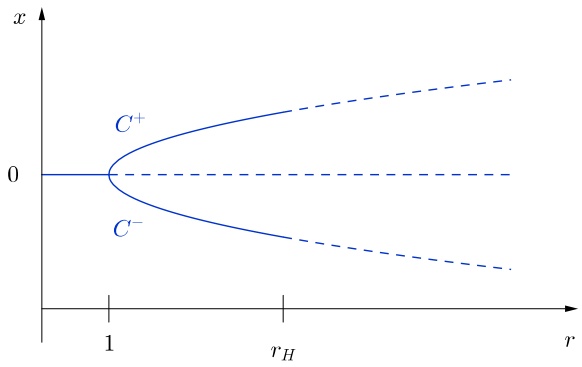
  $$A_{C^+} = \begin{bmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -\sqrt{b\,(r-1)} \\ \sqrt{b\,(r-1)} & \sqrt{b\,(r-1)} & -b \end{bmatrix}$$

sowie

$$A_{C^-} = \begin{bmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & \sqrt{b\,(r-1)} \\ -\sqrt{b\,(r-1)} & -\sqrt{b\,(r-1)} & -b \end{bmatrix}.$$

Wegen $\det(\lambda I - A_{C^+}) = \lambda^3 + (\sigma + b + 1)\,\lambda^2 + b\,(\sigma + r)\,\lambda + 2\sigma b\,(r-1) = \det(\lambda I - A_{C^-})$ haben beide Matrizen die selben Eigenwerte. Es genügt folglich, A_{C^+} zu untersuchen. Für diese lässt sich zeigen, dass für $\sigma - b - 1 \leq 0$ die Ruhelage C^+ stets stabil und für $\sigma - b - 1 > 0$ stabil ist, sofern $1 < r < r_H = \frac{\sigma(\sigma+b+3)}{\sigma-b-1}$. Gilt $\sigma - b - 1 > 0$ und $r > r_H$ ist sie instabil, für $r = r_H$ gilt $\lambda_{1,2} = \pm\sqrt{b(\sigma + r)}i$, $\lambda_3 = -(b + \sigma + 1)$. Sie ist folglich stabil, eine Projektion des Phasenporträts auf die (x, y)-Ebene ergibt das Bild eines Zentrums.

Trägt man die x-Koordinate der Ruhelage gegen r auf, ergibt sich folgendes Gesamtbild für die Stabilität:



Ist für ein r der Punkt $(r, x(r))$ Teil einer durchgezogenen Kurve, so ist die Ruhelage mit der entsprechenden x-Koordinate stabil, ist er Teil einer gestrichelten Kurve, ist diese instabil.

Gilt nun $r > r_H$ so ist wie bereits gesehen, keine der drei Ruhelagen stabil. Numerische Simulationen wie die Folgende weisen jedoch auch keinen Grenzzyklus auf, was insofern seltsam anmutet, da im 2-dimensionalen dies die einzig verbleibende Möglichkeit wäre, nach dem die Ruhelagen allesamt instabil, die Trajektorien jedoch allesamt beschränkt sind.

Simulation 2.5.1: In dieser Simulation wurden die von Lorenz gewählten Parameter verwendet, d. h. $\sigma = 10$, $b = \frac{8}{3}$ und $r = 28$. Die Anfangswerte waren $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ und $\begin{pmatrix} 0,9999998 \\ 1,0000002 \\ 1,0000002 \end{pmatrix}$.
Das verwendete Verfahren war das in Matlab verfügbare Verfahren ode113, d. h. ein Verfahren erster Ordnung mit einem Kontrollverfahren 13. Ordnung. (Der verwendete Code befindet sich in Anhang B).
Da $r_H = \frac{\sigma(\sigma+b+3)}{\sigma-b-1} = \frac{470}{19} < 28$ ist, entspricht diese Parameterwahl dem noch unbekannten Fall. Die Trajektorien weisen zwei wesentliche Eigenschaften auf:

1. Wie vorher errechnet, sind die Trajektorien beschränkt, jedoch - trotz wiederkehrender Verhaltensmuster - stark aperiodisch, wodurch die Existenz eines Grenzzyklus sehr unwahrscheinlich ist.

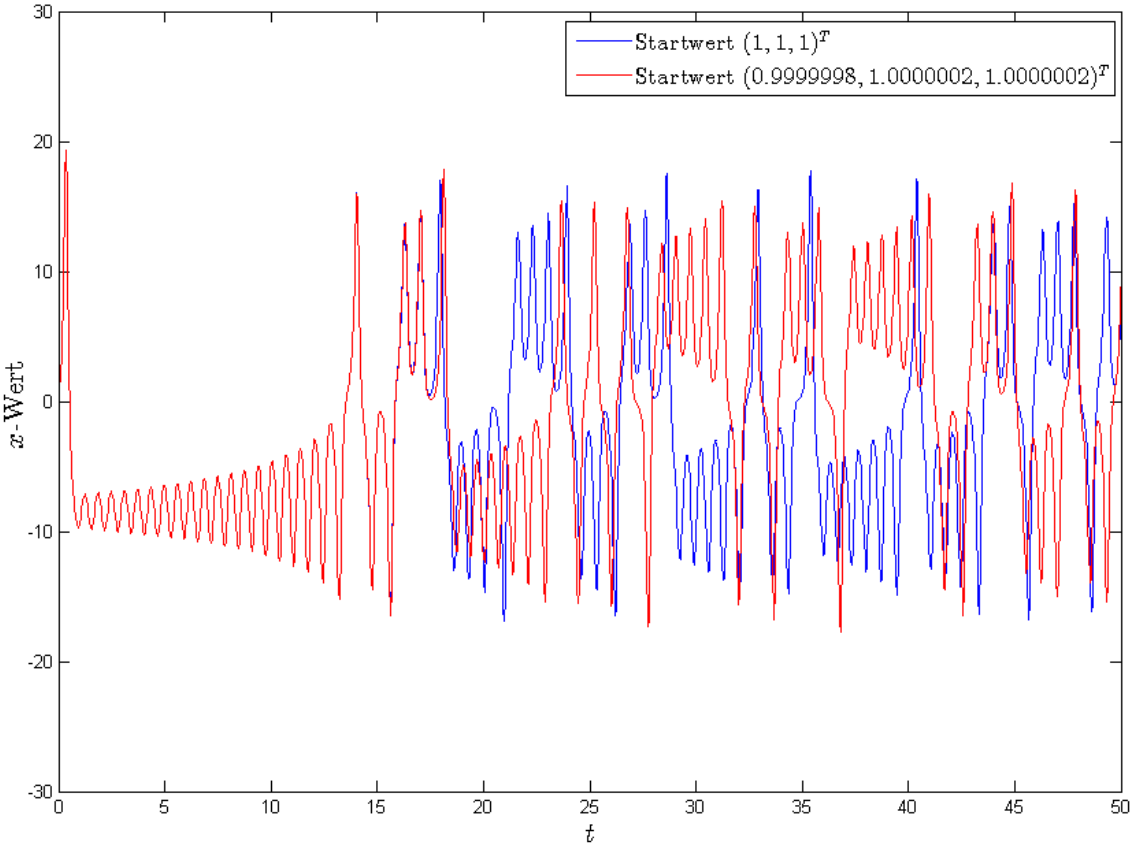


Abbildung 2.5.1: x-Werte der Trajektorien gegen die Zeit aufgetragen. Auf den ersten Blick scheint ein sich wiederholendes Muster zu ergeben, bei genauerem Hinsehen erweisen sich diese als zu unregelmäßig, um von periodischer Oszillation zu sprechen.

2. Der „Verlauf“ der Trajektorie „reagiert“ äußerst sensibel auf Störungen in den Anfangswerten: Die in der Simulation gewählten waren $\begin{pmatrix}1\\1\\1\end{pmatrix}$ und $\begin{pmatrix}0,9999998\\1,0000002\\1,0000002\end{pmatrix}$, somit gilt für den absoluten Fehler

$$\|x_0 - y_0\|_2 = \|x_0 - y_0\|_2 = 2 \cdot 10^{-7} \cdot \left\| \begin{pmatrix}-1\\1\\1\end{pmatrix} \right\|_2 = 2 \cdot 10^{-7} \cdot \sqrt{3} \approx 3,5 \cdot 10^{-7},$$

der relative Fehler ist in beiden Fällen ca. $2 \cdot 10^{-7}$. Somit sind sowohl absoluter als auch relativer Fehler sehr gering. Nichtsdestotrotz weisen, wie in Abbildung 2.5.1 und Abbildung 2.5.2 ersichtlich, die zugehörigen Trajektorien starke Abweichungen voneinander auf.

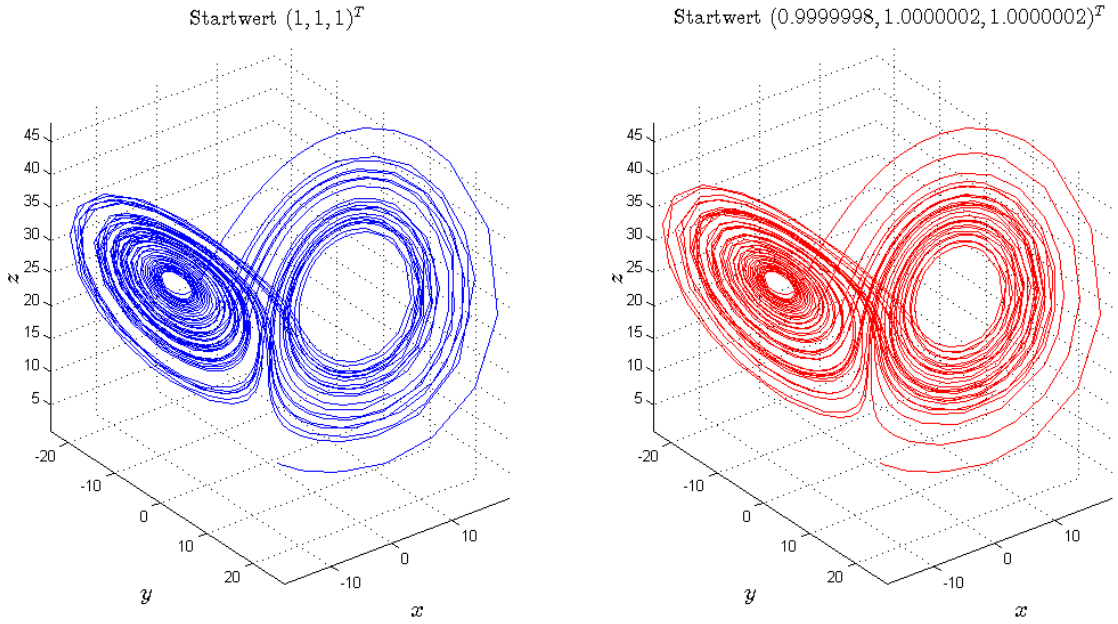


Abbildung 2.5.2: Trajektorien im (dreidimensionalen Phasenraum). U. a. anhand der Gitterlinien der $y - z$-Ebene im Hintergrund lassen sich die Unterschiede mit bloßem Auge erkennen.

Auf der anderen Seite jedoch nähern sich alle Trajektorien trotz aperiodischen Verhaltens und starken Abweichungen einem komplizierten „Gebilde“ bzw. „Muster“ an, welches zu ehren seines Entdeckers ***Lorenz-Attraktor*** genannt wird. Die grundlegende Erkenntnis, dass kleine Änderungen in den Anfangsbedingungen gravierende Auswirkungen auf das Verhalten einzelner Partikel haben können, wurde unter dem Namen ***Schmetterlingseffekt*** bekannt. Und auch wenn keine kanonische Definition von „Chaos“ existiert, so wird die mit diesem Effekt zum Ausdruck gebrachte Sensibilität gegenüber den Anfangsbedingungen allgemein als (ein) wesentliches Merkmal von Chaos angeführt.

2.6 Mechanik

Zum Abschluss des Modellierungskapitels sollen noch Einblicke in die Mechanik gewährt werden. Bei den Resultaten handelt es sich hauptsächlich um eine Zusammenstellung von Ergebnissen aus [13], [14] und [15].

Eines der Grundprinzipien der klassischen Mechanik ist die von Newton entdeckte Gleichheit von Kraft und dem Produkt aus Masse und Beschleunigung. Als Formel geschrieben

$$F = m \cdot a,$$

wobei die Masse m eine reelle Zahl ist und $a \in \mathbb{R}^n$ der Beschleunigungsvektor. (Die Beschleunigung ist eine gerichtet Größe.) F kann hierbei auch als ortsabhängige Funktion aufgefasst werden. In diesem Falle entspricht $F(x)$ der Kraft, die auf einem Körper der Masse m im Punkt x wirkt. Das Vektorfeld F beschreibt hierbei ein ***Kraftfeld***.
Als ***Arbeit*** entlang eines Weges $W = [w_0, w_1]$ wird das Integral $\int_{w_0}^{w_1} F(r)\mathrm{d}r$ bezeichnet. Ein Kraftfeld F heißt ***konservativ***, wenn für jede geschlossene Kurve C die Gleichung

$$\oint_C F(x)\mathrm{d}x = 0$$

gilt. In diesem Falle ist die geleistete Arbeit wegunabhängig. D. h., für zwei Wege W, W', welche die von w_0 nach w_1 gehen, gilt

$$\int_{W'} F(x)\mathrm{d}x = \int_W F(x)\mathrm{d}r$$

Somit ist für beliebige $V_0 \in \mathbb{R}, x_0 \in \mathbb{R}^n$ durch

$$V(x) \coloneqq V_0 - \int_{x_0}^{x} F(r)\mathrm{d}r$$

eine skalarwertige Funktion gegeben. Durch Differentiation folgt

$$F = -\frac{\partial V}{\partial x}. \tag{2.6.1}$$

Für ein Objekt der Masse m, welches sich mit Geschwindigkeit $v(t)$ bewegt, ist die ***kinetische Energie*** zum Zeitpunkt t durch $T(v) = \frac{1}{2} \cdot \|v\|_2^2$ gegeben. Bewegt es sich hierbei durch ein Kraftfeld entlang des Weges $W = [w_0, w_1]$, so ist der Ort x zeitabhängig. Mittels der Beziehung $a = \ddot{x}$ lässt sich die Kraft ebenfalls als zeitabhängige Funktion $F(x(t))$ modellieren. Aus $w_0 = x(t_0)$, $w_1 = x(t_1)$ sowie der Substitutionsregel ergibt sich für den Weg die Gleichung

$$\int_{x(t_0)}^{x(t_1)} F(r)\mathrm{d}r = \int_{t_0}^{t_1} F(x(t)) \cdot v(t)\mathrm{d}t,$$

wobei $v(t) = \dot{x}(t)$ die Geschwindigkeit bezeichnet. Hieraus folgt wegen $F(x(t)) = m \cdot a(t) = m \cdot \frac{\mathrm{d}}{\mathrm{d}t}(v(t))$ die Beziehung

$$\int_{x(t_0)}^{x(t_1)} F(r)\mathrm{d}r = \frac{m}{2} \cdot \int_{t_0}^{t_1} \frac{\mathrm{d}}{\mathrm{d}t}(v(t)^T \cdot v(t))\mathrm{d}t = \frac{m}{2}\left(\|v(t_2)\|_2^2 - \|v(t_1)\|_2^2\right)$$

Die entlang des Weges verrichtete Arbeit entspricht also der Differenz der kinetischen Energien zu End- und Anfangszeitpunkt. Andererseits ergibt sich aus (2.6.1) die Gleichung

$$\int_{w_0}^{w_1} F(x)\mathrm{d}x = -\int_{w_0}^{w_1} \frac{\partial}{\partial x} V(x)\mathrm{d}x = -V(w_2) + V(w_1).$$

Es folgt

$$0 = \int_{w_0}^{w_1} F(x)\mathrm{d}x - \int_{w_0}^{w_1} F(x)\mathrm{d}x = T(t_2) - T(t_1) + V(x(t_2)) - V(x(t_1))$$

beziehungsweise, durch Addition von $V(x(t_1))), T(t_1)$ auf beiden Seiten

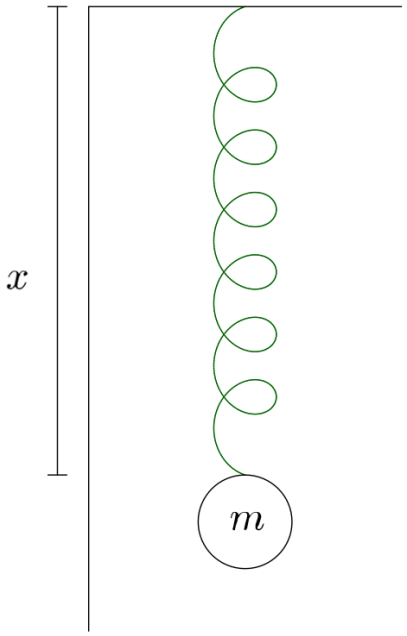
$$T + V \equiv \text{const.}$$

In konservativen Systemen gilt also der ***Energieerhaltungssatz***, $\frac{\mathrm{d}}{\mathrm{d}t}(T+V) = 0$.

Beispiel (Feder mit Gewicht):

An einer Feder mit Federhärte k ist ein Gewicht der Masse m aufgehängt. x beschreibe den Abstand des Gewichts zur Aufhängepunkt der Feder. Zur Vereinfachung finde dies in einer Umgebung mit vernachlässigbarer Gravitation und Reibung statt. Das harmonische Potenzial ist gegeben durch

$$V(x) = \frac{k}{2}x^2$$



Somit gilt für die Kraft, welche die Feder auf das Gewicht ausübt:

$$F = -\frac{\partial V}{\partial x} = -kx$$

und somit

$$m\ddot{x} = -kx$$

Dies entspricht der Differentialgleichung

$$\ddot{x} + \omega^2 x = 0$$

mit $\omega = \sqrt{\frac{k}{m}}$. Löst man dies mithilfe des Eulerschen Ansatzes $x = e^{\lambda t}$, so folgt mittels der Bedingung $\lambda^2 + \omega^2 = 0$ $\lambda = \pm\omega i$. Reellwertige Lösungen sind folglich sin und cos, d. h. die Lösungen oszillieren periodisch mit Frequenz ω.

Die Newtonsche Mechanik, so einfach sie auf den ersten Blick wirkt, hat jedoch verschiedene Nachteile. Einer davon ist, dass die Form der Gleichungen sehr stark von der Wahl des Koordinatensystems abhängt. Einen allgemeinerer Ansatz stellt der ***Lagrange-Formalismus*** dar. Hierzu sei zuallererst die Lagrangefunktion

$$L\left(x(t), \dot{x}(t), t\right) = T\left(x(t), \dot{x}(t), t\right) - V\left(x(t), \dot{x}(t)\right)$$

definiert. Mittels dieser Definition lassen sich Bewegungen mithilfe des ***Hamiltonsches Prinzips*** beschreiben. (Dieses wird auch Prinzip der kleinsten/stationären Wirkung, engl. principle of least/stationary action, genannt.) Bewegt sich ein Objekt zwischen zwei Zeitpunkten t_A und t_B von A nach B, gibt es hierfür verschiedene Pfade. Jedem dieser Pfade lässt sich hierbei eine Wirkung

$$\tilde{A} = \int_{t_A}^{t_B} L\left(x(t), \dot{x}(t), t\right) \mathrm{d}t$$

zuweisen. Das genannte Prinzip besagt nun, dass von allen diesen Pfaden derjenige genommen wird, entlang dem die Wirkung $\tilde{A}$ stationär ist, d. h. die erste Variation des Wirkungsfunktional

$\int_{t_A}^{t_B} L\left(x(t), \dot{x}(t)\right) \mathrm{d}t$ in der Lösung ist 0. Bei der ersten Variation handelt es sich hierbei um eine Verallgemeinerung der Richtungsableitung für Funktionale (linearen, stetigen Abbildungen auf Funktionenräumen). Genaueres wird in Kapitel 3.3 erklärt, an dieser Stelle folgen nur zwei Beispiele, welche die Berechnung der ersten Variation veranschaulichen.

Beispiel 2.6.1 (Minimierung eines Funktionals):

Gegeben sei das Funktional

$$F : C([0,T]) \to \mathbb{R}, F(y) = \int_0^T (y(t) - c)^2 \, \mathrm{d}t,$$

welches minimiert werden soll. Sei hierzu $y \in C([0,T]), \delta y$ eine zulässige Variation, d. h. $y(0) + \alpha\delta y(0) = y(0)$, $y(T) + \alpha\delta y(T) = y(T)$ sowie $y + \alpha\delta y \in C([0,T])$ für alle α (genaueres hierzu folgt in den späteren Kapiteln). Somit gilt für $\alpha > 0$:

$$\frac{F\left(y + \alpha\delta y\right) - F\left(b\right)}{\alpha} = \frac{1}{\alpha}\left[\int_0^T \left(y\left(t\right) + \alpha\delta y\left(t\right) - c\right)^2 \mathrm{d}t - \int_0^T \left(y\left(t\right) - c\right)^2 \mathrm{d}t\right] =$$

$$= \frac{1}{\alpha}\left[\int_0^T \left(y\left(t\right) - c\right)^2 \mathrm{d}t + \alpha^2 \int_0^T \left(\delta y\left(t\right)\right)^2 \mathrm{d}t + 2\alpha \int_0^T \delta y\left(t\right)\left(y\left(t\right) - c\right) \mathrm{d}t - \int_0^T \left(y\left(t\right) - c\right)^2 \mathrm{d}t\right] =$$

$$= \left[\alpha \int_0^T \left(\delta y\left(t\right)\right)^2 \mathrm{d}t + 2 \int_0^T \delta y\left(t\right)\left(y\left(t\right) - c\right) \mathrm{d}t\right]$$

Grenzwertbildung liefert:

$$\lim_{\alpha \to 0} \frac{F\left(y + \alpha\delta y\right) - F\left(b\right)}{\alpha} = \int_0^T 2\delta y\left(t\right)\left(y\left(t\right) - c\right) \mathrm{d}t$$

Dies ist die erste Variation von F nach y. Das Funktional kann nur durch diejenigen Funktionen y minimiert werden, für welche $\int_0^T 2\delta y\left(t\right)\left(y\left(t\right) - c\right) \mathrm{d}t = 0$ für alle (zulässigen!) Variationen δy gilt. Da $\delta y(t)$ abgesehen von den Anfangsbedingungen beliebig gewählt werden darf, lässt sich $y(t) \equiv c$ folgern: Aus der Positivität des Integrals ergäbe sich für $y(t) \neq c$, $\delta(t) \coloneqq \sqrt{\frac{T^2}{4} - (t - \frac{T}{2})^2}(y(t) - c)$ ein Widerspruch. Wegen $F(\tilde{y}) \geq 0 \forall \tilde{y} \in C\left([0,T]\right)$ ist unmittelbar klar, dass $y(t) \equiv c$ tatsächlich ein Minimum (und nicht nur stationärer Punkt) ist.

Bemerkung: *Ist $F : H \to \mathbb{R}$ ein lineares Funktional auf einem reellen Hilbertraum H, so existiert ein $r \in H$, sodass*

$$F(y) = \langle r, y \rangle_H \quad \forall y \in H$$

gilt. Versieht man den Funktionenraum $C\left([0,T]\right)$ mit dem Skalarprodukt $\langle u, v \rangle = \int_0^T u\left(t\right) v\left(t\right) \mathrm{d}t$, so existiert für ein darauf definiertes, differenzierbares Funktional F zu jedem $y \in C\left([0,T]\right)$ eine Funktion $\nabla F \in C\left([0,1]\right)$ mit

$$\frac{\mathrm{d}F(y)}{\mathrm{d}y} \circ v \coloneqq \langle \nabla F, v \rangle_H \quad \forall v \in C\left([0,T]\right)$$

Diese wird Gradient genannt. (Die genauere Definition von Differenzierbarkeit von Funktionalen folgt in den Kapiteln zu Optimierung.) Im vorherigen Beispiel ist $\nabla F(y) = 2(y - c)$.

Beispiel 2.6.2:

Gegeben sei das Funktional

$$F(y) = \int_0^T \left(y(t)^2 + 2y(t)\right) \mathrm{d}t$$

Differenzieren liefert:

$$\lim_{\alpha \to 0} \frac{F(y + \alpha\delta y) - F(y)}{\alpha} = \int_0^T (2y(t) + 2)\, \delta y(t)\, \mathrm{d}t$$

Somit lautet der Gradient $\nabla F(y) = 2y(t) + 2$

Nun soll erneut die Lagrangefunktion betrachtet werden. Um diese von der bildlichen Vorstellung in kartesischen Koordinaten zu lösen, sei diese im Folgenden von den verallgemeinerten Koordinaten q und $\dot{q}$ abhängig. Das Wirkungsfunktional lautet also

$$\tilde{A} : X \to \mathbb{R}, F(q) = \int_0^T L(q, \dot{q})\, \mathrm{d}t,$$

mit $X = \{q \in C^1([0,T]), q(0) = q_0, q(T) = q_T\}$. Gesucht ist nun $\bar{q} \in X$ mit $\frac{\mathrm{d}}{\mathrm{d}q} F(\bar{q}) \circ \delta q = 0$ für alle zulässigen Variationen δq, d. h. $\delta q \in C^1([0,T])$, $\delta q(0) = \delta q(T) = 0$, $q + \alpha\delta q \in X \forall \alpha$. Differentiation liefert:

$$\lim_{\alpha \to 0} \frac{F(q + \alpha\delta q) - F(q)}{\alpha} = \lim_{\alpha \to 0} \frac{1}{\alpha} \left[\int_0^T L(q + \alpha\delta q, \dot{q} + \alpha\delta\dot{q})\, \mathrm{d}t - \int_0^T L(q, \dot{q})\, \mathrm{d}t \right] =$$

$$= \lim_{\alpha \to 0} \frac{1}{\alpha} \left[\int_0^T \left(L(q, \dot{q}) + \frac{\partial L}{\partial q}(q, \dot{q})(\alpha\delta q) + \frac{\partial L}{\partial \dot{q}}(q, \dot{q})(\alpha\delta\dot{q}) + O(\alpha^2) \right) \mathrm{d}t - \int_0^T L(q, \dot{q})\, \mathrm{d}t \right] =$$

$$= \int_0^T \left(\frac{\partial L}{\partial q}(q, \dot{q})\, \delta q + \frac{\partial L}{\partial \dot{q}}(q, \dot{q})\, \delta\dot{q} \right) \mathrm{d}t =$$

$$\overset{\text{part. Int.}}{=} \int_0^T \left(\frac{\partial L}{\partial q}(q, \dot{q})\, \delta q - \left(\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{q}}(q, \dot{q}) \right) \delta q \right) \mathrm{d}t + \left[\frac{\partial L}{\partial \dot{q}}(q, \dot{q})\, \delta q \right]_0^T =$$

$$= \int_0^T \left(\frac{\partial L}{\partial q}(q, \dot{q}) - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{q}}(q, \dot{q}) \right) \delta q \mathrm{d}t = \langle \nabla F(q), \delta q \rangle$$

Aus der Bedingung $\langle \nabla F(q), \delta q \rangle = 0$ ergeben sich schließlich die ***Euler-Lagrange-Gleichungen***:

$$\frac{\partial L}{\partial q}(q, \dot{q}) - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{q}}(q, \dot{q}) = 0 \tag{2.6.2}$$

Im Falle der Feder ist $L(q, \dot{q}, t) = T(q, \dot{q}, t) - V(q, \dot{q}, t)$. mit $T(q, \dot{q}, t) = T(\dot{q}) = \frac{m}{2}(\dot{q})^2$. Hieraus folgt $-\frac{\partial V}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}(m\dot{q}) = 0$. Dies ist äquivalent zu $m\ddot{q} = -\frac{\partial V}{\partial q} = F$.

2.7 Hamiltonsche Mechanik

Eine andere Möglichkeit, die klassische Mechanik allgemeiner zu formulieren, besteht in der ***Hamiltonfunktion*** (engl. ***Hamiltonian***): Für die verallgemeinerte Koordinate $q = (q_1, \ldots, q_d)^T$ ist der ***verallgemeinerte (konjugierte) Impuls*** p gegeben durch $p_k = \frac{\partial}{\partial \dot{q}} L(q, \dot{q})$, $k = 1, \ldots, d$. Besagte Funktion ist dann definiert durch

$$H(q, p) = p^T \dot{q} - L(q, \dot{q})$$

mit $p = (p_1, \ldots, p_d)^T$.

Bemerkung: *Um H in Abhängigkeit von p schreiben zu können, sei hier und im Folgenden stets angenommen, dass die durch $p_k = \frac{\partial L(q,\dot{q})}{\partial \dot{q}_i}, k = 1, ..., d$ beschriebene Abbildung $\dot{q}_k \mapsto p_k$, Legendre-Transformation genannt, für jedes q ein Diffeomorphismus ist. (Damit lässt sich $\dot{q}$ auch als Funktion $\dot{q}(q, p)$ auffassen.)*

Satz 2.7.1:
Die Euler-Lagrange-Gleichungen in (2.6.2) sind äquivalent zu folgender Hamilton-Gleichung:

$$\begin{aligned} \dot{q}_k &= \frac{\partial H}{\partial p_k}(q, p) \\ \dot{p}_k &= -\frac{\partial H}{\partial q_k}(q, p) \end{aligned}, \quad k = 1, \ldots, d$$

Bevor dies bewiesen wird, sollen jedoch einige Eigenschaften des Hamilton-Systems aufgelistet werden:

1. Die Ruhelagen (a_q, a_p) entsprechen den kritischen Punkten der Hamiltonfunktion. (Eine Ruhelage (a_q, a_p) wird hierbei ***nicht-degeneriert*** genannt, falls $\det\left(\frac{\partial^2}{\partial (q,p)^2} H\left(a_q, a_p\right)\right) \neq 0$.)

2. Hängt die Hamiltonfunktion nicht explizit von der Zeit ab, dann erhält das Hamilton-System die Energie (ist konservativ), d. h. $H\left(q, p\right)$ bleibt konstant entlang der Trajektorien. Also liegen die Trajektorien des Systems in der Niveaumenge

$$H\left(q, p\right) = \text{const.}$$

3. Falls die zweite Ableitung von H ausgewertet an einer Ruhelage nur Eigenwerte mit positivem Realteil hat, dann ist die Ruhelage (Lyapunov)-stabil.

4. Viele mechanische Systeme werden durch eine Hamiltonfunktion der Form

$$H\left(q, p\right) = \frac{1}{2} \sum_{i=1}^{d} \frac{p_i^2}{m} + V\left(q\right)$$

beschrieben, wobei sich das Kraftfeld aus einem Potenzial ableitet, d. h. $F = -\frac{\partial}{\partial q} V(q)$.

Beweis: Für $k = 1, \ldots, d$ gilt zum einen:

$$\frac{\partial}{\partial p_k} H(q,p) = \frac{\partial}{\partial p_k} \sum_{i=1}^{d} p_i \dot{q}_i - \frac{\partial}{\partial p_k} L(q,\dot{q}) =$$

$$= \dot{q}_k + \underbrace{\sum_{i=1}^{d} p_i \frac{\partial}{\partial p_k} \dot{q}_i}_{= p^T \frac{\partial}{\partial p_k} \dot{q}} - \underbrace{\frac{\partial}{\partial \dot{q}} L(q,\dot{q})}_{= p^T} \frac{\partial}{\partial p_k} \dot{q} = \dot{q}_k$$

Zum anderen gilt:

$$\frac{\partial}{\partial q_k} H(q,p) = p^T \frac{\partial \dot{q}}{\partial q_k} - \frac{\partial}{\partial q} L(q,\dot{q}) - \frac{\partial}{\partial \dot{q}} L(q,\dot{q}) \frac{\partial \dot{q}}{\partial q} =$$

$$= -\frac{\partial}{\partial q} L(q,\dot{q}) = -\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial}{\partial \dot{q}} L(q,\dot{q}) = -\frac{\mathrm{d}}{\mathrm{d}t} p$$

□

Hieraus folgt Eigenschaft 2: Es gilt

$$\frac{\mathrm{d}}{\mathrm{d}t} H\,(q,p) = \frac{\partial}{\partial q} H(q,p)\dot{q} + \frac{\partial}{\partial p} H(q,p)\dot{p} =$$

$$= \frac{\partial}{\partial q} H(q,p) \left(-\frac{\partial}{\partial p} H(q,p) \right) + \frac{\partial}{\partial p} H(q,p) \cdot \frac{\partial}{\partial q} H(q,p) = 0$$

Beispiel 2.7.2:

Gegeben seien N Partikel mit Massen m_i, welche sich entfernungsabhängig beeinflussen. Für $1 \leq i \leq N$ bezeichne $q_i \in \mathbb{R}^3$ den Ort, $p_i \in \mathbb{R}^3$ den Impuls und $V_{ij} = V_{ij}(\|q_i - q_j\|)$ für $1 \leq j < i \leq N$ das Interaktionspotenzial zwischen dem i-ten und dem j-ten Partikel. Dies ist ein Hamilton-System mit

$$\begin{cases} H(q,p) = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{m_i} p_i^T p_i + \sum_{i=2}^{N} \sum_{j=1}^{i-1} V_{ij}(\|q_i - q_j\|) \\ \dot{p}_k = -\frac{\partial}{\partial q_k} H(q,p) \\ \quad = -\sum_{j=1}^{k-1} \frac{\partial V_{kj}\left(\|q_k - q_j\|\right)}{\partial \|q_k - q_j\|} \cdot \frac{(q_k - q_j)}{\|q_i - q_j\|} + \sum_{i=k+1}^{N} \frac{\partial V_{ik}\left(\|q_i - q_k\|\right)}{\partial \|q_i - q_k\|} \cdot \frac{(q_i - q_k)}{\|q_i - q_k\|} \\ \dot{q}_i = \frac{p_i}{m_i} \end{cases}$$

Betrachtet man nun den Gesamtimpuls $P = \sum_{i=1}^{N} p_i$, und verwendet kurz $d_{i,j} = \|q_i - q_j\|$, so ergibt sich wegen $\frac{\partial V_{kj}(d_{kj})}{\partial d_{kj}} = \frac{\partial V_{ii}(d_{jj})}{\partial d_{ji}}, d_{ik} = d_{ki}$

$$\dot{p} = \sum_{i=k}^{N} \dot{p}_k = -\sum_{k=1}^{N} \sum_{j=1}^{k-1} \frac{\partial V_{kj}(d_{kj})}{d_{kj}} \frac{q_k - q_j}{d_{kj}} + \sum_{k=1}^{N} \sum_{i=k+1}^{N} \frac{\partial V_{ik}(d_{ik})}{d_{ik}} \frac{q_k - q_i}{d_{ik}} =$$

$$= -\sum_{k=2}^{N} \sum_{j=1}^{k-1} \frac{\partial V_{kj}(d_{kj})}{d_{kj}} \frac{q_k - q_j}{d_{kj}} + \sum_{i=2}^{N} \sum_{k=1}^{N-1} \frac{\partial V_{ik}(d_{ik})}{d_{ik}} \frac{q_i - q_k}{d_{ik}} = 0$$

Andererseits ergibt sich für den Gesamtdrehimpuls $L := \sum_{k=1}^{N} q_k \times p_k$:

$$\begin{aligned}\dot{L} &= \sum_{k=1}^{N} \frac{\mathrm{d}}{\mathrm{d}t}(q_k \times p_k) = \sum_{k=1}^{N} (\dot{q}_k \times p_k + q_k \times \dot{p}_k) = \\ &= \sum_{k=1}^{N} \left(\frac{1}{m_k} p_k \times p_k + q_k \times \dot{p}_k \right) = \sum_{k=1}^{N} (q_k \times \dot{p}_k) = \\ &= -\sum_{k=2}^{N} \sum_{j=1}^{k-1} \frac{\partial V_{kj}(d_{kj})}{d_{kj}} \frac{-q_k \times q_j}{d_{kj}} + \sum_{i=2}^{N} \sum_{k=1}^{N-1} \frac{\partial V_{ik}(d_{ik})}{d_{ik}} \frac{q_k \times q_i}{d_{ik}} = 0\end{aligned}$$

Somit bleiben beide Gesamtimpulse erhalten!

Beispiel 2.7.3 (Zweikörperproblem):

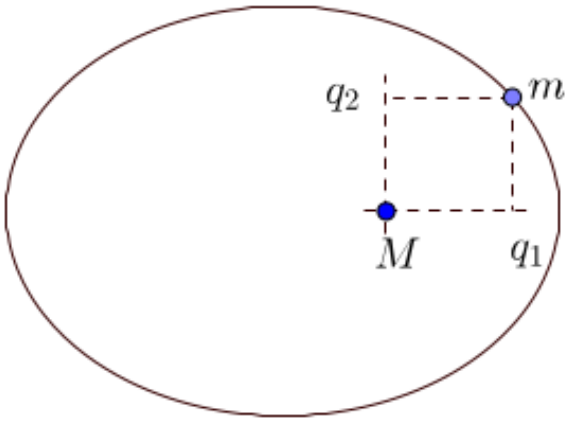
Betrachtet werden zwei Körper (Sonne, Planet) mit Massen $M \gg m$, welche sich gegenseitig anziehen. Wählt man den Ort der Sonne als Ursprung des Koordinatensystems, so genügen für die Beschreibung der Position des zweiten Körpers zweidimensionale Koordinaten $q = (q_1, q_2)$. Die entsprechende Hamiltonfunktion lautet bei geeigneter Skalierung

$$H(q_1, q_2, p_1, p_2) = \frac{1}{2}\left(p_1^2 + p_2^2\right) - \frac{1}{\sqrt{q_1^2 + q_2^2}}$$

Das entsprechende Hamilton-System lautet:

$$\begin{aligned}\dot{q}_i &= \frac{\partial}{\partial p_i} H(q_1, q_2, p_1, p_2) = p_i \\ \dot{p}_i &= -\frac{\partial}{\partial q_i} H(q_1, q_2, p_1, p_2) = -\frac{q_i}{\sqrt{q_1^2 + q_2^2}^3}\end{aligned}, \quad i = 1, 2$$

Die Trajektorie der Ortskoordinate ist eine Ellipse:



Hierbei entsprechen die Gleichungen für $\dot{p}_1$, $\dot{p}_2$ dem Newtonschen Gravitationsgesetz. Gemäß diesem ist die auf den Planeten wirkende Kraft gegeben durch

$$\begin{pmatrix} F_x \\ F_y \end{pmatrix} = G \cdot M \cdot m \begin{pmatrix} \frac{x_S - x_P}{\sqrt{(x_S - x_P)^2 + (y_S - y_P)^2}^2} \\ \frac{y_S - y_P}{\sqrt{(x_S - x_P)^2 + (y_S - y_P)^2}^3} \end{pmatrix},$$

wobei (x_S, y_S) die Koordinaten der Sonne und (x_P, y_P) die Koordinaten des Planeten sind. Mit $x_S = y_S = 0$ ergeben sich aus den Newtonschen Gesetzen, $F_x = m_1 \cdot \ddot{x}_S$, und der Identität $\ddot{q}_i = \dot{p}_i$ durch geeignete Normierung obige Gleichungen.

Beispiel 2.7.4 (Harmonischer Oszillator):

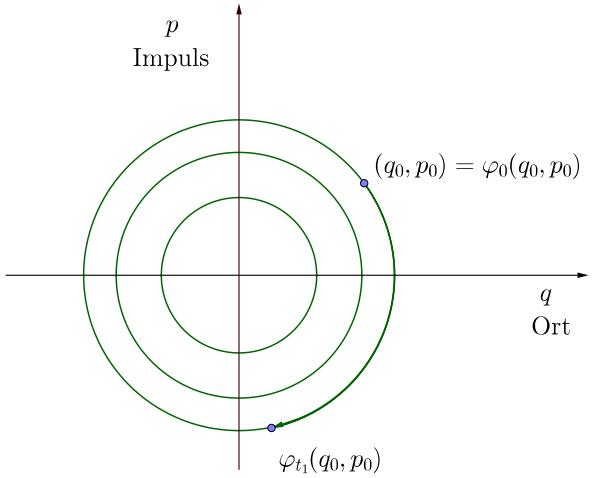


Abbildung 2.7.1: Exemplarischer Phasenraum eines harmonischen Oszillators

Die Hamiltonfunktion eines harmonischen Oszillators wie im Eingangsbeispiel zum Federpendel ist gegeben durch

$$H(q,p) = \frac{1}{2}\frac{p^2}{m} + \frac{1}{2}Dq^2.$$

Dies entspricht dem Hamilton-System

$$\begin{pmatrix}\dot{q}\\ \dot{p}\end{pmatrix} = \begin{pmatrix}\frac{p}{m}\\ -Dq\end{pmatrix} = \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{pmatrix}Dq\\ \frac{p}{m}\end{pmatrix}$$

Der Fluss $\varphi_t(q_0, p_0)$ gibt hierbei nicht nur eine reine Ortsbewegung an, sondern auch den Impuls.

Abschließend sollen noch zwei Modifizierungen der Hamiltonfunktion betrachtet werden:

1) Reibungskräfte: Häufig sind mechanische Systeme geschwindigkeitsabhängigen Reibungskräften unterworfen. Diese werden in der Lagrangefunktion jedoch nicht berücksichtigt, da sie sich nicht aus einem Potenzial ableiten. In diesem Falle erweitert man die Euler-Lagrange-Gleichungen mittels der sogenannten ***(Rayleighschen) Dissipationsfunktion***

$$D(q,\dot{q},t) = \frac{1}{2}\sum_{k=1}^{d}\frac{\gamma_k}{2}\dot{x}_k\left(q,\dot{q},t\right)^T\dot{x}_k\left(q,\dot{q},t\right)$$

(γ_k beschreibt die Proportinalitätskonstante der Reibung zur Geschwindigkeit in kartesischen Koordinaten $\dot{x}_k$.)
Die Euler-Lagrange-Gleichungen lauten dann:

$$\frac{\partial}{\partial q_k}L(q,\dot{q},t) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial \dot{q}_k}L(q,\dot{q},t) - \frac{\partial}{\partial \dot{q}_k}D(q,\dot{q},t) = 0, \qquad k = 1,...,d$$

2) Zwangsbedingungen: Das System ist sogenannten holonomen ***Zwangsbedingungen*** unterworfen, dass heißt, die kartesischen Koordinaten genügen Gleichungen der Form

$$e_i(x_1,...,x_d,t) = 0, \qquad i = 1,...,m$$

Dies ist z. B. bei einem Fadenpendel mit fester Schnurlänge der Fall, das Gewicht bewegt sich im dreidimensionalen Fall in einer Ebene auf einer Kreisbahn. In solch einem Fall nennt man die Differenz $3d - m$ die ***Anzahl der Freiheitsgrade***. Schreibt man die kartesischen Koordinaten in Abhängigkeit von den generalisierten Koordinaten, so werden Zwangskräfte durch einen Lagrangemultiplikatorenansatz in der Lagrange- bzw.

Hamiltonfunktion berücksichtigt:

$$\begin{aligned} L_c(q,\dot{q},t,\lambda) &= L(q,\dot{q},t) + \sum_{i=1}^{m} \lambda_i e_i(q,\dot{q},t) \\ H_c(q,\dot{q},t,\lambda) &= H(q,\dot{q},t) - \sum_{i=1}^{m} \lambda_i e_i(q,\dot{q},t) \end{aligned} \tag{2.7.1}$$

Analog zu obigen ergibt sich

$$\begin{aligned} \dot{q_k} &= \frac{\partial H_c}{p_k} = \frac{\partial H}{\partial p_k} - \sum_{i=1}^{m} \lambda_i \frac{\partial e_i}{\partial p_k} \\ \dot{p_k} &= -\frac{\partial H_c}{p_k} = -\frac{\partial H}{\partial p_k} + \sum_{i=1}^{m} \lambda_i \frac{\partial e_i}{\partial p_k} \end{aligned}$$

Im Falle von Zwangsbedingungen erweist es sich als praktisch, die generalisierten Koordinaten q_d so zu wählen, dass die Zwangsbedingungen für alle $x_k(q_1, ..., q_d)$ automatisch erfüllt sind. Wählt man für das Fadenpendel mit fester Länge l den Aufhängepunkt als Ursprung des Koordinatensystems, so wären eine geeignete Wahl der Auslenkungswinkel θ von der x_3-Achse. (Der Auslenkungswinkel von der $x_1 - x_3$- oder der $x_2 - x_3$-Ebene, von der die kartesischen Koordinaten ebenfalls abhängen, ist durch den Winkel der jeweiligen Ebene zu der Ebene, in der sich das Fadenpendel bewegt, festgelegt und konstant.)

3 Optimierung, Variationsrechnung und Optimale Steuerung

In diesem Kapitel soll ein kleiner Einblick in Optimierungs-, Variations-, und optimale Steuerungsprobleme gewährt werden. Definitionen, Eigenschaften sowie sonstige Resultate zum erstgenannten Gebiet wurden hierbei, soweit nicht anders vermerkt, dem Skript [16] von S. Volkwein entnommen.

3.1 Grundlagen der endlich-dimensionalen Optimierung

Für die (endlich-dimensionale) Optimierung, sind einige (technische) Definitionen und Eigenschaften von Nöten, welche im Folgenden ohne weitere Einleitung/Beweis aufgelistet werden.

Definition 3.1.1 (Konvexe Menge):
Eine Teilmenge eines komplexen Vektorraums V heißt ***konvex***, wenn für alle $a, b \in M$ und alle $d \in [0,1]$ stets gilt:

$$da + (1-d)\, b \in M$$

Lemma 3.1.2 (Eigenschaften konvexer Mengen):

1. Wenn $C \subseteq \mathbb{R}^n$ eine konvexe Menge ist und $\beta \in \mathbb{R}$ gilt, dann ist auch die Menge $\beta C = \{x : x = \beta c, c \in C\}$ konvex.
2. Seien C und D konvexe Mengen, dann ist auch $C + D = \{x \in \mathbb{R}^n \,|\, x = c + d, c \in C, d \in D\}$ konvex.
3. Der Schnitt von beliebig vielen konvexen Mengen ist wieder konvex.

Definition 3.1.3:
Eine Menge $C \subseteq V$ wird ***Kegel*** genannt, wenn aus $x \in C$ $\alpha x \in C$ für alle $\alpha > 0$ folgt.

Definition 3.1.4 (Lineare Mannigfaltigkeit):
$V \in \mathbb{R}^n$ heißt eine ***lineare Mannigfaltigkeit***, wenn $dx_1 + (1-d)\, x_2 \in V$ für alle $d \in \mathbb{R}$ und $x_1, x_2 \in V$ gilt.

Definition 3.1.5 (Hyperebene):
Eine ***Hyperebene*** im $\mathbb{R}^n$ ist eine $(n-1)$-dimensionale Mannigfaltigkeit.

Lemma 3.1.6 (Identifikation einer Hyperebene):
Sei $a \in \mathbb{R}^n \setminus \{0\}$ und $\gamma \in \mathbb{R}$. Die Menge $H = \left\{x \in \mathbb{R}^n : a^T x = \gamma\right\}$ ist eine Hyperebene.

Lemma 3.1.7:
Sei H eine Hyperebene im $\mathbb{R}^n$. Dann existieren ein Vektor $a \in \mathbb{R}^n \setminus \{0\}$ und ein $\gamma \in \mathbb{R}$, sodass $H = \left\{x \in \mathbb{R}^n : a^T x = \gamma\right\}$.

Definition 3.1.8 (Extrempunkt):
Ein Punkt x in einer konvexen Menge C heißt ***Extrempunkt*** von C, wenn es keine zwei Punkte x_1 und x_2 aus C gibt, so dass $x = \alpha x_1 + (1-\alpha)\, x_2$ für ein α mit $0 < \alpha < 1$ gilt.

3.1.1 Probleme ohne Nebenbedingungen

Betrachtet werden soll ein Problem der Form

$$\text{Minimiere } f(x) \text{ mit } x \in \Omega \subseteq \mathbb{R}^n, \tag{3.1.1}$$

wobei f in eine reellwertige Funktion $f : \Omega \to \mathbb{R}$ und Ω die sogenannte ***zulässige Menge*** ist. Im Folgenden bezeichne $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$.

Definition 3.1.9 (Relative Minimalstelle):
Der Punkt $x^* \in \Omega$ heißt ***relative*** oder ***lokale Minimalstelle*** von f über Ω, wenn es ein $\varepsilon > 0$ gibt, sodass $f(x) \geq f(x^*)$ für alle $x \in \Omega$ mit $\|x - x^*\| < \varepsilon$ gilt. Gilt $f(x) > f(x^*)$ für alle $x \in \Omega$ mit $\|x - x^*\| < \varepsilon$, heißt x^* ***strikte relative Minimalstelle*** von f über Ω.

Definition 3.1.10:
Der Punkt $x^* \in \Omega$ heißt ***globale Minimalstelle*** von f über Ω, wenn $f(x) \geq f(x^*)$ für alle $x \in \Omega$ gilt. Analog zur vorherigen Definition heißt x^* ***strikte globale Minimalstelle***, wenn $f(x) > f(x^*)$ für alle $x \in \Omega$ gilt.

Definition 3.1.11:
Sei $x \in \Omega$. Ein Vektor d heißt ***zulässige Richtung*** in x, wenn es ein $\bar{\alpha} > 0$ gibt, sodass $x + \bar{\alpha} d \in \Omega$ für alle α mit $0 \leq \alpha \leq \bar{\alpha}$ gilt.

Mit obigen Definitionen lässt sich bereits eine Bedingung angeben, die in einem Minimum erfüllt sein muss. (Da es im allgemeinen sehr schwierig ist, ein globales Minimum zu finden, begnügt man sich häufig damit, Stellen zu finden, die die notwendigen Bedingungen erfüllen.)

Satz 3.1.12 (Notwendige Bedingung erster Ordnung):
Seien $\Omega \subset \mathbb{R}^n$ und f eine Funktion aus $C^1(\Omega)$. Wenn x^* eine relative Minimalstelle von f über Ω ist, dann folgt für jedes $d \in \mathbb{R}^n$, das eine zulässige Richtung in x^* angibt, die Ungleichung $\nabla f(x^*)^T d \geq 0$, wobei $\nabla f(x^*) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)(x^*)$ der Gradient ist.

Beweis: Für beliebiges α mit $0 \leq \alpha \leq \bar{\alpha}$ sei $x(\alpha) = x^* + \alpha d \in \Omega$ und $g(\alpha) = f(x(\alpha))$. Dann hat

g an $\alpha = 0$ ein relatives Minimum. Ferner folgt

$$g(\alpha) - g(0) = \underbrace{g'(0)}_{=\nabla f(x^*)^T d} \alpha + o(\alpha)$$

Angenommen, es sei $g'(0) < 0$. Für hinreichend kleines $\alpha > 0$ gilt dann die Ungleichung $g(0) > g(\alpha)$, was ein Widerspruch zum relativen Minimum von g an der Stelle $\alpha = 0$ ist. Also: $0 \leq g'(0) = \nabla f(x^*)^T d$. □

Bemerkung 3.1.13: *Betrachtet man den Spezialfall $\Omega = \mathbb{R}^n$, so gilt $\nabla f(x^*)^T d$ für alle $d \in \mathbb{R}^n$. Also folgt $\nabla f(x^*) = 0$.*

Korollar 3.1.14 (unrestringierter Fall): *Seien $\Omega \subseteq \mathbb{R}^n$ und $f \in C^1(\Omega)$. Wenn x^* ein innerer Punkt von Ω und eine relative Minimalstelle von f über Ω ist, dann folgt*

$$\nabla f(x^*) = 0.$$

Bemerkung 3.1.15: *Notwendige Bedingungen führen im nicht-restringierten Fall auf n Gleichungen in n Unbekannten.*

Unter zusätzlichen Glattheitsvoraussetzungen lässt sich (u. U.) genauer überprüfen, ob es sich bei einem Punkt um einen Kandidaten für eine Minimalstelle handelt.

Lemma 3.1.16:
Sei $f \in C^2(\Omega)$, $\Omega \subseteq \mathbb{R}^n$. Wenn x^* eine relative Minimalstelle von f über Ω ist, dann gilt für jede zulässige Richtung in x^*:

a) $\nabla f(x^*)^T d \geq 0$

b) Aus $\nabla f(x^*)^T d \geq 0$ folgt $d^T \nabla^2 f(x^*) d \geq 0$ (D. h. die Hessematrix $\nabla^2 f(x^*)$ ist positiv semidefinit.)

Beweis:

a) Siehe Satz 3.1.12

b) Sei $x(\alpha) = x^* + \alpha d, g(\alpha) = f(x(\alpha))$. Aus $\nabla f(x^*)^T d = 0$ folgt $g'(\alpha)|_{\alpha=0} = 0$ (da $\frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) = \nabla f(x(\alpha))^T d$). Taylorentwicklung liefert

$$g(\alpha) = g(0) + g'(\alpha)|_{\alpha=0} \alpha + \frac{1}{2} g''(\alpha)\Big|_{\alpha=0} \alpha^2 + o(\alpha^2)$$

Hierbei ist

$$g''(\alpha)\Big|_{\alpha=0} = \frac{\mathrm{d}}{\mathrm{d}\alpha} g'(\alpha)\Big|_{\alpha=0} = \frac{\mathrm{d}}{\mathrm{d}\alpha} \left(\nabla f(x(\alpha))^T d \right)\Big|_{\alpha=0} = d^T \nabla^2 f(x^*) d$$

Somit gilt für die Taylorentwicklung:

$$f(x(\alpha)) = f(x^*) + 0 + \left(\frac{1}{2} d^T \nabla^2 f(x^*) d \right) \alpha^2 + o(\alpha^2)$$

Angenommen, es gilt $d^T \nabla^2 f(x^*) d < 0$. Dann ist $f(x(\alpha)) < f(x^*)$ für α hinreichend klein. Da x^* Minimalstelle ist, ergibt sich hieraus ein ein Widerspruch.

□

Korollar 3.1.17: *Sei $x^* \in \text{int}(\Omega)$ relative Minimalstelle von $f \in C^2(\Omega)$ über Ω. Dann gilt:*

a) $\nabla f(x^*) = 0$

b) $\forall d \in R^n$ *folgt* $d^T \nabla^2 f(x^*)\, d \geq 0$

Satz 3.1.18 (Hinreichende Bedingungen zweiter Ordnung):
Sei $f \in C^2(\Omega), \Omega \subseteq R^n, x^* \in \text{int}(\Omega)$ und weiter gelten

a) $\nabla f(x^*) = 0$

b) Die Hessematrix $\nabla^2 f(x^*)$ ist positiv definit

Dann ist x^* eine strikte Minimalstelle von f.

Beweis: $F(x^*) = \nabla^2 f(x^*)$ ist positiv definit, folglich gibt es ein $\alpha > 0$ mit $d^T \nabla^2 f(x^*)\, d \geq \alpha \|d\|^2$ für alle d. (äquivalenz von Normen.) Taylorentwicklung liefert wegen $\nabla f(x^*) d = 0$

$$f(x^* + d) - f(x^*) = \frac{1}{2} d^T \nabla^2 f(x^*)\, d + o\left(\|d\|^2\right)$$

und somit

$$f(x^* + d) - f(x^*) \geq \frac{\alpha}{2} \|d\|^2 + o\left(\|d\|^2\right) \geq 0.$$

Für $d \neq 0$ mit $\|d\|$ hinreichend klein gilt somit $f(x^* + d) \geq f(x^*)$. □

Konvexe Funktionen

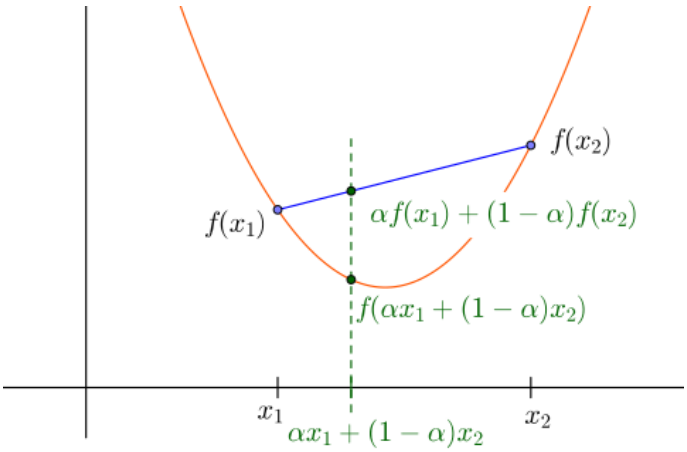
Nun sollen konvexe Funktionen betrachtet werden, welche für die Optimierung einige schöne Eigenschaften haben.

Definition 3.1.19 ((strikt) konvex):

Eine Funktion $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$ konvex, wird *(strikt) konvex* genannt, wenn für jedes $0 < \alpha < 1$ und $x_1, x_2 \in \Omega$, $x_1 \neq x_2$ gilt:

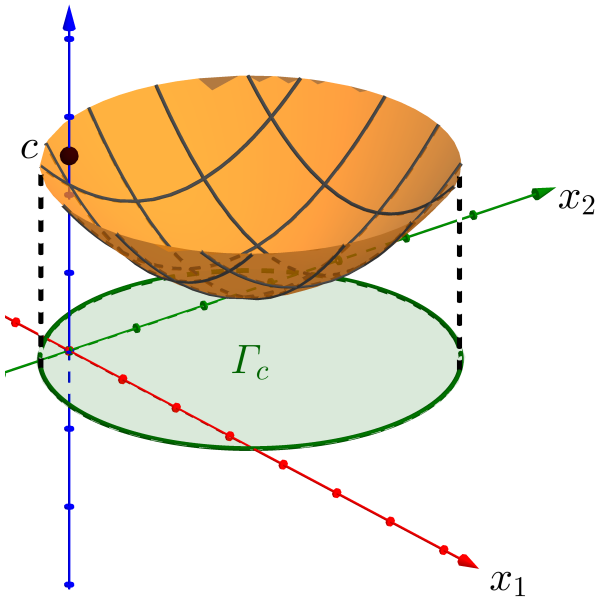
$$f(\alpha x_1 + (1-\alpha) x_2) \overset{(\leq)}{\leq} \alpha f(x_1) + (1-\alpha) f(x_2)$$

(Strikt) Konvexe Funktionen zeichnen sich dadurch aus, dass Graph der Funktion entlang der Verbindungsstrecke von x_1 nach x_2 unterhalb der Geraden durch die Punkte $(x_1, f(x_1))$ und $(x_2, f(x_2))$ verläuft (s. Bild rechts).



Beispiel für eine konvexe Funktion mit $\Omega \subseteq \mathbb{R}$

Lemma 3.1.20:
Seien f_1, f_2 konvexe Funktionen auf einer konvexen Menge, dann ist $f_1 + f_2$ konvex auf Ω, weiter ist αf_1 konvex auf Ω für alle $\alpha \geq 0$.



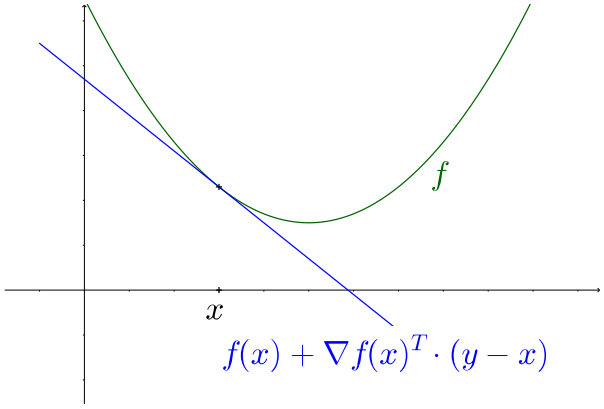
Veranschaulichung von Lemma 3.1.21.

Lemma 3.1.21:
Sei f konvexe Funktion auf einer konvexen Menge Ω. Die Menge $\Gamma_c := \{x \in \Omega \,|\, f(x) \leq c\}$ ist dann ebenfalls konvex für alle $c \in \mathbb{R}$.

Lemma 3.1.22:
Sei $f \in C^1(\Omega)$ mit $\Omega \subseteq \mathbb{R}^n$ konvex. Eine Funktion $f : \Omega \to \mathbb{R}$ ist genau dann konvex, wenn

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

für alle $x, y \in \Omega$ gilt. Für $f \in C^2(\Omega)$ gilt: Eine Funktion f ist genau dann konvex, wenn sie eine strikt positive Krümmung hat.



Veranschaulichung von Lemma 3.1.22 mit $\Omega \subseteq \mathbb{R}$. Die Tangente im Punkt $(x, f(x))$ verläuft unterhalb des Graphen von f.

Satz 3.1.23:
Seien $\Omega \subseteq \mathbb{R}^n$ und $f \in C^1(\Omega)$ konvex. Gibt es einen Punkt $x^* \in \Omega$, sodass für alle $y \in \Omega$

$$\nabla f(x^*)^T (y - x^*) \geq 0$$

gilt, dann ist x^* eine globale Minimalstelle von f auf Ω.

Beweis: Für alle $y \in \Omega$ ist $y - x^* =: d$ eine zulässige Richtung in x^*. (Die Bedingung $\nabla f(x^*)^T (y - x^*) \geq 0$ ist identisch mit der notwendigen Bedingung 1. Ordnung.) Ferner gilt nach Lemma 3.1.22

$$f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x^*)$$

Somit gilt für alle $y \in \Omega$ $f(y) \geq f(x^*)$, also ist x^* globale Minimalstelle. □

3.1.2 Probleme mit Nebenbedingungen

In diesem Kapitel sollen ***restringierte Minimierungsprobleme***, d. h. Probleme der Form

$$\text{Minimiere } f(x) \text{ unter den Nebenbedingungen } \begin{array}{c} h_1(x) = 0,\ g_1(x) \leq 0 \\ \vdots \qquad \vdots \\ h_m(x) = 0,\ g_p(x) \leq 0 \end{array}, x \in \Omega \subseteq \mathbb{R}^n \tag{3.1.2}$$

behandelt werden. Hierbei sei $m \leq n$. Ferner seien im Folgenden h_i, $1 \leq i \leq m$ sowie g_j, $1 \leq j \leq p$ zweimal stetig differenzierbar. Fässt man die einzelnen Funktionen zu vektorwertigen Funktionen $h = (h_1, \ldots, h_m)^T$ und $g = (g_1, \ldots, g_p)^T$ zusammen, so lässt sich Problem (3.1.2) schreiben als

$$\text{Minimiere } f(x) \text{ unter den Nebenbedingungen } h(x) = 0,\ g(x) \leq 0 \text{ und } x \in \Omega \subseteq \mathbb{R}^n$$

(wobei die Ungleichung komponentenweise zu verstehen ist.)

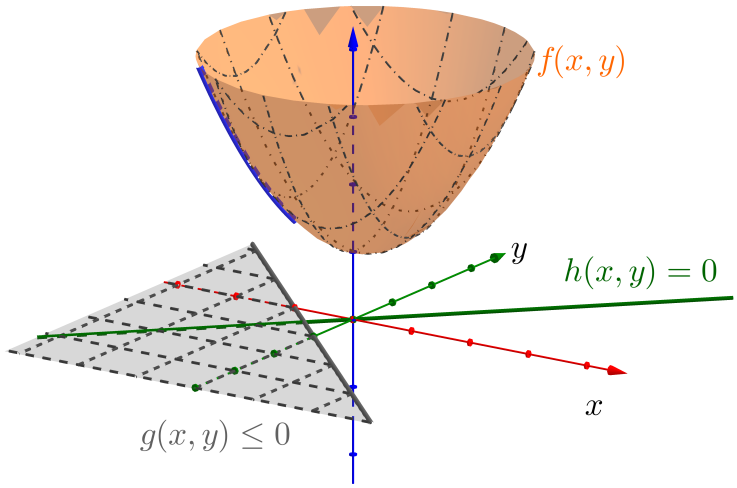


Abbildung 3.1.1: Beispielhafte Skizze eines Minimierungsproblems. Die blaue Kurve enthält alle Punkte des Graphen, deren x- und y-Koordinaten die Nebenbedingungen erfüllen.

Die Restriktionen $h(x) = 0$ und $g(x) \leq 0$ heißen ***funktionale Nebenbedingungen***. Erfüllt ein

Punkt $x \in \Omega$ sämtliche Nebenbedingungen, so heißt dieser ***zulässig***, die ***zulässige Menge*** ist die Menge all dieser Punkte. Ist für einen zulässigen Punkt x der Wert $g_i(x)$ gleich 0, so nennt man diese Ungleichungsnebenbedingung ***aktiv*** an diesem Punkt, andernfalls ***inaktiv***. Aufgrund der Stetigkeit erfüllen in letzterem Falle alle Punkte $\tilde{x}$ in einer hinreichend kleinen Umgebung von y ebenfalls $g_i(\tilde{x}) < 0$. In dieser Umgebung wird die zulässige Menge durch ebendiese Restriktion folglich nicht beeinflusst.

Vereinbarung: Im Folgenden bezeichne ∇f den Gradienten von f, sowie $\nabla h = \nabla h_1$. Im Falle $h = (h_1, \ldots, h_m)$ mit $m > 1$ bezeichne ∇h die Matrix, deren Spalten aus den Gradienten $\nabla h_j, 1 \leq j \leq m$ besteht, d. h. $\nabla h = (\nabla h_1, \nabla h_2, \ldots, \nabla h_m) \in \mathbb{R}^{n \times m}$. (Anders ausgedrückt ist ∇h folglich die transponierte Jacobimatrix.) Analoges gelte für g. Mit $\nabla^2 f$ wird im Folgenden die Hessematrix von f bezeichnet, mit $\nabla^2 h_i$ und $\nabla^2 g_j$ die Hessematrizen von h_i bzw. g_j bezeichnet.

3.1.3 Restringierte Probleme ohne Ungleichungsnebenbedingungen

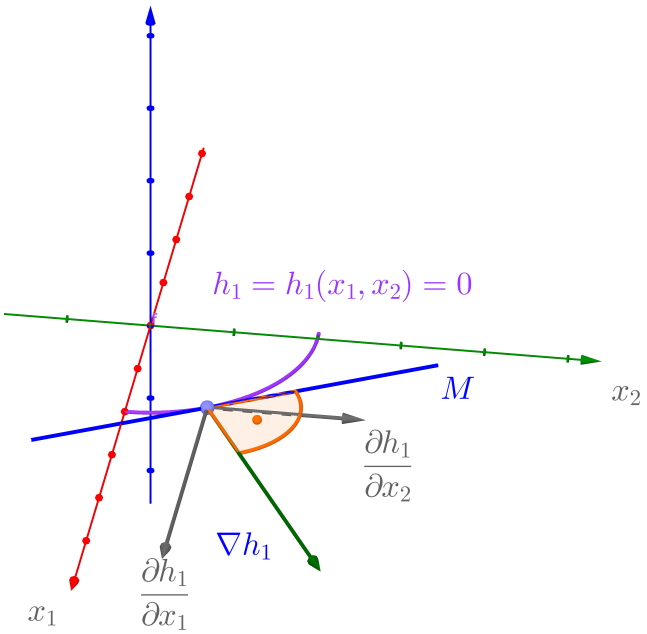
Vorerst wird soll der Fall ohne Ungleichungsnebenbedingungen, also $h(x) = 0$ sowie $x \in \Omega$ als einzigen Nebenbedingungen betrachtet werden.

Definition 3.1.24 (Regulärer Punkt):

Ein Punkt x^* mit $h(x^*) = 0$ heißt ***regulärer Punkt***, wenn die Gradienten $\nabla h_1(x^*)$, …, $\nabla h_m(x^*)$ linear unabhängig sind.

Satz 3.1.25:
An einem regulären Punkt x^* der Fläche, die durch $h(x) = 0$ gegeben ist, ist die Tangentialebene gleich der Menge $M = \left\{ y : \nabla h(x^*)^T y = 0 \right\}$.

Im Bild rechts ist so eine Tangentialebene für $h : \mathbb{R}^2 \to \mathbb{R}$ exemplarisch dargestellt. (In diesem Falle entspricht diese einer Tangente.)



Lemma 3.1.26:
Sei x^* ein regulärer Punkt der Nebenbedingung $h(x) = 0$ und eine lokale Extremstelle. Dann gilt

$$\nabla f(x^*)^T y = 0$$

für alle y aus der Tangentialebene, d. h. alle $y \in \mathbb{R}^n$ mit $\nabla h(x^*)^T y = 0$.

Beweis: Sei y ein Vektor in der Tangentialebene an x^* und $x(t)$ eine glatte Kurve auf der durch $h(x) = 0$ definierten Fläche, welche durch x^* geht, d. h. $h(x(t)) = 0$ und $x(0) = x^*, \dot{x}(0) = y$ für $t \in [-\alpha, \alpha]$. Da x^* regulärer Punkt ist, ist die Menge der $y \in \mathbb{R}^n$ mit $\nabla h(x^*)^T y = 0$ identisch zur Tangentialebene. Weiter gilt

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t))\Big|_{t=0} = 0$$

sowie

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t))\Big|_{t=0} = \nabla f(x^*)^T \dot{x}(0) = \nabla f(x^*)^T y$$

Hieraus folgt

$$\nabla f(x^*)^T y = 0.$$

□

Bemerkung 3.1.27: *Lemma 3.1.26 hat eine geometrische Interpretation: In einer lokalen Minimalstelle, welche regulärer Punkt ist, steht der Gradient $\nabla f(x^*)$ senkrecht auf der Tangentialebene.*

Satz 3.1.28 (notwendige Bedingungen erster Ordnung):

Sei x^* eine lokale Extremalstelle von f unter der Nebenbedingung $h = 0$. Weiter sei x^* ein regulärer Punkt. Dann existiert $\lambda \in \mathbb{R}^m$, so dass

$$\nabla f(x^*) + \nabla h(x^*) \cdot \lambda = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0$$

gilt.

Beweisidee/-skizze: Lineare Programmierung, starker Dualitätssatz:
Das Problem

$$\min c^T x \quad \text{u. d. Nb. } Ax = b, \ \left(c, x \in \mathbb{R}^n, b \in \mathbb{R}^m, A \in \mathbb{R}^{m \times n}\right)$$

besitzt genau dann eine endliche Optimallösung, wenn das duale Problem

$$\max b^T y \quad \text{u. d. Nb. } A^T y = c$$

eine endliche Optimallösung besitzt. In dem Falle gilt

$$\min c^T x = \max b^T y.$$

Im vorliegenden Falle lautet das Maximierungsproblem

$$\max_{y \in \mathbb{R}^n} \nabla f(x^*)^T y \quad \text{u. d. Nb. } \nabla h(x^*)^T y = 0,$$

das Minimierungsproblem lautet

$$\min_{w \in \mathbb{R}^m} 0 \quad \text{u. d. Nb. } \nabla h(x^*) w = \nabla f(x^*).$$

Nach dem vorherigen Lemma ist $\nabla f(x^*)^T y = 0$ für die gesamte Nebenbedingungsmenge. Folglich existiert mit dem Nullvektor eine Lösung. $\Rightarrow$ Das duale Minimierungsproblem ist lösbar, es existiert ein $w \in \mathbb{R}^m : \nabla h(x) \cdot w = \nabla f(x)$. Hieraus folgt die Behauptung ($\lambda = -w$). □

Beispiel:

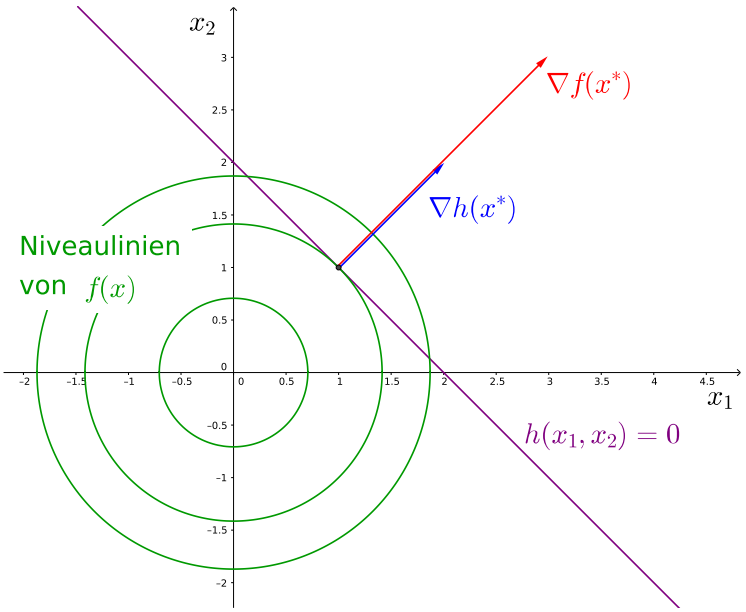
Gegeben sei das Problem

$$\min_{x\in\mathbb{R}^2} f\left(x_1,x_2\right),\ f(x_1,x_2)=x_1^2+x_2^2$$

unter der Nebenbedingung

$$0=h\left(x_1,x_2\right)=x_1+x_2-2$$

In diesem Falle ist $\nabla h(x_1,x_2)=\begin{pmatrix}1\\1\end{pmatrix}$ für alle $(x_1,x_2)^T\in\mathbb{R}^2$ sowie $\nabla f(x_1,x_2)=\begin{pmatrix}2x_1\\2x_2\end{pmatrix}$. Für die lokale Extremstelle $x^*=(1,1)^T$ gilt folglich $\nabla f(x^*)=\begin{pmatrix}2\\2\end{pmatrix}$. Hier ist also $\lambda=2$. (Herleitung der Extremstelle: Wegen h lässt sich das Minimierungsproblem schreiben als $\min f(x_1,x_2)=\min\tilde{f}(x_1)$ mit $\tilde{f}(x_1)=x_1^2+(2-x_1)^2=2\cdot(x_1-1)^2+2$.)



Bemerkung 3.1.29: *Der Faktor λ heißt **Lagrange-Multiplikator**. Mit Hilfe der **Lagrangefunktion***

$$\mathcal{L}\left(x,\lambda\right)=f\left(x\right)+\lambda^T h\left(x\right)$$

lassen sich die notwendigen Bedingungen des restringierten Minimierungsproblems in der Form

$$\nabla_x\mathcal{L}\left(x,\lambda\right)=\nabla f(x)+\nabla h(x)\lambda=0$$
$$\nabla_\lambda\mathcal{L}\left(x,\lambda\right)=h(x)^T=0$$

schreiben.

Satz 3.1.30 (Notwendige Bedingungen zweiter Ordnung):

Seien f,h zweimal stetig differenzierbar. Sei x^* eine lokale Minimalstelle von f mit $h\left(x^*\right)=0$ und ein regulärer Punkt. Dann existiert $\lambda\in\mathbb{R}^m$ sodass

$$\nabla f\left(x^*\right)+\nabla h\left(x^*\right)\cdot\lambda=0$$

und die Matrix $\nabla^2 f\left(x^*\right)+\nabla^2 h\left(x^*\right)\cdot\lambda:=\nabla^2 f(x^*)+\sum_{i=1}^m\lambda_i\nabla^2 h_i(x^*)$ positiv semidefinit auf der Tangentialebene $M=\left\{y\in\mathbb{R}^n,\nabla h\left(x^*\right)^T y=0\right\}$ ist.

Beweis: Sei $x(t)$ eine zweimal stetig differenzierbare Funktion mit $x(0) = x^*$, $\dot{x}(0) = y$ für ein $y \in M$, welche überall die Nebenbedingung $h(x) = 0$ erfüllt. (D. h. $h(x(t)) \equiv 0$). Da x^* Minimalstelle für $f(x(t))$ ist, gilt:

$$0 \leq \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x(t))\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla f(x(t))^T \dot{x}(t)\right)\Big|_{t=0} = \\ = \dot{x}(0)^T \nabla^2 f(x^*)\dot{x}(0) + \nabla f(x^*)^T \ddot{x}(0) \tag{I}$$

Ferner folgt aus $h(x(t)) \equiv 0$ die Beziehung $\lambda^T h(x(t)) \equiv 0$ und somit:

$$0 = \frac{\mathrm{d}^2}{\mathrm{d}t^2} h(x(t))\lambda\Big|_{t=0} = \sum_{i=1}^{m} \lambda_i \frac{\mathrm{d}^2}{\mathrm{d}t^2} h_i(x(t))\Big|_{t=0} = \\ = \dot{x}(0)^T \nabla^2 h(x^*)\lambda\dot{x}(0) + \lambda^T \nabla h(x^*)\ddot{x}(0) \tag{II}$$

Addition von (I) und (II) ergibt schließlich:

$$0 \leq \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x(t))\Big|_{t=0} = \dot{x}(0)^T \left(\nabla^2 f(x^*) + \nabla^2 h(x^*)\lambda\right)\dot{x}(0) + \left(\nabla f(x^*)^T + \lambda^T \nabla h(x^*)^T\right)\ddot{x}(0)$$

Da der zweite Summand infolge der notwendigen Bedingung 1. Ordnung 0 ist, folgt wegen $\dot{x}(0) = y$ die Behauptung. □

Satz 3.1.31 (Hinreichende Bedingungen zweiter Ordnung):

Sei x^* ein Punkt mit $h(x^*) = 0$ und es existiere $\lambda \in \mathbb{R}^m$ mit $\nabla f(x^*) + \nabla h(x^*)\lambda = 0$. Weiter sei die Matrix

$$\nabla^2 f(x^*) + \nabla^2 h(x^*)\lambda = \nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*)$$

positiv definit auf der Tangentialebene. Dann ist x^* eine strikte Minimalstelle von f unter der Nebenbedingung $h = 0$.

Beispiel:

Gegeben sei das restringierte Minimierungsproblem

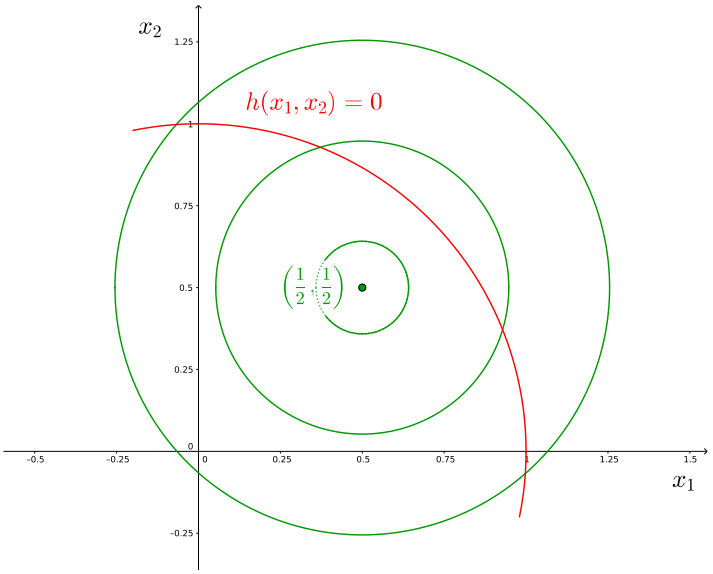
$$\min_{x\in\mathbb{R}^2} f(x) \quad \text{u. d. Nb. } h(x)=0$$

mit zu minimierender Funktion $f(x_1,x_2)=x_1^2+x_2^2-x_1-x_2+1$ sowie $h(x_1,x_2)=x_1^2+x_2^2-1$. Die zugehörigen Gradienten sind

$$\nabla f\left(x_1,x_2\right)=\begin{pmatrix}2x_1-1\\2x_2-1\end{pmatrix}$$

sowie

$$\nabla h\left(x_1,x_2\right)=\begin{pmatrix}2x_1\\2x_2\end{pmatrix}$$



Exemplarische Niveaulinien der zu minimierenden Funktion f (grün) mit Ortskurve der Nebenbedingungsmenge.

Sei nun

$$0=\begin{pmatrix}2x_1-1\\2x_2-1\end{pmatrix}+\lambda\begin{pmatrix}2x_1\\2x_2\end{pmatrix}$$

Dies ist äquivalent zu der Gleichung

$$(1+\lambda)\begin{pmatrix}2x_1\\2x_2\end{pmatrix}=\begin{pmatrix}1\\1\end{pmatrix},$$

aus welcher

$$x_1=x_2=\frac{1}{2\left(1+\lambda\right)}$$

folgt. Eingesetzt in die Nebenbedingung ergibt sich aus

$$0=\left(\frac{1}{2\left(1+\lambda\right)}\right)^2+\left(\frac{1}{2\left(1+\lambda\right)}\right)^2-1=\frac{2}{4\left(1+\lambda\right)^2}-1$$

die quadratische Gleichung

$$\lambda^2+2\lambda+1=\frac{1}{2}\Leftrightarrow\lambda^2+2\lambda+\frac{1}{2}=0,$$

deren Lösungen durch $\lambda_1=\frac{-2-\sqrt{4-2}}{2}=$ und $\lambda_2=-1+\frac{\sqrt{2}}{2}$ gegeben sind. Somit haben die einzigen Kandidaten (!) für die Lösung des gegeben Problems die Koordinaten $x_1^*=x_2^*=\pm\frac{1}{\sqrt{2}}$. Die Diagonalmatrix

$$\nabla^2 f\left(x_1,x_2\right)+\lambda_{1/2}\nabla^2 h\left(x_1,x_2\right)=\begin{bmatrix}2&0\\0&2\end{bmatrix}+\lambda_{1/2}\begin{bmatrix}2&0\\0&2\end{bmatrix}=\begin{bmatrix}2\left(1+\lambda_{1/2}\right)&0\\0&2\left(1+\lambda_{1/2}\right)\end{bmatrix}$$

ist genau dann positiv semidefinit, wenn die Eigenwerte ≥ 0 sind, was nur für $\lambda_2\geq -1+\frac{\sqrt{2}}{2}$ gilt. Da sie in diesem Falle sogar positiv definit ist, ist folglich die eindeutige Lösung des Minimierungsproblems durch $x_1=x_2=\frac{1}{\sqrt{2}}$ gegeben.

Häufig werden Optimierungsprobleme numerisch gelöst. Daher ist es sinnvoll, zu untersuchen wie sensibel die Lösung aufÄnderungen in den Nebenbedingungen reagiert. über diese sogenannte ***Sensitivität*** gibt folgender Satz Auskunft:

Satz 3.1.32 (Sensitivitätssatz):
Gegeben sei das restringierte Minimierungsproblem

$$\min f(x) \quad \text{u. d. Nb. } h(x) = c \tag{$*$}$$

mit $f, h \in C^2$ sowie $c \in \mathbb{R}^m$. Für $c = 0$ gebe es eine lokale Lösung, die regulärer Punkt ist und mit dem zugehörigen Lagrange-Multiplikatoren die hinreichende Bedingung zweiter Ordnung für eine strikte Minimalstelle erfülle. Dann gibt es für jedes $c \in \mathbb{R}^m$ in einer hinreichend kleinen Umgebung von 0 ein $x(c)$, welches stetig von c abhängt, so dass $x(0) = x^*$ gilt und $x(c)$ lokale Lösung von $(*)$ ist. Ferner gilt

$$\nabla_c f(x(c))|_{c=0} = -\lambda$$

3.1.4 Restringierte Probleme mit Ungleichungsnebenbedingungen

Nun sollen Probleme mit Ungleichungsnebenbedingungen ($g(x) \leq 0$) betrachtet werden, d. h. Probleme der Form

$$\min f(x) \quad \text{u. d. Nb. } h(x) = 0,\ g(x) \leq 0. \tag{3.1.3}$$

Definition 3.1.33 (Regulärer Punkt):

Sei x^* ein Punkt, der die Nebenbedingungen erfüllt, d. h. $h(x^*) = 0, g(x^*) \leq 0$ und $J(x^*)$ die Menge der Indizes j mit $g_j(x^*) = 0$. Dann heißt x^* regulärer Punkt bezüglich der Nebenbedingung, wenn die Vektoren $\nabla h_1(x^*), \nabla h_2(x^*), \ldots, \nabla h_m(x^*)$ und $\nabla g_j(x^*), j \in J$ linear unabhängig sind.

Bemerkung: *Die Indexmenge $J(x^*) = \{j : 1 \leq j \leq p, g_j(x^*) = 0\}$ ist die Menge der Indizes der **aktiven Nebenbedingungen** im Punkt x^*. Inaktive Nebenbedingung im Punkt x^* (d. h. $g(x^*) < 0$) werden hierbei nicht beachtet, da sie aufgrund der Stetigkeit der g_j die Nebenbedingungsmenge lokal (!) nicht einschränken.*

Satz 3.1.34 (Notwendige Bed. erster Ordnung/Karush-Kuhn-Tucker-Bed.):
Sei x^* eine relative Minimalstelle für das Problem (3.1.3) und ein regulärer Punkt. Dann gibt es ein $\lambda \in \mathbb{R}^m$ und ein $\mu \in R^p, \mu \geq 0$ (komponentenweise), sodass

$$\begin{aligned} &\nabla f(x^*) + \nabla h(x^*)\lambda + \nabla g(x^*)\mu = \\ &= \nabla f\left(x^*\right) + \sum_{i=1}^{m} \lambda_i \nabla h_i\left(x^*\right) + \sum_{j=1}^{p} \mu_j \nabla g_j\left(x^*\right) = 0 \end{aligned} \tag{3.1.34.a}$$

sowie

$$\mu^T g\left(x^*\right) = 0 \tag{3.1.34.b}$$

gelten. Erfüllt ein Punkt $\left(\tilde{x}, \tilde{\lambda}, \tilde{\mu}\right) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ diese Bedingungen, so wird er KKT-Punkt genannt.

Satz 3.1.35 (Notwendige Bedingungen zweiter Ordnung):
Seien $f, g, h \in C^2$ und x^* ein regulärer Punkt der Nebenbedingung $h\left(x^*\right) = 0, g\left(x^*\right) \leq 0$. Wenn x^* eine relative Minimalstelle für (3.1.3) ist, dann gibt es ein $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p, \mu \geq 0$ so dass

$$\begin{aligned} \nabla f\left(x^*\right) + \nabla h\left(x^*\right)\lambda + \nabla g\left(x^*\right)\mu &= 0 \\ \mu^T g\left(x^*\right) &= 0 \end{aligned}$$

gelten und die Matrix

$$\nabla^2 f\left(x^*\right) + \nabla^2 h\left(x^*\right)\lambda + \nabla^2 g\left(x^*\right)\mu \coloneqq \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*) + \sum_{j=1}^{p} \mu_j \nabla^2 g_j(x^*) = 0$$

positiv semidefinit auf der Tangentialebene der Nebenbedingungen an x^* ist.

Satz 3.1.36 (Hinreichende Bedingungen):
Falls die notwendigen Bedingungen 1. und 2. Ordnung in x^* erfüllt sind und die Hessematrix $\nabla^2 \mathcal{L}(x^*) = \nabla^2 + \nabla^2 h(x^*)\lambda^* + \nabla^2 g(x)\mu$ positiv definit ist für alle y in

$$M' \coloneqq \left\{ y \in \mathbb{R}^n | \nabla h_i(x^*)^T y = 0, \nabla g_j(x^*)^T y = 0 \forall j \in J_1, \nabla g_j(x^*)^T y \leq 0 \forall j \in J_2 \right\},$$

mit $J_1 \coloneqq \left\{ j \in x^* | \mu_j > 0 \right\}, J_2 \coloneqq \left\{ j \in x^* | \mu_j = 0 \right\}$, dann ist x^* eine strikte Minimalstelle.

Auch für Probleme mit Ungleichungsnebenbedingungen lassen sich Aussagen bezüglich der Sensitivität, ähnlich den Resultatetn bei restringierten Problemen ohne Ungleichungsnebenbedingungen, treffen:

Satz 3.1.37 (Sensitivitätssatz):

Gegeben sei das Problem

$$\min f(x) \quad \text{u. d. Nb. } h(x) = c,\ g(x) \leq d \tag{$**$}$$

Hierbei seien $f, g, h \in C^2, c \in R^m, d \in R^p$. Für $c = 0, d = 0$ existiere eine lokale Minimalstelle x^* und für λ, μ gelten die hinreichende Bedingungen für eine strikte Minimalstelle. Dann gibt es für alle $c \in \mathbb{R}^m, d \in \mathbb{R}^p$ eine Umgebung von $0 \in R^{m+p}$, in der eine Lösung $x(c,d)$ von $(**)$ existiert, die stetig von (c,d) abhängt und $x^* = x(0,0)$ erfüllt. Ferner gelten

$$\nabla_c f\left(x\left(c,d\right)\right)\big|_{(0,0)} = -\lambda$$
$$\nabla_d f\left(x\left(c,d\right)\right)\big|_{(0,0)} = -\mu$$

3.1.5 Verfahren zur Lösung eines Minimierungsproblem

Im Folgenden sollen einige Verfahren zur computergestützten Lösung von Optimierungsproblemen der Form

$$\min f(x) \quad \text{u. d. Nb. } x \in S \tag{3.1.4}$$

vorgestellt werden, wobei S eine Restriktionsmenge ist. (Bei Problemen der Form (3.1.2) z. B. die Menge aller $x \in \mathbb{R}^n$ mit $h(x) = 0$ und $g(x) \leq 0$).

Penalty-Methode:

Bei der Penalty-Methode wird das restringierte Problem (3.1.4) durch ein unrestringiertes Problem der Form

$$\min\left(f\left(x\right) + \nu P\left(x\right)\right) \tag{3.1.5}$$

mit $\nu > 0, \nu \in \mathbb{R}$ und $P : \mathbb{R}^n \to \mathbb{R}$ ersetzt, wobei P folgende Eigenschaften aufweist:

1. P ist stetig
2. Für alle $x \in \mathbb{R}^n$ gilt $P\left(x\right) \geq 0$
3. Es gilt $P\left(x\right) = 0$ genau dann, wenn $x \in S$ ist.

Beispiel:

Gegeben sie das Minimierungsproblem (3.1.4) mit $S = \{x \,|g_j\left(x\right) \leq 0, j = 1, \ldots, p\}$. Eine geeignete Funktion p ist beispielsweise durch

$$P\left(x\right) = \frac{1}{2}\sum_{j=1}^{p}\left(\max\left(0, g_j\left(x\right)\right)\right)^2$$

gegeben. Für große Werte für ν wirken sich große Funktionswerte von P stärker aus, als große Funktionswerte von f. Entsprechend wird der Wert $P(x^*)$ für Lösung x^* von (3.1.5) klein sein und die Lösungen nähern sich für $\nu \to \infty$ der Restriktionsmenge S an.

Sei nun

$$q(x,\nu) := f(x) + \nu P(x)$$

sowie (ν_k) eine strikt monoton steigende Folge in $\mathbb{R}$ mit $\nu_0 \geq 0$. Dieses q wird auch als ***Penalty-Funktion*** bezeichnet. Bei der Penalty-Methode werden nun im k. Schritt die Lösungen des entsprechenden Minimierungsproblems (3.1.5) berechnet, d. h.

$$x_k = \arg\min q(x,\nu_k)$$

bestimmt. Aus den Definitionen ergibt sich folgendes Lemma:

Lemma 3.1.38:
Vorausgesetzt, für jedes k existiere eine Lösung x_k des Problems $\min q(x,\nu_k)$, so gelten folgende Ungleichungen:

(i) $q(x_k,\nu_k) \leq q(x_{k+1},\nu_{k+1})$

(ii) $P(x_k) \geq P(x_{k+1})$

(iii) $f(x_k) \leq f(x_{k+1})$

Ferner gilt für eine Lösung x^* von (3.1.4) sowie jedes k:

(iv) $f(x^*) \geq q(\nu_k, x_k) \geq f(x_k)$

Beweis:

(i) $f(x_k) + \nu_k P(x_k) \leq f(x_{k+1}) + \nu_k P(x_{k+1}) \leq f(x_{k+1}) + \nu_{k+1} P(x_{k+1})$

(ii) $f(x_k) + \nu_k P(x_k) \leq f(x_{k+1}) + \nu_k P(x_{k+1}), f(x_{k+1}) + \nu_{k+1} P(x_{k+1}) \leq f(x_k) + \nu_{k+1} P(x_k) \Rightarrow$
$(\nu_{k+1} - \nu_k) P(x_{k+1}) \leq (\nu_{k+1} - \nu_k) P(x_k) \overset{\nu_{k+1} > \nu_k}{\Rightarrow} P(x_{k+1}) \leq P(x_k)$

(iii) $f(x_{k+1}) + \nu_k P(x_{k+1}) \geq f(x_k) + \nu_k P(x_k) \geq f(x_k) + \nu_k P(x_{k+1})$

(iv) $f(x^*) = f(x^*) + \nu_k P(x^*) \geq f(x_k) + \nu_k P(x_k) \geq f(x_k)$

□

Mit Hilfe dieses Lemmas lässt sich zeigen, dass der Algorithmus der Penalty-Methode gutartig in dem Sinne ist, dass sich mit ihm tatsächlich eine Lösung des Optimierungsproblem (3.1.4) finden lässt:

Satz 3.1.39:
Sei zu einer gegebenen Folge $(\nu_k)_{k\in\mathbb{N}}$, welche obige Voraussetzungen erfüllt, $(x_k)_{k\in\mathbb{N}}$ eine durch die Penalty-Methode generierte Folge von Lösungen der entsprechenden Minimierungsprobleme. Dann ist jeder Grenzwert dieser Folge eine Lösung des Problems (3.1.4).

Barriere-Methode:

Bei der Methode wird ähnlich vorgegangen wie bei der Penalty-Methode, jedoch mit einigen wesentlichen Unterschieden: Statt durch ein unrestringiertes Problem wird das gegeben Problem der Form (3.1.4) durch ein restringiertes Problem der Form

$$\min_{x\in \mathrm{int}(S)} f(x) + \frac{1}{\nu} B(x)$$

ersetzt. Hierbei habe die Restriktionsmenge S ein nicht leeres Inneres und jeder Randpunkt lasse sich durch Punkte im Inneren approximieren. Ferner sei B eine Funktion $B : \mathbb{R}^n \to \mathbb{R}$ mit folgenden Eigenschaften:

1. B ist stetig
2. $B(x) \geq 0$
3. $B(x) \to \infty$ für $x \to \partial S$

Eine solche Funktion wird ***Barriere-Funktion*** genannt. Da $B(x)$ am Rand unendlich wird, bleibt eine entsprechende Lösung des Minimierungsproblems auch dann im Inneren von S, wenn sie mit Verfahren für unrestringierte Probleme gelöst werden, d. h. die Lösungen müssen nicht dahingehend überprfüft werden, ob sie den Nebenbedingungen genügen.

Beispiel:

Seien $g_j, 1 \leq j \leq p \in C(\mathbb{R}^n, \mathbb{R})$ und $S = \{x \,|g_i(x) \leq 0, j = 1, \ldots, p\}$, dergestalt, dass $\text{int}(S) = \{x \in \mathbb{R}^n | g_j(x) \leq 0, 1 \leq j \leq p\}$ gilt und es zu jedem Punkt $\bar{x} \in \partial S$ eine Folge $(x_k)_{k \in \mathbb{N}_0}$ mit $x_k \to \bar{x}$ gibt. In diesem Fall wäre eine geeignete Barriere-Funktion

$$B(x) = -\sum_{j=1}^{p} \frac{1}{g_j(x)}$$

Das sonstige Vorgehen erfolgt jedoch, wie bereits angedeutet, völlig analog zur Penalty-Methode: Zu einer gegebenen, unbeschränkten, strikt monoton steigenden Folge $(\nu_k)_{k \in \mathbb{N}}$ werden die Lösungen des entsprechenden Minimierungsproblems,

$$x_k = \arg\min f(x) + \frac{1}{\nu_k} B(x),$$

bestimmt, mit welchen sich wiederum eine Lösung des ursprünglichen Problems bestimmen lässt.

Satz 3.1.40:
Jeder Grenzwert einer durch die Barriere-Methode generierten Folge $(x_k)_{k \in \mathbb{N}}$ ist eine Lösung des Problems (3.1.4) mit den genannten Einschränkungen für S.

Vor- und Nachteile der beiden Verfahren

+ Anwendbar für nichtlineare Probleme
+ Restriktionsmenge S kann sehr allgemein sein, f, g können auch nichtlinear sein
+ Penalty- bzw. Barriere-Funktion bei funktionalen Nebenbedingungen ($h(x) = 0$, $g(x) \leq 0$) sehr einfach angebbar (s. obige Beispiele)
- Für die meisten Algorithmen sind stetige erste Ableitungen erforderlich
- Penalty-Verfahren schlecht konditioniert

Augmentiertes Lagrange-Verfahren:

Zum Schluss soll noch ein Verfahren vorgestellt werden, welches sich sehr gut für restringierte Probleme der Form

$$\min f(x) \quad \text{u. d. Nb. } h(x) = 0 \tag{3.1.6}$$

eignet. Dieses Verfahren heißt ***augmentiertes Lagrange-Verfahren*** und beruht auf der Nutzung der ***augmentierten Lagrangefunktion***

$$\mathcal{L}_c(x, \lambda) = f(x) + \lambda^T h(x) + \frac{c}{2} \|h(x)\|^2, \; c > 0. \tag{3.1.7}$$

Ausgehend von einem Vektor λ_k wird die Minimalstelle x_k von $\mathcal{L}_c(x, \lambda_k)$ bestimmt. Anschließend wird, beispielsweise durch

$$\lambda_{k+1} = \lambda_k + ch(x_k) \tag{3.1.8}$$

ein neuer Multiplikationsvektor bestimmt und die Prozedur wiederholt.
Begründung: Sei λ^* der Lagrange-Multiplikator des Problems (3.1.6). Für festes λ_k kann die Minimierung der augmentierten Lagrangefunktion als Minimierungsproblem

$$\min f(x) + \lambda_k^T h(x) \quad \text{u. d. Nb. } h(x) = 0 \tag{3.1.9}$$

aufgefasst werden. Der entsprechende Lagrange-Multiplikator ist $\lambda^* - \lambda_k$. (3.1.7) kann jedoch auch als Penalty-Funktion für (3.1.9) interpretiert werden, womit $ch(x_k) \approxeq \lambda^* - \lambda_k$ gesetzt werden kann. (3.1.8) liefert folglich eine gute Näherung für λ^*.
Der große Vorteil dieses Verfahrens gegenüber der Penalty-Methode besteht darin, dass die Konstante c nicht gegen unendlich streben muss, wodurch sich die bei ebendieser Methode auftretende schlechte Kondition vermeiden lässt.

Bemerkung 3.1.41: *In Anhang C.2 ist eine mögliche Implementierung des Lagrange-Verfahrens zu finden. Zur Lösung der unrestringierten Probleme wird das sogenannte* ***Gradienten-Verfahren*** *(engl. auch* ***steepest descent****) verwendet (s. Anhang C.1). Dieses sucht eine bessere Lösung in Richtung des negativen Gradienten, welcher den steilsten Abstieg repräsentiert.*

3.2 Grundlagen der Unendlich-dimensionalen Optimierung

Bisher wurden ausschließlich Optimierungsprobleme im $\mathbb{R}^n$ behandelt. Um auf ähnliche Weise in den folgenden Kapiteln Optimierungsprobleme in unendlich-dimensionale Räumen lösen zu können, müssen zuallererst die gängigen Ableitungsbegriffe entsprechend verallgemeinert werden. Seien hierzu X, Y Banachräume, $\mathcal{L}(X, Y)$ der Raum der linearen, beschränkten Abbildungen von X nach Y.

Definition 3.2.1:
Sei $D \subseteq X$ offen. Die Abbildung $J : D \to Y$ ist ***Fréchet-differenzierbar*** in $x_0 \in D$, wenn ein Operator $A_{x_0} \in \mathcal{L}(X, Y)$ existiert, sodass

$$\lim_{x \to x_0} \frac{\|J(x) - J(x_0) - A_{x_0}(x - x_0)\|_Y}{\|x - x_0\|_X} = 0$$

gilt.

Hierbei gelten folgende Äquivalenzen:

1. J ist (Fréchet-)differenzierbar in x_0.

2. Es existiert ein Operator $A_{x_0} \in \mathcal{L}(X,Y)$ und eine in x_0 stetige Abbildung $r_{x_0} : X \to Y$ mit $r_{x_0}(x_0) = 0$
$$J(x) = J(x_0) + A_{x_0}(x - x_0) + r_{x_0}(x) \|x - x_0\|_X$$

3. Es existiert eine Abbildung $A_{x_0} \in \mathcal{L}(X,Y)$ mit
$$J(x) = J(x_0) + A_{x_0}(x - x_0) + o(\|x - x_0\|_X)$$

Ferner ist J in x_0, wenn J in x_0 differenzierbar ist, auch stetig. Der Operator A_{x_0} ist eindeutig bestimmt, er wird kurz mit $\partial J(x_0)$ bezeichnet: Die ***Fréchet-Ableitung*** von J in x_0. Aus 3 folgt für $g : X \to Y, x \mapsto J(x_0) + \partial J(x_0)(x - x_0)$

$$\lim_{x \to x_0} = \frac{\|J(x) - g(x)\|}{\|x - y_0\|} = 0,$$

g stellt ist eine lineare Approximation an J in x_0. Existiert der einseitige Grenzwert

$$\delta J(x;h) = \lim_{t \searrow 0} \frac{J(x+th) - J(x)}{t},$$

so heißt dieser ***Richtungsableitung*** oder ***(erste) rechte Variation*** von J in y.

Definition 3.2.2:
Die Abbildung J heißt ***Gâteaux***-differenzierbar in $x \in X$, wenn $\delta J(x;h)$ für alle $h \in X \setminus \{0\}$ existiert und ferner
$$J'(x) : X \to Y, h \mapsto \delta J(x;h)$$
linear und beschränkt ist. $J'(x)$ heißt ***Gâteaux-Differential*** von J in x.

Ähnlich wie beim Fréchet-Differential lässt sich J in x_0 mittels $\tilde{g} : X \to Y, x \mapsto J(x_0) + J'(x_0)(x - x_0)$ approximieren, jedoch ist der Fehler hier im Allgemeinen nur in $O(\|x - x_0\|)$, wobei im Falle des Fréchet-Differentials der Fehler wie bereits gesehen in $o(\|y - y_0\|)$ liegt. Ist J Gáteaux-differenzierbar, so kann die Gáteaux-Ableitung mittels

$$J'(x)h = \frac{\mathrm{d}}{\mathrm{d}t} J(x+th)\bigg|_{t=0}$$

berechnet werden, was auch der ersten Variation von J in Richtung h entspricht. Die Ableitung

$$\delta^2 J(x;h) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} J(x+th)\bigg|_{t=0} \tag{3.2.1}$$

heißt - sofern existent - ***zweite Variation*** von J in x in Richtung h.

Ist $Y = \mathbb{R}$, so ist die Gáteaux-Ableitung von J ist ein Element des Dualraums $X^* = \mathcal{L}(X, \mathbb{R})$. Betrachtet man das Gáteaux-Differential jedoch auf einem Hilbert-Raum H, so erlaubt der Rieszsche Darstellungssatz die Identifizierung von $J'(x) \in X^*$ mit einem ***Gradient*** genanntem Element $\nabla J(x) \in H$, sodass

$$J'(x)h = \langle \nabla J(x), h \rangle_H \, \forall h \in H$$

gilt. Ferner genügt selbiges $J'(x)$, betrachtet auf einem kleineren Hilbertraum $\tilde{H} \hookrightarrow H$ ebenfalls einer Bedingung der Form

$$J'(x)\tilde{h} = \left\langle \nabla J(x), \tilde{h} \right\rangle_{\tilde{H}} \, \forall \tilde{h} \in \tilde{H}$$

Der Gradient $\nabla \tilde{J}(x)$ bzgl. $\langle \cdot, \cdot \rangle_{\tilde{H}}$ kann sich hierbei vom Gradienten $\nabla J(x)$ unterscheiden.

Ist J konvex und Gâteaux-differenzierbar auf $V \subseteq H$, so gelten:

(1) $J(y) \geq J(x) + \langle \nabla J(x), y - x \rangle_H \, \forall y, x \in V$

(2) $\langle J(y) - J(x), y - x \rangle_H \geq 0$

Analog zu Funktionen im $\mathbb{R}^n$ ist ein Funktional hierbei (strikt) konvex, wenn

$$J(\lambda x + (1 - \lambda y)) \overset{(<)}{\leq} \lambda J(x) + (1 - \lambda) J(y) \; \forall x, y \in V, \lambda \in (0, 1)$$

gilt.

Satz 3.2.3:
Sei $(X, \|\cdot\|)$ normierter Raum und $a \in X$. Das Funktional

$$f_a : X \to \mathbb{R}, x \mapsto \|x - a\|$$

ist stetig und konvex.

Beweisidee/-skizze: Nachrechnen. □

Das Konzept des Gradienten kann auch verallgemeinert werden:

Definition 3.2.4:
Sei $J : X \to \mathbb{R}$ Fréchet-differenzierbar. Das Element $\partial J(x)$ aus dem Dualraum X^* heißt ***Gradient*** von f in x und wird auch mit $\nabla f(x)$ bezeichnet.

Bemerkung 3.2.5:

- *Aus den Definitionen folgt sofort: Existiert das Fréchet-Differential von J in x_0, dann auch das Gâteaux-Differential und die beiden Differentiale stimmen überein.*
- *Aus der Definition folgt für Fréchet-differenzierbare Funktionale J_1 und J_2 unmittelbar die Summenregel: Für beliebige in x Fréchet-differenzierbare Funktionen $\alpha_1, \alpha_2 \in \mathbb{R}$ ist $\alpha_1 J_1 + \alpha_2 J_2$ ebenfalls in x_0 Fréchet-differenzierbar, es gilt*

$$\partial \left(\alpha_1 J_1 + \alpha_2 J_2 \right)(x_0) = \alpha_1 \partial J_1(x_0) + \alpha_2 \partial J_2(x_0)$$

- *Sind $D \subseteq X, E \subseteq Y$ offen, Z Banachraum, $J_1 : D \to Y$ mit $J_1(D) \subseteq E$ in x_0 sowie $J_2 : E \to Z$ in $y_0 = J_1(x_0)$ Fréchet-differenzierbar, so ist $J_2 \circ J_1$ ebenfalls Fréchet-Differenzierbar in x_0 mit $\partial(J_2 \circ J_2)(x) = \partial J_2'(f(x_0)) J_1'(x)$.*

In den folgenden Kapiteln ist auch das Konzept der schwachen Konvergenz von großer Bedeutung:

Definition 3.2.6:
Eine Folge $(x_k)_k \subset X$ ***konvergiert schwach*** gegen $x \in X$, $(x_k \rightharpoonup x)$, wenn für alle linearen, stetigen Funktionale $l : X \to \mathbb{R}$ gilt:

$$\lim_{k \to \infty} l(x_k) = l(x)$$

.

Ebenso fällt immer wieder der Begriff der Reflexivität:

Definition 3.2.7:
Ein normierter Raum X heißt ***reflexiv***, wenn er isomorph zu seinem Bidualraum ist, also

$$X^{**} \cong X,$$

wobei $X^* = \mathcal{L}(X, \mathbb{R})$ und $X^{**} = (X^*)^*$ ist.

Bemerkung 3.2.8: *Hilberträume, wie z. B. der $L^2(a, b)$ sind reflexiv.*

Für eine schwach konvergente Folge gilt nach Definition stets $l(x_k) \to l(x)$ für lineares und stetiges l, jedoch kann durch aus $x_k \nrightarrow x$ gelten.

Definition 3.2.9:
Sei B Banachraum, $J : B \to \mathbb{R}$ ein Funktional, welches von unten beschränkt ist. Eine Folge $(x_k)_k \subset B$ heißt ***minimierende Folge***, wenn $\lim\limits_{k\to\infty} J(x_k) = \inf\limits_{y\in B} J(y)$ gilt.

3.3 Gewöhnliche Differentialgleichungen und Variationsrechnung

Bei vielen Modellierungsproblemen, welche mit gewöhnlichen Differentialgleichungen in Zusammenhang stehen, handelt es sich um Optimierungsprobleme in Funktionenräumen, welche Differentialrechnung für Funktionale in normierten Vektorräumen umfassen. In der Mechanik beispielsweise könnte man (wie bereits gesehen) daran interessiert sein, eine Trajektorie (einen Pfad) $y(x), x \in [a, b]$ zu finden, sodass ein problemspezifisches Funktional $J(y)$ minimiert wird. Unter gewissen Voraussetzungen kann der optimale Pfad als Lösung einer gewöhnlichen Differentialgleichung charakterisiert werden.

Mit Blick auf die Anwendung fokussiert sich das folgende Kapitel auf den reellen Banachraum $B = C^1([a, b])$ $(a < b)$ sowie den reellen Hilbertraum $H = H^1(a, b)$, sodass $B \subseteq H$ mit stetiger, injektiver Einbettung $\|v\|_H \leq c \|v\|_B \forall v \in B$ und ein $c > 0$ gilt. Ferner sei mit $J : V \to \mathbb{R}$ mit nichtleerer Teilmenge $V \subseteq B$ das Funktional bezeichnet, welches das Optimierungsproblem repräsentiert, d. h. dass ein $y^* \in V$ gesucht ist, welches

$$J(y^*) = \min_{y\in V} J(y) \tag{3.3.1}$$

erfüllt. Dies führt auf die folgende Definition:

Definition 3.3.1:
$y^* \in V$ heißt ***schwaches lokales Minimum*** (alternativ Minimierer) (SLM) von J in V, falls ein $\rho > 0$ existiert mit

$$J(y^*) \leq J(y) \qquad \forall y \in V, \|y - y^*\|_B < \rho$$

Gilt zusätzlich für alle $y \in V \setminus \{y^*\}$ mit $\|y - y^*\|_B < \rho$ $J(y^*) < J(y)$, so heißt y ***striktes SLM***. y^* heißt starkes lokaler Minimierer, wenn es ebenfalls Minimierer bezüglich einer Norm ist, welche eine größere Klasse von Vergleichselementen erlaubt, zum Beispiel $C([a, b])$.

Die Existenz von Lösungen für (3.3.1) kann gezeigt werden, sofern gewisse Voraussetzungen erfüllt sind. Hierzu ist der der Begriff der Halbstetigkeit von Nöten:

Definition 3.3.2:
Ein Funktional J ist ***(folgen)unterhalbstetig***, in $y \in V$, wenn

$$J(y) \leq \liminf_{k\to\infty} J(y_k) \tag{3.3.2}$$

für alle Folgen $(y_k)_k \subseteq V$ mit $y_k \to y$. J ist ***schwach (folgen)unterhalbstetig***, falls (3.3.2) für alle Folgen $(y_k)_k \subseteq V$, welche schwach gegen y konvergieren ($y_k \rightharpoonup y$), das heißt für die $f(y_k) \to f(y)$ für alle linearen, stetigen Funktionale $f : V \to \mathbb{R}$ gilt.

Satz 3.3.3:
Sei $J : V \to \mathbb{R}$ unterhalbstetig sowie die Niveaumenge

$$\mathcal{L}_J(C) \coloneqq \{y \in V | J(y) \leq C\} \tag{3.3.3}$$

nichtleer und kompakt für ein $C \in \mathbb{R}$. Dann existiert ein globales Minimum von J in V.

Beweis: Wegen $J(y) > C \forall y \notin \mathcal{L}_J(C)$ ist klar, dass ein für die Niveaumenge globaler Minimierer ein für ganz V globaler Minimierer ist. Es bleibt zu zeigen, das ersterer existiert. Hierzu sei die Familie von Mengen $(\{y \in V | J(y) > c\})_{c\in\mathbb{R}}$ gegeben, welche $\mathcal{L}_J(C)$ überdecken. Da aus der Unterhalbstetigkeit folgt, dass die Mengen $(\{y \in V | J(y) \geq c\})$ für alle $c \in \mathbb{R}$ abgeschlossen sind, sind alle Mengen dieser Familie offen. Aufgrund der Kompaktheit von $\mathcal{L}_J(C)$ lässt sich aus dieser Familie eine endliche Teilüberdeckung auswählen, wodurch $\alpha = \inf_{y\in\mathcal{L}_J(C)} J(y) = \inf_{y\in V} J(y)$ wohldefiniert ist. Insbesondere ist J von unten beschränkt für alle $y \in V$. Sei nun $(y_k)_k \subseteq V$ mit $J(y_k) \to \alpha$. Ist $\alpha = C$, so ist jedes Element der Niveaumenge globaler Minimierer. Für $\alpha < C$ gilt für hinreichend großes k_0 $J(y_k) \leq C \forall k \geq k_0$, für $\alpha = C$, daher sei o. E. $(y_k)_k \subseteq \{y \in V | J(y) \leq C\}$. Aufgrund der Kompaktheit existiert eine konvergente Teilfolge $(y_{k_l})_l$ mit Grenzwert $\tilde{y} \in V$. Für diesen gilt

$$\alpha \leq J(\tilde{y}) \leq \liminf_{l\to\infty} J(y_{k_l}) = \alpha,$$

folglich ist $\tilde{y}$ ein globaler Minimierer. □

Ist V Teilmenge eines endlich-dimensionalen Raumes, ist für die Kompaktheit der Niveaumengen Beschränktheit hinreichend. (Abgeschlossen sind sie nach Definition.) Dieser Fall kann zum Beispiel dann auftreten, wenn ein minimierender kubischer Spline gesucht wird. Für ein analoges Resultat in unendlich-dimensionalen Räumen ist folgendes Theorem hilfreich:

Satz 3.3.4:
Sei H ein Hilbert-Raum sowie $(y_k)_k \subseteq H$ eine beschränkte Folge. Dann existiert eine schwach konvergente Teilfolge, d. h. eine Folge $(k_l)_l$ und ein $\tilde{y}$ mit

$$\lim_{l\to\infty} \langle v, y_{k_l}\rangle = \langle v, \tilde{y}\rangle \qquad \forall v \in H$$

(Dies gilt auch allgemein für beliebige reflexive Banachräume, wobei v entsprechend ein lineares, stetiges Funktional bezeichnet.)

Die Existenz eines Minimierers lässt sich damit analog zum Beweis von Satz (3.3.3) auch dann zeigen, wenn J schwach folgenunterhalbstetig ist und für beliebiges $(y)_k$ die Bedingung $J(y) \leq C$ Beschränktheit von y impliziert. Letztere Bedingung führt auf das Prinzip der ***Koerzitivität*** von J.

Definition 3.3.5:
Ein Funktional $J : B \to \mathbb{R}$ ist ***koerzitiv***, wenn

$$\lim_{\|y\|\to\infty} \frac{J(y)}{\|y\|} = \infty \tag{3.3.4}$$

gilt. Gilt hingegen nur

$$\lim_{\|y\|\to\infty} J(y) = \infty \tag{3.3.5}$$

so ist J ***schwach koerzitiv***.

Koerzitivität

Satz 3.3.6:
Sei K nichtleere, abgeschlossene, konvexe Teilmenge eines Hilbert-Raumes H, $J : H \to \mathbb{R}$ ein konvexes, Gâteaux-differenzierbares Funktional. Wenn K beschränkt ist oder J koerzitiv, dann existiert mindestens ein Minimum von J in K, d. h. ein $y \in K$ mit

$$J(y) = \inf_{v\in K} J(v)$$

Beweis: Sei $(y_k)_k \subseteq K$ eine minimierende Folge von J in K, d. h. $J(y_k) \to \inf\limits_{y\in K} J(y)$. Aus formalen Gründen sei hierbei auch $\inf\limits_{y\in K} J(y) = -\infty$ zulässig, im Verlauf des Beweis wird klar, dass das Infimum eine reelle Zahl ist. Da $(y_k)_k$ beschränkt ist (K beschränkt bzw. J koerzitiv), existiert eine Teilfolge $(y_{k_l})_l$ mit $y_{k_l} \rightharpoonup y \in H$. Es gilt:

- $y \in K$: Sei $\bar{y} = \mathcal{P}_K(y)$ die Projektion von y auf K, dass heißt die Lösung $v^* \in K$ von $\min\limits_{v\in K} \|y - v\|$. Da K konvex ist, gilt $\langle y - \bar{y}, w - \bar{y}\rangle \leq 0$ für alle $w \in K$, inbesondere für $y_{k_l} = w$. Folglich gilt wegen $\langle y - \bar{y}, y_{k_l}\rangle \to \langle y - \bar{y}, y\rangle$ und $\langle y - \bar{y}, y_{k_l} - \bar{y}\rangle \leq 0$ $0 \leq \langle y - \bar{y}, y - \bar{y}\rangle \leq 0$, also $y = \bar{y}$ und insbesondere $y \in K$.
- $J(y) \leq \liminf\limits_{l\to\infty} J(y_{k_l})$: Da J konvex und Gâteaux-differenzierbar ist, gilt $J(y_{k_l}) \geq J(y) + \langle J'(y), y_{k_l} - y\rangle$. Aus $\langle J'(y), y_{k_l} - y\rangle \to 0$ für $l \to \infty$ folgt, dass J schwach unterhalbstetig ist.

Es gilt also: $y \in K$ und $J(y) \leq \liminf\limits_{l\to\infty} J(y_{k_l}) = \inf_{v\in K} J(v)$, d. h. y ist ein Minimum in K. □

Proposition 3.3.7:
Sei J konvex. Dann ist J genau dann unterhalbstetig, wenn J schwach unterhalbstetig ist.

Beweisidee/-skizze: Für eine beliebige schwach konvergent Folge $(y_k)_k$ mit (schwachem) Grenzwert $y \in B$ wird eine Teilfolge $(y_{k_n})_n$ mit $\lim\limits_{n\to\infty} J(y_{k_n}) = \liminf\limits_{k\to\infty} J(y_k) = \alpha$ sowie eine Nullfolge $(\varepsilon_m)_m \subseteq \mathbb{R}^+$ gewählt. (Man beachte $\alpha \in \mathbb{R}$.) Die Niveaumengen $\mathcal{L}_J(\alpha + \varepsilon_m) = \{y \in B | J(y) \leq \alpha + \varepsilon_m\}$ sind abgeschlossen und konvex. Wegen $y_{k_n} \rightharpoonup y$ und $y_{k_n} \in \mathcal{L}_J(\alpha+\varepsilon_m)$, n hinreichend groß, folgt aus letzterem $y \in \bigcap_{m\in\mathbb{N}} \mathcal{L}_J(\alpha + \varepsilon_{k_n})$. Nun folgt die Behauptung aus $\varepsilon_k \to 0$. □

Satz 3.3.8:
Sei B reflexiver Banachraum sowie

a) $U \subset B$ nichtleer, abgeschlossen und konvex

b) $J : U \to \mathbb{R}$ konvex und unterhalbstetig.

c) U beschränkt oder J koerzitiv.

Dann existiert ein $y^* \in U$, welches das Minimierungsproblem

$$\min_{y\in U} J(y)$$

löst. Ist J darüberhinaus strikt konvex, so ist die Lösung sogar eindeutig.

Ein Minimierungsproblem kann gelöst werden, wenn es eine minimierende Folge $(y_k)_k$ ermöglicht, für die die Folge $(J(y_k))_k$ monoton fällt und gegen einen reellen oder einen uneigentlichen Grenzwert $(-\infty)$ konvergiert.

Problematisch hierbei ist, dass die Menge der differenzierbaren Funktionen offen ist und die Grenzwerte von Folgen differenzierbarer Funktionen nicht zwangsweise differenzierbar sein müssen. So konvergiert zum Beispiel die Folge $(y_k)_k, y_k = \frac{k}{\sqrt{\pi}} \exp(-k^2x^2)$ gegen die Delta-Distribution (welche nicht mal L^2-Funktion ist). Allerdings ist $(y_k)_k$ auch keine Cauchyfolge im Banachraum $C^1([a, b])$.

Im Folgenden sollen Funktionale der Form

$$J(y) = \int_a^b f(x, y(x), y'(x))\mathrm{d}x \tag{3.3.6}$$

mit stetigem f betrachtet werden. In diesem Falle garantieren die folgenden Bedingungen die Existenz einer Lösung des Variationsproblems $\min\limits_{v\in V} J(v)$:

a) Die Funktion f wächst superlinear bzgl. y', d. h.

$$\lim_{|y'|\to\infty} \frac{f(x, y, y')}{|y'|} = \infty \qquad \forall x, y$$

Durch diese Bedingung sind keine Sprünge des Minimums möglich.

b) Das Funktional J wächst unbegrenzt für $|y| \to \infty$. (Dies verhindert ein „Explodieren“ der Lösung.)

c) Die Funktion f ist konvex bezüglich y' für alle x, y.

Hierbei garantieren a) und b) die Beschränktheit jeder minimierenden Folge. Nach Satz (3.3.4) für einen Hilbertraum H existiert eine in H schwach konvergente Teilfolge.

Nun wird der Fall $J : H \to \mathbb{R}$ mit $H = H^1(a,b)$ betrachtet. (Ein solcher Raum wird ***Sobolevraum*** genannt, es gilt $H^1(a,b) = \left\{ f \in C\left((a,b)\right) | \exists g : (a,b) \to \mathbb{R} \text{ mit } f(x) = \int_a^x g(t)\mathrm{d}t \right\}$.) Wenn f stetig und von unten beschränkt ist sowie Bedingung c) genügt, dann ist J schwach unterhalbstetig auf $H^1(a,b)$, womit die Existenz einer Lösung des Minimierungsproblems gezeigt werden kann.

Gegenbeispiele

1) Gegeben sei das Funktional $J(y) = \int_{-1}^{1} x^2 (y')^2 \mathrm{d}x$ sowie die Menge

$$V = \left\{ v \in C^1([a,b]) | v(-1) = -1, v(1) = 1 \right\}$$

Hier ist f konvex in y', von unten beschränkt sowie stetig. Nichtsdestotrotz gilt a) nicht: Für $x = 0$ ist $\lim \lim_{|y'| \to \infty} \frac{x^2 (y')^2}{|y'|} = 0$. Das Funktional hat keinen Minimierer in V, nicht mal in C^1. Jedoch wird es von $y(x) = -1 + 2H(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$ minimiert, wobei H die Heaviside-Funktion bezeichnet, und $y'(x) = 2\delta(x)$. Die Delta-Distribution ist hierbei die Ableitung der Heaviside-Funktion im Sinne der Distributionen und nicht die schwache Ableitung! (Letztere existiert nicht.)

2) Das Funktional des minimalen Flächeninhaltes (s. u.) ,

$$J(y) = \int_a^b y(x) \sqrt{1 + (y'(x))^2} \mathrm{d}x, \quad y(a) = 1, y(b) = 1$$

genügt ebenfalls nicht der Bedingung a) ($\lim_{|y'| \to \infty} \frac{f(x,y,y')}{|y'|} = y$). Überschreitet die Differenz $b - a$ einen gewissen Schwellenwert, so existiert keine klassische Lösung. Eine Aushilfe bietet hier die sogenannte ***Goldschmidt-Lösung***: Sie besteht aus den beiden Kreisscheiben, die durch die Liniensegmente zwischen $(a,0)$ und $(a,y(a))$ bzw. $(b,0)$ und $(b,y(b))$ erzeugt werden.

3) Gegeben sei das Funktional

$$J(y) = \int_0^1 \left(1 - (y'(x))^2\right)^2 \mathrm{d}x, \qquad y(0) = 0,\ y(1) = 0.$$

Der minimale Wert ist 0, und er wird von „Zick-Zack-Funktionen“ angenommen, deren Ableitung fast überall ± 1 ist, und welche an den Randpunkten den Wert 0 annehmen. Diese liegen nicht in $C^1([0,1])$, aber in $H^1(0,1)$. Das Funktional erfüllt a) und b), aber nicht c).
J ist nicht schwach unterhalb stetig: Sei $y(x) = \begin{cases} x,\ 0 \leq x \leq \frac{1}{2} \\ 1 - x,\ \frac{1}{2} < x \leq 1 \end{cases}$, $\bar{y}$ die periodische Fortsetzung auf $\mathbb{R}$. Die Folge $(y_k)_k$, $y_k(x) = \frac{1}{k}\bar{y}(kx)$ liegt in H^1, ist eine minimierende Folge mit $\lim_{k \to \infty} J(y_k) = 0$ und konvergiert in $H^1(0,1)$ schwach gegen die Nullfunktion. Jedoch gilt $J(0) = 1 > \lim_{k \to \infty} J(y_k)$.

Diese Beispiele zeigen, dass einige Variationsprobleme keine klassichen (d. h. C^1) Lösungen zulassen, jedoch solche, die zu einer großeren Klasse von Funktionen gehören. Um diese Schwierigkeiten zu beheben, gibt es im Wesentlichen zwei Lösungsansätze:

1. Vergrößerung der zulässigen Lösungsmenge, zum Beispiel durch Zulassen stückweise glatter Funktionen. Hierfür wird die sogenannte ***Weierstraß-Erdmann-Bedingung*** genauer untersucht werden.

2. Modifizierung des Variationsproblems, sodass „irreguläre“ Lösung vermieden, jedoch durch reguläre approximiert werden.

Der erste Ansatz wird ***Relaxierung*** genannt und später untersucht werden. Der zweite Ansatz nennt sich ***Regularisierung***, zu seiner Veranschaulich soll erneut das erste Gegenbeispiel betrachtet werden. Hierzu wird der Integrand um eine strikt in y' konvexe Funktion sowie ein Gewicht ε wie folgt erweitert:

$$\tilde{J}(y) = \int_{-1}^{1} x^2(y'(x))^2 + \varepsilon(y'(x))^2 \mathrm{d}x$$

(Die Anfangsbedingungen bleiben gleich.) Hierdurch ist Bedingung a) erfüllt für alle x und y. Daher hat das Regularisierungsproblem eine Lösung, welche sogar C^1 ist:

$$y_\varepsilon(x) = c \cdot \arctan\left(\frac{x}{\sqrt{\varepsilon}}\right)$$

mit $c = \left(\arctan\left(\frac{1}{\sqrt{\varepsilon}}\right)\right)^{-1}$. Die Wahl des Regularisierungsterms ist in gewissem Rahmen willkürlich. In der Praxis kann und sollte dieser daher so gewählt werden, dass die Regularisierungslösung problemspezifischen Anforderungen genügt. Ist beispielsweise b) gefordert, aber nicht erfüllt, liese sich εy^2 als Regularisierungsterm verwenden. Alternativ liese sich auch $\varepsilon\,(y - \bar{y})^2$ verwenden, wobei $\bar{y}$ ein gewünschtes „Profil “ der Lösung beschreibt. Soll eine zu starke Krümmung der Lösung bestraft werden, wäre $\varepsilon\,(y'')$ denkbar. Diese Form der Regularisierung nennt sich ***Thikonov-Regularisierung***.

3.3.2 Optimalitätsbedingungen

Nachdem untersucht wurde, wann unter welchen Bedingungen Minimierer existieren, sollen diese nun charakterisiert werden, basierend auf den Ableitungen der Funktion $J : V \to \mathbb{R}$, $V \subseteq B$.

Existiert der einseitige Grenzwert

$$\delta^2 J(y; v, w) = \lim_{t \searrow 0} \frac{J'(y + tw)h - J'(y)h}{t}$$

für $y \in V, h, w \in B \setminus \{0\}$, so heißt er ***zweite Variation*** von J in y in Richtung h und w.

Definition 3.3.9:
Das Funktional ist ***zweimal Gâteaux-differenzierbar*** in $y \in V$, wenn

i $J'(z)h$ existiert, linear und stetig in h für alle z in einer Umgebung von y ist.

ii $\delta^2 J(y; h, w)$ für alle $h, w \in B \setminus \{0\}$ existiert und $J''(y) : B \times B \to \mathbb{R}, (h, w) \mapsto \delta^2 J(y; h, w)$ eine stetige, bilineare und symmetrische Abbildung in y und w darstellt.

J'' heißt zweites Gâteaux-Differential von J in y.

Proposition 3.3.10:
Ist das Funktional J zweimal Gâteaux-differenzierbar in $y + \theta h$, $\forall \theta \in [0, 1]$, so existiert ein $\theta_0 \in (0, 1)$ sodass

$$J(y + h) = J(y) + J'(y)h + \frac{1}{2} J''(y + \theta_0)(h, h)$$

Definiert man $\phi(t) \coloneqq J(y+th)$ und nimmt $\phi \in C^2(-\varepsilon, \varepsilon)$ an, dann gilt

$$J(y+th) = J(y) + tJ'(y)h + \frac{t^2}{2}J''(y)(h,h) + r(th)$$

für ein geeignetes r mit $\lim_{t\to 0} \frac{r(th)}{t^2} = 0$ für festes h.
Angenommen, die stetige Bilinearform $J''(y)(\cdot,\cdot)$ auf $B \times B$ lässt sich zu einer stetigen, symmetrischen Bilinearform auf $H \times H$ ergänzen. Dann existiert ein beschränkter linearer Operator $\nabla^2 J(y) : H \times H$, sodass

$$J''(y)(h,w) = \left\langle \nabla^2 J(y)h, w \right\rangle \forall h, w \in H$$

gilt.

Nun soll untersucht werden, unter welchen Bedingungen eine Funktion y überhaupt als Kandidat für ein Optimum in Frage kommt. Hierzu ist die folgende Definition notwendig:

Definition 3.3.11:
Der ***lokale Tangentialkegel*** $T(V,y)$ von V in y ist die Menge aller $w \in H$, sodass eine Folge $(y_n)_{n\in\mathbb{N}} \subseteq V$ sowie eine Nullfolge $(t_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^+$ existieren, sodass $y_n \to y$ in B und $\frac{y_n - y}{t_n} \to w$ in H gelten:

$$T(V,y) = \left\{ w \in H | \exists\, (y_n)_{n\in\mathbb{N}} \subseteq V, \exists\, (t_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^+ : y_n \to_B y, t_n \searrow 0, \frac{y_n - y}{t_n} \to_H w \right\}$$

Der Name ist nicht willkürlich, tatsächlich ist $T(V,y)$ ein Kegel (mit Scheitelpunkt 0): Gilt $d \in T(V,y)$ für ein $d \in H$, so gilt $\alpha d \in T(V,y)$ für alle $\alpha > 0$. $T(V,y)$ ist schwach abgeschlossen in H und konvex, wenn V konvex ist. Einfach gesagt kann $T(V,y)$ als lokale Approximation an V in y angesehen werden. Bildlich lässt sich der Tangentialkegel schwer veranschaulichen. Einfacher ist dies mit der Menge der ***zulässigen Richtungen***:

$$F(V,y) = \{ w \in B | \exists \varepsilon_0 > 0 : y + \varepsilon w \in V \forall \varepsilon \in (0, \varepsilon_0) \}.$$

$T(V,y)$ ist der Abschluss von $F(V,y)$ in H.
Für $y, h \in B$ gelte

$$J(y+h) = J(y) + \langle \nabla J(y), h \rangle + \eta(\|h\|_B) \cdot \|h\|_H,\ \|h\|_B \to 0, \tag{3.3.7}$$

mit beschränktem, linearen Funktional $\langle \nabla J(y), \cdot \rangle$ und stetigem η, welches $\eta(t) \to 0$ für $t \to 0$ erfüllt. (Dies impliziert Fréchet-Differenzierbarkeit: Man wähle $r_y(z) = \eta(\|z-y\|_B)$.)

Satz 3.3.12 (Notwendige Optimalitätsbedingungen):
Sei $V \subseteq B$ nichtleer, $y \in V$ ein schwaches lokales Minimum von J in V. Dann gilt

$$\langle \nabla J(y), w \rangle \geq 0 \qquad \forall w \in T(V,y)$$

Beweis: Seien $(t_n)_n$, $(y_n)_n$ die Folgen aus der Definition des Tangentialkegels, für die $y_n \to y$, $\frac{y_n - y}{t_n} \to w$ gilt. Dann gilt für hinreichend großes n

$$\begin{aligned} J(y) \leq J(y_n) &= J(y + (y_n - y)) = \\ &= J(y) + \langle \nabla J(y), y_n \rangle + \eta(\|y_n - y\|_B) \|y_n - y\|_H \end{aligned}$$

Somit gilt für hinreichend kleines $t_n < 1$

$$0 \leq \left\langle \nabla J(y), \frac{y_n - y}{t_n} \right\rangle + \eta(\|y_n - y\|_B) \left\| \frac{y_n - y}{t_n} \right\|_H$$

Übergang zum Grenzwert $n \to \infty$ liefert die Behauptung. □

Ein klassisches Problem der Variationsrechnung ist

$$\min_{y \in V} J(y) \coloneqq \int_a^b f(x, y(x), y'(x)) \mathrm{d}x \tag{3.3.8}$$

mit

$$V = \left\{ v \in C^1([a,b]) | v(a) = y_a, v(b) = y_b \right\} \tag{3.3.9}$$

($-\infty < a < b < \infty$, f hinreichend glatt).
Es sei angemerkt, dass für beliebige $\hat{y}, \tilde{y} \in V$, $t_n \coloneqq \frac{1}{n}, n = 1, 2, ...$ sowie $y_n \coloneqq y + t_n (\tilde{y} - \hat{y} - y)$ einerseits $y_n \to y$, andererseits $\frac{y_n - y}{t_n} \to \tilde{y} - \hat{y} =: w$ gilt. Anders ausgedrückt ist $V^0 \coloneqq V - V \subset T(V, y)$. Eine andere Schreibweise für V^0 ist

$$V^0 = \left\{ v \in C^1([a,b]) | v(a) = 0, v(b) = 0 \right\}$$

Somit wird aus der notwendigen Bedingung

$$\langle \nabla J(y), w \rangle = 0 \forall w \in V^0 \tag{3.3.10}$$

Als nächstes soll die erste Variation von J wie in (3.3.8) mit $h \in V^0$ untersucht werden unter der Annahme, dass f C^2 in all seinen Argumenten ist.

$$\delta J(y; h) = \lim \lim_{\alpha \to 0^+} \frac{1}{\alpha} \left(\int_a^b f(x, y + \alpha h, y' + \alpha h') \right) - f(x, y, y') \mathrm{d}x \tag{3.3.11}$$

Da $f \in C^2$ ist, lässt sich die Taylor-Entwicklung betrachten:

$$\begin{aligned} f(x, y + \alpha h, y' + \alpha h') = f(x, y, y') + \alpha h \frac{\partial f}{\partial y}(x, y, y') + \alpha h' \frac{\partial f}{\partial y'}(x, y, y') \\ + \frac{\alpha^2}{2} \left(h \frac{\partial}{\partial y} + \frac{\partial}{\partial (y')^2} \right)^2 f(x, y + \theta h, y' + \theta h'),\ \theta \in (0, \alpha) \end{aligned}$$

Substituiert man dies in (3.3.11) und betrachtet den Grenzwert $\lim_{\alpha \searrow 0}$, führt dies auf

$$\delta J(y; h) = \int_a^b \left(\frac{\partial}{\partial y} f(x, y(x), y'(x)) h(x) + \frac{\partial}{\partial y'}(x, y(x), y'(x) h'(x) \right) \mathrm{d}x$$

Ist $y \in V$ eine Lösung von (3.3.8)-(3.3.9), muss diese daher

$$\int_a^b \left(\frac{\partial}{\partial y} f \cdot h + \frac{\partial}{\partial y'} f \cdot h' \right) \mathrm{d}x = 0 \forall h \in V - V \tag{3.3.12}$$

Eine Funktion $y \in V$, welche (3.3.12) erfüllt, wird ***schwache Extremale*** genannt. Integriert man (3.3.12) partiell und verwendet $h(a) = h(b) = 0$, so ergibt dies ein notwendiges Kriterium der Form

$$\int_a^b \left(\frac{\partial}{\partial y} f(x, y, y') - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial}{\partial y'} f(x, y, y') \right) h(x) \mathrm{d}x = 0 \forall h \in V^0 \tag{3.3.13}$$

Mit Hilfe der Funktion $A(t) = \int_a^t \frac{\partial}{\partial y} f(x, y, y') \mathrm{d}x$ ergibt sich

$$\int_a^b \frac{\partial}{\partial y} f(x, y(x), y'(x)) h(x) \mathrm{d}x = - \int_a^b A(x) h'(x) \mathrm{d}x.$$

Folglich kann (3.3.12) zu

$$\int_a^b \left(-A(x) + \frac{\partial}{\partial y'} f(x, y(x), y'(x)\right) h(x) \mathrm{d}x \qquad h \in V - V \tag{3.3.14}$$

umgeschrieben werden. Es gilt das folgende Lemma:

Lemma 3.3.13 (du Bois-Reymond):
Sei $\varphi \in C([a,b])$ und $\int_a^b \varphi(x) h'(x) = 0 \ \forall h \in C^1([a,b])$ mit $h(a) = h(b) = 0$. Dann ist φ konstant, es existiert ein $c \in \mathbb{R}$ mit $\varphi(x) = c \forall x \in [a,b]$.

Beweisidee/-skizze: Es lässt sich leicht nachrechnen, dass die Wahl $h(t) = \int_a^t \varphi(x) - c \, \mathrm{d}x$ mit $c = \frac{1}{b-a} \int_a^b \varphi(x) \mathrm{d}x$ zulässig ist und das gewünschte liefert. □

Angewandt auf (3.3.14) ergibt sich aus diesem Lemma

$$-A(x) + \frac{\partial}{\partial y} f(x, y(x), y'(x)) = c, \ x \in [a,b],$$

woraus sich wiederum durch Differentiation die bereits bekannte Euler-Lagrange-Gleichung

$$\frac{\partial}{\partial y} f(x, y(x), y'(x)) - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial}{\partial y'} f(x, y, y') = 0, \ x \in [a,b] \tag{3.3.15}$$

ergibt. Trivialerweise genügt eine Lösung y, welche Gleichung (3.3.15) sowie $y(a) = y_a$ und $y(b) = y_b$ erfüllt, auch der Gleichung (3.3.13).

An dieser Stelle sei noch Folgendes angemerkt: In (3.3.13) wird der Gradient bezüglich des $L^2(a,b)$-Skalarprodukts mit

$$\nabla J(y) = \frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial f}{\partial y}$$

identifiziert. Diese Darstellung unterscheidet sich von der Darstellung mittels (3.3.15),

$$\nabla J(y) = \int_a^x -A(t) + \frac{\partial}{\partial y'} f(t, y(t), y'(t)) \mathrm{d}t$$

Letzteres ist der Gradient bezüglich des Skalarproduktes $\langle v, w \rangle = \int_a^b v' w' \mathrm{d}x$ auf dem Hilbert-Raum $\{v \in C^1([0,1]) | v(a) = v(b) = 0\}$. Ferner gibt es folgende Spezialfälle:

(a) Hängt f nicht explizit von y ab, lautet die Euler-Lagrange-Gleichung

$$\frac{\partial}{\partial y'} f(x, y'(x)) = c.$$

(b) Hängt f hingegen nicht explizit von y' ab, lautet die Euler-Lagrange-Gleichung

$$\frac{\partial}{\partial y} f(x, y(x)) = 0.$$

(c) Hängt f nicht explizit von x ab und gilt ferner $y \in C^2([a,b])$, so entspricht die Euler-Lagrange-Gleichung der Gleichung

$$f(y(x), y'(x)) - y'(x) \frac{\partial}{\partial y'} f(y(x), y'(x)) = c.$$

3.3.3 Anwendungsbeispiel: Minimale Rotationsfläche

Gesucht ist die/eine Kurve y zwischen zwei Punkten (a, y_a) und (b, y_b), sodass die zugehörige Rotationsfläche, d. h. die Fläche die durch Rotation einer Kurve um die x-Achse entsteht, minimal wird. Der gesuchte Flächeninhalt $J(y)$ lässt sich hierbei wie folgt bestimmen:

Man unterteilt die Kurve in infinitesimale Bogenstücke der Länge ds und bezeichnet mit $y(s)$ die Höhe auf dem Mittelpunkt zwischen den beiden Endpunkten des Bogenstücks (vgl. Abbildung 3.3.1). Der entsprechende Teil der Rotationsfläche wird dann durch ein Band mit Flächeninhalt $2\pi y(s) \cdot ds$ approximiert. Grenzwertübergang $ds \to 0$ liefert

$$J(y) = 2\pi \int_0^L y(s)\mathrm{d}s,$$

wobei L die Länge der Kurve ist.

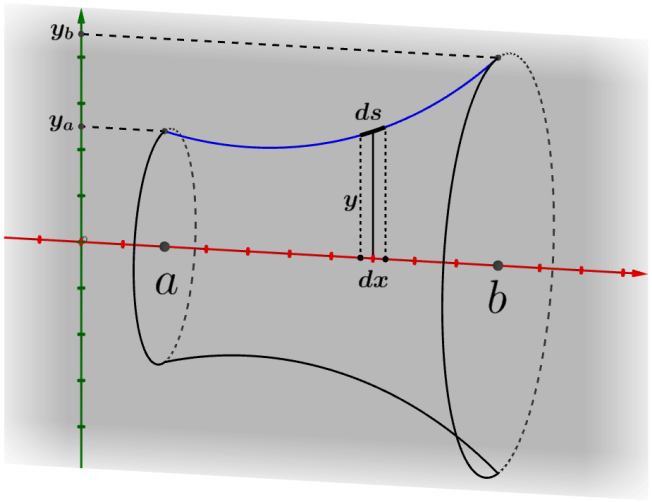


Abbildung 3.3.1: Kettenlinie

Für ds gilt wiederum näherungsweise (Pythagoras) $ds = \sqrt{(dx)^2 + (dy)^2}$, wobei dx der x-Abstand der Endpunkte ist, dy der entsprechende Höhenunterschied. Anders ausgedrückt, ist $ds = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}dx$, woraus sich

$$J(y) = 2\pi \int_a^b y(x)\sqrt{1 + (y'(x))^2}\mathrm{d}x \tag{3.3.16}$$

als Flächeninhalt ergibt. Dieses Minimal gilt es über dem Raum V aus (3.3.9) zu minimieren. Da $f(y, y') = y\sqrt{1 + (y')^2}$ nicht explizit von x abhängt, gilt der Spezialfall (c) und die Euler-Lagrange-Gleichung lautet wie folgt:

$$y\sqrt{1 + (y')^2} - y\frac{(y')^2}{\sqrt{1 + (y')^2}} = c, \tag{3.3.17}$$

Hieraus ergibt sich $y\sqrt{1 + (y')^2} = c$. Die allgemeine Lösung dieser Differentialgleichung ist durch $y(x) = c \cdot \cosh\left(\frac{x+c_1}{c}\right)$ gegeben. Die zugehörige Kurve heißt ***Kettenlinie***, sie entspricht der Form, welche eine zwischen den Punkten (a, y_a) und (b, y_b) aufgehängte Kette mit beweglichen Gliedern annimmt. Die entsprechende Rotationsfläche heiß ***Katenoid***.
Die Konstanten werden durch die Bedingungen

$$y_a = c \cdot \cosh\left(\frac{a + c_1}{c}\right), \qquad y_b = c \cdot \cosh\left(\frac{b + c_1}{c}\right)$$

bestimmt. Allerdings hat dieses nichtlineare Gleichungssystem u. U. keine Lösung: Sei hierzu $a = -b$ und $y_a = y_b = \bar{y}$. Wegen $\cosh(z) = \cosh(-z)$ gilt

$$\bar{y} = c \cdot \cosh\left(\frac{-b + c_1}{c}\right) = c \cdot \cosh\left(\frac{b - c_1}{c}\right) = c \cdot \cosh\left(\frac{b + c_1}{c}\right)$$

Hieraus lässt sich $c_1 = 0$ sowie

$$\bar{y} = c \cdot \cosh\left(\frac{b}{c}\right)$$

folgern. Diese Gleichung ist nur für $\frac{b}{\bar{y}} < \frac{2}{3}$ lösbar. Tatsächlich liefert Gleichung (3.3.17) nur ein notwendiges, aber kein hinreichendes Kriterium für ein Minimum von 3.3.16.

Obiges Beispiel benötigt, ebenso wie die Spezialfälle (a)-(c), nicht die zweite Ableitung y''. Im Allgemeinen benötigt die Euler-Lagrange-Gleichung jedoch ebendiese, was Problematisch ist, da eine solche möglicherweise nicht existiert. In einem solchen Falle kann folgendes Theorem herangezogen werden:

Satz 3.3.14:
Sei $f \in C^2([a,b])$ und $y' \in C^1([a,b])$ ein schwaches Extremum von (3.3.7). Gilt ferner

$$\frac{\partial^2}{\partial y'^2} f(x,y,y') \neq 0, x \in [a,b],$$

so ist $y \in C^2([a,b])$.

3.3.4 Gleichungssysteme

Betrachtet wird erneut (3.3.8) und (3.3.9), jedoch diesmal mit $y \in C^1\left([a,b], \mathbb{R}^m\right), y(x) = (y_1(x), ..., y_m(x))^T$ sowie $y_a, y_b \in \mathbb{R}^{>} \ (m \in \mathbb{N})$. Ist $y \in C^2$ ein lokales Minimum von $J(y)$, so muss y dem Gleichungssystem

$$\frac{\partial f}{\partial y_j}(x,y,y') - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y'_j}(x,y,y') = 0, \ j = 1, ..., m$$

genügen. Ähnlich zu 3.3.14 ist ein schwaches Extremum $y' \in C^1$ hierbei in C^2, wenn für $f \in C^2$

$$\frac{\partial^2}{\partial y_j} f(x,y,y') \neq 0, \ j,k = 1, ..., m$$

auf $[a,b]$ gilt. Das Optimalitätssystem (3.3.12) ist hierbei durch

$$\int_a^b \sum_{j=1}^m \left(\frac{\partial f}{\partial y'_k} h'_k + \frac{\partial f}{\partial y_k} h_k \right) \mathrm{d}x = 0 \ \forall v \in V - V$$

gegeben. Es sollte klar sein, dass für Probleme höherer Ordnung mit

$$J(y) = \int_a^b f\left(x, y, y', y'', ..., y^{(l)}\right) \mathrm{d}x$$

die Optimalitätsbedingung für ein (skalares) schwaches Minimum

$$\int_a^b \sum_{k=0}^l \frac{\partial}{\partial y^{(k)}} f\left(x, y, y', ..., y^{(l)}\right) \cdot h^{(k)} \mathrm{d}x = 0$$

lautet.

3.3.5 Einseitige Beschränkungen

Sei $V \subset B$ nichtleer und $y \in V$ ein lokaler schwacher Minimierer von J in V. Die Optimalitätsbedingung aus Theorem 3.3.12 ist

$$\langle \nabla J(y), w \rangle \geq 0 \quad \forall w \in T(V, y)$$

Ist V konvex, wird diese Bedingung zu

$$\langle \nabla J(y), v - y \rangle \geq 0 \quad \forall v \in V.$$

Eine typische Problemstellung mit einseitiger Beschränkung wäre

$$V = \left\{ v \in C^1([a,b]) \,|\, v(a) = y_a, v(b) = y_b, v(x) \geq \psi(x), x \in (a,b) \right\}.$$

Als Beispiel lässt sich eine Seite betrachten, welche über ein Hindernis ψ gespannt ist. Die stationäre Konfiguration minimiert

$$J(y) = \frac{1}{2} \int_a^b (y'(x))^2 \, \mathrm{d}x, \; y \in V$$

$$V = \left\{ v \in C^1([a,b]) \,\middle|\, \begin{matrix} v(a) = v(b) = 0 \\ v|_{(a,b)} \geq \psi(x)|_{(a,b)} \end{matrix} \right\},$$

wobei $\psi \in C^1([a,b])$ mit $\psi(a) \leq 0, \psi(b) \leq 0$ ist.

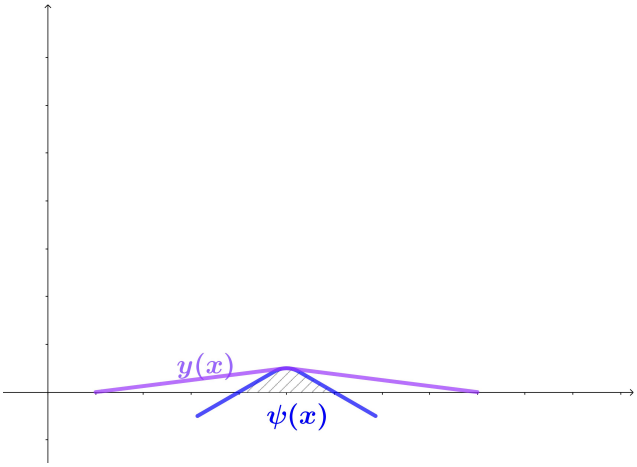


Abbildung 3.3.2: Gespannte Saite

Das Optimalitätskriterium ist dann durch

$$\int_a^b y'(x) \left(v'(x) - y'(x) \right) \mathrm{d}x \geq 0 \qquad \forall v \in V$$

gegeben. Für dieses Problem lässt sich die Existenz einer schwachen Lösung $y \in H_0^1(a,b)$ zeigen. Ferner gilt unter geeigneten Regularitätsbedingungen an ψ, dass diese Lösung y sogar in $C^1([a,b])$ liegt.

3.3.6 Freier Rand

Eine weitere wichtige Klasse von Variationsproblemen ist durch

$$\min_{y \in V} J(y) \coloneqq \int_a^b f(x, y(x), y'(x)) \mathrm{d}x + g(y(a), y(b)), \quad V = C^1([a,b])$$

mit hinreichend glattem $g = g(\alpha, \beta)$ gegeben. Hier ist die Menge, über die minimiert wird, viel größer als in den vorherigen Problemen, da keine Bedingungen an den Rand existieren. In diesem Fall gilt offensichtlich $T(V,y) = V$ und die Optimalitätsbedingung lautet wie folgt:

$$\int_a^b \left(\frac{\partial f}{\partial y} h + \frac{\partial f}{\partial y'} h' \right) \mathrm{d}x + \frac{\partial g}{\partial \alpha}(y(a), y(b)) \cdot h(a) + \frac{\partial g}{\partial \beta}(y(a), y(b)) \cdot h(b) = 0 \forall h \in V \qquad (3.3.18)$$

Da (3.3.18) insbesondere für die h mit $h(a) = h(b) = 0$ gelten soll, führt dies erneut auf Bedingung (3.3.12) und somit die Euler-Lagrange-Gleichung (3.3.15). Im Allgemeinen werden diese durch partielle Integration hergeleitet, womit (3.3.18) zu

$$\int_a^b \left(\frac{\partial f}{\partial y} - \frac{d}{dx} + \frac{\partial f}{\partial y'} h \Big|_a^b + \frac{\partial f}{\partial y'} h' \right) \mathrm{d}x + \frac{\partial g}{\partial \alpha}(y(a), y(b)) \cdot h(a) + \frac{\partial g}{\partial \beta}(y(a), y(b)) \cdot h(b) = 0$$

wird. Da dies für alle $h \in C^1([a,b])$ gelten muss, folgt (durch Wahl von $h(a) = h(b) = 0; h(a) = 1, h(b) = 0$ sowie $h(a) = 0, h(b) = 1$

$$\begin{aligned} \frac{\partial}{\partial y} f(x, y(x), y'(x)) - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial}{\partial y'} f(x, y(x), y'(x)) = 0 \\ -\frac{\partial}{\partial y'} f(a, y(a), y'(a)) + \frac{\partial}{\partial \alpha} g(y(a), y(b)) = 0 \\ \frac{\partial}{\partial y'}(b, y(b), y'(b)) + \frac{\partial}{\partial \beta} g(b, y(b)) = 0 \end{aligned}$$

Die letzten zwei Gleichungen heißen ***freie Randbedingungen***. Sie repräsentieren den einfachen Fall von sogenannten ***Transversalitätsbedingungen***.

3.3.7 Gleichungsnebenbedingungen

Nun soll der Fall betrachtet werden, in welchem das zu minimierende Funktional erneut durch

$$J(y) = \int_a^b f(x, y(x), y'(x)) \mathrm{d}x, \ y \in V \tag{3.3.19}$$

gegeben ist, V hingegen durch

$$V = \left\{ v \in C^1([a,b]) | v(a) = y_a, v(b) = y_b, K(v) = c \right\} \tag{3.3.20}$$

mit einer Konstanten c sowie

$$K : C^1([a,b]) \to \mathbb{R}, y \mapsto \int_a^b \varphi(x, y(x), y'(x)) \mathrm{d}x \tag{3.3.21}$$

Die Gleichung $K(y) = c$ definiert eine Gleichheitsbedingung für y. Seien nun $f, \varphi \in C^2$. Ist $y \in V$ ein schwaches lokales Minimum des Optimierungsproblems (3.3.19)-(3.3.20) und hat das Fréchet-Differential ∂K in einer Umgebung von y eine stetige Inverse, existiert ein $\lambda \in \mathbb{R}$, sodass die Gleichung

$$\frac{\partial}{\partial y} \left(f + \lambda \varphi \right) (x, y(x), y'(x)) - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial}{\partial y'} \left(f + \lambda \varphi' \right) (x, y(x), y'(x)) = 0 \tag{3.3.22}$$

für alle $x \in [a,b]$ erfüllt ist. Die Invertierbarkeit von ∂K bedeutet, dass y kein Extremum von (3.3.21) ist. Hierzu sei $y + \alpha\eta$ eine Variation von y. Diese soll die Gleichheitsbedingung erfüllen, also muss $\delta K(y; \eta) = 0$ für die entsprechende Variation von K gelten, also

$$\int_a^b \left(\frac{\partial}{\partial y} \varphi - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial}{\partial y'} \varphi \right) \eta \mathrm{d}x = 0 \tag{3.3.23}$$

Diese Gleichung selektiert die zulässigen Variationen, und die Optimalitätsbedingung ist durch

$$\delta J(y; \eta) = 0 \quad \forall \eta : \eta \text{ erfüllt (3.3.23)}$$

gegeben. Ist c ein Extremwert von K, so ist η nicht bestimmt und das Problem (3.3.19) -(3.3.21) hat (u. U.) keine Lösung.

Zusammengefasst ist eine Lösung y entweder ein Extremum von K oder von $J + \lambda K$. Also muss y die Euler-Lagrange-Gleichung für $\lambda_0 J + \lambda K$, $\lambda_0, \lambda \in \mathbb{R}$ nicht beide gleich 0, erfüllen. λ_0 heißt ***abnormaler Multiplikator***.

Beispiel 3.3.15:

Unter allen C^1-Kurven in der oberen Halbebene des Koordinatensystems ($y \neq 0$) mit Länge $l = 2\pi r$ ist diejenige gesucht, welche mit dem Intervall $[-r, r]$ die größte Fläche einschließt. Die Fläche unter der Kurve ist durch

$$J(y) = \int_{-r}^{r} y(x)\mathrm{d}x$$

gegeben und es muss $y(-r) = y(r) = 0$ gelten. Die Länge dieser Kurve ist

$$K(y) = \int_{-r}^{r} \underbrace{\sqrt{1 + (y')^2}}_{=\varphi(y')}\mathrm{d}x$$

Die Euler-Lagrange-Gleichung (3.3.22) wird zu

$$1 - \lambda \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y'(x)}{\sqrt{1 + (y'(x))^2}}\right) = 0,$$

woraus

$$\lambda \frac{y'(x)}{\sqrt{1 + (y'(x))^2}} = x + c$$

folgt. Somit ergibt sich

$$\lambda^2 (y'(x))^2 = (x - c)^2(1 + (y'(x))^2) = (x - c)^2 + (x - c)^2 (y'(x))^2$$

Aus Symmetriegründen gilt $y'(0) = 0$, also $c = 0$. Ist nun $\lambda = r$, so ist die Lösung durch $\sqrt{r^2 - x^2}$, also einen Kreis, gegeben.

Eine andere Möglichkeit für Gleichheitsbedingungen ist punktweise Gleichheit. Das Ziel ist es,

$$J(y) = \int_a^b f(x, y(x), y'(x))\mathrm{d}x$$

mit $y = (y_1, ..., y_n) \in C^1([a, b], \mathbb{R}^n)$ und

$$V = \left\{v \in C^1([a, b], \mathbb{R}^n) | v(a) = y_a, v(b) = y_b, e_k(x, v(x)) = 0, x \in [a, b], k = 1, ..., m\right\}$$

zu minimieren. Dabei seien $y_a, y_b \in \mathbb{R}^n$ und $f(x)$ sowie $e(x) \coloneqq (e_1(x), ..., e_m(x))$ C^2-Funktionen, $m < n$. Das Problem $\min\limits_{y \in V} J(y)$ heißt ***Lagrange-Problem***. Die ***Lagrangefunktion*** ist

$$\mathcal{L}(x, y, y', \lambda) = f(x, y, y') + \sum_{k=1}^{m} \lambda_k e_k(x, y)$$

Die λ_k heißen erneut Lagrange-Multiplikatoren, im Gegensatz zur endlichdimensionalen Optimierung handelt es sich hier jedoch um Funktionen $\lambda_k : \mathbb{R}^n \to \mathbb{R}$. Die Bedingungen $e_k(x, y) = 0$ werden ***holonom*** genannt, dies bedeutet, dass e_k nicht von y' abhängt. Es gilt folgendes Theorem:

Satz 3.3.16:
Sei y ein Minimierer von J in V. Ferner existiere eine $m \times m$-Teilmatrix der Jacobimatrix $\frac{\partial e}{\partial y}$, welche für alle $x \in [a,b]$ regulär ist. Dann existieren $\lambda_k \in C^1([a,b])$, sodass folgende Gleichung erfüllt ist:

$$\frac{\partial}{\partial y}\mathcal{L}(x,y,y',\lambda) - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial}{\partial y'}\mathcal{L}(x,y,y',\lambda) = 0, x \in (a,b) \tag{3.3.24}$$

Beispiel 3.3.17 (Geodätische Linie):

Gegeben sei eine durch $\phi(y) = 0, \phi \in C^1(\mathbb{R}^3,\mathbb{R}^3)$ gegebene Mannigfaltigkeit S. $y(x) \in \mathbb{R}^3$ beschreibt hierbei die Koordinaten eines Punktes x. Seien nun $\nabla\phi \neq 0$ und $V = \{v \in C^1([a,b],\mathbb{R}^3) | v(a) = P, v(b) = Q, \phi(v) = 0\}$ die Menge der Wege von P nach Q auf S. Gesucht ist der kürzeste unter diesen Wegen:

$$\min_{y\in V} \int_a^b \underbrace{\sqrt{y'(x)\cdot y'(x)}}_{=\|y'(x)\|_2}\mathrm{d}x$$

Hier ist $\mathcal{L}(x,y,y',\lambda) = \sqrt{y'(x)\cdot y'(x)} + \lambda(x)\cdot\phi(y(x))$ und die Euler-Lagrange-Gleichung (3.3.24) wird zu

$$\lambda(x)\nabla\phi(y(x)) - \frac{\mathrm{d}}{\mathrm{d}x}\frac{y'(x)}{\sqrt{y'(x)\cdot y'(x)}} = 0$$

Bezeichnet s die Bogenlänge, sodass $\tilde{y}(s) = y(x(s))$ die Parametrisierung nach der Bogenlänge ist, so gilt

$$\lambda(s)\phi(y(s)) = y''(s)$$

Hierbei gilt es zu beachten, dass $\nabla\phi \perp S$ und der Hauptnormalenvektor der Kurve durch

$$N(s) = \frac{y''(s)}{\|y''(s)\|}$$

gegeben ist. Der Lagrange-Multiplikator ist also eine Art Skalierungsfaktor, sodass der Normalenvektor der Kurve mit dem der Oberfläche S in jedem Punkt übereinstimmt. Dies stellt eine weitere Möglichkeit der Charakterisierung einer Geodäte dar: $y \in C^2$ beschreibt genau dann den kürzesten Verbindungsweg zwischen zwei Punkten auf der Oberfläche S, wenn parametrisiert nach der Bogenlänge

$$\lambda(s) = \frac{\|y''(s)\|}{\nabla\phi(s)}$$

erfüllt ist.

Betrachtet man das Problem

$$\min_{y\in V} J(y) = \int_a^b f(x,y(x),y'(x))\mathrm{d}x$$

mit $V = \{v \in C^1([a,b]) | v(a) = y_a, v(b) = y_b, e(x,v(x),v'(x)) = 0 \text{ in } (a,b)\}$ mit nicht-holonomen Gleichungsnebenbedingungen, lautet die Lagrangefunktion erneut

$$\mathcal{L}(x,y,y',\lambda) = \lambda_0 f(x,y,y') + \lambda(x)e(x,v(x),v'(x)),$$

wobei wiederum $\frac{\partial e}{\partial v'} \neq 0$ entlang einer gesamten Kurve gelte, welche $e(\cdot,\cdot,\cdot) \equiv 0$ erfüllt. Dann existiert ein $\lambda = \lambda(x)$ (oder nur ein λ_0), sodass die Lösung y (3.3.24) erfüllt.

3.3.8 Optimalitätsbedingungen zweiter Ordnung

Die Optimalitätsbedingungen basierend auf der ersten Variation, wie oben vorgestellt, werden von jedem Extremal J auf V erfüllt, d. h. bei jedem Minimum oder Maximum, aber auch bei jedem Sattelpunkt. Um diese Fälle unterscheiden zu können, muss die zweite Ableitung untersucht werden. Die hierfür nötigen Kriterien werden im Folgenden dargelegt. Zu diesem Zweck wird die zweite Variation (3.2.1) explizit berechnet:

$$\delta^2 J(y;h) = \int_a^b \frac{\partial^2 f}{\partial y^2}(x,y,y')h^2 + 2\frac{\partial^2 f}{\partial y \partial y'}(x,y,y') + \frac{\partial^2 f}{\partial y'^2}(x,y,y')\mathrm{d}x.$$

wobei $f \in C^2$ sei. Ferner sei angenommen, dass eine symmetrische Bilinearform $\nabla^2 J(y)$ auf $H \times H$ mit

$$\left\langle \nabla^2 J(y)h, h \right\rangle = \delta^2 J(y;h)$$

existiere. Eine notwendige Bedingung für ein Minimum im Falle (3.3.8)-(3.3.9) mit $\nabla J(y) = 0$ ist gegeben durch

Satz 3.3.18:
Sei y ein lokaler, schwacher Minimierer, dann gilt

$$\left\langle \nabla^2 J(y)w, w \right\rangle \geq 0 \tag{3.3.25}$$

für alle $w \in V - V$.

Beweis: Sei $\phi(t) \coloneqq J(y + th)$. Dann gilt für hinreichend kleine t

$$\phi(0) \leq \phi(t) = \phi(0) + \phi'(0)t + \frac{1}{2}\phi''(0)t^2 + o(t^2)$$

Somit folgt wegen $\phi'(0) = \langle \nabla J(y), w \rangle = 0$ und $\phi''(0) = \langle \nabla J^2(y)w, w \rangle$

$$0 \leq \frac{1}{2}\left\langle \nabla J^2(y)w, w \right\rangle t^2 + o(t^2)$$

Division durch t^2 und Grenzwertübergang $t \to 0$ liefern schließlich die Behauptung. □

Damit (3.3.25) gilt, muss die sogenannte ***Legendre-Bedingung*** erfüllt sein:

Satz 3.3.19:
Erfüllt $y \in V$ (3.3.25), so gilt

$$\frac{\partial^2 f}{\partial y'^2}(x, y(x), y'(x) \geq 0$$

für alle $x \in [a,b]$. (Soll in (3.3.25) > 0 gelten, muss diese Ungleichung ebenfalls strikt sein.)

Ist $y \in C^1([a,b], \mathbb{R}^m)$, so lautet die Legendre-Bedingung

$$\sum_{l,k=1}^{m} \frac{\partial^2 f}{\partial y_l' \partial y_k'}\xi_l \xi_k \geq 0 \qquad \forall \xi \in \mathbb{R}^m, \xi = (\xi_1, \xi_2, ..., \xi_m)$$

Wie bereits gesehen, ist

$$\delta^2 J(y;h) = \int_a^b \frac{\partial^2 f}{\partial y^2} h^2 + \frac{\partial^2 f}{\partial y \partial y'} hh' + \frac{\partial^2 f}{\partial y'^2}(h')^2 \mathrm{d}x$$

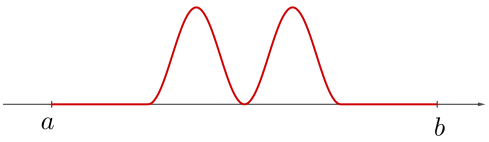
Sei nun $f \in C^3$, dann gilt

$$\int_a^b 2\frac{\partial^2 f}{\partial y \partial y'} hh' \mathrm{d}x = \int_a^b \frac{\partial^2 f}{\partial y \partial y'} \frac{\mathrm{d}}{\mathrm{d}x}(h^2)\mathrm{d}x \overset{\text{part. Int.}}{=} \left.\frac{\partial^2 f}{\partial y \partial y'} h^2\right|_a^b - \int_a^b \left(\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial^2 f}{\partial y \partial y'}\right) h^2 \mathrm{d}x$$

und somit

$$\delta^2 J(y;h) = \int_a^b \frac{\partial^2 f}{\partial y'^2}(h')^2 + \left(\frac{\partial^2 f}{\partial y^2} \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial^2 f}{\partial y \partial y'}\right) h^2 \mathrm{d}x$$

Sei nun $P(x) := \frac{\partial^2}{\partial y'^2} f(x, y(x), y'(x)$ sowie $Q(x) := \left(\frac{\partial^2 f}{\partial y^2} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial^2 f}{\partial y \partial y'}\right)(x, y(x), y'(x))$. Die Legendre-Bedingung ist dann darauf zurückzuführen, dass sich Variationen h konstruieren lassen, für die $P(x)(h'(x))^2$ den Term $Q(x)(y(x))^2$ dominiert. (Umgekehrt ist dies nicht möglich):
Aufgrund der Stetigkeit von $P(x)$ lassen sich für den Fall $P(x_0) < 0$ für ein x $\alpha, \beta \in \mathbb{R}^+$ so wählen, dass $P(x_0) < -2\beta$ und $P(x) < -\beta \forall x \in (x_0 - \alpha, x_0 + \alpha)$ gilt. Wählt man nun beispielsweise



$$h(x) := \begin{cases} \sin^2\left(\frac{\pi(x-x_0)}{\alpha}\right) & x_0 - \alpha \le x \le x_0 + \alpha \\ 0 & \text{sonst} \end{cases},$$

so gilt mit $M := \max_{x \in [a,b]} |Q(x)|$

$$\int_a^b P(x)(h'(x))^2 + Q(x)(h(x))^2 \mathrm{d}x = \int_{x_0-\alpha}^{x_0+\alpha} P(x)(h'(x))^2 + Q(x)(h(x))^2 \mathrm{d}x$$
$$< \int_{x_0-\alpha}^{x_0+\alpha} -\beta\frac{\pi}{\alpha}\sin^2\left(\frac{\pi(x-x_0)}{\alpha}\right)\cos^2\left(\frac{\pi(x-x_0)}{\alpha}\right) + M \mathrm{d}x = -\frac{\beta\pi^2}{\alpha} + 2M\alpha \overset{\alpha \to 0}{\to} -\infty$$

Eine hinreichende(!) Bedingung für einen schwachen (lokalen) Minimierer ist

$$\int_a^b \left(P(h')^2 + Qh^2\right) \mathrm{d}x > 0 \quad \forall h \in V - V \tag{3.3.26}$$

Für beliebiges $h \in V - V, \omega \in C^1([a,b])$ gilt:

$$0 = \omega h^2\big|_a^b = \int_a^b \frac{\mathrm{d}}{\mathrm{d}x}\left(\omega h^2\right) \mathrm{d}x = \int_a^b \omega' h^2 + 2\omega hh' \mathrm{d}x$$

Addiert man dies zu (3.3.26), so erhält man

$$\int_a^b P(h')^2 + Qh^2 \mathrm{d}x = \int_a^b \left(P(h')^2 + 2\omega hh' + (Q + \omega')h^2\right) \mathrm{d}x$$

Gesucht ist nun ein ω, sodass der Integrand auf der rechten Seite ein Quadrat ist, d. h. dass der Integrand die Form $(f_1 h + f_2 h')^2$ mit Funktionen f_1, f_2 hat. Aus den Bedingungen $f_1^2 = Q + \omega'$, $2f_1 f_2 = 2\omega$ und $f_2^2 = P^2$ ergibt sich

$$P(Q + \omega') = \omega^2, \tag{3.3.27}$$

also eine Differentialgleichung vom Riccati-Typ (siehe Kapitel 1.2). Existiert ein solches ω, so ergibt Umstellen der Gleichung $Q + \omega' = \frac{\omega^2}{P}$, wodurch (3.3.27) zu

$$\int_a^b \left(\sqrt{P}h' + \frac{\omega}{\sqrt{P'}h}\right)^2 \mathrm{d}x = \int_a^b P(h' + \frac{\omega}{P}h)^2 \mathrm{d}x$$

wird, wobei $P > 0$ (strikte Legendre-Bedingung). Ist ω regulär, so lässt die Randbedingung $h(a) = 0$ nur eine Lösung der Differentialgleichung $h' + \frac{\omega}{P}h = 0$ zu, nämlich die triviale Lösung. In diesem Fall gilt (3.3.26) für alle $h \in V - V \setminus \{0\}$. Jedoch kann ω sich in $[a, b]$ auch aufblähen. Die Bedingung (3.3.26) ist daher nur dann garantiert, wenn (3.3.27) eine reguläre Lösung hat. Eine mögliche Herangehensweise an das Problem ist, eine neue Funktion $u > 0$ einzuführen, sodass

$$\omega(x) = -\frac{P(x)u'(x)}{u(x)}$$

gilt, und diese in (3.3.27) einzusetzen. Dies führt auf die Gleichung

$$\frac{\mathrm{d}}{\mathrm{d}x}(Pu') = Qu,$$

nach Umstellen ergibt sich die sogenannte ***Jacobi-Gleichung***:

$$-\frac{\mathrm{d}}{\mathrm{d}x}(Pu') + Qu = 0$$

Dies ist eine homogene, lineare Differentialgleichung zweiter Ordnung, d. h. ist u eine Lösung, so ist für jede Konstante c cu ebenfalls eine Lösung. Seien nun die Anfangsbedingungen $u(a) = 0, u'(a) = 1$ gegeben. Ein Punkt $\bar{x} > a$ mit $u(\bar{x}) = u(a) = 0$ heißt - sofern existent - ***konjugierter Punkt*** von a. (Allgemeiner ist ein konjugierter Punkt on a ein $\bar{x} \in (a, b]$ mit $u(a) = u(\bar{x}) = 0$ für eine nicht-triviale Lösung u der Jacobi-Gleichung.) $\bar{x}$ hängt dabei von P und Q ab. Existieren keine konjugierten Punkte in $[a, b]$, so ist ω regulär in diesem Intervall und (3.3.26) ist erfüllt. Dies führt auf folgendes Theorem:

Satz 3.3.20:
Sei $f \in C^3$ in y' und es gelte:

1) $y \in V$ erfüllt die Euler-Lagrange-Gleichung.

2) Die strikte Legendre-Bedingung $\frac{\partial^2 f}{\partial y'^2}(x, y(x), y'(x)) > 0, x \in [a, b]$.

3) $[a, b]$ enthält keine konjugierten Punkte von a.

Dann ist $y \in V$ ein schwaches (striktes) lokales Minimum von J in V.

In Gegensatz zur endlichdimensionalen Optimierung ist die Bedingung $\langle \nabla^2 J(y)w, w\rangle > 0 \forall w \in V - V, w \neq 0$ kein hinreichendes Kriterium für ein lokales Minimum. Stattdessen ist (u. a.) folgendes Theorem hilfreich:

Satz 3.3.21:
Angenommen, für $y \in V$ gilt

1) $\langle \nabla J(y), w \rangle = 0 \forall w \in V - V$

2) $\langle \nabla^2 J(y) w, w \rangle > 0 \forall w \in V - V \setminus \{0\}$

3) $\frac{\partial^2}{\partial y'^2} f(x, y, y') > 0$ auf $[a, b]$

Dann ist y ein striktes lokales Minimum von J in V. (Und $y \in C^2([a, b])$.) Gilt zusätzlich $\frac{\partial^2 f}{\partial p^2} f(x, z, p) \geq 0$ für $p \in \mathbb{R}$ und alle $(x, z) \in D_\delta(y)$ für ein $\delta > 0$, wobei

$$D_\delta(y) = \{(x, z) | x \in [a, b], y(x) - \delta \leq z \leq y(x) + \delta\},$$

so ist y ein starkes, lokales Minimum:

$$J(y) \leq J(v) \forall v \in V, \|y - v\|_{C([a,b])} \leq \rho$$

für ein $\rho > 0$.

Sei $D \subseteq B$ eine offene Teilmenge von B und $y \in D$ sowie $v \in B \setminus \{0\}$. Sei ferner U eine offene Teilmenge von $\mathbb{R}^3$. Offensichtlich ist die Menge

$$D_U = \left\{ y \in C^1([a, b]) | \, (x, y(x), y'(x)) \in U \forall x \in [a, b] \right\}$$

im Banachraum $\left(C^1([a, b]), \|\cdot\|_{C^1([a,b])} \right)$ mit $\|y\|_{C^1([a,b])} \coloneqq \sup_{x \in [a,b]} |y(x)| + \sup_{x \in [a,b]} |y'(x)|$ enthalten.

Ferner ist $D_U \subset C^1([a, b])$ offen, da für jedes feste $\bar{y} \in D_U$ die Menge $\{(x, \bar{y}(x), \bar{y}'(x)) \, | x \in [a, b]\}$ kompakt ist und somit eine offene Umgebung in U hat. Ist $f \in C^2$, dann ist das Funktional

$$J(y) = \int_a^b f(x, y(x), y'(x)) \mathrm{d}x$$

Gâteaux-differenzierbar in D_U.

Satz 3.3.22:
Sei $J : D \to \mathbb{R}$ und $y \in D$. Dann gilt:

1) Notwendige Bedingung: Ist y ein schwacher, lokaler Minimierer und $\delta J(y; h)$ existiert für ein $h \in B \setminus \{0\}$ dann gilt $\delta J(y; h) = J'() = 0$. Insbesondere ist für Gâteaux-differenzierbares J $J'(y) = 0$.

2) Hinreichende Bedingung: J hat ein (lokales) Minimum in y, wenn gilt:
   a) $\delta J(y; h) = 0 \forall h \in B$
   b) (Koerzitivität) Für jedes $\bar{y}$ in einer konvexen Nachbarschaft von y existiert die zweite Variation $\delta^2 J(y; h)$ für alle $h \in B$. Ferner existiert eine Konstante $c > 0$, sodass
   $$\delta^2 J(y; h) \geq c \|h\|^2$$
   für alle $h \in B$ gilt.
   c) Für jedes $\varepsilon > 0$ existiert $\eta > 0$ mit
   $$\left| \delta^2 J(\bar{y}; h) - \delta^2 J(y; h) \right| \leq \varepsilon \|h\|^2$$
   für alle $h \in B$ und $\bar{y}$ mit $\|\bar{y} - y\| < \eta$.

Satz 3.3.23:
Seien $A, C \in C([a,b]), B \in C^1([a,b])$. Das quadratische Funktional

$$Q(y) = \int_a^b A(x)y(x)^2 + 2B(x)y(x)y'(x) + C(x)(y'(x))^2 \mathrm{d}x$$

ist positive definit - das heißt für alle $y \in C^1([a,b]), y \neq 0$ mit $y(a) = 0, y(b) = 0$ ist $Q(y) > 0$ - wenn folgendes gilt:

a) $C(x) > 0 \forall x \in [a,b]$

b) Für jedes $x^* \in (a,b]$ ist die einzige Lösung der Jacobi-Gleichung

$$-\frac{\mathrm{d}}{\mathrm{d}x}(Cu') + (A - B')u \mathrm{d}x = 0, u(a) = u(x^*) = 0$$

die triviale Lösung $u \equiv 0$.

3.3.9 Stückweise C^1-Kurven

Zu Beginn des Kapitels wurden einige beispielhafte Variationsprobleme vorgestellt, deren Lösungen nicht stetig differenzierbar sind. Es wurde auch das Konzept der Relaxierung erwähnt, bei dem der Raum für zulässige Lösungen vergrößert wird. Der erste, naheliegende Schritt in diesem Sinne ist, stückweise C^1-Funktionen zu betrachten.

Definition 3.3.24:
Eine Funktion $y \in C([a,b])$ ist ***stückweise in C^1***, wenn höchstens endlich viele Punkte $a = x_0 < x_1 < ... < x_{N+1} = b$ existieren, sodass $y|_{[x_k, x_{k+1}]} \in C^1([x_k, x_{k+1}])$ für $k = 0, ..., N$ gilt. Der Raum aller solcher Funktionen wird mit $C^1_{pw}([a,b])$ bezeichnet (pw für piecewise).

In Beispiel 3) vom Anfang des Kapitels wurde bereits ein Funktional vorgestellt, welches eine eine abzählbare Menge von C^1_{pw}-Minimierern hat. Es stellt sich die Frage, wie sich die Lösungen in Bezug auf die Euler-Lagrange-Gleichungen charakterisieren lassen. Diese führt auf die ***Weierstraß-Erdmannsche Eckenbedingungen***.

Satz 3.3.25 (1. Weierstraß-Erdmannsche Eckenbedingungen):
Sei $f \in C^2$ und y ein schwaches, lokales Minimum in $C^1_{pw}([a,b])$. Dann gilt für jede Unstetigkeitsstelle x_k der Ableitung y folgende Gleichung:

$$\frac{\partial f}{\partial y'} f(x_k, y(x_k), y'_-(x_k)) = \frac{\partial f}{\partial y'} f(x_k, y(x_k), y'_+(x_k)) \tag{WE1}$$

Hier bezeichnet $y'_-(x_k) = \lim\limits_{\substack{x \to x_k \\ x < x_k}} y(x)$ den linksseitigen, $y'_+(x_k) = \lim\limits_{\substack{x \to x_k \\ x > x_k}} y(x)$ den rechtsseitigen Grenzwert von y' in x_k.

In Gegenbeispiel 3) gilt $y'(x) \in \{-1, 1\}$ und somit $\frac{\partial}{\partial y'} f(x,y,y') = -2(1 - y'(x)^2)y'(x) = 0$. Man beachte, dass keine C^1_{pw}-Minima existieren können, wenn $\frac{\partial^2 f}{\partial y'^2} > 0$, also f strikt kon-

vex in y' ist: In diesem Falle ist $\frac{\partial f}{\partial y'}$ streng monoton steigend, wegen $y'_-(x_k) \neq y'_+(x_k)$ kann (WE1) nicht erfüllt sein. Die Bedingung (WE1) erinnert an die Euler-Lagrange-Gleichung mit freiem Rand, wobei der „Rand“ nun die Unstetigkeitsstelle y_k enthält. Tatsächlich lässt sich die Weierstraß-Erdmannsche Bedingung auf sehr ähnliche Weise herleiten, indem man „$\int_a^b = \int_a^{y_k} + \int_{y_k}^b$“ schreibt.

Zum besseren Verständnis von (WE1) und der daraus abgeleiteten 2. Bedingung soll das allgemeinere Variationsproblem $\min_y J(y) = \int_a^b f(x, y(x), y'(x))\mathrm{d}x, y(a) = y_a, y(b) = y_b$ betrachtet werden, wobei die Randpunkte varieren dürfen. Diese Variation sei Parametrisiert nach $t \in \mathbb{R}$, der Extrempunkt entspreche $t = 0$. Für gegebenes t sei die Kurve $y(x, t)$ mit Endpunkten $a(t)$ und $b(t)$ gegeben. Das Funktional kann nun auf folgende Weise geschrieben werden:

$$J(y, t) = \int_{a(t)}^{b(t)} f\left(x, y(x, t), \frac{\partial}{\partial x} y(x, t)\right) \mathrm{d}x \tag{3.3.28}$$

Die Variation bezüglich x und t führt auf die Gleichung

$$\frac{\mathrm{d}}{\mathrm{d}t} y(x(t), t) = \frac{\partial}{\partial t} y(x, t) + \frac{\partial}{\partial x} y(x, t) x'(t),$$

das heißt, die eigentliche Variation besteht aus zwei Termen: Einem für die Variation in t, welcher $\frac{\partial y}{\partial t}$ entspricht, der andere entspricht der Variation in x:

$$\delta y = \delta_t y + y' \delta x.$$

Entsprechend dieser Variation der Argumente von J lautet eine allgemeine Variationsformel wie folgt:

$$\begin{aligned}
\delta J(y; \delta y) &= \int_{a(t)}^{b(t)} \partial_t f(x, y, y')\mathrm{d}x + f(x, y, y')\delta x|_{x=a}^{x=b} = \\
&= \int_{a(t)}^{b(t)} \frac{\partial f}{\partial y}\delta_t y + \frac{\partial f}{\partial y'}\left(\delta_t y\right)' \mathrm{d}x + f(x, y, y')\delta x|_{x=a}^{x=b} = \\
&\overset{\text{part. Int.}}{=} \int_a^b \left(\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{y'}\right)\left(\delta_t y\right)\mathrm{d}x + \frac{\partial f}{\partial y'}\delta_t y\Big|_a^b + f\delta x|_a^b = \\
&\overset{\delta_t y = \delta y - y'\delta x}{=} \int_a^b \left(\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{y'}\right)\left(\delta_t y\right)\mathrm{d}x + \frac{\partial f}{\partial y'}\delta y\Big|_a^b + \left(f - \frac{\partial f}{\partial y'}y'\right)\delta x\Big|_a^b
\end{aligned}$$

Wendet man dies nun auf (3.3.28), aufgeteilt auf die Integrale über $[a, c]$ und $[c, b], a < b < c$, an und verlangt, dass die erste Variation von J gleich 0 ist, führt dies auf

$$\frac{\partial f}{\partial y'}\bigg|_{c_-} = \frac{\partial f}{\partial y'}\bigg|_{c_+}$$

einerseits sowie

$$f - \frac{\partial f}{\partial y'}y'\bigg|_{c_-} = f - \frac{\partial f}{\partial y'}y'\bigg|_{c_+} \tag{WE2}$$

Ersteres entspricht der ersten, letzteres der zweiten Weierstraß-Erdmannschen Eckenbedingung. Eine Anwenungsmöglichkeit der Weierstraß-Erdmannschen Eckenbedingungen ist das Snelliussche Brechungsgesetz:

Beispiel 3.3.26 (Snelliussches Brechungsgesetz):

Betrachtet wird ein Lichtstrahl, welcher von einem Punkt $(a, y_)$ nach (b, y_b) durch zwei verschiedene Medien geht. c_1 beschreibe die Lichtgeschwindigkeit im ersten, c_2 die im zweiten Medium. Sei $(z, y(z))$ der Punkt, an dem der Lichtstrahl von einem Medium ins andere übertritt, sowie

$$f_j(x, y(x), y'(x)) = \frac{1}{c_j}\sqrt{1 + (y')^2},\ j \in \{1, 2\}.$$

Das Fermatsche Prinzip besagt, dass Licht von allen möglichen Wegen denjenigen annimmt, welcher die geringste Zeit beansprucht. Diese berechnet sich durch den Kehrwert der Geschwindigkeit multipliziert mit der Länge des Weges. Unter Berücksichtigung der unterschiedlichen Ausbreitungsgeschwindigkeiten ist die gesuchte Trajektorie also diejenige, welche das Funktional

$$J(y) = \frac{1}{c_1}\int_a^z \sqrt{1 + (y')^2}\mathrm{d}x + \frac{1}{c_2}\int_z^b \sqrt{1 + (y')^2}\mathrm{d}x$$

minimiert. In diesem Zusammenhang sei auf folgendes hingewiesen: Die notwendigen Bedingungen für ein Optimum bedingen $\delta J(y; \delta y) = 0$ für jede zulässige Variation δy, d. h.

$$\begin{aligned}
0 = &\int_a^z \left(\frac{\partial f_1}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f_1}{\partial y'}\right)\delta y \mathrm{d}x + \left(\delta y \frac{\partial f_1}{\partial y'}\right)\Big|_a^z \\
&+ \int_z^b \left(\frac{\partial f_2}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f_2}{\partial y'}\right)\delta y \mathrm{d}x + \left(\delta y \frac{\partial f_2}{\partial y'}\right)\Big|_z^b = \\
= &\int_a^z \left(\frac{\partial f_1}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f_1}{\partial y'}\right)\delta y \mathrm{d}x + \int_z^b \left(\frac{\partial f_2}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f_2}{\partial y'}\right)\delta y \mathrm{d}x \\
&+ \left(\delta y \frac{\partial f_1}{\partial y'}\right)(z) - \left(\delta y \frac{\partial f_2}{\partial y'}\right)(z)
\end{aligned}$$

Durch (WE1) sowie geschickte Wahl der Variation lässt sich $\frac{\partial}{\partial y} f_j - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial}{\partial y'} f_j = 0$ folgern. Wegen $\frac{\partial}{\partial y} f_j = 0$ gilt also $\frac{\partial}{\partial y'} f_j = c$ für ein $c \in \mathbb{R}$. Hieraus ergibt sich die Differentialgleichung

$$c = \frac{y'(x)}{c_j\sqrt{1 + (y'(x))^2}},$$

aus welcher sich wiederum ergibt, dass $y'(x)$ konstant sein muss. Dies bedeutet, dass Licht sich innerhalb eines Mediums geradlinig fortbewegt!
Gemäß (WE1) gilt $\frac{\partial}{\partial y'} f_1\Big|_{z_-} = \frac{\partial}{\partial y'} f_2\Big|_{z_+}$, d. h.

$$\frac{1}{c_1}\frac{y'}{\sqrt{1 + (y')^2}}\Bigg|_{z_-} = \frac{1}{c_2}\frac{y'}{\sqrt{1 + (y')^2}}\Bigg|_{z_+} \qquad (*)$$

Sei nun θ_1 der Einfallswinkel, mit welchem das Licht auf die Grenze zwischen den Medien trifft, θ_2 der Brechungswinkel.

Es gilt

$$\tan\theta_1 = y'|_{z_-}\,, \qquad \tan\theta_2 = y'|_{z_+}$$

Aus $(*)$ resultiert nun das Snelliussche Brechungsgesetz:

$$\frac{\sin\theta_1}{c_1} = \frac{\sin\theta_2}{c_2}$$

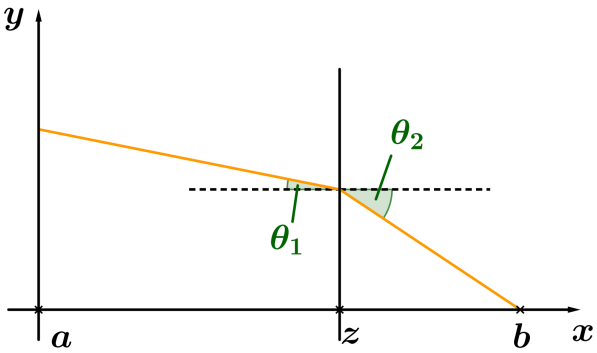


Abbildung 3.3.3: Gebrochener Strahl

3.3.10 Abschließende Bemerkungen

In der Mechanik betrachtet man häufig zeitabhängige Probleme, weshalb x durch $t \in [0,T]$ ersetzt wird. Der Integrand f von J wird durch $L = L\,(q(t),\dot{q}(t),t)$ ersetzt, die sogenannte ***Lagrangefunktion***. (**Achtung**: Hier handelt es sich um eine Bezeichnung aus der Physik! Sie ist nicht zu verwechseln mit der Lagrangefunktion im Sinne der Optimierung!)

Seien nun L und y C^2-Funktionen sowie

$$p = \frac{\partial}{\partial\dot{q}} L\,(q(t),\dot{q}(t),t)\,,$$

was wiederum $\dot{p} = \frac{\partial}{\partial y} L$ erfüllt. Dies kann ausgenutzt werden, um unter der Voraussetzungen $L \in C^2, \frac{\partial^2}{\partial\dot{q}^2} L$ strikt positiv definit, $\dot{p} = \dot{p}\,(q(t),p(t),t)$ zu schreiben. Dies führt auf die aus dem vorherigen Kapitel bekannte Hamiltonfunktion

$$H\,(q(t),p(t),t) = p\dot{q} - L\,(q(t),\dot{q}(t),t)\,,$$

welches die Legendre-Transformierte der Lagrangefunktion ist. Wie bereits gesehen ist dies äquivalent zu

$$\dot{q} = \frac{\partial}{\partial p} H(q,p,t)$$
$$\dot{p} = -\frac{\partial}{\partial q} H(q,p,t)$$

Hierbei ist zu beachtem, dass gemäß den Definitionen $\frac{\partial}{\partial\dot{q}} = p - \frac{\partial}{\partial\dot{q}} = 0$ gilt. Aufgrund der vorausgesetzen strikten Definitheit von $\frac{\partial^2}{\partial\dot{q}^2} L$ ist die Trajektorie (q,p) dergestalt, dass H als Funktion von $\dot{q}$ einen Extremwert annimmt.

Eine Verallgemeinerung der Legendre-Transformation ist die ***Legendre-Fenchel-Transformation***, mit ihrer Hilfe lässt sich die Hamiltonfunktion verallgemeinert als

$$H\,(q,p,t) = \sup_{\dot{q}\in\mathbb{R}^n} \left(p\dot{q} - L\,(q,\dot{q},t)\right)$$

schreiben. Ist L ist konvex und schwach unterhalbstetig in $\dot{q}$, so gilt umgekehrt

$$L\,(q,\dot{q},t) = \sup_{p\in\mathbb{R}^n} \left(p\dot{q} - H\,(q,p,t)\right).$$

In diesem Fall wird durch die Legendre-Fenchel-Transformation P eine Bijektion $L \mapsto H$ beschrieben.

Betrachtet man ein mechanisches System mit kinetischer Energie $T(\dot{q})$, welches aus einem Potenzial $V(q)$ resultierenden Kräften unterworfen ist, so ist die Lagrangefunktion $L = T - V$ (vgl. 2.6, Federpendel.) Für ein Partikel mit Masse m, welches sich zum Zeitpunkt t am (verallgemeinerten) Ort q befindet, gilt $T(\dot{q}) = \frac{m}{2}\dot{q}^2$. Wie bereits gesehen, resultiert aus dem Prinzip der stationären Wirkung (die erste Variation des Funktionals $J(q) = \int_0^T L(q(t), \dot{q}(t))\mathrm{d}t$ gleich 0) das Newtonsche Gesetz $m\ddot{q} = -\frac{\partial V}{\partial q}(q) = F(q)$, welches den Euler-Lagrange-Gleichungen entspricht. In diesem Falle entspricht $p = \frac{\partial L}{\partial \dot{q}} = m\dot{q}$ dem Impuls des Partikels und es gilt:

$$\begin{aligned} H(q,p,t) = p\dot{q} - L(q,\dot{q},t) = m\dot{q}^2 - \frac{m}{2}\dot{q}^2 + V(q) = \\ = \frac{m}{2}\dot{q}^2 + V(q) = T(\dot{q}) + V(q) \end{aligned}$$

Die Hamiltonfunktion spiegelt also die Gesamtenergie wieder.

3.4 Optimale Steuerung von Modellen mit Differentialgleichungen

Ein Schwerpunkt der Variationsrechnung ist die Lösung von Problemen der Art

$$\min_{y \in V} J(y) \coloneqq \int_a^b l(t, y(t), \dot{y}(t))\mathrm{d}t, \tag{3.4.1}$$

wobei

$$V = \left\{ v \in C^1([a,b]) | v(a) = y_a, v(b) = y_b \right\} \tag{3.4.2}$$

ist (s. Kapitel 3.3). Das unrestringierte Problem (3.4.1)+(3.4.2) lässt sich auch als restringiertes schreiben. Mit Hilfe einer neuen Funktion $u \in C([a,b])$ lautet eine alternative Formulierung:

$$\begin{aligned} &\min J(y,u) \coloneqq \int_a^b l(t, y(t), u(t))\mathrm{d}t \\ &\text{u. d. Nb.} \quad \dot{y}(t) = u(t) \\ &\text{und} \quad y(a) = y_a,\ y(b) = y_b \end{aligned} \tag{3.4.3}$$

Dieses Problem kann nun auch als ein ***optimales Steuerungsproblem*** angesehen werden, wobei u die ***Kontrollfunktion*** ist. (y heißt ***Zustandsvariable***.)

Definition 3.4.1:
Sei U Funktionenraum. Eine Steuerung $\tilde{u} \in U$ ist ***optimal*** und $\tilde{y} = \tilde{y}(\tilde{u})$ der ***assoziierte optimale Zustand***, wenn

$$J(\tilde{y}, \tilde{u}) \leq J(y(u), u) \quad \forall u \in U.$$

gilt.

Obwohl das Problem (3.4.1)+(3.4.2) und das Problem (3.4.3) äquivalent erscheinen, sind sie es nicht: In (3.4.1) ist $\dot{y}$ die Ableitung der gesuchten Lösung. In (3.4.3) hingegen resultiert y aus der Lösung der Differentialgleichung $\dot{y} = u$ mit den gegeben Anfangs- und Endbedingungen. In anderen Worten liegt das Hauptaugenmerk bei optimalen Steuerungsproblemen auf der Kontrollfunktion. Im Hinblick auf diesen Paradigmenwechsel kann das Problem (3.4.3) auf viele Weisen verallgemeinert werden, für welche es wiederum entsprechende Konfigurationen in der Variationsrechnung gibt.

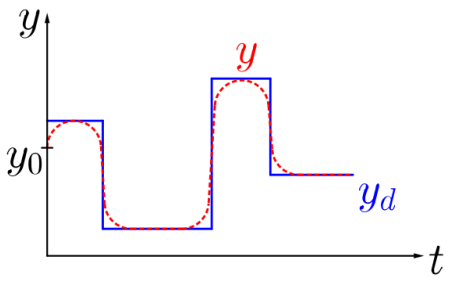
Sei nun das allgemeinere Problem

$$\begin{aligned} \min J(y,z,u) &:= \int_a^b l(t,y(t),u(t))\mathrm{d}t + g(b,y(b)) \\ \text{u. d. Nb.} \quad \dot{y}(t) &= f(t,y(t),u(t)), \quad y(a) = y_a \end{aligned} \tag{3.4.4}$$

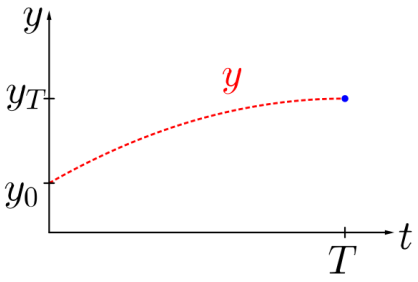
gegeben. J ist das ***Kosten-*** oder auch ***Zielfunktional***, l sind ***laufende Kosten*** und g sind die ***Endkosten***.

Beinhaltet J beide Kostenarten, so ist es in ***Bolza-Form***. Sind hingegen nur die laufenden Kosten enthalten, so ist es in ***Lagrange-Form***, sind es nur die Endkosten, so ist das Problem in ***Mayer-Form***. Anbei sind einige Beispiele, wie solche Funktionale aussehen können:

- $J(y,u) = \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\nu}{2}\|u\|_{L^2}^2$: In diesem Falle ist $\frac{1}{2}\|y - y_d\|_{L^2}^2$ eine Art „Spurfunktional“, durch welches y_d das (grobe) Soll-Verhalten der Zustandsfunktion vorgibt. $\|u\|_{L^2}^2$ hingegen spiegelt die Kosten der Kontrolle wieder.



- $J(y,u) = \frac{1}{2}|y(T) - y_T|^2 + \frac{\nu}{2}\|u\|_{L^2}^2$: In diesem Falle beschreibt $\frac{1}{2}|y(T) - y_T|_{L^2}^2$ die Abweichung von einem gewünschtem Endwert.



- $J(y,u) = \frac{1}{2}|y(T) - y_T|^2 + \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\nu}{2}\|u\|_{L^2}^2$: Kombination obiger Fälle

Man kann ein Problem von der Lagrange- in die Mayer-Form bringen, indem man eine zusätzliche Variable z einführt, sodass

$$\dot{z}(t) = l(t,y(t),u(t)), \quad z(a) = 0 \tag{3.4.5}$$

gilt. Dann ist

$$\int_a^b l(t,y(t),u(t))\mathrm{d}t = z(b),$$

folglich wird (3.4.4) zu

$$\begin{aligned} \min J(y,u,z) &= z(b) + g(b,y(b)) \\ \text{u. d. Nb.} \quad \dot{y} &= f(t,y(t),u(t)), \quad y(a) = y_a \\ \dot{z}(t) &= l(t,y(t),u(t)), \quad z(a) = 0 \end{aligned}$$

Vice versa kann, ausgehend von einem Problem in Mayer-Form, ein Problem in Lagrange-Form wie folgt konstruiert werden:

$$\begin{aligned} g(b, y(b)) &= g(a, y_a) + \int_a^b \frac{\mathrm{d}}{\mathrm{d}y} g(t, y(t)) \mathrm{d}t \\ &= g(a, y_a) + \int_a^b \left(\frac{\partial g}{\partial t} + \frac{\partial g}{\partial y} \dot{y} \right) \mathrm{d}t = \\ &= g(a, y_a) + \int_a^b \left(\frac{\partial g}{\partial t} + \frac{\partial g}{\partial y} f \right) \mathrm{d}t, \end{aligned}$$

wobei $g(a, y_a)$ eine Konstante ist, welche das Optimierungsproblem nicht beeinflusst und ignoriert werden kann, $l = \frac{\partial g}{\partial t} + \frac{\partial g}{\partial y} f$ sind die laufenden Kosten.

Im Allgemeinen können b und a fixiert sein oder auch nicht, je nach Situation liegt entsprechend ein Problem mit fester oder freier Endzeit vor. Auch die Endbedingung $y(b) = y_b$ ist optional, entsprechend handelt es sich dann um ein Problem mit festem beziehungsweise freiem Endpunkt. Im Folgenden soll der Fokus auf Problemen der Form wie in (3.4.4) mit freiem Endpunkt, aber fester Endzeit liegen.

3.4.1 Existenz einer optimalen Steuerung

Die erste Frage, die sich stellt, ist, ob zumindest eine Kontrollfunktion existiert, die (3.4.4) löst. Dies hängt von der Struktur des Problems sowie dem Funktionenraum ab, in welchem die Kontrolle gesucht wird.

Im Folgenden sollen lineare, quadratische Kontrollprobleme betrachtet werden, welche es ermöglichen, einige klassische Techniken zum Nachweis der Existenz anzuwenden. Sei hierzu das lineare Dgl.-System

$$\dot{y} = Ay + Bu, \quad y(a) = y_0 \tag{3.4.6}$$

gegeben, wobei $y(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times n}$ sowie $B \in \mathbb{R}^{n \times m}$ sind. Die Lösung von (3.4.6) für ein gegebenes u lautet

$$y(t) = e^{tA} y_0 + e^{tA} \int_0^t e^{-sA} B u(s) \mathrm{d}s. \tag{3.4.7}$$

Ist $u \in L^2(a, b)$, so ist wegen $\sup_{s \in [a,b]} \left\| e^{-sA} \right\| < \infty$ der Integrand ebenfalls ein Element des $L^2(a, b)$ und die Lösung $y \in C([a, b])$. (3.4.7) definiert eine Funktion $S : L^2(a, b) \to C([a, b]), u \mapsto y = S(u)$, im Englischen auch ***control-to-state map*** genannt. Wie in (3.4.6) ersichtlich ist, ist diese affin linear. Da für $\tilde{S}(u) = S(u) - S(0)$

$$\left\| \tilde{S} \right\|_\infty = \max_{t \in [a,b]} \left\| (Su)(t) - (S0)(t)) \right\|_2 \leq c \sqrt{b - a} \left\| u \right\|_{L^2(a,b)}$$

gilt, ist $\tilde{S}$ und somit S stetig. Als nächstes sei die Menge der zulässigen Kontrollen wie folgt definiert:

$$U_{ad} = \left\{ u \in L^2(a, b) | u(t) \in K_{ad}, u \text{ stetig in } (a, b) \right\},$$

wobei $K_{ad} \subseteq \mathbb{R}^m$ nichtleer, abgeschlossen und konvex sei. Das Kostenfunktional kann wie folgt aussehen:

$$J(y, u) = J_1(y) + J_2(u)$$

mit $J_1 : C([a, b]) \to \mathbb{R}, J_2 : L^2(a, b) \to \mathbb{R}$. Ein klassisches Beispiel wäre

$$J_1(y) = \frac{1}{2} \int_a^b \left\| y(t) - y_d \right\|_2^2 \mathrm{d}t,$$

wobei $y_d \in L^2(a,b)$ ein gewünschtes Profil (oder eine gewünschte Trajektorie) angibt. Ferner sei

$$J_2(u) = \frac{\nu}{2} \int_a^b \|u(t)\|_2^2 \, \mathrm{d}t, \nu \geq 0.$$

Dies repräsentiert die L^2-Kosten der Kontrolle. Im Allgemeinen wird gefordert, dass J_1, J_2 konvex, von unten beschränkt und J_2 koerzitiv auf U_{ad} ist. Gilt jedoch $J_2 = 0$, so soll U_{ad} beschränkt sein. Das ***reduzierte Kostenfunktional*** ist durch

$$\hat{J}(u) = J(S(u), u) = J_1(S(u)) + J_2(u)$$

definiert. Nun gilt die Äquivalenz

$$\begin{cases} \min J(y,u) \\ \text{u. d. Nb. } \dot{y} = Ay + Bu \\ \qquad\quad y(a) = y_a \\ \qquad\quad\ u \in U_{ad} \end{cases} \Leftrightarrow \min_{u \in U_{ad}} \hat{J}(u) \tag{3.4.8}$$

Die rechte Seite wird hierbei mitunter auch als ***reduziertes Problem*** bezeichnet. Wie im Beispiel sei J_1 stetig, konvex und von unten beschränkt. Ferner sei angenommen, dass J_2 stetig und koerzitiv in U_{ad} ist. Nun, da S affin linear und stetig ist, ist $J_1(S(u))$ ebenfalls konvex und stetig in u. Daher ist $\hat{J}(u)$ ein konvexes, stetiges, koerzitives Funktional auf U_{ad}, welches wiederum eine abgeschlossene, konvexe Teilmenge des $L^2(a,b)$ darstellt. Daher hat, wie bereits im Kapitel 3.3 dargelegt, das optimale Steuerungsproblem (3.4.8) eine eindeutige Lösung. Ein ähnliches Resultat (mit Ausnahme der Eindeutigkeit) erhält man für ein Modell $\dot{y} = f(t,y,u)$ immer dann, wenn das entsprechende Cauchy-Problem eine Kontroll-Zustands-Funktion ermöglicht, welche wie oben stetig ist, und man $\left\|\dot{y}_{L^2(a,b}\right\| \leq c(\|y_d\| + \|u\|_{L^2(ab)})$ zeigen kann. Dann resultiert schwache L^2-Konvergenz einer minimierenden Folge $u_k \rightharpoonup u$ in starker Konvergenz der entsprechenden Zustände $y(u_k) \to y(u)$. Sei nun allgemeiner

$$\begin{aligned} \min J(y,u) &:= J_1(y(b)) + \alpha J_2(y) + \frac{\nu}{2} \|u\|_U \\ \text{u. d. Nb.} \quad \dot{y} &= f(t,y,u), \quad y(a) = y_a \\ u &\in U_{ad} \subset U \end{aligned} \tag{3.4.9}$$

Annahme 1: J_1, J_2 stetig, konvex und differenzierbar mit $J_1 \geq 0, J_2 \geq 0$ sowie $\nu > 0$. Ferner seien im Folgenden $H = H^1(a,b)$ und $U = L^2(a,b)$.

1) Zur Lösung des Cauchy-Problems

Annahme 2: f sei Lipschitz-stetig in y, stetig in (t,y,u), linear in u und messbar, es gelte $|f(t,y(t),u(t))| \leq \alpha(t)\|y\|$ mit $\alpha \in L^1(a,b)$.

Der Satz von Caratheodory garantiert nun die Existenz einer Lösung für $u \in L^2(a,b)$, welche stetig ist, (d. h. $y \in C([a,b])$ und für die $\dot{y} \in L^1(a,b)$ ist. Um dieses Resultat zu verbessen, sodass $y \in H^1(a,b)$ gilt, wird eine Stabilitätsabschätzung für $\dot{y}$ benötigt, weshalb die folgende Annahme gelte.
Annahme 3: Für das Modellproblem gilt die Abschätzung

$$\|\dot{y}\|_{L^2(a,b)} \leq c(\|y_a\| + \|u\|_{L^2(a,b)})$$

für ein $c > 0$.

2) Zur Kontroll-Zustands-Funktion

Mit dem letzten Resultat ist die Funktion

$$S : L^2(a,b) \to H^1(a,b), u \mapsto y = S(u)$$

wohldefiniert.

Definition:
Die Abbildung $u \mapsto y = S(u)$ von U nach H heißt ***schwach folgenstetig***, wenn für jede Folge von Kontrollfunktionen $(u_k)_k \subset U$, welche schwach gegen einen Grenzwert $u \in U$ konvergiert, die korrespondierende Folge von Zuständen $(y_k)_k \subset H, y_k = S(u_k)$ schwach gegen $y = S(u)$ konvergiert, d. h. $y_k \rightharpoonup S(u)$.

Gelten die Annahmen 2&3, so ist die Abbildung S schwach folgenstetig. Um dies zu zeigen, sie (u_k) eine schwach konvergente Folge von Kontrollfunktionen in $L^2(a,b)$ mit Grenzwert $u \in L^2(a,b)$. Dann ist die Folge $(y_k) \subset H^1(a,b)$ beschränkt, es existiert folglich eine schwach konvergente Teilfolge $y_{k_j} \rightharpoonup y \in H^1(a,b)$, da H Hilbertraum ist. Aufgrund der kompakten Einbettung $H^1(a,b) \subset\subset C([a,b])$ gilt $y_{k_j} \to y$ in $C([a,b])$ und der Grenzwert ist eindeutig. Nun muss gezeigt werden, dass $y = S(u)$ ist. Zu diesem Zweck sei $v \in H^1(a,b)$ eine Testfunktion $(v(a) = v(b) = 0)$. Es gilt

$$\int_a^b \left(\dot{y}_{k_j} - f(t, y_{k_j}, u_{k_j})\right) v \mathrm{d}t = 0 \forall v \in H, k_j$$

Unter dem Integral kann zum schwachen Grenzwert übergegangen werden, folglich

$$\int_a^b (\dot{y} - f(t,y,u))\, v \mathrm{d}t = 0 \forall v \in H$$

Also gilt tatsächlich $y = S(u)$.

3) Zur minimierenden Folge

Zur Erinnerung: Eine Folge $(y_k)_k \subset B$ ist minimierend, wenn $\lim\limits_{k\to\infty} J(y_k) = \inf\limits_{y\in B} J(y)$ gilt.

Aus der Definition des Infimums wird sofort klar, dass Beschränktheit von unten hinreichend für die Existenz einer solchen Folge ist.

Lemma 3.4.2:
Sei $J : U \to \mathbb{R}$ von unten beschränkt, U nichtleere Teilmenge eines reflexiven Banachraums. Ist J (schwach) koerzitiv, dann ist eine minimierende Folge $(u_k)_k \subseteq U$ von J beschränkt.

Beweisidee/-skizze: Folgt aus den Definitionen. □

5) Zur zulässigen Menge U_{ad}

Lemma 3.4.3:
Sei B Banachraum und C eine nichtleere, konvexe, abgeschlossene Teilmenge von B. Dann ist C schwach folgenabgeschlossen, d. h., für jede schwach konvergente Folge in C ist der Grenzwert ebenfalls in C.

Lemma 3.4.4:
Sei B Banachraum und C eine nichtleere, konvexe, abgeschlossene und beschränkte Teilmenge von B. Dann ist C schwach folgenabgeschlossen, d. h., für jede schwach konvergente Folge in C ist der Grenzwert ebenfalls in C. Dann ist C schwach folgenkompakt, d. h., jede Folge in C hat eine Teilfolge, welche schwach in C konvergiert.

Ist U_{ad} konvex und abgeschlossen (und nichtleer), soll eine minimierende Folge schwach in U_{ad} konvergieren. Hierfür genügt es, dass die minimierende Folge beschränkt ist, wenn der Hilbertraum reflexiv ist (Lemma 3.4.2). Ist U_{ad} ebenfalls beschränkt, dann hat eine minimierende Folge in U_{ad} eine konvergente Teilfolge mit schwachem Grenzwert in U_{ad}. In diesem Falle wird Koerzitivität nicht benötigt und in (3.4.9) kann der Fall $\nu = 0$ betrachtet werden.

Beweis der Existenz eines Minimierers

Betrachtet wird das Problem (3.4.9) unter den Annahmen 1-3. Da J von unten beschränkt ist, existiert eine minimierende Folge $(y_n, u_n)_k \subset H \times U_{ad}, y_k = S(u_n)$ mit

$$\lim_{n\to\infty} J(y_n, u_n) = \inf_{(y,u)\in H\times U_{ad}} J(y, u)$$

Sei U_{ad} konvex und abgeschlossen in U, $\nu > 0$, dann gilt

$$\hat{J}(u) = J(S(u), u) = J_1(S(u)(b)) + \alpha J_2(S(u)) + \frac{\nu}{2}\|u\|_U^2 \geq \frac{\nu}{2}\|u\|_U^2 .$$

Folglich ist $\hat{J}$ koerzitiv bezüglich U in der U-Norm. Somit ist die minimierende Folge beschränkt. Da $U = L^2(a, b)$ Hilbertraum ist, existiert eine schwach konvergente Teilfolge $u_{n_j} \rightharpoonup u$ in U_{ad}. Diese ist beschränkt in H, wie sich aus Annahme 3 ergibt. Es lässt sich also eine schwach konvergente Teilfolge $y_{n_{j_k}} \rightharpoonup y \in H$ extrahieren.

Es bleibt zu zeigen, dass $y = S(u)$ gilt, wie in (3.4.9). Hierzu muss jedoch der schwache Grenzwert $f(t, y_n, u_n) \rightharpoonup f(t, y, u)$ Sinn ergeben. Ist zum Beispiel das Kontrollproblem bilinear, bspw. $f(y_n, u_n) = y_n u_n$, so muss mindestens eine der Folgen stark konvergieren. Um diese starke Konvergenz zu erhalten, nutzt man gewöhnlich Einbettungstheoreme, im vorliegenden Fall $H^1(a, b) \subset\subset C([a, b])$ und $y_{n_{j_k}} \to y$ in $C([a, b])$. Dieser Schritt ist essentiell für den Kostenfunktionalterm J_1, denn nun gilt $y_{n_j j_k}(b) \to y(b)$. In Annahme 1 wurde vorausgesetzt, dass J_1 und J_2 stetig und konvex sind, was ebenfalls für $\frac{\nu}{2}\|u\|_U^2$ gilt. Damit ist J schwach folgenunterhalbstetig. Es gilt (der Einfachheit halber wird n_{j_k} durch n ersetzt)

$$\begin{aligned} J(y, u) &= J_1(y(b)) + \alpha J_2(y) + \frac{\nu}{2}\|u\|_U^2 = \\ &\leq \lim_{n\to\infty} J_1(y_n(b)) + \liminf_{n\to\infty}\left(\alpha J_2(y_n, u_n) + \frac{\nu}{2}\|u_n\|_U^2\right) = \\ &= \liminf_{n\to\infty} J(y_n, u_n) = \inf_{u\in U_{ad}} J(S(u), u), \end{aligned}$$

also

$$J(y, u) \leq \inf_{u\in U_{ad}} J(S(u), u),$$

was bedeutet, das (y, u) die optimale Lösung von (3.4.9) ist.

Zum Abschluss sei noch angemerkt, dass der Term $\frac{\nu}{2}\|u\|_U^2$ die Rolle eines Regularisierungsterms inne hat, wie bereits im vorherigen Kapitel. Alternativ kann dieser Term auch durch $\frac{\nu}{2}\|u\|_{H^1(a,b)}^2$ ersetzt werden. Dann wäre die minimierende Folge beschränkt in H^1 und hätte eine schwach konvergente Folge in diesem Raum. Aufgrund der Kompakten Einbettung würde dann, wie bereits mehrfach erwähnt, die Teilfolge stark in $C([a, b])$ konvergieren.

Wie im vorherigen Kapitel gezeigt wurde, wird mit einer wohldefinierten Kontroll-Zustands-Funktion aus einem optimalen Steuerungsproblem wie in (3.4.4) ein Problem der Variationsrechnung,

$$\min_{u \in U_{ad}} \hat{J}(u) \tag{3.4.10}$$

Somit können Optimalitätsbedingungen für (3.4.10) aufgestellt werden, welche ggf. Ableitungen enthalten. Insbesondere muss - vorausgesetzt $U_{ad} = U$ ist Hilbertraum und $\hat{J}$ ist differenzierbar - die optimale Steuerung $u^* \in u$ die Gleichung

$$\left\langle \nabla \hat{J}(u^*), w \right\rangle = 0 \forall w \in U \tag{3.4.11}$$

erfüllen. Wegen $\hat{J}(u) = J(S(u), u)$ setzt dies Differenzierbarkeit der Abbildungen $S(\cdot)$ und $J(\cdot, \cdot)$ voraus. Ist J wie in (3.4.4), so sind hierfür $l^1 \in C^1$ und $g \in C^1$ hinreichend. Betrachtet man das lineare, quadratische Problem (3.4.8), so ist S aufgrund der Affinität differenzierbar und das quadratische Funktional

$$J(y, u) = \frac{1}{2} \int_a^b |y(t) - y_d(t)|^2 \, \mathrm{d}t + \frac{\nu}{2} \int_a^b |u(t)|^2 \, \mathrm{d}t$$

ist differenzierbar in y und u.

Sei nun allgemeiner der Operator

$$C : H \times U \to H, (y, u) \mapsto \dot{y} - f(\cdot, y, u)$$

definiert. Offensichtlich definiert $C(y, u) = 0$ die Differentialgleichungsnebenbedingung. Es stellt sich die Frage, ob C differenzierbar ist. Der Kandidat für die (Fréchet-)Ableitung von C ist

$$\partial C(y, u)(\delta y, \delta u) = \delta \dot{y} - \left(\frac{\partial f}{\partial y} \right)_{(y,u)} \delta y - \left(\frac{\partial f}{\partial u} \right)_{(y,u)} \delta u \tag{3.4.12}$$

In der Tat kann für $f \in C^2$ mittels Taylor-Entwicklung gezeigt werden, dass (3.4.12) die Fréchet-Ableitung ist, diese also insbesondere existiert. Es handelt sich bei (3.4.12) um eine sogenannte ***linearisierte Nebenbedingung***.

Im Folgenden sei $J(\cdot, \cdot)$ differenzierbar, wodurch $\nabla \hat{J}$ berechnet werden kann (bzw. zumindest existiert). Hierbei ist $\partial C(y, u)(\delta y, \delta u)$ für gegebenes δu invertierbar bzgl. δy:

$$\delta \dot{y} = \frac{\partial f}{\partial y} \delta y + \frac{\partial f}{\partial u} \delta u, \qquad \delta y(a) = 0$$

hat eine eindeutige Lösung. Für Problem (3.4.4),

$$\begin{aligned} \min J(y, u) &\coloneqq \int_a^b l(t, y(t), u(t)) \mathrm{d}t + g(b, y(b)) \\ \text{u. d. Nb.} \quad \dot{y}(t) &= f(t, y(t), u(t)), \quad y(a) = y_a \end{aligned}$$

gilt folgendes Lemma:

Lemma 3.4.5:
Das Differentialgleichungssystem

$$-\dot{\lambda}(t) = \left(\frac{\partial f}{\partial y}(t, y, u) \right) \lambda(t) - \frac{\partial l}{\partial y}(t, y, u) \tag{3.4.13}$$

mit Endbedingung $\lambda(b) = -\frac{\partial g}{\partial y}(b, y(b))$ und $y \in H^1(a, b)$, $f, l, g \in C^1$ hat eine eindeutige Lösung $\lambda \in H$.

(3.4.13) ist hierbei lineare Dgl, welche „rückwärts“ gelöst wird. Durch die Transformation $\hat{t} := (b+a) - t$ wird es zu einem „vorwärts gerichteten“ Problem mit Anfangsbedingung in a.

Mit dieser Vorbereitung kann nun $\nabla \hat{J}(u)$ berechnet werden. Genauer gesagt soll das Gâteaux-Differential für eine Variation δu der Kontrolle $u \in U$ berechnet werden, es gilt dann

$$\left\langle \nabla \hat{J}(u), \delta u \right\rangle = \lim_{\alpha \to 0^+} \frac{1}{\alpha} \left(\hat{J}(u + \alpha \delta u) - \hat{J}(u) \right),$$

ferner ist $\delta y = S'(u)\delta u$. Also:

$$\begin{aligned} \left\langle \nabla \hat{J}(u), \delta u \right\rangle &= \lim_{\alpha \to 0^+} \frac{1}{\alpha} \left(\hat{J}(u + \alpha \delta u) - \hat{J}(u \right) = \\ &= \lim_{\alpha \to 0^+} \frac{1}{\alpha} \left(\alpha \left(\frac{\partial l}{\partial y} \delta y + \frac{\partial l}{\partial u \delta u} \right) \mathrm{d}t + \alpha \frac{\partial g}{\partial y} \delta y(b) \right) + \text{Terme höherer Ordnung} \end{aligned}$$

Bildet man den Grenzwert, verschwinden die Terme höherer Ordnung. Unter Verwendung von (3.4.13) zur Ersetzung von $\frac{\partial l}{\partial y}$ erhält man

$$\begin{aligned} \left\langle \nabla \hat{J}(u), \delta u \right\rangle &= \int_a^b \left(\dot{\lambda} + \frac{\partial f}{\partial y} \lambda \right) \delta y \mathrm{d}t + \int_a^b \frac{\partial l}{\partial u} \mathrm{d}t + \frac{\partial g}{\partial y} \delta y(b) = \\ &= \int_a^b \left(-\delta \dot{y} + \frac{\partial f}{\partial y} \delta y \right) \lambda \mathrm{d}t + \lambda \delta y|_a^b + \int_a^b \frac{\partial l}{\partial u} \delta u + \frac{\partial f}{\partial y} \delta y(b) \end{aligned}$$

Mithilfe der linearisierten Nebenbedingung, $-\delta \dot{y} + \frac{\partial f}{\partial y} \delta y = -\frac{\partial f}{\partial u} \delta u, \delta y(a) = 0$ ergibt sich

$$\left\langle \nabla \hat{J}(u), \delta u \right\rangle = \int_a^b \left(-\frac{\partial f}{\partial u} \lambda + \frac{\partial l}{\partial u} \right) \delta u \mathrm{d}t + \left(\lambda + \frac{\partial g}{\partial y} \right) \delta y \Big|_b$$

Somit resultiert die notwendige Optimalitätsbedingung (3.4.11) in

$$-\frac{\partial f}{\partial u}(t, y, u)\, \lambda(t) + \frac{\partial l}{\partial u}(t, y, u) = 0$$

λ ist der Lagrange-Multiplikator, welcher (3.4.13) sowie die Endbedingung $p(b) = -\frac{\partial g}{\partial y}(b, y(b))$ löst. Diese Ergebnisse können in einem ***Optimalitätssystem*** zusammengefasst werden, bestehend aus

- ***Zustandsgleichung***:
$$\dot{y} = f(t, y, u), \quad y(a) = y_a \tag{3.4.14}$$
- ***Adjungierter Gleichung***:
$$-\dot{\lambda} = \frac{\partial f}{\partial y} \lambda - \frac{\partial l}{\partial y}, \quad \lambda(b) = -\frac{\partial g}{\partial y}(b, y(b)) \tag{3.4.15}$$
- ***Optimalitätsbedingung***:
$$-\frac{\partial f}{\partial u} \lambda + \frac{\partial l}{\partial u} = 0 \tag{3.4.16}$$

Hierbei ist $\nabla \hat{J}(u) = -\frac{\partial f}{\partial u} + \frac{\partial l}{\partial u}$ der ***reduzierte Gradient***.

Bemerkung 3.4.6:

a) Das Optimalitätssystem (3.4.14)-(3.4.16) *wurde für skalare Funktionen aufgestellt. Im Falle vektorwertiger Funktionen, also* $y : [a, b] \to \mathbb{R}^n$, $u : [a, b] \to \mathbb{R}^m$, $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to$

$\mathbb{R}^n$, $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ *und* $\lambda : [a,b] \to \mathbb{R}^n$ *sind die Ableitungen* $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial u}$, $\frac{\partial l}{\partial y}$, $\frac{\partial l}{\partial u}$ *und* $\frac{\partial g}{\partial y}$ *Jacobi-Matrizen. Das Optimalitätssystem lautet dann*

$$\begin{aligned} \dot{y} &= f(t,y,u), \qquad y(a) = y_a \\ -\dot{\lambda} &= \left(\frac{\partial f}{\partial y}(t,y,u)\right)^T \lambda - \frac{\partial l}{\partial y}(t,y,u), \qquad \lambda(b) = -\frac{\partial g}{\partial y}(b,y(b)) \\ 0 &= -\left(\frac{\partial f}{\partial u}\right)^T \lambda + \frac{\partial l}{\partial u}(t,y,u) = 0 \end{aligned}$$

b) Das Optimalitätssystem kann auch erhalten werden durch die Lagrangefunktion

$$\mathcal{L}(y,u,\lambda) = J(y,u) + \int_a^b \left(\dot{y}(t) - f(t,y(t),u(t))\right)\lambda(t)\mathrm{d}t \tag{3.4.17}$$

Dann entsprechen (3.4.14)-(3.4.16) *den Bedingungen*

$$\nabla_\lambda \mathcal{L}(y,u,\lambda) = 0, \quad \nabla \mathcal{L}_y(y,u,p) = 0 \quad \nabla_u \mathcal{L}(y,u,p) = 0$$

c) Ist $U_{ad} \subset U$ konvex, abgeschlossen und beschränkt, wird die Optimalitätsbedingung (3.4.16) *zu einer Variationsungleichung:*

$$\left\langle -\frac{\partial f}{\partial u}\lambda + \frac{\partial l}{\partial u}, v - u \right\rangle \geq 0 \tag{3.4.18}$$

für alle $v \in U_{ad}$, u optimale Lösung.

d) Ist $U_{ad} = U = H^1(a,b)$ und $l(t,y,u) = \tilde{l}(t,y) + \frac{\nu}{2}\|u\|^2_{H^1(a,b)}$, so können Randbedingungen für u hinzugefügt werden: $u(a) = u_a, u(b) = u_b$. Insbesondere für $u_a = u_b = 0$ und $u \in C^2([a,b])$ ergibt sich

$$\begin{aligned} \int_a^b \frac{\partial l}{\partial u}\delta u &= \nu \int_a^b (u\delta u + \dot{u}\delta\dot{u})\,\mathrm{d}t \\ &= \nu \int_a^b (u - \ddot{u})\,\delta u \mathrm{d}t \end{aligned}$$

Somit lautet der L^2-Gradient

$$\nabla \hat{J}(u) = -\frac{\partial f}{\partial u}\lambda + \nu - \nu\ddot{u}$$

3.4.3 Hamilton-Funktion

In diesem Kapitell soll der Ansatz untersucht werden, Lösungen von optimalen Steuerungsproblem mittels der Hamiltonfunktion darzustellen. Zu diesem Zweck soll erneut das Problem

$$\begin{cases} \min J(y,u) \coloneqq \displaystyle\int_a^b l(t,y(t),u(t))\mathrm{d}t + g(b,y(b)) \\ \text{u. d. Nb. } \dot{y}(t) = f(t,y(t),u(t)), \quad y(a) = y_a \end{cases} \tag{3.4.19}$$

mit $l, f, g \in C^1$ betrachtet werden. Sei u^* eine optimale Steuerung und globales Minimum von $\hat{J}(u)$ in U. Variationen der Kontrollfunktion können beschrieben werden durch

$$u = u^* + \alpha\delta u$$

Diese Variation resultiert in einer Variation des Zustands,

$$y = y^* + \alpha\delta y,$$

wobei $y^* = S(u^*)$ und δy Lösung der linearisierten Nebenbedingung

$$\begin{cases} \delta\dot{y} = \dfrac{\partial f}{\partial y}(t, y^*, u^*)\,\delta y + \dfrac{\partial f}{\partial u}(t, y^*, u^*) \\ \delta y(a) = 0 \end{cases}$$

ist. Die Lagrangefunktion für (3.4.19) ist (s. (3.4.17))

$$\begin{aligned} \mathcal{L}(y, u, \lambda) &= J(y, u) + \int_a^b (\dot{y}(t) - f(t, y(t), u(t)))\,\lambda(t)\mathrm{d}t \\ &= g(b, y(b)) + \int_a^b l(t, y, u) + (\dot{y} - f(t, y, u))\,\lambda(t)\mathrm{d}t \end{aligned}$$

Die ***Hamilton-Pontryagin-Funktion*** ist definiert durch

$$H(t, y, u, \lambda) = f(t, y, u)\lambda - l(t, y, u).$$

Mit ihr kann die Lagrangefunktion umgeschrieben werden zu

$$\mathcal{L}(y, u, \lambda) = g(b, y(b)) + \int_a^b \dot{y}\lambda - H(t, y, u, \lambda)\mathrm{d}t$$

Im Folgenden soll nun die Variation von $\mathcal{L}$ in (y^*, u^*, λ) entlang der Nebenbedingung berechnet werden. Es gilt:

$$\begin{aligned} \mathcal{L}(S(u), u, \lambda) - \mathcal{L}(S(u^*), u^*, \lambda) = \alpha\frac{\partial g}{\partial y}(b, y^*(b))\delta y(b) + \alpha\int_a^b \Bigg(\delta\dot{y}\lambda - \frac{\partial H}{\partial y}(t, y^*, u^*, \lambda)\delta y \\ -\frac{\partial H}{\partial u}(t, y^*, u^*, \lambda)\delta u\Bigg)\,\mathrm{d}t + \text{Terme höherer Ordnung} \end{aligned}$$

Partielle Integration ergibt zusammen mit $\delta y(a) = 0$

$$\begin{aligned} \mathcal{L}(y, u, \lambda) = \alpha\left(\frac{\partial g}{\partial y}(b, y^*(b)) + \lambda(b)\right)\delta y(b) + \alpha\int_a^b \Bigg(-\left(\dot{\lambda} + \frac{\partial H}{\partial y}(t, y^*, u^*, \lambda)\right)\delta y \\ -\frac{\partial H}{\partial u}(t, y^*, u^*, \lambda)\delta u\Bigg)\,\mathrm{d}t + \text{Terme höherer Ordnung} \end{aligned}$$

Da u^* optimal ist, muss die erste Variation von $\mathcal{L}$ in u^* gleich 0 sein. Folglich muss gelten:

$$\begin{aligned} \dot{\lambda} &= -\frac{\partial H}{\partial y}(t, y^*, u^*, \lambda) \\ \lambda(b) &= -\frac{\partial g}{\partial y}(b, y^*(b)) \end{aligned} \tag{3.4.20}$$

Sei nun angenommen, dass eine Lösung $\lambda \in C^1([a, b])$ von (3.4.20) existiert, λ^* genannt. Nach Konstruktion lässt sich y als Lösung des Systems

$$\begin{aligned} \dot{y} &= \frac{\partial H}{\partial \lambda}(t, y^*, u^*, \lambda^*) \\ y(a) &= y_a \end{aligned}$$

schreiben. Des Weiteren erhält man durch Variation in der u-Komponente die Optimalitätsbedingung

$$\frac{\partial H}{\partial u}(t, y^*, u^*, \lambda^*) = 0. \tag{3.4.21}$$

Das bedeutet, dass für die Optimallösung (y^*, u^*, λ^*) die Funktion $H(t, y^*, \cdot, \lambda^*)$ für alle $t \in [a, b]$ ein Extremum in u^* hat. Der letzte Umstand ist sehr wichtig: Er führt auf das sogenannte ***Pontryagin'sche Maximumsprinzip*** , welches im folgenden Kapitel behandelt werden soll.

Bemerkung 3.4.7:

a) Gegeben sei das Problem

$$\min J(y, u) = \int_a^b l(t, y, u)\mathrm{d}t$$
$$\dot{y} = u, \quad y(a) = y_a, \quad y(b) = y_b$$

Dann ist $y(t) = (Su)(t) = \int_a^t u(\tau)\mathrm{d}\tau$ *und die zweite Variation von* $\hat{J}(u)$ *ist gegeben durch*

$$\delta^2 \hat{J}(u; \delta u) = \int_a^b \left(\frac{\partial^2 f}{\partial y^2}\delta y^2 + 2\delta y \delta \frac{\partial^2 f}{\partial y \partial u} + \frac{\partial^2 l}{\partial u^2}\delta u^2 \right) \mathrm{d}t$$

Die Legendrebedingung resultiert in $\frac{\partial^2 l}{\partial u^2} \geq 0$*, einer notwendigen Bedingungen für ein lokales Minimum, vgl. Sätze 3.3.18, 3.3.19. Dies bedeutet, dass* $\frac{\partial^2 H}{\partial u^2}(t, y^*, u^*, \lambda^*) \leq 0$ *ist. Daher ist das Extremum von* H *ein Maximum.*

b) Es liese sich auch ein anderer Lagrange-Multiplikator $\tilde{\lambda} = -\lambda$ *definieren. Dieser entspricht* $\mathcal{L} = J - \int_a^b (\dot{y} - f)\lambda \mathrm{d}t$*. In diesem Falle ist* $\tilde{H}$ *entsprechend durch* $\tilde{H} = f\tilde{p} + l = -H$ *definiert und* $\tilde{H}$ *hätte ein Minimum in* u^**.*

3.4.4 Pontryagin'sches Maximumsprinzip

Im vorherigen Kapitel wurde eine äquivalente Formulierung der Optimalitätsbedingungen (3.4.20)-(3.4.21) erarbeitet, welche zur Erkenntnis führte, dass die Hamilton-Pontryagin-Funktion ein Extremum bezüglich u in der Optimalen Lösung hat. Dies ist der Ausgangspunkt der Optimalen Steuerungstheorie, welche von *L. S. Pontryagin* und seinem Forschungsteam in den 50ern entwickelt wurde.

Hierbei ist zu beachten, dass der Ansatz im vorherigen Kapitel C^2-Regularität und eine unbeschränkte Menge von zulässigen Kontrollen voraussetzte. Wie in (3.4.18) gesehen, muss im anderen Fall für $u^* \in \partial U_{ad}$ die Bedingung $\frac{\partial H}{\partial u} = 0$ modifiziert werden. Auch kann sie nicht angewandt werden, wenn U eine diskrete Menge ist oder H nicht differenzierbar in U. Pontryagin's Ansatz umgeht diese Beschränkungen, indem einfach

$$H(t, y^*(t), u^*(t), \lambda^*(t)) \geq H(t, y^*(t), u(t), \lambda^*(t))$$

für alle u nahe an u^*, $t \in [a, b]$ gefordert wird.

Darüber hinaus sind im Folgenden auch Variationen der Kontrolle mit großer Norm erlaubt, solange die entsprechenden Trajektorien nahe bleiben. Dies entspricht dem Konzept von starken Minima.

Bevor jedoch das neue Prinzip hergeleitet wird, sollen einige Transformationen gezeigt werden, mit denen man optimale Steuerungsprobleme in eine gewünschte Form bringen kann. Betrachtet werde hierzu das Funktional mit Freien Endpunkten

$$J(y, u) = \int_a^b l(t, y, u)\mathrm{d}t + g_1(a, y(a)) + g_2(b, y(b))$$

Die Variable z kann, wie in (3.4.5), als Lösung von $\dot{z} = l(t, y, u)$, $z(a) = 0$ definiert werden. Somit wird eine Zustandsvariable addiert und das Funktional wird zu

$$J(y, z, u) = g_1(a, y(a)) + g_2(b, y(b)) + z(b)$$

Allgemeiner lässt sich dies nun als

$$J(a, y(a), b, y(b))$$

schreiben, wobei das neue y alle Zustandsvariablen beinhaltet. Gleichheitsnebenbedingungen in den Endpunkten können wie folgt formuliert werden:

$$K(a, y(a), b, y(b)) = 0 \tag{3.4.22}$$

Dies können zum Beispiel Anfangsbeidnungen $y(a) - y_a = 0$ und Endbedingungen $y(b) - y_b = 0$ sein. Im Folgenden bezeichne $d(K)$ die Dimension - besser gesagt die Anzahl von Gleichungen - von (3.4.22), in vorliegenden Beispiel also $d(K) = 2$. Unter Umständen können auch Ungleichungsnebenbedingungen vorliegen (z. B. $y(b) \leq B$), welche in

$$I(a, y(a), b, y(b)) \leq 0$$

zusammengefasst sind. ($d(I)$ bezeichne analog die Anzahl an Ungleichungen.) Das die Zustandsvariablen bestimmende Modell sei $\dot{y} = f(t, y, u)$ und es gelte $u(t) \in K_{ad}$. Nun kann das folgende optimale Steuerungsproblem (im englischen ***canonical Pontryagin type problem***) aufgestellt werden.

$$\begin{aligned} &\min\ J(a,y(a), b, y(b)) \\ \text{u. d. Nb.}\quad &\dot{y} = f(t, y(t), u(t), \quad u(t) \in K_{ad} \\ &K(a, y(a), b, y(b)) = 0 \\ &I(a, y(a), b, y(b)) \leq 0 \end{aligned}, \tag{3.4.23}$$

mit $y(t) \in \mathbb{R}^n$ und $u(t) \in \mathbb{R}^m$. Hierbei seien J, K, I in C^1 und f stetig, ebenso wie die Ableitungen $\frac{\partial f}{\partial t}$ und $\frac{\partial f}{\partial y}$. Die Menge K_{ad} ist beliebig. Die Lösung von (3.4.23) soll in der Menge der absolut stetigen Funktionen y und der messbaren Funktionen u gesucht werden. Ein Funktionenpaar $w = (y, u)$ zusammen mit einem Zeitabschnitt $[a, b]$, auf dem diese definiert, sind heißt ***Prozess***. Erfüllt ein Prozess alle gegebenen Nebenbedingungen, so ist er ***zulässig***. Existiert für einen zulässigen Prozess $w^* = (y^*(t), u^*(t)|t \in [a^*, b^*])$ ein $\varepsilon > 0$ mit $J(w) \geq J(w^*)$ für alle zulässigen Prozesse $w = ((y(t), u(t))|t \in [a, b])$, welche den Bedingungen

$$\begin{gathered} |a - a^*| < \varepsilon, \quad |b - b^*| < \varepsilon \\ \|y(t) - y^*(t)\|_2 < \varepsilon \quad \forall t \in [a, b] \cap [a^*, b^*] \end{gathered}$$

genügen, so stellt besagter Prozess ein ***starkes Minimum*** dar. Die zugehörige Hamilton-Pontryagin-Funktion ist (man beachte $\lambda(t) \in \mathbb{R}^n$

$$H(t, y, u, p) = \lambda^T f(t, y, u),$$

die zugehörige Endpunkt-Lagrange-Funktion lautet

$$l(a, y_a, b, y_b) = \left(\alpha_0 J + \alpha^T I + \beta^T K\right)(a, y_a, b, y_b)$$

mit $\alpha_0 \in \mathbb{R}$, $\alpha \in \mathbb{R}^{d(I)}$, $\beta \in \mathbb{R}^{d(K)}$. (Die Abhängigkeit der Funktion von von α_0, α und β wird aus Gründen der Übersichtlichkeit weggelassen.) Sei nun $w = ((y, u)|t \in [a, b])$ ein zulässiger Prozess von (3.4.23).

Definition 3.4.8:
w genügt dem ***Pontryagin'schen Maximumsprinzip***, falls $\alpha_0 \in \mathbb{R}$, $\alpha \in \mathbb{R}^{d(I)}$ und $\beta \in \mathbb{R}^{d(K)}$ sowie absolut stetige Funktionen $\lambda : \mathbb{R} \to \mathbb{R}^n$, $\mu : \mathbb{R} \to \mathbb{R}$ existieren, sodass Folgendes gilt:

(i) Nichtnegativität: $\alpha_0 \geq 0$, $\alpha \geq 0$ (komponentenweise)

(ii) Nichttrivialität: $\alpha_0 + \|\alpha\|_2 + \|\beta\|_2 > 0$ ($\Leftrightarrow$ mindestens ein Parameter ungleich 0)

(iii) Komplementaritätsbedingung: $\alpha^T I(a, y(a), b, y(b)) = 0$ (α_i ist 0 oder in der i. Ungleichung gilt Gleichheit)

(iv) Adjungierte Gleichung: $-\dot{\lambda} = \frac{\partial H}{\partial y}(t, y, u, \lambda)$, $-\dot{\mu} = \frac{\partial H}{\partial t}(t, y, u, \lambda)$

(v) Transversalitätsbedingung: $\lambda(a) = \frac{\partial l}{\partial y_a}(a, y(a), b, y(b))$, $\lambda(b) = -\frac{\partial l}{\partial y_b}(a, y(a), b, y(b))$

(vi) $H(t, y(t), u(t), \lambda(t)) + \mu(t) = 0$ für fast alle $t \in [a, b]$

(vii) $H(t, y(t), v, \lambda(t)) + \mu(t) \leq 0$ für alle $t \in [a, b]$, $v \in K_{ad}$.

Bedingung (vi) kann hierbei als „Energieentwicklungsgesetz“ betrachtet werden: Sie beschreibt die Gleichung $\dot{H} = \frac{\partial H}{\partial t}$, wobei H in mechanischen Systemen der Energie entspricht. (Ist H zeitunabhängig, wird dies zum Energieerhaltungssatz, da $\dot{H} \equiv 0$ $H \equiv$ konstant bedingt, vgl. Kapitel 2.6). Aus den Bedingungen (vi) und (vii) lässt sich schließlich die Maximalitätsbedingung für die Hamilton-Pontryagin-Funktion folgern:

$$\max_{v \in K_{ad}} H(t, y(t), v, \lambda(t)) = H(t, y(t), u(t), \lambda(t))$$

für fast alle $t \in [a, b]$, einem Umstand, dem (i)-(vii) ihren Namen verdanken.

Bemerkung 3.4.9: *Die Gleichung für μ liese sich direkt aus den anderen Bedingungen ableiten. Allerdings ist die explizite Forderung nützlicher für die Anwendung.*

Satz 3.4.10:
Stellt ein Prozess $w^* = ((y^*, u^*)|t \in [a, b])$ ein starkes Minimum von (3.4.23) dar, so genügt er dem Pontryagin'schen Maximumsprinzip

Bemerkung 3.4.11:

1) *Ist das Intervall $[a, b]$ fest und das System autonom, d. h. $\dot{y} = f(y, u)$, so werden die Bedingungen (i)-(vii) zu*

(i') $\alpha_0 \geq 0$, $\alpha \geq 0$

(ii') $\alpha_0 + \|\alpha\|_2 + \|\beta\|_2 > 0$

(iii') $\alpha I(y(a), y(b)) = 0$

(iv') $-\dot{\lambda} = \frac{\partial H}{\partial y} H(y, u, \lambda)$

(v') $\lambda(a) = \frac{\partial l}{\partial y_a}(y(a), y(b))$, $\lambda(b) = -\frac{\partial l}{\partial y_b}(y(a), y(b))$

(vi') $H(y(t), u(t), \lambda(t)) = c$ *fast überall in* $[a, b]$

(vii') $H(y(t), v, \lambda(t)) \leq c$ *für alle* $t \in [a, b], v \in K_{ad}$, $t \in [a, b]$

2) *Ein Problem mit einem nichtfestem Zeitintervall $[\hat{a}, \hat{b}]$ kann auf eines mit einem festen Intervall $[a, b]$ abgebildet werden: Sei τ so dass $t = t(\tau)$ und $\frac{\mathrm{d}x}{\mathrm{d}\tau} = v(t)$ gilt. Definiert man*

$\tilde{y}(\tau) \coloneqq y(t(\tau))$, $\tilde{u}(\tau) \coloneqq u(t(\tau))$, *so erhält man*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\tau}\tilde{y}(\tau) = f(t(\tau), \tilde{y}(\tau), \tilde{u}(\tau))^T v(\tau) \\ \dfrac{\mathrm{d}}{\mathrm{d}\tau}x(\tau) = v(\tau) \end{cases},$$

wobei v die Rolle einer neuen Kontrollvariable inne hat.

3.4.5 Beispiele

Beispiel 3.4.12 (Kontrolle von Produktion & Verbrauch):

Sei $y(t)$ der Arbeitsertrag der Wirtschaft und $u(t)$ der Anteil an reinvestiertem Kapital (jeweils zum Zeitpunkt t). Die Wirtschaft entwickle sich nach dem (illusorischen) Modell

$$\dot{y}(t) = u(t)y(t), \qquad y(0) = y_0 > 0$$

Sei $K_{ad} = [0, 1]$, $0 \leq u(t) \leq 1$. Das Ziel der Kontrolle u ist es, den Konsum zu maximieren:

$$\min J(y, u) \coloneqq -\int_0^T (1 - u(t))y(t)\mathrm{d}t$$

In dieser Form ist die Hamilton-Pontryagin-Funktion durch

$$H(t, y, u, \lambda) = (yu)\lambda + (1 - u)y = y + uy(\lambda - 1)$$

gegeben. Die adjungierte Gleichung ist $\dot{\lambda} = -\frac{\partial H}{\partial y}$, also

$$\dot{\lambda} = -1 + u(1 - \lambda), \quad \lambda(T) = 0$$

Gemäß dem Pontryagin'schen Maximumsprinzip erfüllt die optimale Lösung

$$H(t, y, u, \lambda) = \max_{0 \leq e \leq 1} \{y(t) + ey(t)(\lambda(t) - 1)\}, \quad t \in [0, 1],$$

und da $y > 0$ für ein sinnvolles Modell sein sollte, folgt

$$u(t) = \begin{cases} 1 & \lambda(t) > 1 \\ 0 & \lambda(t) \leq 1 \end{cases}$$

Für die adjungierte Gleichung gilt $\lambda(T) = 0$, nach Stetigkeit also $\lambda(t) \leq 1$ für $t \in \left(\tilde{t}, T\right]$. (Dieses Intervall sei so groß wie möglich.) In diesem Intervall gilt also $u(t) = 0$ und $\dot{\lambda} = -1$. Daher ist $\lambda(t) = T - t$ in $\left(\tilde{t}, T\right]$. In $\tilde{t}$ gilt $\lambda(\tilde{t}) = 1$, also $\tilde{t} = T - 1$. Betrachtet man nun die Zeit $t \leq T - 1$, so ist $\lambda(t) > 1$, $u(t) = 1$ und mit der adjungierte Gleichung

$$\dot{\lambda} = -1 + 1 - \lambda = -\lambda, \quad \lambda(T - 1) = 1$$

erhält man $\lambda(t) = e(T-1) - t, 0 \leq t \leq T-1$, und keine weiteren Sprünge in u treten auf. Somit ist die optimale Kontrolle gegeben durch

$$u(t) = \begin{cases} 1 & 0 \leq t < T-1 \\ 0 & T-1 \leq t \leq T \end{cases}$$

Nun soll diese Analyse mit einer Transformation des Funktionals J wiederholt werden. Sei hierzu $\dot{z} = -(1-u)y$, $z(0) = 0$ sowie $J = z(T)$. J soll minimiert werden unter den Nebenbedingungen

$$\begin{aligned} \dot{y} &= uy, \quad y(0) = y_0 > 0 \\ \dot{z} &= -(1-u)y, \quad z(0) = 0 \end{aligned}$$

Die HP-Funktion lautet nun, wegen $z(T) = \int_0^T 0 \mathrm{d}t + \underbrace{\mathrm{id}}_{g} z(T)$

$$H(t, (y,z), u, \lambda) = \lambda_1(uy) + \lambda_2 \left(-(1-u)y\right)$$

Die adjungierten Gleichungen sind gegeben durch

$$\begin{aligned} \dot{\lambda}_1 &= -\frac{\partial H}{\partial y} = -u\lambda_1 + (1-u)\lambda_2 \\ \dot{\lambda_2} &= -\frac{\partial H}{\partial z} = 0 \quad \Rightarrow \ \lambda_2(t) = \text{konstant} \end{aligned}$$

H hängt ferner nicht explizit von t ab. Die Lagrange-Funktion für den Endzeitpunkt lautet

$$l = \alpha_0 z(T) + \beta_1(y(0) - y_0) + \beta_2(z(0)),$$

es ergibt sich

$$\begin{aligned} \lambda_1(0) &= \beta_1 \quad & \lambda_1(T) &= 0 \\ \lambda_2(0) &= \beta_2 \quad & \lambda_2(T) &= -\alpha_0 \end{aligned}$$

Durch Skalierung, sodass $\alpha_0 = 1$ ist, erhält man $\lambda_2(t) = -1 = \beta_2$. Daher:

$$\dot{\lambda}_1 = -u\lambda_1 + u - 1, \quad \lambda_1(T) = 0,$$

analog zum vorherigen ist

$$\lambda_1(t) = \begin{cases} T-t & T-1 \leq t \leq T \\ e^{(T-1)-t} & 0 \leq t < T-1 \end{cases},$$

λ_2 erfüllt $\lambda_2(0) = \beta_1 = e^{T-1}$. H ist entlang der optimalen Trajektorie konstant, $H = y_0 e^{T-1}$.

Beispiel 3.4.13 (Linear-quadratisches Problem):

Gegeben sei das Problem

$$\begin{aligned} \min J(y,u) &:= \int_0^T \left(y^2 + u^2\right) \mathrm{d}t \\ \text{u. d. Nb.} \quad \dot{y} &= y + u, \quad y(0) = y_0 \end{aligned}$$

Mit Hilfe der Hamilton-Funktion $H(y,u,\lambda) = \lambda(y+u) - (y^2+u^2)$ ergibt sich als adjungierte

Gleichung

$$\dot{\lambda} = -\frac{\partial H}{\partial y} = -\lambda + 2y$$

Da H in der optimalen Lösung ein Extremum bzgl. u hat, lässt sich das optimale u mittels $0 = \frac{\partial H}{\partial u} = \lambda - 2u$ finden. Also ist $u = \frac{\lambda}{2}$, womit sich das System

$$\begin{aligned} \dot{y} &= y + \frac{\lambda}{2} \\ \dot{\lambda} &= -2y - \lambda \end{aligned}$$

ergibt. Die Lösung hierzu lautet

$$\begin{aligned} y(t) &= \frac{e^{-\sqrt{2}t} \cdot \left(-1 + e^{2\sqrt{2}t}\right)}{4\sqrt{2}} c_1 + \frac{e^{-\sqrt{2}t} \cdot \left(2 - \sqrt{2} + \left(2 + \sqrt{2}\right) e^{2\sqrt{2}t}\right)}{4} c_2 \\ \lambda(t) &= -\frac{e^{-\sqrt{2}t} \cdot \left(-2 - \sqrt{2} - \left(2 - \sqrt{2}\right) e^{2\sqrt{2}t}\right)}{4} c_1 + \frac{e^{-\sqrt{2}t} \cdot \left(-1 + e^{2\sqrt{2}t}\right)}{\sqrt{2}} c_2 \end{aligned}$$

Aus $y(0) = y_0$ folgt $c_2 = y_0$, aus $\lambda(T) = 0$ ergibt sich schließlich $c_1 = \left(\frac{1}{2} - \frac{\coth \sqrt{2}T}{\sqrt{2}}\right) y_0$.

3.4.6 Lineare und nichtlineare Steuerungsprobleme

Betrachtet werden soll nun folgender Kontroll-Mechanismus

$$\begin{cases} \dot{y} = f(u) + Bu \\ y(0) = y_0 \end{cases} \tag{3.4.24}$$

mit zeitabhängiger Zustandsvariable $y : \mathbb{R} \to E_1$ und Kontrollfunktion $u : \mathbb{R} \to E_2$ (E_1, E_2 reelle Vektorräume) sowie linearem Operator B. Ferner sei

$$J : (y, u) \mapsto J(y, u) \in \mathbb{R}$$

ein gegebenes Zielfunktional. Das zugehörige optimale Steuerungsproblem lautet dann:

$$\min_{y,u} J(y, u), \quad \text{so, dass (3.4.24) erfüllt ist.} \tag{3.4.25}$$

Wie bereits gesehen, wird durch das System (3.4.25) eine Kontroll-Zustands-Funktion $u \mapsto y$ beschrieben, wodurch sich das Problem (3.4.25) auch in der Form

$$\min_u \hat{J}(u) \coloneqq J(y(u), u) \tag{3.4.26}$$

schreiben lässt. Im Falle der Differenzierbarkeit von $\hat{J}(u)$ reduziert sich das Problem somit auf die Lösung von

$$\nabla_u \hat{J}(u) = \nabla_u J(y(u), u) = \nabla_y J(y, u) \nabla_u y(u) + \nabla_u J(y, u) = 0$$

Im Folgenden sei zu (3.4.24) die Steuerung $u \in L^2(0, T)$ gegeben.

Lemma 3.4.14:
Gegeben sei das System $\dot{y} = f(y) + Bu$, mit $y(t) \in D \subset \mathbb{R}^n \, \forall t$, $u \in L^2((0,T);\mathbb{R}^n)$ und $f : D \to \mathbb{R}^n$ sei lokal Lipschitz-stetig im Gebiet D. Dann existiert für alle $t \in [0,T]$ und alle Anfangswerte $y_0 \in D$ eine eindeutige Lösung $y = y(u)$.

Beweis: Sei $F(t,y) \coloneqq f(y) + Bu(t)$, F is messbar für jedes feste y und stetig für jedes feste t. Ferner gelten

$$\|F(t,x) - F(t,y)\| = \|f(x) + Bu(t) - f(y) - Bu(t)\| = \|f(x) - f(y)\| \leq c\,\|x - y\|$$

sowie

$$\|F(t,y)\| = \|f(y) + Bu(t)\| \leq \|f(x)\| + \|B\|\,\|u\| \leq c_y + \|B\|\,\|u\| = \beta(t).$$

Somit folgt die Behauptung aus dem Satz von Caratheodory. □

Das optimale Steuerungsproblem soll nun wie folgt genauer spezifiziert werden:

$$\begin{cases} \min J(y,u) \coloneqq \dfrac{1}{2}\|y - y_d\|^2_{L^2(0,T)} + \frac{\alpha}{2}|y(T) - y_T|^2 + \frac{\nu}{2}\|u\|^2_{L^2(0,T)} \\ \text{u. d. Nb. } \dot{y} = f(y) + Bu, \quad y(0) = y_0 \end{cases} \tag{3.4.27}$$

mit $\nu, \alpha > 0$ und $y, u \in L^2(0,T)$.

Satz 3.4.15:
Sei X normierter Raum. Dann ist eine konvexe, stetige Funktion $J : X \to \mathbb{R}$ schwach unterhalbstetig.

Zur Erinnerung: Ist J koerzitiv und unterhalbstetig, sowie die U als Teilmenge eines reflexiven Banachraums nichtleer und konvex, so existiert eine minimierende Folge für das Problem $\min\limits_{u \in U} J(u)$.

Lemma 3.4.16:

a) Das Funktional

$$J(y,u) = \tfrac{1}{2}\|y - y_d\|^2_{L^2(0,T)} + \tfrac{\alpha}{2}|y(T) - y_T|^2 + \tfrac{\nu}{2}\|u\|^2_{L^2(0,T)}$$

ist konvex und stetig, daher auch schwach unterhalbstetig sowie koerzitiv in U.

b) Es existiert eine minimierende Folge $(u^m)_m$ mit $u^m \in L^2(0,T)$ und $y^m = y(u^m)$ so dass

$$\lim_{m \to \infty} J(y^m, u^m) = \inf_{u \in U} J(y(u), u)$$

gilt.

Es wurde bereits gezeigt, dass für eine Folge $(u^{m_l})_l$ mit schwachem Grenzwert $\tilde{u}$ in $L^2(0,T)$, die entsprechende Folge $y^{m_l} := y(u^{m_l})$ gegen $\tilde{y} = y(\tilde{u})$ konvergiert. Ist $\nu > 0$, $U = L^2(0,T)$, so hat das Problem (3.4.27) folglich eine Lösung.

Beispiel 3.4.17:

Gegeben sei das Problem

$$\begin{cases} \min J(y,u),\ J(y,u) \coloneqq \frac{1}{2}\underbrace{\int_0^T (y(t))^2\,\mathrm{d}t}_{=\|y\|^2_{L^2(0,T)}} + \frac{\alpha}{2}\underbrace{\int_0^T (u(t))^2\,\mathrm{d}t}_{=\|u\|^2_{L^2(0,T)}} \\ \dot{y} = y + u, \quad y(0) = y_0 \end{cases}$$

Hier ist

$$f\left(t, y\left(t\right), u\left(t\right)\right) = y\left(t\right) + u\left(t\right)$$

sowie

$$l\left(t, y\left(t\right), u\left(t\right)\right) = \frac{1}{2}\left(y\left(t\right)\right)^2 + \frac{\alpha}{2}\left(u\left(t\right)\right)^2.$$

Somit ist

$$\dot{\lambda} = -\frac{\partial f}{\partial y}\lambda + \frac{\partial l}{\partial y} = -\lambda + y\left(t\right)$$

(Gemäß Kapitel 1.1 hat λ also die Form $\lambda(t) = e^{-t}\left(c + \int_0^t y\left(t\right)e^t dt\right)$). Ferner muss gelten:

$$0 = -\lambda\frac{\partial f}{\partial u} + \frac{\partial l}{\partial u} = -\lambda + \alpha u\left(t\right)$$

Die Lagrangefunktion ist:

$$\mathcal{L}\left(y, u, \lambda\right) = \frac{1}{2}\int_0^T \left(y\left(t\right)\right)^2 dt + \frac{\alpha}{2}\int_0^T \left(u\left(t\right)\right)^2 dt + \int_0^T \left(\dot{y} - y - u\right)\lambda dt$$

Für die Gradienten sowie (zulässige) Störungen gilt somit:

- $\nabla_u \mathcal{L}(y,u,\lambda)$:

$$\begin{aligned} \left\langle \nabla_u \mathcal{L}(y,u,\lambda), \delta u\right\rangle_{L^2(0,T)} &= \lim_{\mu\to 0}\frac{1}{\mu}\left(\frac{\alpha}{2}\int_0^T (u+\mu\delta u)^2\,\mathrm{d}t + \int_0^T (\dot{y} - y - u - \mu\delta u)\lambda \mathrm{d}t\right. \\ &\qquad \left. -\frac{\alpha}{2}\int_0^T (u(t))^2\,\mathrm{d}t - \int_0^T (\dot{y} - y - u)\lambda \mathrm{d}t\right) = \\ &= \lim_{\mu\to 0}\frac{1}{\mu}\left(\frac{\alpha}{2}\int_0^T 2\mu\cdot u\delta u + \mu^2(\delta u)^2\,\mathrm{d}t - \int_0^T (\mu\delta u)\lambda \mathrm{d}t\right) = \\ &= \int_0^T \alpha u\delta u \mathrm{d}t - \int_0^T \delta u\lambda \mathrm{d}t = \left\langle \alpha u - \lambda, \delta u\right\rangle_{L^2(0,T)} \end{aligned}$$

- $\nabla_y \mathcal{L}(y,u,\lambda)$:

$$\begin{aligned}\langle \nabla_y \mathcal{L}(y,u,\lambda), \delta y\rangle_{L^2(0,T)} &= \lim_{\mu\to 0}\frac{1}{\mu}\left(\frac{1}{2}\int_0^T (y+\mu\delta y)^2\,\mathrm{d}t + \int_0^T (\dot{y}+\mu\dot{\delta y} - y - \mu\delta y - u))\lambda \mathrm{d}t\right.\\ &\qquad \left. -\frac{1}{2}\int_0^T (y(t))^2\,\mathrm{d}t - \int_0^T (\dot{y}-y-u)\lambda \mathrm{d}t\right) =\\ &= \lim_{\mu\to 0}\frac{1}{\mu}\left(\frac{1}{2}\int_0^T 2\mu\cdot y\delta y + \mu^2(\delta y)^2\,\mathrm{d}t + \int_0^T \mu(\dot{\delta y}-\delta y)\lambda \mathrm{d}t\right) =\\ &= \int_0^T y\delta y \mathrm{d}t + \int_0^T (\dot{\delta y}-\delta y)\lambda \mathrm{d}t =\\ &= \int_0^T y\delta y \mathrm{d}t + \underbrace{\delta y(t)\lambda(\,t)|_0^T}_{\delta y(0)=0} - \int_0^T \delta y\dot{\lambda} + \delta y\lambda \mathrm{d}t\\ &= \left\langle y-\dot{\lambda}-\lambda, \delta y\right\rangle_{L^2(0,T)} + \underbrace{\lambda(T)\delta y(T)}_{=0\forall\delta y\Rightarrow\lambda(T)=0}\end{aligned}$$

- $\nabla_\lambda \mathcal{L}(y,u,\lambda)$:

$$\begin{aligned}\langle \nabla_\lambda \mathcal{L}(y,u,\lambda), \delta\lambda\rangle_{L^2(0,T)} &= \lim_{\mu\to 0}\frac{1}{\mu}\left(\int_0^T (\dot{y}-y-u)\,(\lambda+\mu\delta\lambda)\,\mathrm{d}t - \int_0^T (\dot{y}-y-u)\lambda \mathrm{d}t\right) =\\ &= \int_0^T (\dot{y}-y-u)\,\delta\lambda \mathrm{d}t = \langle \dot{y}-y-u, \delta\lambda\rangle_{L^2(0,T)}\end{aligned}$$

Die Gradienten sind also: $\nabla_u \mathcal{L}(y,u,\lambda) = \alpha u - \lambda$, $\nabla_y \mathcal{L}(y,u,\lambda) = y - \dot{\lambda} - \lambda y$ und $\nabla_\lambda \mathcal{L}(y,u,\lambda) = \dot{y} - y - u$, das Optimalitätssystem lautet:

$$\begin{cases} \dot{y}\,(t) = y + u, & y\,(0) = y_0\\ \dot{\lambda}\,(t) = y - \lambda y, & \lambda\,(T) = 0\\ -\lambda + \alpha u = 0, & t\in(0,T)\end{cases}$$

Beispiel 3.4.18:

Betrachtet werden soll das Problem:

$$\begin{cases} \min J\,(y,u),\ J\,(y,u) = \dfrac{\alpha}{2}\,\|y-y_d\|^2_{L^2(0,T)} + \dfrac{\beta}{2}\,|y\,(T)-y_T|^2 + \dfrac{\nu}{2}\,\|u\|^2_{L^2(0,T)}\\ \text{u. d. Nb. } \dot{y} = f\,(y,u)\,, y\,(0) = y_0\end{cases}$$

Die Lagrangefunktion lautet:

$$\mathcal{L}\,(y,u,\lambda) = \frac{\alpha}{2}\,\|y-y_d\|^2_{L^2(0,T)} + \frac{\beta}{2}\,|y\,(T)-y_T|^2 + \frac{\nu}{2}\,\|u\|^2_{L^2(0,T)} + \int_0^T (\dot{y}-f(y,u))\,\lambda \mathrm{d}t$$

Das Optimalitätssystem ergibt sich aus :

$$\begin{cases} \nabla_\lambda \mathcal{L}\,(y,u,\lambda) = 0\\ \nabla_y \mathcal{L}\,(y,u,\lambda) = 0\\ \nabla_u \mathcal{L}\,(y,u,\lambda) = 0\end{cases}$$

Sei nun δy eine beliebige zulässige Richung, wegen $(y+\delta y)(0) = y_0$ muss also folglich $\delta y(0) = 0$ gelten.

$$\begin{aligned}
\langle \nabla_y \mathcal{L}, \delta y\rangle_{L^2(0,T)} &= \lim_{s\to 0} \frac{1}{s}\left[\mathcal{L}(y+\mu\delta y, u, \lambda) - \mathcal{L}(y,u,\lambda)\right] = \\
&= \lim_{s\to 0} \frac{1}{s}\left[\frac{\alpha}{2}\|y+s\delta y - y_d\|^2_{L^2(0,T)} + \frac{\beta}{2}|y(T)+s\delta y(T) - y_T|^2_{L^2(0,T)}\right. \\
&\qquad + \int_0^T \left(\dot{y} + s\dot{\delta y} - f(y+s\delta y, u)\right)\lambda \mathrm{d}t - \frac{\alpha}{2}\|y-y_d\|^2 \\
&\qquad \left. -\frac{\beta}{2}|y(T)-y_T|^2 - \int_9^T (\dot{y} - f(y,u))\lambda \mathrm{d}t\right] = \\
&= \lim_{s\to 0} \frac{1}{s}\left[\frac{\alpha}{2}s^2\|\delta y\|^2_{L^2(0,T)} + \frac{\alpha}{2}2s\langle y-y_d, \delta y\rangle_{L^2(0,T)} + \frac{\beta}{2}s^2|\delta y(T)|^2\right. \\
&\qquad + \frac{\beta}{2}2s(y(T)-y_T)\delta y(T) \\
&\qquad + \int_0^T \left(s\dot{\delta y} - f(y,u) - \frac{\partial f}{\partial y}(y,u)(s\delta y) + O\left(s^2\right)\right)\lambda \mathrm{d}t \\
&\qquad \left. - \int_0^T (-f(y,u))\lambda \mathrm{d}t\right] = \\
&= \alpha\langle y-y_d, \delta y\rangle_{L^2(0,T)} + \beta(y(T)-y_T)\delta y(T) \\
&\qquad + \int_0^T \left(\dot{\delta y} - \frac{\partial f}{\partial y}(y,u)\delta y\right)\lambda \mathrm{d}t = \\
&\overset{\substack{\text{part. Int.}\\ \delta y(0)=0}}{=} \int_0^T \left(-\dot{\lambda} - \frac{\partial f}{\partial y}(y,u)\lambda + \alpha(y-y_d)\right)\delta y \mathrm{d}t \\
&\qquad + (\beta(y(T)-y_T) + \lambda(T))\delta y(T) \overset{!}{=} 0
\end{aligned}$$

Da dies für alle zulässigen δu gelten soll, folgt

$$-\dot{\lambda} - \frac{\partial f}{\partial y}(y,u)\lambda + \alpha(y-y_d) \equiv 0\,, \lambda(T) + \beta(y(T)-y_T) = 0$$

δu sei beliebige Störung. Es muss gelten:

$$\begin{aligned}
\langle \nabla_u \mathcal{L}, \delta u\rangle_{L^2(0,T)} &= \lim_{s\to 0} \frac{1}{s}\left(\mathcal{L}(y, u+s\delta u, \lambda) - \mathcal{L}(y,u,\lambda)\right) = \\
&= \lim_{s\to 0} \frac{1}{s}\left[\frac{\nu}{2}\|u+s\delta u\|^2_{L^2(0,T)} + \int_0^T (\dot{y} - f(y, u+\delta u))\lambda \mathrm{d}t\right. \\
&\qquad \left. -\frac{\nu}{2}\|u\|^2_{L^2(0,T)} - \int_0^T (\dot{y} - f(y,u))\lambda \mathrm{d}t\right] = \\
&= \nu\langle u, \delta u\rangle_{L^2(0,T)} + \int_0^T \left(-\frac{\partial f}{\partial u}(y,u)\right)\delta u \lambda \mathrm{d}t = \\
&= \int_0^T \left(\nu u - \frac{\partial f}{\partial u}(y,u)\lambda\right)\delta u \mathrm{d}t = 0
\end{aligned}$$

(In diesem Falle sind alle δu zulässig). Also muss

$$\nu u - \frac{\partial f}{\partial u}(y,u)\lambda = 0$$

gelten. Für beliebiges $\delta\lambda$ muss $\nabla_\lambda \mathcal{L}$ die Gleichung

$$\begin{aligned}\langle \nabla_\lambda \mathcal{L}, \delta\lambda \rangle_{L^2(0,T)} &= \lim_{s\to 0} \frac{1}{s} \left(\mathcal{L}\left(y, u, \lambda + s\delta\lambda\right) - \mathcal{L}\left(y, u, \lambda\right)\right) = \\ &= \ldots = \\ &= \int_0^T \left(\dot{y} - f(y,u)\right) \delta\lambda \mathrm{d}t = 0\end{aligned}$$

erfüllen. Somit ergibt sich folgendes Optimalitätssystem:

$$\begin{cases} \dot{y} = f\left(y,u\right), \quad y\left(0\right) = y_0 \\ -\dot{\lambda} = \dfrac{\partial f}{\partial y}\left(y,u\right)\lambda - \alpha\left(y - y_d\right), \quad \lambda\left(T\right) = -\beta\left(y\left(T\right) - y_T\right) \\ \nu u - \dfrac{\partial f}{\partial u}\left(y,u\right)\lambda = 0 \end{cases}$$

Beispiel 3.4.19:

Abschließend soll noch das Problem

$$\begin{cases} \min J\left(y,u\right), \quad J\left(y,u\right) \coloneqq \dfrac{1}{2}\left\|y - y_d\right\|^2_{L^2(0,T)} + \dfrac{\nu}{2}\left\|u\right\|^2_{L^2(0,T)} \\ \text{u. d. Nb.} \begin{cases} -\ddot{y} = u \\ y\left(0\right) = 0 \\ y\left(T\right) = 0 \end{cases} \end{cases}$$

betrachtet werden, bei dem die Nebenbedingung eine DGL. 2 Ordnung ist. Die zugehörige Lagrangefunktion lautet

$$\mathcal{L}\left(y,u,\lambda\right) = \frac{1}{2}\left\|y - y_d\right\|^2_{L^2(0,T)} + \frac{\nu}{2}\left\|u\right\|^2_{L^2(0,T)} + \int_0^T \left(\ddot{y} + u\right)\lambda \mathrm{d}t$$

Analog zu obigen Beispielen ergibt sich aus

$$\begin{aligned}\left\langle \nabla_\lambda \mathcal{L}\left(y,u,\lambda\right), \delta\lambda \right\rangle_{L^2(0,T)} &= \lim_{s\to 0} \frac{\mathcal{L}\left(y,u,\lambda + s\delta\lambda\right) - \mathcal{L}\left(y,u,\lambda\right)}{s} = \\ &= \int_0^T \left(\dot{y} + u\right)\delta\lambda \mathrm{d}t \overset{!}{=} 0 \quad \forall \delta\lambda\end{aligned}$$

die Bedingung

$$\nabla_\lambda \mathcal{L}(y,u,\lambda) = \ddot{y} + u = 0.$$

Sei δy zulässige Störung, d. h. es gilt $\delta y(0) = \delta y(T) = 0$. Es muss

$$\begin{aligned}
\langle \nabla_y \mathcal{L}(y,u,\lambda), \delta y\rangle_{L^2(0,T)} &= \lim_{s\to 0} \frac{\mathcal{L}(y+s\delta y,u,\lambda) - \mathcal{L}(y,u,\lambda)}{s} = \\
&= \lim_{s\to 0} \frac{1}{s}\left(\frac{1}{2} 2s \langle y - y_d, \delta y\rangle_{L^2(0,T)} + s^2 \|\delta y\|^2_{L^2(0,T)} + \int_0^T s\ddot{\delta y}\lambda \mathrm{d}t\right) = \\
&= \langle y - y_d, \delta y\rangle_{L^2(0,T)} + \int_0^T \ddot{\delta y}\lambda \mathrm{d}t = \\
&\overset{\text{part. Int.}}{=} \langle y - y_d, \delta y\rangle_{L^2(0,T)} + \dot{\delta y}\lambda\Big|_0^T - \int_0^T \delta\dot{y}\dot{\lambda}\mathrm{d}t = \\
&\overset{\text{part. Int.}}{=} \langle y - y_d, \delta y\rangle_{L^2(0,T)} + \dot{\delta y}\lambda\Big|_0^T - \delta y\dot{\lambda}\Big|_0^T + \int_0^T \delta y\ddot{\lambda}\mathrm{d}t = \\
&= \langle y - y_d, \delta y\rangle_{L^2(0,T)} + \left(\dot{\delta y}(T)\,\lambda(T) - \dot{\delta y}(0)\,\lambda(0)\right) + \int_0^T \delta y\ddot{\lambda}\mathrm{d}t = 0
\end{aligned}$$

gelten. Hieraus folgt die Bedingung

$$y - y_d + \ddot{\lambda} = 0, \quad \left(\dot{\delta y}(T)\,\lambda(T) - \dot{\delta y}(0)\,\lambda(0)\right) = 0$$

3.4.7 Algorithmus zur Lösung optimaler Steurungsprobleme

Gegeben sei das Problem

$$\begin{cases} \min J(y,u),\ J(y,u) = \displaystyle\int_0^T l(t,y(t),u(t))\mathrm{d}t \\ \text{u. d. Nb. } \dot{y} = f(y,u), \quad y(0) = y_0 \end{cases},$$

wobei J stetig differenzierbar sei. Ferner existiere zu jedem u ein eindeutiges $y(u)$ mit $(\dot{u}) = f(y(u),u)$ und $y(u)(0) = y_0$. Das reduzierte Funktional $\hat{J}(u) \coloneqq J(y(u),y)$ sei ferner konvex. (Somit existiert eine eindeutige Lösung). Der folgende Algorithmus, ähnlich der Gradienten-Projektion aus [28], greift auf folgende Subroutinen zurück, welche alle als Input die/eine Steuerung u erhalten:

GB (Gradient berechnen): Berechnet zu gegebener Kontrolle u den Gradienten des reduzierten Funktionals, $\nabla\hat{J}(u) = \nabla J(y(u),u)$ wie folgt:

1) Vorwärtslösung von $\begin{cases} \dot{y}(t) = f(t,y(t),u(t)) \\ y(0) = y_0 \end{cases}$ ($\rightarrow$ Methoden s. Kapitel 1.11)
2) Rückwärtslösung von $\begin{cases} -\dot{\lambda} = \frac{\partial f}{\partial y}(t,y(t),u(t))\lambda - \frac{\partial l}{\partial y}(t,y(t),u(t)) \\ \lambda(T) = 0 \end{cases}$
3) Zusammensetzen: $\nabla\hat{J}(u) = -\frac{\partial f}{\partial u}\lambda + \frac{\partial l}{\partial u}$

LS (Line search/Backtracking): Berechnet zu gegebener Kontrolle u^0 eine bessere Kontrollfunktion wie folgt:

1) Berechnung von $\hat{u} = u^{k-1} - \alpha^k \nabla\hat{J}(u^{k-1})$, wobei $\alpha^k > 0$ am Anfang den Wert α_0 hat.
2) Berechnung der Lösung $\hat{y}$ des AWPs

$$\begin{cases} y' = f(t,y,\hat{u}) \\ y(0) = y_0 \end{cases}$$

3) Berechnung von $J(\hat{y}, \hat{u})$

4) Vergleich: Ist $J(\hat{y}, \hat{u}) \leq \hat{J}(u^{k-1}) - \delta \left\| \nabla \hat{J}(u^{k-1}) \right\|^2$, $\delta \in (0,1)$, so ist α^k „gut“ $\rightarrow$ Als neue Kontrolle wird $u^k = \hat{u}$ genommen. Gilt die Ungleichung nicht, werden die Schritte 1-4 wiederholt, jedoch mit kleinerem α^k (z. B. $\frac{\alpha^k}{2}$ oder $\alpha^k \beta$, $\beta \in (0,1)$). (Diese Bedingung nennt man ***Armijo-Regel***. Andere Bedingungen führen evtl. zu einer höheren Genauigkeit, sind aber u. U. aufwendiger zu berechnen.)

Der eigentliche Algorithmus lautet wie folgt:

```
Input: u^0 = u^0(t) auf Gitter
GB → ∇Ĵ(u^0)
Optimierungsschleife: for k = 1, ..., k_max
1.  LS → α^k
2.  u^k = u^{k-1} − α∇Ĵ(u^{k-1})
3.  GB → ∇Ĵ(u^k)
4.  if ||∇Ĵ(u^k)|| ≤ ε stop %ε > 0: Fehlertoleranz
else goto 1
```

Bemerkung 3.4.20:

a) Obiger Algorithmus ist nur eine mögliche Grobstruktur für die generelle Lösung optimaler Steuerungsprobleme. Bei der tatsächlichen Implementierung sind durch aus Verbesserungen möglich. So ist es beispielsweise sinnvoll, sobald man zu einer Kontrollfunktion den zugehörigen Zustand berechnet hat, diesen zwischenzuspeichern, anstatt ihn in jedem Schritt neu zu berechnen ($\rightarrow$ erspart Rechenzeit). Arbeitet man ferner mit Programmen, die Verweise auf Funktionen und Routinen als Funktionsparameter akzeptieren (wie z. B. MATLAB), kann man auch Zielfunktional und Gradient als Parameter übergeben. Auf diese Weise ist der Algorithmus für weitere Optimierungsprobleme mit anderen Bedingungen wiederverwendbar.

b) Zur Rückwärtslösung: Gegeben sei das Problem

$$\begin{cases} -\dot{\lambda}(t) = g(t, y(t), u(t))\,\lambda + b(t) \\ \lambda(T) = 0 \end{cases}$$

sowie die Gitterschrittweite $h = \frac{T}{N}$. Zur Lösung kann man beispielsweise expliziten,

$$\lambda_n = \lambda_{n-1} + h\left(g(t_{n-1}, y_{n-1}, u(t_{n-1}))\lambda_{n-1} + b(t_{n-1})\right),$$

oder impliziten Euler,

$$\lambda_n = \lambda_{n-1} + h\left(g(t_n, y_n, u(t_n))\,\lambda_n + b(t_n)\right),$$

anwenden. Da bei der Rückwärtslösung jedoch aus λ_n λ_{n-1} berechnet wird, ist in diesem Zusammenhang der implizite Euler ein explizites Verfahren, der explizite hingegen ein implizites. Als Startknoten dient hierbei $t_N = T$ mit $\lambda_N = 0$.

*c) In Anhang C.3 befindet sich eine mögliche Implementierung des obigen Gradientenverfahrens, welches die Linesearch mittels der sogenannten **Armijo-Regel** durchführt (dies ist die oben beschriebene). Des Weiteren befindet sich im Anhang auch ein Testskript für das Problem*

$$\begin{aligned} \min J(y,u) &= \int_0^1 y(t)^2 + u(t)^2 \mathrm{d}t \\ \text{u. d. Nb. } \dot{y} &= y + u^2,\ y(0) = 1 \end{aligned} \tag{3.4.28}$$

An diesem wird auch die Wichtigkeit der Iterationsbeschränkung wichtig: Die berechnete Lösung erfüllt nicht das Abbruchkriterium. Grund hierfür ist, dass sowohl bei der Berechnung des Integrals als auch bei der Lösung der Differentialgleichungen nur approximative Lösungen berechnet werden. Da als Schrittweite $h = \frac{1}{10}$ dient und nur die Eulerverfahren verwendet werden, kann die gewünschte Genauigkeit nicht erreicht werden. Andererseits würde eine zu geringe Genauigkeit evtl. dazuführen, dass der Algorithmus zu früh abbricht.

3.4.8 Abschließende Bemerkungen und Ausblicke

In Kapitel 3.4.6 wurden optimale Steuerungsprobleme, je nach Form, in folgende Klassen eingeteilt:

- Opt.-Strg.-Problem in Lagrange-Form:

$$\begin{cases} \min \int_a^b l\left(t, y(t), u(t)\right) \mathrm{d}t \\ \text{u. d. Nb. } \dot{y} = f(y,u), y(a) = y_a \end{cases}$$

- Opt.-Strg.-Problem in Bolza-Form:

$$\begin{cases} \min J(y,u),\ J(y,u) \coloneqq \varphi(a, y(a), b, y(b)) + \int_a^b l\left(t, x(t), u(t)\right) \mathrm{d}t \\ \text{u. d. Nb. } \dot{y} = f(y,u), y(a) = y_a \end{cases}$$

- Opt.-Strg.-Problem in Meyer-Form:

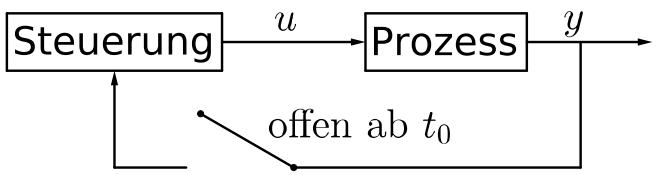
$$\begin{cases} \min J(y,u),\ J(y,u) \coloneqq \varphi(a, y(a), b, y(b)) \\ \text{u. d. Nb. } \dot{x} = f(y,u), y(a) = y_a \end{cases}$$

Stattdessen lassen sich optimale Steuerungsprobleme jedoch auch hinsichtlich der Art der Kontrollvorschrift unterscheiden:

- ***Open-Loop***:

$$u(t) = \omega(t, t_0, y(t_0))$$

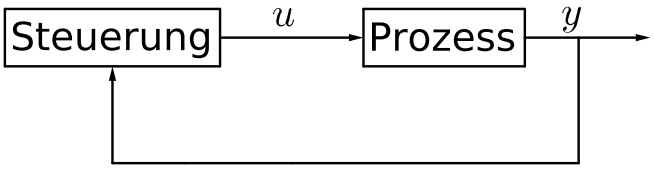
  In diesem Falle hängt die Kontrollfunktion neben der Zeit nur von den Anfangsbedingungen ab. Dies entspricht einer Situation, in der die Kontrolle nach dem Zeitpunkt t_0 keine Rückmeldung über den Zustand erhält, eine Anpassung der Kontrolle ist in diesem Falle nicht mehr möglich.



- ***Closed-Loop***:

$$u(t) = \omega(t, y(t))$$

In diesem Falle hängt der Wert der Kontrollfunktion vom jeweils aktuellen Zustand ab. Dies entspricht einer Situation, in der durch ständige Rückmeldungen über den Zustand die Kontrolle nachjustiert werden kann.



Daneben existiert das Konzept des reduzierten Gradienten auch für endlich-dimensionale Optimierungsprobleme. Hat man ein Problem der Form

$$\begin{cases} \min f(x) \\ \text{u. d. Nb. } e(x) = 0 \end{cases}$$

gegeben, wobei x aus $\mathbb{R}^n$ und die zulässige Menge $Z = \{x \in V | e(x) = 0\}$ ist, so kann für $n > 1$ x auch als Tupel (y, u) mit $y \in \mathbb{R}^m$ und $u \in \mathbb{R}^{n-m}$ für ein $m < n$ aufgefasst werden:

$$\begin{cases} \min f(y, u) \\ \text{u. d. Nb. } e(y, u) = 0 \end{cases} \tag{3.4.29}$$

Erfüllt e nun beispielsweise die Voraussetzungen des Satzes über implizite Funktionen, so kann man y nach u auflösen.

Beispiel 3.4.21:

Sei $n = 3$ und $m = 2$, d. h. $x \in \mathbb{R}^3$, $y \in \mathbb{R}^2$ und $u \in \mathbb{R}$. Die Nebenbedingung sei gegeben durch

$$\begin{aligned} a_1x_1 + a_2x_2 + a_3x_3 &= b_1 \\ a_4x_1 + a_5x_2 + a_6x_3 &= b_2 \end{aligned},$$

mit $x = (y, u)$ also

$$\begin{aligned} a_1y_1 + a_2y_2 + a_3u &= b_1 \\ a_4y_1 + a_5y_2 + a_6u &= b_2 \end{aligned}$$

Hieraus ergibt sich

$$\begin{aligned} a_1y_1 + a_2y_2 &= b_1 - a_3u \\ a_4y_1 + a_5y_2 &= b_2 - a_6u \end{aligned}.$$

Existiert die Inverse von $\begin{pmatrix} a_1 & a_2 \\ a_4 & a_5 \end{pmatrix}$, gilt folglich

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_4 & a_5 \end{bmatrix}^{-1} = \begin{pmatrix} b_1 - a_3u \\ b_2 - a_6u \end{pmatrix},$$

$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ kann also als Funktion von u geschrieben werden.

In diesem Falle genügt es, das reduzierte Optimierungsproblem

$$\min \hat{f}(u)$$

mit reduziertem Funktional $\hat{f}(u) \coloneqq f(\varphi(u), u)$ zu betrachten. Der Satz über implizite Funktionen sagt jedoch nicht nur etwas über die lokale Auflösbarkeit der Funktion y nach u in einer Umgebung eines Punktes (y_0, u_0) aus, sondern auch über die Differenzierbarkeit der Abbildung $u \mapsto y$. Es gilt:

$$\varphi_u(u) = -e_y\left(\varphi(u), u\right)^{-1} e_u\left(\varphi(u), u\right)$$

Aus der Optimalitätsbedingung 1. Ordnung, $\nabla_u \hat{f}\left(u\right) = 0$, ergibt sich

$$\begin{aligned} 0 = \hat{f}_u\left(u\right) &= f_u\left(\varphi\left(u\right), u\right) = f_y\left(\varphi\left(u\right), u\right) \cdot \varphi_u\left(u\right) + f_u\left(\varphi\left(u\right), u\right) \\ &= -f_y\left(\varphi\left(u\right), u\right) \cdot \left[e_y\left(\varphi\left(u\right), u\right)\right]^{-1} \cdot e_u\left(\varphi\left(u\right), u\right) + f_u\left(\varphi\left(u\right), u\right) \end{aligned}$$

Setzt man

$$\begin{aligned} \lambda &= -\left[f_y\left(\varphi\left(u\right), u\right)\left[e_y\left(\varphi\left(u\right), u\right)\right]^{-1}\right]^T \\ &= -\left[\nabla_y e\left(\varphi\left(u\right), u\right)\right]^{-T} \nabla_y f\left(\varphi\left(u\right), u\right)^T, \end{aligned}$$

so folgt

$$0 = \hat{f}_u\left(u\right) = \lambda^T e_u\left(\varphi\left(u\right), u\right) + f_u\left(\varphi\left(u\right), u\right).$$

Dies entspricht der oben eingeführten adjungierten Gleichung. Das entsprechende Optimalitätssystem lautet dann:

- Zustandsgleichung: $e\left(y, u\right) = 0$
- Adjungierte Gleichung: $\nabla_y e\left(y, u\right) \lambda = -\nabla_y f\left(y, u\right)$
- Optimalitätsbedingung: $\lambda^T e_u\left(\varphi\left(u\right), u\right) + f_u\left(\varphi\left(u\right), u\right) = 0$

Gilt hierbei

$$\nabla_u \hat{f}\left(u\right) = 0$$

für ein u, so existiert eine Lösung zu

$$\begin{cases} e\left(y, u\right) = 0 \\ \left(\nabla_y e\left(y, u\right)\right)^T \lambda = -\left(\nabla_y f\left(y, u\right)\right)^T \\ \nabla_u f\left(y, u\right) - \lambda^T \cdot \nabla_u e\left(y, u\right) = 0 \end{cases}$$

Umgekehrt gilt für eine Lösung (y, u, λ) von letzterem $\nabla_u \hat{f}(u) = 0$. Analog wie den vorherigen Kapitel lautet die Lagrangefunktion:

$$\mathcal{L}\left(y, u, \lambda\right) = f\left(y, u\right) + \lambda^T e\left(y, u\right)$$

Für diese gilt:

$$\begin{aligned} \nabla_\lambda \mathcal{L}\left(y, u, \lambda\right) &= e\left(y, u\right) \\ \nabla_y \mathcal{L}\left(y, u, \lambda\right) &= \nabla_y f\left(y, u\right) + \nabla_y e\left(y, u\right) \lambda, \\ \nabla_u \mathcal{L}\left(y, u, \lambda\right) &= \nabla_u f\left(y, u\right) + \nabla_u e\left(y, u\right) \lambda \end{aligned}$$

womit sich das Optimalitätssystem erneut in der kompakten Form

$$\nabla \mathcal{L}(y, u, \lambda) = 0$$

angeben lässt. Sind f und e C^2-Funktionen, so muss in einer lokalen Minimalstelle, die regulärer Punkt ist, gemäß Satz 3.1.30 die Hessematrix $\nabla^2_{xx} \mathcal{L}(x)$ positiv definit auf der Tangentialebene von h in ebendiesem Punkt sein. Hier ist nun $h = e$, die Tangentialebene gegeben durch

$$M = \left\{(\delta y, \delta u) \,|\nabla_y e(y, u)^T \delta y + \nabla_u e(y - u)^T \delta u = 0\right\}$$

und die entsprechende Hessematrix ist

$$\nabla^2_{xx}\mathcal{L} = \begin{bmatrix} \nabla^2_{yy}\mathcal{L} & \nabla^2_{yu}\mathcal{L} \\ \nabla^2_{uy}\mathcal{L} & \nabla^2_{uu}\mathcal{L} \end{bmatrix}.$$

Die Optimalitätsbedingung zweiter Ordnung ist folglich

$$\begin{pmatrix} \delta y^T & \delta u^T \end{pmatrix} \begin{bmatrix} \nabla^2_{yy}\mathcal{L} & \nabla^2_{yu}\mathcal{L} \\ \nabla^2_{uy}\mathcal{L} & \nabla^2_{uu}\mathcal{L} \end{bmatrix} \begin{pmatrix} \delta y \\ \delta u \end{pmatrix} \geq 0 \,\forall\, (y,u) \in M$$

Aus der linearisierten Gleichungsnebenbedingung $\nabla_y e(y,u)^T \delta y + \nabla_u e(y,u)^T \delta u = 0$ folgt wegen der Invertierbarkeit von $\nabla_y e(y,u)^T = e_y(y,u)$ die Gleichung

$$\delta y = -\left(\nabla_y e\left(y,u\right)\right)^{-T} \left(\nabla_u e\left(y,u\right)\right)^T \delta u$$

und somit

$$\begin{pmatrix} \delta y \\ \delta u \end{pmatrix} = \begin{bmatrix} -\left(\nabla_y e\left(y,u\right)\right)^{-T} \left(\nabla_u e\left(y,u\right)\right)^T \\ I \end{bmatrix} \delta u.$$

Folglich gilt

$$\begin{pmatrix} \delta y & \delta u \end{pmatrix} \begin{bmatrix} \nabla^2_{yy}\mathcal{L} & \nabla^2_{yu}\mathcal{L} \\ \nabla^2_{uy}\mathcal{L}L & \nabla^2_{uu}\mathcal{L} \end{bmatrix} \begin{pmatrix} \delta y \\ \delta u \end{pmatrix} =$$

$$\delta u^T \left(\begin{bmatrix} -\left(\nabla_u e\left(y,u\right)\right) \cdot \left(\nabla_y e\left(y,u\right)\right)^{-1} & I \end{bmatrix} \begin{bmatrix} \nabla^2_{yy}\mathcal{L} & \nabla^2_{yu}\mathcal{L} \\ \nabla^2_{uy}\mathcal{L} & \nabla^2_{uu}\mathcal{L} \end{bmatrix} \begin{bmatrix} -\left(\nabla_y e\left(y,u\right)\right)^{-T} \left(\nabla_u e\left(y,u\right)\right)^T \\ I \end{bmatrix} \right) \delta u$$

Die Matrix

$$\begin{bmatrix} -\left(\nabla_u e\left(y,u\right)\right) \cdot \left(\nabla_y e\left(y,u\right)\right)^{-1} & I \end{bmatrix} \begin{bmatrix} \nabla^2_{yy}\mathcal{L} & \nabla^2_{yu}\mathcal{L} \\ \nabla^2_{uy}\mathcal{L} & \nabla^2_{uu}\mathcal{L} \end{bmatrix} \begin{bmatrix} -\left(\nabla_y e\left(y,u\right)\right)^{-T} \left(\nabla_u e\left(y,u\right)\right)^T \\ I \end{bmatrix}$$

heißt ***reduzierte Hessematrix***.

4 Inverse Probleme mit gewöhnlichen Differentialgleichungen

In diesem kurzem Kapitel sollen die Formulierung und Lösung inverser Probleme mit gewöhnlichen Differentialgleichungen etwas genauer untersucht werden. Zu diesem Zwecke soll folgendes Cauchy-Problem mit linearer Differentialgleichung betrachtet werden:

$$\begin{cases} y'(x) = p(x)y(x) + q(x), & x \in [a,b] \\ y(a) = 0 \end{cases} \tag{4.0.1}$$

p und q seien hierbei stetig im Intervall $I = [a,b]$.

Wie aus Kapitel 1.1 bekannt, ist die Lösung von (4.0.1) durch

$$y(x) = e^{\int_a^x p(x)\mathrm{d}x} \cdot \int_a^x q(s) e^{-\int_a^s p(t)\mathrm{d}t} \mathrm{d}s, \quad x \in [a,b] \tag{4.0.2}$$

gegeben. Die Lösung (4.0.2) kann hierbei als Operator $\tilde{K}$ betrachten, welcher die problemspezifischen Daten einschließlich $p, q \in C(I)$ auf die entsprechende Funktion $y \in C^1(I)$ abbildet. $\tilde{K}$ ist hierbei linear in q, jedoch **nicht** in p. Die Anwendung des Operators auf gegebene Daten, um den Zustand y des Systems zu erhalten, wird als ***direktes Problem*** bezeichnet. Unter einem ***inversen Problem*** ist in gewisser Weise die Umkehrung des Ganzen zu verstehen: Ausgehend von dem (durch Beobachung, Messung oderÄhnlichem) bekannten Zustand y sollen die problemspezifischen Daten p und q ermittelt werden.

Im Folgenden soll der Fall betrachtet werden, in dem p gegeben ist. Der betrachtete, vereinfachte Operator sei K genannt, er ist durch $Kq \coloneqq \tilde{K}(p,q)$ gegeben und linear. Gemäß dem Satz von Arcela-Ascoli ist K ein kompakter Operator auf einem unendlich-dimensionalem Raum (in diesem Falle $C(I)$). Dies hat schwerwiegende Konsequenzen: Wäre im Falle, dass K eineindeutig ist, die inverse K^{-1} stetig (und somit beschränkt), wäre auch die Verkettung $K^{-1}K = I$ kompakt. Die Identität ist für unendlich-dimensionale Räume jedoch nie kompakt. Folglich kann K^{-1} nicht stetig sein. Diese Unstetigkeit der Inversen stellt die Hauptschwierigkeit bei der Lösung inverser Probleme dar: Bei Stetigkeit hätte man eine Art „Garantie“, dass man fast die exakte Lösung hat, wenn die in der Regel vorliegende Störung der Daten hinreichend klein ist. Bei Unstetigkeit sind Sprünge möglich, selbst bei kleinsten Störungen können die zu den gestörten Daten gehörigen Parameter weit entfernt von den eigentlichen liegen.

Zur Veranschaulichung dieser Ergebnisse sei das einfachste Cauchy-Problem,

$$\begin{cases} y'(x) = q(x) \\ y(a) = 0 \end{cases} \tag{4.0.3}$$

mit $q \in C(\mathbb{R})$ und $q(x) = 0$, $\forall x \in \mathbb{R} \setminus [a,b]$, gegeben. Die Lösung von (4.0.3) ist

$$y(x) = \int_a^x q(s)\mathrm{d}s, \qquad x \in [a,b]\,. \tag{4.0.4}$$

Eine alternative Darstellung des Problems $Kq = y$ ist mithilfe der sogenannten ***Heaviside-Funktion*** möglich. Diese ist durch $H(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$ definiert, die Lösung (4.0.4) kann somit durch

$$y(x) = \int_{\mathbb{R}} H(x-s) q(s) \mathrm{d}s, \quad x \in [a,b] \tag{4.0.5}$$

dargestellt werden. Nun soll die Lösung dieses inversen Problems untersucht werden. Die inverse Abbildung K^{-1} erhält man durch Differentiation von (4.0.5) (oder (4.0.4)). Da q stetig ist, gilt

$$q = K^{-1}y = y'$$

Es soll gezeigt werden, dass Inverse K^{-1} der linearen Abbildung $K : C(I) \to C(I)$ tatsächlich unstetig ist, wobei $C(I)$ mit der Norm $\|f\|_C \coloneqq \max_{x\in[a,b]} |f(x)|$ ist. ($(C(I), \|\cdot\|_C)$ ist Banachraum.) Offensichtlich ist K stetig: Es gilt $\left|\int_a^x f(x)\mathrm{d}x\right| \leq \int_a^x \|f\|_C \,\mathrm{d}s \leq (b-a)\,\|f\|_C$, somit existiert eine Konstante $M > 0$ mit $\|Kf\|_C \leq M\,\|f\|_C$ für alle $f \in C(I)$. (In unendlich dimensionalen Räumen ist eine lineare Abbildung genau dann stetig, wenn sie auf derartige Weise beschränkt ist.) Um zu zeigen, dass K^{-1} nicht stetig ist, sei die durch

$$y_k(x) = \frac{1}{\sqrt{k}} \sin(kx)$$

definierte Funktionenfolge (y_k) gegeben. Für die Ableitung von y_k im Punkt x folgt somit $y_k'(x) = \sqrt{k}\cos(kx)$. Somit gilt $\|y\|_C \leq \frac{1}{\sqrt{k}}$. Wählt man nun $k_0 \in \mathbb{N}$ so, dass $k_0 \geq \frac{2\pi}{b-a}$ gilt, so nehmen aufgrund der Periodizität $\sin(kx)$ und $\cos(kx)$ jeden Wert aus $[-1, 1]$ mindestens einmal an, es folgt

$$\|y_k\|_C = \frac{1}{\sqrt{k}} \qquad \text{sowie} \qquad \|y_k'\|_C = \sqrt{k}$$

für $k \geq k_0$. Hieraus folgt $\|y_k\|_C \to 0$ und $\|y_k'\|_C \to \infty$ für $k \to \infty$. Für die Stetigkeit von K^{-1} wäre jedoch die Existenz einer Konstanten $\bar{M} > 0$ notwendig, sodass $\|K^{-1}f\|_C \leq \bar{M}\,\|f\|_C$ für alle $f \in C(I)$, insbesondere für die y_k, gilt. Da dies nicht möglich ist, ist K^{-1} unstetig.

Bemerkung: *Die Unstetigkeit bleibt auch im Falle des Wechsels zur L^2-Norm, gegeben durch*

$$\|f\|_{L^2(a,b)} = \left(\int_a^b |f(x)|^2 \,\mathrm{d}x\right)^{\frac{1}{2}},$$

erhalten. Betrachtet man hingegen K als Operator $C^1(I) \to C(I)$ mit Banachraum $(C^1(I), \|\cdot\|_{C^1})$, $\|f\|_{C^1} \coloneqq \max_{x\in I} |f(x)| + \max_{x\in I} |f'(x)|$, so ist K^{-1} durchaus stetig.

Nun soll der Fall betrachtet werden, bei dem q bekannt (im Folgenden exemplarisch $q \equiv 0$) und p die zu bestimmende Funktion ist. (Auch bekannt als Parameteridentifikationsproblem.) Hierzu sei das Anfangswertproblem

$$\begin{cases} y'(x) = p(x)y(x) \\ y(a) = y_0 \end{cases}$$

gegeben. Die aus Kapitel 1.1 bekannte Lösung vereinfacht sich in diesem Falle auf

$$y(x) = y_0 \cdot e^{\int_a^x p(t)\mathrm{d}t}$$

Ist $p(x) = p_0$ konstant, wird dies zu $y(x) = y_0 \cdot e^{p_0(x-a)}$, $\quad x \in [a, b]$. Angenommen, man habe eine Messung von y zu einem Punkt $x_0 \in (a, b)$ mit Wert $\bar{y} = y(x_0)$. Hieraus ergibt sich

$$p_0 = \frac{1}{x_0 - a} \log\left(\frac{\bar{y}}{y_0}\right)$$

Folglich können kleine Störungen von $\bar{y}$ (z. B. durch Messfehler) große Änderungen von p_0 herbeiführen, insbesondere wenn x_0 nahe bei a ist.

Nun sei hingegen angenommen, dass p nicht konstant ist, jedoch die Form $p(x) = p_0(x) - \varphi(x), \quad x \in [a,b]$ hat, wobei $\operatorname{supp}\varphi \subset (a_1,b_1) \subset [a,b]$ und $\int_a^b \varphi(x)\mathrm{d}x = 0$ gilt. Die Lösung des entsprechenden Cauchy-Problems ist durch

$$y(x) = y_0 e^{\int_a^x (p_0(t)+\varphi(t))\,\mathrm{d}t}, \quad x \in [a,b]$$

gegeben. Aus den Voraussetzungen geht hervor, dass für $a \leq x \leq a_1$ oder $b_1 \leq x \leq b$ die Lösung gleich $y_0 e^{\int_a^x p_0(t)\mathrm{d}t}$ ist. Folglich kann ϕ nicht durch Messungen in diesen Intervallen bestimmt werden.

Betrachtet werden soll das Randwertproblem

$$\begin{cases} -y''(x) = q(x), \quad x \in (0,1) \\ \quad y(0) = 0, \quad y(1) = 0 \end{cases} \tag{4.0.6}$$

mit $q \in C([0,1])$.

Das inverse Problem besteht in der Bestimmung von q durch Messungen von y (***inverse source problem***). Es hat jedoch möglicherweise eine nicht eindeutige Lösung. Wählt man ein beliebiges $\varphi \in C_0^\infty(0,1)$ mit $\operatorname{supp}\varphi \subset (a,b) \subset [0,1]$ und definiert $\tilde{q}(x) = -\varphi''(x)$, dann zeigt Nachrechnen, dass

$$\int_0^1 G(x,z)\tilde{q}(z)\mathrm{d}z = \varphi(x)$$

mit Greenscher Funktion G gilt. Folglich ist das Integral 0 außerhalb des Trägers von φ, also in $[0,1] \setminus (a,b)$, für jedes φ mit $\operatorname{supp}\varphi \subset (a,b)$. Dies zeigt die Nichteindeutigkeit des zu (4.0.6) gehörigen Problems.

Wie dargelegt, kann ein inverses Problem also eine eindeutige Lösung haben oder nicht, abhängig von der Teilmenge von beispielsweise $\mathbb{R}^d$, auf der der Zustand des Systems gemessen wird. Ebenfalls ersichtlich wurde, dass die Inverse von K (u. U.) stetig wird, wenn man die Menge, in der die Lösung gesucht wird, einschränkt.

Satz 4.0.1:
Seien X, Y normierte lineare Räume. Wenn $K : D(K) \to Y$ $(D(K) \subseteq X)$ ein stetiger, bijektiver Operator und $C \subseteq D(K)$ eine kompakte Menge. Dann ist die Inverse der Einschränkung auf diese Menge, $(K|_C)^{-1}$ stetig invertierbar.

Beweis: K ist stetig, also ist $K(C)$ kompakt. Für eine in C offene Teilmenge $A \subseteq C$ ist das Komplement $A^C = C \setminus A$ abgeschlossen, folglich ebenfalls kompakt. Damit ist auch $K(A^C)$ kompakt, also abgeschlossen. Da K ist bijektiv ist, ist dies insbesondere die Einschränkung auf C, womit $K(A^C) = K(C) \setminus K(A)$ gilt. Somit ist $K(A)$ als Komplement einer abgeschlossenen Menge offen. Aufgrund der Beliebigkeit der (in C) offenen Teilmenge sind die Urbilder offener Mengen der Abbildung $(K|_C)^{-1}$ offen. Dies impliziert Stetigkeit. □

Seien nun $K : H_1 \to H_2$ ein beschränkter, linearer Operator, H_1, H_2 reelle Hilberträume. Betrachtet man nun die Abbildung

$$Kq = y, \tag{4.0.7}$$

so existiert eine Lösung, sofern $y \in \mathcal{R}(K)$ gilt. Jedoch ist $\mathcal{R}(K)$ im Allgemeinen ein Unterraum von H_2, welcher H_2 nicht vollkommen ausfüllt. Um diese Einschränkung zu überwinden, erweitert man die Klasse der Funktionen y, für die eine Art verallgemeinerte Lösung existiert, auf einen dichten Teilraum von H_2. Dies kann durch die sogenannte ***Methode der kleinsten Quadrate*** (engl. least-squares) erreicht werden. Eine Funktion $q \in H_1$ ist eine Lösung der

kleinsten Quadrate zu (4.0.7), wenn sie der Gleichung

$$\|Kq - y\| = \inf \{\|Ku - y\| : u \in H_1\} \tag{4.0.8}$$

genügt. Eine solche Lösung existiert, wenn $y \in \mathcal{R}(K) + \mathcal{R}(K)^{\perp}$ gilt, was dicht in H_2 liegt. Bezeichnet man mit P die orthogonale Projektion von H_2 auf $\mathcal{R}(K)$, dann genügt die Lösung q von (4.0.8) $Kq = Py$ sowie

$$Kq - y \in \mathcal{R}(K)^{\perp} \tag{4.0.9}$$

Der Hilbertraum-adjungierte Operator von K, $K^* : H_2 \to H_1$, ist mittels

$$\langle Kf, g\rangle_{H_2} = \langle f, K^* g\rangle_{H_1}, \qquad \forall f \in H_1,\, g \in H_2$$

definiert. Für diesen kann gezeigt werden, dass der Nullraum $\mathcal{N}(K^*)$ gleich dem orthogonalen Komplement des Bildes von K, $\mathcal{R}(K)^{\perp}$, ist. Aus (4.0.9) folgt nun wegen $\langle Kq - y, Kf\rangle = 0 \forall Kf \in \mathcal{R}(K)$

$$K^* Kq = K^* y.$$

Aus dieser Gleichung, auch ***(Gaußsche) Normalengleichung*** genannt, wird ersichtlich, dass eine eindeutige Lösung der Methode der kleinsten Quadrate genau dann existiert, wenn

$$\mathcal{N}(K^* K) = \{0\}$$

gilt. Ist dies nicht erfüllt, wählt man im allgemeinen die (eindeutige) Lösung von (4.0.8) mit kleinster Norm als verallgemeinerte Lösung von (4.0.7). Dies führt auf die Definition der ***Moore-Penrose-Inversen*** beziehungsweise ***verallgemeinerten Inversen*** von K, kurz mit $K^{\dagger}$ bezeichnet. Da der selbstadjungierte, kompakte Operator K^*K nichtnegative Eigenwerte hat, hat der Operator

$$K^* K + \alpha I, \qquad \alpha > 0$$

(I Identität auf H_1) strikt positive Eigenwerte und eine beschränkte Inverse. Das Problem

$$(K^* K + \alpha I)\, q_\alpha = K^* y$$

ist daher korrekt gestellt, seine Lösung durch

$$q_\alpha = (K^* K + \alpha I)^{-1} K^* y \tag{4.0.10}$$

gegeben. q_α heißt ***Thikonov-Approximation*** von $K^{\dagger} y$. Für diese lässt sich

$$\lim_{\alpha \to 0} \left\| q_\alpha - K^{\dagger} y \right\| = 0$$

zeigen.

In der Praxis liefern Messungen von y keine exakten, sondern gestörte Daten y^δ, wobei man im günstigen Falle einen Fehlerschätzer der Form

$$\left\| y - y^\delta \right\| \le \delta$$

hat. Ferner erhält man die zugehörige Thikonov-Approximation

$$q_\alpha^\delta = (K^* K + \alpha I)^{-1} K^* y^\delta.$$

Unter Verwendung der Abschätzungen $\left\| K K^* (K^* K + \alpha I)^{-1} \right\| \le 1$ und $\left\| (K^* K + \alpha I)^{-1} \right\| \le \frac{1}{\alpha}$ ergibt sich

$$\left\| q_\alpha^\delta - q_\alpha \right\| \le \frac{\delta}{\sqrt{\alpha}}.$$

Diese Abschätzung zeigt, dass der Regularisierungsparameter α in Abhängigkeit vom Störpegel δ gewählt werden sollte, das heißt $\alpha = \alpha(\delta)$. Die Forderung $q^\delta_{\alpha(\delta)} \to K^\dagger y$ für $\delta \to 0$ führt ferner auf die Bedingung

$$\frac{\delta^2}{\alpha(\delta)} \to 0, \qquad \delta \to 0.$$

Eine Wahl der Form $\alpha = c\delta^\beta$ mit $0 < \beta < 2, c > 0$ ist daher geeignet.
Die Thikonov-Regularisierung hat eine sehr wichtige Interpretation im Sinne der Variationsrechnung: Betrachtet man das Funktional

$$J_\alpha(q) = \left\|Kq - y^\delta\right\|^2 + \alpha \|q\|^2, \tag{4.0.11}$$

so ist dieses konvex und differenzierbar, der zugehörige Gradient lautet

$$\nabla_q J_\alpha(q) = (K^*K + \alpha I)\, q - K^* y^\delta.$$

Der eindeutige Minimierer von (4.0.11), charakterisiert durch $\nabla_q J_\alpha(q) = 0$, ist folglich

$$q^\delta_\alpha = (K^*K + \alpha I)^{-1} K^* y^\delta,$$

dies ist die Thikonov-Approximation aus (4.0.10).

Das Problem $\min\limits_q J_\alpha(q)$ mit J_α aus (4.0.11) kann auch äquivalent in der Form

$$\begin{cases} \min J_\alpha(y,q), \quad J_\alpha(y,q) \coloneqq \left\|y - y^\delta\right\|^2 + \alpha \|q\|^2 \\ \quad \text{u. d. Nb. } y = Kq \end{cases} \tag{4.0.12}$$

angegeben werden. (Die Nebenbedingung $Kq = y$ ist ferner äquivalent zu der Forderung, dass die Lösung (y, q) die K entsprechende Differentialgleichung löst.)

Die Formulierung in (4.0.12) wird auch im Falle eines nichtlinearen Operators K angewandt. Dieses sogenannte ***regularisierte Least-Squares-Problem*** (englisch auch ***Penalized Least Squares Problem***) ist wie folgt formuliert:

$$\min \left\|K(q) - y^\delta\right\|^2 + \alpha \|q\|^2$$

Alternativ kann - vorausgesetzt, K ist Fréchet-differenzierbar - auch eine Startnäherung q_0 sowie die entsprechende Linearisierung

$$K(q_0 + \delta q) = K(q_0) + \partial K(q_0)\delta q + r(q_0, \delta q)$$

mit $\|r(q_0, \delta q)\| = o(\|\delta q\|)$ betrachtet werden. Sei nun $K(q_0) = y_0$ und $K(q_0 + \delta q) = y^\delta$. Dies führt auf die lineare Gleichung

$$\partial K(q_0)\delta q = y^\delta - y_0.$$

Diese Gleichung ist im Allgemeinen schlecht gestellt, folglich ist die Tikhonov-Regularisierung zur Lösung des Problems ratsam. Dies führt auf

$$(\partial K(q_0)^* \partial K(q_0) + \alpha I)\, \delta q_\alpha = \partial K(q_0)^* \left(y^\delta - y_0\right)$$

Die Lösung δq_α wird zur Aktualisierung von q verwendet, das heißt eine bessere Approximation mittels $q_1 = q_0 + \delta q_\alpha$ berechnet. Dies ist das ***Levenberg-Marquardt-Verfahren*** und die Linearisierungsstrategie heißt ***output least squares***.

Bemerkung: *Der obige „Algorithmus“ arbeitet formal mit ganzen Funktionen als Eingabewerte. In der Praxis muss man diese jedoch diskretisieren, was naturgemäß zu Abweichungen führt, vor allem bei der numerischen Lösung der Differentialgleichung, je nachdem, welche Methode*

verwendet wird. Eine mögliche Implementierung des Levenberg-Marquardt-Verfahrens befindet sich in Anhang (D)*. Dort ist auch ein Testskript für das Problem*

$$y'(x) = e^{px}$$
$$y(0) = 1$$

mit $p = -\frac{1}{2}$ *zu finden. Wie gehabt gilt es, dieses* p *aus gestörten Daten zu rekonstruieren. Die Störung der Daten wird durch die Multiplikation der exakten Lösung mit* $(1+r)$ *erreicht, wobei* r *eine normalverteilte Zufallsvariable mit Erwartungswert* 0 *und Varianz gemäß dem angegebenen Störlevel* noise *ist.*

Für einen abschließenden Vergleich der in diesem Kapitel vorgestellten Konzepte sei erneut das einfachste Cauchy-Problem

$$\begin{cases} y'(x) = q(x), & x \in (0,1) \\ y(0) = 0 \end{cases}, \tag{4.0.13}$$

gegeben, wobei q eine auf $[0,1]$ stetige Funktion mit $q(x) = 0$ für $x \in \mathbb{R} \backslash [0,1]$ sei. Wie bereits erwähnt, kann die Lösung zu (4.0.13) in der Form $y = Kq$ mittels

$$y(x) = \int_0^x q(s)\mathrm{d}s, \qquad x \in (0,1)$$

oder äquivalent

$$y(x) = \int_{\mathbb{R}} H(x-s)q(s)\mathrm{d}s, \qquad x \in (0,1) \tag{4.0.14}$$

geschrieben werden. (4.0.14) lässt sich wie folgt in eine allgemeine Form bringen: Sei hierzu der Operator $K : H_1 \to H_2$, welcher auf den Hilberträumen $H_1 = H_2 = L^2(0,1)$ operiert, mittels

$$(Kq)(x) = \int_0^1 k(x,t)q(t)\mathrm{d}t$$

mit $k \in L^2((0,1) \times (0,1))$ definiert. (Ein Operator dieser Gestalt wird auch als *Fredholmscher Integraloperator* bezeichnet.) Der Hilbertraum-adjungierte Operator ist dann mittels

$$(K^*g)(t) = \int_0^1 k(x,t)g(x)\mathrm{d}x, \quad g \in C([0,1])$$

gegeben. Im speziellen Fall (4.0.14) lautet dieser folglich

$$(K^*g)(s) \coloneqq \int_{\mathbb{R}} H(x-s)g(x)\mathrm{d}x, \quad g(x) = 0, x \in \mathbb{R} \backslash (0,1)$$

Dies ist wiederum äquivalent zu

$$(K^*g)(s) = \int_s^1 g(x)\mathrm{d}x, \qquad s \in (0,1).$$

Ferner gilt

$$\mathcal{R}(K) = \left\{ y \in L^2(0,1),\, y \in A \subset [0,1],\, y' \in L^2(0,1),\, y(0) = 0 \right\}$$
$$\mathcal{R}(K^*) = \left\{ y \in L^2(0,1),\, y \in A \subset [0,1],\, y' \in L^2(0,1),\, y(1) = 0 \right\}$$

Wie bereits gezeigt, ist K^{-1} unstetig in $L^2(0,1)$ und es gilt $K^{-1}y = y'$. Gleicherweise gilt $(K^*)\,K^{-1}y = -y'$.
Nun soll die Tikhonov-Approximation von q betrachtet werden, welche durch die Lösung von

$$(K^*K + \alpha I)\, q_\alpha = K^*y \tag{4.0.15}$$

gegeben ist. Wegen der Definition von $\mathcal{R}(K^*)$ muss $q_\alpha(1) = 0$ sein. $(K^*)^{-1}$ von links angewandt auf (4.0.15) ergibt

$$Kq_\alpha + \alpha (K^*)^{-1} q_\alpha = y, \tag{4.0.16}$$

also

$$Kq_\alpha - \alpha q'_\alpha = y.$$

Hiermit folgt aus der Gestalt von $\mathcal{R}(K)$ und der Anfangsbedingung für y die Gleichung $q'_\alpha(0) = 0$. Wendet man nun die Inverse von K von links auf (4.0.16) an, erhält man

$$q_\alpha(x) - \alpha q''_\alpha(x) = y'(x). \tag{4.0.17}$$

Die Lösung dieses Randwertproblems mit Randbedingungen $q_\alpha(1) = 0, q'_\alpha(0) = 0$ stellt die Thikonov-Approximation von (4.0.13) dar, vorausgesetzt, y ist gegeben.
Das selbe Ergebnis erhält man auch mit dem output least squares Ansatz: Sei $\bar{y}$ das Ergebnis der Messung von y. Das Problem lautet

$$\begin{cases} \min\limits_{\alpha} \|y - \bar{y}\|^2_{L^2(0,1)} + \alpha \|q\|^2_{L^2(0,1)} \\ \text{u. d. N.} \quad y'(x) = q(x), \quad y(0) = 0 \end{cases}$$

Das Optimalitätssystem, welches die zugehörige Lösung charakterisiert, ist durch

$$\begin{cases} y'(x) = q(x), & y(0) = 0 \\ -\lambda'(x) = -2 (y(x) - \bar{y}(x)), & \lambda(1) = 0 \\ 2\alpha q(x) - \lambda(x) = 0 \end{cases}$$

gegeben. Substituiert man nun $\lambda(x)$ in der adjungierten Gleichung durch $2\alpha q(x)$ und differenziert anschließend, führt dies auf

$$-\alpha q''(x) = -(y'(x) - \bar{y}'(x)).$$

Die Ersetzung von $y'(x)$ durch $q(x)$ liefert schließlich

$$q(x) - \alpha q''(x) = \bar{y}(x),$$

was (4.0.17) entspricht.

5 Einführung in die Stochastischen Differentialgleichungen

Dieses Kapitel soll einen kurzen Einblick in stochastische Differentialgleichungen (kurz: SDE) gewähren. Der Einfachheit halber wird hierbei jedoch auf eine detaillierte Ausarbeitung der theoretischen Grundlagen verzichtet und, der diesem Kapitel zugrundeliegenden Primärquelle [31] folgend, vor allem die numerische Behandlung solcher Gleichungen betrachtet.
Zum Einsteig sei folgendes Beispiel gegeben:

Beispiel 5.0.1 (random walk):

Betrachtet wird ein betrunkener Mann, welcher nach einer durchzechten Nacht in einer Weinstube ebendiese verlässt und nach Hause geht. Infolge der Trunkenheit schwankt der Mann mit jedem Schritt nach vorne zeitgleich nach links oder rechts. Dieses Schwanken soll nun modelliert werden. Hierzu sei Δt die für einen Schritt benötigte Zeiteinheit sei. Ferner sei $\Delta x > 0$ die Größe einer Schwankbewegung. Schwankt der Mann nach links, so entspricht dies einer Ortsänderung von $-\Delta x$, schwankt er nach rechts, ist die Ortsänderung $+\Delta x$.Ähnlich wie bei gewöhnlichen Differentialgleichungen (vgl. Kapitel 1) besteht hier die Aufgabe darin, die Ortskurve aus Informationen über die Ortsänderung zu rekonstruieren, jedoch mit einem gravierenden Unterschied: War bei gewöhnlichen Differentialgleichungen die Ortsänderung $\dot{x}$ allein durch den aktuellen Ort x und/oder die aktuelle Zeit t bestimmt, hängt dieseÄnderung hier aufgrund der durch die Trunkenheit fehlenden Kontrolle vom Zufall ab.
Die Wahrscheinlichkeit für

- einen Schlenker nach rechts sei $p \geq 0$.
- einen Schlenker nach links sei $q \geq 0$.

Die beiden Wahrscheinlichkeiten müssen hierbei $p+q \leq 1$ erfüllen. (Ein Modell mit $p+q < 1$ beinhaltet die Möglichkeit, dass der Mann einen „geraden“ Schritt nach vorne schafft.) Für $j = 0, 1, ...$ sei s_j der Ort des Mannes zum Zeitpunkt $t_j = j\Delta t$ mit $s_0 = 0$. Der Ort zum Zeitpunkt t_{j+1} ist durch den Ort s_j sowie der Ortsänderung zu diesem Zeitpunkt z_{j+1} mittels

$$s_{j+1} = s_j + z_{j+1}$$

bestimmt. Induktiv ergibt sich:

$$s_n = s_0 + \sum_{j=1}^{n} z_j = \sum_{j=1}^{n} z_j$$

DieÄnderung z_j ist, wie beschrieben, eine Zufallsvariable, es gilt

$$z_j \in \{+\Delta x, -\Delta x\}$$

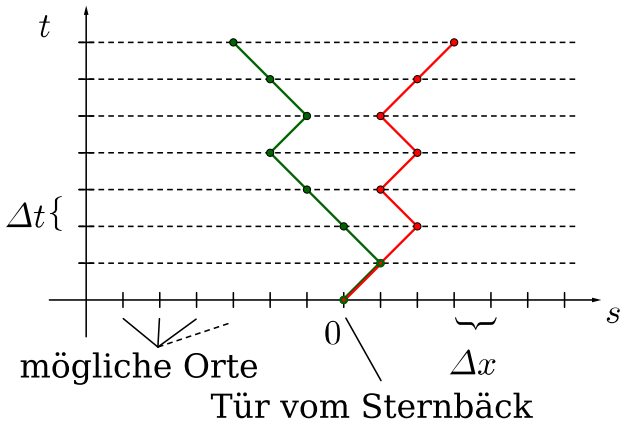
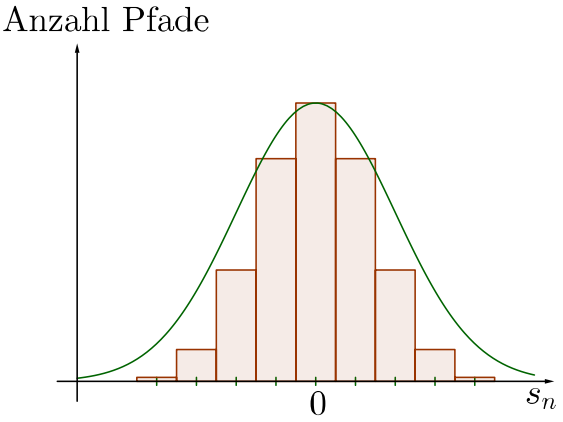


Abbildung linkes: Zwei mögliche Wege des torkelnden Mannes. Möchte man nur Informationen über die möglichen erreichten Orte zu einem Endzeitpunkt t_n haben, so lässt sich dies mit einem ***Histogramm*** verwirklichen. In diesem trägt man die Anzahl der Pfade, die in einem solchen Ort enden, gegen diese Orte auf.

In der Abbildung rechts ist ein solches Histogramm für den Zeitpunkt t_8 sowie die Wahrscheinlichkeiten $p = q = \frac{1}{2}$ dargestellt. Die grüne Kurve ist die skalierte (!) ***Dichtefunktion*** der entsprechenden ***Normal-*** oder ***Gauß-verteilung***, für feste Parameter gegeben durch

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

mit Parametern μ und $\sigma > 0$. Letztere entsprechen hierbei dem Erwartungswert beziehungsweise der Standardabweichung (s. u.). (Die Kurzschreibweise für eine Normalverteilung mit Parametern μ und σ ist $\mathcal{N}(\mu, \sigma^2)$).



Bemerkung 5.0.2: *Der Begriff Dichtefunktion rührt daher, dass in der Stochastik die Wahrscheinlichkeit P eigentlich ein Maß ist. Die Wahrscheinlichkeit dass eine Zufallsvariable X einen Wert $\leq x$ annimmt, wird dann mittels $P(X \leq x) = \int_{-\infty}^{x} f(\xi)\mathrm{d}\xi$ berechnet, wobei f die Dichte des Maßes ist.*

Beispiel (Fortsetzung):

In der Stochastik ist der ***Erwartungswert*** einer Zufallsvariablen X mit möglichen Werten X_j, $j = 1, ..., N$, welche jeweils mit Wahrscheinlichkeit $P(X = X_j)$ angenommen werden, definiert durch

$$\mathbb{E}(X) = \sum_{j=1}^{N} X_j P(X = X_j)$$

Gilt $p = q = \frac{1}{2}$, so ist der Erwartungswert der Ortsänderung z_j folglich

$$\mathbb{E}(z_j) = \frac{1}{2}\Delta x + \frac{1}{2}(-\Delta x) = 0.$$

Der Ort des Mannes nach n Schritten kann ebenfalls als Zufallsvariable s betrachtet werden

mit $s_n \in \{j\Delta x | j = -n, -n+1, ..., n\}$. Da der Erwartungswert linear ist, gilt

$$\mathbb{E}(s_n) = \mathbb{E}\left(\sum_{j=1}^{n} z_j\right) = \sum_{j=1}^{n} \mathbb{E}(z_j) = 0.$$

Dies lässt sich wie folgt interpretieren: Gemittelt über alle möglichen Orte, an denen der Mann sich zum Zeitpunkt t_n befinden kann, ist er überhaupt nicht getorkelt.
Die ***Varianz*** einer Zufallsvariablen X ist definiert durch

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right),$$

sie erfüllt die Gleichung

$$\mathrm{Var}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2.$$

Im vorliegenden Fall gilt also ($\mathbb{E}(s_n) = 0$)

$$\mathrm{Var}(s_n) = \mathbb{E}\left(s_n^2\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}(z_i z_j)$$

Da für $i \neq j$

$$\mathbb{E}(z_i z_j) = \frac{1}{2}\cdot\frac{1}{2}\Delta x\Delta x + \frac{1}{2}\cdot\frac{1}{2}\Delta x(-\Delta x) + \frac{1}{2}\cdot\frac{1}{2}(-\Delta x)\Delta x + \frac{1}{2}\cdot\frac{1}{2}(-\Delta x)(-\Delta x) = 0$$

gilt, folgt mittels $\mathbb{E}(z_i^2) = \frac{1}{2}\Delta x^2 + \frac{1}{2}(-\Delta x)^2$

$$\mathrm{Var}(s_n) = \sum_{i=1}^{n} \mathbb{E}\left(z_i^2\right) = n\cdot(\Delta x)^2$$

Im Gegensatz zum konstanten Erwartungswert steigt die Varianz im vorliegenden Fall linear mit der Anzahl an Schritten. Die ***Standardabweichung*** einer Zufallsvariablen ist definiert durch

$$\sigma(X) \coloneqq \sqrt{\mathrm{Var}(X)},$$

und beschreibt die Stärke der Streuung der Häufigkeiten um den Mittelwert. Im vorliegenden Falle gilt $\sigma(s_n) = \sqrt{n}\Delta x$.
Gilt hingegen $p \neq q$ (z. B. $p = \frac{3}{4}$, $q = \frac{1}{4}$), so führt dies im Mittel zu einem Drift des betrunkenen Mannes nach rechts, es gilt:

$$\mathbb{E}(s_n) = \sum_{j=1}^{n} \mathbb{E}(z_j) = \frac{n}{2}\Delta x,$$

die (mittlere) Abweichung vom Ursprungsort steigt also linear mit der Zeit, es kommt zu einem „Rechtsdrift“:

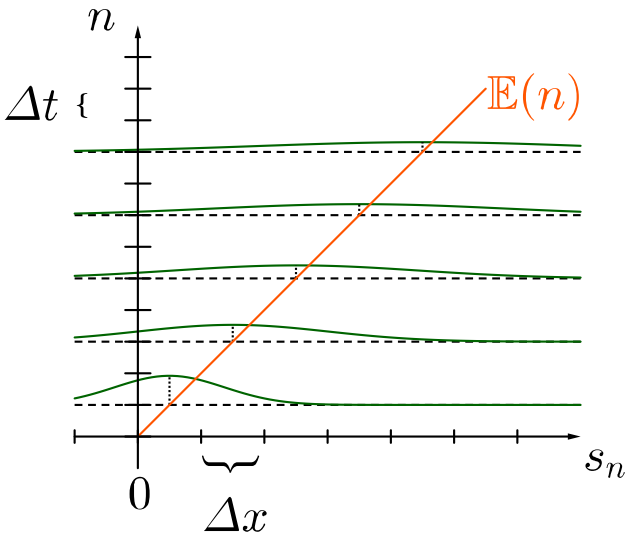


Abbildung 5.0.1: Verschiebung des Mittelwerts bei fortschreitender Zeit. Die grünen Kurven sind die zugehörigen Dichtefunktionen der Ortsvariablen.

Für die Varianz folgt somit mittels

$$\mathbb{E}(z_i z_j) = \left(\frac{3}{4}\right)^2 (\Delta x)^2 + \frac{3}{4}\cdot\frac{1}{4}\Delta x\,(-\Delta x) + \frac{1}{4}\cdot\frac{3}{4}(-\Delta x)\,\Delta x + \left(\frac{1}{4}\right)^2 (-\Delta x)^2 = \frac{1}{4}(\Delta x)^2$$

für $i \neq j$ sowie $\mathbb{E}(z_i^2) = \frac{3}{4}(\Delta x)^2 + \frac{1}{4}(-\Delta x)^2 = (\Delta x)^2$:

$$\begin{aligned}
\mathrm{Var}(s_n) =& \mathbb{E}\left(s_n^2\right) - \left(\mathbb{E}(s_n)\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(z_i z_j) - \frac{n^2}{4}(\Delta x)^2 = \\
&= \sum_{i=1}^{n}\mathbb{E}\left(z_i^2\right) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathbb{E}(z_i z_j) - \frac{n^2}{4}(\Delta x)^2 = \\
&= n(\Delta x)^2 + n(n-1)\frac{1}{4}(\Delta x)^2 - \frac{n^2}{4}(\Delta x)^2 = \\
&= \frac{3}{4}n(\Delta x)^2
\end{aligned}$$

5.1 Brownsche Bewegung

Definition 5.1.1:
Eine skalare ***Standard Brownsche Bewegung*** über dem Intervall $[0, T]$, auch ***Standard Wiener-Prozess*** genannt, ist eine zeitabhängige Zufallsvariable $W(t)$, welche stetig auf $[0, T]$ ist und folgenden Bedingungen genügt:

1. $W(0) = 0$ mit Wahrscheinlichkeit 1.
2. Für $0 \leq s < t \leq T$ ist das ***Wiener-Inkrement*** $W(t) - W(s)$ normalverteilt mit Erwartungswert 0 und Varianz $t - s$, kurz $W(t) - W(s) \sim \sqrt{t-s}\,\mathcal{N}(0,1)$.
3. Für $0 \leq s < t < u < v \leq T$ sind die Inkremente $W(t) - W(s)$ und $W(v) - W(u)$ stochastisch unabhängig.

Definition 5.1.2:
Eine skalare ***Brownsche Bewegung mit Drift μ und Diffusionskonstante $\sigma^2 > 0$*** über dem Intervall $[0, T]$ist eine zeitabhängige Zufallsvariable $W(t)$, welche stetig auf $[0, T]$ ist und folgenden Bedingungen genügt:

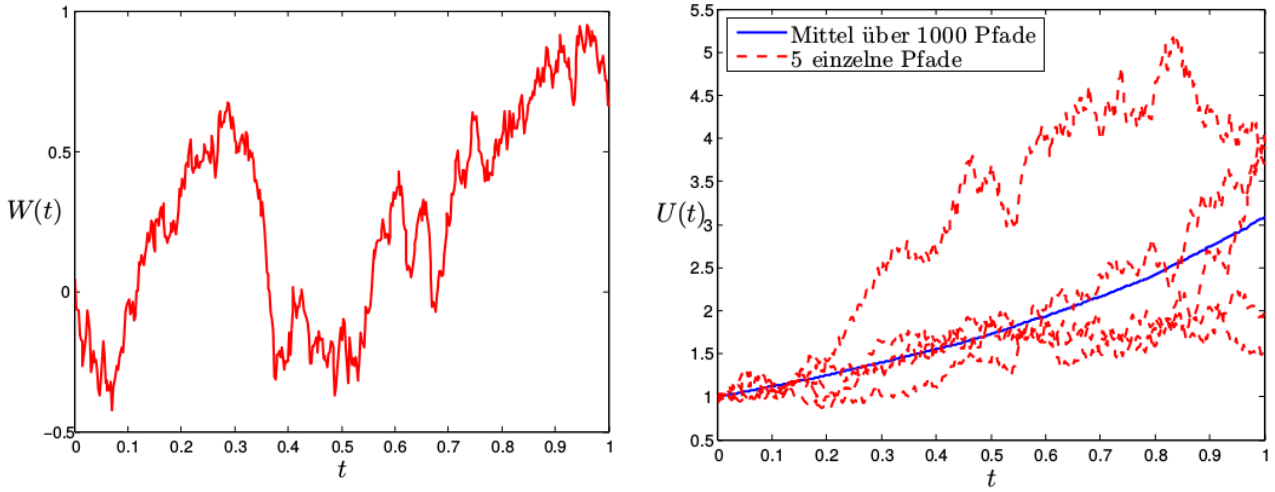
1. $W(0) = 0$ mit Wahrscheinlichkeit 1.
2. Für $0 \leq s < t \leq T$ ist das Wienerinkrement $W(t) - W(s)$ normalverteilt mit Erwartungswert $\mu(t-s)$ und Varianz $\sigma^2(t-s)$, kurz $W(t) - W(s) \sim \mathcal{N}(\mu(t-s), \sigma^2(t-s))$.
3. Für $0 \leq s < t < u < v \leq T$ sind die Inkremente $W(t) - W(s)$ und $W(v) - W(u)$ stochastisch unabhängig.

Bemerkung 5.1.3: *Obwohl eine Brownsche Bewegung auf $[0, T]$ in jedem Punkt $t_0 \in [0, T]$ stetig ist, ist sie in jedem dieser Punkte mit Wahrscheinlichkeit 1 nicht differenzierbar.*

Zur numerischen Simulation lässt sich eine Brownsche Bewegung wie folgt diskretisieren: Für eine Anzahl $N > 0$ an Knotenpunkten sei $\delta t := \frac{T}{N}$.Ähnlich wie im Einführungsbeispiel sei $t_j = j\delta t$ und $W_j = W(t_j)$. Aus der ersten Bedingung ergibt sich $W_0 = 0$ (mit Wahrscheinlichkeit 1), aus den anderen beiden die Gleichung

$$W_j = W_{j-1} + \mathrm{d}W_j,\ j = 1, 2, ..., N,$$

wobei $\mathrm{d}W_j$ eine Zufallsvariable der Form $\sqrt{\delta t}\mathcal{N}(0, 1)$ ist. Das Ergebnis einer solchen Simulation ist in Abbildung 5.1.1a



(a) **(b)**

Abbildung 5.1.1

zu sehen. Der zugehörige Code befindet sich in Anhang E.1. Selbstverständlich lässt sich eine Brownsche Bewegung auch mit anderen Funktionen verknüpfen. So sind in Abbildung 5.1.1b für die Funktion

$$u(t) = \exp\left(t + \frac{1}{2}W(t)\right)$$

fünf Pfade sowie das Mittel von 1000 Pfaden dargestellt. (Auch hier befindet sich der Code in Anhang E.1.)

Beispiel (Fortsetzung zum Einführungsbeispiel):

Wie bereits gesehen, kommt es bei $P(z_j = \Delta x) > P(z_j = -\Delta x)$ zu einem Drift. Der Drift ist gegeben durch

$$\mu = \frac{\mathbb{E}(z_j)}{\Delta t},$$

die Diffusionskonstante durch

$$\sigma = \sqrt{\frac{\mathrm{Var}(z_j)}{\Delta t}}.$$

Es handelt sich hierbei folglich um eine diskretisierte Brownsche Bewegung mit Drift $\frac{\mathbb{E}(z_j)}{\Delta t}$ und Diffusion $\frac{\mathrm{Var}(z_j)}{\Delta t}$.

Sei nun $W(t)$ eine Brownsche Standardbewegung, $X(t), f(t, X(t)), \sigma(t, X(t))$ seien reellwertige Funktionen. Eine ***stochastischen Differentialgleichung*** (kurz SDE) ist eine Differentialgleichung der Form

$$\mathrm{d}x(t) = f(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}W(t).$$

(Die Bedeutung von $\mathrm{d}W(t)$ wird in den folgenden Kapiteln erklärt). Eine Schreibweise in der Form $\dot{x} = ...$ ist hierbei jedoch unzulässig. In diesem Falle stünde auf der rechten Seite $\frac{\mathrm{d}W(t)}{\mathrm{d}t}$, was jedoch der fast sicheren Nicht-Differenzierbarkeit widerspricht. Gilt hingegen $\sigma(t, X(t)) \equiv 0$, so ist dies durchaus zulässig, in diesem Falle entspricht die stochastische Differentialgleichung einer gewöhnlichen. Ähnlich wie bei diesen lassen sich auch bei stochastischen Differentialgleichungen Anfangsbedingungen der Form $X(0) = X_0$ stellen. Gilt $\sigma \not\equiv 0$, so ist $X(t)$ für jedes t eine Zufallsvariable. Ein Antwortproblem der Form

$$\begin{cases} \mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}W(t) \\ X(0) = X_0 \end{cases}$$

kann hierbei eine eindeutige Lösung, aber dennoch mehrere verschiedene Trajektorien haben: Der Begriff der Eindeutigkeit bezieht sich in diesem Kontext auf die Gesamtheit der Trajektorien sowie ihren Wahrscheinlichkeiten.

5.2 Stochastische Integration

Ähnlich wie bei gewöhnlichen Differentialgleichungen lässt sich eine Lösung X in Integralform angeben:

$$X(t) = X_0 + \int_0^t f(\tau, X(\tau))\mathrm{d}\tau + \int_0^t \sigma(\tau, X(\tau))\mathrm{d}W(\tau), \quad 0 \le t \le T$$

Bei dem zweiten Integral handelt es sich um ein stochastisches Integral. In diesem Kapitel soll ein kurzer Einblick in die Berechnung solcher Integrale gegeben werden.
Sei hierzu $\{t_0, t_1, ..., t_N\}$ eine Zerlegung des Intervalls $[0, T]$, so dass $t_0 = 0 < t_1 < ... < t_N = T$ gilt. Für eine Riemann-integrierbare Funktion h lässt sich das Riemann-Integral $\int_0^T h(\tau)\mathrm{d}\tau$ bekanntlich mittels

$$\int_0^T g(\tau)\mathrm{d}\tau \approx \sum_{i=0}^{N-1} h(t_j)(t_{j+1} - t_j)$$

approximieren, genauer gilt:

$$\lim_{N\to 0} \sum_{i=0}^{N-1} h(t_j)(t_{j+1} - t_j) = \lim_{\max_{1\le i\le N-1}|t_{j+1}-t_j|} \sum_{i=0}^{N-1} h(t_j)(t_{j+1} - t_j) = \int_0^T h(\tau)\mathrm{d}\tau$$

Analog hierzu lässt sich die Summe

$$\sum_{i=0}^{N-1} h(t_j)(W(t_{j+1}) - W(t_j))$$

betrachten, welche als Approximation von $\int_0^T h(t)\mathrm{d}W(t)$ betrachtet werden kann. Auf diese Weise ist das Itô-Integral (benannt nach Kiyoshi Itō, dem Begründer der stochastischen Analysis) definiert:

Definition 5.2.1:
Sei $\sigma(t)$ ein stochastischer Prozess, $W(t)$ ein Wienerprozess. S heißt ***Itô-Integral*** des stochastischen Prozesses σ mit Wienerprozess W, wenn

$$\lim_{N\to\infty} \mathbb{E}\left(\left|S - \sum_{i=0}^{N-1} \sigma(t_j)\left(W(t_{j+1}) - W(t_j)\right)\right|^2\right) = 0$$

gilt.

Wird an Stelle der Summe $\sum_{i=0}^{N-1} h(t_j)(W(t_{j+1}) - W(t_j))$ die dem Mittelpunktsverfahren ähnelnde Summe

$$\sum_{i=0}^{N-1} h\left(\frac{t_{j+1}+t_j}{2}\right)\left(W(t_{j+1}) - W(t_j)\right)$$

verwendet, führt dies auf das sogenannte ***Stratonovich-Integral***. Während es bei hinreichend glatten, deterministischen Prozessen h irrelevant ist, welche der beiden Summen für die Grenzwertbildung verwendet wird, ist dies bei stochastischen Prozessen nicht egal, wie das folgende Beispiel zeigt.

Beispiel 5.2.2:

Sei W ein Wienerprozess. Für beliebiges t_i gilt

$$\begin{aligned} W(t_i)\left(W(t_{i+1}) - W(t_i)\right) &= W(t_i)W(t_{i+1}) - W(t_i)^2 = \\ &= -\frac{1}{2}\left(W(t_{i+1}) - W(t_i)\right)^2 + \frac{1}{2}W(t_{i+1})^2 + \frac{1}{2}W(t_i)^2 - W(t_i)^2 = \\ &= \frac{1}{2}\left(W(t_{i+1})^2 - W(t_i)^2\right) - \frac{1}{2}\left(W(t_{i+1}) - W(t_i)\right)^2 \end{aligned}$$

Somit ergibt sich

$$\begin{aligned} \int_0^T W(t)\,\mathrm{d}W(t) &= \lim_{N\to\infty} \sum_{i=0}^{N-1} W(t_i)\left(W(t_{i+1}) - W(t_i)\right) = \\ &= \lim_{N\to\infty} \left[\frac{1}{2}\underbrace{\sum_{i=0}^{N-1}\left(W(t_{i+1})^2 - W(t_i)^2\right)}_{=W(T)^2 - W(0)^2 = W(T)^2} - \frac{1}{2}\sum_{i=0}^{N-1}\left(W(t_{i+1}) - W(t_i)\right)^2\right] = \\ &= \frac{1}{2}W(T)^2 - \frac{1}{2}\lim_{N\to\infty}\sum_{i=0}^{N-1}\left(W(t_{i+1}) - W(t_i)\right)^2 \end{aligned}$$

Da ferner

$$\mathbb{E}\left(\lim_{N\to\infty}\sum_{i=0}^{N-1}\left(W\left(t_{i+1}\right)-W\left(t_{i+1}\right)\right)^2\right)=\lim_{N\to\infty}\sum_{i=1}^{N-1}\mathbb{E}\left(\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)^2\right)=$$
$$=\lim_{N\to\infty}\sum_{i=1}^{N-1}\underbrace{\mathrm{Var}\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)}_{=t-s}+\left(\underbrace{\mathbb{E}\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)}_{=0}\right)^2=T$$

gilt und sich ferner $\mathrm{Var}\left(\left(W(t_{i+1})-W(t_i)\right)^2\right)=2\left(t_{i+1}-t_i\right)^2$ zeigen lässt, ergibt sich insgesamt

$$\int_0^T W\left(t\right)\mathrm{d}W(t)=\frac{1}{2}W\left(T\right)^2-\frac{1}{2}T$$

Im Falle des Stratonovich-Integrals lautet die entsprechende Summe

$$\sum_{i=0}^{N-1}W\left(\frac{t_i+t_{i+1}}{2}\right)\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)$$

Addiert man zu $\frac{W(t_i)+W(t_{i+1})}{2}$ eine $\mathcal{N}(0,\frac{\Delta t}{4})$-verteilte Zufallsvariable ΔZ_j, so genügen die so erhaltenen Werte den Bedingungen von Definition 5.1.1. Es ergibt sich

$$\sum_{i=0}^{N-1}\left(\frac{W(t_i)+W(t_{i+1})}{2}+\Delta Z_i\right)\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)=$$
$$=\frac{1}{2}\sum_{i=0}^{N-1}W\left(t_{i+1}\right)^2-W(t_i)^2+\sum_{i=0}^{N-1}\Delta Z_i\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)=$$
$$=\frac{1}{2}W(T)^2-\frac{1}{2}W(0)^2+\sum_{i=0}^{N-1}\Delta Z_i\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)$$

Da

$$\mathbb{E}\left(\sum_{i=0}^{N-1}\Delta Z_i\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)\right)=\sum_{i=0}^{N-1}\mathbb{E}\left(\Delta Z_i\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)\right)$$
$$=\sum_{i=0}^{N-1}\mathbb{E}\left(\Delta Z_i\right)\mathbb{E}\left(\left(W\left(t_{i+1}\right)-W\left(t_i\right)\right)\right)=0$$

gilt und sich auch hier $\mathrm{Var}\left(\Delta Z_j\left(W(t_{i+1}-W(t_i)\right)\right)\overset{N\to\infty}{\to}0$ zeigen lässt, ergibt sich als Stratonovich-Integral

$$\int_0^T W(t)\mathrm{d}W(t)=\frac{1}{2}W(T)^2$$

Beispiel 5.2.3:

Sei σ konstant. Für das Itô-Integral gilt:

$$\int_0^T\sigma\mathrm{d}W(t)=\lim_{N\to\infty}\sum_{i=0}^{N-1}\sigma\left(W(t_{i+1})-W(t_i)\right)=\lim_{N\to\infty}\sigma(W(T)-W(0))=\sigma W(T)$$

Aus der Definition des Itô-Integrals ergibt sich desweiteren folgende Eigenschaft: Ist zu jedem beliebigen Zeitpunkt t die zu integrierende Funktion $\sigma(t)$ stochastisch unabhängig von $W(t^*)$

für alle $t^* \geq t$, so ist der Erwartungswert 0, da in diesem Falle

$$\mathbb{E}\left(\int_0^T \sigma(t)\mathrm{d}W(t)\right) = \mathbb{E}\left(\lim_{N\to\infty}\sum_{i=0}^{N-1}\sigma(t_i)\left(W\left(t_{i+1}\right) - W\left(t_{i+1}\right)\right)\right) =$$
$$= \lim_{N\to\infty}\sum_{i=1}^{N-1}\mathbb{E}\left(\sigma(t_i)\right)\cdot\underbrace{\mathbb{E}\left(W(t_{i+1}) - W(t_i)\right)}_{=0}$$

gilt.

Lemma 5.2.4 (Lemma von Itô):
Gegeben sei eine stochastische Differentialgleichung

$$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}W(t)$$

mit $\sigma \not\equiv 0$. Ferner sei

$$Y(t) = \varphi(t, X(t)),$$

Ist φ zweimal stetig differenzierbar, so genügt Y der stochastischen Differentialgleichung

$$\mathrm{d}Y(t) = \tilde{f}(t, X(t))\mathrm{d}t + \tilde{\sigma}(t, X(t))\mathrm{d}W(t)$$

mit

$$\tilde{f}(t, X(t)) \coloneqq \varphi_t\left(t, X\left(t\right)\right) + \varphi_X\left(t, X\left(t\right)\right) f\left(t, X\left(t\right)\right) + \frac{1}{2}\varphi_{XX}\left(t, X\left(t\right)\right)\sigma^2\left(t, X\left(t\right)\right)$$

und

$$\tilde{\sigma}(t, X(t)) \coloneqq \varphi_X\left(t, X\left(t\right)\right)\sigma\left(t, X\left(t\right)\right)$$

Beweisidee/-skizze: Taylorentwicklung liefert:

$$\mathrm{d}Y(t) = \varphi_t(t, X(t))\mathrm{d}t + \varphi_X(t, X(t))\mathrm{d}X(t) + \frac{1}{2}\varphi_{tt}(t, X(t)) + \varphi_{XX}\left(\mathrm{d}X(t)\right) + ...$$

$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}W(t)$ eingesetzt:

$$\begin{aligned}\mathrm{d}Y\left(t\right) = {} & \varphi_t\left(t, X\left(t\right)\right)\mathrm{d}t + \varphi_X\left(t, X\left(t\right)\right)\left(f\left(t, X\left(t\right)\right)dt + \sigma\left(t, X\left(t\right)\right)\mathrm{d}W(t)\right) \\ & + \frac{1}{2}\varphi_{tt}\left(t, X\left(t\right)\right)\left(\mathrm{d}t\right)^2 + \frac{1}{2}\varphi_{XX}\left(t, X\left(t\right)\right)\left[f^2\left(t, X\left(t\right)\right)\left(\mathrm{d}t\right)^2\right. \\ & \left. + \sigma^2\left(t, X\left(t\right)\right)\left(\mathrm{d}W(t)\right)^2 + 2f\left(t, X\left(t\right)\right)\sigma\left(t, X\left(t\right)\right)\mathrm{d}t\mathrm{d}W(t)\right]\end{aligned}$$

Beachtet man bei der Vernachlässigung der Terme höherer Ordnung, dass infolge der $\mathcal{N}(0, \mathrm{d}t)$-Verteilung von $W(t)$ der Term $\left(\mathrm{d}W(t)\right)^2$ durch $\mathrm{d}t$ zu ersetzen ist, folgt die Behauptung. □

Bemerkung 5.2.5: *Aus dem in der Beweisskizze verwendeten Argument, dass der* $\left(\mathrm{d}W(t)\right)^2$ *Term den* $\mathrm{d}t$*-Termen hinzuzurechnen ist, ergibt sich für zwei stochastische Prozesse* X, Y *mit* $\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}W(t)$ *und* $\mathrm{d}Y(t) = \tilde{f}(t, Y(t))\mathrm{d}t + \tilde{\sigma}(t, Y(t))\mathrm{d}W(t)$ *als Produktregel die Gleichung*

$$\mathrm{d}\left(X(t)Y(t)\right) = X(t)\cdot\mathrm{d}Y(t) + \mathrm{d}X(t)\cdot Y(t) + \sigma(t, X(t))\cdot\tilde{\sigma}(t, Y(t))\mathrm{d}t$$

Betrachtet man anstelle einer Funktion $\varphi(t, X(t))$ eine Funktion $V(X)$, welche mit $X(t)$ verknüpft wird, so ergibt sich mit Itôs Lemma ($\varphi_t = V_t \equiv 0, \varphi_X = V_X$) als stochastische Kettenregel die Gleichung

$$\mathrm{d}V(X(t)) = \left(V_X(X(t))f(t, X(t)) + \frac{1}{2}V_{XX}(X(t))\sigma(t, X(t))^2\right)\mathrm{d}t + V_X(X(t))\sigma(t, X(t))\mathrm{d}W(t)$$

beziehungsweise ($f(t,X(t))\mathrm{d}t + \sigma(t,X(t))\mathrm{d}W(t) = \mathrm{d}X$)

$$\mathrm{d}V(X(t)) = V_X(X(t))\mathrm{d}X + \frac{1}{2}\sigma(t,X(t))^2 V_{XX}(X(t))\mathrm{d}W(t).$$

Mit Hilfe dieses Lemmas lassen sich einige stochastische Differentialgleichungen sehr elegant lösen, wie in den folgenden Beispielen dargelegt wird.

Beispiel 5.2.6 (Geometrische Brownsche Bewegung):

Gegeben sei die stochastische Differentialgleichung

$$\mathrm{d}S(t) = \mu S(t)\mathrm{d}t + \sigma S(t)\mathrm{d}W(t)$$

mit Zufallsprozess S sowie Konstanten μ und σ. Es sei

$$Y(t) = \varphi(t,S(t)) \coloneqq \log(S(t))$$

Gemäß dem Lemma von Itô gilt

$$\begin{aligned}\mathrm{d}Y(t) &= \left(\varphi_t + \varphi_S \mu S + \frac{1}{2}\varphi_{SS}\sigma^2 S^2\right)\mathrm{d}t + \varphi_S \sigma S \mathrm{d}W(t) = \\ &= \left(\frac{1}{S}\mu S - \frac{1}{2}\frac{1}{S^2}\sigma^2 S^2\right)\mathrm{d}t + \frac{1}{S}\sigma S \mathrm{d}W(t) = \\ &= \left(\mu - \frac{1}{2}\sigma^2\right)\mathrm{d}t + \sigma \mathrm{d}W(t)\end{aligned}$$

Also gilt

$$Y(t) = Y_0 + \int_0^t \left(\mu - \frac{1}{2}\sigma^2\right)\mathrm{d}t + \int_0^t \sigma \mathrm{d}W(t) = \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma W(t)$$

Wegen $Y(t) = \log(S(t))$ folgt

$$S(t) = e^{Y(t)} = \underbrace{S_0}_{=e^{Y_0}} e^{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)}$$

Die Lösung $S(t)$ heißt ***Geometrische Brownsche Bewegung***.

Beispiel 5.2.7 (Lineare stochastische Differentialgleichung):

Gegeben sei das AWP

$$\begin{cases} \mathrm{d}X(t) = f(t,X(t))\mathrm{d}t + \sigma(t,X(t))\mathrm{d}W(t) \\ X(0) = X_0 \end{cases} \qquad (*)$$

mit

$$f(t,X(t)) = A(t)X(t) + a(t), \qquad \sigma(t,X(t)) = B(t)X(t) + b(t)$$

Bei diesem Problem kann man einen ähnlichen Ansatz wie im Falle einer linearen, gewöhnlichen Differentialgleichung (Kapitel 1.1, Fall (6)) verfolgen. Man teilt die Lösung $X(t)$ in

einen homogenen Teil X_h, welcher

$$\begin{cases} \mathrm{d}X_h(t) = f(t, X_h(t))\mathrm{d}t + \sigma(t, X_h(t))\mathrm{d}W(t) \\ X_h(0) = 1 \end{cases}$$

löst, und eine spezielle Lösung $X_p(t)$ von $(*)$ auf.
Setzt man $Y_h(t) = \varphi(t, X_h(t)) = \log(X_h(t))$, so ergibt sich, analog zu obigen Beispiel, aus dem Lemma von Itô:

$$\begin{aligned} \mathrm{d}Y_h\,(t) &= \left(\frac{1}{X_h(t)}A(t)X_h(t) - \frac{1}{2}\frac{1}{X_h(t)^2}B(t)^2X_h(t)^2\right)\mathrm{d}t + \frac{1}{X_h(t)}B(t)X_h(t)\mathrm{d}W(t) = \\ &= \left(A(t) - \frac{B(t)^2}{2}+\right)\mathrm{d}t + B(t)\mathrm{d}W(t) \end{aligned}$$

Hieraus folgt

$$X_h(t) = e^{Y_h(t)} = \exp\left(\int_0^t A(\tau) - \frac{B(\tau)^2}{2}\mathrm{d}\tau + \int_0^t B(\tau)\mathrm{d}W(\tau)\right)$$

Der Ansatz $X_p(t) = c(t)X_h(t)$ liefert mit $\mathrm{d}X_p(t) = \mathrm{d}c(t)\cdot X_h(t) + c(t)\mathrm{d}X_h(t) + \mathrm{d}c(t)\cdot \mathrm{d}X_h(t)$ Gleichung

$$\begin{gathered} \mathrm{d}c(t)\cdot X_h(t) + c(t)\mathrm{d}X_h(t) + \mathrm{d}c(t)\cdot \mathrm{d}X_h(t) \\ = \\ (A(t)c(t)X_h(t) + a(t))\mathrm{d}t + (B(t)c(t)X_h(t) + b(t))\mathrm{d}W(t) \end{gathered}$$

Wegen $c(t)\mathrm{d}X_h(t) = A(t)c(t)X_h(t)\mathrm{d}t + B(t)c(t)X_h(t)\mathrm{d}W(t)$ folgt

$$\mathrm{d}c(t)\cdot X_h(t) + \mathrm{d}c(t)\mathrm{d}X_h(t) = a(t)\mathrm{d}t + b(t)\mathrm{d}W(t)$$

und somit

$$\mathrm{d}c(t)\,(X_h(t) + A(t)X_h(t)\mathrm{d}t + B(t)X(t)\mathrm{d}W(t)) = a(t)\mathrm{d}t + b(t)\mathrm{d}W(t)$$

Schreibt man $\mathrm{d}c(t)$ als Itô-Prozess $\mathrm{d}c(t) = \tilde{f}(t, c(t))\mathrm{d}t + \tilde{\sigma}(t, c(t))\mathrm{d}W(t)$, so ergibt sich bei Einsetzen in obiges

$$\begin{aligned} a(t)\mathrm{d}t + b(t)\mathrm{d}W(t) = {} & X_h(t)\left(\tilde{f}(t, c(t))\mathrm{d}t + \tilde{\sigma}(t, c(t))\mathrm{d}W(t)\right) \\ & + A(t)X_h(t)\mathrm{d}t\left(\tilde{f}(t, c(t))\mathrm{d}t + \tilde{\sigma}(t, c(t))\mathrm{d}W(t)\right) \\ & + B(t)X_h(t)\mathrm{d}W(t)\left(\tilde{f}(t, c(t))\mathrm{d}t + \tilde{\sigma}(t, c(t))\mathrm{d}W(t)\right) = \\ = {} & X_h(t)\tilde{f}(t, c(t))\mathrm{d}t + B(t)X_h(t)\tilde{\sigma}(t, c(t))(\mathrm{d}W(t))^2 \\ & + \tilde{\sigma}(t, c(t))X_h(t)\mathrm{d}W(t) + A(t)X_h(t)\tilde{f}(t, c(t))(\mathrm{d}t)^2 \\ & + \left(A(t)X_h(t)\tilde{\sigma}(t, c(t)) + B(t)X_h(t)\tilde{f}(t)\right)\mathrm{d}t\mathrm{d}W(t) \end{aligned}$$

Durch Vernachlässigung der Terme höherer Ordnung ergibt sich hieraus durch Koeffizientenvergleich (man beachte $(\mathrm{d}W(t))^2 \cong \mathrm{d}t$):

$$\begin{aligned} a(t) &= X_h(t)\tilde{f}(t, c(t)) + B(t)X_h(t)\tilde{\sigma}(t, c(t)) \\ b(t) &= \tilde{\sigma}(t, c(t))X_h(t) \end{aligned}$$

Aus der zweiten Gleichung ergibt sich $\sigma(t, c(t)) = \frac{b(t)}{X_h(t)}$. Einsetzen in die erste Gleichung liefert

$$a(t) = X_h(t)\tilde{f}(t, c(t)) + B(t)X_h(t)\frac{b(t)}{X_h(t)}$$

und somit ergibt Auflösen nach $\tilde{f}(t, c(t))$:

$$\tilde{f}(t, c(t)) = \frac{1}{X_h(t)}\left(a(t) - B(t)b(t)\right)$$

Hieraus folgt

$$\begin{aligned} c(t) &= c(0) + \int_0^t \tilde{f}(\tau, c(\tau))\mathrm{d}\tau + \int_0^t \tilde{\sigma}(\tau, c(\tau))\mathrm{d}W(\tau) = \\ &= c_0 + \int_0^t X_h^{-1}(\tau)\left(a(\tau) - B(t)b(t)\right)\mathrm{d}\tau + \int_0^t X_h^{-1}(t)b(\tau)\mathrm{d}W(\tau) \end{aligned}$$

Wegen $X(0) \overset{!}{=} X_0$ lautet die allgemeine Lösung von $(*)$ folglich

$$X(t) = \phi(t) \cdot \left(X_0 + \int_0^t \phi^{-1}(\tau)a(\tau)\mathrm{d}\tau + \int_0^t \phi^{-1}(\tau)b(\tau)\mathrm{d}W(\tau)\right)$$

mit

$$\phi(t) = \exp\left(\int_0^t A(\tau) - \frac{B(\tau)^2}{2}\mathrm{d}\tau + \int_0^t B(\tau)\mathrm{d}W(\tau)\right).$$

Gilt $\sigma \equiv 0$, muss insbesondere $b(t) \equiv 0$ und $B(t) \equiv 0$ sein. In diesem Falle fallen alle Teile obiger Lösung, welche $B(t)$ und $b(t)$ enthalten, weg (insbesondere die stochastischen Integrale). Die übrigbleibende Gleichung entspricht genau der in Kapitel 1.1 berechneten Lösung.

Statt einem festen Wert X_0 wie im Eingangsbeispiel, kann X_0 auch eine normalverteilte Zufallsvariable sein. In beiden Fällen ist der Mittelwert $m(t) := \mathbb{E}(X(t))$ Lösung des AWPs

$$\begin{cases} \dot{m}(t) = A(t)m(t) + a(t) \\ m(0) = \mathbb{E}(X_0) \end{cases}$$

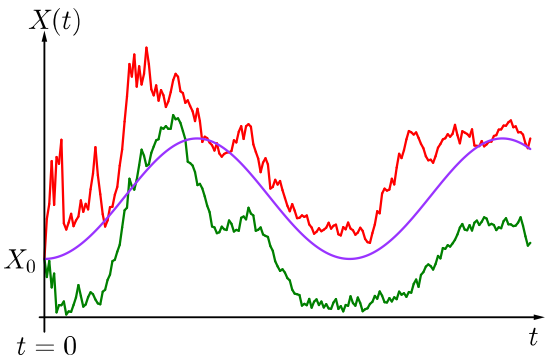


Abbildung 5.2.1: Exemplarische Trajektorien einer linearen stochastischen Differentialgleichung mit periodischem Mittelwert (violett).

Um dies zu zeigen, sei $m(t) = \mathbb{E}(X(t))$. Aus der Linearität des Mittelwerts ergibt sich

$$\begin{aligned} m(t) &= \mathbb{E}\left(X_0 + \int_0^t (t, X(t))\mathrm{d}t + \int_0^t \sigma(t, X(t))\mathrm{d}W(t)\right) = \\ &= \mathbb{E}(X_0) + \mathbb{E}\left(\int_0^t A(t)X(t)a(t)\mathrm{d}t\right) + \underbrace{\mathbb{E}\left(\int_0^t \sigma(t, X(t))\mathrm{d}W(t)\right)}_{=0} = \\ &= \mathbb{E}(X_0) + \int_0^t A(t)\mathbb{E}(X(t)) + a(t)\mathrm{d}t = \mathbb{E}(X_0) + \int_0^t A(t)m(t) + a(t)\mathrm{d}t \end{aligned}$$

Differentiation nach t liefert schließlich

$$\dot{m}(t) = A(t)m(t) + a(t)$$

Somit gilt gemäß Kapitel 1.1 für den Mittelwert

$$\mathbb{E}\left(X(t)\right) = e^{\int_0^t A(\tau)\mathrm{d}\tau} \cdot \left(\mathbb{E}\left(X_0\right) + \int_0^t a(\tau) e^{-\int_0^\tau A(s)\mathrm{d}s}\mathrm{d}\tau\right)$$

Sei nun $P(t) = \mathbb{E}\left(X(t)^2\right)$. Nach Lemma 5.2.4 genügt $X(t)^2$ der stochastischen Differentialgleichung

$$\begin{aligned}\mathrm{d}(X(t)^2) = &\left[\left(2A\left(t\right) + B\left(t\right)\right) X^2\left(t\right) + 2\left(B\left(t\right) b\left(t\right) + a\left(t\right)\right) X\left(t\right) + b^2(t)\right] \mathrm{d}t \\ &+ 2X\left(t\right)\left(B\left(t\right) X\left(t\right) + b\left(t\right)\right) \mathrm{d}W(t)\end{aligned}$$

(Hier ist $\varphi(t, X(t)) = X^2(t)$ und somit $\varphi_t(t, X(t)) = 0, \varphi_X(t, X(t)) = 2X$ sowie $\varphi_X X(t, X(t)) = 2$.) Es ergibt sich

$$\begin{aligned}P(t) &= \mathbb{E}\left(\mathbb{E}\left(X_0^2\right) + \int_0^t \left(2A\left(t\right) + B\left(t\right)\right) X^2\left(t\right) + 2\left(B\left(t\right) b\left(t\right) + a\left(t\right)\right) X\left(t\right) + b^2(t)\mathrm{d}t \right.\\ &\qquad \left. + \int_0^t 2X\left(t\right)\left(B\left(t\right) X\left(t\right) + b\left(t\right)\right) \mathrm{d}W(t)\right) \\ &= \mathbb{E}\left(X_0^2\right) + \int_0^t \left(2A\left(t\right) + B\left(t\right)\right) \mathbb{E}\left(X(t)\right)^2 + 2\left(B\left(t\right) b\left(t\right) + a\left(t\right)\right) \mathbb{E}\left(X(t)\right) + b^2(t)\mathrm{d}t\end{aligned}$$

Hieraus folgt

$$\dot{P}(t) = \left(2A\left(t\right) + B\left(t\right)\right) P(t) + 2\left(B\left(t\right) b\left(t\right) + a\left(t\right)\right) m(t) + b^2(t)$$

Für die Varianz gilt:

$$\mathrm{Var}\left(X(t)\right) = \mathbb{E}\left(\left(X(t) - \mathbb{E}\left(X(t)\right)\right)^2\right) = \mathbb{E}\left(X(t)^2\right) - \mathbb{E}\left(X(t)\right)^2$$

Wegen

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(t) - \frac{\mathrm{d}}{\mathrm{d}t}(m^2(t)) &= \dot{P}(t) - 2m(t)\dot{m}(t) = \\ &= \left(2A\left(t\right) + B\left(t\right)\right) P(t) + 2\left(B\left(t\right) b\left(t\right) + a\left(t\right)\right) m(t) + b^2(t) \\ &\quad - 2m(t)\left(A(t)m(t) + a(t)\right) = \\ &= 2A(t)\left(P(t) - m^2(t)\right) + B(t)P(t) + 2B(t)b(t)m(t) + b^2(t) = \\ &= \left(2A(t) + B(t)\right)\left(P(t) - m^2(t)\right) + B(t)m^2(t) + 2B(t)b(t)m(t) + b^2(t) = \\ &= \left(2A(t) + B(t)\right)\left(P(t) - m^2(t)\right) + B(t)m^2(t) + 2B(t)b(t)m(t) \\ &\quad + B(t)b^2(t) - B(t)b^2(t) + b^2(t) = \\ &= \left(2A(t) + B(t)\right)\left(P(t) - m^2(t)\right) + B(t)\left(m(t) + b(t)\right)^2 + b^2(t)(1 - B(t))\end{aligned}$$

genügt die Varianz $\mathrm{Var}\left(X(t)\right)$ folglich der Differentialgleichung

$$\dot{v}(t) = \left(2A(t) + B(t)\right) v(t) + B(t)\left(m(t) + b(t)\right)^2 + b^2(t)\left(1 - B(t)\right)$$

sowie der Anfangsbedingung $v(0) = \mathrm{Var}\left(X_0\right)$.

5.3 Numerik stochastischer Differentialgleichungen

5.3.1 Die Eulyer-Maruyama-Methode

Wie bereits gesehen, kann eine stochastische Differentialgleichung

$$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}W(t), \quad X(0) = X_0, \quad 0 \leq t \leq T$$

auch in Integralform geschrieben werden:

$$X(t) = X_0 + \int_0^t f(t, X(t))\mathrm{d}t + \int_0^t \sigma(t, X(t))\mathrm{d}W(t)$$

Wird das Intervall $[0, T]$ durch Zeitpunkte $\tau_i = i\Delta t$ diskretisiert (mit $\Delta t = \frac{T}{L}$ für eine positive Konstante L), so lassen sich zur numerischen Berechnung einer Lösung ähnliche Verfahren wie im Falle gewöhnlicher Differentialgleichungen anwenden. Die sogenannte ***Euler-Maruyama-Methode*** (kurz EM) berechnet ausgehend von einer Näherung X_j zum Zeitpunkt τ_j den Wert zum nächsten Zeitpunkt mittels

$$X_{i+1} = X_i + f(\tau_i, X_i)\Delta t + \sigma(\tau_i, X_i) \cdot (W(\tau_{i+1}) - W(\tau_i)) \,. \tag{5.3.1}$$

Es handelt sich hierbei um eine Näherungslösung der Integralgleichung

$$X(\tau_{i+1}) = X(\tau_i) + \int_{\tau_i}^{\tau_{i+1}} f(t, X(t))\mathrm{d}t + \int_{\tau_i}^{\tau_{i+1}} \sigma(t, X(t))\mathrm{d}W(t).$$

(Im Falle $\sigma \equiv 0$, das heißt einer gewöhnlichen Differentialgleichung, entspricht dies dem aus Kapitel 1.11 bekannten expliziten Euler-Verfahren.) Es empfiehlt sich hierbei, Δt als ganzzahliges Vielfaches von δt zu wählen, da hierdurch die Zeitpunkte t_j, für die die Brownsche Bewegung simuliert wird, die Zeitpunkte τ_i enthalten, an denen die Näherungslösung der EM berechnet wird. Wird diese Vielfachheit mit R bezeichnet, das heißt gilt $\Delta t = R\delta t$, so ist das Inkrement $W(\tau_{i+1}) - W(\tau_i)$ durch

$$W(\tau_{i+1}) - W(\tau_i) = W((i+1)R\delta t) - W(iR\delta t) = \sum_{j=iR+1}^{(i+1)R} \mathrm{d}W_j$$

gegeben, wobei $\mathrm{d}W_j$ die Wiener-Inkremente aus Kapitel 5.1 sind.

Beispiel 5.3.1:

Gegeben sei das stochastische Anfangswertproblem

$$\begin{cases} \mathrm{d}X(t) = \lambda X(t)\mathrm{d}t + \mu X(t)\mathrm{d}W(t) \\ X(0) = X_0 \end{cases}$$

mit Konstanten $\lambda, \mu \in \mathbb{R}$. Die zugrundeliegende stochastische Differentialgleichung ist linear. Wie in Beispiel 5.2.7 gesehen, ist die allgemeine Lösung einer solchen durch

$$X(t) = \phi(t) \cdot \left(X_0 + \int_0^t \phi^{-1}(\tau)a(\tau)\mathrm{d}\tau + \int_0^t \phi^{-1}(\tau)b(\tau)\mathrm{d}W(\tau) \right)$$

mit

$$\phi(t) = \exp\left(\int_0^t A(\tau) - \frac{B(\tau)^2}{2}\mathrm{d}\tau + \int_0^t B(\tau)\mathrm{d}W(\tau)\right)$$

gegeben. Hier ist $A(t) \equiv \lambda, B(t) \equiv \mu$ sowie $a(t) \equiv 0$ und $b(t) \equiv 0$. Folglich gilt

$$\begin{aligned}\phi(t) &= \exp\left(\int_0^t (\lambda - \frac{\mu^2}{2})\mathrm{d}t + \int_0^t \mu\mathrm{d}W(t)\right) = \\ &= \exp\left(\left(\lambda - \frac{\mu^2}{2}\right)t + \int_0^t \mu\mathrm{d}W(t)\right).\end{aligned}$$

Wegen

$$\lim_{N\to\infty}\sum_{i=0}^{N-1}\mu\cdot\left(W\left((i+1)\frac{t}{N}\right) - W\left(i\frac{t}{N}\right)\right) = \lim_{N\to\infty}\mu\left(W(t) - W(0)\right) = \mu W(t)$$

folgt

$$\phi(t) = e^{\left(\lambda - \frac{\mu^2}{2}\right)t + \mu W(t)}$$

und somit

$$X(t) = X_0 \cdot e^{\left(\lambda - \frac{\mu^2}{2}\right)t + \mu W(t)}$$

In Abbildung 5.3.1 sind die Approximationen der Euler-Maruyama sowie die Werte der „exakten“ Lösung (besser gesagt, einer möglichen Trajektorie von dieser) auf dem Gitter $\{0, \Delta t, 2\Delta t, ..., T\}$ dargestellt. Die Simulation wurde mit MATLAB ausgeführt, als Zufallszahlengenerator wurde hierbei der Legacy MATLAB 5.0 normal generator mit seed 100 verwendet. Die Parameterwerte waren $\delta t = 2^{-8}$ sowie $R = 4$ der Endfehler betrug in etwa $|X(T) - X_T| \approx 0.6907$. Der Code befindet sich in Anhang E.2

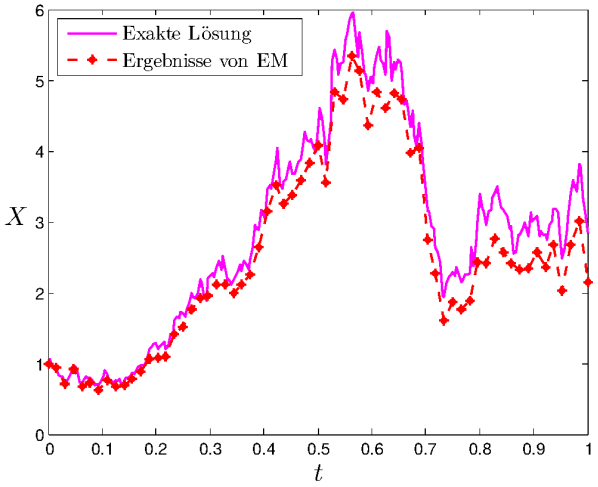


Abbildung 5.3.1

Im Falle gewöhnlicher Differentialgleichungen wurde die „Güte“ eines numerischen Approximationsverfahrens anhand der Konsistenzordnung ($\left\| \frac{y(t_{n+1})-y(t_n)}{h} - \frac{y_{n+1}-y_n}{h} \right\| \leq c \cdot h^p$) bestimmt (vgl. Kapitel 1.11). Bei stochastischen Differentialgleichung sind jedoch sowohl die „exakten“ Werte $X(\tau_n)$ als auch deren Approximationen X_n Zufallsvariablen. Um das Konzept der Konsistenzordnung übetragen zu können, bedarf es einer geeigneten Metrik.
Eine Möglichkeit hierfür ist der Erwartungswert der Differenz $|X_n - X(\tau_n)|$, das entsprechende Konzept ist das Konzept der ***starken Konvergenz***.

Definition 5.3.2:
Ein Verfahren hat ***starke Konvergenzordnung*** γ, falls es eine konstante C gibt, so dass

$$\mathbb{E}\left(|X_n - X(\tau_n)|\right) \leq C \cdot \Delta t^{\gamma}$$

für jedes feste $\tau_n = n\Delta t \in [0, T]$ und alle hinreichend kleinen Δt gilt.

Unter passenden Voraussetzungen ist die starke Konvergenzordnung der Euler-Maruyama-Methode $\gamma = \frac{1}{2}$.
Sei nun

$$e_{\Delta t}^{\text{strong}} := \mathbb{E}\left(|X_L - X(T)|\right), \qquad L\Delta t = T \tag{5.3.2}$$

der Fehler im Endzeitpunkt T. Gilt die Schranke aus 5.3.2 mit $\gamma = \frac{1}{2}$ für jeden festen Punkt in $[0, T]$, dann gilt sie insbesondere im Endpunkt, das heißt es gilt

$$e_{\Delta t}^{\text{strong}} \leq C\Delta t^{\frac{1}{2}} \tag{5.3.3}$$

für hinreichend kleine Zeitschritte Δt. Gilt in (5.3.3) näherungsweise Gleichheit, so ergibt Logarithmieren

$$\log e_{\Delta t}^{\text{strong}} \approx \log C + \frac{1}{2} \log \Delta t \tag{5.3.4}$$

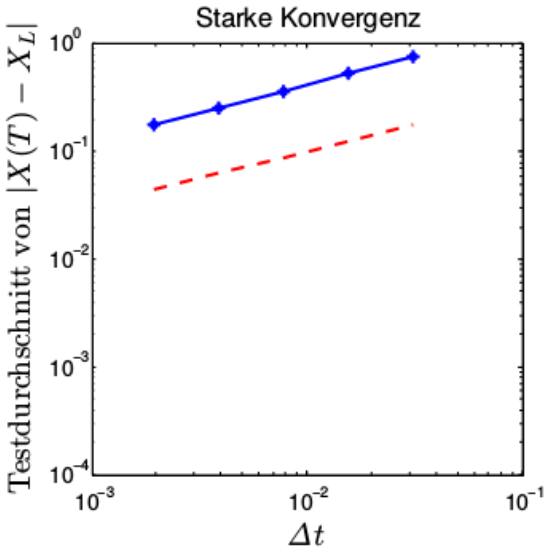


Abbildung 5.3.2

In Abbildung 5.3.2 ist der Fehler $e_{\Delta t}^{\text{strong}}$ (blau), gemittelt über 1000 Pfade gegen die gewählten Schrittweiten $\Delta t = 2^{p-1}\delta t$, $1 \leq p \leq 5$, $\delta t = 2^{-9}$ in einem doppelt-logarithmischen Plot aufgetragen. Die getestete stochastische Differentialgleichung war hierbei

$$\mathrm{d}X(t) = \lambda X(t)\mathrm{d}t + \mu X(t)\mathrm{d}W(t)$$

mit $\lambda = 2, \mu = 1$ sowie $X(0) = 1$. (Der zugehörige Code befindet sich in Anhang E.3.) Die Parallelität der Ergebniskurve zur roten Referenz-Geraden mit Steigung $\frac{1}{2}$ weist darauf hin, dass die annähernde Gleichheit aus (5.3.4) gilt, (5.3.3) also scharf ist. Neben der Diskretisierung existieren weitere Fehlerquellen, wie zum Beispiel:

- Stichprobenfehler: Fehler, der durch die Approximation des Erwartungswertes mittels des Mittelwerts der Stichproben entsteht.
- Verzerrung der Zufallszahlen: Fehler, die auf die Pseudo-Zufälligkeit des Zufallszahlengenerators zurückzuführen sind.
- Rundungsfehler: Genauigkeitsverlust durch Gleitkommadarstellung der Daten.

(Ersterer fällt hierbei wie $\frac{1}{\sqrt{M}}$, wobei M die Anzahl der Probepfade ist.)
Gemäß der ***Markov-Ungleichung*** ist für eine Zufallsvariable X mit endlichem Erwartungswert für jedes $a > 0$ die Wahrscheinlichkeit von $|X| \geq a$ nach oben durch $\frac{\mathbb{E}(|X|)}{a}$ beschränkt, es gilt also

$$P\left(|X| \geq a\right) \leq \frac{\mathbb{E}\left(|X|\right)}{a}.$$

Wählt man nun $a = \Delta t^{\frac{1}{4}}$, so ergibt sich für die Euler-Mayurama-Methode die Abschätzung

$$P\left(|X_n - X(\tau)| \geq \Delta t^{\frac{1}{4}}\right) \leq \frac{\overbrace{\mathbb{E}\left(|X_n - X(\tau)|\right)}^{\leq C\cdot\Delta t^{\frac{1}{2}}}}{a} \leq C\Delta t^{\frac{1}{4}}$$

beziehungsweise

$$P\left(|X_n - X(\tau)| < \Delta t^{\frac{1}{4}}\right) \geq 1 - C\Delta t^{\frac{1}{4}}.$$

Folglich ist für einen festen Punkt in $[0, T]$ der Fehler klein mit Wahrscheinlichkeit nahe 1.
Eine Alternative zum Konzept der starken Konvergenz ergibt sich, wenn man anstelle des Mittels der Fehler den Fehler der Mittel betrachtet:

Definition 5.3.3:
Ein Verfahren hat ***schwache Konvergenzordnung*** $\boldsymbol{\gamma}$, falls konstante C existiert, so dass für alle Funktionen p einer bestimmten Klasse von Funktionen

$$\left|\mathbb{E}\left(p(X_n)\right) - \mathbb{E}\left(p(X(\tau_n))\right)\right| \leq C \cdot \Delta t^{\gamma}$$

für jedes feste $\tau_n = n\Delta t \in [0, T]$ und alle hinreichend kleinen Δt gilt.

Die Funktionen p aus Definition müssen hierbei üblicherweise bestimmten Glattheits- und polynomiellen Wachstumsbedingungen genügen. Unter geeigneten Voraussetzungen kann gezeigt werden, dass die Euler-Maruyama-Methode schwache Konvergenzordnung $\gamma = 1$ hat. Zu Testzwecken sei p im Folgenden die Identität. Analog zu (5.3.2) sei

$$e_{\Delta t}^{\text{weak}} \coloneqq \left|\mathbb{E}\left(X_L\right) - \mathbb{E}\left(X(T)\right)\right|, \qquad L\Delta t = T \tag{5.3.5}$$

der schwache Fehler der Euler-Maruyama-Methode im Endpunkt. Aus Definition 5.3.3 folgt für $p(X) \equiv X$ mit $\gamma = 1$

$$e_{\Delta t}^{\text{weak}} \leq C\Delta t$$

für hinreichend kleine Δt. Da die schwache Konvergenz sich auf den Mittelwert der Lösung bezieht, kann in (5.3.1) für das Wiener-Inkrement $W(\tau_j+1)-W(\tau_j)$ in jedem Zeitschritt eine beliebige Stichprobe einer $\sqrt{\Delta t}\mathcal{N}(0,1)$-verteilten Zufallsvariablen gewählt werden. Ersetzt man das Inkrement durch eine Zufallsvariable $\sqrt{\Delta t}V_j$, wobei V_j die Werte $+1$ und -1 jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ annimmt, so bleibt die schwache Konvergenzordnung der Euler-Maruyama-Methode erhalten. ($\sqrt{\Delta t}V_j$ hat hierbei selben Mittelwert und selbe Varianz wie $\sqrt{\Delta t}\mathcal{N}(0,1)$.) Diese Variante der Euler-Maruyama-Methode heißt ***schwache Euler-Maruyama-Methode*** (weak Euler-Maruyama, WEM). In Abbildung 5.3.3 ist der Fehler $e_{\Delta t}^{\text{weak}}$ der Euler-Maruyama-

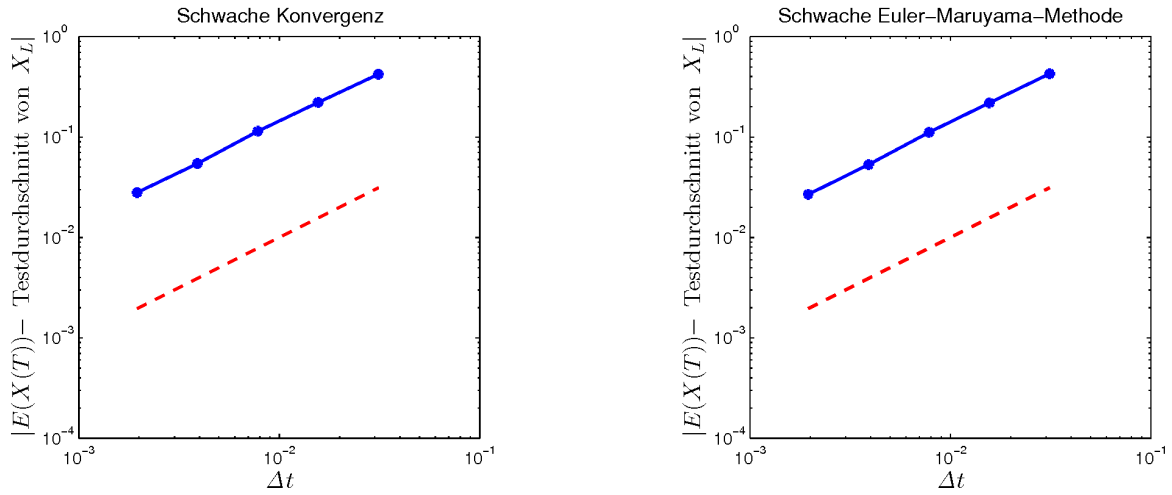


Abbildung 5.3.3: $e_{\Delta t}^{\text{weak}}$ für normale und schwache Euler-Maruyama-Methode

methode (blau) so wie je eine Referenzgerade mit Steigung 1 gegen die Schrittweite Δt aufgetragen. (Die getesteten Schrittweiten waren erneut $\Delta = 2^{p-1}\delta t$, $1 \leq p \leq 5$, mit $\delta t = 2^{-9}$). Das zugrundeliegende Anfangswertproblem war wie beim Test der starken Konvergenz von der Form

$$\begin{cases} \mathrm{d}X(t) = \lambda X(t)\mathrm{d}t + \mu X(t)\mathrm{d}W(t) \\ X(0) = X_0 \end{cases},$$

die problemspezifischen Parameter waren $X_0 = 1$, $\lambda = 2$, $\mu = \frac{1}{10}$. Der Testdurchschnitt wurde über 50000 Pfade gemittelt. (Der verwendete Code befindet sich in Anhang E.4.)

5.3.3 Stabilität

In diesem Kapitel soll die Stabilität stochastischer Differentialgleichungen untersucht werden. Wie bei gewöhnlichen Differentialgleichungen steht hierbei das Verhalten von Lösungen für $t \to \infty$ im Fokus (vgl. Kapitel 1.8) . Im allgemeinen Fall gilt für die Konvergenzschranken aus 5.3.2 und 5.3.3, dass die Konstante C mit der Endzeit T unbeschränkt wächst.
Sei nun erneut die lineare stochastische Differentialgleichung

$$\mathrm{d}X(t) = \lambda X(t)\mathrm{d}t + \mu X(t)\mathrm{d}W(t) \tag{5.3.6}$$

mit Anfangsbedingung $X(0) = X_0$ gegeben, wobei λ und μ komplexe Zahlen sind. Im deterministischen Fall, dass heißt $\mu = 0$, ist die Nulllösung $X(t) \equiv 0$ des Systems asymptotisch stabil genau dann, wenn $\operatorname{Re}\lambda < 0$ ist (vgl. Kapitel 1.8, Satz 1.8.2). Gemäß der der ursprünglichen Definition bedeutet Stabilität der Nulllösung, dass für alle Lösungen $X(t)$ mit $X(0) = X_0$ und $\|X_0 - 0\| \leq \delta$ für ein δ $\lim_{t\to\infty} \|X(t)\| = 0$ oder äquivalent $\lim_{t\to\infty} X(t) = 0$ gilt. Im Falle $\mu \neq 0$ ist diese Definition jedoch nicht ohne weiteres übertragbar, da erst eine geeignete Definition von

$\lim_{t\to\infty} X(t) = 0$ gefunden werden muss. Stattdessen sollen im Folgenden zwei der gebräuchlichsten Stabilitätsmaße vorgestellt werden:

Definition 5.3.4:
Gegeben sei eine stochastische Differentialgleichung mit Anfangsbedingung $X(0) = X_0$, wobei $X_0 \neq 0$ mit Wahrscheinlichkeit 1 gelte. Die Lösung $X(t)$ heißt ***stabil im quadratischen Mittel***, falls

$$\lim_{t\to\infty} \mathbb{E}\left(|X(t)|^2\right) = 0$$

gilt.
Sie heißt ***asymptotisch stabil***, falls

$$\lim_{t\to\infty} |X(t)| = 0 \qquad \text{mit Wahrscheinlichkeit 1}$$

gilt.

Ähnlich wie bei gewöhnlichen Differentialgleichungen lässt für die stochastische Differentialgleichung (5.3.6) das Stabilitätsverhalten aus den Parametern λ und μ genau bestimmen:

Satz 5.3.5:
Die Lösungen von (5.3.6) sind

- stabil im quadratischen Mittel genau dann wenn

$$\text{Re}\,(\lambda) + \frac{1}{2}|\mu|^2 < 0$$

- asymptotisch stabil genau dann wenn

$$\text{Re}\left(\lambda - \frac{1}{2}\mu^2\right) < 0$$

Hieraus folgt unmittelbar, dass aus Stabilität im quadratischen Mittel auch asymptotische Stabilität folgt, die umgekehrte Implikation gilt jedoch nicht (als Gegenbeispiel genügen $\lambda = \frac{1}{4}$ und $\mu = 1$).
Im Folgenden soll untersucht werden, unter welchen Bedingungen die Approximationen der Euler-Maruyama-Methode an die Lösung von (5.3.6) analoge Stabilitätskriterien erüllen. Seien hierzu λ und μ dergestalt, dass die exakten Lösungen von (5.3.6) stabil im quadratischen Mittel oder asymptotisch stabil sind. Für die Stabilität im quadratischen Mittel ergibt sich aus (5.3.1) mit $f(t, X(t)) = \lambda X(t)$, $\sigma(t, X(t)) = \mu X(t)$

$$X_{i+1} = X_i + \lambda X_i \Delta t + \mu X_i \left(W(\tau_{i+1}) - W(\tau_i)\right)$$

und somit

$$\begin{aligned} |X_{i+1}|^2 &= |(1 + \Delta t\lambda)X_i + \mu X_i \left(W(\tau_{i+1}) - W(\tau_i)\right)|^2 = \\ &= |1 + \Delta t\lambda|^2 X_i + 2\text{Re}\left((1 + \Delta t\lambda) X_i\right) \cdot \left(W(\tau_{i+1}) - W(\tau_i)\right) + \\ &\quad \left|\mu^2\right| \left(W(\tau_{i+1}) - W(\tau_i)\right)^2 \end{aligned}$$

Wegen $\mathbb{E}\left(W(\tau_{i+1}) - W(\tau_i)\right) = 0$ sowie

$$\mathbb{E}\left(\left(W(\tau_{i+1}) - W(\tau_i)\right)^2\right) = \text{Var}\left(W(\tau_{i+1}) - W(\tau_i)\right) + \mathbb{E}\left(W(\tau_{i+1}) - W(\tau_i)\right)^2 = \Delta t$$

folgt aus der Linearität des Erwartungswertes

$$\mathbb{E}\left(|X_{i+1}|^2\right) = \left(|1+\Delta t\lambda|^2 + \Delta t\,|\mu|^2\right)\mathbb{E}\left(|X_i|^2\right)$$

und somit induktiv

$$\mathbb{E}\left(|X_i|^2\right) = \left(|1+\Delta t\lambda|^2 + \Delta t\,|\mu|^2\right)^i\mathbb{E}\left(|X_0|^2\right)$$

Es gilt folglich

$$\lim_{i\to\infty}\mathbb{E}\left(|X_i|^2\right) = 0 \Leftrightarrow |1+\Delta t\lambda|^2 + \Delta t\,|\mu|^2 < 1$$

Im Falle der asymptotischen Stabilität ergibt sich Mittels starkem Gesetz der großen Zahlen und dem Gesetz des iterierten Logarithmus

$$\lim_{i\to\infty}|X_i| = 0 \quad \text{mit Wahrscheinlichkeit } 1 \Leftrightarrow \mathbb{E}\left(\log\left|1+\Delta t\lambda+\sqrt{\Delta t}\mu\mathcal{N}(0,1)\right|\right) < 0.$$

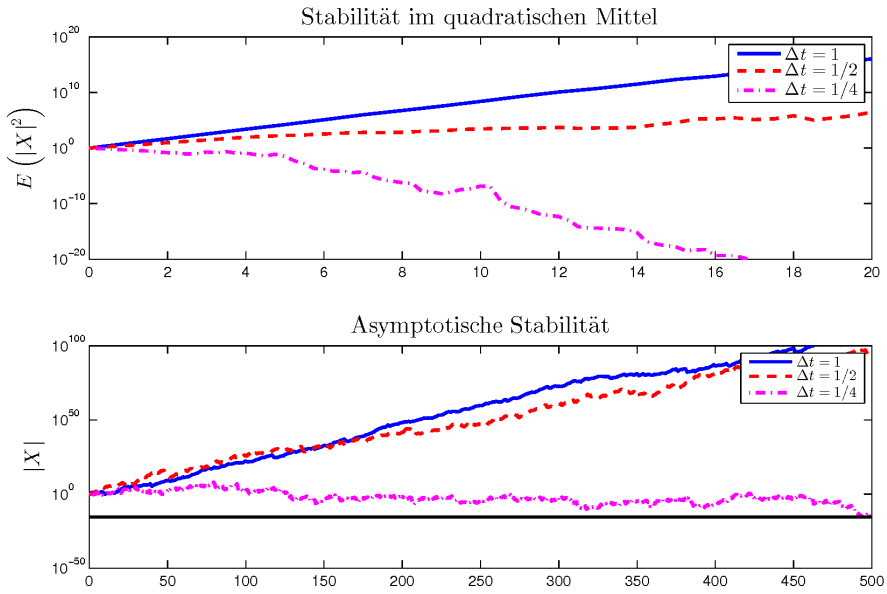


Abbildung 5.3.4

In Abbildung 5.3.4 sind die Ergebnisse numerischer Tests dargestellt. Für die Stabilität im quadratischen Mittel wurde $\lambda = -3, \mu = \sqrt{3}$ und $X_0 = 1$ gewählt, der Erwartungswert $\mathbb{E}\left(|X(t)|^2\right)$ wurden für jede der untersuchten Schrittweiten $\Delta t \in \left\{\frac{1}{4}, \frac{1}{2}, 1\right\}$über 50000 Pfade gemittelt. Für den Test der asymptotische Stabilität wurden drei einzelne Pfade mit denselben Schrittweiten gebildet, die Parameter wurden hier auf $\lambda = \frac{1}{2}$ und $\mu = \sqrt{6}$ gesetzt. (Der entsprechende Matlabcode befindet sich in Anhang E.5). Die schwarze Linie zeigt die Maschinengenauigkeit an.

5.4 Ausblick: Stückweise deterministische Prozesse

Dieses Kapitel soll einen abschließenden, kurzen Ausblick auf das Forschungsfeld der stückweise deterministischen Markov-Prozesse darstellen, es basiert im Wesentlichen auf [32]. Einfach ausgedrückt, handelt es sich bei einem stückweise deterministischen Markov-Prozess um einen Prozess, dessen Änderung im Wesentlichen deterministisch ist, jedoch zu zufälligen (diskreten) Zeitpunkten die zugrundeliegende deterministische Struktur zufällig ändert. Zur Veranschaulichung sei folgender Modellwettbewerb gegeben:

Beispiel:

Zwei Sportler (S_1, S_2)treten in einem Rennen nacheinander gegeneinander an. Das Rennen ist hierbei wie folgt aufgebaut:

- Jeder Sportler hat die selbe Grundzeit T zur Verfügung, in der er soweit wie möglich kommen muss.
- Für jeden Sportler S ($S \in \{S_1, S_2\}$) wird eine Folge von zufälligen Zeitpunkten $0 < t_S^{(1)} < t_S^{(2)} < ... < t_S^{(n)} < ...$ bestimmt, welche abbricht, sobald der Zeitpunkt über T hinausgeht.
- Gemessen wird die zurückgelegte Strecke $X_S(t)$ vom Start $X_S(0) = 0$.
- Jeder Sportler S beginnt im Start und fängt zum Zeitpunkt $t_S^{(0)}$ an, mit einer Vespa zu fahren. Ab dem Zeitpunkt $t_S^{(0)}$ wird auch die Zeit gemessen.
- Zum Zeitpunkt $t_S^{(1)}$ unterbricht Sportler S seinen Lauf und spielt ein Glücksspiel mit zwei möglichen Ergebnissen: Vespa (V) oder Ferrari (F). (Die Ergebnismenge ist also $\Omega = (V, F)$. Nachdem das Ergebnis feststeht, setzt der Sportler den Weg entsprechend dem Ergebnis des Glücksspiels entweder auf der Vespa oder mit dem Ferrari fort. Für die Dauer des Glücksspiels sowie (ggf.) des Wechsels wird die Zeitmessung ab dem Zeitpunkt $t_S^{(1)}$ unterbrochen. Sobald der Sportler seinen Weg fortsetzt, wird auch die Zeit weiter gemessen. Zum Zeitpunkt $t_S^{(2)}$ unterbricht der Sportler erneut das Rennen, wiederholt das Glücksspiel und setzt seinen anschließend seinen Weg entsprechend dem Ergebnis fort. Analog wird zu den restlichen Zeitpunkten $t_S^{(i)}, i = 3, ..., n_S$ verfahren.

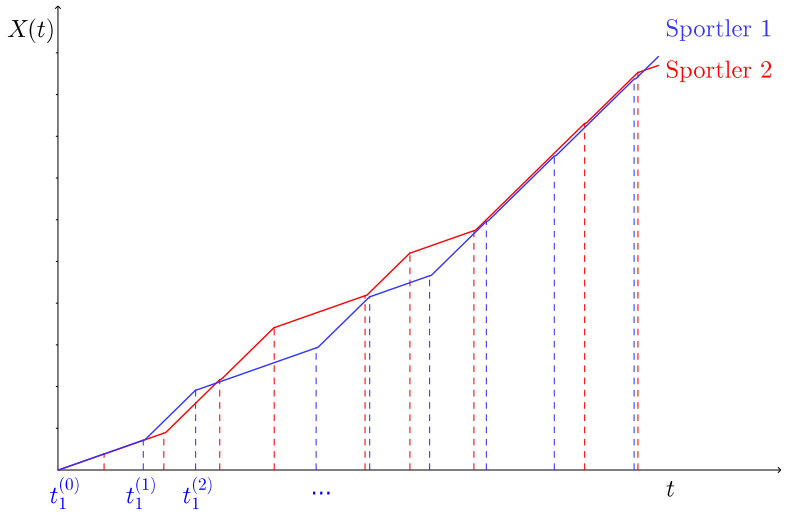


Abbildung 5.4.1

Die für das Wettbewerbsergebnis interessante Größe ist nun die zurückgelegte Strecke $X(t)$. Diese ist zufällig bestimmt, wie es im Einführungsbeispiel (random walk) der Fall war, jedoch fließt der Zufall auf eine ganz andere Art und Weise ein: Im (diskreten) Einführungsbeispiel, ebenso wie beim kontinuierlichen Pendant (Brownsche Bewegung), war die Ortsänderung zu jedem Zeitpunkt zufällig bestimmt. Im vorliegenden Modell hingegen ist die Ortsänderung zwischen zwei Zeitpunkten $t_S^{(i)}$ und $t_S^{(i+1)}$ deterministisch bestimmt. Ist die Art des im entsprechenden Zeitabschnitts verwendeten Fahrzeugs erstmal festgelegt, so sind Ortsänderung und zurückgelegte Strecke durch Beschleunigung, maximal möglicher Geschwindigkeit,

Luftwiderstand und Reibung ebenso festgelegt. Für das vereinfachte Modell geradliniger Bewegungen ohne Reibung und Luftwiderstand ist die Situation in Abbildung 5.4.1 dargestellt. Zufällig bestimmt ist die Art der Dynamik sowie der Zeitpunkt des Wechsels: Es gilt $\dot{X}_S(t) = f_{S,\omega(t)}(t, X_S)$, wobei $\omega(t) \in \Omega$ die Fortbewegungsart des Sportlers S zum Zeitpunkt t ist, welche wiederum vom Ergebnis des zum letzten Zeitpunkt durchgeführten Glücksspiels abhängt.

Dieses Beispiel soll nun verallgemeinert werden. Hierzu sei die Zustandsfunktion $X(t), X : [t_0, \infty) \to \Omega, \Omega \subseteq \mathbb{R}^d$ durch folgende Eigenschaften definiert:

a) Ihre Dynamik ist für $t \in [t_0, \infty)$ durch

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = A_{\mathcal{S}(t)}(X) \tag{5.4.1}$$

gegeben. $\mathcal{S} : [t_0, \infty) \to \mathbb{S}$ ist hierbei ein Markov-Prozess mit diskreten Zuständen $\mathbb{S} = \{1, ..., S\}$, welcher in c) und d) genauer beschrieben wird. Ist ein $s \in \mathbb{S}$ gegeben, so befindet sich die Dynamik in Zustand s, gesteuert durch die Funktion $A_s : \Omega \to \mathbb{R}^d$, welche zur Funktionenmenge $\{A_1, ..., A_S\}$ gehört. Fordert man, dass für alle $s \in \mathbb{S}$ die Funktion A_s Lipschitz-stetig ist, so ist für festes s die Lösung $X(t)$ eindeutig und beschränkt.

b) Die Zustandsfunktion genügt der Anfangsbedingung $X(t_0) = X_0 \in \Omega$ und ihre Dynamik befindet sich im Startzustand $s_0 = \mathcal{S}(t_0)$.

c) $\mathcal{S}(t)$ wird durch exponentielle Dichtefunktionen $\psi_s : \mathbb{R}^+ \to \mathbb{R}^+$ für die übergangsereignisse charakterisiert. Hierbei sei für jedes feste $s \in \mathbb{S}$ ψ_s gegeben durch

$$\psi_s(t) = \mu_s e^{-\mu_s t}, \quad \text{mit } \int_0^\infty \psi_s(t)\mathrm{d}t = 1. \tag{5.4.2}$$

Bezeichnet T die zufällige Zeitspanne, die sich der Prozess im Zustand s befindet, so ist die Wahrscheinlichkeit, dass T kleiner oder gleich einer bestimmten Zeit T ist, gegeben durch

$$P(T \leq t) = \int_0^t \mu_s e^{-\mu_s \tau}\mathrm{d}\tau = 1 - e^{-\mu_s t},$$

die Wahrscheinlichkeit, dass $T > t$ ist, durch

$$P(T > t) = e^{-\mu_s \tau}.$$

Anders ausgedrückt entspricht also die Zeit, für die das System in einem Zustand s bleibt, der Wartezeit zwischen zwei aufeinanderfolgenden Ereignissen eines Poisson-Prozesses.

d) Tritt ein übergangsereignis auf, so wechselt das System ünverzüglich von gegenwärtigen Zustand $j \in \mathbb{S}$ in einen neuen Zustand $i \in \mathbb{S}$ mit Wahrscheinlichkeit q_{ij}. $\mathbb{S}(t)$ kann folglich durch die übergangsmatrix $Q = [q_{ij}]$ beschrieben werden, wobei

$$0 \leq q_{ij} \leq 1 \forall i, j \in \mathbb{S} \tag{5.4.3}$$

und $\sum_{i=1}^{S} q_{ij} = 1 \forall j \in \mathbb{S}$ gilt.

Der durch c) und d) definierte Markov-Erneuerungs-Prozess $\mathcal{S}(t)$ erzeugt eine zeitliche Abfolge von Übergangsereignissen, oder Sprungzeiten, $(t_0, t_1, ..., t_k, t_{k+1}, ...)$, und entsprechenden Zuständen $(s_0, s_1, ..., s_k, s_{k+1}, ...)$. Der Zustandsraum des stückweise deterministischen Prozesses besteht aus S disjunkten Kopien von $\mathbb{R}^d$. (Anders ausgedrückt ist er ein Teilraum von $\underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{S \text{ mal}}$.) Das Vektorfeld der j. Komponente ist hierbei durch A_j gegeben. Zu einer Sprungzeit t_k tritt nun ein Wechsel zu einer anderen Komponenten A_i mit zufälligem i ein beziehungsweise kann eintreten, je nachdem, ob $q_{jj} = 0$ oder $q_{jj} > 0$ ist. $X(t)$ ist hierbei durchgehend stetig, insbesondere in den Sprungzeitpunkten t_0, t_1, usw.

6 Von gewöhnlichen zu partiellen Differentialgleichungen: Fokker-Planck und Liouville

In diesem Kapitel sollen ***partielle Differentialgleichungen*** (englisch partial differential equation, PDE) kurz eingeführt werden. Dies sind Gleichungen, welche eine von mehreren Variablen abhängige Funktion u zusammen mit partiellen Ableitungen von u nach - sinnvollerweise - mindestens zwei Variablen enthält. Als ***Ordnung*** der Differentialgleichung wird hierbei die höchste Ordnung der enthaltenen partiellen Ableitungen angesehen.
Anstelle jedoch die Wärmeleitungsgleichung als Einstieg zu nehmen, wie es in der Literatur häufig der Fall ist, soll der Bogen hierbei über stochastische Differentialgleichungen und die Focker-Planck-Gleichung geschlagen werden.

Hierzu sei die zufällige Bewegung eines Partikels betrachtet, welches während seiner Vorwärtsbewegung, ähnlich des betrunkenen Mannes im Einführungsbeispiel zu Kapitel 5, zufällig einen Sprung nach links ($-\Delta x$) oder rechts (Δx) in einem Zeitintervall Δt macht. Es sei auch zulässig, dass der Partikel keinen Seitwärtssprung macht. Bezeichne $X(t)$ den Ort, an dem sich das Partikel zum Zeitpunkt t befindet. Im Gegensatz zum erwähnten Beispiel soll diesmal nicht der Ort betrachtet werden, an dem das Partikel nach einer bestimmten Zeit ist, sondern die Wahrscheinlichkeit, dass das Partikel zu einer bestimmten Zeit an einem bestimmten Ort ist. (Andere Interpretationsmöglichkeit: Der Anteil an der gesamten Partikelmenge, der sich zu einem festen Zeitpunkt an einem bestimmten Ort befindet.) Bezeichne

- $p = p(x,t)$ die Wahrscheinlichkeit der Bewegung $x \to x + \Delta x$ zum Zeitpunkt t
- $q = q(x,t)$ die Wahrscheinlichkeit der Bewegung $x \to x - \Delta x$ zum Zeitpunkt t
- $f = f(x,t)$ die Dichte des stochastischen Prozesses

Aus ersteren beiden folgt, dass die Wahrscheinlichkeit der „Bewegung“ $x \to x$ gleich der Differenz $1 - p(x,t) - q(x,t)$ ist, immerhin muss die Summe der Wahrscheinlichkeiten über alle möglichen Bewegungen 1 ergeben. $f(x,t)\Delta x$ beschreibt (näherungsweise) die Wahrscheinlichkeit, dass der Ort des Partikels zum Zeitpunkt t im Intervall $\left[x - \frac{1}{2}\Delta x, x + \frac{1}{2}\Delta\right]$ liegt. (Dies ergibt sich durch die Mittelpunktsformel, angewandt auf $P(x - \frac{\Delta x}{2} \leq X \leq x + \frac{\Delta x}{x}, t) = \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} f(x,t)\mathrm{d}x$.) Geht man davon aus, dass die Bewegung eines Teilchens unabhängig von der vorherigen Bewegung ist, so berechnet sich die Wahrscheinlichkeit $f(x, t+\Delta t)\Delta x$ mittels

$$\begin{aligned} f(x,t+\Delta t)\Delta x =& p(x-\Delta x,t)f(x-\Delta x,t)\Delta x + q(x+\Delta x,t)f(x+\Delta x,t) \\ &+ (1 - p(x,t) - q(x,t))\, f(x,t)\Delta x \end{aligned}$$

Die Wahrscheinlichkeitsdichte an x zum Zeitpunkt $t + \Delta t$ lässt sich für kleine $\Delta t, \Delta x$ durch mittels

$$f(x,t+\Delta t) = p(x-\Delta x,t)f(x-\Delta x,t) + q(x+\Delta x,t)f(x+\Delta x) + (1 - p(x,t) - q(x,t))\, f(x,t)$$

(näherungsweise) berechnen. Sei nun $p(t,x) + q(t,x) = 1$, d. h. der Partikel macht auf jeden Fall eine Seitwärtsbewegung. Sind f, p, q hinreichend glatt, ergibt Taylor-Entwicklung

- $f(x,t+\Delta t) - f(x,t) = f(x,t) + \Delta t \frac{\partial}{\partial t} f(x,t) + o(\Delta t) - f(x,t) = \Delta t \frac{\partial}{\partial t} f(x,t) + o(\Delta t)$
- $p(x-\Delta x,t) = p(x,t) - \Delta x \frac{\partial}{\partial x} p(x,t) + \Delta x^2 \frac{1}{2} \frac{\partial^2}{\partial x^2} p(x,t) + o(\Delta x^2)$

- $f(x-\Delta x,t) = f(x,t) - \Delta x \frac{\partial}{\partial x} f(x,t) + \Delta x^2 \frac{1}{2} \frac{\partial^2}{\partial x^2} f(x,t) + o(\Delta x^2)$
- $f(x+\Delta x,t) = f(x,t) + \Delta x \frac{\partial}{\partial x} f(x,t) + \Delta x^2 \frac{1}{2} \frac{\partial^2}{\partial x^2} f(x,t) + o(\Delta x^2)$
- $q(x+\Delta x,t) = q(x,t) + \Delta x \frac{\partial}{\partial x} q(x,t) + \Delta x^2 \frac{1}{2} \frac{\partial^2}{\partial x^2} q(x,t) + o(\Delta x^2)$

Somit gilt

$$p(x-\Delta x,t)f(x-\Delta x,t) = p(x,t)f(x,t) - \Delta x \left(\frac{\partial}{\partial x} p(x,t) \cdot f(x,t) + p(x,t) \cdot \frac{\partial}{\partial x} f(x,t) \right)$$
$$+ \frac{1}{2} \Delta x^2 \left(\frac{\partial^2}{\partial x^2} p(x,t) \cdot f(x,t) - 2 \frac{\partial}{\partial x} p(x,t) \cdot \frac{\partial}{\partial x} f(x,t) + p(x,t) \cdot \frac{\partial^2}{\partial x^2} f(x,t) \right) + o(\Delta x^2)$$

sowie

$$q(x+\Delta x,t)f(x+\Delta x,t) = q(x,t)f(x,t) + \Delta x \left(\frac{\partial}{\partial x} q(x,t) \cdot f(x,t) + q(x,t) \cdot \frac{\partial}{\partial x} f(x,t) \right)$$
$$+ \frac{1}{2} \Delta x^2 \left(\frac{\partial^2}{\partial x^2} q(x,t) \cdot f(x,t) + 2 \frac{\partial}{\partial x} q(x,t) \cdot \frac{\partial}{\partial x} f(x,t) + q(x,t) \cdot \frac{\partial^2}{\partial x^2} f(x,t) \right) + o(\Delta x^2).$$

Eingesetzt in die Gleichung $f(x,t+\Delta t) = p(x-\Delta x,t)f(x-\Delta x,t) + q(x+\Delta x,t)f(x+\Delta x,t)$ ergibt dies:

$$\Delta t \frac{\partial}{\partial t} f(x,t) + o(\Delta t) = - \left(\frac{\partial}{\partial x} (p-q)(x,t) \Delta x f(x,t) + (p-q)(x,t) \Delta x \frac{\partial}{\partial x} f(x,t) \right)$$
$$+ \left(\frac{\partial^2}{\partial x^2} (p+q)(x,t) \Delta x^2 f(x,t) + 2 \frac{\partial}{\partial x} (p+q)(x,t) \Delta x \frac{\partial}{\partial x} f(x,t) + (p+q) \frac{\partial^2}{\partial x^2} \right) + o(\Delta x^2)$$

beziehungsweise

$$\frac{\partial}{\partial t} f(x,t) = - \left(\frac{\partial}{\partial x} (p-q)(x,t) \frac{\Delta x}{\Delta t} f(x,t) + (p-q)(x,t) \frac{\Delta x}{\Delta t} \frac{\partial}{\partial x} f(x,t) \right)$$
$$+ \frac{1}{2} \left(\frac{\partial^2}{\partial x^2} (p+q)(x,t) \frac{\Delta x^2}{\Delta t} f(x,t) + 2 \frac{\partial}{\partial x} (p+q)(x,t) \frac{\Delta x}{\Delta t} \frac{\partial}{\partial x} f(x,t) \right. \tag{6.0.1}$$
$$\left. + (p+q)(x,t) \frac{\Delta x^2}{\Delta t} \frac{\partial^2}{\partial x^2} f(x,t) \right) + O(\Delta t) + O(\frac{\Delta x^3}{\Delta t})$$

Es gelte $\lim\limits_{\Delta x, \Delta t \to 0} \frac{\Delta x}{\Delta t} \left(p(x,t) - q(x,t) \right) = \mu(x,t)$ und $\lim\limits_{\Delta x, \Delta t \to 0} \frac{\Delta x^2}{\Delta t} \left(p(x,t) + q(x,t) \right) = \sigma^2(x,t)$. Aus (6.0.1) wird durch Grenzwertbildung $\lim\limits_{\Delta x, \Delta t \to 0}$

$$\frac{\partial}{\partial t} f(x,t) + \frac{\partial}{\partial x} \left(\mu(x,t) f(x,t) \right) - \frac{1}{2} \frac{\partial^2}{\partial x^2} \left(\sigma^2(x,t) f(x,t) \right) = 0 \tag{6.0.2}$$

Eine Gleichung dieser Form wird ***Fokker-Planck-Gleichung*** genannt. $\mu(x,t)$ heißt ***Drift***, $\sigma^2(x,t)$ heißt ***Diffusion***. Startet das Partikel zum Zeitpunkt $t = 0$ im Ort $x = x_0$, so ist die Anfangsbedingung durch

$$f(x,0) = \delta(x - x_0)$$

gegeben. Für diese Gleichung gibt es einige Spezialfälle:

1. Gilt $\mu = 0$ und $0 \neq \sigma(x,t) =$ konst., so geht die Fokker-Planck-Gleichung über in die eingangs erwähnte Wärmeleitungsgleichung

$$\frac{\partial}{\partial t} f(x,t) = \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} f(x,t)$$

2. Im Falle $\sigma(x,t) \equiv 0, \mu(x,t) = \text{konst.}$ führt dies auf die Transportgleichung

$$\frac{\partial}{\partial t} f(x,t) + \mu \frac{\partial}{\partial x} f(x,t) = 0$$

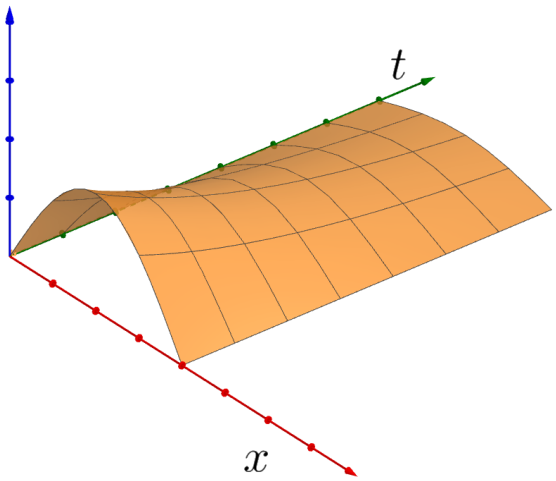


Abbildung 6.0.1: Wärmeleitungsgleichung

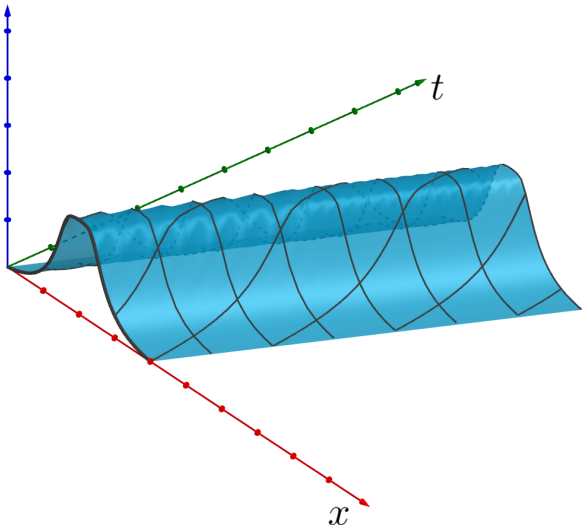


Abbildung 6.0.2: Transportgleichung

Bemerkung: *Im Falle einer allgemeinen partiellen Differentialgleichung sind Anfangs- oder Randbedingungen nicht durch einzelne Werte, sondern durch Funktionen gegeben. Welche (ggf. Kombination von) Bedingungen man wählen kann, sodass eine Lösung existiert, hängt hierbei jedoch von der Art der Gleichung ab. Für die abgebildeten Lösungen wurden beispielsweise im Falle der Wärmeleitungsgleichung die Bedingungen* $f(0,t) = f(2,t) = 0$ *sowie* $f(x,0) = \sin(x \cdot \frac{\pi}{2})$ *gestellt, im Falle der Transportgleichung* $u(x,0) = \sin(x \cdot \frac{\pi}{2}) \cdot \left(1 - \exp\left(-\frac{1}{10(x-1)^2}\right)\right)$.

Anhang A: Anhang A: Matlab-Codes zu Kapitel 2.4 (Populationsmodelle)

Anhang A.1: Anhang A.1: Simulation 2.4.1 (Klassisches Lotka-Volterra-Modell)

Code für die Berechnung derÄnderung der Populationsgrößen N, P zu einem Zeitpunkt t in Abhängigkeit der beiden Populationsgrößen

```
%Berechnet das Populationswachstum
%zum Zeitpunkt t
%gemaess dem klassischen Lotka-Volterra-Modell
%    dN/dt = N(a-bP)
%    dP/dt = P(cN-d)
function xdot = lotkaVolt(t,x,a,b,c,d)
    %ggf. Korrektur (-> nicht negative Populationsgroessen)
    if x(1) < 0
        x(1) = 0;
    end
    if x(2) < 0
        x(2) = 0;
    end
    xdot = [x(1)*(a-b*x(2));
            x(2)*(c*x(1)-d)];
end
```

Datei: lotkaVolt.m

Code für die eigentliche Simulation

```
%Simulation des Lotka-Volterra-Modells
clear all; close all;

t = [];                 %Speicher fuer die Zeitpunkte
x = [];                 %Speicher fuer die Populationsgroessen,
                        % x=[N,P]

a = 0.8;                % Wachstumsrate der Beutepopulation
b = 0.02;               % Jeder Raeuber toetet 2% der
                        % Beutepopulation (durchschnittlich)
c = 0.001;              % Wachstumsrate der Raeuberpopulation
                        % pro vorhandenem Beutetier
d = 0.3;                % Todesrate der Raeuberpopulation

t0 = 0;                 %Startzeitpunkt
T = 100;                %Modellierung ueber 100 Jahre
x0 = [1000;100];        %Anfangsgroessen: 1000 Beutetiere,
                        %100 Raeuber

%Verwendung des vorimplementierten Dormand-Prince-Verfahrens
[t,x] =  ode45(@(t,x) lotkaVolt(t,x,a,b,c,d),[t0 T],x0);

%ggf. Korrektur de letzten Werte
if x(1,end) < 0
    x(1,end) = 0;
end
if x(2,end) < 0
    x(2,end) = 0;
end

%Grafische Ausgabe
plot(t,x(:,1),'-b',t,x(:,2),'-r','LineWidth',1.5);
xlabel('Zeit in Jahren','Fontsize',14',....
        'Interpreter','LaTex');
ylabel('Populationsgr{\"o}{\ss}e in Individuen',...
        'Fontsize',14','Interpreter','LaTex');
legend({'Beute','R{\"a}uber'},'Fontsize',14',...
        'Interpreter','LaTex');
```

Datei: LotkaVoltSim.m

Anhang A.2: Anhang A.2: Simulationen 2.4.2 und 2.4.3 (Modifiziertes Lotka-Volterra-Modell)

```
%Berechnet das Populationswachstum
%zum Zeitpunkt t
%gemaess dem  modifizierten Lotka-Volterra-Modell
%    dN/dt = N(r(1-N/K)-kP/(N+D))
%    dP/dt = P(s(1-hP/N))
function xdot = lotkaVoltMod(t,x,D,K,h,k,s,r)
    %ggf. Korrektur (-> nicht negative Populationsgroessen)
    if x(1) < 0
        x(1) = 0;
    end
    if x(2) < 0
        x(2) = 0;
    end
    xdot = [x(1)*(r*(1-x(1)/K)-k*x(2)/(x(1)+D));
            x(2)*s*(1-h*x(2)/x(1))];
    if x(1) == 0
        %Keine Beute -> Aussterben der Raeuber
        %(Bis jetzt waere xdot(2)=+/-infty)
        xdot(2) = -x(2);
    end
end
```

Datei: lotkaVoltMod.m

Code für die eigentliche Simulation (Hier für Simulation 2.4.3, für die Ergebnisse von Simulation 2.4.2 genügt es, die entsprechenden Zeilen von Kommentar- in Code-Zeilen umzuwandeln.)

```
%Simulation des modifizierten Lotka-Volterra-Modells
clear all; close all;

t = [];                 %Speicher fuer die Zeitpunkte
x = [];                 %Speicher fuer die Populationsgroessen,
                        % x=[N,P]

K = 2000; D = 200;
k = 0.02; r = 0.8;
s = 0.5; h = 5/3;

t0 = 0;
T = 100;                %Wachstum ueber 100 Jahre
x0 = [1000;100];        %Anfangsgroessen: 1000 Beutetiere,
                        %100 Raeuber
% Werte fuer Grenzzyklus
% K = 5000;
% D = 10;
% k = 0.44;
% r = 0.4;
% s = 0.32;
% h = 1.1;
% T = 500;

%Verwendung des vorimplementierten Dormand-Prince-Verfahrens
```

```
[t,x] = ode45(@(t,x) lotkaVoltMod(t,x,D,K,h,k,s,r) ,[t0 T],x0);

%ggf. Korrektur de letzten Werte
if x(1,end) < 0
    x(1,end) = 0;
end
if x(2,end) < 0
    x(2,end) = 0;
end

%Grafische Ausgabe
figure(1)
plot(t,x(:,1),'-b',t,x(:,2),'-r','LineWidth',1.5);
xlabel('Zeit in Jahren','Fontsize',14,...
        'Interpreter','LaTex');
ylabel('Populationsgr{\"o}{\ss}e in Individuen',...
        'Fontsize',14,'Interpreter','LaTex');
legend({'Beute','R{\"a}uber'},'Fontsize',14,...
        'Interpreter','LaTex');

%Berechnung des Phasenportraets
[X,Y] = meshgrid(50:150:3500,50:150:2000);
X = X';
Y = Y';
M = length(X(:,1));
N = length(X(1,:));
U = zeros(M,N);
V = zeros(M,N);
l = 0;
for m = 1:M
    for n = 1:N
        xdot = lotkaVoltMod(0,[X(m,n);Y(m,n)],D,K,h,k,s,r);
        l = sqrt(xdot(1)^2+xdot(2)^2);
        U(m,n) = xdot(1)*100/l; %->Skalierung
        V(m,n) = xdot(2)*100/l; %->Skalierung
    end
end

%Grafische Ausgabe des Phasenportraets
figure(2)
plot(x(:,1),x(:,2),'Color',[1 .5 0],'LineWidth',1.5);
hold on
quiver(X,Y,U,V,'AutoScale','off');
axis([0 3500 0 2000])
xlabel('Anzahl Beutetiere','Fontsize',14,...
        'Interpreter','LaTex');
ylabel('Anzahl R{\"a}uber','Fontsize',14,...
        'Interpreter','LaTex');
```

Datei: LotkaVoltModSim.m

Anhang A.3: Anhang A.3: Simulationen 2.4.4 und 2.4.5 (Kompetitives Modell)

Code für die Berechnung der Änderung der Populationsgrößen N_1, N_2 zu einem Zeitpunkt t in Abhängigkeit der beiden Populationsgrößen

```
%Berechnet das Populationswachstum
%zum Zeitpunkt t
%gemaess kompetitiven Modell
%    dN1/dt = r1*N1*(1-N1/K1-b12*N2/K2)
%    dN1/dt = r2*N2*(1-N2/K2-b21*N1/K1)
function xdot = competitive(t,x,K1,K2,b12,b21,r1,r2)
    if x(1) < 0
        x(1) = 0;
    end
    if x(2) < 0
        x(2) = 0;
    end
    xdot = [r1*x(1)*(1-x(1)/K1-b12*x(2)/K2);
            r2*x(2)*(1-x(2)/K2-b21*x(1)/K2)];
end
```

Datei: competitive.m

Code für die eigentliche Simulation (Hier für Simulation 2.4.4, für die Ergebnisse von Simulation 2.4.5 genügt es, die entsprechenden Zeilen von Kommentar- in Code-Zeilen umzuwandeln.)

```
%Simulation des kompetitiven Modells
clear all; close all;

t = [];                 %Speicher fuer die Zeitpunkte
x = [];                 %Speicher fuer die Populationsgroessen

K1 = 2000; K2 = 3000;
b12 = 1; b21 = 1;
r1 = 0.8; r2 = 0.8;

t0 = 0;
T = 100;                %Wachstum ueber 100 Jahre
x0 = [1000;500];        %Anfangsgr.: 1000 Individuen von Spez. 1,
                        %500 von Spez. 2
% Werte fuer stabiles Gleichgewicht
% x0 = [1000;1000];
% r1 = 0.5;
% b12 = 1/3;
[t,x] =  ode45(@(t,x) competitive(t,x,K1,K2,b12,b21,r1,r2),...
                  [t0 T],x0);

%ggf. Korrektur de letzten Werte
if x(1,end) < 0
    x(1,end) = 0;
end
if x(2,end) < 0
    x(2,end) = 0;
end

%Grafische Ausgabe
plot(t,x(:,1),'-b',t,x(:,2),'-r','LineWidth',1.5);
xlabel('Zeit in Jahren','Fontsize',14,...
        'Interpreter','LaTex');
ylabel('Populationsgr{\"o}{\ss}e in Individuen',...
        'Fontsize',14,'Interpreter','LaTex');
legend({'Spezies 1','Spezies 2'},'Fontsize',14,...
        'Interpreter','LaTex');
```

Datei: CompetitiveSim.m

Anhang A.4: Anhang A.4: Simulationen 2.4.6 (Symbiose-Modell)

Code für die Berechnung derÄnderung der Populationsgrößen N_1, N_2 zu einem Zeitpunkt t in Abhängigkeit der beiden Populationsgrößen

```
%Berechnet das Populationswachstum
%zum Zeitpunkt t
%gemaess symbiotischen Modell
%    dN1/dt = r1*N1*(1-N1/K1+b12*N2/K1)
%    dN1/dt = r2*N2*(1-N2/K2+b21*N1/K2)
function xdot = symbiotic(t,x,K1,K2,b12,b21,r1,r2)
    %ggf. Korrektur (-> nicht negative Populationsgroessen)
    if x(1) < 0
        x(1) = 0;
    end
    if x(2) < 0
        x(2) = 0;
    end
    xdot = [r1*x(1)*(1-x(1)/K1+b12*x(2)/K2);
            r2*x(2)*(1-x(2)/K2+b21*x(1)/K2)];
end
```

Datei: symbiotic.m

Code für die eigentliche Simulation

```
%Simulation des symbiotischen Modells
clear all; close all;

t = [];                 %Speicher fuer die Zeitpunkte
x = [];                 %Speicher fuer die Populationsgroessen

K1 = 2000; K2 = 3000;
b12 = 1/2; b21 = 1/2;
r1 = 0.5; r2 = 0.8;

t0 = 0;
T = 100;                %Wachstum ueber 100 Jahre
x0 = [1000;500];        %Anfangsgr.: 1000 Individuen von Spez. 1,
                        %500 von Spez. 2

[t,x] = ode45(@(t,x) symbiotic(t,x,K1,K2,b12,b21,r1,r2),...
                  [t0 T],x0);

%Grafische Ausgabe
plot(t,x(:,1),'-b',t,x(:,2),'-r','LineWidth',1.5);
xlabel('Zeit in Jahren','Fontsize',14',...
        'Interpreter','LaTex');
ylabel('Populationsgr{\"o}{\ss}e in Individuen',...
        'Fontsize',14','Interpreter','LaTex');
legend({'Spezies 1','Spezies 2'},'Fontsize',14',...
        'Interpreter','LaTex');
```

Datei: SymbioticSim.m

Anhang B: Anhang B: Matlab-Codes zu Kapitel 2.5 (Lorenz-Modell)

Code für die Berechnung der Änderung der x,-,y- und z-Koordinaten gemäß dem Lorenz-Modell.

```
%Lorenz-Modell
% d/dt x = sigma(y-x)
% d/dt y = rx-y-xz
% d/dt z = xy - bz
% x = [x;y;z]
function xdot = lorenz(x,sigma,r,b)
    xdot = [0; 0; 0];
    xdot(1) = sigma*(x(2)-x(1));
    xdot(2) = r*x(1)-x(2)-x(1)*x(3);
    xdot(3) = x(1)*x(2)-b*x(3);
end
```

Datei: lorenz.m

Code für die eigentliche Simulation. Aufgrund der Sensibilität gegenüber den Anfangsbedingungen wird hier, anstatt des in den vorherigen Simulationen verwendetet Dormand-Prince-Verfahrens, das vorimplementierte Verfahren ode113 verwendet, bei welchem man auf aufgrund der höheren Ordnung des Kontrollverfahrens geringere Abweichungen von der tatsächlichen Lösung erwarten kann.

```
%Simulation des Lorenz-Modells
close all; clear all;

%"Klassische" Parameter
sigma = 10;
b = 8/3;
r = 28;

%Fuer die Matlabsolver benoetigte odefun
f=@(t,x) lorenz(x,sigma,r,b);

t0 = 0;          %Startzeitpunkt
T = 50;          %Endzeitpunkt
%Startwerte
x0 = [1;1;1];
y0 = [0.9999998;1.0000002;1.0000002];

t1 = [];         %Speicher fuer die Zeitpunkte
t2 = [];
x1 = [];         %Speicher fuer die Datenpunkte,
                  %x(1) = x, x(2) = y, x(3) = z
x2 = [];


%Chaotisches Verhalten -> Kontrollverfahren 13. Ordnung
%aussagekraeftiger als Kontrollverfahren 5. Ordnung

[t1,x1]=  ode113(@(t,x) f(t,x),[t0 T],x0);
[t2,x2]=  ode113(@(t,x) f(t,x),[t0 T],y0);

%Vorbereitung fuer die Synchronisation
```

```
%der Achsenverhaeltnisse
xMax = max(max(x1(:,1)),max(x2(:,1)));
yMax = max(max(x1(:,2)),max(x2(:,2)));
zMax = max(max(x1(:,3)),max(x2(:,3)));
xMin = min(min(x1(:,1)),min(x2(:,1)));
yMin = min(min(x1(:,2)),min(x2(:,2)));
zMin = min(min(x1(:,3)),min(x2(:,3)));

%Grafische Ausgabe

fig1 = figure(1);
set(fig1, 'position', [200,200,1000,700]);
plot(t1,x1(:,1),'-b',t2,x2(:,1),'-r','LineWidth',1.2);
xlabel('$t$','FontSize',16,'Interpreter','LaTex');
ylabel('$x$-Wert','FontSize',16,'Interpreter','LaTex');
legend({'Startwert $(1,1,1)^T$',...
    'Startwert $(0.9999998,1.0000002,1.0000002)^T$'},...
    'FontSize',13,'Interpreter','LaTex');
axis([0 50 -30 30]);

fig2 = figure(2);
set(fig2, 'position', [200,200,1200,700]);
sub1 = subplot(1,2,1);
plot3(x1(:,1),x1(:,2),x1(:,3),'-b','LineWidth',1.2);
xlabel('$x$','FontSize',16,'Interpreter','LaTex');
ylabel('$y$','FontSize',16,'Interpreter','LaTex');
zlabel('$z$','FontSize',16,'Interpreter','LaTex');
title('Startwert $(1,1,1)^T$','FontSize',14,...
        'Interpreter','LaTex');
ax = gca;
axis equal;
grid on;
axis([xMin, xMax, yMin, yMax, zMin, zMax]);
set(ax,'Ydir','reverse');

sub2 = subplot(1,2,2);
plot3(x2(:,1),x2(:,2),x2(:,3),'-r','LineWidth',1.2);
xlabel('$x$','FontSize',16,'Interpreter','LaTex');
ylabel('$y$','FontSize',16,'Interpreter','LaTex');
zlabel('$z$','FontSize',16,'Interpreter','LaTex');
title('Startwert $(0.9999998,1.0000002,1.0000002)^T$',...
        'FontSize',14,'Interpreter','LaTex');
ax = gca;
axis equal;
grid on;
axis([xMin, xMax, yMin, yMax, zMin, zMax]);
set(ax,'Ydir','reverse');
```

Datei: LorenzSim.m

Anhang C: Anhang C: Matlab-Codes zu Kapitel 3 (Optimierung und Optimale Steuerung)

Anhang C.1: Anhang C.1: Steepest-Descent

Code für das Gradienten-Verfahren.

```
% f, gradf: Function-handels von Funktion und Gradient
% beta: Positiver Parameter < 1 fuer die Armijo-Regel
% sigma: Parameter fuer die Abstiegungsbedingung.
% epsilon: Abbruchbedingung
% maxIter: Maximale Iterationszahl
function [x,iter] = gradientMethod(f,gradf,x0,beta, ...
                        sigma, epsilon,maxIter)
    k = 0;
    xk = x0;
    %Solange Abbruchbedingung nicht erfuellt
    while norm(gradf(xk))> epsilon && maxIter>k
        %Nehme steilsten Abstieg als Suchrichtung
        dk = -gradf(xk);
        %Armijo-Linesearch
        tk = 1;
        while f(xk+tk*dk)>f(xk)-sigma*tk*dk'*dk
            tk = tk*beta;
        end
        %Aktualisierung von xk
        xk = xk+tk*dk;
        k = k+1;
    end
    x = xk;
    iter = k;
end
```

Datei: gradientMethod.m

Anhang C.2: Anhang C.2: Augmentiertes Lagrange-Verfahren

Code für das augmentierte Lagrange-Verfahren.

```
% f, h, gradf, jacobh: Function-handels der zu minimierenden
% Fkt., der Nebenbedingung, sowie der zugehoerigen
% Ableitungen
% x0, lambda0 sind Startwerte
% eps1 und eps2 bestimmen die Abbruchkriterien
% maxIter ist die maximale Anzahl an Iterationen
function [x, lambda, iter] = augmentedLagrangian(f, h, gradf, ...
        jacobh, x0, lambda0, c0, eps1, eps2, maxIter)
    lambdak = lambda0;
    ck = c0;
    xk = x0;
    k = 0;
    %Solange Abbruchbedingung nicht erfuellt
    while (norm(gradf(xk)+jacobh(xk)'*lambdak)>eps1 || norm(h(xk)
  ) > eps2) && maxIter > k
        %Umwandlung in ein unrestringiertes Problem mittels
```

```
        %augmentierter Lagrange-Funktion
        aLag = @(x) f(x)+lambdak'*h(x)+ck/2*norm(h(x))^2;
        gradLag = @(x) gradf(x)+jacobh(x)'*lambdak+ck*jacobh(x)'*
   h(x);
        x_old = xk;
    %Minimiere Lagrange-Funktion
        xk = gradientMethod(aLag,gradLag,xk,0.5, 10^-4, 10^-8,
   maxIter);
        lambdak = lambdak+ck*h(xk);
    %Wenn Annaeherung an Nebenbedingungsmenge
    %zu gering -> Erhoehe Penalty-Parameter
        if norm(h(xk)) > 0.5*norm(x_old)
            ck = 10*ck;
        end
        k = k+1;
    end
    x = xk;
    lambda = lambdak;
    iter = k;
end
```

Datei: augmentedLagrangian.m

Anhang C.3: Anhang C.3: Lösung eines optimalen Steuerungproblems

Code für die Lösung eines Optimalen Steuerungsproblems mit einem Differentialgleichungsmodell der Form

$$\min J(y,u) = \int_0^T l(t,y,u)\mathrm{d}t$$
$$\text{u. d. Nb. } \dot{y} = f(t,y,u), y(0) = y_0$$

Für die Lösung des Modells wurde jeweils der explizite bzw. implizite Euler angewandt. Problemspezifische Parameter wie Diskretisierung, Modellfunktion (f), Funktional J, sowie partielle Ableitungen werden jeweils mittels einem sogenannten ***structure arry*** (OCP genannt) übergeben, die Anzahl an Parametern einer Funktion vermieden werden.

```
function [y,u] = ocpGradient(u0, OCP, delta, kmax, eps)
    %Anzahl Gitterpunkte
    n = length(OCP.dt);
    %Laufvariable
    k = 0;
    [reducedGrad, y0] = computeGradient(OCP,u0);
    u = u0;
    y = y0;
    %Optimierungssschleife
    while norm(reducedGrad,2) > eps && k < kmax
        [alpha,y,u] = linesearchEx(u, y, reducedGrad, ...
                          delta, 2, OCP);
        [reducedGrad, y0] = computeGradient(u,OCP);
        k = k+1;
    end
end
```

Datei: ocpGradient.m

Für die zugrunde liegende „linesearch“ wurde die Armijo-Regel verwendet:

```
function [alpha,y,u] = linesearch(u0, y0, reducedGrad, delta, ...
                                  alpha0, OCP)
    n = length(OCP.dt)-1;
    alpha = alpha0;
    u = u0-alpha*reducedGrad;
    y = OCP.y0;
    %Loese die Dgl. fuer y
    %mittels explizitem Euler
    for i = 1:n
        h = OCP.dt(i+1)-OCP.dt(i);
        y = [y, y(i)+h*OCP.f(OCP.dt(i),y(i),u(i))];
    end
    k = 0;
    %Suche passendes alpha
    while OCP.J(y,u)>OCP.J(y0,u0)-...
            delta*alpha*norm(reducedGrad,2)^2 && k < 16
        alpha = alpha/2;
        u = u0-alpha*reducedGrad;
        y = OCP.y0;
        for i = 1:n
            h = OCP.dt(i+1)-OCP.dt(i);
            y = [y, y(i)+h*OCP.f(OCP.dt(i),y(i),u(i))];
        end
        k = k+1;
    end
end
```

Datei: linesearch.m

Der Gradient wird mit der folgenden Funktion berechnet:

```
function [reducedGrad, y,lambda] = computeGradient(u,OCP)
    %Anzahl an Gitterpunkten ohne den Startknoten
    n = length(OCP.dt)-1;
    y = OCP.y0;
    %Loese die Dgl. fuer y
    %mittels expliziten Euler
    for i = 1:n
        h = OCP.dt(i+1)-OCP.dt(i);
        y = [y, y(i)+h*OCP.f(OCP.dt(i),y(i),u(i))];
    end
    %Berechne den Lagrange-Multiplikator
    %mittels impliziten Euler
    lambda = zeros(n+1,1);
    for i=n:-1:1
        h = OCP.dt(i+1)-OCP.dt(i);
        d = -OCP.f_y(OCP.dt(i+1),y(i+1),u(i+1))*lambda(i+1)+...
            OCP.l_y(OCP.dt(i+1),y(i+1),u(i+1));
        lambda(n) = lambda(n+1)-h*d;
    end
    %Setze den reduzierten Gradienten zusammen
    reducedGrad = -OCP.f_u(OCP.dt,y,u)*lambda+...
            OCP.l_u(OCP.dt,y,u);
```

```
end
```

Datei: computeGradient.m

Zum Testen kann beispielsweise das folgende Skript verwendet werden. Es entspricht dem Problem (3.4.28).

```
close all; clear all;
%Erstellung des Structure arrays
OCP.dt = [0:0.1:1];
%Dgl-Modell und Integrand
OCP.y0 = 1;
OCP.f = @(t,y,u) y+u^2;
OCP.l = @(t,y,u) y^2+u^2;
%Partielle Ableitungen
OCP.f_y = @(t,y,u) 1;
OCP.f_u = @(t,y,u) 2*u;
OCP.l_y = @(t,y,u) 2*y;
OCP.l_u = @(t,y,u) 2*u;
%Berechnung des Integralls mittels Trapezregel
OCP.J = @(y,u) sum(1/2*(OCP.dt(2:end)-...
            OCP.dt(1:end-1)).*(y(1:end-1).^2+u(1:end-1).^2+...
            y(2:end).^2+u(2:end).^2));
%Maximale Iterationszahl
kmax = 100000;
eps = 10^-7;
delta = 0.8;
u0 = 2*rand(1,length(OCP.dt));
[y,u] = ocpGrad(u0, OCP, delta, kmax, eps);
plot(OCP.dt,y,OCP.dt,u)
legend('Zustand','Kontrolle')
```

Datei: TestOCP.m

Anhang D: Anhang D: Matlab-Codes zu Kapitel 4 (Inverse Probleme)

Code für die Parameter-Identifikation eines Differentialgleichungsmodells der Form

$$y' = f(x, y, p), \quad y(x_0) = y_0$$

Der Code wurde möglichst allgemein gehalten, so können z. B. verschieden ODE-Lösungsmethoden ausprobiert oder die Anzahl an Messdaten verändert werden, ohne dass der Algorithmus geändert werden muss. Die Messpunkte, Messdaten, Dynamik sowie der Anfangswert werden mittels einem structure array MODEL übergeben, ODE enthält Function(handles) für die Abbildung $p \mapsto y$ sowie deren Gradienten. $p0$ beschreibt die Startnäherung für p. Der abgebildete Algorithmus kombiniert die Armijo-Regel mit dem Levenberg-Marquardt-Verfahren, wofür die restlichen Parameter bestimmt sind.

```
% Identifiziert fuer ein Problem der Form
%       y'=f(x,y,p), y(x0)= y0
% den Parameter p zu gegebenen Messdaten
function [y,p] = parameterID(MODEL, ODE, p0, sigma, beta, eps,
   maxIter)
    %Least-Squares-Funktional
    J = @(p) 1/2*norm(ODE.solve(MODEL.xd, MODEL.y0,p, MODEL.f)-
   MODEL.yd,2)^2;
    gradJ = @(p) ODE.grad(MODEL.xd, MODEL.y0,p, MODEL.f)'*...
        (ODE.solve(MODEL.xd, MODEL.y0,p, MODEL.f)-MODEL.yd);
    %Suchrichtung
    pk = p0;
    grad = gradJ(pk);
    k = 0;
    while norm(grad) > eps && k < maxIter
        val = J(pk);
        % Je naeher man der optimalen Loesung ist, desto
        % weniger muss der Regularisierungsparameter
        % eine Nachjustierung erzwingen -> er soll
        % kleiner werden
        lambda = min([1,norm(grad), val]);
        tmp = ODE.grad(MODEL.xd, MODEL.y0,pk, MODEL.f);
        dk = -(tmp'*tmp+lambda*eye(length(pk)))\grad;
        tk = 1;
        while J(pk+tk*dk) > val+sigma*tk*grad'*dk
            tk = tk*beta;
        end
        pk = pk+tk*dk;
        grad = gradJ(pk);
        k = k+1;
    end
    p = pk;
    y = ODE.solve(MODEL.xd, MODEL.y0,p, MODEL.f);
end
```

Datei: parameterID.m

Der Algorithmus lässt sich z. B. mit folgendem Skript testen, welches ihn auf das Problem

$$y' = p \cdot y, \quad y(x_0) = 1,\ p \in \mathbb{R}$$

anwendet.

```
close all; clear all;

%Erstellung der Structure-Arrays
h = 0.1;
MODEL.xd = [0:h:1]';
%Exakte und gestoerte Daten
pExact = -0.5;
yExact = exp(pExact*MODEL.xd);
noise = 0.01;
MODEL.yd = yExact.*(1+noise*randn(length(MODEL.xd),1));
MODEL.y0 = 1;
MODEL.f = @(x,y,p) y*p;

%Dgl-Solver und Gradient
ODE.solve =
    @(grid, y0, p, f) explicitEuler(grid, y0, p, f);
ODE.grad =
    @(grid, y0, p, f) explicitEulerGradient(grid, y0, p, f);

%Maximale Iterationszahl
maxIter = 10000;
eps = 10^-7;
beta = 0.5;
sigma = 10^-3;

p0 = -1;
[y,p] = parameterID(MODEL, ODE, p0, sigma, beta, eps, maxIter);

fig = figure('visible','off');
movegui('center');
plot(MODEL.xd,yExact,'-ob',MODEL.xd,MODEL.yd,'--xr',...
    MODEL.xd,y,'-.xm','LineWidth',1.6);
legend({'Exakte Daten', 'Messdaten', ...
    'Daten zum rekonstruierten $p$'}, 'Interpreter', 'LaTex');
set(fig,'visible', 'on');
```

Datei: TestParameterID.m

Das explizite Eulerverfahren sowie dessen Gradient bzgl. p sind in den folgenden MATLAB-Dateien beschrieben.

```
% Berechnet die Loesung des AWPs
%  y'=f(x,y,p), dim(p) = 1
%  y(x0) = y0
% auf dem vorgegebenen Gitter
% x0 muss selbst Gitterpunkt sein
function y = explicitEuler(grid, y0, p, f)
    n = length(grid);
    steps = grid(2:n)-grid(1:n-1);
    y = y0;
    for i = 1:n-1
        y = [y; y(i)+steps(i)*f(grid(i),y(i),p)];
    end
```

```
end
```

Datei: explicitEuler.m

```
% Berechnet den Gradienten zur Abbildung p -> y,
% wobei y Loesung des AWPs
%  y'=f(x,y,p), dim(p) = 1
%  y(x0) = y0
% auf dem vorgegebenen Gitter ist.
% x0 muss selbst Gitterpunkt sein
function grad = explicitEulerGradient(grid, y0, p, ~)
    steps = grid(2:end)-grid(1:end-1);
    n = length(steps);
    grad = zeros(n+1,1);
    % Es gilt: y_0 = y_0, y_k = y_0*(1+h_0*p)*...*(1+h_(k-1)*p)
    % Daraus laesst sich der Gradient rekursiv mittels
    % Produktregel berechnen, d. h.
    % y_k' = y_(k-1)*h_(k-1)+y_(k-1)'*(1+h_(k-1)*p)
    % prod = Produkt y_0*...*y_(k-1)
    prod = y0;
    for i = 1:n
        tmp = prod*steps(i)+grad(i)*(1+steps(i)*p);
        prod = prod*(1+steps(i)*p);
        grad(i+1) =tmp;
    end
end
```

Datei: explicitEulerGradient.m

Anhang E: Anhang E: Matlab-Codes zu Kapitel 5 (Stochastische Differentialgleichungen)

Bei den nun folgenden MATLAB-Codes handelt es sich größtenteils um die in [31] angegebenen, welche, von einigen kleinen Änderungen (v. a. grafische Ausgabe und Kommentare) abgesehen, unverändert übernommen wurden. Der in [31] angegebene Downloadlink http://www.maths.strath.ac.uk/~aas96106/algfiles.html scheint tot zu sein, jedoch sind sämtliche MATLAB-Dateien des Artikels über https://people.cs.clemson.edu/~steve/CPLSCMODS/SDE-matlab/algfiles.html erhältlich.

Anhang E.1: Anhang E.1: Simulation einer Brownschen Bewegung

Code für die normale Brownsche Bewegung.

```
rng(100,'v5normal')     % Initialisierung des Generators
T = 1; N =500; dt = T/N;


dW = zeros(1,N);
W = zeros(1,N);

dW(1) = sqrt(dt)*randn;
W(1) = dW(1);
for j = 2:N
    dW(j) = sqrt(dt)*randn;
    W(j) = W(j-1)+dW(j)
end


% Alternative, effizientere Implementierung
% dW = sqrt(dt)*randn(1,N);
% W = cumsum(dW);   % Kumulative Summe

plot([0:dt:T],[0,W],'r-', 'LineWidth', 1.5);
xlabel('$t$','FontSize',16,'Interpreter','LaTex')
ylabel('$W(t)$','FontSize',16,'Interpreter','LaTex','Rotation',0)
```

Datei: bpath.m

Code für eine Verkettung der Brownschen Bewegung.

```
rng(100,'v5normal')
T = 1; N = 500; dt = T/N; t = [dt:dt:1];

M = 1000;                    % Anzahl Probepfade
dW = sqrt(dt)*randn(M,N);    % Wiener-Inkremente
W = cumsum(dW,2);
U = exp(repmat(t, [M 1]) +0.5*W);
Umean = mean(U);

plot([0,t],[1,Umean],'b-','LineWidth',1.5), hold on
plot([0,t],[ones(5,1),U(1:5,:)],'r--','LineWidth',1.5)
xlabel('$t$', 'FontSize', 16, 'Interpreter','LaTex')
ylabel('$U(t)$', 'FontSize', 16, 'Interpreter','LaTex',...
      'Rotation',0)
legend({'Mittel \"uber 1000 Pfade', '5 einzelne Pfade'}, ...
```

```
    'FontSize',14,'Interpreter','LaTex','Location','Northwest')
```

Datei: bpath3.m

Anhang E.2: Anhang E.2: Simulation der Euler-Mayurama-Methode (Kapitel 5.3.1)

Das Ergebnis ist in Abbildung 5.3.1 dargestellt.

```
% EM Euler-Maryama-Methode fuer lineare SDE
%
%       dX = lambda*X dt + mu*X dW
% mit lambda = 2, mu = 1 und Xzero = 1
% Diskretisierter Brownscher Pfad auf [0,1] mit dt = 2^(-8)
% Zeitschritt fuer Euler-Maruyama: R*dt

seed = 100;
rng(seed,'v5normal')
lambda = 2; mu = 1; Xzero = 1;  % Problemspezifische Parameter
T = 1; N = 2^8; dt = 1/N;
dW = sqrt(dt)*randn(1,N);       % Wiener-Inkremente
W = cumsum(dW);                 % Diskretisierter Brownscher Pfad

Xtrue = Xzero*exp((lambda-0.5*mu^2)*([dt:dt:T])+mu*W);
plot([0:dt:T],[Xzero, Xtrue],'m-','LineWidth',1.5); hold on

R = 4; Dt = R*dt; L = N/R;
Xem = zeros(1,L);
Xtemp = Xzero;
for j = 0:L-1
    Winc = sum(dW(R*j+1:R*(j+1)));
    Xtemp = Xtemp + Dt*lambda*Xtemp + mu*Xtemp*Winc;
    Xem(j+1) = Xtemp;
end

plot([0:Dt:T],[Xzero,Xem],'r--*','LineWidth',1.5); hold off
xlabel('$t$', 'FontSize',16,'Interpreter','LaTex')
ylabel('$X$','FontSize',16,'Rotation',0','HorizontalAlignment','
   right','Interpreter','LaTex')
legend({'Exakte L{\"o}sung','Ergebnisse von EM'}, 'FontSize',12,'
   Interpreter','LaTex','Location','Northwest')

emerr = abs(Xem(end)-Xtrue(end))
```

Datei: em.m

Anhang E.3: Anhang E.3: Code für den Test der starken Konvergenz

Das Ergebnis ist in Abbildung 5.3.2 dargestellt.

```
%EMSTRONG  Veranschaulichung der starken Konvergenz der EM-
   Methode
% mit Hilfe der Test-SDE
%       dX = lambda*X dt + mu*X dW,    X(0) = Xzero,
%       mit lambda = 2, mu = 1 and Xzer0 = 1.
%
% Die Schrittweite des Brownschen Pfades ist dt = 2^(-9).
% Fuer die EM werden 5 verschiedene Zeitschritte getestet:
% 16dt, 8dt, 4dt, 2dt, dt.
% Der Fehler wird zum Zeitpunkt T=1 betrachtet:  E | X_L - X(T)
   |.

rng(100,'v5normal')

lambda = 2; mu = 1; Xzero = 1;     % Problemparameter
T = 1; N = 2^9; dt = T/N;          %
M = 1000;                          % Anzahl der Probepfade

Xerr = zeros(M,5);
for s = 1:M,
    dW = sqrt(dt)*randn(1,N);
    W = cumsum(dW);
    Xtrue = Xzero*exp((lambda-0.5*mu^2)+mu*W(end));
    for p = 1:5
        R = 2^(p-1); Dt = R*dt; L = N/R;      % EM-Schritt
        Xtemp = Xzero;
        for j = 1:L
             Winc = sum(dW(R*(j-1)+1:R*j));
             Xtemp = Xtemp + Dt*lambda*Xtemp + mu*Xtemp*Winc;
        end
        Xerr(s,p) = abs(Xtemp - Xtrue);
    end
end

figure('position', [686,330,480,330])
Dtvals = dt*(2.^([0:4]));
%subplot(1,2,1)
loglog(Dtvals,mean(Xerr),'b*-', 'LineWidth', 1.5), hold on
loglog(Dtvals,(Dtvals.^(.5)),'r--','LineWidth', 1.5), hold off
axis([1e-3 1e-1 1e-4 1])
axis square
xlabel('$\mathit{\Delta} t$','FontSize',14,'Interpreter','LaTex')
ylabel('Testdurchschnitt von $\left| X(T) - X_L \right|$',...
        'FontSize',14,'Interpreter','LaTex')
title('Starke Konvergenz','FontSize',12)
```

Datei: emstrong.m

Anhang E.4: Anhang E.4: Code für den Test der schwachen Konvergenz

Für den Test der schwachen Konvergenz der schwachen Euler-Maruyama-Methode genügt es, die entsprechenden Zeilen auszukommentieren. Das Ergebnis ist in Abbildung 5.3.3 dargestellt.

```
%EMWEAK Veranschaulichung der schwachen Konvergenz der EM-Methode
% sowie der schwachem EM-Methode anhand der SDE
%
%       dX = lambda*X dt + mu*X dW,    X(0) = Xzero,
%
% mit lambda = 2, mu = 0.1 and Xzero = 1.
% Die Schrittweite des Brownschen Pfades ist dt = 2^(-9).
% Fuer die EM werden 5 verschiedene Zeitschritte getestet:
% 16dt, 8dt, 4dt, 2dt, dt.
% Der Fehler wird zum Zeitpunkt T=1 betrachtet:
% E | X_L - X(T) |.

rng(100,'v5normal')

lambda = 2; mu = 0.1; Xzero = 1; T = 1;
M = 50000;                          % Anzahl der Probepfade

Xem = zeros(5,1);
for p = 1:5
    Dt = 2^(p-10); L = T/Dt;
    Xtemp = Xzero*ones(M,1);
    for j = 1:L
        Winc = sqrt(Dt)*randn(M,1);
        %Winc = sqrt(Dt)*sign(randn(M,1)); %%Schwache Euler-
        %                                  Maruyama-Methode E-M %%
        Xtemp = Xtemp + Dt*lambda*Xtemp + mu*Xtemp.*Winc;
    end
    Xem(p) = mean(Xtemp);
end
Xerr = abs(Xem-exp(lambda));

figure('position', [400,300,1050,390])
Dtvals = 2.^([1:5]-10);
subplot(1,2,1)
%subplot(1,2,2) %%Schwache Euler-Maruyama-Methode E-M %%
loglog(Dtvals,Xerr,'b*-', 'LineWidth', 1.5), hold on
loglog(Dtvals,Dtvals,'r--','LineWidth', 1.5), hold off
axis([1e-3 1e-1 1e-4 1])
axis square
xlabel('$\mathit{\Delta} t$','FontSize',14,'Interpreter','LaTex')
ylabel(['$\left|E(X(T))-\right.$ Testdurchschnitt von '...
        '$\left.X_L\right|$'],'FontSize',14,...
        'Interpreter','LaTex')
title('Euler-Maruyama-Methode','FontSize',12)
%title('Schwache Euler-Maruyama-Methode','FontSize',12)
%                        %%Schwache Euler-Maruyama-Methode E-M %%
```

Datei: emweak.m

Anhang E.5: Anhang E.5: Code für den Test der Stabilität

Das Ergebnis ist in Abbildung 5.3.4 dargestellt.

```
%STAB Stabilitaetstest anhand der Euler-Maruyama-Methode
% anhand der Testgleichung
%
%   dX = lambda*X dt + mu*X dW,   X(0) = Xzero,
%
% mit lambda = -3, mu = sqrt(3) und Xzero = 1.

rng(100,'v5normal')
T = 20; M = 50000; Xzero = 1;
ltype = {'b-','r--','m-.'};              % Linienstile fuer Plot

figure('position', [550   300   900   580])
subplot(211)  %%%%%%%%%%%% Quadratisches Mittel %%%%%%%%%%%%
lambda = -3; mu = sqrt(3);               % Problemparameter
for k = 1:3
    Dt = 2^(1-k);
    N = T/Dt;
    Xms = zeros(1,N); Xtemp = Xzero*ones(M,1);
    for j = 1:N
            Winc = sqrt(Dt)*randn(M,1);
            Xtemp = Xtemp + Dt*lambda*Xtemp + mu*Xtemp.*Winc;
            Xms(j) = mean(abs(Xtemp).^2);
    end
    semilogy([0:Dt:T],[Xzero,Xms],ltype{k},'Linewidth',2), hold
   on
end
legend({'$\Delta t = 1$','$\Delta t = 1/2$',...
    '$\Delta t = 1/4$'},'Interpreter','Latex','FontSize',12)
title('Stabilit{\"a}t im quadratischen Mittel','FontSize',16,...
    'Interpreter','Latex')
ylabel('$E\left(\left|X\right|^2\right)$','FontSize',14,...
    'Interpreter','Latex')
axis([0,T,1e-20,1e+20]), hold off

subplot(212)  %%%%% Asymptotische Stabilitaet %%%%%%%
T = 500;
lambda = 0.5; mu = sqrt(6);              % Problemparameter
for k = 1:3
    Dt = 2^(1-k);
    N = T/Dt;
    Xemabs = zeros(1,N); Xtemp = Xzero;
    for j = 1:N
            Winc = sqrt(Dt)*randn;
            Xtemp = Xtemp + Dt*lambda*Xtemp + mu*Xtemp*Winc;
            Xemabs(j) = abs(Xtemp);
    end
    semilogy([0:Dt:T],[Xzero,Xemabs],ltype{k},'Linewidth',2),
   hold on
end
semilogy([0:Dt:T],eps+[0:Dt:T].*0,'k-','Linewidth',2), hold on
```

```
legend({'$\Delta t = 1$','$\Delta t = 1/2$',...
    '$\Delta t = 1/4$'},'Interpreter','Latex')
title('Asymptotische Stabilit{\"a}t','FontSize',16,...
    'Interpreter','Latex')
ylabel('$\left|X\right|$','FontSize',14,'Interpreter','Latex')
axis([0,T,1e-50,1e+100]), hold off
```

Datei: stab.m

Anhang E.6: Anhang E.6: Funktion für die Euler-Mayurama-Methode (Kapitel 5.3.1)

Code für die Euler-Mayurama-Methode zur Lösung einer SDE. In dieser Version wurde sie als möglichst allgemein gehaltene Funktion implementiert, welche alle problemspezifischen Daten wie Anfangswert, Diskretisierungsschema oder die Differentialgleichung als Parameter enthält, sodass sie vielseitig verwendbar ist. (Die hierdurch nötigen Fehlermeldungen sind in Englisch angegeben, da die MATLAB-eigenen Fehlermeldungen ebenfalls in Englisch sind.)
Die optionalen Parameter `seed` und `generator` sowie die Befehle `rng('default')`, `rng(seed)` und `rng(seed, generator)` dienen hierbei der Initialisierung des Zufallsgenerators. Auf diese Weise kann eine bestimmte Folge von Zufallszahlen reproduziert werden, sofern dies gewünscht ist.

```
% Euler-Maruyama-Methode fuer eine allgemeine SDE der Form
%
% dX = f(t,X(t))dt + sigma(t,X(t)) dW, X(0) = Xzero
%
% dt: Zeitinkrement fuer die Brownsche Bewegung,
% Dt: Zeitinkrement fuer EM, muss ganzzahliges
%     Vielfaches von dt sein
%
% T: Endzeit, muss ganzzahliges Vielfaches von Dt
%    sowie > 0 sein
%
% seed: Vom Zufallszahlengenerator verwendeter Seed (optional),
%       bei Verwendung des selben Seeds erzeugt ein erneuter
%       Aufruf die selbe Folge an Zufallszahlen
%       seed muss eine Ganzzahl aus zwischen 0 und 2^32-1
%       (einschliesslich) sein
% generator: Typ des Zufallsgenerators (optional),
%            benoetigt die Angabe eines Seeds,
%            erlaubt sind:'twister', 'simdTwister',
%            'combRecursive', 'multFibonacci',
%            'v5uniform', 'v5normal', 'v4'
%
% Erzeugt eine Fehlermeldung, wenn mindestens eine
% der Folgenden Bedingungen erfuellt ist:
%    - Weniger als 6 Inputargumente
%    - Mehr als  8 Inputargumente
%    - 'f' ist kein function handle
%    - 'sigma' ist kein function handle
%    - 'dt' ist keine Zahl vom Typ 'double'
%    - 'Dt' ist keine Zahl vom Typ 'double'
%    - 'T' ist keine Zahl vom Typ 'double'
%    - Zufallsgeneratortyp ist angegeben, aber kein seed
%      (sofern angegeben)
%    - 'seed' ist keine Ganzzahl
%      (sofern angegeben)
%    - 'generator' ist kein gueltiger Zufallsgeneratortyp
%      (sofern angegeben)
%    - f ist keine Funktion in 2 Variablen
%    - sigma ist eine Funktion 2 Variablen
%    - dt <= 0
%    - Dt <= 0
```

```
%     - T <=0
%     - Dt/dt  ist keine Ganzzahl
%     - T/Dt  ist keine Ganzzahl
%     - seed < 0 oder seed  >= 2^32
%       (sofern angegeben)

function [t,X] = eulerMaruyama(f, sigma, Xzero, Dt, dt, T, ...
                                          seed, generator)
    if nargin < 6
        error('Not enough input arguments.')
    elseif nargin > 8
        error('Too many input arguments.')
    elseif not(isa(f,'function_handle'))
        error('Input argument ''f'' must be a function handle.')
    elseif not(isa(sigma,'function_handle'))
        error(['Input argument ''sigma'' must be a ' ...
               'function handle.'])
    elseif not(isfloat(dt))
        error(['Input argument ''dt'' must be '...
               'of type ''double''.'])
    elseif not(isfloat(Dt))
        error(['Input argument ''Dt'' must be '...
               'of type ''double''.'])
    elseif not(isfloat(T))
        error(['Input argument ''T'' must be '...
               'of type ''double''.'])
    elseif nargin >= 7
        if nargin == 8
            if not(ischar(generator))
                error('''generator'' must be a character array.')
            elseif not(any(ismember({'twister','simdTwister',...
                    'combRecursive','multFibonacci',...
                    'v5uniform','v5normal','v4'},generator)))
                error(sprintf(['''%s'' is not a valid '...
                               'generator type.'],generator))
            elseif not(isinteger(seed)) && not(isfloat(seed))...
                    || round(seed) ~= seed
                error(['Input argument ''seed'' must be a '...
                       'nonnegative integer less then 2^32 .']);
            elseif seed < 0 || seed >= 2^32
                error(['Input argument ''seed'' must be a '...
                       'nonnegative integer less then 2^32 .']);
            end
        elseif ischar(seed) && any(ismember({'twister',...
                'simdTwister','combRecursive','multFibonacci',...
                'v5uniform','v5normal','v4'},seed))
            error(['To use ''generator'' option,'...
                   ' a valid seed must be specified.']);
        elseif not(isinteger(seed)) && not(isfloat(seed))...
                || round(seed) ~= seed
            error(['Input argument ''seed'' must be a '...
                   'nonnegative integer less then 2^32.'])
        elseif seed < 0 || seed >= 2^32
```

```
            error(['Input argument ''seed'' must be a '...
                   'nonnegative integer less then 2^32.'])
        end
    elseif nargin(f) ~= 2
        error(['Input argument ''f'' must be a '...
               'function of two variables.'])
    elseif nargin(sigma) ~= 2
        error(['Input argument ''sigma'' must be a '...
               'function of two variables.'])
    elseif dt <= 0
        error(sprintf(['Input argument ''dt'' must be '...
                       'nonnegative but was %f.'], dt))
    elseif Dt <= 0
        error(sprintf(['Input argument ''Dt'' must be '...
                       'nonnegative but was %f.'], Dt))
    elseif T <= 0
        error('Input argument ''T'' must be nonnegative.')
    elseif mod(Dt,dt) ~= 0
        error('Input argument ''Dt'' must be nonnegative.')
    elseif mod(T,Dt) ~= 0
        error('Input argument ''dt'' must be nonnegative.')
    end

    if nargin == 7
        rng('default')          % Initialisierung des Zufalls-
        rng(seed)               % zahlengenerators
    elseif nargin == 8
        rng(seed, generator)
    end

    N = T/dt;                   % Anzahl an Wiener-Inkrementen
    R = Dt/dt;                  % Anzahl an Wiener-Inkrementen
                                % fuer Zeitschritt
    L = N/R;                    % Anzahl an Zeitschritten

    dW = sqrt(dt)*randn(1,N);   % Wiener-Inkremente
    t = [0:Dt:T];               % Diskretisierung von [0,T]
    X = zeros(1,L+1);           % Vorinitialisierung der Loesung
                                % zwecks Effizienz
    Xtemp = Xzero;
    X(1) = Xzero;
    for j = 0:L-1
        Winc = sum(dW( R*j+1:R*(j+1)));
        Xtemp = Xtemp + f(t(j+1), Xtemp)*Dt ...
                + sigma(t(j+1), Xtemp)*Winc;
        X(2+j) = Xtemp;
    end
end
```

Datei: eulerMaruyama.m

Referenzen und weiterführende Literatur

Kapitel 1

[1] Heuser, Harro: *Gewöhnliche Differentialgleichungen: Einführung in Lehre und Gebrauch.* Stuttgart: Teubner 1989.

[2] Hubbard, John H. & West, Beverly H.: *Differential equations: A Dynamical Systems Approach. Ordinary Differential Equations.* (Vol. I&II) Springer, New York 1991.

[3] Walter, Wolfgang: *Gewöhnliche Differentialgleichungen.* 4. Auflage. Springer: Berlin/Heidelberg 1990.

[4] Borzi, Alfio & Wogrin, Melani: *Equazioni differenziali ordinarie.* Hevelius Edizioni: Benevento 2009.

[5] : Arrowsmith, David K. & Place, Colin M.: *Dynamical systems. Differential equations, maps and chaotic behaviour.* Chapman & Hall: London 1992.

[6] Hairer, Ernst; Wanner, Gerhard & Lubich, Christian: *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer: Berlin/Heidelberg 2006.

Kapitel 2

[7] Eck, Christof; Garcke, Harald & Knabner, Peter: *Mathematische Modellierung* [elektronische Version]. Springer: Berlin/Heidelberg 2008.

[8] Murray, James D.: *Mathematical Biology.* 2. Auflage. Springer: Berlin [u. a.] 1993

[9] Yur'evna, Riznichenko G.: *Mathematical models in biophysics.* Erhalten über: https://www.biophysics.org/Portals/1/PDFs/Education/galina.pdf.

[10] Slawig, Thomas: *Klimamodelle und Klimasimulationen* [Elektronische Version]. Springer: Berlin/Heidelberg 2015.

[11] Yang, Jianke: *Chaos: Lorenz Equations, Strange Attractors and Fractals* [Elektronisches Skript]. Erhalten über https://www.emba.uvm.edu/~jxyang/teaching/Math266notes13.pdf.

[12] Nolting, Wolfgang: *Grundkurs Theoretische Physik 2. Analysierte Mechanik.* [Kapitel 2, elektronische Version.] Springer: Berlin/Heidelberg 2014.

[13] : Hairer, Ernst: *Lecture 1: Hamiltonian systems* [Elektronisches Skript]. Erhalten über http://www.unige.ch/~hairer/poly_geoint/week1.pdf.

[14] Tu, Pierre N. V.: Springer: Berlin/Heidelberg 2014. *Dynamical Systems. An Introduction with Applications in Economics and Biology.* [Kapitel 8.] Springer: Berlin/Heidelberg 1994.

[15] Marsden Jerrold, E.: *Gradient and Hamiltonian Systems* [Elektronisches Skript]. Erhalten über https://www.cds.caltech.edu/~murray/wiki/images/7/7d/Cds140a-wi11-Week6NotesHamGrad.pdf.

Kapitel 3-4

[16] Volkwein, Stefan: *Grundlagen der Optimierung* [Elektronisches Skript].

[17] Botelho, Fabio: *Functional Analysis and Applied Optimization in Banach Spaces. Applications to Non-Convex Variational Models.* [Elektronische Version.] Springer: Cham [u. a.] 2014.

[18] Clarke, Francis: *Functional Analysis, Calculus of Variations and Optimal Control.* [Elektronische Version.] Springer: London [u. a.] 2014.

[19] Kielhöfer, Hansjörg: *Variationsrechnung. Eine Einführung in die Theorie einer unabhängigen Variablen mit Beispielen und Aufgaben.*[Elektronische Version.] Vieweg+Teubner: Wiesbaden 2010.

[20] Kot, Mark: *A first course in the calculus of variations.* American Mathematical Society: USA 2014.

[21] van Brunt, Bruce: *Calculus of Variations.* Springer: New York 2004.

[22] Rockafeller, Ralph T. & Wets, Roger J. B.: *Variational Analysis.* Grundlehren der mathematischen Wissenschaften. Springer: Berlin [u. a.] 1998.

[23] Borzi, Alfio: *Numerics for the optimal control of models with differential equations* [Elektronisches Skript].

[24] Chernous'ko, Felix L.; Ananievski, I. M. & Reshmin S. A. *Control of Nonlinear Dynamical Systems. Methods and Applications.* Springer: Berlin/Heidelberg 2008.

[25] Evans, Lawrence C.: *An Introduction to Mathematical Optimal Control Theory Version 0.2.*[Elektronisches Skript]. Erhalten über https://math.berkeley.edu/~evans/control.course.pdf.

[26] Dmitruk, Andrei & Osmolovskii, Nikolai: *On the proof of Pontryagin's maximum principle by means of needle variations.* [Elektronisches Paper.] In: Arxiv e-prints, `arXiv:1412.2363v2 [math.OC]`, Dezember 2014.

[27] López, Eric D.; Molgado, Alberto & Vallejo, José A.: Nikolai: *The principle of stationary action in the calculus of variations.* [Elektronisches Paper.] In: Arxiv e-prints, `arXiv:1205.0865v2 [math-ph]`, Mai 2012.

[28] Graichen, Knut, Käpernick, Bartosz & Zheng, Tao (ed.): *A Real-Time Gradient Method for Nonlinear Model Predictive Control.* [Elektronische Version.] In: *Frontiers of Model Predictive Control.*, IntechOpen, DOI: 10.5772/37638, Februar 2012. Erhalten über https://www.intechopen.com/books/frontiers-of-model-predictive-control/a-real-time-gradient-method-for-nonlinear-model-predictive-control.

[29] Kirsch, Andreas: *An Introduction to the Mathematical Theory of Inverse Problems.* [Elektronische Version.] Springer: New York [u. a.] 2011.

[30] Vogel, Curtis R. *Computational Methods for Inverse Problems.* Society for Industrial and Applied Mathematics: Philadelphia 2002.

Kapitel 5-6

[31] Higham, Desmond J.: *An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations.* In: SIAM Review, Band 43 (3), 2013, S. 525 – 546.

[32] Annunziato, Mario & Borzi, Alfio: *Optimal Control of a Class of Piecewise Deterministic Processes.* In: European Journal of Applied Mathematics, Band 25(1), S. 1 - 25.

[33] Behrends, Ehrhard: *Markovprozesse und stochastische Differentialgleichungen. Vom Zufallsspaziergang zur Black-Scholes-Formels.* [Elektronische Version.] Springer: Wiesbaden 2013.

[34] Davis, M.: *Piecewise-Deterministic Markov Processes. A General Class of Non-Diffusion Stochastic Models.* In: Journal of the Royal Statistical Society, Band 46 (3), S. 353-388. Erhalten über http://www.jstor.org/stable/2345677.

Bildquellen:
Logo auf der Titelseite: https://upload.wikimedia.org/wikipedia/commons/4/4e/Universit%C3%A4t_W%C3%BCrzburg_Logo.svg
Alle anderen Bilder wurden mit Geogebra, MATLAB oder Mathematica erstellt.